Learning Facial Representations from the Cycle-consistency of Face

Jia-Ren Chang Yong-Sheng Chen Wei-Chen Chiu
National Yang Ming Chiao Tung University, Hsinchu, Taiwan
{followwar.cs00g, yschen, walon}@nctu.edu.tw

Abstract

Faces manifest large variations in many aspects, such as identity, expression, pose, and face styling. Therefore, it is a great challenge to disentangle and extract these characteristics from facial images, especially in an unsupervised manner. In this work, we introduce cycle-consistency in facial characteristics as free supervisory signal to learn facial representations from unlabeled facial images. The learning is realized by superimposing the facial motion cycle-consistency and identity cycle-consistency constraints. The main idea of the facial motion cycle-consistency is that, given a face with expression, we can perform de-expression to a neutral face via the removal of facial motion and further perform re-expression to reconstruct back to the original face. The main idea of the identity cycle-consistency is to exploit both de-identity into mean face by depriving the given neutral face of its identity via feature re-normalization and re-identity into neutral face by adding the personal attributes to the mean face. At training time, our model learns to disentangle two distinct facial representations to be useful for performing cycle-consistent face reconstruction. At test time, we use the linear protocol scheme for evaluating facial representations on various tasks, including facial expression recognition and head pose regression. We also can directly apply the learnt facial representations to person recognition, frontalization and image-to-image translation. Our experiments show that the results of our approach is competitive with those of existing methods, demonstrating the rich and unique information embedded in the disentangled representations. Code is available at https://github.com/JiaRenChang/FaceCycle.

1. Introduction

Face perception is vital for human beings and is also essential in the field of computer vision. Neuroimaging studies of both human and monkey [13, 15, 43] reveal the neuroanatomical dissociation between expression and identity representations in face perception. Their findings suggest that these facial characteristics are processed in different brain areas. With the renaissance of deep learning in recent years, computer vision research field follows this thread of thinking and progresses in the direction of disentangling the face characteristics into separated low-dimensional latent representations, such as identity [40], expression [45, 48], shape/appearance [35, 44], intrinsic images [36], and fine-grained attributes (age, gender, wearing glasses, etc.) [34].

Figure 1. We propose an unsupervised framework based on cycle-consistency for learning face disentanglement. We define that all the variations between a face and its corresponding neutral face of the same identity as *expression*. Similarly, all the variations between a neutral face and the global mean face are defined as *identity*. The input face is sequentially deprived of expression (R_{exp}) and identity (R_{id}) representations by networks to become the neutral face and mean face, respectively, which can be transformed back to the original face in reverse order.

Several supervised methods have been proposed to disentangle face characteristics for image manipulation by conditioning generative models on a pre-specified face representations, including landmarks [47], action units [32], or facial attributes [27]. Particularly, these methods are able to manipulate faces while preserving the identity. Other studies incorporate head pose information to disentangle pose-invariant representations for robust identity [40]/expression [48] recognition. Provided with neutral face, moreover, de-expression residue learning [45] can facilitate the model to learn identity-invariant expression representations for performing facial expression recognition.

The 3D Morphable Model (3DMM) [2, 4] for face shape modeling incorporates a similar thinking of dissociation for expression and identity. The most widely-used form of 3DMM is that a face shape $\mathbf{S}$ is a linear combination of mean shape $\bar{\mathbf{S}}$ and identity and expression vectors ($\mathbf{z}_{\text{id}}, \mathbf{z}_{\text{exp}}$): $\mathbf{S} = \bar{\mathbf{S}} + \mathbf{A}_{\text{id}}\mathbf{z}_{\text{id}} + \mathbf{A}_{\text{exp}}\mathbf{z}_{\text{exp}}$, where $\mathbf{A}_{\text{id}}$ and

$\mathbf{A}_{\text{exp}}$ are the identity and expression PCA bases, respectively. Jiang *et al*. [20] introduce a variational autoencoder approach for learning latent representations of expression mesh and identity mesh in the framework of 3DMM. However, they provided strong supervision for the disentanglement of identity and expression representations, including ground truths of shape meshes for expression, identity, and the mean face [20]. It is difficult to generalize such methods to 2D facial images without being given any ground truth.

In addition to the aforementioned works which are mostly based on supervised learning, recently a few studies begin to exploit the unsupervised learning framework to disentangle facial characteristics [26, 41, 42, 44]. These methods focus on extracting a part of facial characteristics. For example, FAb-Net [41] learns representations that encode information about pose and expression, [26, 41] introduce frameworks to learn representations for action unit detection, and Zhang *et al*. [49] propose an autoencoder to locate facial landmarks. Some unsupervised methods [35, 44] attempt to separate two independent representations of face images, including shape and appearance. However, these unsupervised methods can only disentangle a part of information of facial images, but not yet investigate a more general generative procedure of a human face, that is, simultaneous disentanglement of expression and identity representations for a wider usage.

In this paper, we propose a novel framework that is able to simultaneously disentangle expression and identity representations from 2D facial images in an unsupervised manner. Particularly, the definition of the **expression** factor in our proposed method contains all the variations between an arbitrary face image and its corresponding neutral face of the same identity, including the facial expression and head pose. While for the **identity** factor, we define it to contain all the variations between a neutral face and the global mean face, including the facial identity and other subject-specific features such as hair style, age, gender, beard, glasses, *etc*. Based on these definitions, we propose two novel cycle-consistency constraints to drive our model learning, as illustrated in Figure 1.

The first cycle-consistency constraint stems from the idea of action unit [9] in which the head poses and facial expressions are the results of the combined and coordinated action of facial muscles. Therefore, the head poses and expression can be treated as the optical flow [28] between a neutral face and any face of the same identity. To this end, a decoder is trained to learn the optical flow field of the input face *without* the ground truth neutral face. This is achieved by applying the proposed idea called ***facial motion cycle-consistency***, which is able to perform both the **de-expression** and **re-expression** operations.

The second cycle-consistency constraint originates from Eigenfaces [38], in which a facial image is represented by adding a linear combination of eigenfaces to the mean face, suggesting that the face identity is embedded in the linear combination of eigenfaces. Instead of representing the identity as the residue of neutral facial image relative to the mean face [38], we model the adding and depriving of identity as a renormalization procedure, analogues to the feed-forward style transfer tasks [18]. To this end, decoders are trained to learn the renormalized features *without* the ground truth mean face. This is achieved by applying the proposed idea called ***identity cycle-consistency***, which is able to perform identity deprivation as **de-identity** and the identity styling as **re-identity**.

The main contributions of our work are summarized as follows:

- We propose a novel framework for unsupervised learning of facial representations from a single facial image, based on the novel ideas of facial motion cycle-consistency and identity cycle-consistency.
- The disentangled expression and identity features obtained by our proposed method can be easily utilized for various downstream tasks, such as facial expression recognition, head pose regression, person recognition, frontalization, and image-to-image translation.
- We demonstrate that the performance of the learned representations in different downstream tasks is competitive with the state-of-the-art methods.

2. Unsupervised Learning of Facial Representations

As motivated previously, in this paper we aim at disentangling the identity and expression representations from a single facial image. Our proposed method is mainly based on an important assumption that: a facial image F, from high-level perspective, can be decomposed as follows:

$$F = \bar{F} + \text{id} + \text{exp} = \hat{F} + \text{exp}\,, \tag{1}$$

where $\bar{F}$ is the global mean face shared among all the faces, **id** and **exp** are the identity and expression factors respectively, and $\hat{F}$ is the neutral face of a particular identity specified by **id**. Therefore, our proposed model is trained to learn the expression and identity representations, denoted as R_{exp} and R_{id} respectively, for indicating the facial characteristics of facial images. We introduce four processes based on cycle-consistency for learning these representations, as shown in Figure 1:

- **de-expression.** We define the *de-expression* as removing R_{exp} from the input facial image F, in which we can obtain the neutral face $\hat{F}$ accordingly.

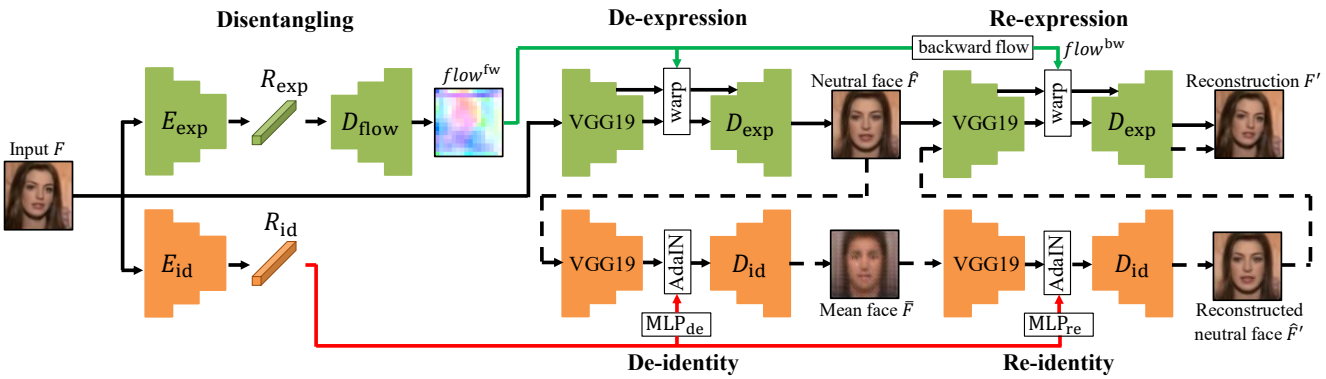


Figure 2. Overall architecture of the proposed model. The E_{exp} and E_{id} are trained to extract expression and identity representations respectively by using our unsupervised disentangling method. By exploring the disentangled representations, networks D_{flow}, D_{exp}, MLPs and D_{id} are trained to generate the representation-removed images, the neutral face $\hat{F}$ and the mean face $\bar{F}$, and to reconstruct the representation-added images, the input face $F^{'}$ and the neutral face $\hat{F}^{'}$. Please note that the proposed method needs two images to train the model as described in Sec. 2.2 and 2.4, and we only show a single image forwarding here for simplicity.

- **re-expression.** The *re-expression* is defined as assigning R_{exp} to the neutral face $\hat{F}$ for reconstructing face with expressions $F^{'}$.
- **de-identity.** We define *de-identity* as an operation for removing R_{id} from the input neutral face $\hat{F}$ in order to obtain the mean face $\bar{F}$.
- **re-identity.** The *re-identity* is defined as the process of recovering the neutral face $\hat{F}^{'}$ back from the mean face $\bar{F}$ according to R_{id}.

As illustrated in Figure 2, the overall architecture of our proposed model consists of two encoders (E_{exp} and E_{id}) for extracting expression and identity representations respectively, and two decoders (D_{exp}, D_{id}) for learning nonlinear mapping functions of the aforementioned four processes. In the following, we detail the proposed unsupervised learning method to disentangle expression and identity representations.

2.1. Expression Representation

We start with the introduction of facial motion cycle-consistency in the following. We denote that the expression representations R_{exp} are learned by an encoder E_{exp} from the input face image F:

$$R_{\text{exp}} = E_{\text{exp}}(F)\,. \tag{2}$$

As the idea described in the previous section, we model a facial expression as the optical flow field between the neutral face and the face with expression. Therefore, the forward ($F \rightarrow \hat{F}$) optical flow field $flow^{\text{fw}} \in \mathbb{R}^{2\times H\times W}$, where H and W is the height and width of the input image, is learned by the decoder D_{flow} from expression representations. Moreover, according to the well-known *forward-backward flow consistency* [1, 19], we can compute the backward ($\hat{F} \rightarrow F$) optical flow field $flow^{\text{bw}}$ according to $flow^{\text{fw}}$, which basically is inverse $flow^{\text{fw}}$ by a warp function $\mathcal{W}$:

$$\begin{aligned} flow^{\text{fw}} &= D_{\text{flow}}(R_{\text{exp}})\,, \\ flow^{\text{bw}} &= -\,\mathcal{W}(flow^{\text{fw}}, flow^{\text{fw}})\,. \end{aligned} \tag{3}$$

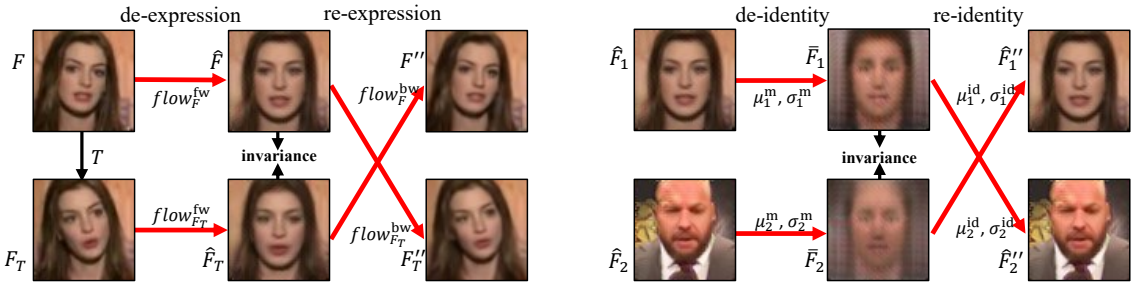
We use bilinear interpolation to implement the warping operation $\mathcal{W}$ as in [39]. By using the forward optical flow field $flow^{\text{fw}}$ we can warp F pixel-wisely to obtain an intermediate facial image, denoted as $\tilde{F}$. Followed by using the corresponding backward optical flow field $flow^{\text{bw}}$, we are able to warp back from $\tilde{F}$ to reconstruct F. This procedure straightforwardly lead to a reconstruction loss $\mathcal{L}_{\text{flow}}$ which is defined as:

$$\mathcal{L}_{\text{flow}} = |F - \mathcal{W}(\mathcal{W}(F, flow^{\text{fw}}), flow^{\text{bw}})|\,. \tag{4}$$

Furthermore, we exploit a general image feature extraction to represent a face image, that is, the coarse-to-fine feature maps $feat_F$ obtained from layers `conv2_1`, and `conv3_1` of VGG19 network pre-trained on ImageNet [37]. Given a forward flow field $flow^{\text{fw}}$, we simply use the bilinear interpolation function $d_s(\cdot)$ to obtain $d_s(flow^{\text{fw}})$ of the size equal to $feat_F$. The **de-expression** is then achieved by first warping $feat_F$ with $d_s(flow^{\text{fw}})$ and then adopting a decoder D_{exp} to generate neutral face image $\hat{F}$:

$$\hat{F} = D_{\text{exp}}(\mathcal{W}(feat_F, d_s(flow^{\text{fw}})))\,. \tag{5}$$

Moreover, we argue that the image features $feat_{\hat{F}}$ of a neutral face obtained by VGG19 could be warped back via the downsampled backward flow $d_s(flow^{\text{bw}})$ and then be fed into the decoder D_{exp} for reconstructing a face with expression, denoted as $F^{'}$, which ideally should be identical to the



(a) Facial motion cycle-consistency (b) Identity cycle-consistency

Figure 3. Illustration of (a) facial motion cycle-consistency for learning expression representations, and (b) identity cycle-consistency for learning identity representations.

original face F. This process is exactly the **re-expression**:

$$F^{'} = D_{\text{exp}}(\mathcal{W}(feat_{\hat{F}}, d_s(flow^{\text{bw}}))) \,. \tag{6}$$

2.2. Facial Motion Cycle-Consistency: Invariance for Learning Expression Representation

The change on a face image F caused by a facial motion can be expressed in terms of a spatial image transformation T, where we denote the corresponding face image with different motion but the same identity as F_T. As both F and F_T are with the same identity, their corresponding neutral faces should be also identical. That is, their decoded neutral faces after performing de-expression are **invariant** to each other, which leads to the constraint:

$$\hat{F} = \hat{F_T} \,. \tag{7}$$

Following the concept of this invariance, we should be able to apply the re-expression operation on $feat_{\hat{F_T}}$ (the features for the decoded neutral face of F_T) via the downsampled backward flow $d_s(flow^{\text{bw}}_F)$ of F (related to the expression of F) to reconstruct a face image denoted as $F^{''} = D_{\text{exp}}(\mathcal{W}(feat_{\hat{F_T}}, d_s(flow^{\text{bw}}_F)))$, which ideally is quite similar to the original F due to the hypothesis that $feat_{\hat{F}} = feat_{\hat{F_T}}$ as $\hat{F} = \hat{F_T}$. The similar story holds for performing re-expression on $feat_{\hat{F}}$ by $d_s(flow^{\text{bw}}_{F_T})$ to reconstruct $F^{''}_T = D_{\text{exp}}(\mathcal{W}(feat_{\hat{F}}, d_s(flow^{\text{bw}}_{F_T})))$, which is almost identical to F_T. The illustration of this invariance, also named as **facial motion cycle-consistency**, is shown in Figure 3(a).

The reconstruction derived from the invariance (that is $F^{''}$ versus F and $F^{''}_T$ versus F_T) builds up the objectives $\mathcal{L}_{\text{exp}}$ for learning the expression representations R_{exp}, where we utilize both the L1 loss and the perceptual loss [11, 21] to evaluate the error of reconstruction:

$$\begin{aligned}\mathcal{L}_{\text{exp}}(F, F_T) =& |F - F^{''}| + |F_T - F^{''}_T| \\ &+\lambda(\Phi(F, F^{''}) + \Phi(F_T, F^{''}_T)) \,,\end{aligned} \tag{8}$$

where λ is set to 0.05 to balance the L1 and perceptual losses. The perceptual loss is defined as $\Phi(F, F^{''}) = \sum_l \|\phi^l(F) - \phi^l(F^{''})\|_2 + \sum_l \|\mathcal{G}(\phi^l(F)) - \mathcal{G}(\phi^l(F^{''}))\|_2$. The function $\phi^l(\cdot)$ extracts VGG19 features from layer l, in which the `conv2_1`, `conv3_1`, and `conv4_1` layers are used here. The function $\mathcal{G}(\cdot)$ calculates the Gram matrix of the feature map.

2.3. Identity Representation

In terms of identity representation R_{id}, we utilize the encoder E_{id} to extract R_{id} from the input face image F:

$$R_{\text{id}} = E_{\text{id}}(F) \,. \tag{9}$$

Based on the idea described previously, we argue that the identity representation could be deprived from the neutral face to obtain the mean face. To implement the **de-identity** operation, we design a decoder D_{id} to generate the mean face $\bar{F}$ from the modulated VGG features $feat_{\hat{F}}$ of a neutral face $\hat{F}$, which is similar to the idea of feature modulation idea proposed in the AdaIN paper [18]:

$$\bar{F} = D_{\text{id}}(\frac{feat_{\hat{F}} - \mu(feat_{\hat{F}})}{\sigma(feat_{\hat{F}})}\sigma^{\text{m}} + \mu^{\text{m}}) \,, \tag{10}$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ are used to compute the mean and standard deviation respectively, and μ^{m} and σ^{m} are learned from R_{id} by the multi-layer perceptron MLP_{de}:

$$\mu^{\text{m}}, \sigma^{\text{m}} = \text{MLP}_{\text{de}}(R_{\text{id}}) \,. \tag{11}$$

Furthermore, the **re-identity** can be achieved in a similar manner but reversely with the decoder D_{id}:

$$\hat{F}^{\prime} = D_{\text{id}}(\frac{feat_{\bar{F}} - \mu(feat_{\bar{F}})}{\sigma(feat_{\bar{F}})}\sigma^{\text{id}} + \mu^{\text{id}}) \,, \tag{12}$$

where the μ^{id} and σ^{id} are also learned from R_{id} but by another multi-layer perceptron MLP_{re}.

2.4. Identity Cycle-Consistency: Invariance for Learning Identity Representation

We hypothesize that the mean face is global for all the faces. In other words, no matter starting from which neural face of any identity, we should always obtain the same

mean face after performing de-identity operation. Given the neutral faces $\hat{F}_1$ and $\hat{F}_2$ of different identities, we can derive the **invariance** related to identity as:

$$\bar{F}_1 = \bar{F}_2 \,. \tag{13}$$

Therefore, we should be able to reconstruct $\hat{F}_1$ by using its corresponding $\{\mu_1^{\text{id}}, \sigma_1^{\text{id}}\}$ to apply the re-identity operation on the mean face obtained from $\hat{F}_2$. The result of this reconstruction is denoted as $\hat{F}_1^{''}$:

$$\hat{F}_1^{''} = D_{\text{id}}(\frac{feat_{\bar{F}_2} - \mu(feat_{\bar{F}_2})}{\sigma(feat_{\bar{F}_2})}\sigma_1^{\text{id}} + \mu_1^{\text{id}}) \,. \tag{14}$$

Again, similar story holds to perform re-identity operation (with $\{\mu_2^{\text{id}}, \sigma_2^{\text{id}}\}$) on the mean face obtained from $\hat{F}_1$ to reconstruct $\hat{F}_2$. We denote the reconstruction result as $\hat{F}_2^{''}$:

$$\hat{F}_2^{''} = D_{\text{id}}(\frac{feat_{\bar{F}_1} - \mu(feat_{\bar{F}_1})}{\sigma(feat_{\bar{F}_1})}\sigma_2^{\text{id}} + \mu_2^{\text{id}}) \,. \tag{15}$$

The illustration of this invariance related to identity representations, also named as **identity cycle-consistency**, is shown in Figure 3(b).

As the way that $\mathcal{L}_{\text{exp}}$ is defined, the reconstruction derived from the invariance (that is, $\hat{F}_1^{''}$ versus $\hat{F}_1$ and $\hat{F}_2^{''}$ versus $\hat{F}_2$) leads to the objectives $\mathcal{L}_{\text{id}}$ for learning the identity representations R_{id}:

$$\begin{aligned}\mathcal{L}_{\text{id}}(\hat{F}_1, \hat{F}_2) =& |\hat{F}_1 - \hat{F}_1^{''}| + |\hat{F}_2 - \hat{F}_2^{''}| \\ &+\lambda(\Phi(\hat{F}_1, \hat{F}_1^{''}) + \Phi(\hat{F}_2, \hat{F}_2^{''})) \,.\end{aligned} \tag{16}$$

Moreover, we additionally introduce a margin loss $\mathcal{L}_{\text{m}}$ to constrain the mean face:

$$\mathcal{L}_{\text{m}}(\bar{F}, \hat{F}) = \max(\left\|\bar{F} - \hat{F}\right\| - \alpha, 0) \,, \tag{17}$$

where we set $\alpha = 0.1$ in all experiments. The main motivation behind this margin loss is that we would like to constrain the difference between the mean face and the neutral face to be within a margin. Otherwise the obtained mean face could potentially become an arbitrary image far from a face image.

3. Experiments

We report experimental results for a model trained on the combination of VoxCeleb1 [29] and VoxCeleb2 [5] from scratch. The trained representations are evaluated on several tasks, including facial expression recognition, head pose regression, person recognition, frontalization, and image-to-image translation. Through various experiments, we show that the acquired representation generalizes to a range of facial image processing tasks.

3.1. Training Procedures

The facial motion cycle-consistency described in Section 2.2 involves an image pair of faces with different expressions/poses but of the same identity. Fortunately, this type of data can be easily available from the video recording of human faces, for instance, the video of interview or talk-show, which exists widely on the Internet nowadays. Given any two frames in this type of video clip of a person, we can easily obtain a pair of facial images showing different expressions. Therefore, we can take the advantage of this type of video sequences (as the dataset described in the following) and collect training data for learning both the expression and identity representations in an unsupervised manner.

Dataset. The proposed model is trained on the combination of VoxCeleb1 [29] and VoxCeleb2 [5] datasets, in which both datasets are built upon videos of interviews. VoxCeleb1 has in total 153,516 video clips of 1,251 speakers, while VoxCeleb2 has 145,569 video clips of 5,994 speakers. Video frames were extracted at 6 fps, cropped to have faces shown in the center of frames, and then resized to the resolution of 64×64. We adopted VoxCeleb2 test dataset for visualizing the intermediate results of our disentanglement process.

Stage-wise Training Procedure. We introduce a stage-wise training procedure for our model learning. There are two main stages for sequentially training different parts of the proposed model, in order to disentangle the expression and identity representations.

– Stage 1: training of E_{exp}, D_{flow} and D_{exp}

For training the subnetworks related to the de-expression and re-expression parts, as the green-shaded components shown in Figure 2, the objectives of $\mathcal{L}_{\text{flow}}$ and $\mathcal{L}_{\text{exp}}$ are utilized to update $\{E_{\text{exp}}, D_{\text{flow}}, D_{\text{exp}}\}$. The transformation T needed for the use of $\mathcal{L}_{\text{exp}}$ can be simply obtained by having the horizontal flipping (that is, F_T is the horizontally flipped version of F) or taking any arbitrary pair of faces from different frames (that is, two faces of the same person shown at different times in a video). We provide an ablation study in the supplementary materials.

– Stage 2: training of E_{id}, D_{id}, MLP_{re}, and MLP_{de}

For training the subnetworks related to the de-identity and re-identity parts, as the orange-shaded components shown in Figure 2, both the objectives of $\mathcal{L}_{\text{id}}$ and $\mathcal{L}_{\text{m}}$ are applied to update all of these subnetworks.

Implementation Details. Our proposed model is implemented based on PyTorch framework and trained with the Adam optimizer ($\beta_1 = 0.5$, and $\beta_2 = 0.999$). The batch

size is set to 32 for all the training stages. The initial learning rate is 0.00005 in the Stage 1 and 0.0001 in the Stage 2. The Stage 1 and Stage 2 are trained for 40 and 20 epochs respectively. The learning rate is decreased by a factor of 10 at half of total epochs. Moreover, both representation encoders (*i.e.* E_{exp} and E_{id}) adopt the same network architecture which is a 16-layer CNN. We leverage a VGG-19 [37] for the general feature extraction (denoted as VGG19 component in the Figure 2), where the VGG-19 encoded facial features can be further passed through our decoders (*i.e.* D_{exp} or D_{id}) to generate new facial images. The model architectures are detailed in the supplementary materials.

Baselines. We adopt the following baselines for making evaluations and comparisons in terms of the quality and representativeness of the extracted facial features:

– HoG descriptor [6]: We follow the same setting as in [23], where the facial images are first rescaled to the size of 100×100, then the HoG feature of 3,240 dimensions is extracted for each image.

– LBP descriptor [30]: Similar to the HoG descriptor, we follow the same setting as in [23] to extract 1,450 dimensional LBP feature vector from each of the facial images which are resized to 100×100.

– MoCo [16]: We adopt the state-of-the-art self-supervised representation learning method, MoCo, as a strong baseline for us to compare with. We follow the MoCo algorithm to train the feature extractor (in which its network architecture is the same as our encoders) based on the same training dataset as ours (*i.e.* VoxCeleb1 and VoxCeleb2). The training runs for 40 epochs with SGD optimizer, batch size of 128, momentum 0.999, and 65,536 negative keys.

– Self-supervisely learnt facial representations: Three state-of-the-art self-supervised frameworks of facial representation learning [24, 26, 41] are utilized to compare with our work. We directly adopt the models officially released by their authors (which are all pretrained on the Voxceleb dataset) for experimenting the downstream tasks of expression classification and head pose regression. Please note that we apply a linear protocol on their learnt features to have a fair comparison.

3.2. Intermediate Results of Our Model

Figure 4 illustrates several examples of the intermediate results obtained from our model, including the input faces, forward flow fields, neutral faces, backward flow fields, mean faces, and the faces reconstructed from their neutral ones. We demonstrate that the proposed method can handle face images with large variation in poses and can preserve facial attributes such as wearing glasses or beard.

Visualization of the facial motion flows presents both the head motion and the movement of facial muscles. The neutral faces are deprived of facial motions in comparison to their original facial images. Moreover, the mean faces obtained from different input images are almost identical to each other, which is in line with our assumption of identity invariance.

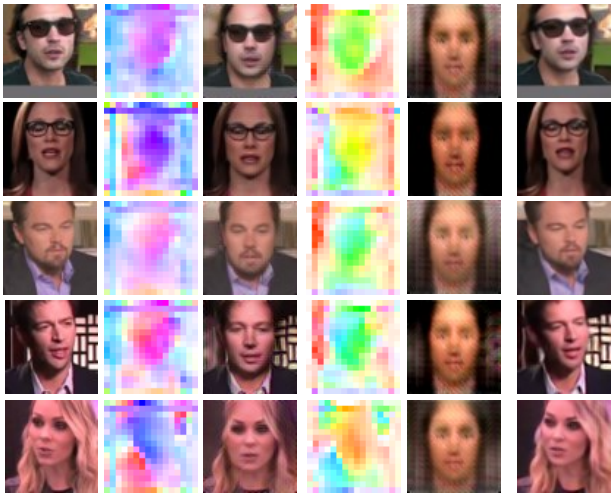


Figure 4. Visualization of intermediate results of our model. Input images are from the test set of the VoxCeleb2 dataset [5]. From left to right, the columns sequentially show the input faces, forward flow fields, neutral faces, backward flow fields, mean faces, and the faces reconstructed from the corresponding neutral ones.

3.3. Evaluation for Expression Representation

Given the trained model, we investigate the learnt expression representation by evaluating the performance of its applications on expression recognition and head pose regression. The goal is to verify whether the expression representation successfully encodes the information related to the facial motions and poses as our definition (*i.e.* the expression factor contains all the variations between a face image and its corresponding neutral face of the same identity, including facial motions and head poses). We conduct *linear-protocol* evaluation scheme for demonstrating the effectiveness of our method.

3.3.1 Expression Recognition

Two datasets are used in the experiments of expression recognition, i.e. FER-2013 [12] and RAF-DB [23]. FER-2013 dataset [12] consists of 28,709 training and 3,589 testing images, while RAF-DB dataset consists of around 30K diverse facial images downloaded from the Internet. Please note that for the RAF-DB dataset, we follow the experimental setup as [23] to particularly use the basic emotion subset of RAF-DB, which includes 12,271 training and 3,068 testing images. For the evaluation scheme of linear-protocol, in order to directly verify the capacity of the expression fea-

Method	FER-2013 Accuracy (%)	RAF-DB Accuracy (%)
Fully supervised		
FSN [50]	67.60	81.10
ALT [10]	69.85	84.50
Linear classification protocol		
LBP	37.89	52.17
HoG	45.47	63.53
FAb-Net [41]	46.98	66.72
TCAE [24]	45.05	65.32
BMVC'20 [26]	47.61	58.86
MoCo	47.24	68.32
Ours	**48.76**	**71.01**

Table 1. Evaluation on the task of expression classification based on the FER-2013 dataset [12] and RAF-DB dataset [23].

Method	Yaw	Pitch	Roll	MAE
Fully supervised				
FAN [3]	6.36	12.3	8.71	9.12
FSA-Net [46]	5.27	6.71	5.28	5.75
Linear regression protocol				
Dlib (68 points) [22]	23.10	13.60	**10.50**	15.80
LBP	23.58	14.86	16.36	18.27
HoG	13.94	13.17	14.92	14.00
FAb-Net [41]	13.92	13.25	14.51	13.89
TCAE [24]	21.75	14.57	14.83	17.39
BMVC'20 [26]	22.06	13.50	15.14	16.90
MoCo	28.49	16.29	15.55	20.11
Ours	**11.70**	**12.76**	12.94	**12.47**

Table 2. Evaluation on the task of head pose regression, where MAE stands for the mean absolute error.

tures extracted by different models, we construct the *linear* classifier upon the *frozen* expression representations to perform the expression recognition, as in [16]. We follow the same procedure as [16] to train the linear layer (as the classifier) for 300 epochs, where the learning rate starts from 30 and decreases by a factor of 10 for every 80 epochs. The classifiers are trained by the SGD optimizer with cross-entropy objective and 256 batch size.

The quantitative results shown in Table 1 demonstrate that the expression representation extracted from our proposed method is able to provide superior performance with respect to all the baselines. These results suggest that our proposed method can be used as a pretext task for expression recognition, where the rich information of facial expression is well learnt in a self-supervised manner.

3.3.2 Regression of Head Pose

Our definition indicates that the information of head poses would be also encoded into the expression representations. Obviously the calculated flow fields using the proposed method contain not only the local facial motion but also the global head motion, suggesting that our expression representation can also be used in the task of head pose regression. We adopt the 300W-LP [33] dataset and the AFLW2000 [52] dataset as the training and testing sets respectively, for experimenting the head pose regression. For the evaluation scheme of *linear-protocol*, we construct a *linear* regressor on top of the *frozen* expression representations E_{exp}. The training runs for 300 epochs for the inear-protocol with SGD optimizer and batch size set to 16.

As shown in Table 2, for the linear-protocol evaluation scheme, the regressor based on our expression representations achieves 12.47 in terms of mean absolute error (MAE), which outperforms all the baselines. These results demonstrate the effectiveness of our proposed method for well capturing the head pose information into expression representations.

3.4. Evaluating Identity Representations

We also investigate the applications of identity representations learned by using the proposed method on the VoxCeleb dataset. Good performance of person recognition demonstrates that our identity representations do contain rich information related to identities.

3.4.1 Person Recognition

In this work we adopt LFW [17] and CPLFW [51] dataset for the evaluation of person recognition, particularly on person verification. The LFW dataset comprises of 13,233 face images from 5,749 identities and has 6,000 face pairs for evaluating person verification. The CPLFW dataset is similar to LFW but includes larger head pose variation. We *directly* extract the identity representations for all of the images in the face pairs from two datasets by using the encoder E_{id} and then compute the cosine similarity between identity representations of each pair of face images. Please note that, the features from baselines (*i.e.* LBP, HoG, and MoCo) are also *directly* applied to perform verification for a fair comparison. As shown in Table 3, our identity representations can achieve 73.72% in accuracy on LFW, which outperforms the unsupervised state-of-the-art method [7].

3.5. Frontalization

Frontalization is the process of synthesizing the frontal facing view of a single facial image. In this work, there are two ways to obtain the neutral face in the frontal view: de-expression and re-identity. The de-expression operation removes the head motion and facial expressions from facial images and thus generates the neutral faces with frontal view. On the other hand, the re-identity operation recovers the neutral face by adding the identity to the mean

	LFW	CPLFW
Method	Accuracy(%)	Accuracy(%)
Fully supervised		
VGG-Face [31]	98.95	84.00
SphereFace [25]	99.42	81.40
ArcFace [8]	99.53	92.08
Unsupervised or hand-crafted features		
VGG [7]	71.48	-
LBP	56.90	51.50
HoG	62.73	51.73
MoCo	65.88	55.12
Ours	**73.72**	**58.52**

Table 3. Evaluation on the task of person recognition based on the LFW [17] and CPLFW [51] dataset. We compare the performance of state-of-the-art methods in both supervised and unsupervised categories.

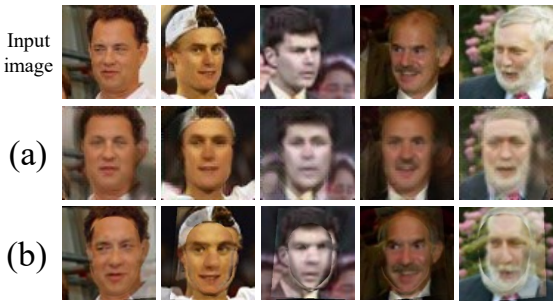


Figure 5. The frontalization results from (a) the proposed method and (b) the method in [14]. These results clearly demonstrates the capacity of frontalization for facial images with various poses by using our method.

face which is already in frontal view. As shown in Figure 5(a), the proposed method is able to synthesize the neutral faces from the facial images with various poses by the de-expression operation. The input images are from the LFW dataset [17] which are never seen during the training of our model. We also show a state-of-the-art approach in Figure 5(b) which additionally uses facial landmarks [14] for qualitative comparison.

We notice that the synthesized images from the proposed method are a little bit blurry, we hypothesize that it might be caused by the plenty blurry training images in the Voxceleb dataset. We believe that further improvements can be obtained by using other high-quality datasets.

3.6. Image-to-image Translation

The proposed model can naturally be used to perform image-to-image translation by transferring the facial motion of the source image into the target one. To this end, we simply calculate and then apply the backward flow field of the source image to warp the neutral face of the target image via the re-expression operation. As shown in Figure 6,

Figure 6. Example results of the proposed method on image-to-image translation in comparison to the X2Face [42] baseline. X2Face shows artifacts when performing large pose transfer. Notice that the proposed method does not include adversarial training to disentangle facial motion and to improve image quality.

our method can transfer the head pose and expression from the source to the target without noticeable artifacts. On the other hand, the results of X2Face method [42] reveal visible artifacts when the pose difference between source and target is large.

4. Conclusions

In this work, we propose novel cycle-consistency constraints for disentangling of identity and expression representations from a single facial image, that is, facial motion cycle-consistency and the identity cycle-consistency. The proposed model can be trained in an unsupervised manner by superimposing the proposed cycle-consistency constraints. We perform extensive qualitative and quantitative evaluations on multiple datasets to demonstrate the efficacy of our proposed method on learning disentangled facial representations. These representations contain rich and distinct information of identity and expression, and can be used to facilitate a variety of applications, such as facial expression recognition, head pose estimation, person recognition, frontalization, and the image-to-image translation.

Acknowledgement. This project is supported by MOST 108-2221-E-009-066-MY3, MOST 110-2636-E-009-001, and MOST 110-2634-F-009-018. We are grateful to the National Center for High-performance Computing for providing computing services and facilities.

References

[1] Michael J Black and Paul Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Computer vision and image understanding*, 63(1):75–104, 1996. 3

[2] Volker Blanz, Thomas Vetter, et al. A morphable model for the synthesis of 3d faces. In *ACM Transactions on Graphics (TOG)*, 1999. 1

[3] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 7

[4] Baptiste Chu, Sami Romdhani, and Liming Chen. 3d-aided face recognition robust to expression and pose variations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 1

[5] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*, 2018. 5, 6

[6] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005. 6

[7] Samyak Datta, Gaurav Sharma, and CV Jawahar. Unsupervised learning of face representations. In *IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, 2018. 7, 8

[8] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 8

[9] Paul Ekman and Wallace V Friesen. *Facial action coding system: Investigator's guide*. Consulting Psychologists Press, 1978. 2

[10] Corneliu Florea, Laura Florea, Mihai-Sorin Badea, Constantin Vertan, and Andrei Racoviteanu. Annealed label transfer for face expression recognition. In *British Machine Vision Conference (BMVC)*, 2019. 7

[11] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 4

[12] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *International conference on neural information processing*, 2013. 6, 7

[13] Michael E Hasselmo, Edmund T Rolls, and Gordon C Baylis. The role of expression and identity in the face-selective responses of neurons in the temporal visual cortex of the monkey. *Behavioural Brain Research*, 1989. 1

[14] Tal Hassner, Shai Harel, Eran Paz, and Roee Enbar. Effective face frontalization in unconstrained images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 8

[15] James V Haxby, Elizabeth A Hoffman, and M Ida Gobbini. The distributed human neural system for face perception. *Trends in Cognitive Sciences*, 2000. 1

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 6, 7

[17] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, University of Massachusetts, Amherst, 2007. 7, 8

[18] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 2, 4

[19] Junhwa Hur and Stefan Roth. Mirrorflow: Exploiting symmetries in joint optical flow and occlusion estimation. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 3

[20] Zi-Hang Jiang, Qianyi Wu, Keyu Chen, and Juyong Zhang. Disentangled representation learning for 3d face shape. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[21] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision (ECCV)*, pages 694–711. Springer, 2016. 4

[22] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 7

[23] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 6, 7

[24] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Self-supervised representation learning from videos for facial action unit detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 6, 7

[25] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 8

[26] Liupei Lu, Leili Tavabi, and Mohammad Soleymani. Self-supervised learning for facial action unit recognition through temporal consistency. In *British Machine Vision Conference (BMVC)*, 2020. 2, 6, 7

[27] Yongyi Lu, Yu-Wing Tai, and Chi-Keung Tang. Attribute-guided face generation using conditional cyclegan. In *European Conference on Computer Vision (ECCV)*, 2018. 1

[28] Kenji Mase. Recognition of facial expression from optical flow. *IEICE TRANSACTIONS on Information and Systems*, 1991. 2

[29] A. Nagrani, J. S. Chung, and A. Zisserman. Voxceleb: a large-scale speaker identification dataset. In *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2017. 5

[30] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern*

Analysis and Machine Intelligence (TPAMI), 24(7):971–987, 2002. 6

[31] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. 2015. 8

[32] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *European Conference on Computer Vision (ECCV)*, 2018. 1

[33] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *IEEE International Conference on Computer Vision Workshops*, 2013. 7

[34] Wei Shen and Rujie Liu. Learning residual images for face attribute manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[35] Zhixin Shu, Mihir Sahasrabudhe, Riza Alp Guler, Dimitris Samaras, Nikos Paragios, and Iasonas Kokkinos. Deforming autoencoders: Unsupervised disentangling of shape and appearance. In *European Conference on Computer Vision (ECCV)*, 2018. 1, 2

[36] Zhixin Shu, Ersin Yumer, Sunil Hadap, Kalyan Sunkavalli, Eli Shechtman, and Dimitris Samaras. Neural face editing with intrinsic image disentangling. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[37] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *ArXiv:1409.1556*, 2014. 3, 6

[38] Lawrence Sirovich and Michael Kirby. Low-dimensional procedure for the characterization of human faces. *Journal of the Optical Society of America A*, 1987. 2

[39] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[40] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[41] Olivia Wiles, A Koepke, and Andrew Zisserman. Self-supervised learning of a facial attribute embedding from video. In *British Machine Vision Conference (BMVC)*, 2018. 2, 6, 7

[42] Olivia Wiles, A Sophia Koepke, and Andrew Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In *European Conference on Computer Vision (ECCV)*, 2018. 2, 8

[43] Joel S Winston, RNA Henson, Miriam R Fine-Goulden, and Raymond J Dolan. fmri-adaptation reveals dissociable neural representations of identity and expression in face perception. *Journal of Neurophysiology*, 2004. 1

[44] Xianglei Xing, Tian Han, Ruiqi Gao, Song-Chun Zhu, and Ying Nian Wu. Unsupervised disentangling of appearance and geometry by deformable generator network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2

[45] Huiyuan Yang, Umur Ciftci, and Lijun Yin. Facial expression recognition by de-expression residue learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[46] Tsun-Yi Yang, Yi-Ting Chen, Yen-Yu Lin, and Yung-Yu Chuang. Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 7

[47] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 1

[48] Feifei Zhang, Tianzhu Zhang, Qirong Mao, and Changsheng Xu. Joint pose and expression modeling for facial expression recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[49] Yuting Zhang, Yijie Guo, Yixin Jin, Yijun Luo, Zhiyuan He, and Honglak Lee. Unsupervised discovery of object landmarks as structural representations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[50] Shuwen Zhao, Haibin Cai, Honghai Liu, Jianhua Zhang, and Shengyong Chen. Feature selection mechanism in cnns for facial expression recognition. In *British Machine Vision Conference (BMVC)*, 2018. 7

[51] T. Zheng and W. Deng. Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments. Technical Report 18-01, Beijing University of Posts and Telecommunications, February 2018. 7, 8

[52] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 7