

LEED: Label-Free Expression Editing via Disentanglement

Rongliang Wu^[0000–0002–5586–0628] and Shijian Lu^[0000–0002–6766–2506]

Nanyang Technological University
ronglian001@e.ntu.edu.sg, shijian.lu@ntu.edu.sg

Abstract. Recent studies on facial expression editing have obtained very promising progress. On the other hand, existing methods face the constraint of requiring a large amount of expression labels which are often expensive and time-consuming to collect. This paper presents an innovative label-free expression editing via disentanglement (LEED) framework that is capable of editing the expression of both frontal and profile facial images without requiring any expression label. The idea is to disentangle the identity and expression of a facial image in the expression manifold, where the neutral face captures the identity attribute and the displacement between the neutral image and the expressive image captures the expression attribute. Two novel losses are designed for optimal expression disentanglement and consistent synthesis, including a mutual expression information loss that aims to extract pure expression-related features and a siamese loss that aims to enhance the expression similarity between the synthesized image and the reference image. Extensive experiments over two public facial expression datasets show that LEED achieves superior facial expression editing qualitatively and quantitatively.

Keywords: Facial Expression Editing, Image Synthesis, Disentangled Representation Learning

1 Introduction

Facial expression editing (FEE) allows users to edit the expression of a face image to a desired one. Compared with facial attribute editing which only considers appearance modification of specific facial regions [58, 31, 42], FEE is much more challenging as it often involves large geometrical changes and requires to modify multiple facial components simultaneously. FEE has attracted increasing interest due to the recent popularity of digital and social media and a wide spectrum of applications in face animations, human-computer interactions, etc.

Until very recently, this problem was mainly addressed from a graphical perspective in which a 3D Morphable Model (3DMM) was first fitted to the image and then re-rendered with a different expression [17]. Such methods typically involve tracking and optimization to fit a source video into a restrictive set of facial poses and expression parametric space [54]. A desired facial expression can be generated by combining the graphical primitives [30]. Unfortunately, 3DMMs

can hardly capture all subtle movements of face with the pre-defined parametric model and often produce blurry outputs due to the Gaussian assumption [17].

Inspired by the recent success of Generative Adversarial Nets (GANs) [19], a number of networks [41, 45, 12, 39, 10, 16, 17, 53] have been developed and achieved very impressive FEE effects. Most of these networks require a large amount of training images with different types of expression labels/annotations, e.g. discrete labels [10, 12], action units intensity [39, 50, 53] and facial landmarks [41, 45, 16], whereas labelling a large amount of facial expression images is often expensive and time-consuming which has impeded the advance of relevant research on this task. At the other end, the ongoing research [39, 41, 45, 17, 53] is largely constrained on the expression editing of frontal faces due to the constraint of existing annotations, which limits the applicability of FEE in many tasks.

This paper presents a novel label-free expression editing via disentanglement (LEED) framework that can edit both frontal and profile expressions without requiring any expression label or annotation by humans. Inspired by the manifold analysis of facial expressions [7, 14, 24] that different persons have analogous expression manifolds, we design an innovative disentanglement network that is capable of separating the identity and expression of facial images of different poses. The label-free expression editing is thus accomplished by fusing the identity of an input image with an arbitrary expression and the expression of a reference image. Two novel losses are designed for optimal identity-expression disentanglement and identity-preserving expression editing in training the proposed method. The first loss is a mutual expression information loss that guides the network to extract pure expression-related features from the reference image. The second loss is a siamese loss that enhances the expression similarity between the synthesized image and the reference image. Extensive experiments show that our proposed LEED even outperforms supervised expression editing networks qualitatively and quantitatively.

The contributions of this work are threefold. First, we propose a novel label-free expression editing via disentanglement (LEED) framework that is capable of editing expressions of frontal and profile facial images without requiring any expression label and annotation by humans. Second, we design a mutual expression information loss and a siamese loss that help extract pure expression-related features and enhance the expression similarity between the edited and reference facial images effectively. Third, extensive experiments show that the proposed LEED is capable of generating high-fidelity facial expression images and even outperforms many supervised networks.

2 Related Work

Facial Expression Editing: FEE is a challenging task and existing works can be broadly grouped into two categories. The first category is more conventional which exploits graphic models for expression editing. A typical approach is to first fit a 3D Morphable Model to a face image and then re-render it with a different expression. A pioneering work of Blanz and Vetter [5] presents the first public

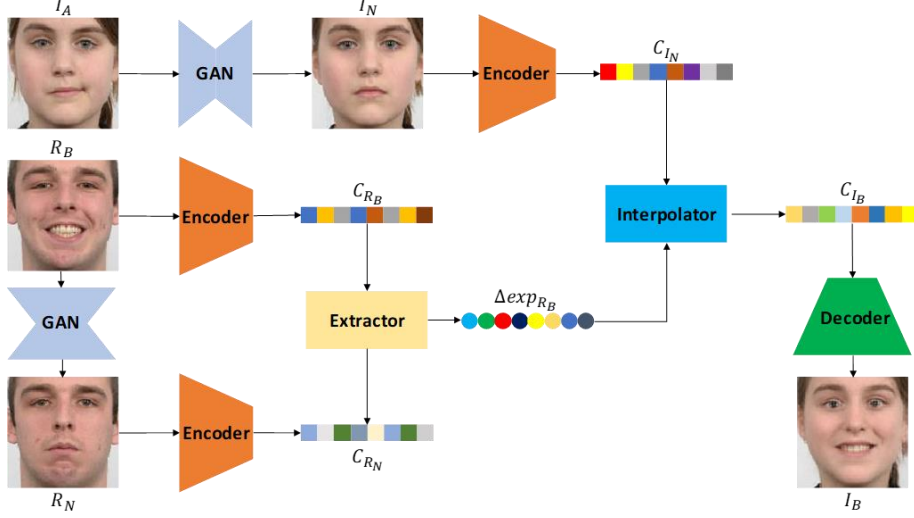


Fig. 1. The framework of LEED: given an input image I_A and a reference image R_B , the corresponding neutral faces I_N and R_N are first derived by a pre-trained GAN. The *Encoder* maps I_N , R_N , and R_B to latent codes C_{I_N} , C_{R_N} , and C_{R_B} in an embedded space which capture the identity attribute of I_A , the identity attribute of R_B , and the identity and expression attributes of R_B , respectively. The *Extractor* then extracts the expression attribute of R_B (Δexp_{R_B}) from C_{R_B} and C_{R_N} , and the *Interpolator* further generates the target latent code C_{I_B} from Δexp_{R_B} and C_{I_N} . Finally, the *Decoder* projects C_{I_B} to the image space to generate the edited image I_B .

3D Morphable Model. Vlastic et al. [49] proposes a video based multilinear model to edit the facial expressions. Cao et al. [6] introduces a video-to-image facial re-targeting application that requires user interaction for accurate editing. Thies et al. [47] presents Face2Face for video-to-video facial expression re-targeting which assumes the target video contains sufficient visible expression variation.

The second category exploits deep generative networks [19, 28]. For example, warp-guided GAN [16] and paGAN [36] are presented to edit the expression of frontal face images with neutral expression. G2-GAN [45] and GCGAN [41] adopt facial landmarks as geometrical priors to control the generated expressions, where ground-truth images are essential for extracting the geometrical information. [17] proposes a model that combines 3DMM and GAN to synthesize expressions on RGB-D images. ExprGAN [12] introduces an expression controller to control the intensity of the generated expressions conditioned on the discrete expression labels. StarGAN [10] generates new expression through identity-preservative image-to-image translation and it can only generate discrete expressions. GANimation [39] adopts Action Units [15] as expression labels and can generate expressions in continuous domain. Cascade EF-GAN [53] also uses Action Units and introduces local focuses and progressive editing strategy

to suppress the editing artifacts. Recently, [40] proposes AF-VAE for face animation where expressions and poses can be edited simultaneously according to the landmarks boundary maps extracted offline with other tools.

The works using graphics models for expression editing require 3D face scans and/or video sequences as well as efforts and dedicated designs for complex parametric fitting. Additionally, they cannot model invisible parts in the source image such as teeth of a closed mouth. The work using generative networks are more flexible but they require a large amount of labelled expressive images to train the models. Besides, the existing deep generative networks also require suitable expression labels/annotations for guiding the model to synthesize desired expression, where the annotations are either created by humans or extracted from reference images by offline tools. The proposed LEED employs generative networks which can hallucinate missing parts of input face images and it just requires a single photo for the input image and reference image which makes it much simpler to implement. At the same time, it can edit the expression of both frontal facial images and profile images without requiring any expression label/annotation by either humans or other tools.

Disentangled Representations: The key of learning disentangled representation is to model the distinct, informative factors of variations in the data [4]. Such representations have been applied successfully to image editing [24, 37, 43], image-to-image translation [57] and recognition tasks [51, 38, 55]. However, previous works achieve disentangled learning by training a multi-task learning model [24, 57, 51], where labels for each disentangled factors are essential. Recently, the unsupervised setting has been explored [8, 23]. InfoGAN [8] achieves disentanglement by maximizing the mutual information between latent variables and data variation and β -VAE [23] learns the independent data generative factors by introducing an adjustable hyper-parameter β to the original VAE objective function. But these methods suffer from the lack of interpretability, and the meaning of each learned factor is uncontrollable. Based on the expression manifold analysis [7], our proposed method seeks another way to disentangle the identity and expression attributes from the facial images.

3 Proposed Method

3.1 Overview

Our idea of label-free expression editing via disentanglement (LEED) is inspired by the manifold analysis of facial expressions [7, 14, 24] that the expression manifold of different individuals is analogous. On the expression manifold, similar expressions are points in the local neighborhood with a ‘neutral’ face as the central reference point. Each individual has its neutral expression that corresponds to the original point in its own expression manifold and represents the identity attribute. The displacement of an expressive face and its neutral face gives the expression attribute.

Our proposed method achieves label-free expression editing by learning to disentangle the identity and expression attributes and fusing the identity of

the input image and the expression of the reference image for synthesizing the desired expression images. As illustrated in Fig. 1, our network has five major components: an extractor for extracting expression attribute; an interpolator for fusing the extracted expression attribute and the identity attribute of the input image; an encoder for mapping the facial images into a compact expression and identity embedded space; a decoder for projecting the interpolated code to image space and a pre-trained GAN for synthesizing the neutral faces.

3.2 Extractor and Interpolator

Learning Expression Attribute Extractor: Given an input image with arbitrary expression A (denoted as I_A) and a reference image with desired expression B (denoted as R_B), our goal is to synthesize a new image I_B that combines the identity attribute of I_A and expression attribute of R_B . Without the expression labels, our proposed method needs to address two key challenges: 1) how to extract identity attribute from the input image and expression attribute from the reference image, and 2) how to combine the extracted identity and expression attributes properly to synthesize the desired expression images. We address the two challenges by learning an expression attribute extractor \mathcal{X} and an interpolator \mathcal{I} , more details to be shared in the following texts.

The label-free expression editing is achieved by disentangling the identity and expression attributes. Given I_A and R_B , LEED first employs a pre-trained GAN to generate their corresponding neutral faces I_N and R_N . An encoder \mathbf{E} is then employed to map all the images to a latent space, producing C_{I_A} , C_{I_N} , C_{R_B} and C_{R_N} , where C_{I_A}/C_{R_B} and C_{I_N}/C_{R_N} are the latent codes of the input/reference image and its neutral face, respectively. More details of the pre-trained GAN and \mathbf{E} are to be discussed in Sec. 3.4.

According to [7], the latent code of the neutral face (C_{I_N}) represents the identity attribute of the input image, and the displacement between C_{R_B} and C_{R_N} represents the expression attribute of the reference image:

$$\Delta exp_{R_B}^* = C_{R_B} - C_{R_N}. \quad (1)$$

On the other hand, $\Delta exp_{R_B}^*$ depends on the embedded space, and the residual between C_{R_B} and C_{R_N} may contain expression-unrelated information such as head-poses variations that could lead to undesired changes in the synthesized images. We therefore propose to learn the expression attribute with an extractor \mathcal{X} rather than directly using $\Delta exp_{R_B}^*$.

Formally, we train an expression extractor \mathcal{X} to extract the expression Δexp_{R_B} from C_{R_B} and C_{R_N} with $\Delta exp_{R_B}^*$ as the pseudo label:

$$\min_{\mathcal{X}} \mathcal{L}_{exp} = \|\Delta exp_{R_B} - \Delta exp_{R_B}^*\|^2, \quad (2)$$

where $\Delta exp_{R_B} = \mathcal{X}(C_{R_B}, C_{R_N})$.

In addition, we design a mutual expression information loss to encourage the extractor to extract pure expression-related information. Specifically, we first

use a pre-trained facial expression classification model Ψ , i.e. the ResNet [21] pre-trained on Real-world Affective Faces Database [32], to extract the features from R_B . As such a model is trained for classification task, the features of the last layers contain rich expression-related information [48]. We take the features from penultimate layer as the representation of the expression attribute of R_B and denote it as F_{R_B} , where $F_{R_B} = \Psi(R_B)$. As Ψ is used for extracting the features, we do not update its parameters in the training process.

In information theory, the mutual information between A and B measures the reduction of uncertainty in A when B is observed. If A and B are related by a deterministic, invertible function, the maximal mutual information is attained [8]. By maximizing the mutual information between Δexp_{R_B} and F_{R_B} , the extractor will be encouraged to extract pure expression-related features and ignore expression-unrelated information. However, directly maximizing the mutual information is hard as it requires access to the posterior distribution. We follow [8, 12] to impose a regularizer \mathcal{Q} on top of the extractor to approximate it by maximizing its derived lower bound [3]:

$$\min_{\mathcal{Q}, \mathcal{X}} \mathcal{L}_{\mathcal{Q}} = -\mathbb{E}[\log(\mathcal{Q}(\Delta exp_{R_B} | F_{R_B}))], \quad (3)$$

By combining Eqs. (2) and (3), the overall objective function of \mathcal{X} is

$$\mathcal{L}_{\mathcal{X}} = \mathcal{L}_{exp} + \lambda_{\mathcal{Q}} \mathcal{L}_{\mathcal{Q}}, \quad (4)$$

where $\lambda_{\mathcal{Q}}$ is the hyper-parameter to balance the terms.

Learning Interpolator: With the identity attribute C_{I_N} of the input image and the expression attribute Δexp_{R_B} of the reference image, we can easily obtain the latent code C_{I_B} for the target image through linear interpolation

$$C_{I_B}^* = C_{I_N} + \Delta exp_{R_B}. \quad (5)$$

On the other hand, the linearly interpolated latent code may not reside on the manifold of real facial images and lead to weird editing (e.g. ghost faces) while projected back to the image space. Hence, we train an interpolator \mathcal{I} to generate interpolated codes and impose an adversarial regularization term on it (details of the regularization term to be discussed in Sec. 3.4) as follows:

$$\min_{\mathcal{I}} \mathcal{L}_{interp} = \mathcal{L}_{adv_{E, \mathcal{I}}} + \|\mathcal{I}(C_{I_N}, \alpha \Delta exp_{R_B}) - (C_{I_N} + \alpha \Delta exp_{R_B})\|^2, \quad (6)$$

where $\alpha \in [0, 1]$ is the interpolated factor that controls the expression intensity of the synthesized image. We can obtain a smooth transition sequences of different expressions by simply changing the value of α once the model is trained.

In addition, the interpolator should be able to recover the original latent code of the input image given his/her identity attribute and the corresponding expression attribute. The loss term can be formulated as follows:

$$\min_{\mathcal{I}} \mathcal{L}_{idt} = \|\mathcal{I}(C_{I_N}, \Delta exp_{I_A}) - C_{I_A}\|^2, \quad (7)$$

where $\Delta exp_{I_A} = \mathcal{X}(C_{I_A}, C_{I_N})$.

The final objective function for the interpolator \mathcal{I} is

$$\mathcal{L}_{\mathcal{I}} = \mathcal{L}_{interp} + \mathcal{L}_{idt}. \quad (8)$$

3.3 Expression Similarity Enhancement

To further enhance the expression similarity between the synthesized image I_B and the reference image R_B , we introduce a siamese network to encourage the synthesized images to share similar semantics with the reference image. The idea of siamese network is first introduced in natural language processing applications [18] that learns a space where the vector that transforms the word *man* to the word *woman* is similar to the vector that transforms *hero* to *heroine* [1]. In our problem, we define the difference between an expression face and its corresponding neutral face as the expression transform vector. And we minimize the difference between the expression transform vector of R_B and R_N and that of I_B and I_N . The intuition is that the transformation that turns a similar expressive face into neutral face should be analogous for different identities, which is aligned with the analysis of expression manifold [7].

Specifically, given reference image with expression B (R_B), its corresponding neutral face (R_N), synthesized image with expression B (I_B) and its corresponding neutral face (I_N), we first map them into a latent space by the siamese network \mathcal{S} and obtain the transform vectors:

$$v_R = \mathcal{S}(R_B) - \mathcal{S}(R_N), \quad (9)$$

$$v_I = \mathcal{S}(I_B) - \mathcal{S}(I_N), \quad (10)$$

then we minimize the difference between v_R and v_I :

$$\min_{\mathcal{S}} \mathcal{L}_{\mathcal{S}} = \text{Dist}(v_R, v_I), \quad (11)$$

where Dist is a distance metric. We adopt cosine similarity as the distance measurement and incorporate the siamese loss in learning the encoder.

3.4 Encoder, Decoder and GAN

Learning Encoder and Decoder: Given a collection of facial images I , we train an encoder to map them to a compact expression and identity embedded space to facilitate the disentanglement. We aim to obtain a flattened latent space so as to generate smooth transition sequences of different expressions by changing the interpolated factor (Sec. 3.2). This is achieved by minimizing the Wasserstein distance between the latent codes of real samples and the interpolated ones.

Specifically, a discriminator \mathcal{D} is learned to distinguish the real samples and the interpolated ones and the encoder \mathcal{E} and interpolator \mathcal{I} are trained to fool the discriminator. We adopt the WGAN-GP [20] to learn the parameters. The adversarial loss functions are formulated as

$$\begin{aligned} \min_{\mathcal{D}} \mathcal{L}_{adv_{\mathcal{D}}} = & \mathbb{E}_{\hat{C} \sim P_{\hat{I}}} [\log \mathcal{D}(\hat{C})] - \mathbb{E}_{C \sim P_{data}} [\log \mathcal{D}(C)] \\ & + \lambda_{gp} \mathbb{E}_{\tilde{C} \sim P_{\tilde{C}}} [(\|\nabla_{\tilde{C}} \mathcal{D}(\tilde{C})\|_2 - 1)^2], \end{aligned} \quad (12)$$

$$\min_{\mathcal{E}, \mathcal{I}} \mathcal{L}_{adv_{\mathcal{E}, \mathcal{I}}} = -\mathbb{E}_{\hat{C} \sim P_{\hat{I}}} [\log \mathcal{D}(\hat{C})], \quad (13)$$

where $C = \mathbf{E}(I)$ stands for the code generated by the encoder, \hat{C} the interpolated code generated by the interpolator \mathcal{I} , P_{data} the data distribution of the codes of real images, $P_{\hat{I}}$ the distribution of the interpolated ones and $P_{\hat{C}}$ the random interpolation distribution introduced in [20].

The model may suffer from ‘mode collapse’ problem if we simply optimize the parameters with Eqs. (12) and (13). The encoder learns to map all images to a small latent space where the real and interpolated codes are closed that yields a small Wasserstein distance. To an extreme, the Wasserstein distance could be 0 if the encoder maps all images to a single point [9]. To avoid this trivial solution, we train a decoder \mathbf{D} to project the latent codes back to the image space. We follow [9, 33] to train the decoder with perceptual loss [25] as Eq. (14), and impose an reconstruction constraint on the encoder as Eq. (15).

$$\min_{\mathbf{D}} \mathcal{L}_{\mathbf{D}} = \mathbb{E}(\|\Phi(\mathbf{D}(C)) - \Phi(I)\|^2), \quad (14)$$

$$\min_{\mathbf{E}} \mathcal{L}_{recon} = \mathbb{E}(\|\Phi(\mathbf{D}(\mathbf{E}(I))) - \Phi(I)\|^2), \quad (15)$$

where Φ is the VGG network [44] pre-trained on ImageNet [11].

The final objective function of the encoder can thus be derived as follows:

$$\mathcal{L}_{\mathbf{E}} = \mathcal{L}_{GAN_{\mathbf{E}, \mathbf{I}}} + \lambda_{recon} \mathcal{L}_{recon} + \lambda_S \mathcal{L}_{\mathbf{S}}, \quad (16)$$

where λ_{recon} and λ_S are the hyper-parameters. \mathbf{E} and \mathbf{S} are updated in an alternative manner.

Pre-training GAN: We generate the neutral face of the input and reference images by using a pre-trained GAN which can be adapted from many existing image-to-image translation models [59, 10, 56, 26]. In our experiment, we adopt the StarGAN [10] and follow the training strategy in [10] to train the model. The parameters are fixed once the GAN is trained.

4 Experiments

4.1 Dataset and Evaluation Metrics

Our experiments are conducted on two public datasets including Radboud Faces Database (RaFD) [29] and Compound Facial Expressions of Emotions Database (CFEED) [13]. RaFD consists of 8,040 facial expression images collected from 67 participants. CFEED [13] contains 5,060 compound expression images collected from 230 participants. We randomly sample 90% images for training and the rest for testing. All the images are center cropped and resized to 128×128 .

We evaluate and compare the quality of the synthesized facial expression images with different metrics, namely, Fréchet Inception Distance (FID) [22], structural similarity (SSIM) index [52], expression classification accuracy and the Amazon Mechanical Turk (AMT) user study results. The FID scores are calculated between the final average pooling features of a pre-trained inception model [46] of the real faces and the synthesized faces, and the SSIM is computed over synthesized expressions and corresponding expressions of the same identity.

Table 1. Quantitative comparison with state-of-the-art methods on datasets RaFD and CFEED by using FID (lower is better) and SSIM (higher is better).

	RaFD		CFEED	
	FID↓	SSIM↑	FID↓	SSIM↑
StarGAN [10]	62.51	0.8563	42.39	0.8011
GANimation [39]	45.55	0.8686	29.07	0.8088
Ours	38.20	0.8833	23.60	0.8194

Table 2. Quantitative comparison with state-of-the-art methods on datasets RaFD and CFEED by using facial expression classification accuracy (higher is better).

Dataset	Method	R	G	R + G
RaFD	StarGAN [10]	92.21	82.37	88.48
	GANimation [39]		84.36	92.31
	Ours		88.67	93.25
CFEED	StarGAN [10]	88.23	77.80	81.87
	GANimation [39]		79.46	84.42
	Ours		84.35	90.06

4.2 Implementation Details

Our model is trained using Adam optimizer [27] with $\beta_1 = 0.5$, $\beta_2 = 0.999$. The detailed network architecture is provided in the supplementary materials. For a fair comparison, we train StarGAN [10] and GANimation [39] using the implementations provided by the authors. In all the experiments, StarGAN [10] is trained with discrete expression labels provided in the two public datasets, while GANimation [39] is trained with AU intensities extracted by OpenFace toolkit [2]. Our network does not use any expression annotation in training.

4.3 Quantitative Evaluation

We evaluate and compare our expression editing technique with state-of-the-art StarGAN [10] and GANimation [39] quantitatively by using FID, SSIM, expression classification accuracy and user study evaluation on RaFD and CFEED.

FID and SSIM: Table 1 shows the evaluation results of all compared methods on the datasets RaFD and CFEED by using the FID and SSIM. As Table 1 shows, our method outperforms the state-of-the-art methods by a large margin in FID, with a 7.35 improvement on RaFD and a 5.47 improvement on CFEED. The achieved SSIMs are also higher than the state-of-the-art by 1.5% and 1.1% for the two datasets. All these results demonstrate the superior performance of our proposed LEED in synthesizing high fidelity expression images.

Expression Classification: We perform quantitative evaluations with expression classification as in StarGAN [10] and ExprGAN [12]. Specifically, we first train expression editing models on the training set and perform expression editing on the corresponding testing set. The edited images are then evaluated by

Table 3. Quantitative comparison with state-of-the-art methods on RaFD and CFEED by Amazon-Mechanical-Turk based user studies (higher is better for both metrics).

	RaFD		CFEED	
	Real or Fake	Which’s More Real	Real or Fake	Which’s More Real
Real	78.82	-	72.50	-
StarGAN [10]	31.76	7.06	14.37	8.75
GANimation [39]	47.06	15.29	31.87	9.38
Ours	74.12	77.65	70.63	81.87

expression classification – a higher classification accuracy means more realistic expression editing. Two classification tasks are designed: 1) train expression classifiers by using the training set images (real) and evaluate them over the edited images; 2) train classifiers by combining the real and edited images and evaluate them over the test set images. The first task evaluates whether the edited images lie in the manifold of natural expressions, and the second evaluates whether the edited images help train better classifiers.

Table 2 shows the classification accuracy results (only seven primary expressions evaluated for CFEED). Specifically, **R** trains classifier with the original training set images and evaluates on the corresponding testing set images. **G** applies the same classifier (in **R**) to the edited images. **R + G** trains classifiers by combining the original training images and the edited ones, and evaluates on the same images in **R**. As Table 2 shows, LEED outperforms the state-of-the-art by 4.31% on RaFD and 4.89% on CFEED, respectively. Additionally, the LEED edited images help to train more accurate classifiers while incorporated in training, where the accuracy is improved by 1.04% on RaFD and 1.83% on CFEED, respectively. They also outperform StarGAN and GANimation edited images, the latter even degrade the classification probably due to the artifacts within the edited images as illustrated in Fig. 2. The two experiments demonstrate the superiority of LEED in generating more realistic expression images.

User Studies: We also evaluate and benchmark the LEED edited images by conducting two Amazon-Mechanical-Turk (AMT) user studies under two evaluation metrics: 1) Real or Fake: subjects are presented with a set of expression images including real ones and edited ones by LEED, GANimation, and StarGAN, and tasked to identify whether the images are real or fake; 2) Which’s More Real: subjects are presented by three randomly-ordered expression images edited by the three methods, and are tasked to identify the most real one. Table 3 shows experimental results, where LEED outperforms StarGAN and GANimation significantly under both evaluation metrics. The two user studies further demonstrate the superior perceptual fidelity of the LEED edited images.

4.4 Qualitative Evaluation

Fig. 2 shows qualitative experimental results with images from RaFD (cols 1-5) and CFEED (cols 6-10). Each column shows an independent expression edit-

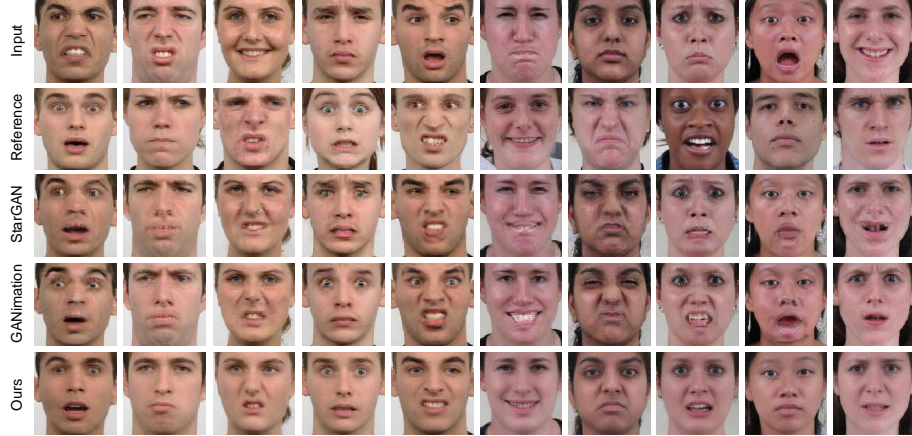


Fig. 2. Expression editing by LEED and state-of-the-art methods: Columns 1-5 show the editing of RaFD images, and columns 6-10 show the editing of CFEED images. Our method produces more realistic editing with better details and less artifacts.

ing, including an input image and a reference image as well as editing by StarGAN [10], GANimation [39] and our proposed LEED.

As Fig. 2 shows, StarGAN [10] and GANimation [39] tend to generate blurs and artifacts and even corrupted facial regions (especially around eyes and mouths). LEED can instead generate more realistic facial expressions with much less blurs and artifacts, and the generated images are also clearer and sharper. In addition, LEED preserves the identity information well though it does not adopt any identity preservation loss, largely due to the identity disentanglement which encodes the identity information implicitly.

4.5 Ablation Study

We study the two designed losses by training three editing networks on RaFD: 1) a network without the mutual expression information loss (regularizer \mathbf{Q}) as labelled by ‘w/o \mathbf{Q} ’; 2) a network without the siamese loss as labelled by ‘w/o \mathbf{S} ’; and 3) a network with both losses as labelled by ‘Final’ in Fig. 3. As Fig. 3 shows, the mutual expression information loss guides the extractor to extract pure expression-relevant features. When it is absent, the extracted expression is degraded by expression-irrelevant information which leads to undesired editing such as eye gazing direction changes (column 1), head pose changes (columns 2, 3, 5 and 8), and identity attribute changes (missing mustache in column 4). The siamese loss enhances the expression similarity of the edited and reference images without which the expression intensity of the edited images becomes lower than that of the reference image (columns 2, 3, 6 and 7) as illustrated in Fig. 3.

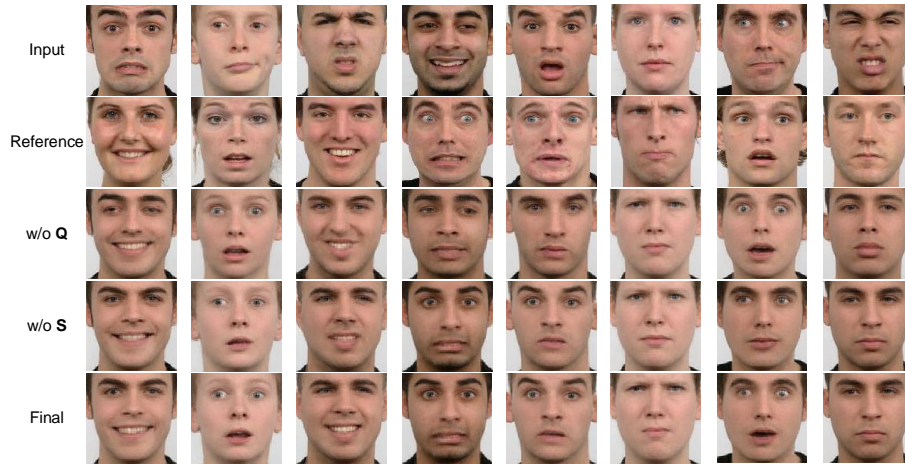


Fig. 3. Ablation study of LEED over RaFD: From top to bottom: input image, reference image, editing without mutual expression information loss, editing without siamese loss, final result. The graphs show the effectiveness of our designed losses.

4.6 Discussion

Feature Visualization: We use t-SNE [34] to show that LEED learns the right expression features via disentanglement. Besides the *Extractor features*, we also show the *Encoder features* (i.e. the dimension reduced representation of original image) and the *Residual features* (i.e. the difference between the Encoder features of expressive and neutral faces) as illustrated in Fig. 4 (learnt from the RaFD images). As Fig. 4 shows, the *Encoder features* and *Residual features* cannot form compact expression clusters as the former learns entangled features and the latter contains expression-irrelevant features such as head-poses variations. As a comparison, the *Extractor features* cluster each expression class compactly thanks to the mutual expression information loss.

Expression Editing on Profile Images: LEED is capable of ‘transferring’ expression across profile images of different poses. As illustrated in Fig. 5, LEED produces realistic expression editing with good detail preservation whereas StarGAN introduces lots of artifacts (GANimation does not work as OpenFace cannot extract AUs accurately from profile faces). The capability of handling profile images is largely attributed to the mutual expression information loss that helps extract expression related features in the reference image. Note AF-VAE [40] can also work with non-frontal profile images but it can only transfer expressions across facial images of the same pose.

Robustness to Imperfect Neutral Expression Images: LEED uses a pre-train GAN to generate neutral expression images for the disentanglement but the generated neutral face may not be perfect as illustrated in Fig. 6 (row 3). LEED is tolerant to such imperfection as shown in Fig. 6 (row 4), largely because of

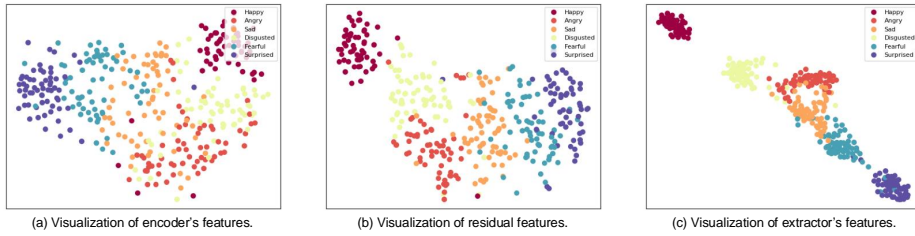


Fig. 4. Expression feature Visualization with t-SNE: The *Extractor* learns much more compact clusters for expression features of different classes. Best view in colors.

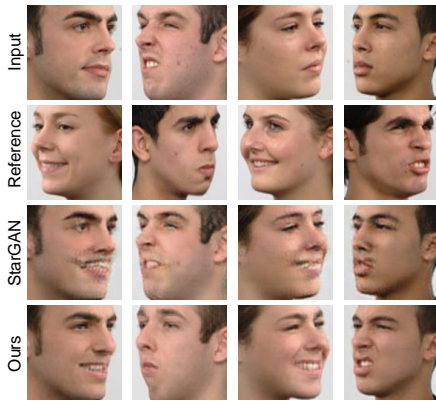


Fig. 5. LEED can ‘transfer’ expressions across profile images of different poses whereas state-of-the-art StarGAN tends to produce clear artifacts.

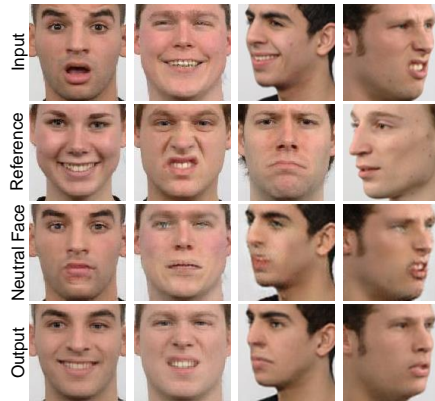


Fig. 6. Though the GAN-generated neutral faces may not be perfect, LEED still generate sharp and clear expression thanks to our adopted adversarial loss.

the adversarial loss that is included into the interpolated latent code. However, the imperfect neutral face of reference image may contain residual of the original expression and lead to lower expression intensity in the output. This issue could be mitigated by adopting a stronger expression normalization model.

Continuous Editing: Our method can generate continuous expression sequences by changing the interpolated factor α (Sec. 3.2) as shown in Fig. 7. Besides interpolation, we show that the extrapolation can generate extreme expressions. This shows our method could uncover the structure of natural expression manifolds.

Facial Expression Editing on Wild Images: FEE for wild images is much more challenging as the images have more variations in complex background, uneven lighting, etc. LEED can adapt to handle wild images well as illustrated in Fig. 8, where the model is trained on expressive images sampled from AffectNet [35]. As Fig. 8 shows, LEED can transform the expressions successfully while maintaining the expression-unrelated information unchanged.

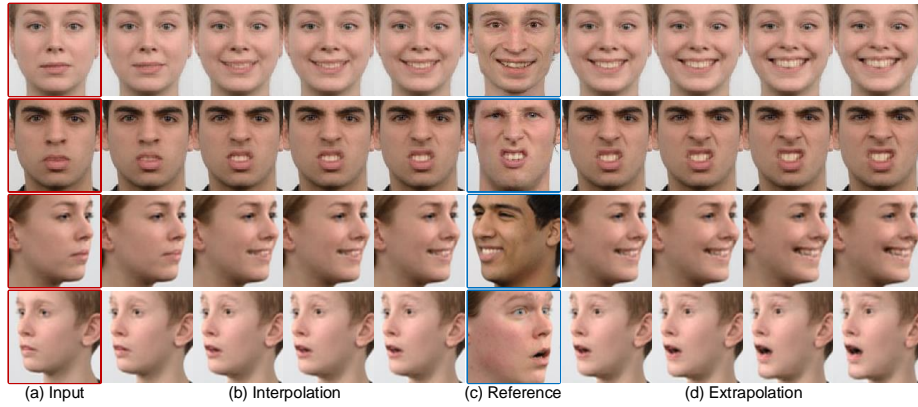


Fig. 7. Expression editing via interpolation/extrapolation: Given input images in (a) and reference images in (c), LEED can edit expressions by either interpolation ($\alpha < 1$) or extrapolation ($\alpha > 1$) as shown in (b) and (d).



Fig. 8. Facial expression editing by LEED on wild images: In each triplet, the first column is input facial image, the second column is the image with desired expression and the last column is the synthesized result.

5 Conclusion

We propose a novel label-free expression editing via disentanglement (LEED) framework for realistic expression editing of both frontal and profile facial images without any expression annotation. Our method disentangles the identity and expression of facial images and edits expressions by fusing the identity of the input image and the expression of the reference image. Extensive experiments over two public datasets show that LEED achieves superior expression editing as compared with the state-of-the-art techniques. We expect that LEED will inspire new insights and attract more interests for better FEE in the near future.

Acknowledgement

This work is supported by Data Science & Artificial Intelligence Research Centre, NTU Singapore.

References

1. Amodio, M., Krishnaswamy, S.: Travelgan: Image-to-image translation by transformation vector learning. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 8983–8992 (2019)
2. Baltrusaitis, T., Zadeh, A., Lim, Y.C., Morency, L.P.: Openface 2.0: Facial behavior analysis toolkit. In: *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. pp. 59–66. IEEE (2018)
3. Barber, D., Agakov, F.V.: The im algorithm: a variational approach to information maximization. In: *Advances in neural information processing systems*. p. None (2003)
4. Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence* **35**(8), 1798–1828 (2013)
5. Blanz, V., Vetter, T., et al.: A morphable model for the synthesis of 3d faces. In: *Siggraph*. vol. 99, pp. 187–194 (1999)
6. Cao, C., Weng, Y., Zhou, S., Tong, Y., Zhou, K.: Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics* **20**(3), 413–425 (2013)
7. Chang, Y., Hu, C., Feris, R., Turk, M.: Manifold based analysis of facial expression. *Image and Vision Computing* **24**(6), 605–614 (2006)
8. Chen, X., Duan, Y., Houthoofd, R., Schulman, J., Sutskever, I., Abbeel, P.: Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In: *Advances in neural information processing systems*. pp. 2172–2180 (2016)
9. Chen, Y.C., Xu, X., Tian, Z., Jia, J.: Homomorphic latent space interpolation for unpaired image-to-image translation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 2408–2416 (2019)
10. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 8789–8797 (2018)
11. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: *2009 IEEE conference on computer vision and pattern recognition*. pp. 248–255. Ieee (2009)
12. Ding, H., Sricharan, K., Chellappa, R.: Exprgan: Facial expression editing with controllable expression intensity. In: *Thirty-Second AAAI Conference on Artificial Intelligence* (2018)
13. Du, S., Tao, Y., Martinez, A.M.: Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences* **111**(15), E1454–E1462 (2014)
14. Ekman, P., Friesen, W., Hager, J.: Facial action coding system (facs) a human face. Salt Lake City (2002)
15. Friesen, E., Ekman, P.: Facial action coding system: a technique for the measurement of facial movement. Palo Alto **3** (1978)
16. Geng, J., Shao, T., Zheng, Y., Weng, Y., Zhou, K.: Warp-guided gans for single-photo facial animation. *ACM Transactions on Graphics (TOG)* **37**(6), 1–12 (2018)
17. Geng, Z., Cao, C., Tulyakov, S.: 3d guided fine-grained face manipulation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 9821–9830 (2019)

18. Goldberg, Y., Levy, O.: word2vec explained: deriving mikolov et al.’s negative-sampling word-embedding method. arXiv preprint arXiv:1402.3722 (2014)
19. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672–2680 (2014)
20. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein gans. In: Advances in Neural Information Processing Systems. pp. 5767–5777 (2017)
21. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
22. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in Neural Information Processing Systems. pp. 6626–6637 (2017)
23. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A.: beta-vae: Learning basic visual concepts with a constrained variational framework. ICLR **2**(5), 6 (2017)
24. Jiang, Z.H., Wu, Q., Chen, K., Zhang, J.: Disentangled representation learning for 3d face shape. arXiv preprint arXiv:1902.09887 (2019)
25. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European conference on computer vision. pp. 694–711. Springer (2016)
26. Kim, T., Cha, M., Kim, H., Lee, J.K., Kim, J.: Learning to discover cross-domain relations with generative adversarial networks. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. pp. 1857–1865. JMLR. org (2017)
27. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
28. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
29. Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D.H., Hawk, S.T., Van Knippenberg, A.: Presentation and validation of the radboud faces database. Cognition and emotion **24**(8), 1377–1388 (2010)
30. Li, H., Weise, T., Pauly, M.: Example-based facial rigging. Acm transactions on graphics (tog) **29**(4), 1–6 (2010)
31. Li, M., Zuo, W., Zhang, D.: Deep identity-aware transfer of facial attributes. arXiv preprint arXiv:1610.05586 (2016)
32. Li, S., Deng, W., Du, J.: Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2584–2593. IEEE (2017)
33. Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., Yang, M.H.: Universal style transfer via feature transforms. In: Advances in neural information processing systems. pp. 386–396 (2017)
34. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(Nov), 2579–2605 (2008)
35. Mollahosseini, A., Hasani, B., Mahoor, M.H.: Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing **10**(1), 18–31 (2017)
36. Nagano, K., Seo, J., Xing, J., Wei, L., Li, Z., Saito, S., Agarwal, A., Fursund, J., Li, H.: pagan: real-time avatars using dynamic textures. ACM Transactions on Graphics (TOG) **37**(6), 1–12 (2018)

37. Narayanaswamy, S., Paige, T.B., Van de Meent, J.W., Desmaison, A., Goodman, N., Kohli, P., Wood, F., Torr, P.: Learning disentangled representations with semi-supervised deep generative models. In: *Advances in Neural Information Processing Systems*. pp. 5925–5935 (2017)
38. Peng, X., Yu, X., Sohn, K., Metaxas, D.N., Chandraker, M.: Reconstruction-based disentanglement for pose-invariant face recognition. In: *Proceedings of the IEEE international conference on computer vision*. pp. 1623–1632 (2017)
39. Pumarola, A., Agudo, A., Martinez, A.M., Sanfeliu, A., Moreno-Noguer, F.: Ganimation: Anatomically-aware facial animation from a single image. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. pp. 818–833 (2018)
40. Qian, S., Lin, K.Y., Wu, W., Liu, Y., Wang, Q., Shen, F., Qian, C., He, R.: Make a face: Towards arbitrary high fidelity face manipulation. In: *Proceedings of the IEEE International Conference on Computer Vision*. pp. 10033–10042 (2019)
41. Qiao, F., Yao, N., Jiao, Z., Li, Z., Chen, H., Wang, H.: Geometry-contrastive gan for facial expression transfer. *arXiv preprint arXiv:1802.01822* (2018)
42. Shen, W., Liu, R.: Learning residual images for face attribute manipulation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 4030–4038 (2017)
43. Shu, Z., Yumer, E., Hadap, S., Sunkavalli, K., Shechtman, E., Samaras, D.: Neural face editing with intrinsic image disentangling. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 5541–5550 (2017)
44. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014)
45. Song, L., Lu, Z., He, R., Sun, Z., Tan, T.: Geometry guided adversarial facial expression synthesis. In: *2018 ACM Multimedia Conference on Multimedia Conference*. pp. 627–635. ACM (2018)
46. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: *Thirty-First AAAI Conference on Artificial Intelligence* (2017)
47. Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., Nießner, M.: Face2face: Real-time face capture and reenactment of rgb videos. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 2387–2395 (2016)
48. Upchurch, P., Gardner, J., Pleiss, G., Pless, R., Snavely, N., Bala, K., Weinberger, K.: Deep feature interpolation for image content changes. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 7064–7073 (2017)
49. Vlasic, D., Brand, M., Pfister, H., Popovic, J.: Face transfer with multilinear models. In: *ACM SIGGRAPH 2006 Courses*, pp. 24–es (2006)
50. Wang, J., Zhang, J., Lu, Z., Shan, S.: Dft-net: Disentanglement of face deformation and texture synthesis for expression editing. In: *2019 IEEE International Conference on Image Processing (ICIP)*. pp. 3881–3885. IEEE (2019)
51. Wang, Y., Gong, D., Zhou, Z., Ji, X., Wang, H., Li, Z., Liu, W., Zhang, T.: Orthogonal deep features decomposition for age-invariant face recognition. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. pp. 738–753 (2018)
52. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., et al.: Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* **13**(4), 600–612 (2004)
53. Wu, R., Zhang, G., Lu, S., Chen, T.: Cascade ef-gan: Progressive facial expression editing with local focuses. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 5021–5030 (2020)

- 54. Wu, W., Zhang, Y., Li, C., Qian, C., Change Loy, C.: Reenactgan: Learning to reenact faces via boundary transfer. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 603–619 (2018)
- 55. Wu, X., Huang, H., Patel, V.M., He, R., Sun, Z.: Disentangled variational representation for heterogeneous face recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 9005–9012 (2019)
- 56. Xiao, T., Hong, J., Ma, J.: Elegant: Exchanging latent encodings with gan for transferring multiple face attributes. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 168–184 (2018)
- 57. Yang, L., Yao, A.: Disentangling latent hands for image synthesis and pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9877–9886 (2019)
- 58. Zhang, G., Kan, M., Shan, S., Chen, X.: Generative adversarial network with spatial attention for face attribute editing. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 417–432 (2018)
- 59. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2223–2232 (2017)