# Facial Expression Recognition with Identity and Emotion Joint Learning

Ming Li, *Member, IEEE*, Hao Xu, Xingchang Huang, Zhanmei Song, Xiaolin Liu, and Xin Li, *Fellow, IEEE*

**Abstract**—Different subjects may express a specific expression in different ways due to inter-subject variabilities. In this work, besides training deep-learned facial expression feature (emotional feature), we also consider the influence of latent face identity feature such as the shape or appearance of face. We propose an identity and emotion joint learning approach with deep convolutional neural networks (CNNs) to enhance the performance of facial expression recognition (FER) tasks. First, we learn the emotion and identity features separately using two different CNNs with their corresponding training data. Second, we concatenate these two features together as a deep-learned Tandem Facial Expression (TFE) Feature and feed it to the subsequent fully connected layers to form a new model. Finally, we perform joint learning on the newly merged network using only the facial expression training data. Experimental results show that our proposed approach achieves 99.31 and 84.29 percent accuracy on the CK+ and the FER+ database, respectively, which outperforms the residual network baseline as well as many other state-of-the-art methods.

**Index Terms**—Facial expression recognition, emotion recognition, face recognition, joint learning, transfer learning

## 1 Introduction

FACIAL Expression Recognition (FER) is a well defined task, aiming to recognize facial expressions with discrete categories (e.g., neutral, sad, contempt, happy, surprise, angry, fear, disgust, etc.) or continuous levels (e.g., valance, arousal) from still images or videos. Although many recent works focus on video or image sequence based FER tasks [1], [2], still image based FER still remains as a challenging problem. First, the differences between some facial expressions might be subtle and thus difficult to classify them accurately in some cases [3]. Second, different subjects express the same specific facial expression in different ways due to the inter-subject variability and their facial biometric shapes [4], [5], [6], etc.

These two challenging problems can be visualized in the following examples in Fig. 1. The left part of Fig. 1 contains two representative faces and both of them are labeled with the "sad" facial expression. However, their eight-category classification scores have great differences in our initial experiment, which is shown in Table 1. The prediction for the left subject is reasonable as the "sad" emotion ranks the top. However, for the person on the right, the score of "angry" is a little higher than "sad", which did not correctly predict her emotion.

Generally, the inter-speaker variability of the faces could potentially lead to errors in emotion classification because the neutral face of one subject could already be very similar to the typical faces of other emotion categories (e.g., the right subject in Fig. 1 and the "angry" emotion). Furthermore, neural science studies also show that the facial expression and identity representations in human cortex are closely connected to each other [7]. Therefore, we believe that if we add a compact description of the subject's facial identity or biometric information as an auxiliary input to our model, the FER system can become more subject adapted and robust against the inter-speaker variability just as the role of speaker adaptation technique in speech recognition tasks [8], [9].

Previous works show that deep neural network based methods have achieved excellent performance in face related recognition tasks [1], [10], [11], [12], [13], [14], [15]. In face recognition, Deep Convolutional Neural Networks (CNNs) outperform traditional methods with hand-crafted features [16], [17], [18], [19], and even perform better than human beings[10], [11], [12], [20]. However, in FER tasks, the system performance still needs to be further enhanced. Lack of large scale labeled training data, inconsistent and unreliable emotion labels and inter-subject variabilities all limit the performance of CNN on the FER task. Therefore, in this work, we aim to utilize additional face recognition training data to perform identity and emotion joint learning for FER.

Related to our work, Xu et al. [21] proposed a transfer learning method from face recognition to FER using CNN directly. Also, Jung et al. [1] proposed a joint fine-tuning method that jointly learns the parameters from image sequences. However, unlike these two methods, we do not transfer the network structures and parameters from face recognition to FER directly. Instead, we extract the high-level identity feature from the face recognition network and consider it as an auxiliary input feature for our FER model. As shown in Fig. 2, we concatenate both the high-level emotion and identity features as Tandem Facial Expression (TFE) features and feed it to the subsequent fully connected layers to form a new network.

In this paper, we adopt the CNN architecture to discover latent identity and emotion features. First, we pre-train latent emotion and identity features separatively using two different CNNs (ResNet [22] for emotion and DeepID [10] for identity) with their own training data. Furthermore, we merge these two networks together by concatenating the deep-learned features and feed to a new fully connected layer. Finally, we use FER training data to jointly learn the parameters of the merged new network. To the best of our knowledge, there is no previous work using auxiliary deep identity feature with deep emotion feature together for joint facial expression learning.

## 2 Related Work

In this section, we will introduce two main types of features used in the FER task, namely hand-crafted features and deep-learned features.

### 2.1 Hand-Crafted Feature Based Method

Before deep learning based approaches dominate face recognition and FER tasks, many works have been conducted based on the hand-crafted features [3]. These approaches usually perform front-end feature extraction and backend classification separately [15]. During the stage of feature extraction, traditional features, such as Local Binary Patterns (LBP) [16], [18], Gabor wavelet coefficients [17], Scale-Invariant Feature Transform (SIFT) [19], and Gaussian Face [20] are designed with prior domain knowledge. Moreover, supervised classifiers, such as SVM [23], feedforward Neural Network (NN) [24] and Extreme Learning Machine [25], [26] are adopted for the subsequent modeling.

- *M. Li is with the Data Science Research Center, Duke Kunshan University, Kunshan 215316, China. E-mail: ming.li369@dukekunshan.edu.cn.*
- *H. Xu and X. Huang are with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510275, China. E-mail: {379548839, 767967354}@qq.com.*
- *Z. Song, X. Liu, and X. Li are with the School of Preschool Education, Shandong Yingcai University, Jinan 250104, China. E-mail: songzhanmei@126.com, liuxiaolin0531@qq.com, xinli.ece@duke.edu.*

*Manuscript received 21 Dec. 2017; revised 4 Sept. 2018; accepted 20 Oct. 2018. Date of publication 9 Nov. 2018; date of current version 28 May 2021.*
*(Corresponding author: Zhanmei Song).*
*Recommended for acceptance by M. Valstar.*
*Digital Object Identifier no. 10.1109/TAFFC.2018.2880201*

Fig. 1. Two example face images with "sad" facial expressions. The left part is the original faces of these two subjects and the right part is the faces with detected landmarks.

### 2.2 Deep-Learned Feature Based Method

Generally, deep learning based methods outperform hand-crafted feature based approaches and achieve state-of-the-art performance on both face recognition [10], [11], [12] and FER [1], [2], [15], [27] tasks. For example, Sun et al. [12] proposed CNN model and DeepID features for face recognition. In order to boost the performance, Sun et al. [10] proposed a Siamese network to train in face pairs. Furthermore, CNN models have been widely used in the FER task. Tang et al. [13] replace the softmax layer with SVM in the CNN framework and achieved the best accuracy on FER2013 dataset [28] in the ICML 2013 Representation Learning Challenge. Emad et al. [29] then proposed a FER+ dataset with more accurate labels and produced a benchmark on this dataset with VGG13 network. Jung et al. [1] and Zhao et al. [2] both consider temporal structures on top of CNNs to model the image sequences.

Recent high performance models normally accompany with deep architectures and a large number of convolutional kernels. Alex et al. [30] has proposed AlexNet for the ImageNet challenge and achieve superior performance at that time. Subsequently, Karen et al. [31] presented their deeper networks with 16 and 19 layers respectively and found that deeper structures can achieve better performance. Furthermore, He et al. [22] proposed deep residual network (ResNet) and trained a CNN with 152 layers. ResNet can converge faster and perform more accurately due to its residual learning mechanism, shortcut connection [22] and batch normalization [32]. Also, Christian et al. [33] proposed GoogleNet and its inception_v4 architecture. They further combined the residual network architecture with Inception-v4 as Inception-ResNet [34] and achieve better performance on the ImageNet dataset [35]. In this work, our approach adopts deep ResNet and CNN for their high performance and less chance of overfitting on image related tasks.

### 2.3 Transfer Learning with CNN

A recent work proposed by Xu et al. [21] used transfer learning with CNN for FER. The difference between our work and their work is that we learn both the emotion and identity features using two separate deep convolutional neural networks and construct a deep-learned Tandem Facial Expression feature in the merged model instead of just transferring the weights of the pre-trained face recognition networks and fine-tuning. Compared with the work by Jung et al. [1], we use feature-level concatenation of deep-learned identity and emotion features to form a new network with joint learning rather than fine-tuning two softmax layers with the landmark-based features.

TABLE 1
The Predicted Scores on 8 Facial Expression Categories for the Two Example Images in Fig. 1

| Figure | Neutral | Angry | Contempt | Disgust | Fear | Happy | Sad | Surprise |
|---|---|---|---|---|---|---|---|---|
| Left | **0.2536** | 0.0042 | 0.0038 | 0.0004 | 0.0025 | 0.0003 | **0.7332** | 0.0020 |
| Right | 0.0002 | **0.5110** | 0.0 | 0.0009 | 0.0001 | 0.0001 | **0.4876** | 0.0001 |

## 3 Our Approach

In this section, we will introduce our proposed method in details.

### 3.1 Overview of our Network Architecture

As shown in Fig. 2, our model consists of two CNNs. The left one is the DeepID network proposed in [10], [12], containing four convolutional layers. Actually, the architecture we use is the same as the ConvNet structure proposed in [10], which learns fully-connected layer (DeepID2 feature) from both the third and fourth convolutional layers, generating a compact feature representation with 160 dimensions. The right network is constructed according to the deep ResNet [22] and we choose the ResNet18 structure and also build a shallower network called ResNet12 for different tasks based on the size of the input images and the size of datasets. In ResNet, we use shortcut connection for deep residual learning and batch normalization layers [32] for faster convergence and better accuracy. We do not use fully-connected layer (FC) to flatten the feature maps. Instead, we use global average pooling (Gap) to generate a compact representation with reduced number of parameters [22]. But for convenience, we still use fully-connected layer to model the concatenated TFE features.

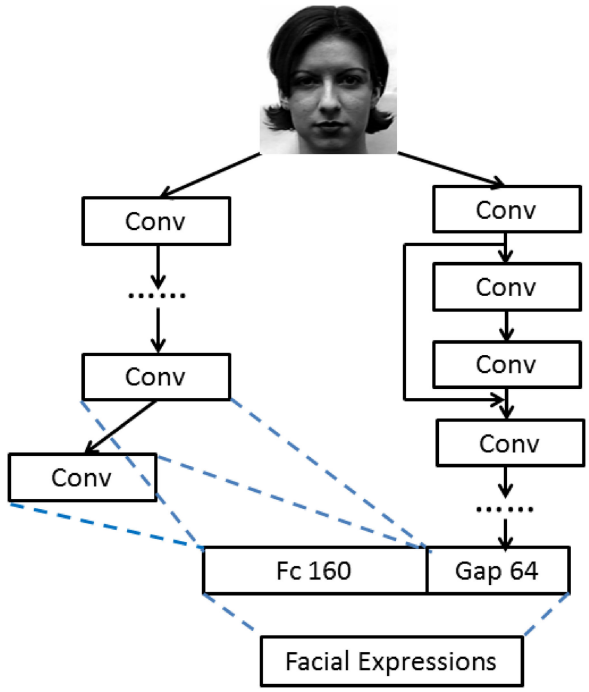


Fig. 2. Our model consists of two convolutional neural networks. The left one represents the DeepID network learning the identity features. The right deep residual network is trained with facial expression databases. After training separatively, the identity feature and the deep-learned emotion feature are concatenated as the TFE features and feed to the subsequent fully connected layers. Finally, we perform joint learning on the new merged network using only the facial expression database.

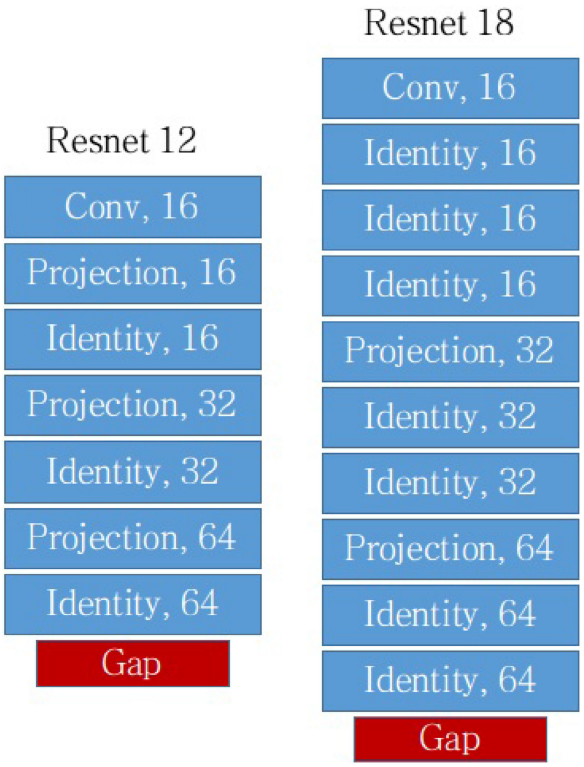


Fig. 3. Our ResNet12 and ResNet18 architectures. "project", "identity" and "gap" are projection block, identity block and global average pooling, respectively.

Actually, the left and the right networks are considered as a merged joint network in Fig. 2. During training, the features of DeepID network and ResNet are concatenated as the input for fully-connected layers and jointly learn the entire network for FER tasks. In order to guide this merged network to extract identity and emotion features, we do not train from scratch but pre-train the weights of both two sub-networks separately with the corresponding datasets except the fully connected layer.

Deep residual network (ResNet) has achieved great success in multiple challenges [22]. ResNet is based on the deep residual learning framework and it is easier to optimize the residual rather than the original mapping. Therefore, in our work we adopt ResNet for training emotion features. Specifically, the architecture of residual networks is made up of the following two blocks shown in Fig. 4. The left one, called "identity block", contains a shortcut link connecting the input x to the convolution output. The right one, called "projection block", contains a convolution operation in the shortcut connection, aiming to ensure the same output size of feature maps using a $1 \times 1$ convolution with a stride of 2 while doing the element-wise sum at the output.

In our work, the right block is used to increase the number of convolutional kernels and decrease the output size of feature maps by half. The stride for each convolution operation is 2 and there is no max-pooling layer. After the last convolution operation, we use Global Average Pooling layer to generate a 64-dimensional emotion feature as shown in Fig. 2. Considering the size of datasets, we use two ResNet structures. The first one is ResNet18 structure for processing the FER+ dataset [29] with over 35 k $48 \times 48$ images. For the smaller dataset, CK+, we reduce the number of parameters by removing four identity building blocks when $n = 16, 16, 32, 64$ and adding one project building block when $n = 16$. We call this shallower network as ResNet12. We adopt batch normalization after each convolution and before the rectified linear unit (ReLU) activations [22], [36]. The architectures of these two ResNets are shown in Fig. 3.

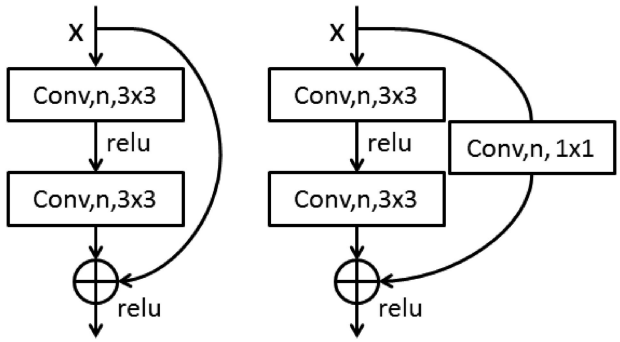


Fig. 4. The building blocks of our residual network. $n$ denotes the number of filters in convolution.

### 3.2 Identity and Emotion Feature Concatenation

Suppose that the identity and emotion features of an arbitrary input image are represented as $Z_i$ and $Z_e$, respectively. Then, we can reconstruct the new TFE representation $Z_{tfe}$ by concatenating $Z_i$ and $Z_e$ together.

However, sometimes the deep-learned $Z_i$ and $Z_e$ features are not in the same scale because both network structure and training data are different. In our work, we first normalize $Z_i$ and $Z_e$ with batch normalization [32]. Then we concatenate these two features together to form the TFE feature $Z_{tfe}$.

## 4 Experimental Results

### 4.1 Datasets

In this work, we evaluate the proposed method on two popular FER datasets, namely Extended Cohn-Kanade (CK+) database [37] and FER+ database [29]. These two FER datasets are used for deep-learned emotion feature extraction and joint learning, while the identity features are learned from the CASIA-WebFace database [38].

- *CASIA-WebFace [38]:* In this dataset, the face images are collected from the Internet, containing 10,575 subjects and 494,414 images. It is usually considered as a standard large scale training dataset for face recognition challenges [39].
- *LFW:* LFW (Labeled Faces in the Wild) dataset contains 13,233 face images from 5,749 identities collected on the Internet. As a benchmark for comparison, LFW suggests reporting performance with 10-fold cross validation using splits they have randomly generated (6,000 pairs) [39]. In this study, the LFW database is adopted purely as a stand alone testing data to evaluate the quality of our identity features trained by the CASIA-WebFace dataset.
- *CK+:* The CK+ database includes 327 image sequences with labeled facial expressions. For each image sequence, only the last frame is provided with an expression label. In order to collect more images for training, we usually selected the last three frames of each sequence for training or validation purpose. Additionally, the first frame from each of the 327 labeled sequences would be chosen as the "neutral" expression. As a result, this dataset can provide totally 1308 images with 8 labeled facial expressions. For testing, we follow the 10-fold cross validation testing protocol on the CK+ database.
- *FER+:* This dataset comes from the face expression recognition challenge [28] in the ICML 2013 Representation Learning Workshop. It consists 28,709 $48 \times 48$ face images for training. The test set has 3,589 images and there are totally 7 discrete facial expressions (anger, disgust, fear, happiness, sadness, and surprise) for classification. However, due to its noisy labels, this dataset is labeled again using crowd-sourced services [29]. In this study, we simply use majority voting to derive the new set of labels for our experiments.

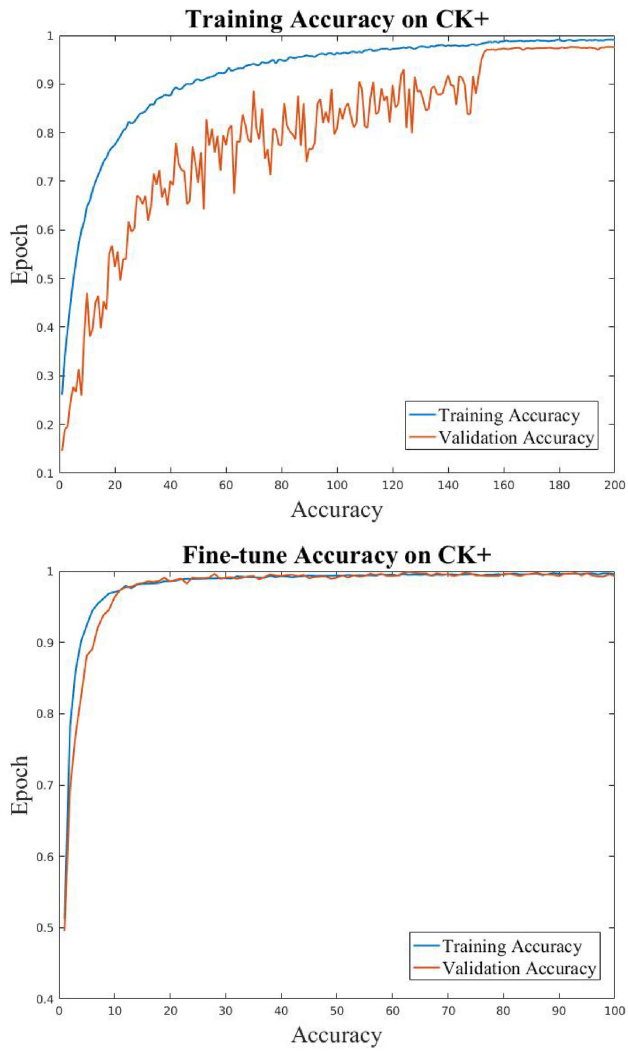


Fig. 5. Training and validation performance on the CK+ database during the training and the joint learning stage.

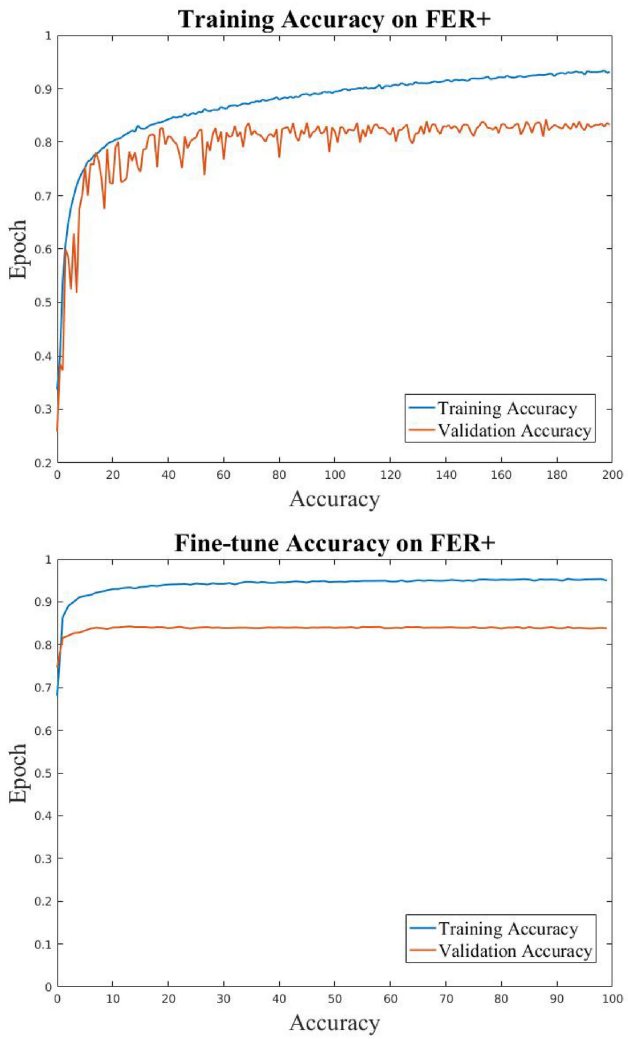


Fig. 6. Training and validation performance on the FER+ dataset during the training and the joint learning stage.

### 4.2 Parameter Settings

For the DeepID network, we follow the same parameter setting in [12], with a 160-dimensional representation in the fully-connected layers. A dropout layer is used after the DeepID layer, with a probability of 0.4 to reduce over-fitting [40].

For the deep ResNet, we have two different settings for the number of layers. We use the ResNet18 architecture for the FER+ dataset but use a shallower network ResNet12 for the CK+ dataset, which has much less data for training and testing.

As for Stochastic Gradient Descent (SGD) method during back propagation [41], we apply different parameters in CK+ dataset and FER+ dataset. For CK+, we initialize the learning rate as 0.16 and 0.01 respectively for ResNet and DeepID network with a mini-batch size of 128 and a momentum of 0.9 [42]. The networks in our model are trained for up to 200 epochs and fine-tuned for up to 100 epochs as well. For FER+, the learning rate is initialized as 0.1 for ResNet training and 0.001 for final joint learning.

### 4.3 Pre-Processing

For CASIA-WebFace dataset, we need to use pre-processing pipeline, including face detection, face landmarks detection face alignments and face cropping. We use the tools from mmlab, CUHK [43] to detect face and landmarks. After these processing steps, those missed faces are removed and there are totally 435,863 faces remaining. Then we use a template (the first image) in the LFW [39] dataset to align the faces in the CASIA-WebFace. Finally, images from the datasets we use (CASIA-WebFace, CK+, LFW) will be cropped in the same way retaining the eye brow and jaw. As LFW provides a deep-funneled version and CK+ is collected containing the frontal whole faces, there is no need to do face alignment. FER+ has been pre-processed and cropped as well.

During training the DeepID network using CASIA-WebFace, we randomly select one image from each person for validation and therefore generate a validation set of 10,575 images, while the remaining images are used for training. It is worth noting that the training set of CASIA-WebFace is augmented with horizontal flipping while the training set in FER+ is augmented with horizontal flipping, shifting and rotation. In addition, all the images are pre-processed with per-pixel mean subtraction and standard deviation normalization [13].

### 4.4 Evaluation of Identity Features on the LFW Database

After the pre-processing step, we train our DeepID network for extracting the auxiliary identity feature on CASIA-WebFace. After training 200 epochs for DeepID network, the accuracy on LFW for face verification can achieve 91 percent using cosine similarity with a 0.15 threshold.

Since our goal is to extract a reasonable good quality identity feature, we may not need to fully optimize the face verification performance on LFW. We use a single DeepID network and single patch for each face image without any ensemble method.

TABLE 2
Average Accuracy on CK+ of our Models Using 10-Fold Cross Validation

| Our Methods | Average Accuracy on CK+ |
|---|---|
| ResNet12 | 97.56% |
| TFE-JL | **99.31%** |

TABLE 3
Comparion with State-of-the-Art Methods

| Methods | Average Accuracy on CK+ |
|---|---|
| DTAGN(Weighted Sum) [1] | 96.9 |
| DTAGN(Joint) [1] | 97.3 |
| IntraFace [44] | 96.4 |
| BDBN [15] | 96.7 |
| CNN [45] | 95.8 |
| PPDN [2] | 97.3% |
| TFE-JL | **99.3%** |

### 4.5 Evaluation of FER on the CK+ Database

During the 10-fold cross validation testing, we train our ResNet from scratch for each rotation with 200 epochs. After the first step training, we combine these two networks, which extract identity features and emotion features respectively, to form a 224-dimensional TFE representation. Finally, we jointly learn the parameters of the merged network with the CK+ training data.

One example of the training and joint learning process on the CK+ database is provided in Fig. 5. The training accuracy (blue) increase gradually while the validation accuracy (orange) increase with fluctuation but finally converge to 97.56 percent. In the joint learning stage using the TFE feature, the accuracy on the validation set converges faster and better than the first training stage and can achieve up to 99.24 percent. As shown in Table 2, the performance of our proposed method outperforms the ResNet baseline by 1.68 percent absolutely.

Besides the comparisons with our own implemented baselines, we also compare our method with other state-of-the-art approaches. Our method can achieve around 2 percent absolute improvement compared with PPDN model [2] as shown in Table 3. Actually, the sequence-based PPDN also can achieve 99.3 percent accuracy on the test set but it was pre-trained on CASIA-WebFace and used the whole sequence of images in CK+ for training, which does not match with our experimental setting (only the first one and the last three frames of each CK+ sequence are used).

TABLE 4
Accuracy on FER+ of Our Proposed Methods

| Methods | Accuracy on FER+ |
|---|---|
| ResNet18 | 83.1% |
| ResNet18 + FC | 83.4% |
| TFE-joint learning | **84.3%** |

TABLE 5
Accuracy on FER2013 with Old and New Labels Compared with the State-of-the-Art Methods

| Labels | Methods | Accuracy on FER2013 |
|---|---|---|
| Old | DLSVM-L2 [13] | 71.2% |
| FER2013 | Zhou et al. [28] | 69.3% |
| | Maxim Milakov [28] | 68.8% |
| | Radu+Marius+Cristi [28] | 67.5% |
| | Our implementation of [21] | 71.1% |
| New | VGG13(MV) [29] | 83.8% |
| FER+ | TFE-JL | **84.3%** |

TABLE 6
The Predicted Scores of Those Representative Images in Fig. 7 Using ResNet and Our TFE Joint Learning Method

| Image | Neutral | Happy | Surprise | Sad | Angry | Disgust | Fear | Contempt |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.0 | 0.0093 | **0.9905** | 0.0767 | 0.0 | 0.0001 | 0.0001 | 0.0 |
| | 0.0 | **0.6180** | 0.3726 | 0.0 | 0.0005 | 0.0012 | 0.00072 | 0.0004 |
| 2 | 0.0023 | 0.3341 | **0.6633** | 0.0 | 0.0001 | 0.0 | 0.0 | 0.0001 |
| | 0.0005 | **0.6191** | 0.3493 | 0.0 | 0.0200 | 0.0076 | 0.0032 | 0.0002 |
| 3 | 0.4771 | 0.0 | 0.0 | **0.5229** | 0.0 | 0.0 | 0.0 | 0.0 |
| | **0.7262** | 0.0 | 0.0 | 0.2735 | 0.0 | 0.0002 | 0.0 | 0.0 |
| 4 | 0.3601 | **0.6396** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0003 |
| | **0.6632** | 0.3219 | 0.0002 | 0.0001 | 0.0058 | 0.0005 | 0.0003 | 0.0081 |
| 5 | 0.4883 | 0.0 | 0.0 | **0.5097** | 0.0020 | 0.0001 | 0.0 | 0.0 |
| | **0.7153** | 0.0 | 0.0 | 0.2837 | 0.0003 | 0.0005 | 0.0001 | 0.0002 |
| 6 | 0.4172 | 0.0 | 0.0 | **0.5798** | 0.0002 | 0.0006 | 0.0 | 0.0022 |
| | **0.7978** | 0.0 | 0.0 | 0.1997 | 0.0001 | 0.0013 | 0.0001 | 0.0009 |
| 7 | 0.4940 | 0.0 | 0.0 | 0.0019 | **0.5037** | 0.0004 | 0.0 | 0.0 |
| | **0.5742** | 0.0 | 0.0006 | 0.0047 | 0.4027 | 0.0097 | 0.0017 | 0.0063 |
| 8 | 0.0059 | 0.0 | 0.0 | 0.1972 | **0.7967** | 0.0 | 0.0 | 0.0001 |
| | 0.0095 | 0.0 | 0.0 | **0.7818** | 0.2000 | 0.0017 | 0.0049 | 0.0020 |
| 9 | 0.3260 | 0.0007 | 0.2076 | 0.0275 | 0.0016 | 0.0020 | **0.4337** | 0.0010 |
| | **0.7291** | 0.0081 | 0.0612 | 0.0347 | 0.0397 | 0.0683 | 0.0420 | 0.0168 |
| 10 | 0.0983 | 0.0 | 0.3654 | 0.0006 | 0.0096 | 0.0 | **0.5260** | 0.0002 |
| | 0.0560 | 0.0 | **0.5034** | 0.0046 | 0.0176 | 0.0106 | 0.4046 | 0.0031 |

*For each face image, the first line is the output score using the ResNet baseline and the second line is the output score of our TFE joint learning method.*

### 4.6 Evaluation FER on the FER+ Database

Similarly, we show the performance of our proposed methods on the FER+ dataset in Table 4 and Fig. 6. We train these two models on the training set and evaluate them on the private test set. We mainly focus on the private test set for direct comparison with VGG13(MV) proposed by [29]. Specifically, we train our ResNet18 on the training set with 200 epochs and the pre-trained accuracy is 83.1 percent on the private test set. Then we jointly learn the network using TFE features generated from DeepID and ResNet, getting an improvement up to 1.2 percent from 83.1 to 84.3 percent.

In Table 5, we compare our methods with the state-of-the-art approaches on FER2013 and FER+. The work DLSVM-L2 has been presented in [13] and ranks the top in the ICML2013 Representation Learning Challenge on the FER2013 dataset. The VGG13(MV) system outperforms the DLSVM-L2 baseline as it used new labels on FER+. Therefore, we just compare our proposed TFE-JL method with VGG13(MV). As shown in Table 5, the proposed TFE-JL method also achieves 0.5 percent accuracy gain compared with the average performance of VGG13(MV) model, which again shows the advantage and effectiveness of the proposed identity and emotion joint learning framework.

We also implement the single network transfer learning method in [21] by directly using the deep-id identity feature learned from CASIA-WebFace as inputs for the subsequent SVM modeling on FER+ database. Results in Table 5 show that our proposed joint learning method also outperforms the single network transfer learning approach.

Furthermore, considering the problem that the face images in FER2013 or FER+ dataset are not aligned as well as in CASIA-WebFace used for pre-training the DeepID network, the gain of our joint learning approach on FER+ may not be as large as in the CK+ database.

In order to demonstrate how our proposed identity and emotion joint learning method improves the FER performance, we list the output scores of 8 facial expression categories on ten representative face images from the test set in Table 6. These face images

Fig. 7. Ten representative face images whose prediction is corrected by our joint learning method. These face images are indexed with number 1 to 10 from left to right, top to bottom.

are misclassified using our baseline (ResNet) model but are corrected by the proposed identity and emotion joint learning method. As shown in Fig. 7, these images were first misclassified, the reason might be their appearances. For example, image (5) looks "sad" and image (7-8) look "angry". By adding their identity information as an auxiliary input to our FER model, the TFE joint learning approach could reduce the inter-subject variability.

## 5 Conclusion and Future Work

In this work, we learn both the emotion and identity features using two separate deep convolutional neural networks and construct a deep-learned Tandem Facial Expression feature by feature level concatenation. We perform fine-tuning on the newly merged model instead of just transferring the weights of the pre-trained face recognition networks. Experimental results show that the proposed approach outperforms the residual network baseline as well as many other state-of-the-art methods on two popular FER databases, namely CK+ and FER+. Future works include investigating the effect of face image format differences or alignment mismatch between face recognition data and FER data as well as exploring other transfer learning and multi-task learning methods for the FER task.

## Acknowledgments

This research was funded in part by the National Natural Science Foundation of China (61773413), Natural Science Foundation of Guangzhou City (201707010363), Science and Technology Program of Guangzhou City and Six talent peaks project in Jiangsu Province (JY-074).

## References

[1] H. Jung, S. Lee, J. Yim, and S. Park, "Joint fine-tuning in deep neural networks for facial expression recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2983–2991.

[2] 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.*, 2016, pp. 425–442.

[3] C. A. Corneanu, M. O. Simón, J. F. Cohn, and S. E. Guerrero, "Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 38, no. 8, pp. 1548–1568, Aug. 2016.

[4] A. Mohammadian, H. Aghaeinia, and F. Towhidkhah, "Incorporating prior knowledge from the new person into recognition of facial expression," *Signal Image Video Process.*, vol. 10, no. 2, pp. 235–242, 2016.

[5] S. Rifai, Y. Bengio, A. Courville, P. Vincent, and M. Mirza, "Disentangling factors of variation for facial expression recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 808–822.

[6] P. Werner, A. Al-Hamadi, K. Limbrecht-Ecklundt, S. Walter, S. Gruss, and H. C. Traue, "Automatic pain assessment with facial activity descriptors," *IEEE Trans. Affective Comput.*, vol. 8, no. 3, pp. 286–299, Jul.-Sep. 2017.

[7] K. Dobs, J. Schultz, I. Bülthoff, and J. L. Gardner, "Task-dependent enhancement of facial expression and identity representations in human cortex," *NeuroImage*, vol. 172, pp. 689–702, 2018.

[8] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *Proc. Autom. Speech Recognit. Understanding*, 2014, pp. 55–59.

[9] V. Gupta, P. Kenny, P. Ouellet, and T. Stafylakis, "I-vector-based speaker adaptation of deep neural networks for french broadcast audio transcription," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2014, pp. 6334–6338.

[10] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," *Adv. Neural Inf. Process. Syst.*, vol. 27, pp. 1988–1996, 2014.

[11] Y. Sun, D. Liang, X. Wang, and X. Tang, "Deepid3: Face recognition with very deep neural networks," arXiv preprint arXiv:1502.00873, 2015.

[12] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1891–1898.

[13] Y. Tang, "Deep learning using support vector machines," in *Proc. ICML Workshop Representational Learn.*, 2013.

[14] Y. Sun, X. Wang, and X. Tang, "Hybrid deep learning for face verification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 1997–2009, Oct. 2016.

[15] 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.*, 2014, pp. 1805–1812.

[16] X. Feng, M. Pietikinen, and A. Hadid, "Facial expression recognition based on local binary patterns," *Comput. Eng. Appl.*, vol. 17, no. 4, pp. 592–598, 2007.

[17] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition," *IEEE Trans. Image Process.*, vol. 11, no. 4, pp. 467–476, Apr. 2002.

[18] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 2037–2041, Dec. 2006.

[19] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[20] C. Lu and X. Tang, "Surpassing human-level face verification performance on lfw with gaussianface," *AAAI*, pp. 3811–3819, 2015.

[21] M. Xu, W. Cheng, Q. Zhao, L. Ma, and F. Xu, "Facial expression recognition based on transfer learning from deep convolutional networks," in *Proc. Int. Conf. Natural Comput.*, 2016, pp. 702–708.

[22] 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.

[23] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proc. Workshop Comput. Learn. Theory*, 1992, pp. 144–152.

[24] L. Ma and K. Khorasani, "Facial expression recognition using constructive feedforward neural networks," *IEEE Trans. Syst. Man Cybern. Part B Cybern.*, vol. 34, no. 3, pp. 1588–1595, Jun. 2004.

[25] S. J. Wang, H. L. Chen, W. J. Yan, Y. H. Chen, and X. Fu, "Face recognition and micro-expression recognition based on discriminant tensor subspace analysis plus extreme learning machine," *Neural Process. Lett.*, vol. 39, no. 1, pp. 25–43, 2014.

[26] D. Ghimire and J. Lee, "Extreme learning machine ensemble using bagging for facial expression recognition," *J. Inf. Process. Syst.*, vol. 10, no. 3, pp. 443–458, 2014.

[27] Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning," in *Proc. ACM Int. Conf. Multimodal Interaction*, 2015, pp. 435–442.

[28] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, and D. H. Lee, "Challenges in representation learning: A report on three machine learning contests," *Neural Netw.*, vol. 64, pp. 59–63, 2015.

[29] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution," in *Proc. ACM Int. Conf. Multimodal Interaction*, 2016, pp. 279–283.

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 2, 2012, Art. no. 2012.
[31] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv: 1409.1556*, 2014.
[32] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.
[33] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.
[34] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 4278–4284.
[35] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein, "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.
[36] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proc. Int. Conf. Mach. Learn.*, 2010, pp. 807–814.
[37] T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2002, Art. no. 46.
[38] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014.
[39] G. B. Huang, M. A. Mattar, H. Lee, and E. Learned-Miller, "Learning to align from scratch," in *Proc. Neural Inf. Processing Syst.*, 2012, pp. 764–772.
[40] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
[41] Y. L. Cun, B. Boser, J. S. Denker, R. E. Howard, W. Habbard, L. D. Jackel, and D. Henderson, "Handwritten digit recognition with a back-propagation network," in *Proc. Neural Inf. Process. Syst.*, 1990, pp. 396–404.
[42] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1139–1147.
[43] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3476–3483.
[44] F. D. L. Torre, W. S. Chu, X. Xiong, and F. Vicente, "Intraface," in *Proc. IEEE Int. Conf. Workshops Autom. Face Gesture Recognit.*, 2015, pp. 1–8.
[45] A. T. Lopes, E. D. Aguiar, A. F. D. Souza, and T. Oliveira-Santos, "Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order," *Pattern Recognit.*, vol. 61, pp. 610–628, 2017.