Identity- and Pose-Robust Facial Expression Recognition through Adversarial Feature Learning

Can Wang, Shangfei Wang* and Guang Liang
canwang@mail.ustc.edu.cn,sfwang@ustc.edu.cn,xshmlgy@mail.ustc.edu.cn
Key Lab of Computing and Communication Software of Anhui Province
University of Science and Technology of China
Hefei, Anhui

ABSTRACT

Existing facial expression recognition methods either focus on pose variations or identity bias, but not both simultaneously. This paper proposes an adversarial feature learning method to address both of these issues. Specifically, the proposed method consists of five components: an encoder, an expression classifier, a pose discriminator, a subject discriminator, and a generator. An encoder extracts feature representations, and an expression classifier tries to perform facial expression recognition using the extracted feature representations. The encoder and the expression classifier are trained collaboratively, so that the extracted feature representations are discriminative for expression recognition. A pose discriminator and a subject discriminator classify the pose and the subject from the extracted feature representations respectively. They are trained adversarially with the encoder. Thus, the extracted feature representations are robust to poses and subjects. A generator reconstructs facial images to further favor the feature representations. Experiments on five benchmark databases demonstrate the superiority of the proposed method to state-of-the-art work.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**;

KEYWORDS

Identity-Robust, Pose-Robust, Facial Expression Recognition, Adversary Features Learning

ACM Reference Format:
Can Wang, Shangfei Wang* and Guang Liang. 2019. Identity- and Pose-Robust Facial Expression Recognition through Adversarial Feature Learning. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19), October 21–25, 2019, Nice, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3343031.3350872

*Dr. Shangfei Wang is the corresponding author.

MM '19, October 21–25, 2019, Nice, France

ACM ISBN 978-1-4503-6889-6/19/10…$15.00
https://doi.org/10.1145/3343031.3350872

1 INTRODUCTION

Facial expression recognition is a research hotspot in computer vision. It can be useful in many applications, such as human-computer interaction, social interaction analysis, and medical treatments. However, most works on facial expression recognition ignore pose variations, so they are less applicable to real-world situations. Subject identity bias is also common in unconstrained facial expression scenarios [9]. Researchers seek to remedy this failure by focusing on identity- and pose-robust facial expression recognition.

Due to head pose and non-rigid facial changes according to expression, the primary challenge for pose-robust facial expression recognition is to decouple the rigid facial changes [4]. To date, only a few pose-robust facial expression recognition methods have been proposed. These methods can be divided into three categories based on how they deal with variations in head poses. The most intuitive method would be to train a single classifier for facial images with multiple poses [24]. However, this method requires large amounts of training data with varying expressions and poses, which is often unavailable. Another intuitive method is to perform pose normalization prior to facial expression recognition [7, 8]. This method has difficulty modeling relationships between different poses. Due to these drawbacks, most recent works focus on pose-invariant features [4, 13, 19, 26]. A pose-invariant features method aims to find the feature representations that allow for better expression recognition while alleviating pose variations. Compared to the first two types of methods, the pose-invariant features method succeeds in capturing correlations between different poses and achieves better performance. However, current pose-invariant features methods fail to cope with large pose variations, thus decreasing recognition performance.

Identity bias is another challenge for facial expression recognition. Existing methods can be divided into two categories. The first tries to learn identity-invariant features [14, 20, 25], and the second tries to reduce identity bias by learning person-specific models [1, 21]. The identity-invariant features method adds extra constraints for identity-related variations in facial expression recognition [14, 20, 25]. However, these constraints depend on identity-related image pairs that are not always accessible in real-world situations. The person-specific method trains the model for each subject, but it is unfeasible due to a lack of annotated data. Most works tend

to apply domain adaptation to overcome this. Domain adaptation improves learning in the target domain (i.e., an unseen subject) by extracting knowledge from source domains (i.e., many subjects in the training data) [17]. However, limited data on the target subject makes it difficult to adapt the distribution of the target subject to the source subjects. In addition, it is time-consuming and laborious to train different models for different subjects.

Previous facial expression recognition methods either focus on pose variations or identity bias; there is no work that considers both at the same time. To this end, we propose a novel feature representation method that uses adversarial learning to overcome the challenges of both pose variations and identity bias. As shown in Figure 1, the framework of the proposed method consists of five components. Specifically, an encoder extracts feature representations, an expression classifier tries to perform facial expression recognition, a pose discriminator and a subject discriminator classify the pose and the subject respectively for the representations, and a generator reconstructs facial images to further favor the feature representations. Through joint training and adversarial learning, the feature representations perform superior expression recognition without differentiating poses or subjects. Thus the feature representations can preserve expression information but avoid pose and subject variations.

2 RELATED WORK

2.1 Pose-Robust FER

As mentioned before, pose-robust facial expression recognition methods can be divided into three categories: (1) methods that train a single classifier for facial images with multiple poses [24]; (2) methods that perform pose normalization prior to facial expression recognition [7, 8]; (3) and methods that try to find pose-robust feature representations that allow for better expression recognition while alleviating pose variations [4, 13, 19, 26].

For the single classifier method, Zhang *et al.* [24] proposed a joint pose and expression modeling method (JPEM) for facial expression recognition. By coding the pose and expression information into the generative adversarial network (GAN), JPEM can synthesize face images with different expressions under arbitrary poses to enlarge and enrich the training set. The performance of the single expression classifier is improved by the large number of generated face images. The method, however, loses sight of variations in large poses and relations between different poses, so the learned features of the single classifier are rarely robust enough for pose-robust facial expression recognition. In contrast, the proposed method is able to capture the relations between different poses to learn a pose-robust feature representation.

For the pose normalization method, Lai *et al.* [8] proposed an emotion-preserving representation learning method (EPRL) via GAN for multi-view facial expression recognition. Based on GAN, this method converts a profile image into a frontal image while preserving the expression content. During image conversion, an expression classifier is learned for expression recognition. However, this method degenerates as pose difference increases, since the synthesized face image contains more significant stretching artifacts. Unlike Lai *et al.*'s work which considers pose normalization at the pixel level, Jampour *et al.* [7] proposed a kernel-based pose specific non-linear mapping method (KPSNM) to map the features of profile images to those of frontal images with the same facial expression. However, KPSNM requires paired images during training. Compared to these methods, the proposed method aims to learn an optimized feature representation and therefore avoids stretching artifacts and image pairs.

For the pose-robust feature representations method, Wu *et al.* [19] proposed a locality-constrained linear coding based bi-layer (LLCBL) model. In this model, they first estimated head pose and adopted bag-of-features to encode view-specific local features. They then used a classifier to perform facial expression recognition. Eleftheriadis *et al.* [4] proposed a discriminative shared Gaussian process latent variable (DS-GPLVM) model for pose-robust facial expression recognition. In this model, a discriminative manifold shared by multiple views of a facial expression is learned and then facial expression classification is performed in the expression manifold. Zhang *et al.* [26] proposed a deep neural network-driven feature learning method (DNND) for multi-view facial expression recognition. The SIFT descriptors are first extracted from the detected landmarks. Then, based on the structure of low-level features, the projection layer and convolutional layer are designed to adaptively learn spatially discriminative robust high-level features for facial expression recognition. Mao *et al.* [13] proposed a pose-based hierarchical Bayesian theme model (HBTM) to jointly learn the intermediate face representation and the expression classifier. This model can learn the relationships among different poses by sharing the pool of features. However, for an unseen sample, pose estimation is indispensable for LLCBL, DS-GPLVM and HBTM. Any errors in pose estimation can then propagate to expression recognition. DNND depends on SIFT descriptions, which would not perform well when the angle is larger than 30° [18]. By comparison, the proposed method avoids these shortcomings by learning high-level pose-agnostic feature representations, augmenting the performance of multi-pose facial expression recognition.

2.2 Identity-Robust FER

As mentioned before, identity-roust facial expression recognition methods can be divided into two categories: the identity-invariant features method and the person-specific method.

For the identity-invariant features method, Meng *et al.* [14] proposed an identity-aware convolutional neural network (I-ACNN). In this method, an identity-sensitive contrastive loss is proposed to learn identity-related information from identity labels to achieve identity-robust expression recognition. However, the contrastive loss suffers from drastic data expansion when constructing image pairs from the training set, while the proposed method does not depend on these image pairs for training. Like Meng *et al.*, Liu *et al.* [11] proposed

an adaptive deep metric learning method (ADML). In this method, they proposed a generalized adaptive triplet loss function together with identity-aware hard-negative mining and online positive mining scheme to perform identity-robust facial expression recognition. However, this triplet loss also depends on identity-related image pairs and perform poorly in the case of limited identities [10]. Instead of using the identity-aware loss, Yang *et al.* [20] proposed a de-expression residue learning method (DeRL) based on conditional GANs (cGANS). DeRL tries to explore expression information, which is filtered out during the de-expression process but still embedded in the generator. The model directly extracted this information from the generator to mitigate the influence of subject variations. DeRL assumes that a face expression consists of both a neutral component and an expressive component. This assumption, however, does not always hold true in real-world situations. Compared to these methods, the proposed method is not limited to image pairs and does not rely on any assumptions, thus achieving better performance.

For the person-specific method, Yang *et al.* [21] proposed a two-step identity-adaptive GAN-based model (IA-gen). It used cGANs to generate images of the same subject with different expressions. The person-specific classifier was trained to conduct facial expression recognition. Similarly, Wang *et al.* [17] proposed a generative adversarial recognition network (GARN) to generate a large amount of facial images that are similar to the target domain but retain AU patterns of the source domain. This enriches and enlarges the target dataset. However, these two GAN-based methods are overly dependent on generated images for classifier training; unrealistic generated images would add noise to the classifier, resulting in poorer facial expression recognition. Unlike Yang *et al.* and Wang *et al.*'s works, which try to enlarge the training set for the target domain, Chu *et al.* [1] proposed a selective transfer machine method (STM) to simultaneously optimize sampling weights and classifier parameters, thus adapting information from source subjects to the target subject. However, they assumed that the data distribution of a new subject can be approximated by weighting the data distribution of other subjects in the training set. This assumption may not hold true in real-life scenarios. In contrast, the proposed method focuses on identity bias at the feature level and does not depend on any assumptions.

Existing facial expression recognition methods either focus on pose variations or identity bias, but not both simultaneously. The proposed method addresses both of these issues. By exploiting adversarial learning and learning all networks jointly, the proposed method learns a feature representation that preserves expression content but avoids subject and pose variations, thus boosting facial expression recognition.

3 PROBLEM STATEMENT

Let $\mathcal{D} = \{x_i, y_i, v_i, s_i\}_{i=1}^{N}$ denote the training set, where $x_i \in X$ represents a training facial image, $y_i \in \{0,1\}^K$ represents the ground truth expression label, $v_i \in \{0,1\}^M$ represents the pose label, and $s_i \in \{0,1\}^W$ represents the subject label.

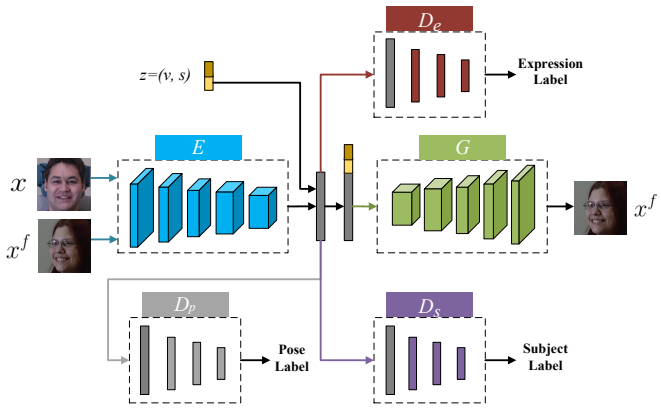

Figure 1: The structure of the proposed method. It consists of an encoder E, an expression classifier D_e, a pose discriminator D_p, a subject discriminator D_s, and a generator G.

The goal is to learn an encoder E to extract identity and pose-robust feature representations, and simultaneously learn an expression classifier D_e to predict expression labels for the feature representations. Let X^t denote the unlabeled testing images. Given an unseen sample $x^t \in X^t$, we can apply the final classifier $R = E \circ D_e$ to predict its expression label.

4 METHODOLOGY

Figure-1 illustrates the framework of our proposed approach. The framework consists of five components, including the encoder E, the generator G, the expression classifier D_e, the pose classifier D_p, and the subject classifier D_s. The encoder E maps a training sample $x \in X$ to its latent feature representation $E(x)$. Then, the expression classifier D_e tries to predict its expression label $D_e(E(x))$, the pose classifier D_p tries to predict its pose label $D_p(E(x))$, and the subject classifier D_s tries to predict its subject label $D_s(E(x))$. The generator G takes the latent feature representation $E(x)$ and a class embedding z as inputs, and outputs a synthesized facial image $x^f = G(E(x), z)$, where $z = (v, s)$ is the concatenation of pose indicator v and subject indicator s. x^f is expected to share the same expression content as the input x. Therefore, $D_e(E(x^f))$, the predicted expression label for x^f, should be the same as that of x. As well as having the pose, subject, and expression labels of x^f, E also takes x^f as an input. The goal is to encode an image to a feature representation $E(x)$ that preserves the facial expression information while removing pose and subject variations. The encoder E should extract a feature that is easily recognized by the expression classifier D_e but difficult for the pose classifier D_p and the subject classifier D_s to recognize. After training, we can compute the final classifier $R = E \circ D_e$.

4.1 Adversarial Feature Learning

The expression classifier D_e is applied to the latent feature representation $E(x)$ to encourage the encoder E to extract a

feature representation that preserves expression information. E and D_e work cooperatively to minimize the K-expression recognition loss $\mathcal{L}_e$, which is a cross-entropy loss between ground truth expression label y and prediction $D_e(E(x))$, defined as Equation-1.

$$\mathcal{L}_e(E, D_e) = -\mathbb{E}_{(x,y)} \sum_{k=1}^{K} \mathbb{1}_{[k=y]} \log(D_e(E(x))) \quad (1)$$

The encoded feature representation $E(x)$ is also used as the input to the pose classifier D_p and the subject classifier D_s. The goal is to learn $E(x)$ that is least likely to be distinguished by D_p and D_s according to the pose and subject labels. Adversarial learning is introduced to accomplish this. Specifically, E and D_p play an adversarial game in which E tries to minimize the divergence of feature distributions for different pose classes so that D_p fails to correctly recognize the pose of a sample. Similarly, E and D_s also play an adversarial game in which E tries to minimize the divergence of feature distributions for different subject classes so that D_s has difficulty classifying the subject of a sample. Here are the adversarial multi-task training objectives:

$$\begin{aligned} &\min_{E,D_e} \mathcal{L}_e(E, D_e) \\ &\min_{D_p} \max_{E} \mathcal{L}_p(E, D_p) \\ &\min_{D_s} \max_{E} \mathcal{L}_s(E, D_s) \end{aligned} \quad (2)$$

where $\mathcal{L}_p$ represents the pose classification loss and $\mathcal{L}_s$ represents the subject classification loss. The M-pose classification loss on a training sample x with pose label v can be defined as the cross-entropy between predicted class distribution $D_p(E(x))$ and ground truth pose label v:

$$\mathcal{L}_p(E, D_p) = -\mathbb{E}_{(x,v)} \sum_{m=1}^{M} \mathbb{1}_{[v=m]} \log(D_p(E(x))) \quad (3)$$

While it is not always possible to have accurate pose labels for a facial data set, it is easy to determine if a sample is frontal or not. Therefore, in this situation, we can convert the multi-class adversarial loss $\mathcal{L}_p(E, D_p)$ to a "real/fake" version. Specifically, we treat the frontal pose as a positive class set Z^+ and other non-frontal poses as a negative class set Z^-. D_p is applied to discriminate whether $E(x)$ is frontal or not. Then the adversarial loss $\mathcal{L}_p(E, D_p)$ can be defined as Equation-4.

$$\begin{aligned} \mathcal{L}_p(E, D_p) = &-\mathbb{E}_{(x,v)} \mathbb{1}_{[v \in Z^+]} \log(D_p(E(x))) \\ &-\mathbb{E}_{(x,v)} \mathbb{1}_{[v \in Z^-]} \log(1 - D_p(E(x))) \end{aligned} \quad (4)$$

Similarly, the W-subject classification loss on a training sample x with subject label s can be defined as the cross-entropy between predicted class distribution $D_s(E(x))$ and ground truth pose label s:

$$\mathcal{L}_s(E, D_s) = -\mathbb{E}_{(x,s)} \sum_{w=1}^{W} \mathbb{1}_{[s=w]} \log(D_s(E(x))) \quad (5)$$

The goal is to optimize the following minmax objective, which is defined as Equation-6:

$$\min_{E,D_e} \max_{D_p,D_s} \mathcal{L}_e(E, D_e) - \alpha \mathcal{L}_p(E, D_p) - \beta \mathcal{L}_s(E, D_s) \quad (6)$$

where α and β are weights that control the interaction of the losses.

4.2 Reconstruction Learning

By optimizing the expression classification loss and the adversarial loss, the learned feature $E(x)$ can retain expression information and remove pose and subject variations. To further ensure $E(x)$ contains a full description of the expression information, a generator branch produces a facial image $x^f = G(E(x), z)$ that shares the same expression as the input image given an arbitrary pose and subject indicator $z = (v, s)$. Specifically, for a training sample $x_i \in X$, the generated image $x_i^f = G(E(x_i), z)$ should be similar to $x_j \in X$ in that it has the same pose and subject indicator z to x_i^f and the same expression label to x_i. An $\mathcal{L}_2$ similarity loss is used to guide the training for the generator branch.

$$\mathcal{L}_{sim}(E, G) = -\mathbb{E}_{(x_i,z)} \sum_{i,j} \|G(E(x_i), z) - x_j\|_2 \quad (7)$$

When the z of x_i^f equals to that of x_i, then $x_j = x_i$ and G will try to reconstruct the original input x_i. Also, to enforce smooth spatial color transformation for the generated face image, we perform a regularization over x_i^f:

$$\begin{aligned} \mathcal{L}_{smo}(E, G) = -\mathbb{E}_{(x_i,z)} \sum_{h,w}^{H,W} [&(G(E(x_i), z)_{h+1,w} \\ -G(E(x_i), z)_{h,w})^2 + (G(E(x_i), z)_{h,w+1} - G(E(x_i), z)_{h,w})^2] \end{aligned} \quad (8)$$

where H and W represent the height and the width of a facial image, respectively.

x_i^f is expected to have the same expression content as the original input x_i, so the classifier $R = E \circ D_e$ can be used to predict its expression label. The expression recognition loss for x_i^f, defined as Equation-9, can then be obtained.

$$\mathcal{L}_{ef}(E, D_e) = -\mathbb{E}_{(x_i^f, y_i)} \sum_{k=1}^{K} \mathbb{1}_{[k=y_i]} \log(D_e(E(x_i^f))) \quad (9)$$

where y_i is the expression label of x_i. Similarly, we can define the pose classification loss $\mathcal{L}_{pf}(E, D_p)$ and subject classification loss $\mathcal{L}_{sf}(E, D_s)$ for x_i^f, that are same to Equation-3 and Equation-5, respectively. Through G, more samples can be generated to enlarge the original training set. The generated training set can be defined as $\mathcal{D}^f = \left\{x_i^f, y_i, v, s\right\}_{i=1}^{F}$, so the actual training set is $\Omega = \mathcal{D} \cup \mathcal{D}^f$.

4.3 Overall Learning

The joint objective function with the loss terms can be formulated as follows:

$$\begin{aligned}\min_{E,D_e}\max_{D_p,D_s} \mathcal{L}_e(E,D_e) - \alpha\mathcal{L}_p(E,D_p) - \beta\mathcal{L}_s(E,D_s) \\ +\gamma\mathcal{L}_{ef}(E,D_e) - \alpha_f\mathcal{L}_{pf}(E,D_p) - \beta_f\mathcal{L}_{sf}(E,D_s) \\ +\lambda\mathcal{L}_{sim}(E,G) + \delta\mathcal{L}_{smo}(E,G)\end{aligned} \tag{10}$$

where α, β, γ, α_f, β_f, λ and δ are weighted coefficients. In our experiments, $\gamma = 1, \alpha_f = \alpha$, and $\beta_f = \beta$. The minmax optimization problem of Equation-10 can be solved by applying an iterative algorithm between two steps. During the first step, D_s and D_p are fixed and E, D_e and G are updated; during the second step, E, D_e and G are fixed and D_s and D_p are updated. The process repeats until convergence. This is shown in Algorithm-1.

Algorithm 1 The learning algorithm of the proposed method.

Input: The training set: $\mathcal{D} = \{x_i, y_i, v_i, s_i\}_{i=1}^N$;
Training steps: K_1, K_2 and K_3;
The threshold to start using the faked images: S;
The pose and subject indicators: $\mathcal{Z}$;
Batch size m and the number of training epochs T.
Output: The final classifier $R = E \circ D_e$.

for $t = 1$ to T **do**
  for $k = 1$ to K_1 **do**
    Randomly sample $\{x_i, y_i, v_i, s_i\}_{i=1}^m \sim \mathcal{D}$.
    Update E and D_e jointly: $\bigtriangledown\theta_{E\cup D_e} := \frac{\partial\mathcal{L}_e(E,D_e)}{\partial\theta_{E\cup D_e}}$.
    Update E: $\bigtriangledown\theta_E := -(\frac{\partial\mathcal{L}_p(E,D_p)}{\partial\theta_E} + \frac{\partial\mathcal{L}_s(E,D_s)}{\partial\theta_E})$.
  end for
  for $k = 1$ to K_2 **do**
    Randomly sample indicators $\{z_i\}_{i=1}^m \sim \mathcal{Z}$.
    Update E and G jointly:
    $\bigtriangledown\theta_{E\cup G} := \frac{\partial\mathcal{L}_{sim}(E,G)}{\partial\theta_{E\cup G}} + \frac{\partial\mathcal{L}_{smo}(E,G)}{\partial\theta_{E\cup G}}$.
    if $t > S$ **then**
      Add faked images $\mathcal{D}^f = \left\{x_i^f, y_i, z_i\right\}_{i=1}^m$ to $\mathcal{D}$:
      $\mathcal{D} = \mathcal{D} \cup \mathcal{D}^f$.
    end if
  end for
  for $k = 1$ to K_3 **do**
    Randomly sample $\{x_i, y_i, v_i, s_i\}_{i=1}^m \sim \mathcal{D}$.
    Update D_p: $\bigtriangledown\theta_{D_p} := \frac{\partial\mathcal{L}_p(E,D_p)}{\partial\theta_{D_p}}$.
    Update D_s: $\bigtriangledown\theta_{D_s} := \frac{\partial\mathcal{L}_s(E,D_p)}{\partial\theta_{D_s}}$.
  end for
end for

5 EXPERIMENTS

We conduct experiments on five publicly available facial expression databases: Multi-PIE [6], BU-3DFE [22], SFEW [2], AffectNet [15] and FER2013 [5].

5.1 Experimental Conditions

The Multi-PIE database contains 755,370 images from 337 subjects under 15 viewpoints. Each facial image is labeled with one of six expressions: disgust, neutral, scream, smile, squint, or surprise. In our experiments, nine different poses ($\pm30°, \pm15°, 0°, 45°, 60°, 75°$, and $90°$ pan angles) of 100 subjects are used for a total of 13,779 facial images. Subjects are randomly divided into a training set of 80 subjects and a testing set of 20 subjects.

The BU-3DFE database is a synthetic database. This database contains 100 subjects including 56 females and 44 males. For each subject, six universal facial expressions (anger, disgust, fear, happiness, sadness and surprise) are elicited by various manners with multiple intensities. In our experiments, $\pm45°, \pm30°, \pm15°, 0°, 60°$, and $90°$ pan angles of 100 subjects are used, yielding 21,600 facial images. Subjects are randomly divided into a training set of 80 subjects and a testing set of 20 subjects.

The SFEW database is created by selecting static frames from the AFEW database after computing key frames based on facial point clustering. The most commonly used version is SFEW 2.0. It has been divided into three sets: Train (958 samples), Val (436 samples) and Test (372 samples). Each image is assigned to one of seven expression categories, i.e., anger, disgust, fear, neutral, happiness, sadness, and surprise.

The AffectNet database contains 420,299 images annotated with 11 categories. In our experiments, images from six categories (anger, disgust, fear, neutral, happiness, sadness, and surprise) are selected and 283,901 images are obtained. We use 3,500 images as the testing set. The experimental setting is the same as IPA2LT [23]. Because AffectNet lacks pose annotations, we train a ResNet on the BU-3DFE and Multi-PIE databases to annotate AffectNet with one of two views (frontal or non-frontal).

The FER2013 database consists of 28,709 samples. The annotation is the same as the SFEW. As FER2013 also lacks pose annotations, we use the same ResNet to annotate FER2013 with one of two views (frontal or non-frontal).

We perform within-database experiments on the Multi-PIE, BU-3DFE, and AffectNet databases. Specially, following Wu *et al.* [19] and Zhang *et al.* [24]'s works, we employ two experimental settings on the Multi-PIE and BU-3DFE databases: $(0°, 90°)$ and $(-30°, 30°)$ pan angles on the Multi-PIE data set and $(0°, 90°)$ and $(-45°, 45°)$ pan angles on the BU-3DFE data set.

Due to limited data on the SFEW database, we perform three cross-database experiments: training on the BU-3DFE data set and testing on the SFEW data set; training on the AffectNet data set and testing on the SFEW data set; training on the FER2013 dataset and fine-turning on the SFEW dataset. For the first two cross-database experiments, the Train and Val sets of SFEW are used as the testing set. For the last cross-database experiment, we fine-tune the model on the Train set of SFEW and test it on the Val set of SFEW, as ADML [11] and IACNN [14] did.

We compare our method to four other techniques. The first, named IPFR$_{D_e}$, trains E and D_e without training D_p, D_s and G. The second method, IPFR$_{D_e+D_p}$, trains E, D_e and D_p without training D_s and G. The third method, named IPFR$_{D_e+D_s}$, trains E, D_e and D_s without training D_p and G. The last method, IPFR$_{D_e+D_p+D_s}$, trains E, D_e, D_p and

Table 1: Experimental results on the Multi-PIE, BU-3DFE, AffectNet, and SFEW databases.

Method	Multi-PIE		BU-3DFE		AffectNet↦SFEW	BU-3DFE↦SFEW	FER2013↦SFEW	AffectNet
	$(0°, 90°)$	$(-30°, 30°)$	$(0°, 90°)$	$(-45°, 45°)$				
IPFR_{D_e}	80.5	84.7	72.2	76.0	54.3	21.3	48.7	53.9
$\text{IPFR}_{D_e+D_p}$	86.1	89.5	78.1	83.6	57.1	26.9	55.1	57.4
$\text{IPFR}_{D_e+D_s}$	85.6	88.0	77.4	82.1	-	26.5	-	-
$\text{IPFR}_{D_e+D_p+D_s}$	87.8	91.3	80.9	84.0	-	27.3	-	-
IPFR	88.4	92.6	81.8	85.1	-	28.1	-	-

D_s without training G. Because the AffectNet and Fer2013 databases lack subject annotations, we are unable to compare our method to $\text{IPFR}_{D_e+D_s}$ and $\text{IPFR}_{D_e+D_p+D_s}$ on them.

5.2 Implementation Details

For images on all databases, face parts are recognized using Opencv, and then cropped and resized to 128×128 pixels. All networks are implemented by Pytorch, and the encoder has an input layer, an Inception-ResNet-V2 block [16], four down-sampling layers, and three residual blocks. The generator has an input layer, four up-sampling layers, and an output layer. Other networks consist of fully connected layers. After flattening the output of the encoder, we can obtain the feature representation $E(x) \in \mathbb{R}^{512}$. The pose and subject information are concatenated to the output feature maps of the encoder and used as the input for the generator. For all experiments, the batch size is set to 32. Due to the vast number of combinations of pose and subject information, the generator is optimized six steps for each iteration, i.e., the value of K_2 in Algorithm-1 is six. Moreover, $K_1 = 1$ and $K_3 = 1$ in Algorithm-1. It takes several iterations before the generator can generate realistic enough data sets D^f to enlarge the training set. Therefore, in our experiments, D^f is not used to train the networks for the first 20 epochs, i.e., the value of S in Algorithm-1 is 20. Coefficients are decided by the validation set. For all experiments, we adopt average results on five randomly selection times and adopt accuracy classification score (ACC) as the performance metric.

5.3 Experimental Results and Analyses

The experimental results of expression recognition are shown in Table 1. From this table, we observe the following:

First, $\text{IPFR}_{D_e+D_p}$ and $\text{IPFR}_{D_e+D_s}$ achieve better results than the baseline method that does not consider pose or subject variations (IPFR_{D_e}). Specifically, $\text{IPFR}_{D_e+D_p}$ outperforms IPFR_{D_e} by 7.6% on the BU-3DFE database with $(-45°, 45°)$ pan angles, and $\text{IPFR}_{D_e+D_p}$ outperforms IPFR_{D_e} by 3.5% on the AffectNet database, and $\text{IPFR}_{D_e+D_p}$ outperforms IPFR_{D_e} by 8.4% in the cross-database experiment FER2013↦SFEW. These results indicate that training the pose discriminator branch can alleviate the pose variations, resulting in better performance. Similarly, $\text{IPFR}_{D_e+D_s}$ outperforms IPFR_{D_e} by 6.1% on the BU-3DFE database with $(-45°, 45°)$ pan angles and by 5.2% in the cross-database experiment BU-3DFE↦SFEW. This indicates that training the subject discriminator branch alleviates the identity variations, resulting in better performance.

Secondly, the method $\text{IPFR}_{D_e+D_p+D_s}$ outperforms IPFR_{D_e}, $\text{IPFR}_{D_e+D_p}$, and $\text{IPFR}_{D_e+D_s}$. For example, $\text{IPFR}_{D_e+D_p+D_s}$ outperforms $\text{IPFR}_{D_e+D_p}$ and $\text{IPFR}_{D_e+D_s}$ by 2.8% and 3.5% respectively on the Multi-PIE database. It also outperforms IPFR_{D_e} by 6.6% on the Multi-PIE database with $(-30°, 30°)$ pan angles. Similar results are obtained on the cross-database experiment BU-3DFE↦SFEW. These results suggest that training E, D_e, D_s, and D_p collaboratively learns a pose-robust and identity-invariant feature representation that achieves superior performance.

Thirdly, the proposed method IPFR achieves best performance. For example, IPFR outperforms $\text{IPFR}_{D_e+D_p+D_s}$ by 1.3% on the Multi-PIE database with $(-30°, 30°)$ pan angles and it outperforms $\text{IPFR}_{D_e+D_p+D_s}$ by 1.1% on the BU-3DFE database with $(-45°, 45°)$ pan angles. It also achieves superior performance on the cross-database experiments. For example, for BU-3DFE↦SFEW, IPFR shows a 0.8% improvement compared to $\text{IPFR}_{D_e+D_p+D_s}$ and shows a 1.2% improvement compared to $\text{IPFR}_{D_e+D_p}$. These results suggest that the generator branch can favor the learned feature representations. To conclude, after combining D_e, D_s, D_p and G, the proposed method can learn a better optimized feature representation, leading to improved recognition.

5.4 Feature Analysis

To further demonstrate how effectively our method learns accurate representations for expression recognition, we use t-SNE [12] to visualize the original facial images and the features learned by IPFR_{D_e} and IPFR. Specifically, we randomly select 8 subjects from the testing set of the Multi-PIE data set, in which there are 55×8 samples. As shown in Figure-2-(a) on the left, the original raw images are randomly distributed and expression cannot be distinguished. Samples of the same pose but different expressions tend to converge into a cluster (e.g, two close-ups enclosed by the red rectangles). This indicates that pose variations can be very influential for facial expression recognition. In contrast, in Figure-2-(b) on the left, samples of the same expression tend to cluster. However, there are still many samples of the same pose but different expressions that are close to each other (e.g, four close-ups enclosed by the red rectangles), which indicates that the samples do not distinguish one expression from another. That is one reason for the baseline's poor performance in facial expression recognition. Compared to the original images and the features learned by the baseline IPFR_{D_e}, the features learned by IPFR are clearly expression-distinguished, as shown in Figure-2-(c) on the left. In this figure, samples of different expressions merge into different clusters, while samples of the same pose are randomly distributed within the clusters. This demonstrates that the proposed IPFR can

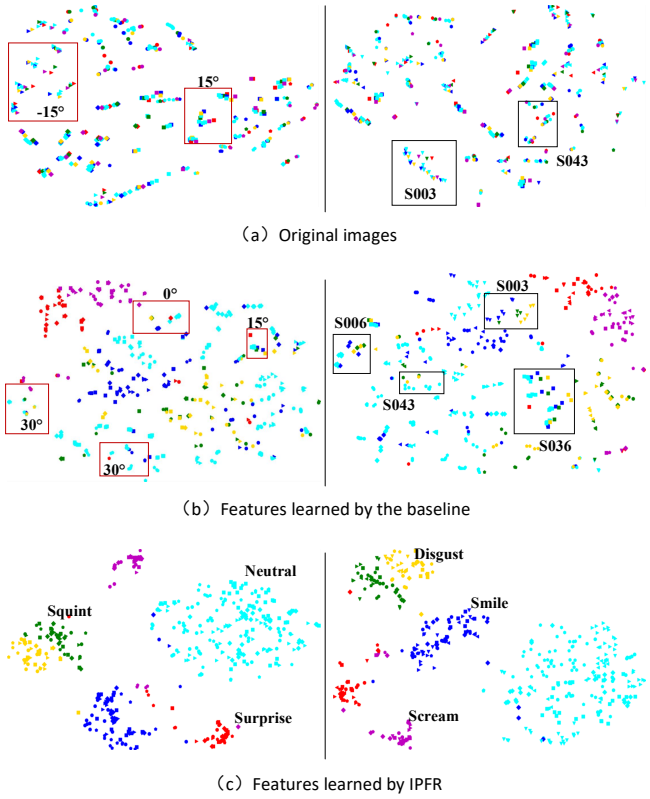

Figure 2: A visualization study. Different colors represent different expressions. In the left column, different shapes represent different poses, while in the right column, different shapes represent different subjects. (a)the original raw images; (b)the features leaned by IPFR$_{D_e}$; (c)the features learned by IPFR.

learn a pose-robust feature representation for facial expression recognition. A similar manifestation about subject bias is illustrated in the right column of the Figure-2. To conclude, the proposed method IPFR has the ability to learn a pose-robust and identity-invariant feature representation that allows for better facial expression recognition.

5.5 Generator Analysis

The generator G is trained to generate face images with the same expression content as the input face image, controlled by the pose and subject indicator z. Figure-3 shows some generated images. For example, on the Multi-PIE database, given an input image with expression smile and a pose and subject indicator of 30° and S004, the corresponding image can be generated. This indicates that the learned feature representation can preserve most of the expression-relevant information. Adding pose and subject information to this feature representation allows G to reconstruct a corresponding image that is realistic enough to enlarge the training data set. Realistic images can be synthesized for large poses,

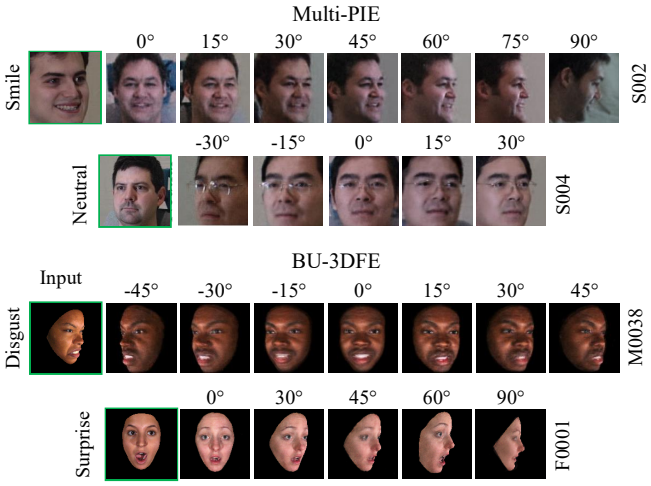

Figure 3: Generated images on the Multi-PIE and BU-3DFE databases. S002, S004, M0038 and F0001 represent different subjects. The image in green is the input.

Table 2: Comparison with state-of-the-art methods on the Multi-PIE and BU-3DFE database.

Methods	Multi-PIE			BU-3DFE		
	Poses	Pose Pan	Acc	Poses	Pose Pan	Acc
KPSNM	7	$(0°, 90°)$	83.1	5	$(0°, 90°)$	78.8
LLCBL	7	$(0°, 90°)$	86.3	5	$(0°, 90°)$	74.6
DNND	7	$(0°, 90°)$	85.2	5	$(0°, 90°)$	80.1
EPRL	7	$(0°, 90°)$	87.1	5	$(0°, 90°)$	73.1
GARN	7	$(0°, 90°)$	80.0	5	$(0°, 90°)$	76.8
IPFR	7	$(0°, 90°)$	**88.4**	5	$(0°, 90°)$	**81.8**
DS-GPLVM	5	$(-30°, 30°)$	90.6	-	-	-
HBTM	5	$(-30°, 30°)$	90.2	7	$(-45°, 45°)$	79.1
JPEM	5	$(-30°, 30°)$	91.8	7	$(-45°, 45°)$	81.2
GARN	5	$(-30°, 30°)$	83.0	7	$(-45°, 45°)$	78.3
DeRL	-	-	-	1	$0°$	84.2
IA-gen	-	-	-	1	$0°$	76.8
IPFR	5	$(-30°, 30°)$	**92.6**	7	$(-45°, 45°)$	**85.1**

demonstrating the superiority of the proposed method in the case of large pose variations.

5.6 Comparison to Related Works

Approaches for comparison are divided into two categories: pose-robust methods and identity-robust methods. For pose-robust methods, we compare the proposed method to JPEM [24], LLCBL [19], EPRL [8], KPSNM [7], DS-GPLVM [4], DNND [26] and HBTM [13]. For identity-robust methods, we compare the proposed method to IACNN [14], ADML [11], DeRL [20], IA-gen [21] and GARN [17]. STM [1] does not provide results on the aforementioned databases and the codes are not available; therefore, we do not compare to this method. To further validate the effectiveness of the proposed method on the in-the-wild database, we also compare the proposed method to Zeng *et al.*'s inconsistent pseudo annotations to latent truth (IPA2LT) framework [23], which does not explicitly address the various poses. In this framework, they proposed an end-to-end trainable network LTNet embedded

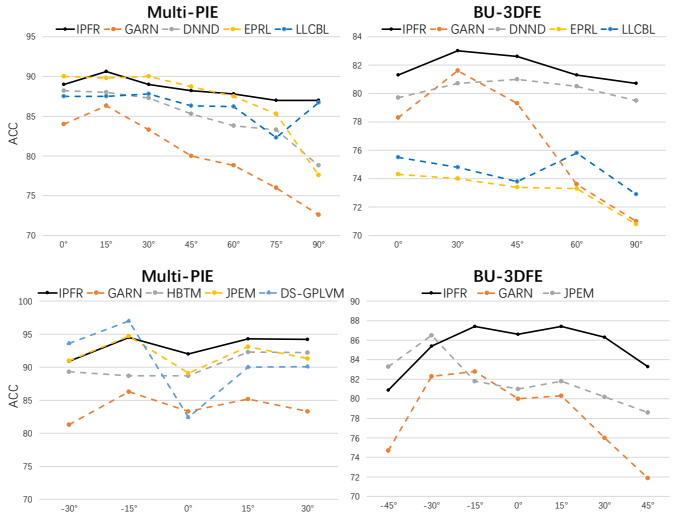

Figure 4: Comparison to related works with poses.

with a scheme of discovering the latent truth from multiple inconsistent labels and the input images.

Table 2 and 3 report experiment results of related works.

For the single classifier method, the proposed method outperforms JPEM. JPEM fails to emphasize variations in poses, causing a single classifier that is not robust enough for pose-robust facial expression recognition. In contrast, the proposed method leverages the relations between different poses, achieving better performance.

The proposed method is also superior to pose normalization methods (KPSNM and EPRL). EPRL tries to convert profiles into frontal images, so the faked images contain more significant stretching artifacts. The proposed method learns an optimized feature representation and avoids these stretching artifacts. KPSNM relies on image pairs for training. The number of image pairs is limited and insufficient for training.

As Figure-4 shows, ACC values of related works drop rapidly when large poses are present, while the proposed method shows stable results. In the vast majority of cases, all works obtain best results when the head pose is $\pm 15°$, a finding which is in agreement with past research [3]. To conclude, Figure-4 demonstrates that the learned feature representation of the proposed method is more robust to pose variations than other state-of-the-art methods.

The proposed method also obtains superior results compared to pose-robust feature representations methods (DS-GPLVM, DNND, and HBTM). For an unseen sample, pose estimation is indispensable for DS-GPLVM and HBTM. Errors in pose estimation then propagate to expression recognition. DNND depends on SIFT descriptions, which do not perform well when the angle is greater than $30°$ [18]. By comparison, the proposed method does not estimate poses and can learn a high-level pose-agnostic feature representation, improving the performance of multi-pose facial expression recognition.

Furthermore, the proposed method achieves better performance than identity-invariant features methods (IACNN, ADML, and DeRL). The contrastive loss of IACNN suffers from drastic data expansion when constructing image pairs from the training set. The triplet loss of ADML does not perform well in the case of limited identities [10]. DeRL assumes that a face expression consists of both a neutral component and an expressive component, which may not hold true in real-life situations. Weaknesses of these methods account for their inferior results in comparison to the proposed method.

Table 3: Comparison with state-of-the-art methods on the in-the-wild databases.

Methods	Train	Test	Emotions							Avg.
			AN	DI	FE	HA	NE	SA	SU	
DS-GPLVM	BU-3DFE	SFEW	25.9	28.2	17.2	43.0	14.0	33.3	11.0	24.7
JPEM	BU-3DFE	SFEW	30.9	22.0	19.6	50.9	19.2	28.0	15.5	26.6
GARN	BU-3DFE	SFEW	25.9	23.8	16.2	38.1	12.6	26.8	12.0	22.2
IPFR	BU-3DFE	SFEW	27.3	28.9	24.3	38.7	19.7	26.2	31.4	**28.1**
ADML	FER2013	SFEW	66.2	4.4	6.4	87.8	57.5	40.4	73.3	54.2
IACNN	FER2013	SFEW	70.7	0	8.9	70.4	60.3	58.8	28.9	54.3
IPFR	FER2013	SFEW	73.7	8.9	8.9	89.0	69.9	61.8	47.1	**55.1**
IPA2LT	AffectNet	SFEW	-	-	-	-	-	-	-	55.6
IPFR	AffectNet	SFEW	68.3	28.7	32.0	90.2	58.6	53.3	68.8	**57.1**
IPA2LT	AffectNet	AffectNet	-	-	-	-	-	-	-	56.5
IPFR	AffectNet	AffectNet	58.6	19.0	24.0	80.6	44.4	55.0	63.0	**57.4**

The proposed method also performs better than the person-specific methods (IA-gen and GARN). These GAN-based methods are overly dependent on generated images for classifier training; unrealistic images add noise to the classifier, resulting in poorer performance on the facial expression recognition task. In contrast, the proposed method exploits adversarial learning to learn a better feature representation that excels at expression recognition without differentiating pose or subject. Thus, the feature representations can preserve expression information but avoid pose and subject variations. Therefore, the proposed method achieves better results.

6 CONCLUSION

In this paper, we proposed an adversarial feature learning method that can simultaneously address pose variations and subject bias. Specifically, an encoder extracts feature representations, an expression classifier tries to perform facial expression recognition, a pose discriminator and a subject discriminator classify the pose and the subject respectively for the representations, and a generator reconstructs facial images to further favor the feature representations. Leveraging adversarial leaning and training these networks jointly yields feature representations that are good for expression recognition but do not differentiate the pose or the subject. Thus, the feature representations can preserve expression information but avoid subject and pose variations. Experiments on five databases demonstrate that the proposed method outperforms the state-of-the-art methods.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under Grant 91748129, and Grant 61727809, and in part by the Project from Anhui Science and Technology Agency under Grant 1804a09020038.

REFERENCES

[1] Wen-Sheng Chu, Fernando De la Torre, and Jeffrey F Cohn. 2017. Selective transfer machine for personalized facial expression analysis. *IEEE transactions on pattern analysis and machine intelligence* 39, 3 (2017), 529–545.

[2] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. 2011. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. IEEE, 2106–2112.

[3] Changxing Ding and Dacheng Tao. 2016. A comprehensive survey on pose-invariant face recognition. *ACM Transactions on intelligent systems and technology (TIST)* 7, 3 (2016), 37.

[4] Stefanos Eleftheriadis, Ognjen Rudovic, and Maja Pantic. 2015. Discriminative shared gaussian processes for multiview and view-invariant facial expression recognition. *IEEE transactions on image processing* 24, 1 (2015), 189–204.

[5] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. 2013. Challenges in representation learning: A report on three machine learning contests. In *International Conference on Neural Information Processing*. Springer, 117–124.

[6] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. 2010. Multi-pie. *Image and Vision Computing* 28, 5 (2010), 807–813.

[7] Mahdi Jampour, Vincent Lepetit, Thomas Mauthner, and Horst Bischof. 2017. Pose-specific non-linear mappings in feature space towards multiview facial expression recognition. *Image and vision computing* 58 (2017), 38–46.

[8] Ying-Hsiu Lai and Shang-Hong Lai. 2018. 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)*. IEEE, 263–270.

[9] Shan Li and Weihong Deng. 2018. Deep facial expression recognition: A survey. *arXiv preprint arXiv:1804.08348* (2018).

[10] Yanwei Li, Xingang Wang, Shilei Zhang, Lingxi Xie, Wenqi Wu, Hongyuan Yu, and Zheng Zhu. 2018. Identity-Enhanced Network for Facial Expression Recognition. *arXiv preprint arXiv:1812.04207* (2018).

[11] Xiaofeng Liu, BVK Vijaya Kumar, Jane You, and Ping Jia. 2017. Adaptive deep metric learning for identity-aware facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 20–29.

[12] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, Nov (2008), 2579–2605.

[13] Qirong Mao, Qiyu Rao, Yongbin Yu, and Ming Dong. 2017. Hierarchical Bayesian theme models for multipose facial expression recognition. *IEEE Transactions on Multimedia* 19, 4 (2017), 861–873.

[14] Zibo Meng, Ping Liu, Jie Cai, Shizhong Han, and Yan Tong. 2017. Identity-aware convolutional neural network for facial expression recognition. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE, 558–565.

[15] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. 2017. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *arXiv preprint arXiv:1708.03985* (2017).

[16] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-First AAAI Conference on Artificial Intelligence*.

[17] Can Wang and Shangfei Wang. 2018. Personalized Multiple Facial Action Unit Recognition through Generative Adversarial Recognition Network. In *2018 ACM Multimedia Conference on Multimedia Conference*. ACM, 302–310.

[18] Jian Wu, Zhiming Cui, Victor S Sheng, Pengpeng Zhao, Dongliang Su, and Shengrong Gong. 2013. A Comparative Study of SIFT and its Variants. *Measurement science review* 13, 3 (2013), 122–131.

[19] Jianlong Wu, Zhouchen Lin, Wenming Zheng, and Hongbin Zha. 2017. Locality-constrained linear coding based bi-layer model for multi-view facial expression recognition. *Neurocomputing* 239 (2017), 143–152.

[20] Huiyuan Yang, Umur Ciftci, and Lijun Yin. 2018. Facial expression recognition by de-expression residue learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2168–2177.

[21] Huiyuan Yang, Zheng Zhang, and Lijun Yin. 2018. Identity-adaptive facial expression recognition through expression regeneration using conditional generative adversarial networks. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 294–301.

[22] Lijun Yin, Xiaochen Chen, Yi Sun, Tony Worm, and Michael Reale. 2008. A high-resolution 3D dynamic facial expression database. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*. IEEE, 1–6.

[23] Jiabei Zeng, Shiguang Shan, and Xilin Chen. 2018. Facial expression recognition with inconsistently annotated datasets. In *Proceedings of the European conference on computer vision (ECCV)*. 222–237.

[24] Feifei Zhang, Tianzhu Zhang, Qirong Mao, and Changsheng Xu. 2018. Joint pose and expression modeling for facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3359–3368.

[25] Kaihao Zhang, Yongzhen Huang, Yong Du, and Liang Wang. 2017. Facial expression recognition based on deep evolutional spatial-temporal networks. *IEEE Transactions on Image Processing* 26, 9 (2017), 4193–4203.

[26] Tong Zhang, Wenming Zheng, Zhen Cui, Yuan Zong, Jingwei Yan, and Keyu Yan. 2016. A deep neural network-driven feature learning method for multi-view facial expression recognition. *IEEE Transactions on Multimedia* 18, 12 (2016), 2528–2536.