Generative Neutral Features-Disentangled Learning for Facial Expression Recognition

Zhenqian Wu
School of Computer Science and Engineering, University of Electronic Science and Technology of China
russius@163.com

Yazhou Ren*
School of Computer Science and Engineering, University of Electronic Science and Technology of China
yazhou.ren@uestc.edu.cn

Xiaorong Pu
School of Computer Science and Engineering, University of Electronic Science and Technology of China
puxiaor@uestc.edu.cn

Zhifeng Hao
College of Science, Shantou University
haozhifeng@stu.edu.cn

Lifang He
Department of Computer Science and Engineering, Lehigh University
lih319@lehigh.edu

ABSTRACT

Facial expression recognition (FER) plays a critical role in human-computer interaction and affective computing. Traditional FER methods typically rely on comparing the difference between an examined facial expression and a neutral face of the same person to extract the motion of facial features and filter out expression-irrelevant information. With the extensive use of deep learning, the performance of FER has been further improved. However, existing deep learning-based methods rarely utilize neutral faces. To address this gap, we propose a novel deep learning-based FER method called Generative Neutral Features-Disentangled Learning (GNDL), which draws inspiration from the facial feature manifold. Our approach integrates a neutral feature generator (NFG) that generates neutral features in scenarios where the neutral face of the same subject is not available. The NFG uses fine-grained features from examined images as input and produces corresponding neutral features with the same identity. We train the NFG using a neutral feature reconstruction loss to ensure that the generative neutral features are consistent with the actual neutral features. We then disentangle the generative neutral features from the examined features to remove disturbance features and generate an expression deviation embedding for classification. Extensitive experimental results on three popular databases (CK+, Oulu-CASIA, and MMI) demonstrate that our proposed GNDL method outperforms state-of-the-art FER methods.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; **Image representations**.

*Corresponding author.

MM '23, October 29–November 3, 2023, Ottawa, ON, Canada.

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3612014

KEYWORDS

Facial expression recognition, neutral feature generator, facial feature manifold, disturbance-disentangling

ACM Reference Format:
Zhenqian Wu, Yazhou Ren, Xiaorong Pu, Zhifeng Hao, and Lifang He. 2023. Generative Neutral Features-Disentangled Learning for Facial Expression Recognition. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29–November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3581783.3612014

1 INTRODUCTION

Facial expression is a dominant and all-encompassing channel of communication that humans use to convey their emotions and intentions [6], and it provides a crucial insight into our psychological and physiological well-being [17]. To understand and interpret the complex nature of facial expressions, researchers have developed the Facial Action Coding System (FACS) [9]. FACS categorizes facial movements into 44 action units that represent distinct movements of facial muscles, such as the inner brow raiser, nose wrinkler, and upper lip raiser, among others. These action units can be combined to describe a range of facial expressions, thereby marking a significant milestone in the field of facial expression recognition (FER).

The facial action units play a critical role in intuitive recognition of facial expressions. Traditional methods for feature extraction in facial expression recognition mainly rely on contrasting these units from an expressionless neutral face to that of an expressive face. For example, Pantic *et al.* [28] use a multidetector to extract landmark points from facial expression images, while Cohn *et al.* [5] utilize a hierarchical optical flow method [23]. These points localize the contours of eyes, nose, and mouth, etc. By contrasting these landmark points with those in neutral face of the same person, they correspond to different facial muscle movements to recognize facial expressions. Huang *et al.* [12] calculate the difference between the model feature parameters of the examined facial expression and the neutral face of the same person to generate action parameters for FER. Kimura *et al.* [15] propose a Potential Net to model a neutral face as reference. By comparing the Potential Net of an image showing expression with the reference, they can extract the motion flow of expression. Kotsia *et al.* [16] define the difference of each Candide grid node coordinates between the examined expression

and the neutral face as geometrical displacement, which is used as an input to a classifier.

There are two main benefits for these methods to compare the representations of examined images with those of the same subject's neutral images. Firstly, they can filter out information that is unrelated to the facial expression. Secondly, they can extract facial feature motion that conforms to the FACS. However, despite the significant performance improvement in FER task [40, 44] with the rise of convolutional neural network (CNN) [18, 36], most deep learning-based methods focus on individual images or image sequences [20, 31], with little attention paid to neutral faces.

Inspired by these traditional methods, in this paper, we introduce a deep learning-based FER method called Generative Neutral Features-Disentangled Learning (GNDL), which aims to disentangle neutral features from the examined facial features. Specifically, we use a backbone CNN to extract feature vectors from the examined and neutral images with the same identity. Then, the neutral feature vector will be disentangled from the examined feature vector by a simple subtraction. Based on the prior study on facial feature manifold [3, 34], the features of examined and neutral facial images with the same identity distribute on the same manifold in the feature space, and they are composed of their respective expression-related features and shared disturbance features (shown in Figure 1). Thus, the final disentangled feature vector contains no disturbance features and captures the difference between the initial state and the apex state of the expression. However, the same subject's neutral face cannot always be obtained, to address which we propose a neutral feature generator (NFG). The NFG takes the examined feature vector as input and generates a neutral feature vector. A neutral feature reconstruction loss will be used for quantifying the distance between the generative neutral feature vector and the real neutral feature vector. With the trained NFG, the neutral feature disentangling will be practiced between the examined feature vector and the generative neutral feature vector, since only the examined images are fed into the model during validation. The main contributions of this paper can be summarized as follows:

1) Based on the manifold of facial features in the feature space, we propose a novel method to filter out expression-irrelevant features and extract expression deviation features, which is achieved by disentangling the generative neutral features from the examined feature vector.
2) Our approach involves a generator specifically designed to produce neutral features. This feature-level generator takes facial expression images as input and outputs their corresponding neutral feature vector, which can be fed into FER models as an additional source of information at a relatively low cost.
3) We achieved 99.69%, 90.14%, and 89.08% recognition accuracies respectively on three popular databases (*i.e.*, CK+, Oulu-CASIA, and MMI), which outperforms the state-of-the-art FER methods.

2 RELATED WORKS

In this section, we will introduce some recent deep learning-based FER methods, including those based on static facial images and image sequences. Additionally, we will present several methods that utilize neutral facial images.

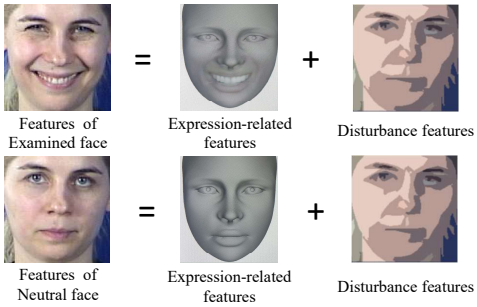


Figure 1: Features of examined and neutral faces are composed of their respective expression-related features and shared disturbance features. We use facial images to represent the corresponding features.

Methods for static images. These methods usually take the apex frame of a facial expression as the input. Devries *et al.* [7] employ a multi-task learning approach to classify facial expressions and predict facial landmark points simultaneously. The landmark points guide the network to focus on expression-related regions. Ali *et al.* [1] use features of a facial expression image and an one-hot encoding that represents identity, to modify the identity of the facial expression image. This approach will guide the network to extract features disentangled from identity information. Ruan *et al.* [32] train a network on databases for face recognition, facial pose recognition, etc., and transfer the learned knowledge to another network to disentangle multiple disturbing factors from facial expression images. Zhang *et al.* [42] propose an identity-disentangled model. They freeze a network pre-trained on a face recognition database to extract the identity information, which is then disentangled from the facial feature vector of the same image. Shao *et al.* [35] pre-train a generator to output label distributions of facial expressions, which are used as ground-truth for label distribution learning. A self-paced learning strategy is also employed. They address the issues of label ambiguity and label noise in FER.

Methods for image sequences. These methods use the entire image sequence of a facial expression from the initial state to the apex state as input. Liu *et al.* [21] convolve the input sequence with 3D filters and then use a set of filters corresponding to 13 facial parts that are manually defined to learn part-based representations. They then apply spatial constraints to refine the representations and achieve more reasonable results. Jung *et al* [13] employ 3D filters varying in importance over time to extract features from image sequences, while using a multi-layer perceptron (MLP) to extract features from concatenated coordinates of landmark points of each frame. They propose a joint fine-tuning method to combine these two models. To distinguish more representative frames during training, Meng *et al.* [25] propose self-attention weights and relation-attention weights for each frame's feature vector, which is extracted by a CNN. Wang *et al.* [38] propose a dual path network to represent sparsely sampled frames, which uses CNNs to extract weighted consecutive frame-level features, and further applies a dual path long short-term memory (LSTM) module to learn and aggregate channel-aware and temporal-aware features.

Methods using neutral facial images. A few deep learning-based methods, like ours, use neutral expressions during training and exclude them during validation. Kim *et al.* [14] propose an encoder-decoder network to generate neutral expressions from examined images and calculate element-wise distance between their features for classification. They use three reconstruction losses to ensure the effectiveness of the extracted features: one to generate the neutral expression from the input image, another to reconstruct the input image from its own feature map, and a third to reconstruct the generative neutral expression from its own feature map. Yang *et al.* [39] introduce an adversarial generative network (GAN) that generates neutral expressions from input facial expression images. They emphasize the importance of recording individual-specific expression information in the intermediate layers of the generator. Therefore, they employ a local CNN model to extract features from the generator's intermediate layers. Unlike the aforementioned methods, which extract neutral features from generative neutral expression images, our proposed GNDL reconstructs the neutral features directly, rather than the neutral images. The GNDL is more efficient, straightforward, and easier to train.

3 PROPOSED METHOD

3.1 Overview

Figure 2 presents an overview of our Generative Neutral Features-Disentangled Learning (GNDL) method, which showcases the frameworks employed in both the training and validation stages.

The training model takes as input both examined expression images to be recognized and their corresponding neutral expression images. They are simultaneously fed into a facial feature extraction network, and normalization will be performed between them at each convolutional layer, aiming to make them distributed on the same manifold in the feature space. The network outputs examined feature vectors $\boldsymbol{F}_{exam}$ and neutral feature vectors $\boldsymbol{F}_{neu}$ separately. Meanwhile, we train a Neutral Feature Generator (NFG) to transform the examined feature vectors into their corresponding neutral feature vectors, which are represented by $\boldsymbol{F}'_{neu}$. NFG's architecture will be introduced in Section 3.3. To train NFG, we propose a neutral feature reconstruction loss, which measures the distance between $\boldsymbol{F}_{neu}$ and $\boldsymbol{F}'_{neu}$. Then, we achieve neutral feature disentangling, which will be explained in Section 3.2, by subtracting $\boldsymbol{F}'_{neu}$ from $\boldsymbol{F}_{exam}$. At last, expressive deviation feature vectors that are free from disturbance features will be obtained for classification.

The facial feature extraction network used for the validation model does not involve any normalization as it only processes examined expression images. The expression deviation feature vectors are obtained by subtracting the neutral feature vector, which is generated by the trained NFG.

3.2 Neutral Feature Disentangling

The features or feature vector $\boldsymbol{F}_{exam}$ of an examined facial image can be defined as a linear combination of expression-related features $\boldsymbol{F}_{exp}$ and expression-unrelated (disturbance) features $\boldsymbol{F}_{dis}$ [2, 4]:

$$\boldsymbol{F}_{exam} = \boldsymbol{F}_{exp} + \boldsymbol{F}_{dis}. \tag{1}$$

Removing disturbance features and utilizing features highly correlated with facial expressions for classification leads to better accuracy. However, it remains challenging to eliminate $\boldsymbol{F}_{dis}$ or obtain $\boldsymbol{F}_{exp}$ directly. Therefore, we focus our study on the early analysis of the facial expression manifold, as proposed in [3, 34].

An N-dimensional feature vector of a facial image can be regarded as a point in an N-dimensional feature space. For instance, Figure 3 illustrates this concept for $N = 3$, but it can be extended to higher dimensions. In this feature space, all facial feature vectors of an individual are situated on a smooth manifold, where the neutral expression serves as the central reference point. Moving away from the reference center, sequences of facial images with continuous changes expand outward along a path. When the identities of facial images differ, their feature vectors are positioned on distinct, yet comparable, manifolds with similar shapes.

Figure 4 shows the manifold on which $\boldsymbol{F}_{exp}$ is located. Particularly, we define this manifold's reference center as $\boldsymbol{F}_{init}$, representing the expression-related features of neutral expression, which is the initial state of other expressions. The features of neutral expression $\boldsymbol{F}_{neu}$ are the linear combination of $\boldsymbol{F}_{init}$ and the similar disturbance features $\boldsymbol{F}_{dis}$ shared with other facial features on the same manifold:

$$\boldsymbol{F}_{neu} = \boldsymbol{F}_{init} + \boldsymbol{F}_{dis}. \tag{2}$$

Building on the aforementioned studies, we subtract the facial feature vector of the neutral image from those of the examined image with the same identity for expression recognition purpose:

$$\boldsymbol{F}_{exp-d} = \boldsymbol{F}_{exam} - \boldsymbol{F}_{neu} = \boldsymbol{F}_{exp} - \boldsymbol{F}_{init}. \tag{3}$$

$\boldsymbol{F}_{exp-d}$ is the expression deviation feature vector, which not only excludes disturbance features and is highly correlated with expressive information, but also contains the variation information of the expression from the initial state to the apex state.

Due to the varying illumination, neutral features and examined facial features may be distributed on different manifolds, even with the same identity. Thus, we apply a normalization operation between the output features of examined image and neutral image at each convolutional layer in the facial-features-extracted network. Let $\boldsymbol{F}^{i,j}_{exam}$ be the output features of the i^{th} examined image in the j^{th} convolutional layer. Let $\boldsymbol{F}^{i,j}_{neu}$ be the output features of the i^{th} neutral image in the j^{th} convolutional layer. We first calculate average value μ and variance σ^2:

$$\mu_j = \frac{1}{M}\sum_{i=1}^{M}(\boldsymbol{F}^{i,j}_{exam} + \boldsymbol{F}^{i,j}_{neu}), \tag{4}$$

$$\sigma_j^2 = \frac{1}{M}\sum_{i=1}^{M}\left((\boldsymbol{F}^{i,j}_{exam} - \mu)^2 + (\boldsymbol{F}^{i,j}_{neu} - \mu)^2\right), \tag{5}$$

where M represents the number of paired samples in a batch. Then, we obtain the input features of the next layer by

$$\hat{\boldsymbol{F}}^{i,j}_{\cdot} = g\left(\gamma\frac{\boldsymbol{F}^{i,j}_{\cdot} - \mu}{\sqrt{\sigma^2 + \epsilon}} + \beta\right), \tag{6}$$

where $\boldsymbol{F}^{i,j}_{\cdot}$ represents $\boldsymbol{F}^{i,j}_{exam}$ or $\boldsymbol{F}^{i,j}_{neu}$, $\hat{\boldsymbol{F}}^{i,j}_{\cdot}$ represents normalized features, γ and β are reconstruction parameters learned by the networks, ϵ is a constant that prevents the denominator from being zero, and $g(\cdot)$ denotes activation function. Eq. (4) to Eq. (6) are all

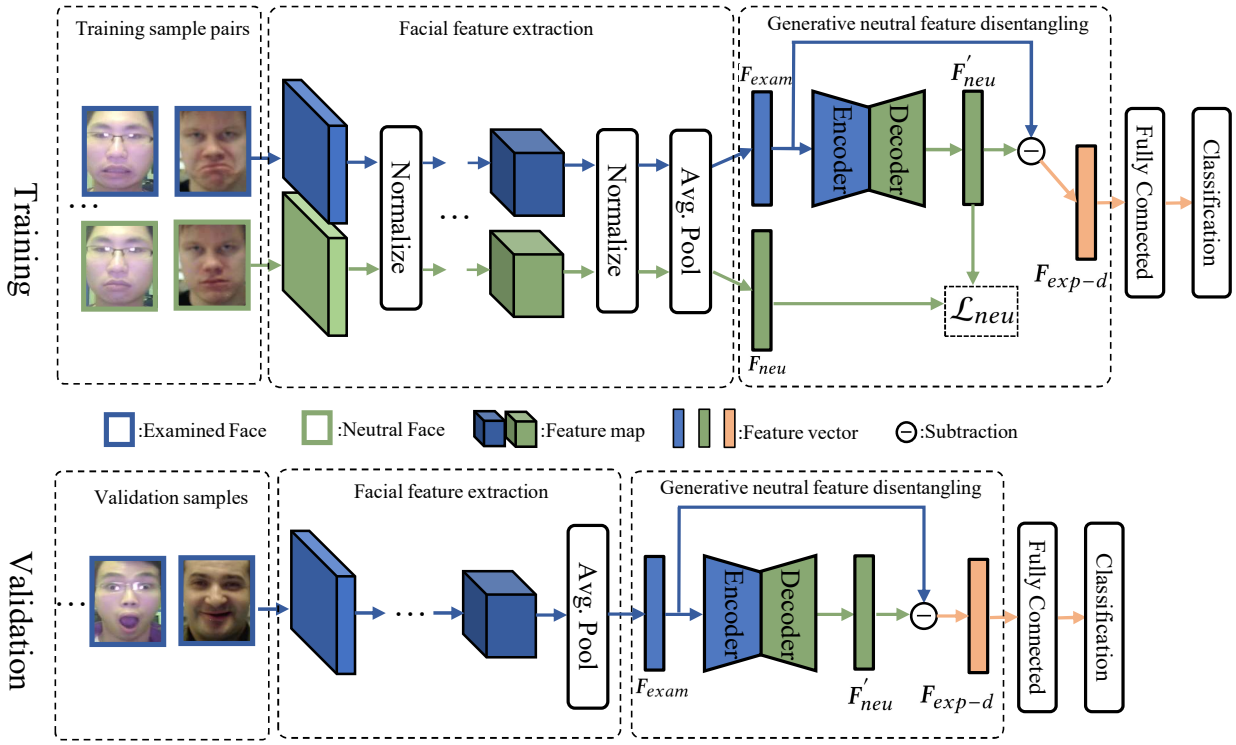


Figure 2: Overview of our proposed Generative Neutral Features-Disentangled Learning (GNDL) method. The training model and the validation model are presented separately. F_{exam} represents the feature vector of the examined expression image to be recognized, F_{neu} represents the feature vector of the neutral expression image with the same identity, F'_{neu} represents the generative neutral feature vector, and $\mathcal{L}_{neu}$ is our proposed neutral feature reconstruction loss.

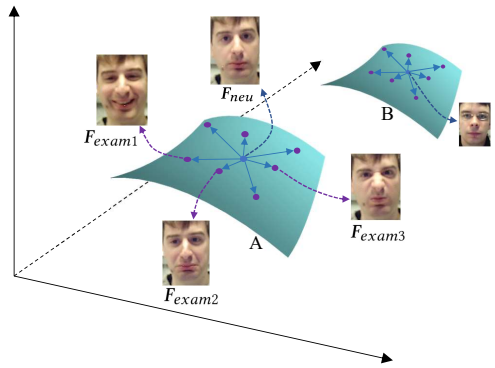


Figure 3: Two examples of manifolds with different identities in the facial feature space. Manifold A and manifold B have similar shapes. The center points of manifold A and manifold B are both neutral expression features (represented by F_{neu}). The features of other expressions (F_{exam1}, F_{exam2}, and F_{exam3}) are distributed around.

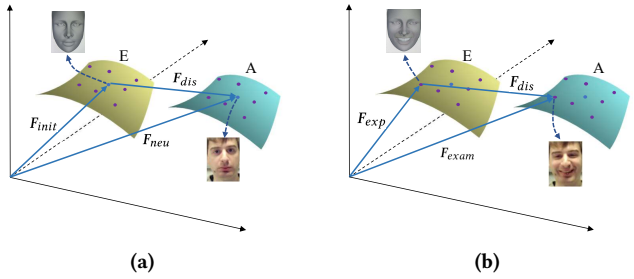


Figure 4: The manifold of expression-related features. The gray facial expressions represent expression-related features. By linearly combining with the similar disturbance features (F_{dis}), expression-related features (F_{exp} or F_{init}) on manifold E can be transformed into the facial image features on manifold A (F_{exam} or F_{neu}).

calculated separately for each channel of the features. At the end of the network, we obtain the examined feature vector and the neutral feature vector through an average pooling layer, then the disentangling operation is performed.

3.3 Neutral Feature Generator

During validation, it is possible that there is no corresponding neutral image available to match a given facial expression image. We

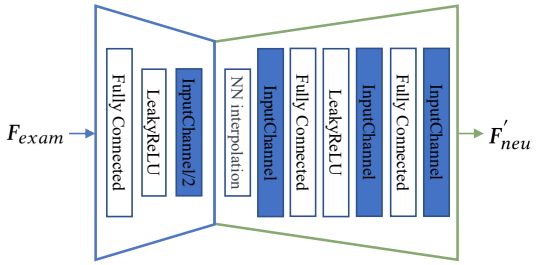


Figure 5: Structure of the proposed NFG. The "InputChannel" represents the number of channels of the input feature vector, the blue rectangle represents feature vector with its channel number, and the "NN interpolation" means nearest neighbor interpolation.

here propose a neutral feature generator (NFG) which can learn the mapping from facial features of any expression to neutral features.

As shown in Figure 5, NFG consists of an encoder and a decoder. In the encoding stage, the input examined feature vector is transformed into an intermediate vector through a fully connected layer and a LeakyReLU activation function, with the channel number being halved. The decoder emulates the typical upsampling process used in image generation task [10, 27] and image segmentation task [30], which involves bilinear interpolation of feature maps followed by convolution. Instead, NFG uses a nearest neighbor interpolation on the low-dimensional vector to double the channel number, followed by fully connected layers and LeakyReLu activation function to achieve a non-linear transformation. At last, a neutral feature vector with the same channel number as the input is generated. Let $\mathcal{G}$ be the generator, then the proposed disentangling method in Section 3.2 can be replaced by:

$$F_{exp-d} = F_{exam} - \mathcal{G}(F_{exam}) = F_{exam} - F'_{neu}. \tag{7}$$

During the training of NFG, a neutral feature reconstruction loss is designed to quantify the distance between facial feature vector and neutral feature vector. Let $\mathcal{X} = \{(\mathbf{x}^1_{exam}, \mathbf{x}^1_{neu}), \cdots, (\mathbf{x}^N_{exam}, \mathbf{x}^N_{neu})\}$ be the data set where $\mathbf{x}^i_{exam} \in \mathbb{R}^{W\times H}$ is the i^{th} examined image, $\mathbf{x}^i_{neu} \in \mathbb{R}^{W\times H}$ is the i^{th} corresponding neutral expression image, N is the total number of samples. The neutral feature reconstruction loss can be formulated as:

$$\mathcal{L}_{neu} = \sum_{i=1}^{N} ||F^i_{exam} - F'^i_{neu}||_1. \tag{8}$$

3.4 Loss Function

We use the crossentropy loss as the facial expression classification loss:

$$\mathcal{L}_{cls} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} y^k_i \log(p^k_i). \tag{9}$$

Here, y^k_i is the value of the k^{th} position in the one-hot label of the i^{th} sample, K is the number of classes ($K \geq 2$), N is the total number of samples, and p^k_i is the probability that the i^{th} sample belongs to the k^{th} class, defined as:

$$p^k_i = \frac{e^{\theta_k^T V^i_{exp-d}}}{\sum_{l=1}^{K} e^{\theta_l^T V^i_{exp-d}}}, \tag{10}$$

where θ_k denotes the parameter vector of the k^{th} class in the linear fully-connected layer, and V^i_{exp-d} defined in Section 3.2 is the expression deviation feature vector.

All parts of our method are jointly trained in an end-to-end manner, and the total loss function is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda\mathcal{L}_{neu}, \tag{11}$$

where $\mathcal{L}_{neu}$ is the neutral feature reconstruction loss defined in Section 3.3, and λ is the regularization parameter to balance the importance of $\mathcal{L}_{cls}$ and $\mathcal{L}_{neu}$.

4 EXPERIMENTS

In this section, we evaluate the effectiveness of our proposed method by conducting experiments on three distinct databases and comparing with the existing methods.

4.1 Experimental Setup

Databases. Our experiments are carried out on three facial expression databases: **Extended Cohn-Kanade Database (CK+)** [24] contains 593 image sequences capturing the facial expressions of 123 subjects, but only 327 sequences with 118 subjects are labeled with one of the seven expressions: anger, contempt, disgust, fear, happiness, sadness, and surprise. We select the first neutral frame and three peak expression frames from each image sequence, resulting in a total of 981 paired samples. **Oulu-CASIA Database** [43] is comprised of 480 image sequences captured from 80 subjects under three different illumination conditions. We use only the sequences captured under strong illumination condition with the VIS camera, which are labeled as one of the six expressions: anger, disgust, fear, happiness, sadness, and surprise. We select the first neutral frame and three peak expression frames from each sequence, resulting in a total of 1440 paired samples. **MMI Facial Expression Database (MMI)** [29] is a more challenging database containing 213 facial expression videos from 32 subjects. It has a relatively small sample size and the subjects have various poses, as well as occlusion of their faces by glasses or hair. Each video is labeled with one of the six expressions: happiness, surprise, sadness, anger, disgust, and fear. For our experiments, we select 205 frontal view sequences and choose three peak frames and one neutral frame from each sequence, resulting in a total of 615 image pairs.

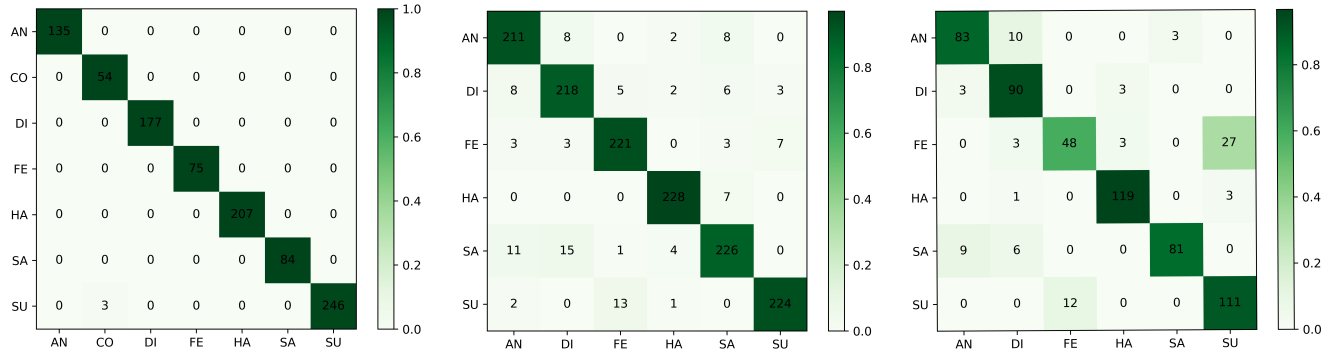
Evaluation metrics. The evaluation of all compared methods is based on recognition accuracy (ACC). As CK+, Oulu-CASIA, and MMI databases do not have pre-defined training and test sets, we follow the same protocol as the compared methods, and use a 10-fold subject-independent cross-validation. The reported final results are the average ACC across all 10 folds.

4.2 Implementation Details

The facial images in all three databases are preprocessed by first detecting and cropping the face region, which is then resized to a standardized size of 224×224 pixels. To prevent overfitting during training, data augmentation techniques such as random horizontal

Table 1: Recognition accuracy (%) on the CK+, Oulu-CASIA, and MMI databases for expressions classification. Methods marked with an asterisk (*) in the setting column indicate that, similar to our method, they use neutral facial expressions during training but not during validation.

Method	Setting	CK+	Oulu-CASIA	MMI
STM-E[22](2014)	sequence-based	94.19	74.59	75.12
3DCNN-DAP[21](2014)	sequence-based	92.40	—	63.40
DTAGN[13](2015)	sequence-based	97.25	81.46	—
IACNN[26](2017)	image-based	95.37	—	71.55
DLP-CNN[19](2017)	image-based	95.78	—	78.46
FN2EN[8](2017)	image-based	98.60	87.71	—
DESTN[41](2017)	sequence-based	98.50	86.25	81.18
$GCNet_{S1R1}$[14](2017)	image-based*	97.93	86.11	81.53
DeRL[39](2018)	image-based*	97.30	88.0	73.23
DDL[32](2020)	image-based	99.16	88.26	83.67
FDRL[33](2021)	image-based	99.54	88.26	85.23
Baseline(ResNet-18)	image-based	97.55	87.57	82.56
GNDL	image-based*	**99.69**	**90.07**	**86.46**



(a) CK+ **(b) Oulu-CASIA** **(c) MMI**

Figure 6: Confusion matrices on the CK+, Oulu-CASIA, and MMI databases. The vertical axis represents the ground truth labels, while the horizontal axis represents the predicted labels (Ha=Happiness, Sa=Sadness, Su=Surprise, Fe=Fear, Di=Disgust, An=Anger, Co=Contempt).

flipping or adding Gaussian noise are applied to increase the amount of training data. Importantly, the same augmentation operation is applied to both the examined and neutral images in each paired sample to maintain consistency.

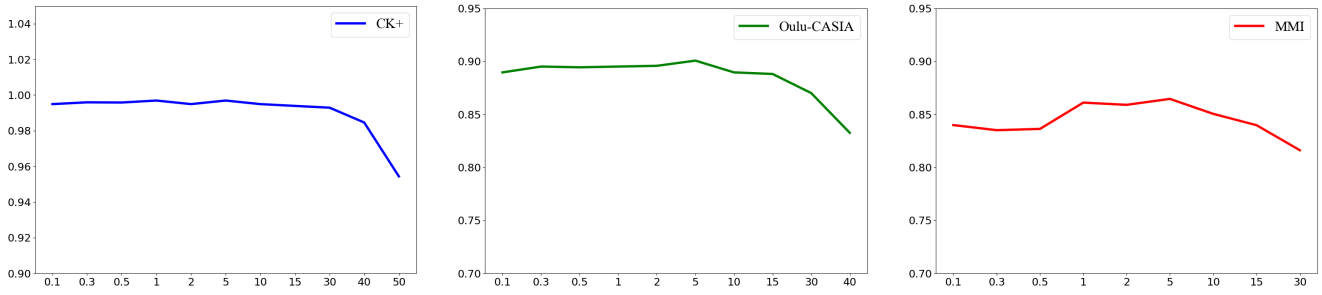
We use the ResNet-18 [11] model as the facial-features-extracted network in independent experiments. We set the channel number of the examined feature vector as 512. The normalization between the examined and neutral features is achieved with the batch normalization layer. The value of λ in Eq. (11) is set to 5.0. The proposed GNDL model is trained in an end-to-end manner for 150 epochs. The batch size is set to 128. We adopt Adam as the optimization method, with learning rate $lr = 0.001$, first-order momentum $\beta_1 = 0.9$, and second-order momentum $\beta_2 = 0.999$. All experiments are implemented with the Python 3.8 and Pytorch 1.7.1 on a Linux server equipped with 2.5GHz CPU and 16GB RAM, a single RTX3070 GPU is used to accelerate the training stage.

4.3 Comparison with State-of-the-Art

Table 1 shows the comparison results between the proposed GNDL and the state-of-the-art methods mentioned in Section 4.1 on the CK+, Oulu-CASIA, and MMI databases.

The CK+ database is widely used in the field of facial expression recognition due to its relatively simple and high-quality samples. However, achieving high accuracy on this database is still a challenging task. Despite the high performance of most of the compared methods on this database, our proposed GNDL method still manages to improve the recognition accuracy to 99.69%. This demonstrates the effectiveness of our proposed method in learning discriminative features from facial expression images. The confusion matrix of the CK+ database is shown in Figure 6a, which indicates that only three "surprise" samples are misclassified as "contempt".

Our proposed GNDL method outperforms the baseline on the Oulu-CASIA database, which is known for its challenging conditions such as partial occlusions and variations in illumination. The



(a) CK+ (b) Oulu-CASIA (c) MMI

Figure 7: Experimental results of sensitivity analysis for the hyperparameter λ in Eq. (11). The horizontal axis represents different values of λ, and the vertical axis represents the recognition accuracy. On each database, λ is varied from 0.1 until a significant decrease in accuracy is observed.

recognition accuracy is improved to 90.07%. Figure 6b displays the confusion matrix of the Oulu-CASIA database. The "sadness" expression is relatively difficult to recognize and is frequently misclassified as "anger" and "disgust".

The MMI database is a more challenging database due to the high intra-class inconsistency, which also contains a wide range of facial expressions under varying conditions. However, Our proposed GNDL still outperforms the baseline by a significant margin and achieves a higher accuracy of 86.46%, demonstrating its robustness in handling challenging conditions. The improvement in accuracy on this database can be attributed to the ability of our method to remove disturbance information and capture expression deviation features, which is crucial in handling the variations in the input images. Figure 6c shows the confusion matrix. It indicates that our method performs very well in recognizing "happiness" and "disgust", while "fear" is relatively hard for recognition, and is mostly confused with "surprise".

In these compared methods, FDRL has the closest performance to ours. It decomposes facial images into different facial-action feature vectors to represent various facial actions, and then fuses them based on their intra-feature and inter-feature relations to identify subtle differences between facial expressions. However, our proposed method disentangles the generative fine-grained semantic features of the neutral expression, which can obtain the deviation information of expression according to the manifold study. This is similar to the effect of obtaining dynamic by using sequence-based methods. Moreover, our method can also remove disturbance information such as identity, occlusion, and illumination, which is also the purpose of IACNN and DDL.

In comparing our method with GCNet$S1R1$ and DeRL, it is important to note that we share the same input setting. Our methods have all achieved good results on the Oulu-CASIA database, as we have all captured the variation information from the neutral expression to the examined expression. Only our method and GCNet$S1R1$ have achieved better results on the MMI database, as we both directly extract the differences between the examined expression and neutral expression, while removing shared disturbance information. However, DeRL uses the intermediate features of the network when mapping the examined expression to the neutral expression, which cannot guarantee the filtering out of disturbance information. Additionally, GCNet$_{S1R1}$ and DeRL require generating neutral expression images within the network, while we propose a feature-level generator that directly generates neutral features, making the process more direct and efficient.

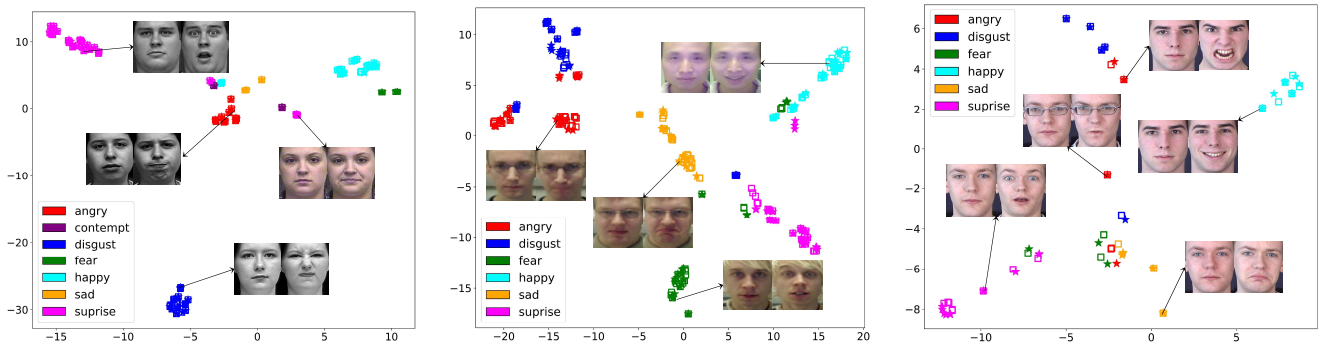
4.4 Parameter Sensitivity Analysis

To verify the robustness of our proposed method, we evaluate the performance of the GNDL with different values of the hyperparameter λ in Eq. (11) on the CK+, Oulu-CASIA, and MMI databases. Starting with an initial value of λ set to 0.1, we gradually increased it to examine its impact on the recognition accuracy.

In more details, on the CK+ database, the recognition accuracy achieved by our method remains consistently high with λ set to values between 0.1 and 10.0, indicating that the proposed GNDL is robust to the variation of this hyperparameter. However, when λ is set to 10.0, the accuracy starts to decrease, suggesting that the regularization effect introduced by λ becomes too strong and may lead to overfitting of the model. On the Oulu-CASIA database, the recognition accuracy gradually increases as λ grows from 0.1 to 5.0, and then decreases when λ further increases. On the MMI database, the recognition accuracy fluctuates around 0.84 when λ is small. As λ grows, the recognition accuracy gradually increases and reaches its peak around $\lambda = 5.0$. As with the Oulu-CASIA database, when λ becomes too large, the regularization effect becomes too strong and can lead to a decrease in recognition accuracy. This phenomenon can be explained as follows: with a small value of λ, the weight of the neutral feature loss is not enough to guide the NFG to generate effective neutral features, resulting in a slow learning speed. However, when λ is set to around 5.0, the NFG can quickly converge and generate effective neutral features, which can significantly improve the recognition accuracy. As λ continues to increase, the weight of the neutral feature reconstruction loss becomes too strong, leading to a lack of balance between the losses.

4.5 Visualization

To visually demonstrate the effectiveness of our proposed NFG, we disentangle the real neutral feature vector and the generative



(a) CK+ (b) Oulu-CASIA (c) MMI

Figure 8: Distributions of the disentangled feature vectors. We respectively present the distributions of disentangled feature vectors obtained using real neutral features (represented by squares) and generative neutral features (represented by asterisks). The data of each database is from a randomly selected validation fold of the 10-fold cross-validation.

Neutral	Examined	Truth	Pedicted (baseline)	Pedicted (GNDL)
		Disgust	Angry	Disgust
		Angry	Disgust	Angry
		Angry	Sad	Angry
		Fear	Sad	Fear
		Disgust	Angry	Disgust
		Sad	Happy	Sad
		Fear	Surprise	Fear

Figure 9: Some hard samples misclassified by the baseline method but correctly classified by the proposed GNDL on the MMI database.

neutral feature vector from the examined feature vector for comparison, and then visualize the distributions of their disentangled feature vectors via t-SNE [37] in Figure 8. The real disentangled feature vectors are represented by squares, while the generative disentangled feature vectors are represented by asterisks. Different colors represent different types of facial expressions. It can be observed that in all databases, the two types of disentangled feature vectors are similar (the coordinates of the asterisks and squares are similar), indicating that the generative neutral feature vectors are consistent with the real ones. Additionally, the disentangled feature vectors can effectively differentiate between different facial expressions in all databases.

Figure 9 illustrates some challenging samples that are misclassified by the baseline ResNet-18 method but correctly classified by our proposed GNDL on the MMI database. The ResNet-18 method fails to detect facial wrinkles caused by facial movements, as shown in the first and fifth rows of samples. Furthermore, wrinkles that are part of the facial anatomy can be mistaken as facial movements, as demonstrated in the sixth row, leading to incorrect classification. Additionally, the presence of glasses, bangs, or beards can obstruct some facial regions and impede the extraction of facial features, as shown in the samples from the second to the fifth row. However, our proposed GNDL takes advantage of the differences between neutral features and examined features, which effectively circumvents these influences and successfully recognizes these hard samples.

5 CONCLUSION AND FUTURE WORK

In this paper, we have proposed a generative neutral features-disentangled learning (GNDL) model for facial expression recognition, which achieves state-of-the-art performance on the CK+, Oulu-CASIA, and MMI databases. Our approach disentangles neutral features based on the expression manifold and incorporates a neutral feature generator (NFG) to output generative neutral feature vectors. By disentangling the neutral feature vector from the examined expression feature vector, our model can remove disturbance features and extract expression deviation features for more accurate classification. In addition, our model does not require the input of a neutral expression during the validation phase, enabling the handling of situations where obtaining the same subject's neutral expression is not possible. We conduct neutral feature disentangling at the vector-level in this paper. In future work, we will explore the effects of disentangling neutral features of different granularities.

ACKNOWLEDGMENTS

This work was supported in part by National Key Research and Development Program of China (2020YFC2004300 and 2020YFC2004302), National Natural Science Foundation of China (61971052), Lehigh's grants (S00010293 and 001250), and National Science Foundation (MRI 2215789 and IIS 1909879).

REFERENCES

[1] Kamran Ali and Charles E. Hughes. 2019. Facial Expression Recognition Using Disentangled Adversarial Learning. *arXiv preprint arXiv:1909.13135* (2019).

[2] Volker Blanz and Thomas Vetter. 1999. A Morphable Model for the Synthesis of 3D Faces. In *SIGGRAPH*. 187–194.

[3] Y. Chang, C. Hu, R. Feris, and M. Turk. 2006. Manifold based analysis of facial expression. *IMAVIS* 24, 6 (2006), 605–614.

[4] Baptiste Chu, Sami Romdhani, and Liming Chen. 2014. 3D-Aided Face Recognition Robust to Expression and Pose Variations. *CVPR* (2014), 1907–1914.

[5] J.F. Cohn, A.J. Zlochower, J.J. Lien, and T. Kanade. 1998. Feature-point tracking by optical flow discriminates subtle differences in facial expression. In *IEEE FG*. 396–401.

[6] Charles Darwin. 1872. *The Expression of the Emotions in Man and Animals*. John Murray, London.

[7] Terrance Devries, Kumar Biswaranjan, and Graham W. Taylor. 2014. Multi-Task Learning of Facial Landmarks and Expression. In *CPV*. 98–103.

[8] Hui Ding, Shaohua Kevin Zhou, and Rama Chellappa. 2017. FaceNet2ExpNet: Regularizing a Deep Face Recognition Net for Expression Recognition. In *IEEE FG*. 118–126.

[9] Paul Ekman and Wallace V. Friesen. 1978. Facial action coding system: A technique for the measurement of facial movement. *CPP* 12 (1978).

[10] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *NIPS*. 2672–2680.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR*. 770–778.

[12] Chung-Lin Huang and Yu-Ming Huang. 1997. Facial Expression Recognition Using Model-Based Feature Extraction and Action Parameters Classification. *JVCIR* 8, 3 (1997), 278–290.

[13] Heechul Jung, Sihaeng Lee, Junho Yim, Sunjeong Park, and Junmo Kim. 2015. Joint Fine-Tuning in Deep Neural Networks for Facial Expression Recognition. In *ICCV*. 2983–2991.

[14] Youngsung Kim, ByungIn Yoo, Youngjun Kwak, Changkyu Choi, and Junmo Kim. 2017. Deep generative-contrastive networks for facial expression recognition. *arXiv preprint arXiv:1703.07140* (2017).

[15] Satoshi Kimura and Masahiko Yachida. 1997. Facial Expression Recognition and Its Degree Estimation. In *CVPR*. 295–300.

[16] I. Kotsia and I. Pitas. 2007. Facial Expression Recognition in Image Sequences Using Geometric Deformation Features and Support Vector Machines. *TIP* 16, 1 (2007), 172–187.

[17] Sylvia D. Kreibig. 2010. Autonomic nervous system activity in emotion: A review. *BIOL PSYCHOL* 84, 3 (2010), 394–421.

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*. 1097–1105.

[19] Shan Li and Weihong Deng. 2019. Reliable Crowdsourcing and Deep Locality-Preserving Learning for Unconstrained Facial Expression Recognition. *TIP* 28, 1 (2019), 356–370.

[20] Shan Li and Weihong Deng. 2022. Deep Facial Expression Recognition: A Survey. *TAFFC* 13, 3 (2022), 1195–1215.

[21] Mengyi Liu, Shaoxin Li, S. Shan, Ruiping Wang, and Xilin Chen. 2014. Deeply Learning Deformable Facial Action Parts Model for Dynamic Expression Analysis. In *ACCV*. 143–157.

[22] Mengyi Liu, Shiguang Shan, Ruiping Wang, and Xilin Chen. 2014. Learning Expressionlets on Spatio-temporal Manifold for Dynamic Facial Expression Recognition. In *CVPR*. 1749–1756.

[23] Bruce D. Lucas and Takeo Kanade. 1981. An Iterative Image Registration Technique with an Application to Stereo Vision. In *IJCAI*. 674–679.

[24] Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. 2010. The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action uniand emotion-specified expression. In *CVPR Workshop*. 94–101.

[25] Debin Meng, Xiaojiang Peng, Kai Wang, and Yu Qiao. 2019. Frame Attention Networks for Facial Expression Recognition in Videos. In *ICIP*. 3866–3870.

[26] Zibo Meng, Liu Ping, Cai Jie, Shizhong Han, and Tong Yan. 2017. Identity-Aware Convolutional Neural Network for Facial Expression Recognition. In *IEEE FG*. 558–565.

[27] Mehdi Mirza and Simon Osindero. 2014. Conditional Generative Adversarial Nets. *arXiv preprint arXiv:1703.07140* (2014).

[28] M Pantic and L.J.M Rothkrantz. 2000. Expert system for automatic analysis of facial expressions. *IMAVIS* 18, 11 (2000), 881–905.

[29] M. Pantic, Michel Valstar, R. Rademaker, and L. Maat. 2005. Web-based database for facial expression analysis. In *ICME*. 5–15.

[30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *MICCAI*. 234–241.

[31] Philipp V. Rouast, Marc T. P. Adam, and Raymond Chiong. 2021. Deep Learning for Human Affect Recognition: Insights and New Developments. *TAFFC* 12, 2 (2021), 524–543.

[32] Delian Ruan, Yan Yan, Si Chen, Jing-Hao Xue, and Hanzi Wang. 2020. Deep Disturbance-Disentangled Learning for Facial Expression Recognition. In *ACM MM*. 2833–2841.

[33] Delian Ruan, Yan Yan, Shenqi Lai, Zhenhua Chai, Chunhua Shen, and Hanzi Wang. 2021. Feature Decomposition and Reconstruction Learning for Effective Facial Expression Recognition. In *CVPR*. 7660–7669.

[34] Caifeng Shan, Shaogang Gong, and Peter W. McOwan. 2005. Appearance Manifold of Facial Expression. In *HCI/ICCV*. 221–230.

[35] Jianjian Shao, Zhenqian Wu, Yuanyan Luo, Shudong Huang, Xiaorong Pu, and Yazhou Ren. 2022. Self-Paced Label Distribution Learning for In-The-Wild Facial Expression Recognition. In *ACM MM*. 161–169.

[36] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *CVPR*. 1–9.

[37] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *JMLR* 9 (2008), 2579–2605.

[38] Yan Wang, Yixuan Sun, Wei Song, Shuyong Gao, Yiwen Huang, Zhaoyu Chen, Weifeng Ge, and Wenqiang Zhang. 2022. DPCNet: Dual Path Multi-Excitation Collaborative Network for Facial Expression Representation Learning in Videos. In *ACM MM*. 101–110.

[39] Huiyuan Yang, Umur Ciftci, and Lijun Yin. 2018. Facial Expression Recognition by De-expression Residue Learning. In *CVPR*. 2168–2177.

[40] Junho Yim, Heechul Jung, ByungIn Yoo, Changkyu Choi, Dusik Park, and Junmo Kim. 2015. Rotating your face using multi-task deep neural network. In *CVPR*. 676–684.

[41] Kaihao Zhang, Yongzhen Huang, Yong Du, and Liang Wang. 2017. Facial Expression Recognition Based on Deep Evolutional Spatial-Temporal Networks. *TIP* 26, 9 (2017), 4193–4203.

[42] Wei Zhang, Xianpeng Ji, Keyu Chen, Yu Ding, and Changjie Fan. 2021. Learning a Facial Expression Embedding Disentangled From Identity. In *CVPR*. 6759–6768.

[43] Guoying Zhao, Xiaohua Huang, Matti Taini, Stan Z. Li, and Matti Pietikainen. 2011. Facial expression recognition from near-infrared videos. *IMAVIS* 29, 9 (2011), 607–619.

[44] Zhenyao Zhu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2013. Deep learning identity-preserving face space. In *ICCV*. 113–120.