# Self-supervised Representation Learning from Videos for Facial Action Unit Detection

Yong Li[1,2], Jiabei Zeng[1], Shiguang Shan[1,2,3,4] , Xilin Chen[1,2]
[1]Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing 100190, China
[2]University of Chinese Academy of Sciences, Beijing 100049, China
[3]CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai, 200031, China
[4]Peng Cheng Laboratory, Shenzhen, 518055, China
yong.li@vipl.ict.ac.cn, {jiabei.zeng, sgshan, xlchen}@ict.ac.cn

## Abstract

*In this paper, we aim to learn discriminative representation for facial action unit (AU) detection from large amount of videos without manual annotations. Inspired by the fact that facial actions are the movements of facial muscles, we depict the movements as the transformation between two face images in different frames and use it as the self-supervisory signal to learn the representations. However, under the uncontrolled condition, the transformation is caused by both facial actions and head motions. To remove the influence by head motions, we propose a Twin-Cycle Autoencoder (TCAE) that can disentangle the facial action related movements and the head motion related ones. Specifically, TCAE is trained to respectively change the facial actions and head poses of the source face to those of the target face. Our experiments validate TCAE's capability of decoupling the movements. Experimental results also demonstrate that the learned representation is discriminative for AU detection, where TCAE outperforms or is comparable with the state-of-the-art self-supervised learning methods and supervised AU detection methods.*

## 1. Introduction

Facial actions convey varied and nuanced meanings, including a person's intentions, affective and physical states. To study the facial actions comprehensively, Ekman and Friesen developed the Facial Action Coding System (FACS) which defines a unique set of about 40 atomic non-overlapping facial muscle actions called Action Units (AUs) [8]. AU detection has drawn significant interest from computer scientists and psychologists over recent decades, as it holds promise to an abundance of applications, such as affect analysis, mental health assessment, and human computer interaction.

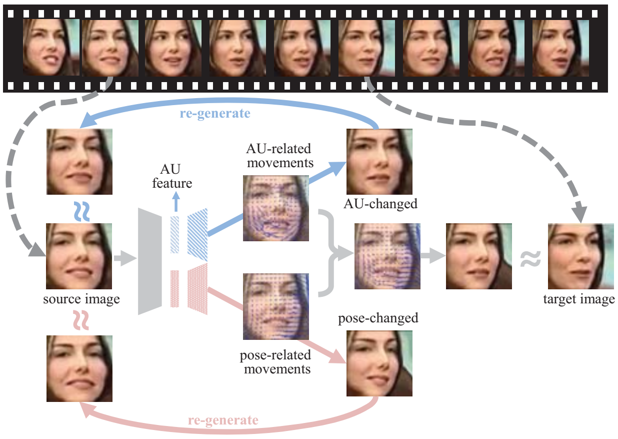


Figure 1. Main idea of the proposed self-supervised learning framework Twin-Cycle Autoencoder (TCAE). TCAE learns AU discriminative features by predicting the disentangled movements that change the AUs and head poses respectively. TCAE ensures the quality of the discovered movements by transforming the AU-changed and pose-changed faces back to the source.

Recently, the development of AU detection is facilitated by the progress in deep learning [47, 22, 21, 32]. However, it is data starved to make full use of these supervised methods, because labelling AUs is time consuming, error prone, and confusing. It takes 30 minutes or more for an FACS expert to manually code an AU for a one-minute video [46].

To alleviate the demand for adequate and accurate annotations, we exploit the practically infinite amount of unlabelled videos to learn discriminative AU representations in a self-supervised manner. Considering that AUs appear as the local movements within the face and the movements are easy be to detected without manual annotations, we propose to use the movements as the supervisory signals in learning the AU representations. However, the detected movements are always caused by both AUs and head motions. In some cases, especially in uncontrolled scenarios, head motions are the dominant contributors to the movements. If we do not remove the movements of head motions from the supervisory signals, the learned features would not be discriminative enough for AU detection, because they would encode more information about head poses than those about AUs.

To address the learning issue from entangled movements, we propose a Twin-Cycle Autoencoder (TCAE) that self-supervisedly learns two embeddings to encode the movements of AUs and head motions, respectively. Fig. 1 illustrates the main idea of the proposed TCAE. We sample two face images (the source image and the target image) of a subject from a video where he/she is talking and moving with varied expressions. TCAE is tasked to change the AUs or head poses of the source frame to those of the target frame by predicting the AU-related and pose-related movements, respectively. Thus, TCAE distills the information required to compute the AU- or pose-related movements separately into the corresponding embeddings. During the training, TCAE enforces the generated face images to be realistic because their quality implies how well the movement is discovered and thus implies how good the representation is. Since we do not have the real face images that merely AUs or poses are changed, we introduce a twin-cycle mechanism to control the quality of the generated face images. In each cycle, either the predicted AU-changed face or the pose-changed face is mapped back to the source. Meanwhile, the AU-related and pose-related movements are combined to map the source to the target. We show that our proposed TCAE can disentangle the movements caused by AUs and head motions.

In summary, our contributions are two folds: 1) We propose a self-supervised learning framework Twin-Cycle Autoencoder (TCAE) to learn AU representations from unlabelled videos. Experimental results show that TCAE outperforms or is comparable to the state-of-the-art self-supervised learning methods and supervised AU detection methods. 2) TCAE can successfully disentangle the AU-related movements from the pose-related ones. It indicates potential applications in editing face images.

## 2. Related work

**Faction unit detection**. AU detection has been studied for decades and various methods have been proposed [24]. To achieve good performance, researchers have designed different features to represent AU. The features include the appearance texture of the whole face [44] or near the facial landmarks [35, 9, 37], or the combination of geometry shape with texture [10]. Most of these features are based on general features in computer vision task, such as SIFT, HOG, LBP, etc. To make the features discriminative for AU, some works considered that AU is tightly correlated to the motions within local regions of the face [13] and thus introduced sparsity-induced algorithms [45, 34] to reduce the influence of uncorrelated facial regions.

Over the last few years, deep learning has become a dominating approach due to their capability and capacity of representation learning. These methods [47, 22, 21, 32] learn rich local features to capture facial deformation. For example, Zhao et al. [47] proposed a locally connected convolutional layer that learns region-specific convolutional filters from sub-areas of the face. JPML [45], EAC-Net [22] and ROI [21] extracted features around facial landmarks that are robust with respect to non-rigid shape changes. JAA-Net [32] proposed to jointly learn AU detection and face alignment in a unified framework. These methods have achieved promising performance on annotated datasets, e.g., CK+ [23], DISFA [25], BP4D [42]. However, these methods depend on accurately labelled images and often overfit on a specific dataset because of insufficient training data.

To alleviate the dependence of AUs annotations, several works start to focus on learning model in a semi-supervised [40, 4], weakly-supervised [46, 31, 43] or self-supervised manner [38]. The semi-supervised learning methods usually incorporate both labelled and unlabelled data by assuming the faces to be clustered by AUs, or to have a smooth label space. The weakly supervised methods exploit noisy, incomplete AU annotations [46]. They usually learn AU classifiers from domain knowledge [43], or naturally existing constraints on AUs [31]. We adopt the self-supervised learning paradigm because it can learn AU discriminative features without AU labels and it is regardless of the assumptions on label distribution.

**Self-supervised learning**. Self-supervised learning adopts supervisory signals that are inferred from the structure of the data itself. The signals include image colorization [41], order of a set of frames [26, 18, 11], camera transformations between pairs of images [1] etc. A typical self-supervised method is SplitBrain [41], which consists of two sub-nets. For the images, each sub-net predicts a subset of the channels according the other subset. The features extracted by the two sub-nets are concatenated and serve as the generic representations.

Since the proposed TCAE is supervised by the disentangled AU-related and pose-related movements, we review the related self-supervised methods that can adopt the motion information as the supervisory signal, or can disentangle different factors.

The former learn visual representations from videos with the help of motion information, e.g., optical flow [36, 29], pixelwise correspondence [12], egomotion [17, 1], or appearance flow [38, 39]. The most related work to TCAE is Fab-Net [38], which is optimized to map a source frame to a target frame by predicting a flow field between them. However, the learned embedding of Fab-Net is not a ded-

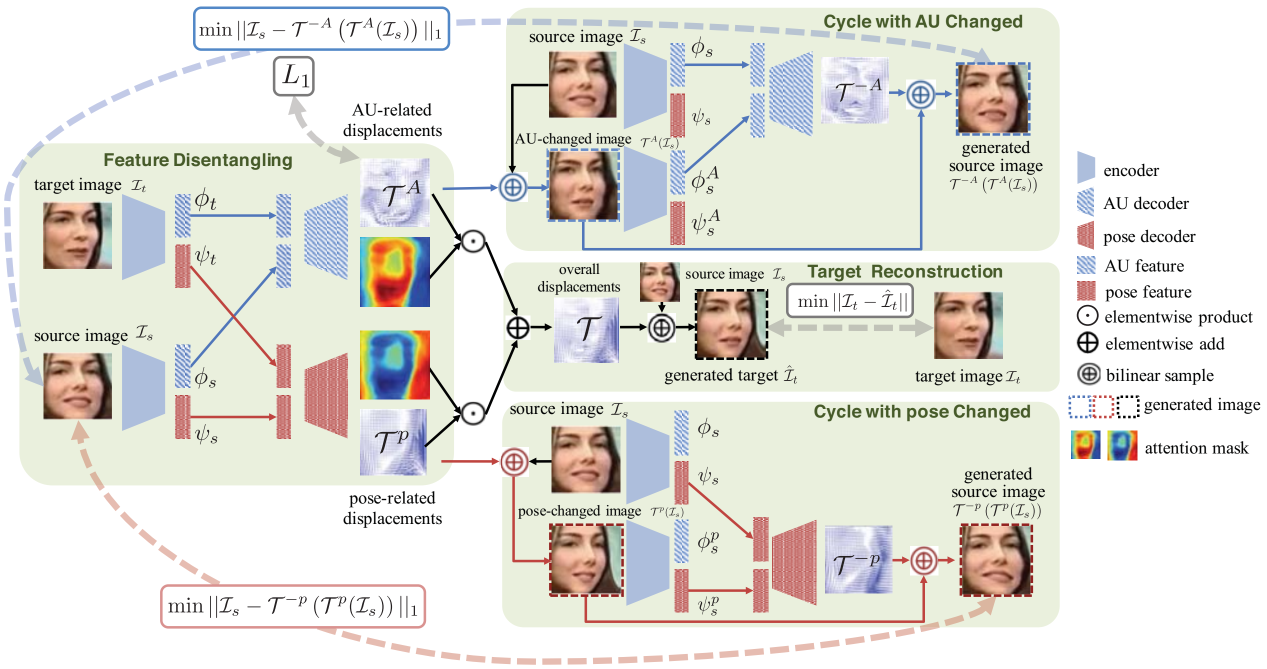

$\min ||\mathcal{I}_s - \mathcal{T}^{-A}(\mathcal{T}^A(\mathcal{I}_s))||_1$
$L_1$
AU-related displacements
Feature Disentangling
target image $\mathcal{I}_t$
$\phi_t$
$\psi_t$
source image $\mathcal{I}_s$
$\phi_s$
$\psi_s$
$\mathcal{T}^A$
$\mathcal{T}^p$
pose-related displacements
$\min ||\mathcal{I}_s - \mathcal{T}^{-p}(\mathcal{T}^p(\mathcal{I}_s))||_1$
Cycle with AU Changed
source image $\mathcal{I}_s$
$\phi_s$
$\psi_s$
AU-changed image $\mathcal{T}^A(\mathcal{I}_s)$
$\phi_s^A$
$\psi_s^A$
$\mathcal{T}^{-A}$
generated source image $\mathcal{T}^{-A}(\mathcal{T}^A(\mathcal{I}_s))$
overall displacements
$\mathcal{T}$
source image $\mathcal{I}_s$
Target Reconstruction
$\min ||\mathcal{I}_t - \hat{\mathcal{I}}_t||$
generated target $\hat{\mathcal{I}}_t$
target image $\mathcal{I}_t$
source image $\mathcal{I}_s$
Cycle with pose Changed
$\phi_s$
$\psi_s$
pose-changed image $\mathcal{T}^p(\mathcal{I}_s)$
$\phi_s^p$
$\psi_s^p$
$\mathcal{T}^{-p}$
generated source image $\mathcal{T}^{-p}(\mathcal{T}^p(\mathcal{I}_s))$
encoder
AU decoder
pose decoder
AU feature
pose feature
elementwise product
elementwise add
bilinear sample
generated image
attention mask

Figure 2. The framework of TCAE. Given a source image $\mathcal{I}_s$ and target image $\mathcal{I}_t$ , TCAE encodes their AU ($\phi_s$ or $\phi_t$) and pose ($\psi_s$ or $\psi_t$) embeddings. The two AU embeddings are decoded into the displacements $\mathcal{T}^A$ reflecting the changing of AUs between $\mathcal{I}_s$ and $\mathcal{I}_t$. Similarly, the two pose embeddings are decoded into the pose-related displacements $\mathcal{T}^p$. In target reconstruction, the integrated displacements are used to transform $\mathcal{I}_s$ to $\mathcal{I}_t$. In the two cycles, TCAE generates an AU-changed face image and a pose-changed face image respectively and then maps them back to the source face image.

icated AU representation. Fab-Net cannot distinguish the information on AUs from that on poses.

The latter disentangle the representation without annotations as the supervisory singals [7, 19, 3, 33]. For example, DRIT [19] factorized an image into the representations in content and attribute space with cross-cycle consistency loss. Dr-Net [7] decomposed a video frame into the stationary component and the temporally varying component by forcing the latter to carry no information about identity. For face analysis, Shu et al. [33] introduced Deforming Autoencoders (DeformAE) that disentangles shape from appearance in a self-supervised manner. In DeformAE, the entangled appearance branch contains an aligned face with no morphing, and the entangled shape branch contains a morphing field that reserve both head pose and facial morphing information.

## 3. Twin-Cycle Autoencoder

We propose a symmetric encoder-decoder architecture called Twin-Cycle Autoencoder (TCAE) to learn AU representations in a self-supervised way. Without the manual annotations, TCAE is trained with pairs of face images of the same person with different facial actions and head poses. Each two face images are sampled from a video where a subject is talking and moving with varied expressions. We denote the two images as the source $\mathcal{I}_s$ and the target $\mathcal{I}_t$. TCAE is trained to respectively change the source face's facial actions or head poses to the ones in the target face.

Fig. 2 illustrates the training framework of TCAE given the two face images. It consists of four parts: feature disentangling, target reconstruction, cycle with pose changed, and cycle with AU changed. In feature disentangling, TCAE learns the features by respectively predicting the AU-related and pose-related movements between two images. In target reconstruction, TCAE integrates the separated movements and uses it in transforming the source image to the target one, ensuring that the two movements are sufficient to represent the changing between the two face images. To make the separated movements realistic, TCAE introduces two cycles with AU or pose changed. It uses the AU-/pose- related movements to generate an AU-/pose-changed face image, which are then transformed back to the source one. TCAE requires the regenerated source image to be consistent to the original one. Below, we present details of the four parts in TCAE.

### 3.1. Feature disentangling

In order to disentangle the information about AUs and poses, TCAE has a nearly symmetric structure with two branches. As can be seen in Fig. 2, TCAE first encodes both the source and target images using the encoder $E$ and gets their embeddings $[\psi_s, \phi_s]$ and $[\psi_t, \phi_t]$, respectively. $\psi_s$ and $\psi_t$ denote the pose-related features. $\phi_s$ and $\phi_t$ denote the AU-related features. Then, TCAE concatenates the two AU-related features $\phi_s$ and $\phi_t$ and passes them into the AU-related decoder $D_A$. $D_A$ decodes how the facial actions in the source face is changed to those in the target face, and where the change happens. Symmetrically, the concatenated pose features $\psi_s$ and $\psi_t$ are passed to the pose-related decoder $D_p$ that decodes how and where the pose is changed. Since the decoders $D_A$ and $D_p$ are in symmetrical branches, we describe one of them in details and the other is similar. Let us take $D_A$ as an example. Since $D_A$ takes the AU-related embeddings of both the source face and target face as the inputs, it is capable to capture the AU-related movement between the two faces. The movement caused by AUs is depicted as the displacements of pixels between the source face and the AU-changed face. The displacements are formulated as a matrix of vectors $\mathcal{T}^A \in \mathbb{R}^{W\times H\times 2}$, where $W$ and $H$ are the width and height of the images. TCAE generates the AU-changed face by sampling the pixels from the source image. $\mathcal{T}^A_{xy} = (\delta x, \delta y)$ is the vector at position $(x, y)$ and it means the offset of the pixel location $(x, y)$ in the source face. That is, the pixel at location $(x, y)$ in the source face is moved to the location $(x+\delta x, y+\delta y)$ in the AU-changed face. For the pixels that do not have corresponding ones in the source face, we adopt bilinear interpolation. Therefore, $\mathcal{T}^A$ serves as an operator $\mathcal{T}^A : \mathcal{I}_s \mapsto \mathcal{I}_A$ which transforms the source face $\mathcal{I}_s$ into the AU-changed face $\mathcal{I}_A$ pixel by pixel. Similar to $\mathcal{T}^A$, we get the pose-related displacements $\mathcal{T}^p$ from the decoder $D_p$.

To distinguish the displacements caused by the change of AU and poses, we add the $L_1$ regularization on $\mathcal{T}^A$ to keep the AU-related movements sparse and subtle, which is formulated as:

$$\mathcal{L}^A_1 = \sum_{x,y} ||\mathcal{T}^A_{xy}||_1, \tag{1}$$

where $x$,$y$ enumerate all the locations in the face images. $\mathcal{T}^A$ should be sparse because facial actions are the movements of one or a group of muscles and they only lead to regional changes of the face pixels. Minimizing the $L_1$-norm of $\mathcal{T}^A$ also enforces the AU-related movements to be subtle. Facial actions appears as motions of smaller range than that of head motions. Therefore, the absolute values of $\mathcal{T}^A$'s elements should be smaller than those in $\mathcal{T}^p$.

### 3.2. Target reconstruction

To ensure that the decoded movements can represent the changing from the source image to the target, we integrate the displacements $\mathcal{T}^A$ and $\mathcal{T}^p$, and then use the integrated displacements to generate a target image from the source.

TCAE integrates $\mathcal{T}^A$ and $\mathcal{T}^p$ by linearly combining each of their elements. Each element of the integrated displacements $\mathcal{T}$ at location $(x, y)$ is computed by:

$$\mathcal{T}_{xy} = \alpha^A_{xy}\mathcal{T}^A_{xy} + \alpha^p_{xy}\mathcal{T}^p_{xy},$$

where $\mathcal{T}_{xy}, \mathcal{T}^A_{xy}, \mathcal{T}^p_{xy} \in \mathbb{R}^2$ are vectors denoting the offsets of the pixel at location $(x, y)$ in the source face. $\alpha^A_{xy}$ and $\alpha^p_{xy}$ are scalers that weighing the contributions of $\mathcal{T}^A_{xy}$ and $\mathcal{T}^p_{xy}$, respectively. They satisfy that $\alpha^A_{xy} + \alpha^p_{xy} = 1$. In the images of size $W \times H$, the weights for all the locations compose an attention mask $\mathbf{A}^A \in \mathbb{R}^{W\times H}$ or $\mathbf{A}^p \in \mathbb{R}^{W\times H}$. The masks are the outputs of decoder $D_A$ and $D_p$, indicating where the AU or pose is changing.

The integrated $\mathcal{T}$ serves as a transforming operator that maps the source image $\mathcal{I}_s$ to the target $\mathcal{I}_t$. Thus, we require the reconstructed target image to be similar to the original one by minimizing their discrepancy. We formulate the reconstruction loss as :

$$\mathcal{L}_{rec} = ||\mathcal{T}(\mathcal{I}_s) - \mathcal{I}_t||_1, \tag{2}$$

where $\mathcal{T}(\mathcal{I}_s) \in \mathbb{R}^{W\times H\times 3}$ denotes the generated RGB images from the source face. $\mathcal{I}_s$ and $\mathcal{I}_t$ are the images of the source and the target face images, respectively.

### 3.3. Twin cycles with AU or pose changed

The quality of the generated AU-changed and pose-changed face images implies how well the movements are disentangled. Since we have no pixel-wise or label level supervisions for the generated face, we exploit the property that the transformation should be "cycle consistent" [48].

TCAE includes two symmetric cycles. In one of them, a face image is generated by changing the facial actions of the source. Then, the AU-changed face is transformed back to the source. In the other, we generate a pose-changed face image and then transform it back to the source.

Fig. 2 illustrates the details in the cycles. Let us take the cycle with AU changed as the example. After we get the AU-related displacements $\mathcal{T}^A$ in the feature disentangling part, we use them to generate the AU-changed face image $\mathcal{T}^A(\mathcal{I}_s)$ from the source. We extract the AU and pose features of the AU-changed face image using the same encoder $E$ in the feature disentangling part. Then, the AU features of the source and the AU-changed face images are concatenated and fed into the AU-related decoder $D_A$. By $D_A$, we get the displacements $\mathcal{T}^{-A}$ which change the facial actions in AU-changed face to those in the source. It is noting that the only differences between the source image $\mathcal{I}_s$ and the AU-changed one $\mathcal{T}^A(\mathcal{I}_s)$ are the facial actions. If we change the facial actions of $\mathcal{T}^A(\mathcal{I}_s)$ using $\mathcal{T}^{-A}$, the new image $\mathcal{T}^{-A}\left(\mathcal{T}^A(\mathcal{I}_s)\right)$ should be similar to $\mathcal{I}_s$. Therefore, we formulate a cycle consistent reconstruction loss to minimize the pixel discrepancy between $\mathcal{T}^{-A}\left(\mathcal{T}^A(\mathcal{I}_s)\right)$ and $\mathcal{I}_s$ as:

$$\mathcal{L}^A_{cyc} = ||\mathcal{T}^{-A}\left(\mathcal{T}^A(\mathcal{I}_s)\right) - \mathcal{I}_s||_1. \tag{3}$$

Similarly, in the cycle with pose changed, we get the displacements $\mathcal{T}^{-p}$ that change the pose in pose-changed face

$\mathcal{T}^p(\mathcal{I}_s)$ to that in the source. we formulate the cycle consistent reconstruction loss to minimize the pixel discrepancy between $\mathcal{T}^{-p}(\mathcal{T}^p(\mathcal{I}_s))$ and $\mathcal{I}_s$ as:

$$\mathcal{L}_{cyc}^p = \ ||\mathcal{T}^{-p}(\mathcal{T}^p(\mathcal{I}_s)) - \mathcal{I}_s||_1, \tag{4}$$

where $\mathcal{T}^{-p}(\mathcal{T}^p(\mathcal{I}_s))$ is the generated face image from the pose-changed face using $\mathcal{T}^{-p}$.

Besides the pixel level consistency, we also exploit the consistency within the embeddings. In the cycle with AU changed, the AU-changed face image and the source image are of the same pose. Thus, their pose embeddings should be similar. Meanwhile, the AU-changed face image and the target image are of the same facial actions. Thus, their AU embeddings should be similar. We minimize the discrepancy of the embeddings for AU-changed face image by

$$\mathcal{L}_{emb}^A = \ ||\psi_s^A - \psi_s||^2 + ||\phi_s^A - \phi_t||^2, \tag{5}$$

where $\psi_s^A$ and $\phi_s^A$ are the pose embeddings and AU embeddings of the AU-changed face image, respectively. $\psi_s$ is the pose embeddings of the source image and $\phi_t$ is the AU embeddings of the target image.

Similarly, in the cycle with pose changed, the pose-changed face image should have similar AU embeddings to the source image, and have similar pose embeddings to the target image. The consistency of the embeddings are constrained by

$$\mathcal{L}_{emb}^p = \ ||\phi_s^p - \phi_s||^2 + ||\psi_s^p - \psi_t||^2, \tag{6}$$

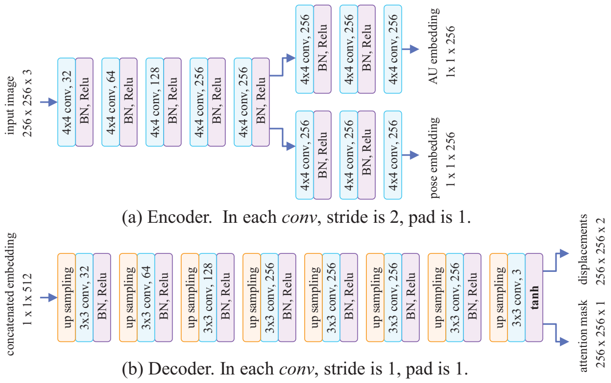
where $\phi_s^p$ and $\psi_s^p$ are the AU embeddings and pose embeddings of the pose-changed face image, respectively. $\phi_s$ is the AU embeddings of the source image and $\psi_t$ is the pose embeddings of the target image.

## 4. Experimental results

In this section, we validated the effectiveness of the proposed TCAE. First, we compared TCAE with other self-supervised methods, descriptors, and supervised AU detection methods on three AU datasets. Then, we analyzed the generated face images and displacements.

### 4.1. Implementation details

**Detailed structures of the encoder and decoders:** Fig. 3 (a) illustrates the encoder used in our experiments. It contains a backbone network followed by two parallel branches. The backbone is shared because the features from early layers are usually general. The encoder takes an RGB image in size of $256\times256$ as the input and outputs two 256-dimensional embeddings that represent AU and head pose, respectively. Fig. 3 (b) shows the decoder which contains eight blocks. In each block, an upsampling layer is placed before the convolution layer to double the width and height of the input feature maps. The last block uses *tanh* as the activation function. The decoder outputs a three-channel feature map with in the size of $256 \times 256 \times 3$. The first two channels serve as the displacements, and the third serves as the attention mask. It is noting that the en/de-coder can be with any suitable blocks (e.g., residual blocks) rather than the vanilla convolutional blocks in our experiments.



(a) Encoder. In each *conv*, stride is 2, pad is 1.

(b) Decoder. In each *conv*, stride is 1, pad is 1.

Figure 3. Structure of the encoder (top) and decoders (bottom) in TCAE. BN denotes batch normalization.

**Training of TCAE:** TCAE was trained end-to-end by minimizing the combination of losses in Eq. (1)∼(6). The full loss is formulated as $\mathcal{L} = \frac{1}{W\times H\times 3}\mathcal{L}_{rec} + \frac{\lambda_1}{W\times H\times 2}\mathcal{L}_1^A + \frac{\lambda_2}{W\times H\times 3}(\mathcal{L}_{cyc}^A + \mathcal{L}_{cyc}^p) + \frac{\lambda_3}{256}(\mathcal{L}_{emb}^A + \mathcal{L}_{emb}^p)$, where $W$ and $H$ denote the width and height of the image. The weights $\lambda_1$, $\lambda_2$ and $\lambda_3$ are set as $0.01$, $0.1$, $0.1$. We implemented TCAE[1] using PyTorch [30] and optimized it by SGD with an initial learning rate of 0.001, and batch size of 64. It took around $1000$ epochs to reach the convergency.

TCAE was trained on the union of VoxCeleb1 [28] and VoxCeleb2 [6] datasets. The two datasets consist of videos of interviews containing around 7,000 subjects. The identities were randomly split as train/val/test with percentages of 75/15/10. During training, TCAE was fed with face pairs that were randomly sampled from a video in the merged VoxCeleb. The faces were detected by Cadcade-CNN [20] and aligned according to the facial landmarks[16]. Each face was cropped to a $256 \times 256$ image.

We adopted the curriculum learning [2] strategy to train the model with progressively difficult samples, because the randomly sampled image pairs may contain large deviations, which are too challenging to learn. Given a batch of image pairs, TCAE executed a forward pass to obtain the loss for each sample. The samples were sorted by their losses in an ascending order. Then the loss $\mathcal{L}$ was only back-propagated to the samples ranking in the top 50% within the batch in the beginning. It was back-propagated to the samples ranking between top 10% and 60% when $\mathcal{L}$ on the validation set saturated, and to the ones ranking between top 20% and 70% when another saturation reached. We kept

[1] Code available at https://github.com/mysee1989/TCAE

Table 1. F1 on BP4D dataset.

| | Methods/AU | 1 | 2 | 4 | 6 | 7 | 10 | 12 | 14 | 15 | 17 | 23 | 24 | ave |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Descriptor | Handcrafted [40]* | 43.4 | 40.7 | 43.3 | 59.2 | 61.3 | 62.1 | 68.5 | 52.5 | 36.7 | 54.3 | 39.5 | 37.8 | 50.0 |
| | ResNet-80 face | 39.3 | 40.6 | 38.5 | 64.2 | 67.5 | 71.0 | 65.3 | 57.2 | 37.8 | 51.3 | 35.1 | 32.6 | 49.9 |
| | VGG emotion | 46.4 | 36.3 | 49.6 | 76.0 | 77.6 | 80.2 | 87.8 | 60.8 | 40.4 | 59.1 | 43.7 | 48.2 | 58.8 |
| Supervised | AlexNet [5]* | 40.3 | 39.0 | 41.7 | 62.8 | 54.2 | 75.1 | 78.1 | 44.7 | 32.9 | 47.3 | 27.3 | 40.1 | 48.6 |
| | DRML [47]* | 36.4 | 41.8 | 43.0 | 55.0 | 67.0 | 66.3 | 65.8 | 54.1 | 33.2 | 48.0 | 31.7 | 30.0 | 48.3 |
| | EAC-Net [22]* | 39.0 | 35.2 | 48.6 | 76.1 | 72.9 | 81.9 | 86.2 | 58.8 | 37.5 | 59.1 | 35.9 | 35.8 | 55.9 |
| | ROI [21]* | 36.2 | 31.6 | 43.4 | 77.1 | 73.7 | 85.0 | 87.0 | 62.6 | 45.7 | 58.0 | 38.3 | 37.4 | 56.4 |
| | JAA-Net [32]* | 47.2 | 44.0 | 54.9 | 77.5 | 74.6 | 84.0 | 86.9 | 61.9 | 43.6 | 60.3 | 42.7 | 41.9 | 60.0 |
| Self-supervised | SplitBrain [41] | 39.0 | 32.0 | 39.7 | 72.9 | 70.6 | 78.2 | 83.7 | 57.8 | 37.3 | 53.6 | 32.3 | 45.1 | 53.5 |
| | DeformAE [33] | 39.5 | 34.5 | 40.8 | 70.5 | 68.4 | 76.3 | 82.9 | 60.7 | 23.1 | 54.1 | 34.3 | 43.1 | 52.3 |
| | Fab-Net [38] | 43.3 | 35.7 | 41.6 | 72.9 | 63.0 | 75.9 | 83.5 | 57.7 | 26.5 | 48.2 | 33.6 | 42.4 | 52.0 |
| | **TCAE (ours)** | **43.1** | **32.2** | **44.4** | **75.1** | **70.5** | **80.8** | **85.5** | **61.8** | **34.7** | **58.5** | **37.2** | **48.7** | **56.1** |

* means that the values are reported in the original papers.

Table 2. F1 on DISFA dataset.

| | Methods/AU | 1 | 2 | 4 | 6 | 9 | 12 | 25 | 26 | ave |
|---|---|---|---|---|---|---|---|---|---|---|
| Descriptor | ResNet-80 face | 24.9 | 17.9 | 49.5 | 41.2 | 26.2 | 48.6 | 56.4 | 32.8 | 37.2 |
| | VGG emotion | 35.5 | 25.5 | 58.1 | 53.8 | 32.4 | 74.4 | 79.0 | 55.7 | 51.8 |
| Supervised | DRML [47]* | 17.3 | 17.7 | 37.4 | 29.0 | 10.7 | 37.7 | 38.5 | 20.1 | 26.7 |
| | EAC-Net [22]* | 41.5 | 26.4 | 66.4 | 50.7 | 80.5 | 89.3 | 88.9 | 15.6 | 48.5 |
| | JAA-Net [32]* | 43.7 | 46.2 | 56.0 | 41.4 | 44.7 | 69.6 | 88.3 | 58.4 | 56.0 |
| Self-supervised | SplitBrain [41] | 13.1 | 10.6 | 35.7 | 40.2 | 30.2 | 57.5 | 77.4 | 40.3 | 38.1 |
| | DeformAE [33] | 17.6 | 12.3 | 46.7 | 43.5 | 26.0 | 62.7 | 64.8 | 47.6 | 40.1 |
| | Fab-Net [38] | 15.5 | 16.2 | 43.2 | 50.4 | 23.2 | 69.6 | 72.4 | 42.4 | 41.6 |
| | **TCAE (ours)** | **15.1** | **15.2** | **50.5** | **48.7** | **23.3** | **72.1** | **82.1** | **52.9** | **45.0** |

* means that the values are reported in the original papers.

changing the samples until the max iteration.

**Evaluation protocols:** After the training process, we obtained the encoder for AU detection. We trained a linear AU classifier from the learned embedding. The linear classifier consists of two layers: a batch-norm layer followed by a linear fully connected layer with no bias. The linear classifier was trained with a binary cross entropy loss for each AU. As all the AU datasets are highly imbalanced, the samples from the under-represented categories were reweighed inversely proportionally to the class frequencies. We adopted F1 score ($F1 = \frac{2RP}{R+P}$) to evaluate the performance of the method, where $R$ and $P$ denote recall and precision, respectively. We also computed the average over all AUs (ave) to measure the overall performance.

**Evaluation datasets:** We evaluated the methods on BP4D [42], GFT [13] and DISFA [25] datasets. **BP4D** contains 41 participants (23 females and 18 males). There are about 146000 frames with available AU labels. **DISFA** consists of 26 participants. The AUs are labelled with intensities from 0 to 5. The frames with intensities greater than 1 were considered as positive, while others were treated as negative. We totally obtained about 130,000 AU-labelled frames. **GFT** contains 96 participants in 32 three-person groups. The moderate out-of-plane head motion and occlusion make AU detection challenging. For BP4D and DISFA dataset, we split the dataset into 3 folds based on subject IDs and conducted a 3-fold cross-validation. We used 12 AUs in BP4D dataset and and 8 AUs in DISFA dataset for evaluation. For GFT dataset, we followed the original train/test splits in [13] (about 108000 facial images for training and 24600 images for evaluation) and used totally 10 AUs for evaluation.

## 4.2. Comparisons with other methods

We compared TCAE with the state-of-the-art self-supervised methods, typical descriptors, and supervised AU detection methods. Table 1, 2, 3 report the F1-score of the methods on BP4D, DISFA, and GFT datasets.

**Comparison with other self-supervised methods:** The TCAE was compared with the state-of-the-art self-supervised learning methods: SplitBrain [41], DeformingAE [33], Fab-Net [38]. We re-trained the three models on the merged VoxCeleb dataset using the released codes. In SplitBrain [41], we used the down-sampled output of conv3

Table 3. F1 on GFT dataset.

| | Methods/AU | 1 | 2 | 4 | 6 | 10 | 12 | 14 | 15 | 23 | 24 | ave |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Descriptor | Handcrafted [13]* | *38* | *32* | *13* | *67* | *64* | *78* | *15* | *29* | *49* | *44* | *42.9* |
| | ResNet-50 face | 24.3 | 50.7 | 18.2 | 39.9 | 44.7 | 41.6 | 17.4 | 27.8 | 31.0 | 25.9 | 32.2 |
| | VGG emotion | 23.9 | 40.6 | 26.4 | 73.6 | 69.3 | 74.4 | 21.1 | 24.9 | 26.0 | 20.2 | 40.0 |
| Supervised | AlexNet [13]* | *44* | *46* | *2* | *73* | *72* | *82* | *5* | *19* | *43* | *42* | *42.8* |
| | ResNet-50 | 23.5 | 37.8 | 3.5 | 79.1 | 70.1 | 82.1 | 20.9 | 11.7 | 49.1 | 40.3 | 41.8 |
| Self-supervised | SplitBrain [41] | 19.0 | 40.6 | 8.7 | 60.2 | 66.6 | 75.4 | 5.6 | 26.7 | 22.9 | 32.3 | 35.8 |
| | DeformAE [33] | 17.3 | 40.1 | 4.8 | 64.1 | 69.1 | 72.1 | 7.8 | 3.9 | 8.0 | 25.2 | 31.2 |
| | Fab-Net [38] | 44.4 | 42.3 | 9.4 | 60.6 | 68.7 | 70.4 | 8.7 | 1.7 | 5.5 | 20.8 | 33.3 |
| | **TCAE (ours)** | **43.9** | **49.5** | **6.3** | **71.0** | **76.2** | **79.5** | **10.7** | **28.5** | **34.5** | **41.7** | **44.2** |

* means that the values are reported in the original papers.

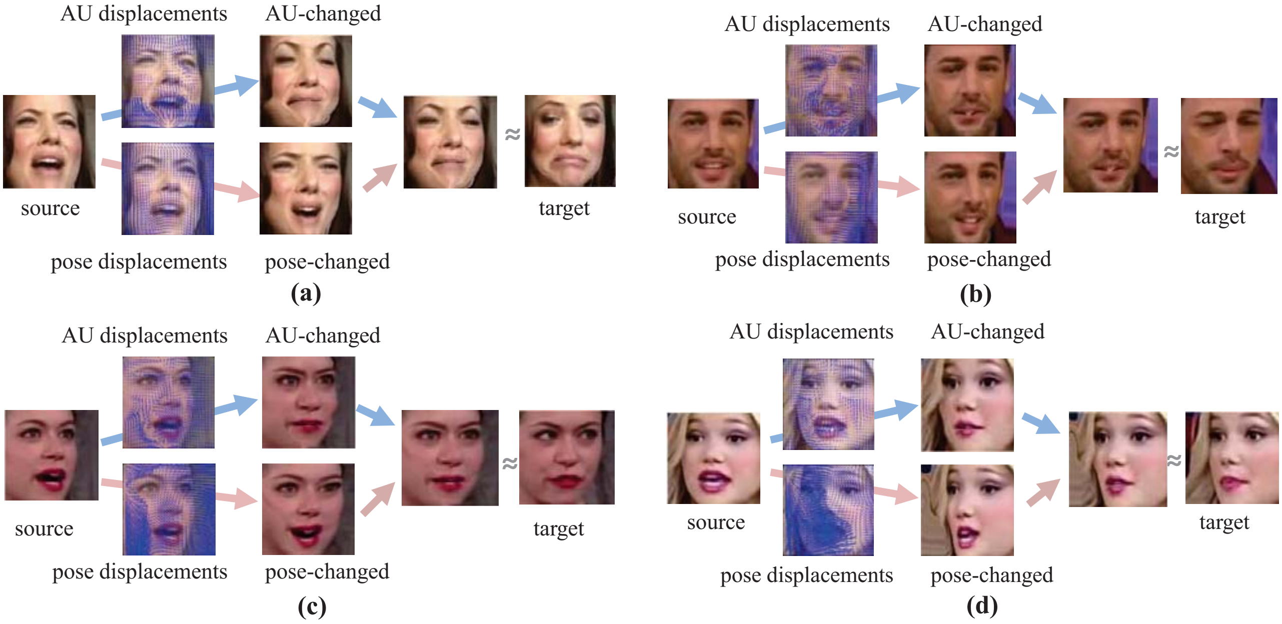


Figure 4. Visualizations of the displacements and the generated face images. The source is transformed to the AU-changed and pose-changed face images through the AU displacements and pose displacements respectively. AU-changed face image should has the same AUs as the target and the same pose as the source. Pose-changed face image should has the same pose as the target and the same AUs as the source. Better viewed in color.

layer as the feature. In DeformAE [33], we changed the input size to 256 × 256 and used the 512 dimensional latent representation as the feature. In Fab-Net, we followed the settings in [38]. All the features were used in training an AU detector as they were used in TCAE.

As shown in table 1, 2, 3, TCAE outperforms other self-supervised methods in the average F1 score. Because the decoupled AU representation can better reflect the facial actions. The advantage of TCAE is the most obvious on GFT dataset. It suggests that the decoupled AU representation is robust against head pose variation, while the entangled representation in [38, 33] is vulnerable to the head pose.

Although TCAE achieved the best average F1 score, it had inconsistent performance on different AUs. TCAE failed on AU9 because AU9 is nose wrinkle and it cannot be generated by moving pixels in the faces without AU9. TCAE showed its success on AU6 (cheek raiser), AU7 (lid tightener), AU10 (upper lip raiser), AU12 (lip corner puller), AU23 (lip tightener), AU24 (lip pressor), AU25 (lips part) etc. It is because these AUs can be easily generated by moving pixels in the source face.

**Comparison with other descriptors:** We compared TCAE with the handcrafted features [40, 13], face and emotion descriptors. The face descriptor was extracted from a 80 layer ResNet [15] trained on MS-Celeb-1M dataset [14] for face recognition. The emotion descriptor was extracted from a VGG-16 net trained on AffectNet dataset [27] for emotion classification.

The TCAE-learned AU representation outperforms the handcrafted features because the handcrafted features are general and are not specially designed for AU detection. TCAE also outperforms the face descriptor because the face

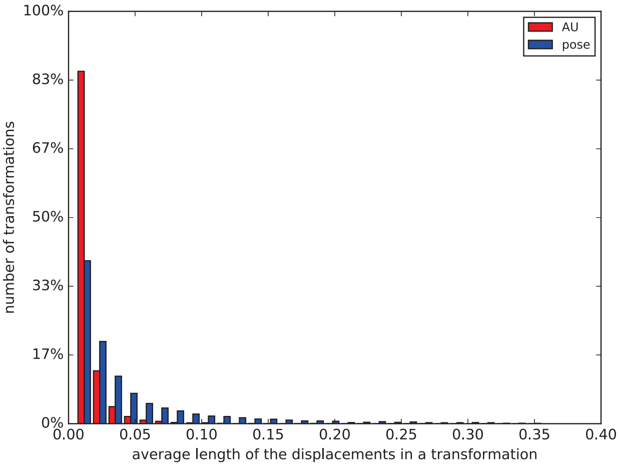


Figure 5. Histogram of the average length of displacements for AU and pose displacements.

descriptor was optimized to be expression invariant. On GFT dataset, TCAE outperforms the emotion descriptor. But TCAE lags behind the emotion descriptor on BP4D and DISFA datasets. It is because the emotion descriptor is emotion discriminative and emotions are tightly correlated with AUs, e.g., activated AU1 (inner brow raiser) often appears in facial expression images labelled with *fear* or *surprise*. Activated AU4 (brow lower) is a partial indicator of *anger* and *sadness*. Thus, emotion descriptor is also AU discriminative. However, the emotion descriptor depends on a large amount of labelled data. It is not easy to annotate emotions.

**Comparison with supervised methods:** We compared TCAE with the state-of-the-art AU detection methods, including DRML [47], EAC-Net [22], ROI [21], JAA-Net [32]. For a fair comparison, TCAE was evaluated under the same protocol as the ones in [47, 22, 21, 32]. On GFT dataset, we trained a 50-layer ResNet [15] from scratch .

TCAE is comparable to supervised methods. It outperforms AlexNet [5], DRML [47], EAC-Net [22] on BP4D dataset, and outperforms DRML [47] on DISFA dataset. On GFT dataset, TCAE outperforms the supervised methods [13] with no exceptions. TCAE is comparable with ROI [21] on BP4D dataset. TCAE lags behind JAA-Net [32] and EAC-Net [22] on DISFA dataset, which both adopt facial landmarks to learn region specific representations.

### 4.3. Analysis

To investigate how well the movements are disentangled, we visualized and analyzed the generated faces and the learned displacements.

**Generated face images:** Fig. 4 visualizes the generated faces and learned displacements during the training of TCAE. TCAE shows its capability in separating the changes caused by facial actions and head motions, as it generated reasonable AU-changed and pose-changed face images. As can be seen in each sub-figure, the AU-changed face images preserve the poses in the source but have similar facial actions to the targets, e.g., the open mouth in (a), the stretched mouth in (b), the lower lip corner in (c), and the closed eyes in (d). Meanwhile, the pose-changed face images preserve the facial actions in the source but have similar poses to the targets. Despite some defections, the generated target face images look similar to the real targets. It indicates that the accumulative effect of the AU and pose displacements compose the changing between the source and the target. All the generated faces look like real ones, thus TCAE shows its potential applications in editing faces.

**Displacements:** We analyse the displacements both quantitatively and qualitatively. Fig. 4 visualizes the AU and pose displacements. The pose displacements are in nearly homogeneous directions while the AU displacements have diverse directions, because pose displacements reflect the rigid motion of the head while AU displacements reflect the non-rigid motion of facial muscles. We randomly sampled 6400 image pairs and calculated the per vector length in AU displacements as $\mathcal{T}_i^A = \sqrt{(\mathcal{T}_{x_i}^A)^2 + (\mathcal{T}_{y_i}^A)^2}$, where $(\mathcal{T}_{x_i}^A, \mathcal{T}_{y_i}^A)$ denotes the AU offset at position $(x_i, y_i)$. The AU displacements between a image pair were averaged as $L_{ave}^A = \frac{1}{N}\sum_{i=1}^{N} \mathcal{T}_i^A$, where $N = W \times H$ denotes product of the image width and height. $L_{ave}^p$ was calculated in a similar manner. Fig. 5 plots the histogram of $L_{ave}^A$ and $L_{ave}^p$ from the overall image pairs. As can be seen, the AU displacements are shorter in the average length than the pose displacements. The overall average length of the pose displacements is $0.044$. It is five times larger than that of the AU displacements, which is 0.008. The reason is that AUs are local movements within the face but head motions are relatively large and global movements.

## 5. Conclusion

This paper presented a Twin-Cycle Autoencoder (TCAE) to learn discriminative representations for AU detection in a self-supervised manner. TCAE successfully disentangled the AU representation from the poses by factorizing the movement between two faces into the AU-related and pose-related displacements. The decoupled AU representation is discriminative for AU detection. Extensive experiments demonstrated that TCAE outperformed or was comparable with the state-of-the-art self-supervised learning methods and supervised AU detection methods. The proposed TCAE can be further used in editing face or decoupling other factors.

**Acknowledgement:** This work is partially supported by National Key R&D Program of China (No. 2017YFA0700800) and National Natural Science Foundation of China (No. 61702481). We also thank Zijia Lu for discussion. J. Zeng, S. Shan, and X. Chen are the corresponding co-authors of this paper.

## References

[1] Pulkit Agrawal, Joao Carreira, and Jitendra Malik. Learning to see by moving. In *CVPR*, 2015.

[2] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *ICML*. ACM, 2009.

[3] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *NIPS*, 2016.

[4] Wen-Sheng Chu, Fernando De la Torre, and Jeffery F Cohn. Selective transfer machine for personalized facial action unit detection. In *CVPR*, 2013.

[5] Wen-Sheng Chu, Fernando De la Torre, and Jeffrey F Cohn. Learning spatial and temporal cues for multi-label facial action unit detection. In *FG*. IEEE, 2017.

[6] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*, 2018.

[7] Emily L Denton et al. Unsupervised learning of disentangled representations from video. In *NIPS*, 2017.

[8] Paul Ekman and Wallace V Friesen. *Manual for the facial action coding system*. Consulting Psychologists Press, 1978.

[9] Stefanos Eleftheriadis, Ognjen Rudovic, and Maja Pantic. Multi-conditional latent variable model for joint facial action unit detection. In *CVPR*, 2015.

[10] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *CVPR*, 2016.

[11] Basura Fernando, Hakan Bilen, Efstratios Gavves, and Stephen Gould. Self-supervised video representation learning with odd-one-out networks. In *CVPR*. IEEE, 2017.

[12] Chuang Gan, Boqing Gong, Kun Liu, Hao Su, and Leonidas J Guibas. Geometry guided convolutional neural networks for self-supervised video representation learning. In *CVPR*, 2018.

[13] Jeffrey M Girard, Wen-Sheng Chu, László A Jeni, and Jeffrey F Cohn. Sayette group formation task (gft) spontaneous facial expression database. In *FG*, pages 581–588. IEEE, 2017.

[14] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *ECCV*. Springer, 2016.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[16] Zhenliang He, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Robust fec-cnn: A high accuracy facial landmark detection system. In *CVPRW*, 2017.

[17] Dinesh Jayaraman and Kristen Grauman. Learning image representations tied to egomotion from unlabeled video. *International Journal of Computer Vision*, 125(1-3):136–161, 2017.

[18] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences. In *ICCV*. IEEE, 2017.

[19] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *ECCV*, 2018.

[20] Haoxiang Li, Zhe Lin, Xiaohui Shen, Jonathan Brandt, and Gang Hua. A convolutional neural network cascade for face detection. In *CVPR*, 2015.

[21] Wei Li, Farnaz Abtahi, and Zhigang Zhu. Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing. In *CVPR*. IEEE, 2017.

[22] Wei Li, Farnaz Abtahi, Zhigang Zhu, and Lijun Yin. Eacnet: A region-based deep enhancing and cropping approach for facial action unit detection. In *FG*, 2017.

[23] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *CVPRW*. IEEE, 2010.

[24] Brais Martinez, Michel F Valstar, Bihan Jiang, and Maja Pantic. Automatic analysis of facial actions: A survey. *IEEE Transactions on Affective Computing*, 2017.

[25] S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013.

[26] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *ECCV*. Springer, 2016.

[27] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 2017.

[28] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: a large-scale speaker identification dataset. In *INTERSPEECH*, 2017.

[29] Joe Yue-Hei Ng, Jonghyun Choi, Jan Neumann, and Larry S Davis. Actionflownet: Learning motion representation for action recognition. In *WACV)*. IEEE, 2018.

[30] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[31] Guozhu Peng and Shangfei Wang. Weakly supervised facial action unit recognition through adversarial training. In *CVPR*, pages 2188–2196, 2018.

[32] Zhiwen Shao, Zhilei Liu, Jianfei Cai, and Lizhuang Ma. Deep adaptive attention for joint facial action unit detection and face alignment. *ECCV*, 2018.

[33] Zhixin Shu, Mihir Sahasrabudhe, Alp Guler, Dimitris Samaras, Nikos Paragios, and Iasonas Kokkinos. Deforming autoencoders: Unsupervised disentangling of shape and appearance. In *ECCV*, 2018.

[34] Sima Taheri, Qiang Qiu, and Rama Chellappa. Structure-preserving sparse decomposition for facial expression analysis. *IEEE Transactions on Image Processing*, 23(8):3590–3603, 2014.

[35] Michel Valstar and Maja Pantic. Fully automatic facial action unit detection and temporal analysis. In *CVPRW*. IEEE, 2006.

[36] Jacob Walker, Abhinav Gupta, and Martial Hebert. Dense optical flow prediction from a static image. In *CVPR*, 2015.

[37] Ziheng Wang, Yongqiang Li, Shangfei Wang, and Qiang Ji. Capturing global semantic relationships for facial action unit recognition. In *ICCV*, 2013.

[38] Olivia Wiles, A Koepke, and Andrew Zisserman. Self-supervised learning of a facial attribute embedding from video. *BMVC*, 2018.

[39] Olivia Wiles, A Sophia Koepke, and Andrew Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In *ECCV*, 2018.

[40] Jiabei Zeng, Wen-Sheng Chu, Fernando De la Torre, Jeffrey F Cohn, and Zhang Xiong. Confidence preserving machine for facial action unit detection. In *CVPR*, 2015.

[41] Richard Zhang, Phillip Isola, and Alexei A Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *CVPR*, 2017.

[42] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, and Peng Liu. A high-resolution spontaneous 3d dynamic facial expression database. In *FG*. IEEE, 2013.

[43] Yong Zhang, Weiming Dong, Bao-Gang Hu, and Qiang Ji. Weakly-supervised deep convolutional neural network learning for facial action unit intensity estimation. In *CVPR*, 2018.

[44] Guoying Zhao and Matti Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):915–928, 2007.

[45] Kaili Zhao, Wen-Sheng Chu, Fernando De la Torre, Jeffrey F Cohn, and Honggang Zhang. Joint patch and multi-label learning for facial action unit detection. In *CVPR*, 2015.

[46] Kaili Zhao, Wen-Sheng Chu, and Aleix M Martinez. Learning facial action units from web images with scalable weakly supervised clustering. In *CVPR*, pages 2090–2099, 2018.

[47] Kaili Zhao, Wen-Sheng Chu, and Honggang Zhang. Deep region and multi-label learning for facial action unit detection. In *CVPR*, pages 3391–3399, 2016.

[48] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *CVPR*. IEEE, 2017.