Pattern Recognition 129 (2022) 108753

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Orthogonal channel attention-based multi-task learning for multi-view facial expression recognition☆

Jingying Chen[a,b], Lei Yang[a], Lei Tan[c], Ruyi Xu[b,*]

[a] National Engineering Research Center for E-learning, Central China Normal University, 430079, China
[b] National Engineering Laboratory For Educational Big Data, Central China Normal University, 430079, China
[c] Department of Artificial Intelligence, School of Informatics, Xiamen University, 361005, China

ARTICLE INFO

Article history:
Received 2 March 2021
Revised 28 March 2022
Accepted 26 April 2022
Available online 27 April 2022

Keywords:
Multi-view facial expression recognition
Orthogonal channel attention
Multi-task learning
Siamese convolutional neural network
Separated channel attention module

ABSTRACT

Multi-view facial expression recognition (FER) is a challenging computer vision task due to the large intra-class difference caused by viewpoint variations. This paper presents a novel orthogonal channel attention-based multi-task learning (OCA-MTL) approach for FER. The proposed OCA-MTL approach adopts a Siamese convolutional neural network (CNN) to force the multi-view expression recognition model to learn the same features as the frontal expression recognition model. To further enhance the recognition accuracy of non-frontal expression, the multi-view expression model adopts a multi-task learning framework that regards head pose estimation (HPE) as an auxiliary task. A separated channel attention (SCA) module is embedded in the multi-task learning framework to generate individual attention for FER and HPE. Furthermore, orthogonal channel attention loss is presented to force the model to employ different feature channels to represent the facial expression and head pose, thereby decoupling them. The proposed approach is performed on two public facial expression datasets to evaluate its effectiveness and achieves an average recognition accuracy rate of 88.41% under 13 viewpoints on Multi-PIE and 89.04% under 5 viewpoints on KDEF, outperforming state-of-the-art methods.

1. Introduction

Facial expression, as one of most important modes of nonverbal communication, conveys ones inner emotions to others. People from different regions, cultures, and races show similar facial expressions when they express several basic emotions: anger, disgust, fear, happiness, sadness, and surprise. Psychological studies have found that humans have the ability to decipher the meanings of various facial expressions and classify a facial expression as one of the basic emotions [1]. In exploring ways to help computers gain a human-like understanding of various facial expressions, automatic facial expression recognition (FER) is a popular research topic in the computer vision field [2].

In the past two decades, FER has progressed significantly, especially with the rise of deep learning [3]. Compared to traditional machine learning approaches, which handle small-scale handcraft features, deep learning approaches employ large-scale cascaded networks with multiple layers or modules similar in structure and function to discover potentially valuable knowledge in millions of raw training data. Various deep learning models have been developed and successfully applied to FER, such as convolutional neural networks (CNN). CNN is a feedforward neural network in which the front end adopts multiple stacked convolutional layers for deep embedding extraction. The deeper embedding contains higher-level visual semantic information, which has a more powerful representation ability for FER.

However, most existing approaches focus primarily on frontal or near-frontal FER, which does not work properly when applied to non-frontal FER. When the view angle changes in a large range, the intra-class distance of facial expressions increases accordingly. Improved recognition accuracy is achieved when combining multiple single viewpoint expression models for multi-view FER [4], but recognition accuracy may decline significantly when a unified model is trained to deal with expressions from all viewpoints. Furthermore, the recognition accuracy of frontal expressions or near-frontal expressions is always higher than that of non-frontal expressions, primarily because, as the viewpoint gradually turns to one side of the face, the region of self-occlusion increases, and discriminative information is lost when the expression is projected from three-dimensional space onto a two-dimensional image plane.

☆ This document is the results of the research project funded by the National Natural Science Foundation of China (No. 61977027) and the Hubei Province Technological Innovation Major Project (No. 2019AAA044).

* Corresponding author.
E-mail address: xuruyi@mail.ccnu.edu.cn (R. Xu).

https://doi.org/10.1016/j.patcog.2022.108753
0031-3203/

Recently, the Siamese network is presented to address the similarity metric learning issue using two identical neural networks that share the same set of weights [5]. For multi-view FER, the Siamese network can be used to learn the similarity of the visual semantic information between the frontal viewpoint and the non-frontal viewpoint to ensure the discriminative ability of deep embedding [6]. However, existing methods based on the Siamese network ignore the effect of head pose information on multi-view FER.

To this end, this paper presents a novel multi-view FER approach called orthogonal channel attention-based multi-task learning (OCA-MTL). The proposed approach adopts a Siamese CNN with two parallel paths. One path is used to learn the expression features from the frontal viewpoint, and the other is a multi-task CNN that learns both the expression features and head pose features from various viewpoints. To enhance the recognition accuracy of non-frontal expressions, the l_2-norm of difference of expression features extracted from two paths is minimized. Moreover, in the multi-task learning path, a separated channel attention module is embedded in the last convolutional layer to learn subtask-specific channel attention. Meanwhile, orthogonal channel attention loss is presented to force the model to employ different feature channels to represent the facial expression and head pose. To verify effectiveness, the proposed approach is performed on two public facial expression databases: the CMU Multi-PIE face (Multi-PIE) database and the Karolinska Directed Emotional Faces (KDEF) database. The contributions of this paper are summarized as follows:

1. **View-independent facial-expression features learning framework.** The proposed framework uses two parallel paths to learn view-independent features from various viewpoints and the corresponding frontal viewpoints, forcing the multi-view expression model to follow the frontal one and learn more discriminative features.
2. **Separated channel attention module.** The proposed module is embedded in the multi-task learning path to improve the performance of two subtasks by introducing a subtask-specific attention mechanism while leveraging the inherent synergy between subtasks within general multi-task learning.
3. **Orthogonal channel attention loss.** The proposed loss supervises the model in learning orthogonal channel attention, which makes two subtasks select different feature channels to represent the facial expression and head pose. Compared to traditional self-attention mechanisms, an orthogonal attention mechanism can further decouple head pose features and facial expression features effectively.

The remainder of the paper is organized as follows. In Section 2, related works on multi-view FER are reviewed. In Section 3, the proposed approach is described in detail. In Section 4, the experimental results are reported and analyzed. Finally, Section 5 summarizes the key findings of this study.

2. Related work

In this section, some related works on multi-view FER, multi-task learning (MTL), and attention mechanisms are introduced and briefly explained.

2.1. Multi-view FER

The existing multi-view FER approaches can be divided into two categories: traditional and deep learning. Traditional approaches focus on the manual design of discriminative features to recognize facial expressions under extreme head pose variations. For example, Moore and Bowden investigated the performance of local binary patterns (LBP) and variations of LBP for multi-view FER [7]. Güney et al. used Gabor features within blocks around the left eye, right eye, and mouth to represent facial expressions [8]. Wu et al. developed the locality-constrained linear coding-based bi-layer model (LLCBL), which extracts dense SIFT features from local patches and constructs an overall bag-of-features model using locality-constrained linear coding [9]. Then, the proposed approach estimates the head pose in the first layer and recognizes the facial expression using a view-dependent model combined with the overall features in the second layer.

Unlike traditional approaches, deep learning approaches learn expression features and classify facial expressions in an end-to-end manner. Hence, feature extraction is completely adaptive to the issue to be addressed without manual intervention. Baddar and Ro investigated the deficiency of long short-term memory (LSTM) in spatio-temporal feature encoding and presented an improved LSTM to encode spatio-temporal features robust to head pose variations [10]. Experiments conducted on a self-collected dataset showed that their approach sustained a high recognition rate over all poses. To compensate for missing features caused by non-frontal views, some works have proposed restoring them by GAN. Lai et al. used GAN to learn emotion-preserving representation in the face frontalization framework, which achieved an average recognition accuracy of 86.76% for six basic expressions and 13 different viewpoints on the Multi-PIE dataset [11]. When the frontal view was available, several works adopted Siamese CNN to learn view-independent features from both the frontal view and non-frontal view simultaneously. Baddar et al. conducted the first study that adopted a Siamese CNN to improve the robustness of features to image variations [5]. On this basis, Luo et al. embedded spatial transformer networks in a Siamese CNN to determine whether using features from both networks makes them more robust to illumination and viewpoint variations and achieved a recognition accuracy rate of 88.13% at 20 illumination levels (0 to 19) and 5 viewpoints (0°, $\pm 15^\circ$, $\pm 30^\circ$) on the Multi-PIE dataset [6]. However, these works ignore the effect of head pose estimation (HPE) in multi-view FER.

2.2. Multi-task learning

MTL involves learning multiple related subtasks simultaneously to explore the synergy among them and improve the individual subtasks generalization performance. Due to its powerful ability to address the problem of insufficient labeled training data, it is widely used to learn deep models. In face-related computer vision tasks, numerous MTL approaches have been developed for face detection, face alignment, and facial attribute classification because the human face contains abundant and diverse information, such as race, age, gender, identity, expression, and pose. For example, Ranjan et al. presented HyperFace, a deep MTL framework for face detection, landmark localization, pose estimation, and gender recognition [12]. In this study, they explored the roles of features in the lowest and deepest layers of the CNN and combined multiple intermediate layer features to improve the performance of MTL. Mao et al. improved facial attribute classification performance with auxiliary facial landmark detection and constructed multiple deep neutral networks by considering the different learning complexities of facial attributes [13]. Chen et al. proposed a residual learning module for MLT to learn the complementary information from the task of facial landmark location to enhance the performance of FER [14].

In view of the advantages of MTL, many multi-view FER methods also adopt the MTL framework, which uses HPE as an auxiliary task to improve expression recognition performance. However, for multi-view FER tasks, the synergy between FER and HPE is utilized to improve shared feature representation, and the dependency between them should be decoupled to enhance the fa-

cial expression features robustness to head pose variations. To this end, Gan et al. incorporated subspace learning into the MTL framework for head pose-insensitive smile detection [15]. Their work regarded the head pose and facial expression as latent variables of observed features and assumed that feature subspaces to which these two variables belonged satisfied orthogonality. Furthermore, they verified the effectiveness of decoupling facial expression and head pose features to improve smile detection. Li et al. assumed that a facial expression image could be divided into four components: an expressive component, an identity component, a head pose component, and a remaining component [16]. They trained the encoder to map the input image to latent space to represent the four components and only decoded the facial expression component within a GAN framework. Their approach achieved an average accuracy rate of 86.9% for six basic facial expressions in head pose variations ranging from 0° to 90°, which showed that it can effectively eliminate the influence of head pose variations. Similarly, Zhang et al. added a geometry-embedding network to the GAN framework, which was used to generate facial images with different expressions and poses in a continuous manner, guided by a set of landmarks [17]. The identity representation was explicitly disentangled from both expression and pose variations through the shape geometry delivered by facial landmarks. Inspired by the previous work, this paper presents a simple yet effective method to disentangle the facial expression and head pose based on an attention mechanism.

2.3. Attention mechanism

In machine learning, an attention mechanism is used to simulate human vision and invest more attention resources in the most informative components of a signal to obtain more detailed information while suppressing useless information. Due to its powerful ability to improve the efficiency of signal processing and the accuracy of signal recognition, it is widely used in image captioning, target location and tracking, and sequence learning.

For FER, a considerable amount of useful information may exist in action unit (AU)-specific local regions. Hence, most previous research on this topic has involved utilizing a spatial attention mechanism to focus on these local regions and improve the performance of FER [18]. Sun et al. embedded spatial self-attention in an 11-layer CNN for FER [19]. Their visualization analysis showed that the approach effectively detected the region of interest (ROI), partly consistent with the emotion-specific AU, even under large head pose variations. Huang et al. developed a region attention network that combined self-attention and relation-attention modules to weight the importance of cropped facial parts while considering the robustness of local patches to pose variations and occlusion [20]. Spatial attention has also been used to segment the facial region from the background to resist the inference of clutter in the environment. Wang et al. developed OAENet, which involved constructing an oriented attention module with an encoder-decoder-style network to highlight global and local facial information and improve FER's classification performance [21].

Channel attention mechanisms have also been used to improve feature representation. Hu et al. described the squeeze-and-excitation network (SE-Net), a channel-wise attention mechanism that uses global information to selectively emphasize informative features and suppress less useful ones [22]. Gan et al. embedded the SE-Net in a multiple attention network for FER and validated the effectiveness of the proposed method in both real-world and controlled environments [23]. In this paper, we attempt to adapt the SE-Net for a multi-task learning framework and multi-view FER.

3. Methodology

In this section, the proposed OCA-MTL approach for multi-view FER is introduced and described in detail. The approach takes advantage of multi-task learning and an attention mechanism.

3.1. Overview

The proposed OCA-MTL approach adopts a multi-task learning framework based on Siamese CNN. An overview of the proposed framework is shown in Fig. 1. The input training data can be described as a four-tuple $T = \left(X_v, X_f, Y_e, Y_h\right) \in \mathcal{T}$, where $\mathcal{T}$ represents the training data set, X_v is a facial expression image with various viewpoints, X_f is the corresponding frontal face image with the same facial expression, Y_e is the facial expression label of X_v or X_f, and Y_h is the head pose label of X_v.

Then, a Siamese CNN is constructed to learn the multi-view FER model from the training dataset $\mathcal{T} = \{T_n\}_{n=1}^{N}$, where N is the total number of training data points. The Siamese CNN has two parallel paths to handle X_v and X_f, respectively. In this study, the backbone networks in two paths with shared parameters follow the setting of AlexNet [24], including five stacked convolutional layers. The path for X_v adopts a multi-task CNN that learns the facial expression and head pose simultaneously A separated channel attention module is proposed to replace the last convolutional layer of the original CNN. The separated channel attention module extracts the facial expression features F_e and head pose features F_h. Next, two subtask-specific classifiers using three fully connected layers are constructed to classify the F_e and F_h, respectively. The prediction loss of this multi-task learning path $\mathcal{L}_m$ can be formulated as follows:

$$\mathcal{L}_m = \mathcal{L}_e + \lambda_h \mathcal{L}_h \tag{1}$$

where $\mathcal{L}_e$ and $\mathcal{L}_h$ are the cross-entropy losses of FER and HPE, respectively. Generally, for the optimization of multi-task learning, it is necessary to weigh the importance of each task to select the appropriate weight to achieve the overall optimization of all tasks. Specific to our task, the FER task should receive more attention, and HPE should be an auxiliary task to enhance the performance of FER. An investigation of the empirical parameter on λ_h will be reported in the next section. The cross-entropy loss is the negative log-likelihood of the ground truth labels given an input sample, which can be represented as follows:

$$\mathcal{L}_t = \frac{1}{N}\sum_{n=1}^{N}\left(- < Y_t, \log\left(\widehat{Y}_t\right) >\right), t \in \{e, h\} \tag{2}$$

where, $\langle \cdot, \cdot \rangle$ represents the operation of the inner product, Y_t is a one-hot label of ground truth, and $\widehat{Y}_t$ is the probability distribution predicted by the multi-task learning CNN.

The path for the frontal face image X_f adopts the original CNN, which extracts facial expression features using the backbone network of AlexNet and a followed fully connected layer. The features extracted from the frontal face image might be more discriminative due to less related information missed. Hence, it is a good reference to guide the model to learn good facial expression features from other views. By minimizing the distance between the frontal expression features and the non-frontal expression features, the model is forced to learn the viewpoint-invariant features with the strongest discriminative ability. To this end, denote the expression features extracted from the frontal face as $F(X_f)$ and the expression features extracted from the non-frontal face as $F(X_v)$. l_2-norm of them is minimized leading a supervised training of parameters in two parallel CNNs:

$$\mathcal{L}_{diff} = \frac{1}{2N}\sum_{n=1}^{N}\left\|F(X_f) - F(X_v)\right\|_2^2 \tag{3}$$

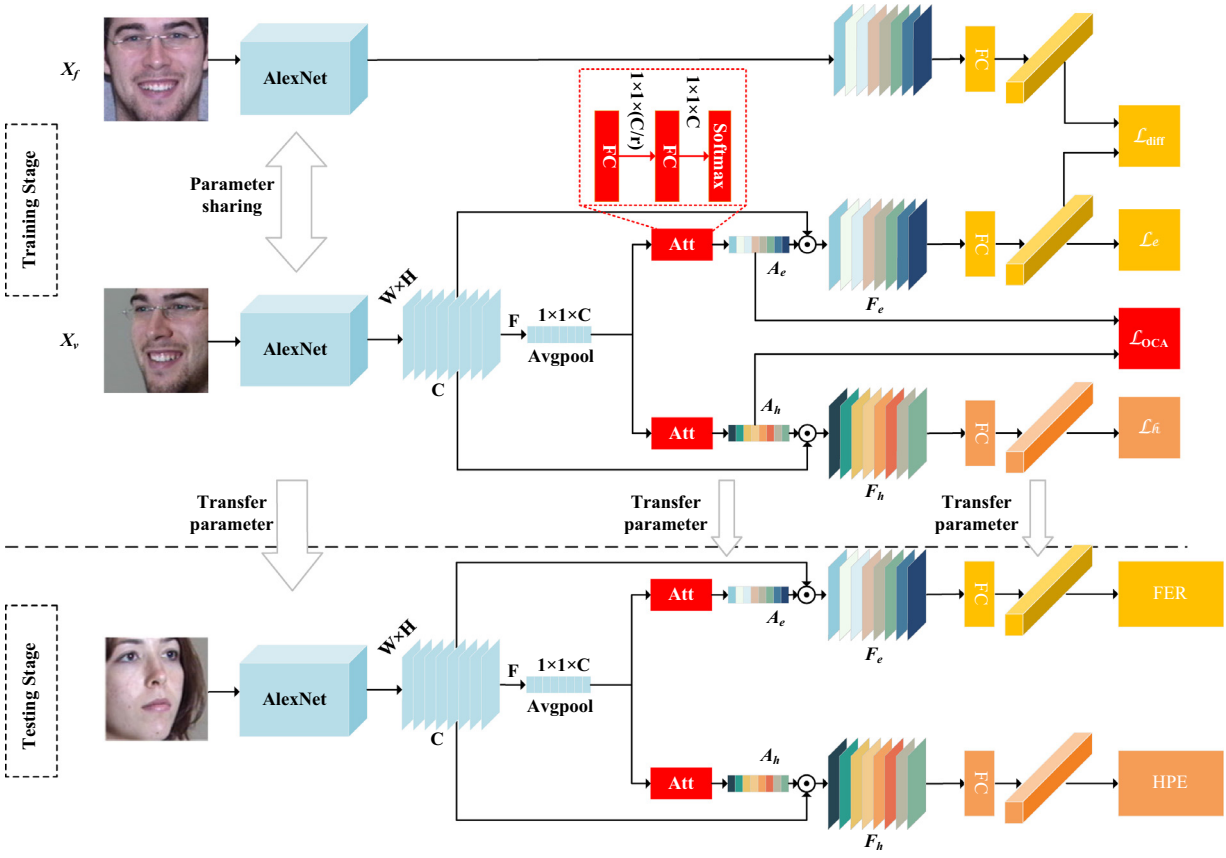


Fig. 1. Overview of the proposed OCA-MTL. In the training stage, pairwise data comprised of a frontal view and non-frontal view of the same expression images is fed into the Siamese network. In the testing stage, the test data are fed to a model whose parameters are transferred from the path for the non-frontal view. FC represents the fully connected layer; Avgpool represents the global average pooling; Att represents the attention module; L_e and L_h are the cross-entropy loss for FER and HPE tasks, respectively; L_{diff} is the Euclidean distance loss of expression features between two paths; and L_{OCA} is the orthogonal channel attention loss.

where $\|\cdot\|_2$ represents the operation of l_2-norm. To reduce the effects caused by scale variability of the training data, $F(X_f)$ and $F(X_v)$ are l_2-norm normalized before the difference is calculated.

3.2. Separated channel attention module

Generally, a deep learning model extracts low-level visual semantic information in the shallow layer, such as edge, texture, and other features, while extracting high-level visual semantic information in the deeper layer, such as expression muscle movement, head rotation, and so on. In the spirit of data augmentation, the proposed model uses a shared backbone to learn two subtasks that aim to utilize more label information to learn better universal features. Hence, such universal low-level visual semantic features are beneficial for both subtasks. However, learning the shared features at a deeper level does not improve the performance of each subtask. To learn the subtask-specific features in the deep layer, a separated channel attention (SCA) module is used to replace the last convolutional layer.

Inspired by SE-Net, the proposed SCA module uses global information to selectively emphasize informative features and suppress less useful features. Unlike SE-Net, SCA is a module with a single input signal and dual output signals. Through SCA, shared lower-level features are mapped into two higher-level features for two subtasks. The SCA module can generate independent feature channel attentions for different subtasks, reduce the performance degradation caused by excessive sharing between subtasks, and enhance the features obtained by the two subtasks.

Denote the shared features $F_{share} \in \mathbb{R}^{W\times H\times C}$ as the input of the SCA module, where W, H and C respectively, represent the features width, height, and channel number. The SCA module contains a global average pooling layer (Avgpool) and two branches, each consisting of two fully connected layers. Avgpool calculates the average value of each feature channel in F_{share} and connects them to a vector $F_{avg} \in \mathbb{R}^{1\times 1\times C}$. Denote the parameters of these two fully connected layers as $W_t^1 \in \mathbb{R}^{C\times \frac{C}{r}}$ and $W_t^2 \in \mathbb{R}^{\frac{C}{r}\times C}$, where r is the reduction ratio; $t \in \{e, h\}$ represents the branch for the FER or HPE subtask. Then, the first fully connected layer with a ReLU activation function is used to reduce the dimension of features to $1 \times 1 \times \frac{C}{r}$. Subsequently, the second fully connected layer with a SoftMax activation function is used to recover the dimension of features to $1 \times 1 \times C$. Formally, the separated channel attention can be formulated as

$$A_t = \sigma\left(W_t^2\delta\left(W_t^1 Avgpool\left(F_{share}\right)\right)\right), t \in \{e, h\} \tag{4}$$

where $\delta(\cdot)$ denotes the ReLU activation function, and $\sigma(\cdot)$ denotes the SoftMax activation function. Compared to the sigmoid activation function used in SE-Net, the attention weight output by the SoftMax activation function forms a probability distribution that reflects the importance across channels. The SCA module contains two outputs: F_e and F_h. The final outputs can be formulated as

$$F_t^i = F_{share}^i \times A_t^i, i = \{1, \ldots, C\}, t \in \{e, h\} \tag{5}$$

where F_t^i is the ith channel of F_t, F_{share}^i is the ith channel of F_{share}, and A_t^i is a scale, which represents the weight of the ith channel.

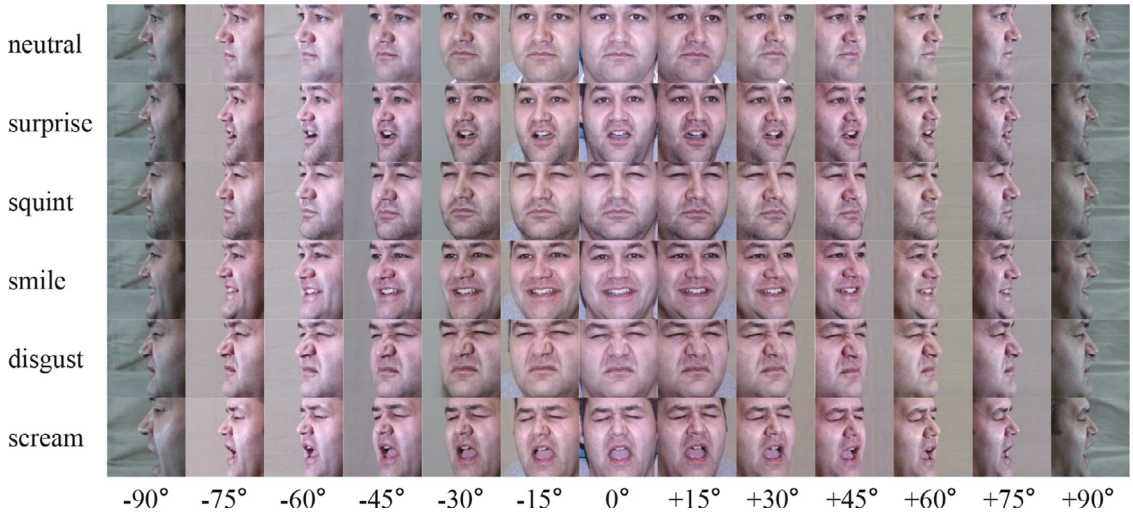


Fig. 2. Exemplars of six facial expressions and thirteen facial views from Multi-PIE Set I. Each row has the same facial expression label, and each column has the same pan angle of head pose.

3.3. Orthogonal channel attention loss

Although the SCA module learns subtask-specified features well, it does not decouple the dependency between two subtasks, which may result in large intra-class variations to limit the performance of multi-view FER. To address this issue, the weights of the channel attention of two subtasks are constrained to be orthogonal, which can be formulated as

$$\mathcal{L}_{\text{OCA}} = \cos\left(A_e, A_h\right) = \frac{\langle A_e, A_h \rangle}{|A_e||A_h|} = A_e A_h^T \tag{6}$$

where $|\cdot|$ represents the modulus of the vector. The sum of attention weights is 1 through SoftMax activation. Hence, the orthogonal channel attention loss is equal to the inner product of separate attention weights of two subtasks.

By minimizing the $\mathcal{L}_{OCA}$, if a channel weight A_e^i in the expression attention is assigned to a non-zero value, then the corresponding channel weight A_h^i in head pose attention should be zero or near-zero value, vice versa. Hence, two subtasks will select mutually exclusive feature channels from shared features to represent the facial expression and head pose, respectively.

3.4. Optimization of the model

To obtain our multi-view FER model, the training process is carried out by optimizing the total loss of the model, which can be formulated as

$$\mathcal{L}_{total} = \frac{1}{2}\|w\|^2 + \lambda_1 \mathcal{L}_m + \lambda_2 \mathcal{L}_{diff} + \lambda_3 \mathcal{L}_{\text{OCA}} \tag{7}$$

where the first term on the right side is a regularization term, and λ_1, λ_2, and λ_3 are trade-off parameters to balance $\mathcal{L}_m$, $\mathcal{L}_{diff}$, and $\mathcal{L}_{\text{OCA}}$.

$\mathcal{L}_m$ is the main loss term for supervised learning to optimize the weights in a whole multi-task learning path. $\mathcal{L}_{diff}$ and $\mathcal{L}_{\text{OCA}}$ can be regarded as two regularization loss terms to optimize the extraction of deep embedding. $\mathcal{L}_{diff}$ aims to optimize the shared weights of two paths to extract the viewpoint-independent facial expression features depending on good frontal facial expression features, while $\mathcal{L}_{\text{OCA}}$ aims to optimize the attention module to decouple the facial expression features from various viewpoints. Hence, there would be collaborative and complementary effect between the two loss items in terms of optimization objectives. However, it is still necessary to set the empirical parameters carefully because they jointly affect the optimization of shared parts in a Siamese network. To illuminate the influence of empirical parameter selection, an investigation of parameter on λ_3 will be reported in the next section. After training, the test data are fed to a model whose parameters are transferred from the path for non-frontal view to predict the facial expression and the head pose.

4. Experiments

In this section, the results achieved after conducting experiments with two public multi-view facial expression datasets (i.e., Multi-PIE and KDEF) are reported. To make a fair comparison, we begin by explaining how to use these two public databases and the concrete settings of the experimental protocol. Then, under the same settings, the proposed method is compared with other state-of-the-art methods. Finally, the effectiveness of the proposed method is demonstrated further by visualized analysis and an ablation study.

4.1. Databases

4.1.1. Multi-PIE database

The Multi-PIE database contains more than 750,000 images of six facial expressions (neutral, scream, smile, surprise, squint, and disgust) recorded from 337 different subjects under 15 view points and 20 illumination conditions in four recording sessions. To make a broad comparison, two datasets (Set I and Set II) are constructed from the original Multi-PIE database.

Multi-PIE Set I is constructed following [25], which selects 100 subjects appearing in all four recordings. It contains 7800 facial expression images under 13 different pan angles (0°, ±15°, ±30,°, ±45°, ±60°, ±75°, ±90°) and the same level of illumination (level 7). Exemplars of one subject with six facial expressions and thirteen viewpoints are shown in Fig. 2.

Multi-PIE Set II consists of 60,000 facial expression images of the same 100 subjects in Multi-PIE Set I under five different pan angles (0°, ±15°, ±30°) and 20 different levels of illumination (0 to 19). Exemplars under 20 different illumination variations from Multi-PIE Set II are shown in Fig. 3.

4.1.2. KDEF Database

The KDEF database contains a set of 4900 images of human facial expressions made by 70 amateur actors (35 females and 35 males) displaying seven different emotional expressions (neutral,

Fig. 3. Exemplars under 20 different illumination variations from Multi-PIE Set II. The first row from left to right is illumination levels 0 to 9, and the second row from left to right is illumination levels 10 to 19.

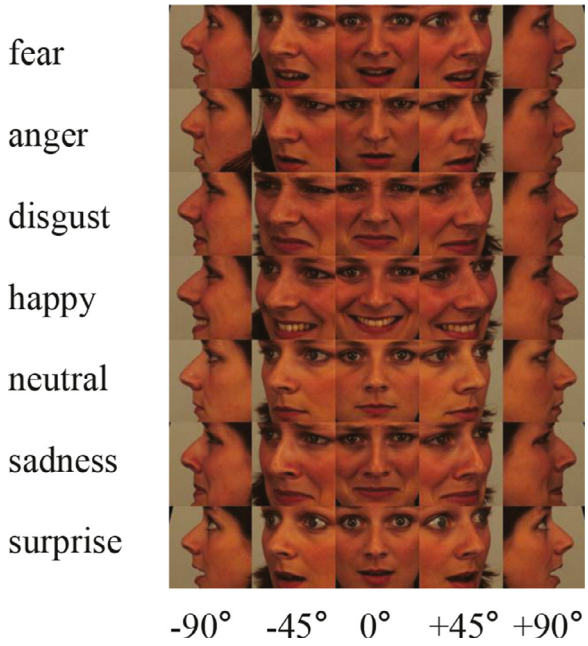


Fig. 4. Exemplars of seven emotional expressions and five head poses from the KDEF database. Each row has the same facial expression label, and each column has the same pan angle of head pose.

happy, anger, fear, disgust, sadness, and surprise). All subjects were asked to rehearse these seven expressions for 1 h, and each expression was photographed twice from five different pan angles (0°, ±45°, ±90°). Exemplars of one amateur actor are shown in Fig. 4. It is noteworthy that 10 images of human faces are missing, and all of them occur in the non-frontal view. Each of the missing images is replaced by a symmetrical view of the image with a mirror flip. Thus, the KDEF dataset used for our experiments consists of a total of 4900 images.

4.2. Protocol settings

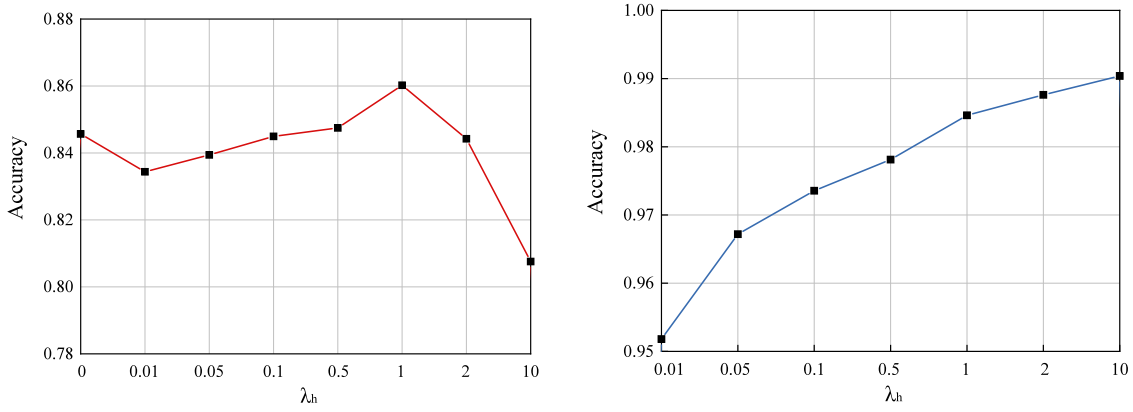
Before inputting the model, pre-processing is applied to all images in the three datasets. First, a multi-task-learning face detector (MTCNN) [26] is utilized to locate the facial region and detect facial landmarks. Subsequently, the face is aligned and resized to a fixed size by a similarity transform that specifies the center of the eyes and mouth in target images, as in [27]. To be consistent with the settings of most methods, images in Multi-PIE Set I and the KDEF dataset are resized to 128 × 128, while the images in Multi-PIE Set II are resized to 64 × 64. Finally, all the images are normalized to have a zero mean and unit variance. We compare different methods with 5-fold subject independent cross-validation without any subjects across the training and testing sets, as in [9,25,28]. Each model is implemented with the TensorFlow framework. The reduction ratio in the SCA module is set to 8. For all experiments, the batch size is set to 32. To keep the optimization algorithm robust and efficient, an Adam optimizer is utilized with the decayed learning rate at the initial learning rate of 1e-4, following a staircase function and decaying every 10 steps with a base of 0.9, and the dropout ratio is set at 0.5. The experiments were conducted on a PC with Intel (R) Core(TM) i7-8700K CPU at 3.70GHz and 64GB memory, and NVIDA GeForce GTX 1080Ti.

4.3. Investigation of parameters

To illuminate the influence of empirical parameters on the performance of the proposed method, two experiments for parameters investigation are conducted on Multi-PIE Set I.

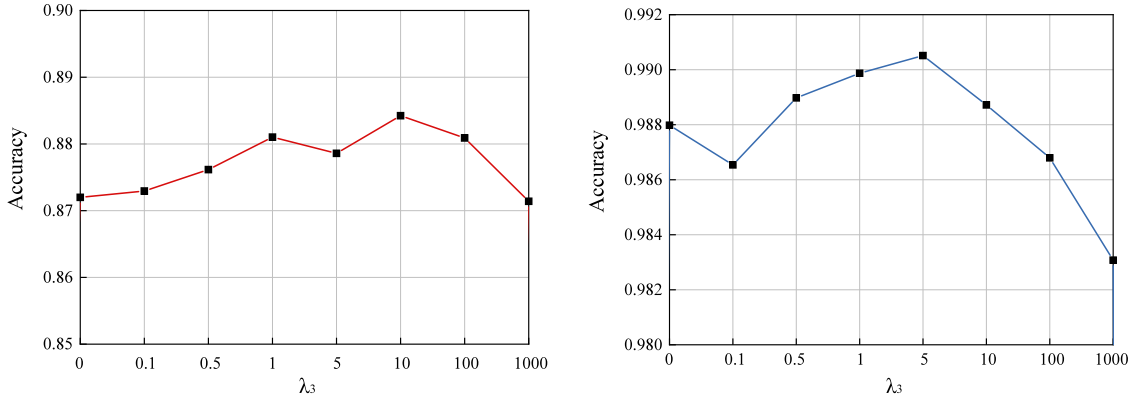
The first experiment trains a multi-task learning network with an objective function defined in Eq. (1). A set of models is obtained by respectively setting λ_h to the value in a list of [0, 0.01, 0.05, 0.1, 0.5, 1, 2, 10]. The performance of each model is evaluated using the average accuracy of FER and the average accuracy of HPE for 5-fold cross-validation. The results are shown in Fig. 5. When $\lambda_h = 0$, it is equivalent to only learning the FER task without HPE task. With the increase of λ_h, the average accuracy of head pose estimation is gradually improved. When $0.1 < \lambda_h < 2$, the accuracy of FER in multi-task learning exceeds the result in single-task learning, indicating that the HPE task plays a role in enhancing the FER task. When λ_h increases to 2, the accuracy of FER starts to a significant decline, showing that if the objective function pays too much attention to the HPE task, the performance of FER will be degraded. Hence, λ_h is set to 1 in all experiments, which may achieve the best FER performance.

The second experiment trains the proposed network with an objective function defined in equation (7). The main aim of the experiment is to investigate how the parameter λ_3 balances the two regularization loss terms for deep embedding extraction. According to the previous parameter investigation, λ_h is set to 1. Meanwhile, both the λ_1 and λ_2 are set to 1 following the work in [5]. A set of models is obtained by respectively setting λ_3 to the value in a list of [0, 0.1, 0.5, 1, 5, 10, 100, 1000]. When $\lambda_3 = 0$, it is equivalent to only learning the multi-task network with SCA module but without OCA loss constraint. Additionally, the performance of each model is evaluated using the average accuracy of FER and the average accuracy of HPE for 5-fold cross-validation. The results are shown in Fig. 6. It can be seen that OCA loss can enhance the performance of FER in a range of [0.1 100], and OCA loss can enhance the performance of HPE in a range of [0.5 10]. This indicates that the optimization of OCA with the appropriate empirical parameter setting improves both subtasks. The best accuracy of 88.42% for FER is obtained at $\lambda_3 = 10$, while the best accuracy of 99.03% for HPE is obtained at $\lambda_3 = 5$. Since the result of FER is more important, λ_3 is finally set to 10.



(a) facial expression recognition (b) head pose estimation

Fig. 5. Investigation of parameter λ_h the Multi-PIE Set I.



(a) facial expression recognition (b) head pose estimation

Fig. 6. Investigation of parameter λ_3 on the Multi-PIE Set I.

Table 1
Recognition accuracies under various head poses on the Multi-PIE Set I.

Acc.(%)	0°	±15°	±30°	±45°	±60°	±75°	±90°	Avg.
Neutral	91.00	90.00	89.00	91.00	89.50	85.50	76.50	87.23
Scream	98.00	98.00	99.00	98.50	98.00	98.00	92.50	97.38
Smile	98.00	96.50	96.50	96.50	96.50	93.00	89.00	94.92
Surprise	98.00	99.50	99.00	98.00	96.50	94.50	88.00	96.08
Squint	80.00	78.50	79.50	80.50	75.50	74.50	75.00	77.46
Disgust	73.00	76.00	77.00	76.50	81.00	75.50	80.50	77.38
Overall	89.67	89.75	90.00	90.17	89.50	86.83	83.58	88.41

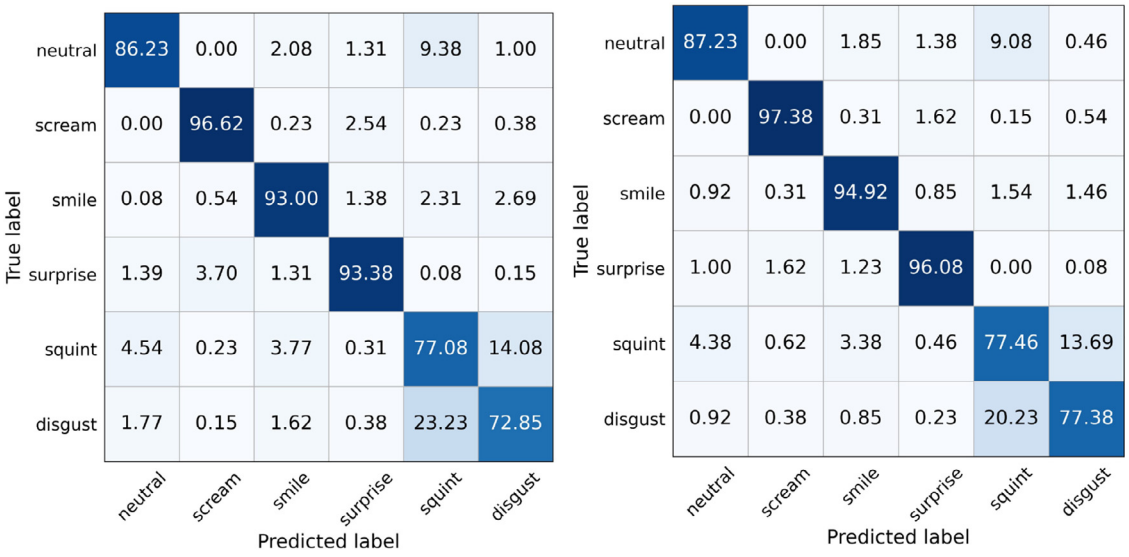
4.4. Comparison to a state-of-the-art method

4.4.1. Experiments on multi-PIE set I

To verify the robustness of our method to viewpoint variations, an experiment using OCA-MTL is conducted on Multi-PIE Set I. Table 1 shows the recognition accuracy of each expression under different viewpoints using OCA-MTL. Each expression achieves the highest recognition accuracy under different pan angles. Regarding the overall recognition accuracy rate, there are no significant differences among pan angles in the range of -60° to $+60^\circ$.

Figure 7 shows the confusion matrix of FER under all the head poses with baseline method and the proposed method. The baseline method is a Siamese CNN, of which the backbone networks adopt two identical AlexNets without multi-task learning and OCA loss. The proposed OCA-MTL approach adopts a multi-task learning framework based on the Siamese CNN, embedded with separated channel attention module and optimized by orthogonal channel attention loss, additionally. Compared with the result of facial expression recognition corresponding to the baseline method, OCA-MTL achieves a higher accuracy of each facial expression, which is an obvious advance on the existing baseline method. Both Table 1 and Fig. 7(b) show that relatively high predictive accuracy is achieved for scream, smile, and surprise expressions (97.38%, 94.92%, and 96.08%, respectively), considering that the overall average recognition accuracy rate is 88.41%. However, squint and disgust are more difficult to recognize, as evidenced by recognition rates below 80% for these expressions. Moreover, disgust and squint expressions are more likely to be confused with each other, which is followed by misclassification between neutral and squint since 9.08% of neutral expressions are recognized as squint expressions. Compared with the baseline method, OCA-MTL can effectively prevent these confusions.

Although most of the previous works used seven poses to conduct the experiments, we use 13 bilateral viewpoints instead and calculate the FER accuracy of our model in this experiment. The recognition accuracy of our method is higher than that of other



(a) Confusion matrix with the baseline method (b) Confusion matrix with the proposed OCA-MTL

Fig. 7. Confusion matrix of FER under all the head poses on the Multi-PIE Set I with baseline method and the proposed method.

Table 2
Performance comparison of state-of-the-art methods on Multi-PIE Set I.

Method	Viewpoints		Expressions number	Illumination number	Overall (%)
	number	pan			
LBPms [7]	7	(0°, +90°)	6	1	73.3
LGBP [7]	7	(0°, +90°)	6	1	80.4
GSRRR [29]	7	(0°, +90°)	6	1	81.7
2D JFDNN [30]	7	(0°, +90°)	6	1	82.9
Single layer LLC [9]	7	(0°, +90°)	6	1	84.4
LLCBL method [9]	7	(0°, +90°)	6	1	86.3
KPSNM [31]	7	(-90°, +90°)	6	1	83.1
KPSNM [31]	13	(-90°, +90°)	6	1	82.6
EPRL(GAN) [11]	13	(-90°, +90°)	6	1	86.7
TP-GAN [28]	13	(-90°, +90°)	6	1	87.0
OCA-MTL	13	(-90°, +90°)	6	1	**88.4**

Table 3
Recognition accuray rates under various illuminations on the Multi-PIE Set II.

Acc.(%)	-30°	-15°	0°	+15°	+30°	Avg.
Neutral	93.10	94.75	94.60	93.85	92.90	93.84
Scream	98.45	98.45	98.75	98.65	98.50	98.56
Smile	95.20	96.35	95.70	95.60	94.45	95.46
Surprise	95.65	95.60	95.95	96.10	96.25	95.91
Squint	77.15	78.70	78.50	79.25	76.15	77.95
Disgust	77.15	75.70	75.70	76.25	72.70	75.50
Overall	89.45	89.93	89.87	89.95	88.49	89.54

methods, demonstrating that OCA-MTL can more effectively address the FER task under viewpoint variations (Table 2).

4.4.2. Experiments on multi-PIE set II

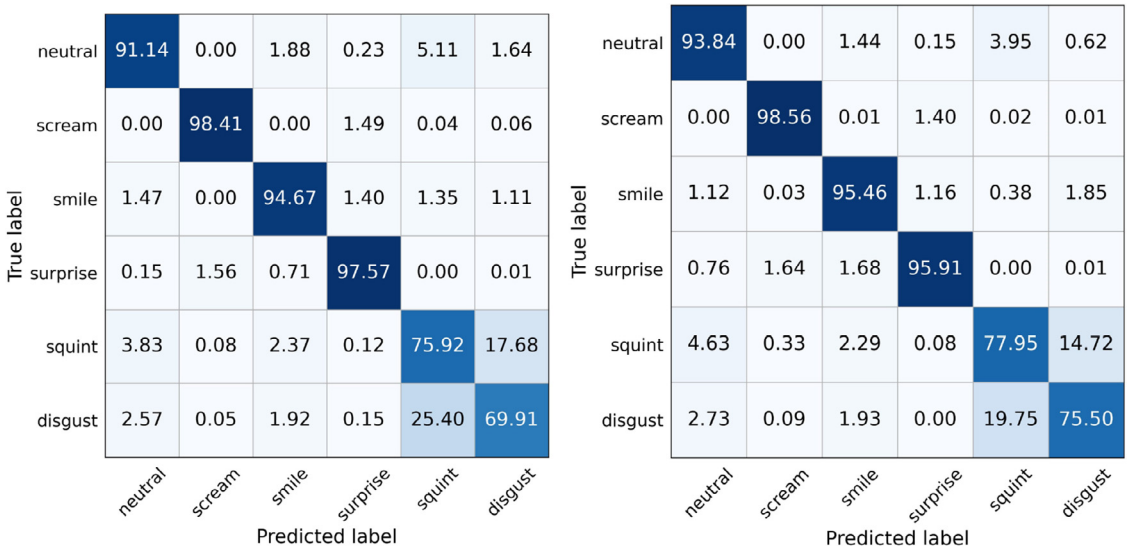
To verify the robustness of our method to illumination variations, an experiment using OCA-MTL is conducted on Multi-PIE Set II. Table 3 shows the recognition accuracy of each expression in Multi-PIE Set II using the OCA-MTL. The recognition accuracy of the same expression under different pan angles does not fluctuate much due to subtle pose variation. Compared to Table 1, there is a decrease in recognition accuracy, indicating that illumination variations affect the performance of the proposed model, but the influence is insignificant.

Figure 8 shows the confusion matrix of FER on Multi-PIE Set II. Even though the recognition accuracy rate of the surprise expression is 1.66% lower than the baseline, OCA-MTL still performs better than the baseline in recognizing the rest expressions. Similar to Fig. 7, disgust and squint expressions are still more likely to be confused with each other, and the proposed method can effectively prevent these confusions. It is inferred that this confusion is consistent to some extent in the two datasets.

Table 4 summarizes the average expression recognition accuracy of our OCA-MTL and the state of the arts on the Multi-PIE Set II, including LDA [32], which uses the 3D geometric shape model and Linear Discriminant Analysis classifier; Bayesian Belief Net [33], 2D + 3D Feature fusion-SVM [34] using 2D and 3D shape descriptors to train the SVM classifier; and Siamese CNN [5], which extracts features robust to image variations with two identical networks. Others are PL-fusion-VGG19 [35], which employs the fusion of different features and pre-trained VGG19 network; and STSN [6], which embeds the spatial transformer into the Siamese network to focus on useful local patches, indicating that OCA-MTL has better recognition performance with illumination variations.

4.4.3. Experiments on the KDEF dataset

To further assess the effectiveness of our method, an experiment using the OCA-MTL is conducted on the KDEF dataset. Table 5 shows the recognition accuracy of each expression under different viewpoints on the KDEF dataset. Recognition accuracy de-



(a) Confusion matrix with the baseline method (b) Confusion matrix with the proposed OCA-MTL

Fig. 8. Confusion matrix of FER under all the head poses on the Multi-PIE Set II with baseline method and the proposed method.

Table 4
Performance comparison with state-of-the-art methods on the Multi-PIE Set II.

Method	Poses		Expressions number	Illumination number	Overall (%)
	number	pan			
LDA [32]	5	(-30°, +30°)	6	20	81.3
Bayesian Belief Net [33]	5	(-30°, +30°)	6	20	81.9
2D+3D Feature fusion-SVM [34]	5	(-30°, +30°)	6	20	85.2
Siamese CNN [5]	5	(-30°, +30°)	6	20	88.0
PL-fusion-VGG19 [35]	5	(-30°, +30°)	6	20	86.7
STSN [6]	5	(-30°, +30°)	6	20	88.1
OCA-MTL	5	(-30°, +30°)	6	20	**89.5**

Table 5
Recognition accuracies under various head poses on the KDEF dataset.

Acc.(%)	-90°	-45°	0°	+45°	+90°	Avg.
Fear	72.86	83.57	86.43	78.57	75.71	79.43
Anger	84.29	82.86	92.14	83.57	86.43	85.86
Disgust	85.71	86.43	93.57	90.71	86.43	88.57
Happy	93.57	99.29	96.43	97.86	95.71	96.57
Neutral	94.29	96.43	95.00	95.71	95.71	95.43
Sadness	86.43	85.00	90.00	87.86	81.43	86.14
Surprise	92.14	90.71	92.14	90.00	91.43	91.28
Overall	87.04	89.18	92.24	89.18	87.55	89.04

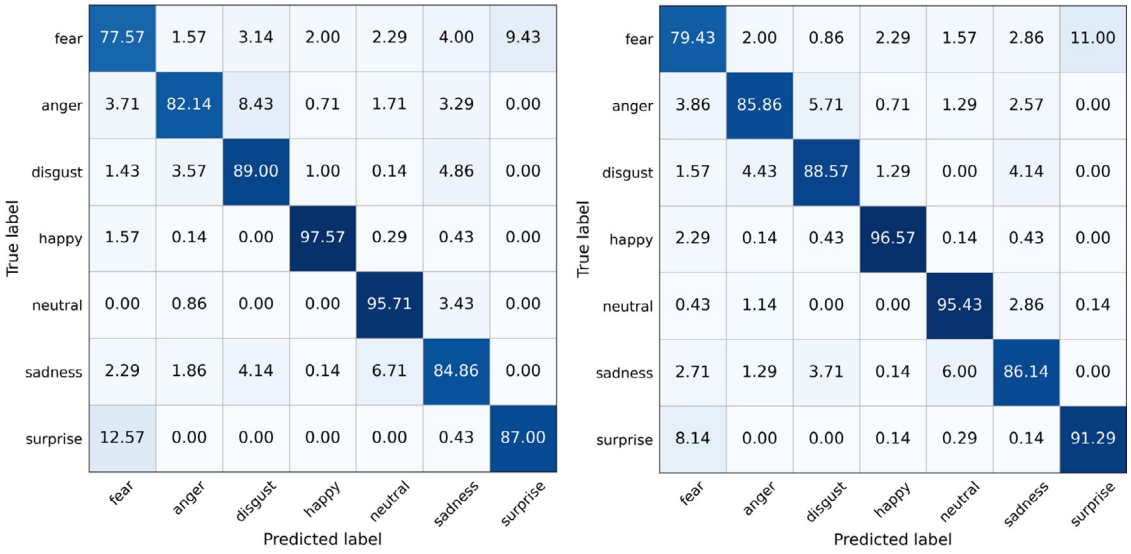
creases dramatically when the viewpoint changes from 0° to 90°. The main reason for this discrepancy is that the KDEF database only provides facial images under five viewpoints, which may be inadequate for learning view-invariant features.

Figure 9 shows the confusion matrix of FER on the KDEF database. Compared with the baseline method, the recognition accuracy rates of three expressions (disgust, happy, neutral) using OCA-MTL are slightly lower (the gap is less than 1%), but the recognition results of other expressions are much better. The overall average recognition accuracy rate of OCA-MTL is 1.35% higher than that of the baseline method. For the proposed OCA-MTL, confusion appears between the fear and surprise expressions, as 8.14% of surprise expressions are recognized as fear expressions, and 11.00% of fear expressions are misclassified as surprise expressions. However, high predictive accuracy rates of 91.29%, 96.57%, and 95.43% are achieved for surprise, happy, and neutral expressions, respectively; these rates are higher than the overall average recognition accuracy rate of 89.04%.

Table 6 lists the FER accuracy of OCA-MTL on the KDEF dataset compared to DenseNet [36], which allows features optimized for early classifiers in later layers of the network, transfer learning based CNN [37] based on action unit selectivity for feature selection, SURF boosting [38] using the feature descriptor of Speeded Up Robust Features and the boosting classifier, the LBP method [7] using the feature descriptor of Local Binary Pattern and SVM classifier, and PhaNet [25], which discovers the most relevant regions to the facial expression by an attention mechanism in hierarchical scales. Our method achieves a recognition accuracy rate of 89.0%, which is higher than the rate of other methods and demonstrates the universality of OCA-MTL on different datasets.

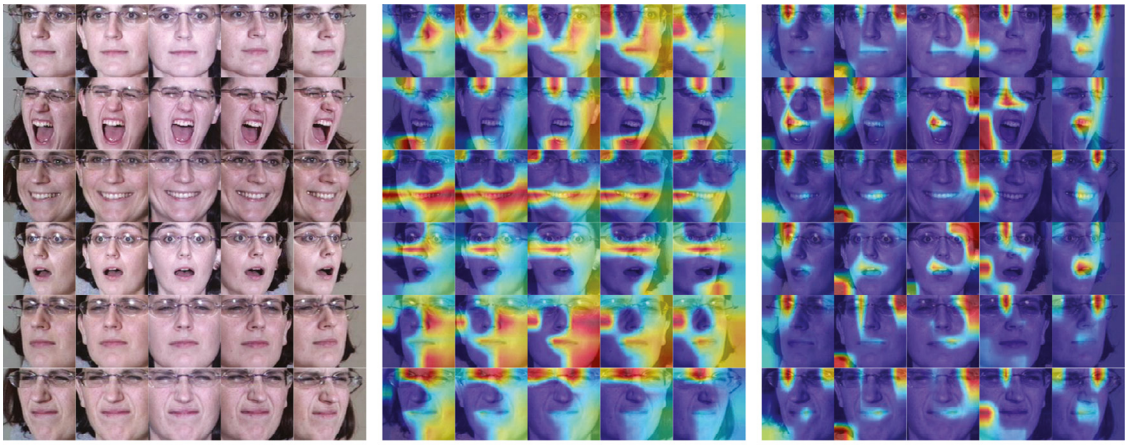
4.5. Visualization analysis

To investigate discrepancies and similarities among different expressions, Grad-CAM [39] is used to conduct a visualization analysis of the proposed OCA-MTL, making the FER results more transparent and explainable. Gradients of two parallel paths are calculated to produce two coarse localization maps highlighting the different important regions of the same image for FER and HPE, respectively. Figures 10 and 11 show input images and the corresponding visualization outputs by Grad-CAM on Multi-PIE Set II and the KDEF dataset, respectively. Red regions represent high score for class decision. The same expressions under different head



(a) Confusion matrix with the baseline method (b) Confusion matrix with the proposed OCA-MTL

Fig. 9. Confusion matrix of FER under all the head poses on KDEF database with baseline method and the proposed method.



(a) Original images (b) Attention region for FER (c) Attention region for HPE

Fig. 10. Visualization analysis using Grad-CAM on Multi-PIE Set II. Red regions indicate a high score for subtask prediction.

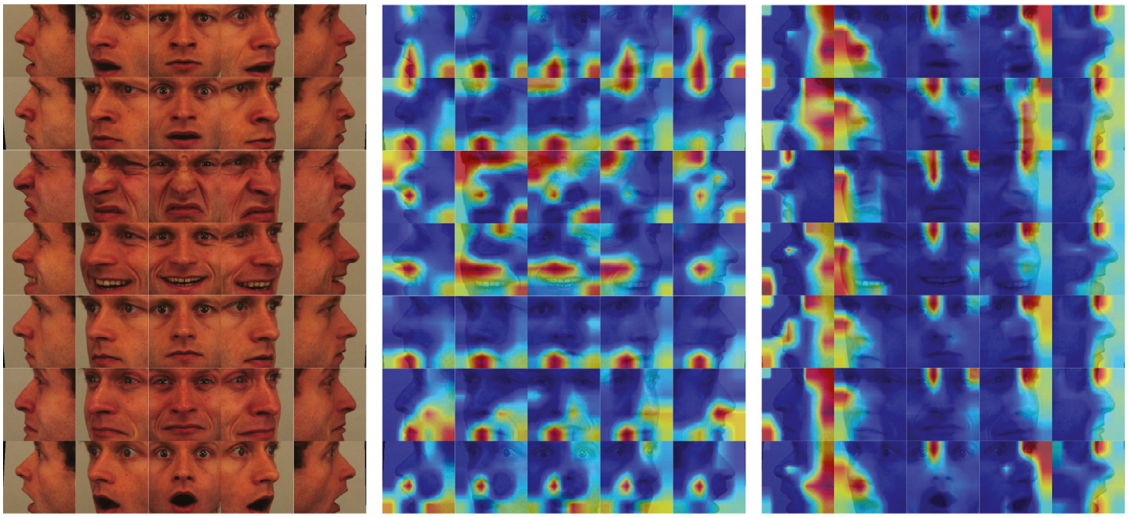
Table 6
Performance comparison with state-of-the-art methods on the KDEF dataset.

Method	Poses		Expressions number	Illumination number	Overall (%)
	number	pan			
LBPms [7]	5	(-90°, +90°)	7	-	70.5
SURF boosting [38]	5	(-90°, +90°)	7	-	74.1
DenseNet [36]	5	(-90°, +90°)	7	-	85.1
TLCNN [37]	5	(-90°, +90°)	7	-	86.4
PhaNet [25]	5	(-90°, +90°)	7	-	88.5
OCA-MTL	5	(-90°, +90°)	7	-	**89.0**

poses are aligned in a row, and the head pose is the same for each column. Figures 10(b) and 11(b) highlight important regions of each face for FER, while Figs. 10(c) and 11(c) highlight important regions of each face for HPE.

Figure 10 (b) shows that the same expression tends to focus on similar regions, even under different head poses, indicating that OCA-MTL can maintain strong robustness to pose variations. For example, the scream expression, composed of frowning brows and an open jaw, draws more attention to the region of the brow and chin, while the disgust expression is composed of a wrinkled nose and squared lips, drawing more attention to the region between the brows.

As mentioned above, disgust and squint expressions are more likely to be confused with each other. It is plausible that some similarities may have influenced the recognition results obtained, while highlighting maps generate visual explanations for more transparency into how the OCA-MTL works. First, facial regions with similar contours are detected in both squint and disgust because these two expressions are objectively similar to each other. Most regions, except for the cheek and nose on the lower left side,



(a) Original image (b) Attention region for FER (c) Attention region for HPE

Fig. 11. Visualization analysis using Grad-CAM on the KDEF dataset. Red regions indicate a high score for subtask prediction.

Table 7
Ablation study on three datasets.

Dataset	Model	Siamese CNN	CAS Module	OCA Loss	FER Acc.(%)	HPE Acc.(%)
Multi-PIE Set I	Siamese CNN	✓			86.53	98.58
	Siamese MTL	✓	✓		87.58	98.79
	OCA-MTL	✓	✓	✓	**88.41**	**98.87**
Multi-PIE Set II	Siamese CNN	✓			87.97	97.93
	Siamese MTL	✓	✓		88.70	98.06
	OCA-MTL	✓	✓	✓	**89.54**	**98.09**
KDEF	Siamese CNN	✓			87.69	99.82
	Siamese MTL	✓	✓		88.16	99.88
	OCA-MTL	✓	✓	✓	**89.04**	**99.90**

seem to be involved in expression classification. Another possible reason for misclassification is that both expressions display squinting eyes. Psychological research has shown that attention to the eye region is essential for the accurate evaluation of others emotional states [40]. Similar appearance variation in the eye regions may lead to confusion in the prediction process. Furthermore, both squint and neutral expressions have very similar highlights in the right cheek and mouth regions, which may explain why 9.08% of neutral expressions are recognized as squint expressions.

As shown in Fig. 10(c), the HPE path focuses on similar regions for faces with the same head pose (each column), which facilitates making consistent predictions. Moreover, HPE pays more attention to facial contours, the teeth, or the nose region, which differ from the attention regions of facial expressions.

The visualization analysis results for the KDEF dataset emphasize the validity of our model. According to Figs. 11(b) and 11(c), anger, neutral, and sadness expressions composed of a closed mouth drew attention to the mouth region. Furthermore, anger expressions focus more on the brow region, and sadness expressions focus more on the contours of nasolabial folds. Expressions of disgust involve brow contortion and the cheek region, so attention is drawn to these regions. The highlighted regions of happy expressions involve the upper lip and mouth; they are detected because of the grinning mouth, which is very different from other expressions.

The same expression under different head poses draws attention to similar regions, and the HPE path focuses on similar regions for faces with the same head pose. Thus, it is proven that the proposed SCA module with OCA loss can effectively decouple facial expression and head pose features, thus improving the performance of FER.

4.6. Ablation study

To investigate the role of an individual loss term for model optimization in expression recognition accuracy, an ablation study is conducted to evaluate various models with different modules or losses on three datasets. Concretely, three models are constructed to analyze the gains of the CAS Module and OCA Loss, namely the Siamese CNN in [5], the Siamese MTL with the CAS Module, and OCA-MTL with both the CAS Module and OCA Loss.

Table 7 records the experimental results of the Siamese CNN, Siamese MTL, and OCA-MTL under the corresponding experimental settings on three datasets. The three models recognition accuracy rates for FER are 86.53%, 87.58%, and 88.41% on Multi-PIE Set I; 87.97%, 88.70%, and 89.54% on Multi-PIE Set II; and 87.69%, 88.16%, and 89.04% on the KDEF dataset. Their recognition accuracy rates for HPE are 98.58%, 98.79%, and 98.87% on Multi-PIE Set I; 97.93%, 98.06%, and 98.09% on Multi-PIE Set II; and 99.82%, 99.88%, and 99.90% on the KDEF dataset. Consistently, for all three datasets, the Siamese MTL is more accurate than the Siamese CNN, and the proposed OCA-MTL is superior to the Siamese MTL for both subtasks. The incremental gains in recognition accuracy due to the contribution of the CAS Module and OCA Loss further confirm the clear advantage of our view-independent facial expression feature-learning framework.

5. Conclusions

In this paper, a multi-view FER approach called OCA-MTL has been presented. The proposed approach learns view-independent facial expression features using a Siamese CNN. The proposed model has two parallel paths for learning facial expression features

from both frontal and non-frontal viewpoints. The two parallel paths share the network parameters for facial expression feature extraction, forcing the path for the non-frontal viewpoint to learn expression features similar to that at the frontal viewpoint. However, different from the traditional Siamese network, the two parallel networks used in our framework are not identical. The HPE as an auxiliary task is introduced into the non-frontal viewpoint path, aimed at learning better low-level visual semantic features leveraging the cooperative training ability of multi-task learning and improving the high-level visual semantic information of each subtask by the proposed SCA module with an orthogonal constraint.

Numerous experiments have been conducted on two public multi-view expression databases, and the results indicate that a separated channel attention module can effectively decouple head pose and facial expression features from various viewpoints, further improving the performance of multi-view FER.

This paper has discussed the fact that the orthogonality of the channel attention level has contributed to multi-task learning, that is, non-overlapping feature channels are selected to represent different subtasks, but it does not discuss the orthogonality of the feature space across channels and its effects on multi-view FER. This will also be covered in our future work.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 61977027), the Hubei Province Technological Innovation Major Project (No. 2019AAA044).

References

[1] P. Ekman, E.L. Rosenberg, What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS) (1997).
[2] C.P. Sumathi, Automatic facial expression analysis a survey, Int. J. Comput. Sci.Eng. Surv. 3 (6) (2012) 47–59.
[3] S. Li, W. Deng, Deep facial expression recognition: a survey, IEEE Trans. Affect. Comput. (2020), doi:10.1109/TAFFC.2020.2981446. 1–1
[4] Y. Liu, X. Yuan, X. Gong, Z. Xie, F. Fang, Z. Luo, Conditional convolution neural network enhanced random forest for facial expression recognition, Pattern Recognit. 84 (2018), doi:10.1016/j.patcog.2018.07.016.
[5] W.J. Baddar, D.H. Kim, Y.M. Ro, Learning features robust to image variations with Siamese networks for facial expression recognition, in: International Conference on Multimedia Modeling, 2017, pp. 189–200.
[6] S. Luo, X. Zhang, Y. Guo, S. Bai, Facial expression recognition based on spatial transformer Siamese networks, in: 2018 IEEE 4th International Conference on Computer and Communications (ICCC), 2018, pp. 1453–1457.
[7] S. Moore, R. Bowden, Local binary patterns for multi-view facial expression recognition, Comput. Vis. Image Understanding 115 (4) (2011) 541–558.
[8] F. Güney, N.M. Arar, M. Fischer, H.K. Ekenel, Cross-pose facial expression recognition, in: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2013, pp. 1–6, doi:10.1109/FG.2013.6553814.
[9] J. Wu, Z. Lin, W. Zheng, H. Zha, Locality-constrained linear coding based bi-layer model for multi-view facial expression recognition, Neurocomputing 239 (2017) 143–152.
[10] W.J. Baddar, Y.M. Ro, Mode variational LSTM robust to unseen modes of variation: application to facial expression recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 3215–3223.
[11] Y.-H. Lai, S.-H. Lai, Emotion-preserving representation learning via generative adversarial network for multi-view facial expression recognition, in: 2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018), 2018, pp. 263–270, doi:10.1109/FG.2018.00046.
[12] R. Ranjan, V.M. Patel, R. Chellappa, HyperFace: a deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition, IEEE Trans. Pattern Anal. Mach. Intell. 41 (1) (2019) 121–135, doi:10.1109/TPAMI.2017.2781233.
[13] L. Mao, Y. Yan, J.-H. Xue, H. Wang, Deep multi-task multi-label CNN for effective facial attribute classification, IEEE Trans. Affect. Comput. (2020), doi:10.1109/TAFFC.2020.2969189. 1–1
[14] B. Chen, W. Guan, P. Li, N. Ikeda, H. Lu, Residual multi-task learning for facial landmark localization and expression recognition, Pattern Recognit. 115 (6) (2021) 107893.
[15] Y. Gan, J. Chen, L. Xu, Learning head pose-insensitive and discriminative deep features for smile detection, J. Electron. Imaging 27 (PT.2) (2018) 1.
[16] D. Li, Z. Li, R. Luo, J. Deng, S. Sun, Multi-pose facial expression recognition based on generative adversarial network, IEEE Access 7 (2019) 143980–143989, doi:10.1109/ACCESS.2019.2945423.
[17] F. Zhang, T. Zhang, Q. Mao, C. Xu, Geometry guided pose-invariant facial expression recognition, IEEE Trans. Image Process. 29 (2020) 4445–4460.
[18] S. Xie, H. Hu, Y. Wu, Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition, Pattern Recognit. 92 (2019) 177–191, doi:10.1016/j.patcog.2019.03.019.
[19] W. Sun, H. Zhao, Z. Jin, A visual attention based ROI detection method for facial expression recognition, Neurocomputing 296 (JUN.28) (2018) 12–22.
[20] K. Wang, X. Peng, J. Yang, D. Meng, Y. Qiao, Region attention networks for pose and occlusion robust facial expression recognition, IEEE Trans. Image Process. 29 (2020) 4057–4069, doi:10.1109/TIP.2019.2956143.
[21] Z. Wang, F. Zeng, S. Liu, B. Zeng, OAENet: oriented attention ensemble for accurate facial expression recognition, Pattern Recognit. 112 (5) (2020) 107694.
[22] J. Hu, L. Shen, S. Albanie, G. Sun, E. Wu, Squeeze-and-excitation networks, IEEE Trans. Pattern Anal. Mach. Intell. 42 (8) (2020) 2011–2023, doi:10.1109/TPAMI.2019.2913372.
[23] Y. Gan, J. Chen, Z. Yang, L. Xu, Multiple attention network for facial expression recognition, IEEE Access 8 (2020) 7383–7393, doi:10.1109/ACCESS.2020.2963913.
[24] Krizhevsky, Alex, Sutskever, Ilya, Hinton, E. Geoffrey, ImageNet classification with deep convolutional neural networks, Commun. ACM (2017).
[25] Y. Liu, J. Peng, J. Zeng, S. Shan, Pose-adaptive hierarchical attention network for facial expression recognition, arXiv preprint arXiv:1905.10059(2019).
[26] K. Zhang, Z. Zhang, Z. Li, Y. Qiao, Joint face detection and alignment using multitask cascaded convolutional networks, IEEE Signal Process. Lett. 23 (10) (2016) 1499–1503.
[27] M. Fischer, H.K. Ekenel, R. Stiefelhagen, Analysis of partial least squares for pose-invariant face recognition, in: 2012 IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS), 2012, pp. 331–338, doi:10.1109/BTAS.2012.6374597.
[28] J. Fan, S. Wang, P. Yang, Y. Yang, Multi-view facial expression recognition based on multitask learning and generative adversarial network, in: 2020 IEEE 18th International Conference on Industrial Informatics (INDIN), vol. 1, 2020, pp. 573–578, doi:10.1109/INDIN45582.2020.9442212.
[29] W. Zheng, Multi-view facial expression recognition based on group sparse reduced-rank regression, IEEE Trans. Affect. Comput. 5 (1) (2014) 71–85.
[30] H. Jung, S. Lee, J. Yim, S. Park, J. Kim, Joint fine-tuning in deep neural networks for facial expression recognition, in: 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2983–2991, doi:10.1109/ICCV.2015.341.
[31] D.C. Luvizon, D. Picard, H. Tabia, 2D/3D pose estimation and action recognition using multitask deep learning, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 5137–5146, doi:10.1109/CVPR.2018.00539.
[32] L. Yin, X. Wei, Y. Sun, J. Wang, M.J. Rosato, A 3D facial expression database for facial behavior research, in: 7th International Conference on Automatic Face and Gesture Recognition (FGR06), 2006, pp. 211–216, doi:10.1109/FGR.2006.6.
[33] X. Yang, D. Huang, Y. Wang, L. Chen, Automatic 3D facial expression recognition using geometric scattering representation, in: 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), vol. 1, 2015, pp. 1–6, doi:10.1109/FG.2015.7163090.
[34] H. Li, H. Ding, D. Huang, Y. Wang, X. Zhao, J.M. Morvan, L. Chen, An efficient multimodal 2D + 3D feature-based approach to automatic facial expression recognition, Comput. Vis. Image Understanding 140 (NOV) (2015) 83–92.
[35] O.K. Oyedotun, G. Demisse, A.E.R. Shabayek, D. Aouada, B. Ottersten, Facial expression recognition via joint deep learning of RGB-depth map latent representations, in: 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), 2017, pp. 3161–3168, doi:10.1109/ICCVW.2017.374.
[36] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2261–2269, doi:10.1109/CVPR.2017.243.
[37] Y. Zhou, B.E. Shi, Action unit selective feature maps in deep networks for facial expression recognition, in: 2017 International Joint Conference on Neural Networks (IJCNN), 2017, pp. 2031–2038, doi:10.1109/IJCNN.2017.7966100.
[38] Q. Rao, X. Qu, Q. Mao, Y. Zhan, Multi-pose facial expression recognition based on SURF boosting, in: 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), 2015, pp. 630–635, doi:10.1109/ACII.2015.7344635.
[39] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: visual explanations from deep networks via gradient-based localization, Int. J. Comput. Vis. 128 (2) (2020) 336–359.
[40] J.K. Hall, S.B. Hutton, M.J. Morgan, Sex differences in scanning faces: does attention to the eyes explain female superiority in facial expression recognition? Cognit. Emotion 24 (4) (2010) 629–637.

JINGYING CHEN received the BS degree in electronic information engineering and the ME degree in pattern recognition and intelligent system from Huazhong University of Science and Technology, Wuhan, China, in respectively 1996 and 1998, and the PhD degree in computer science from Nanyang Technological University, Singapore, in 2001. From 2002 to 2003, she was a Research Scientist with the Singa-

pore Technologies Engineering Ltd, Singapore. From 2003 to 2004, she was a Postdoctor in INRIA, France. Form 2005 to 2010, she was a Research Fellow with University of St. Andrews and University of Edinburgh, U.K. Since 2011, she has been a Professor jointly with the National Engineering Research Center for ELearning and the National Engineering Laboratory for Educational Big Data, Central China Normal University, China. Her research interests include computer vision, pattern recognition, and human-computer interaction. Dr. Chen was a recipient of the Second Prize for Hubei Provincial Excellent Social Science Paper Award in respectively 2018 and 2020, and the First Prize for Hubei Provincial Progress in Science and Technology Award in 2020.

LEI YANG received the BS degree in communication engineering from Wuhan University of Technology, Wuhan, China, in 2019. He is currently working toward the ME degree in computer science and technology at Central China Normal University, Wuhan, China. His research interests include computer vision, computer graphics and machine learning.

LEI TAN received the BS degree in automation from University of Science and Technology Beijing, China, in 2017, and M.S. degree in communication and information system from Central China Normal University, China, in 2020. He is currently pursuing the PhD degree with the School of informatics, Xiamen University. His current research interests include computer vision and machine learning.

RUYI XU received the ME degree in circuits and systems from Huazhong University of Science and Technology, Wuhan, China, in 2016. He is currently pursuing the PhD degree with National Engineering Laboratory For Educational Big Data, Central China Normal University. His current research interests include computer vision and human-computer interaction.