Face Reenactment Based Facial Expression Recognition

Kamran Ali and Charles Hughes

October 14, 2020

Face Reenactment Based Facial Expression Recognition

Kamran Ali and Charles E. Hughes

Synthetic Reality Lab, CECS, University of Central Florida, USA
kamran@knights.ucf.edu, ceh@cs.ucf.edu

Abstract. Representations used for Facial Expression Recognition (FER) are usually contaminated with identity specific features. In this paper, we propose a novel Reenactment-based Expression-Representation Learning Generative Adversarial Network (REL-GAN) that employs the concept of face reenactment to disentangle facial expression features from identity information. In this method, the facial expression representation is learned by reconstructing an expression image employing an encoder-decoder based generator. More specifically, our method learns the disentangled expression representation by transferring the expression information from the source image to the identity of the target image. Experiments performed on widely used datasets (BU-3DFE, CK+, Oulu-CASIA, SEFW) show that the proposed technique produces comparable or better results than state-of-the-art methods.

Keywords: Facial Expression Recognition · Face Reenactment · Disentangled Representation Learning · Image Classification

1 Introduction

Facial expression recognition (FER) has many exciting applications in domains like human-machine interaction, intelligent tutoring system (ITS), interactive games, and intelligent transportation. As a consequence, FER has been widely studied by the computer vision and machine learning communities over the past several decades. Despite this extensive research, FER is still a difficult and challenging task. Most FER techniques developed so far do not consider inter-subject variations and differences in facial attributes of individuals present in data. Hence, the representation used for the classification of expressions contains identity-related information along with facial expression information, as observed in [1], [2]. The main drawback of this entangled representation is that it negatively affects the generalization capability of FER techniques, which, as a result, degrades the performance of FER algorithms on unseen identities. We believe the key to overcoming the challenge of over-fitting of FER models to subjects involved in the training set lies in FER techniques to disentangle the expression features from the identity information.

Face reenactment is an emerging face synthesis task that has attracted the attention of the research community due to its applications in the virtual reality

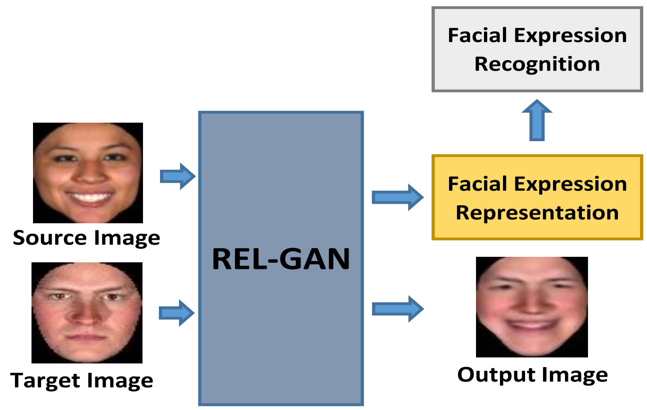

Fig. 1. REL-GAN takes a source image and target image as inputs and generates a synthetic image by transferring the expression of the source image to the identity of the target image. After training, REL-GAN is used to extract disentangled facial expression representation for FER.

and entertainment domain, besides the challenging research problems that it offers. The main goal of face reenactment is to transfer the expression information from the source image to the identity of the target face. Therefore, an ideal face reenactment technique should be able to disentangle expression features from the identity information of the source image and transfer the disentangled expression features to the identity of the target image. In this paper, we employ a novel Reenactment-based Expression-Representation Learning Generative Adversarial Network (REL-GAN) that learns to disentangle expression features from the source image and synthesize an expression image by transferring the disentangled expression features to the identity of the target image. The overall framework of our proposed FER technique is shown in Figure 1.

The architecture of REL-GAN is based on an encoder-decoder based generator G, that contains two encoders, an expression encoder G_{es} and an identity encoder G_{et}. These two encoders are then connected to a decoder G_{de}. The discriminator D of REL-GAN is designed to be a multi-task CNN. During training, the input to REL-GAN is a source image x_s, and a target image x_t, and the output of REL-GAN is a synthesized expression image $\bar{x}$. The goal of G_{es} is to map the source image x_s to an expression representation $f(e)$, while G_{et} is used to project the target image x_t to an identity embedding $f(i)$. The concatenation of the two embeddings,: $f(x) = f(e) + f(i)$, bridges the two encoders with a decoder G_{de}. The objective of decoder G_{de} is to synthesize an expression image $\bar{x}$ having the expression e of the source image and the identity i of the target image: $\bar{x} = G_{de}(f(x))$. The disentangled facial expression representation learned by the encoder G_{es} is mutually exclusive of identity information, which can be best used for FER. Thus, generator, G, is used for two purposes: 1. to disentangle facial expression features employing encoder G_{es}, 2. to reconstruct a facial expression image by transferring the expression information from the source image to the identity of the target image. The discriminator, D, of REL-GAN is used to classify not only between real and fake images but to also perform the classification of identities and facial expressions. The estimation of facial expressions and identities in the discriminator helps in improving the quality of generated images during training.

The main contributions of this paper are as follows:

- To the best of our knowledge, this is the first technique that employs the concept of face reenactment for the task of learning disentangled expression representation.
- We present a novel disentangled and discriminative facial expression representation learning technique for FER by employing concepts of adversarial learning and learning by reconstruction.
- Experimental results show that the proposed framework generalizes well to identities from various ethnic backgrounds and expression images captured both in lab-controlled and in-the-wild settings.

2 Related Work

The main goal of FER is to extract features that are discriminative and invariant to variations such as pose, illumination, and identity-related information. The feature extraction process can be divided into two main categories: human-engineered features and learned features. Before the deep learning era, most of FER techniques involved human-designed features using techniques such as Histograms of Oriented Gradients (HOG) [8], Scale Invariant Feature Transform (SIFT) features [9] and histograms of Local Phase Quantization (LPQ) [10].

The human-crafted features perform well in lab-controlled environment where the expressions are posed by the subjects with constant illumination and stable head pose. However, these features fail on spontaneous data with varying head position and illumination. Recently, deep CNN based methods [11], [12], [13], [14], [40] have been employed to increase the robustness of FER to real-world scenarios. However, the learned deep representations used for FER are often influenced by large variations in individual facial attributes such as ethnicity, gender, and age of subjects involved in training. The main drawback of this phenomenon is that it negatively affects the generalization capability of the model and, as a result, the FER accuracy is degraded on unseen subjects. Although significant progress has been made in improving the performance of FER, the challenge of mitigating the influence of inter-subject variations on FER is still an open area of research. Various techniques [15], [16], [20], [39] have been proposed in the literature to increase the discriminative property of extracted features for FER by increasing the inter-class differences and reducing intra-class variations. Most recently, Identity-Aware CNN (IACNN) [17] was proposed to

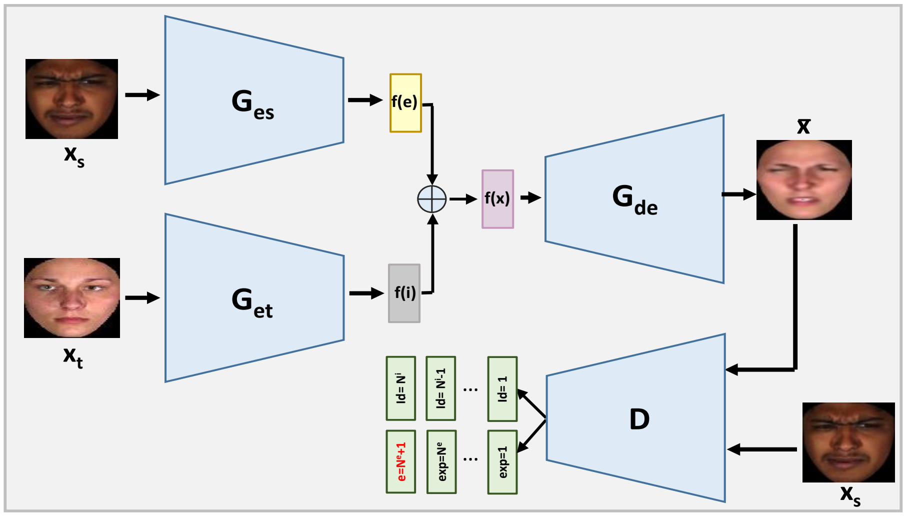

Fig. 2. Architecture of our REL-GAN

enhance FER performance by using an expression-sensitive contrastive loss and an identity-sensitive contrastive loss to reduce the effect of identity related information. However, the effectiveness of contrastive loss is negatively affected by large data expansion, which is caused due to the compilation of training data in the form of image pairs [2]. Similarly, in [7] and [2], the effect of identity-related information is mitigated by generating an expression image with a fixed identity. The main problem with these methods is that the performance of FER depends on the quality of the synthesized images because FER is performed employing the generated images. In [4], person-independent expression representations are learned by using residue learning. However, this technique, apart from being computationally costly, does not explicitly disentangle the expression features from the identity information, because the same intermediate representation is used to generate neutral images of same identities.

3 Proposed Method

The proposed FER technique consists of two stages: during the first stage of learning, a disentangled expression representation $f(e)$ is learned by employing face reenactment, and during the second stage of learning, the disentangled expression representation $f(e)$ is used for facial expression recognition. The overall architecture of REL-GAN is shown in Figure 2. The generator G of REL-GAN is based on an encoder-decoder structure, while the discriminator D is a multi-task CNN [5]. Given a source image $x_s \in X$ and a target image $x_t \in X$, the goal of REL-GAN is to transfer the expression of x_s to the identity of x_t and generate an expression image $\bar{x}$ similar to the ground-truth image x_{t_g}. Specifi-

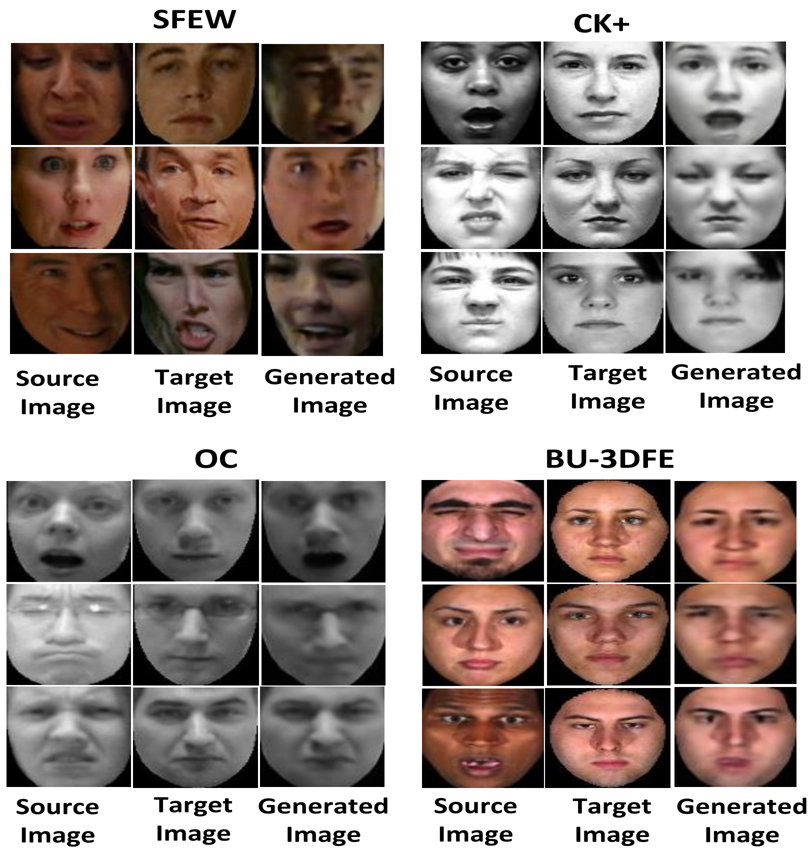

Fig. 3. REL-GAN extracts expression information from source images and transfers the disentangled expression information to the identity of target images.

cally, an encoder $G_{es} : X \rightarrow E$ is used to encode the expression representation $f(e) \in E$ from x_s, and an encoder $G_{et} : X \rightarrow I$ is employed to encode the identity representation $f(i) \in I$ from x_t. A decoder $G_{de} : E \times I \rightarrow X$ is then used to map the expression and the identity latent embedding space back to the face space. The synthesized expression image $\bar{x}$ is generated by computing $\bar{x} = G_{de}(G_{es}(x_s), G_{et}(x_t))$. In order to efficiently transfer the expression of x_s to $\bar{x}$ while preserving the identity of x_t, the expression features of x_s should be captured in $f(e)$ in such a way that it does not contain the identity information of x_s. Thus, by explicitly inputting the extracted identity features $f(i)$ of x_t to G_{de}, we are able to disentangle the expression information of x_s from its identity features in $f(e)$. Figure 3 shows the result of transferring the extracted facial expression features from source images to the identity of the target images by employing our face reenactment technique. During the second stage of learning, encoder G_{es} is detached from REL-GAN after training, and all the filters of G_{es} are fixed. The disentangled expression representation $f(e)$ is extracted from the encoder G_{es} and becomes input into a Multilayer Perceptron (MLP) for facial expression recognition.

3.1 Expression Representation Learning

The architectures of the generator G and the discriminator D of REL-GAN is designed to fulfill two main objectives simultaneously: 1) disentangle and transfer a source facial expression to a target face; while 2) preserving the identity information of the target image.

Discriminator: The main objective of D is three-fold: 1. to classify between real and fake images, 2. to categorize facial expressions, and 3. to recognize the identities of expression images. Therefore, discriminator D is divided into two parts: $D = [D^e, D^i]$, where $D^e \in R^{N^{e+1}}$ corresponds to the part of D that is used for the classification of expressions, i.e N^e denotes the number of expressions, and an additional dimension is used to differentiate between real and fake images. Similarly, $D^i \in R^{N^i}$ is the part of D that is used to classify the identities of images, where N^i denotes the number of identities. The overall objective function of our discriminator D is given by the following equation:

$$\begin{aligned}\max_D \mathcal{L}_{\mathcal{D}}(D,G) = \; & E_{\substack{x_s,y_s \sim p_s(x_s,y_s),\\ x_t,y_t \sim p_t(x_t,y_t)}}[\log(D^e_{y^e_s}(x_s)) + \log(D^i_{y^i_t}(x_t))] + \\ & E_{\substack{x_s,y_s \sim p_s(x_s,y_s),\\ x_t,y_t \sim p_t(x_t,y_t)}}[\log(D^e_{N^e+1}(G(x_s,x_t)))]\end{aligned} \tag{1}$$

Given a real expression image x, the first part of the above equation corresponds to the objective function of D to classify the identity and expression of images. While the second part of the equation represents the objective of D to maximize the probability of a synthetic image $\bar{x}$ being classified as a fake. The expression and identity classification in the discriminator D helps in transferring expression information from a source image to a target face while preserving the identity of x_t. y^e denotes the expression label and y^i represents the identity labels in eq(1).

Generator: The generator G of REL-GAN aims to extract expression and identity features from source image x_s and target image x_t, respectively, and to synthesize an image $\bar{x}$ to fool D to classify it to the expression of x_s and the identity of x_t. Therefore, the generator G contains two encoders and a decoder: $G = (G_{es}, G_{et}, G_{de})$. The objective function of G is given by the following equation:

$$\max_G \mathcal{L}_{\mathcal{G}}(D,G) = E_{\substack{x_s,y_s \sim p_s(x_s,y_s),\\ x_t,y_t \sim p_t(x_t,y_t)}}[\log(D^e_{y^e_s}(G(x_s,x_t)) + \log(D^i_{y^i_t}(G(x_s,x_t))] \tag{2}$$

Pixel Loss: The goal of the generator G is to not only extract the disentangled expression features but also generate realistic-looking expression images. Therefore, to overcome the blurriness of generated images we employ a pixel-wise loss [32] in the raw-pixel space.

$$\mathcal{L}_{pixel} = L_1(G_{de}(G_{es}(x_s), G_{et}(x_t)), x_{t_g}). \tag{3}$$

The total REL-GAN loss is given by:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{adv} + \lambda_2 \mathcal{L}_{pixel} + \lambda_3 \mathcal{L}_{D_e} + \lambda_4 \mathcal{L}_{D_i}. \tag{4}$$

Where $\mathcal{L}_{D_e}$ and $\mathcal{L}_{D_i}$ represents the expression and identity classification loss calculated by the discriminator D.

3.2 Facial Expression Recognition

After the first stage of training, the encoder G_{es} is detached from REL-GAN, and all the filters of G_{es} are kept fixed. To learn facial expressions, the disentangled expression representation $f(e)$ is used to train a simple MLP to classify facial expressions.

4 Experiments

The proposed REL-GAN based FER technique is evaluated on four publicly available facial expression databases: CK+ [18], Oulu-CASIA [19], BU-3DFE[28] and Static Facial Expression in the Wild (SFEW)[33].

4.1 Implementation Details

Face detection is performed by employing the technique proposed in [34]. The detected faces are then aligned using the face alignment method proposed by Hassner et. al [35]. Data augmentation is applied to avoid the over-fitting problem by increasing the number of training images. Therefore, five patches of size 96×96 are cropped-out from five different locations: the center and four corners of each image, and each cropped image are then rotated at four angles i.e $-6°$, $-3°$, $3°$, $6°$. Horizontal flipping is also applied on each rotated image.

REL-GAN is initially pre-trained using the BU-4DFE [22] dataset, which consists of 60,600 images from 101 identities. The architecture of both encoders is designed based on five downsampling blocks consisting of a 3×3 stride 1 convolution. The number of channels are 64, 128, 256, 512, 512 and a one 30-dimensional FC layer for expression feature vector $f(e)$, and a 50-dimensional identity representation $f(i)$, constitute G_{es} and G_{et}, respectively. The decoder G_{de} is built on five upsampling blocks containing a 3×3 stride 1 convolution. The number of channels are 512, 256, 128, 64, 32. The multi-task discriminator D network is designed in such a way that the initial four CNN blocks with 16, 32, 64, 128 channels and a 1024-dimensional FC layer are shared between D^e and

Table 1. CK+: 10-fold Average Accuracy for seven expressions classification.

Method	Setting	Accuracy
HOG 3D [8]	Dynamic	91.44
3DCNN [14]	Dynamic	85.90
STM-Explet [25]	Dynamic	94.19
IACNN [17]	Static	95.37
DTAGN [26]	Static	97.25
DeRL [4]	Static	97.30
CNN(baseline)	Static	90.34
REL-GAN(Ours)	Static	**97.41**

D^i. It is then divided into two branches, where, each branch has two additional FC layers with 512 and 256 channels. D^e then has an expression classification layer and D^i has an identity classification layer. The CNN baseline network used in this paper is the encoder G_{es} network with one additional FC layer for expression classification.

For the optimization of the hyper-parameters, we adopted the optimization strategies presented in [23] as part of our technique. Adam optimizer is used with a batch size of 64, learning rate of 0.0001 and momentum of 0.5. We empirically set $\lambda_1 = 0.7$, $\lambda_2 = 20$, $\lambda_3 = 50$ and $\lambda_4 = 30$. REL-GAN is trained for 300 epochs, and the MLP is trained for 50 epochs. Contrary to conventional GAN training strategies mentioned in [24], in later iterations of REL-GAN, when D reaches a near-optimal solution, G is updated more frequently than D, due to the supervised classification provided by the class labels.

4.2 Experimental Results

The **Extended Cohn-Kanade database CK+** [18] is a popular facial expression recognition database that contains 327 videos sequences from 118 subjects. Each of these sequences corresponds to one of seven expressions, i.e., anger, contempt, disgust, fear, happiness, sadness, and surprise, where each sequence starts from a neutral expression to a peak expression. To compile the training dataset, the last three frames of each sequence are extracted, which results in 981 images in total. To perform 10-fold cross validation, the dataset is divided into ten different sets with no overlapping identities.

The average accuracy of 10-fold cross-validation on the CK+ database is reported in Table 1. It can be seen that the proposed method produces a recognition accuracy of 97.41%, which is higher than the accuracy of previous FER methods. Our image-based technique outperforms the sequence-based methods, where the features for FER are extracted from videos or sequences of images.

The **Oulu-CASIA (OC) dataset** [19] used in this experiment corresponds to the section of the OC dataset that is compiled under strong illumination condition using the VIS camera. It contains 480 sequences from 80 subjects, and each sequence is labeled as one of the six basic expressions. The last three frames of each sequence are selected to create a training and testing dataset.

Table 2. Oulu-CASIA: 10-fold Average Accuracy for six expressions classification.

Method	Setting	Accuracy
HOG 3D [8]	Dynamic	70.63
STM-Explet [25]	Dynamic	74.59
PPDN [27]	Static	84.59
DTAGN [26]	Static	81.46
DeRL [4]	Static	88.0
CNN(baseline)	Static	73.14
REL-GAN(Ours)	Static	**88.93**

Table 3. BU-3DFE database: Accuracy for six expressions classification.

Method	Setting	Accuracy
Wang et al.[29]	3D	61.79
Berretti et al.[30]	3D	77.54
Lopes[31]	Static	72.89
DeRL[4]	Static	**84.17**
CNN(baseline)	Static	72.74
REL-GAN(Ours)	Static	**83.46**

The average of the 10-fold person-independent cross-validation accuracy of the proposed method on Oulu-CASIA dataset, as shown in Table 2, demonstrates that the proposed method outperforms all state-of-the-art techniques with average accuracy of 88.93%. The accuracy obtained using the proposed method is much higher than the accuracy of video-based techniques.

The **BU-3DFE database** [28] is a widely used FER database that contains static 3D face models and texture images from 100 subjects from various ethnic backgrounds with a variety of ages. For each subject, there are expression images corresponding to seven expressions (six basic expressions and a neutral expression) and these images are labeled with four different expression intensity levels. During this experiment, we only use texture images corresponding to the last two highest intensity expressions. We perform a 10-fold cross-validation by dividing the dataset into ten different sets in a person-independent manner.

The average of the 10-fold cross-validation on BU-3DFE dataset is shown in Table 3. The recognition accuracy obtained using the proposed REL-GAN-based method is significantly higher than most of the state-of-the-art techniques. The highest accuracy is produced by the DeRL[4] method that involves the computationally costly residue learning process.

The **SFEW dataset** [33] is the most widely used benchmark for facial expression recognition in an unconstrained setting. The SFEW database contains 1,766 images, i.e. 958 for training, 436 for validation, and 372 for testing. All images are extracted from film clips, and each image has been labeled as one of the seven expression categories, i.e., anger, disgust, fear, neutral, happy, sad, and surprise. We validate our technique on the validation set of SFEW because the labels for the test set are held back by the challenge organizer.

Table 4. SFEW: The Average Accuracy on the Validation Set.

Method	Accuracy
AUDN [36]	26.14
STM-Explet [25]	31.73
Dhall et al. [33] (baseline of SFEW)	35.93
Mapped LBP [37]	41.92
FN2EN [38]	**48.19**
CNN(baseline)	29.75
REL-GAN(Ours)	**45.82**

In Table 4, we compare our result with techniques that use the training set of the SFEW dataset to train their models and do not use extra training data. The average accuracy obtained using our proposed method on the validation set of the SFEW dataset is higher than the accuracy of most of the state-of-the-art techniques. The highest accuracy, however, is obtained by the FN2EN [38] method. We hypothesize that it may be due to the reason that, during the first stage of training, the parameters of the convolutional layers of the FN2EN network are made close to the parameters of convolutional layers of the face-net (VGG-16) [41], which is trained on 2.6M face images. Our REL-GAN, on the other hand, is pre-trained on 60,600 images of the BU-4DFE [22] dataset. Table 4 shows that our proposed method produces promising recognition results not only in a lab-controlled setting but it can also be used to classify facial expressions in real-world scenarios.

5 Conclusions

In this paper, we present a novel FER architecture called REL-GAN that employs the concept of face reenactment to disentangle expression representation from identity features. More specifically, an encoder-decoder based generator is used in REL-GAN, in which the disentangled expression representation is learned by transferring the expression features from a source image to the identity of the target image. After training the REL-GAN architecture for face reenactment, the expression encoder is detached from the rest of the network and is used to extract a disentangled expression representation. A simple MLP is then trained using the disentangled expression features to perform facial expression recognition. The proposed method is evaluated on publicly available state-of-the-art databases, and the experimental results show that the accuracy of FER obtained by employing the proposed method is comparable or even better than the accuracy of state-of-the-art facial expression recognition techniques.

References

1. Yang, H., Zhang, Z., Yin, L.: Identity-adaptive facial expression recognition through expression regeneration using conditional generative adversarial networks. In: International Conference on Automatic Face and Gesture Recognition. (2018) 294–301

2. Cai, J., Meng, Z., Khan, A.S., Li, Z., O'Reilly, J., Tong, Y.: Identity-free facial expression recognition using conditional generative adversarial network. CoRR abs/1903.08051 (2019), https://arxiv.org/abs/1903.08051
3. Bai, M., Xie, W., Shen, L.: Disentangled feature based adversarial learning for facial expression recognition. In: IEEE International Conference on Image Processing. (2019) 31–35
4. Yang, H., Ciftci, U., Yin, L.: Facial expression recognition by de-expression residue learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 2168–2177
5. Tran, L., Yin, X., Liu, X.: Disentangled representation learning GAN for pose-invariant face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 1415–1424
6. Song, B.C., Lee, M.K., Choi, D.Y.: Facial expression recognition via relation-based conditional generative adversarial network. In: International Conference on Multimodal Interaction. (2019) 35–39
7. Ali, K., Isler, I., Hughes, C.E.: Facial expression recognition using human to animated-character expression translation. CoRR abs/1910.05595 (2019), https://arxiv.org/abs/1910.05595
8. Klaser, A., Marszałek, M., Schmid, C.: A spatio-temporal descriptor based on 3d-gradients. In: Proceedings British Machine Vision Conference. (2008)
9. Chu, W.S., Torre, F., Cohn, J.F.: Selective transfer machine for personalized facial expression analysis. IEEE transactions on pattern analysis and machine intelligence. 39 (2016) 529–545
10. Jiang, B., Valstar, M.F., Pantic, M.: Action unit detection using sparse appearance descriptors in space-time video volumes. In: International Conference on Automatic Face and Gesture Recognition. (2011) 314–321
11. Kim, B.K., Lee, H., Roh, J., Lee, S.Y.: Hierarchical committee of deep CNNs with exponentially-weighted decision fusion for static facial expression recognition. In: Proceedings of the International Conference on Multimodal Interaction. (2015) 427–434
12. Yu, Z., Zhang, C.: Image based static facial expression recognition with multiple deep network learning. In: Proceedings of the International Conference on Multimodal Interaction. (2015) 435–442
13. Ng, H.W., Nguyen, V.W., Vonikakis, V., Winkler, S.: Deep learning for emotion recognition on small datasets using transfer learning. In: Proceedings of the international conference on multimodal interaction. (2015) 443–449
14. Liu, M., Li, S., Shan, S., Wang, R., Chen, X.: Deeply learning deformable facial action parts model for dynamic expression analysis. In: Asian conference on computer vision. (2014) 143–157
15. Li, S., Deng, W., Du, J.P.: Reliable crowd sourcing and deep locality-preserving learning for expression recognition in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2017) 2852–2861
16. Cai, J., Meng, Z., Khan, A.S., Li, Z., O'Reilly, J., Tong, Y.: Island loss for learning discriminative features in facial expression recognition. In: IEEE International Conference on Automatic Face and Gesture Recognition. (2018) 302–309
17. Meng, Z., Liu, P., Cai, J., Han, S., Tong Y.: Identity-aware convolutional neural network for facial expression recognition. In: IEEE International Conference on Automatic Face and Gesture Recognition. (2017) 558–565
18. Lucey, P, Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-

specified expression. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops. (2010) 94–101

19. Zhao, G., Huang, X., Taini, M., Li, S.Z., Pietikainen, M. : Facial expression recognition from near-infrared videos. Image and Vision Computing 29 (2011) 607–619
20. Ali, K., and Hughes, C.E.: Facial expression recognition using disentangled adversarial learning. CoRR abs/1909.13135 (2019), https://arxiv.org/abs/1909.13135
21. Zadeh, A., Lim, Y.C., Baltrusaitis, T., Morency, L.P.: Convolutional experts constrained local model for 3d facial landmark detection. In: Proceedings of the IEEE International Conference on Computer Vision. (2017) 2519–2528
22. Yin, L., Chen, X., Sun, Y., Worm, T., Reale, M.: A high-resolution 3d dynamic facial expression database. In: IEEE International Conference on Automatic Face and Gesture Recognition. (2008)
23. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. CoRR abs/1511.06434 (2015), https://arxiv.org/abs/1511.06434
24. Goodfellow, I., Abadie, J., Mirza, M., Xu, B., Farley, D.W., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. (2014) 2672–2680
25. Liu, M., Shan, S., Wang, R., Chen, X.: Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2014) 1749–1756
26. Jung, H., Lee, S., Yim, J., Park, S., Kim, J.: Joint fine-tuning in deep neural networks for facial expression recognition. In: Proceedings of the IEEE international conference on computer vision. (2015) 2983–2991
27. Zhao, X., Liang, X., Liu, L., Li, T., Han, Y., Vasconcelos, N., Yan, S.: Peak-piloted deep network for facial expression recognition. In: European conference on computer vision. (2016) 425–442
28. Yin, L., Wei, X., Sun, Y., Wang, J., Rosato, M.J.: A 3d facial expression database for facial behavior research. In: International conference on automatic face and gesture recognition. (2006) 211–216
29. Wang, J., Yin, L., Wei, X., Sun, Y.: 3d facial expression recognition based on primitive surface feature distribution. In: IEEE Computer Vision and Pattern Recognition. (2006) 1399–1406
30. Berretti, S., Del Bimbo, A., Pala, P., Amor, B.B., Daoudi, M.: A set of selected sift features for 3d facial expression recognition. In: International Conference on Pattern Recognition. (2010) 4125–4128
31. Lopes, A.T., de Aguiar, E., De Souza, A.F., Oliveira-Santos, T.: Facial expression recognition with convolutional neural networks: coping with few data and the training sample order. Pattern Recognition 61 (2017) 610–628
32. Song, L., Lu, Z., He, R., Sun, Z., Tan, T.: Geometry guided adversarial facial expression synthesis. In: Proceedings of the ACM International Conference on Multimedia. (2018) 627–635
33. Dhall, A., Murthy, O.R., Goecke, R., Joshi, J., Gedeon, T.: Video and image based emotion recognition challenges in the wild: Emotiw. In: Proceedings of the ACM on International Conference on Multimodal Interaction. (2015) 423–426
34. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters (2016)
35. Hassner, T., Harel, S., Paz, E., Enbar, R.: Effective face frontalization in unconstrained images. IN: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015), 4295-4304

36. Liu, M., Li, S., Shan, S., Chen, X.: Au-aware deep networks for facial expression recognition. In: IEEE International Conference and Workshops on Automatic Face and Gesture Recognition. (2013) 1–6
37. Levi, G., Hassner, T.: Emotion recognition in the wild via convolutional neural networks and mapped binary patterns. In: Proceedings of the ACM on International Conference on Multimodal Interaction. (2015) 503–510
38. Ding, H, Zhou, S.K., Chellappa, R.: Facenet2expnet: Regularizing a deep face recognition net for expression recognition. In: IEEE International Conference on Automatic Face and Gesture Recognition. (2017) 118–126
39. Halawa, M., Wollhaf, M., Vellasques, E., Sanz, U. S., Hellwich, O.: Learning Disentangled Expression Representations from Facial Images. CoRR abs/2008.07001 (2020), https://arxiv.org/abs/2008.07001
40. Alaghband, M., Yousefi, N., Garibay, I.: FePh: An Annotated Facial Expression Dataset for the RWTH-PHOENIX-Weather 2014 Dataset. CoRR abs/2003.08759 (2020), https://arxiv.org/abs/2003.08759
41. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: Proceedings British Machine Vision Conference. (2015)