Learning Identity-Invariant Motion Representations for Cross-ID Face Reenactment

Po-Hsiang Huang[1], Fu-En Yang[1], Yu-Chiang Frank Wang[1,2]
[1]Graduate Institute of Communication Engineering, National Taiwan University, Taiwan
[2]ASUS Intelligent Cloud Services, Taiwan
{r06942138, r07942077, ycwang}@ntu.edu.tw

Abstract

*Human face reenactment aims at transferring motion patterns from one face (from a source-domain video) to another (in the target domain with the identity of interest). While recent works report impressive results, they are not able to handle multiple identities in a unified model. In this paper, we propose a unique network of **CrossID-GAN** to perform multi-ID face reenactment. Given a source-domain video with extracted facial landmarks and a target-domain image, our CrossID-GAN learns the identity-invariant motion patterns via the extracted landmarks and such information to produce the videos whose ID matches that of the target domain. Both supervised and unsupervised settings are proposed to train and guide our model during training. Our qualitative/quantitative results confirm the robustness and effectiveness of our model, with ablation studies confirming our network design.*

1. Introduction

Video retargeting aims at transferring motion patterns of video from a source-domain video to another in the target domain of interest, while the content (e.g., face identity or body shape) of the latter remains unchanged. For example, converting the blooming process across distinct flowers, or adapting pose and expression from one person to another, can be considered as examples of video retargeting. Due to a wide range of applications such as virtual reality and film production, transferring motion/expression of face videos are of particular interest. With the focus of retargeting face videos, one typically refer to this task as *face reenactment*. To tackle the task of face reenactment, existing approaches typically apply predefined parametric 3D models to represent human faces. Although exhibiting promising ability in describing inner faces [17] [16], movement of the human head might not be sufficiently modeled. Moreover, such methods might not be able to describe subtle movements of human faces. Thus, a large amount of effort and delicacy designs of complex parametric fitting would be required. With the recent advances of Generative Adversarial Networks (GANs) [3], which shows great potential in producing realistic images, many GAN-based methods have been proposed as alternatives to 3D parametric models. For example, CycleGAN [26] and RecycleGAN [1] are among well-known data-driven GAN models for human face reenactment. However, such methods only support 1-to-1 mapping between two face identities. In other words, one needs to train a face reenactment model for each pair of persons. Therefore, its extension of multiple persons with unseen motion would be the limitation.

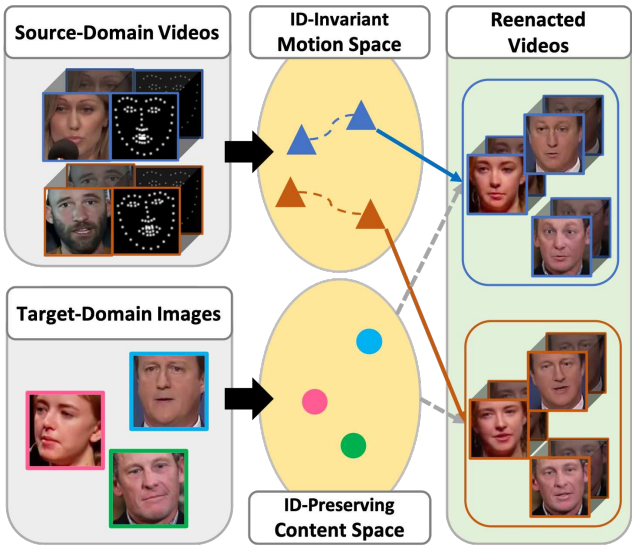

Figure 1. Illustration of our CrossID-GAN for face reenactment. By encoding a sequence of facial landmarks from a source-domain video, we transform a target-domain image (of a different identity) with motion patterns associated with those of the source domain, while the identity of the target-domain is preserved.

On the other hand, another group of researchers incorporate facial landmarks/face boundary, which provide direct

	No Attribute Labels of Required	Landmark Guidance	Unified Network	Target Identity Preservation	Mapping Domain
DR-GAN [19]	-	-	√	√	-
CR-GAN [18]	-	-	√	√	-
GC-GAN [12]	-	√	√	√	-
Cycle-GAN [26]	√	-	-	√	1-to-1
Recycle-GAN [1]	√	-	-	√	1-to-1
GANnotation [13]	√	√	√	-	-
ReenactGAN [23]	√	√	-	√	many-to-1
Talking-Head [25]	√	√	√	-	-
X2Face [22]	√	-	√	-	many-to-many
CrossID-GAN (Ours)	√	√	√	√	many-to-many

Table 1. Comparisons of recent works on face reenactment. While transfer of face attributes needs to be performed, the first three methods are designed to do so in the image level, and thus they cannot produce satisfactory video sequence outputs.

guidance of facial pose and expression, and thus can easily guide the model to generate desirable outputs. Existing techniques of facial landmarks detection are sufficiently precise and efficient, which makes the use of such models practical. However, facial landmarks contain information of the observed input face image (e.g., shape or identity of faces), which cannot be directly utilized to generate the target-domain face image outputs. As a result, how to transform the source-domain face landmarks to the target domain while disregarding identity-dependent information becomes a challenging problem. Wu *et al.* propose a GAN-based model, ReenactGAN [23], for many-to-1 face reenactment, which means that they are able to produce output video of *specific identity* with the motion pattern transferred from arbitrary source videos. With the aids of ID-specific transformer and ID-specific decoder, ReenactGAN is able to transform source-domain face landmarks to the target domain, so that reenactment of images with target-domain identity can be achieved. Despite its promising performance, it cannot generate videos of distinct identities in a unified model. To achieve face reenactment across multiple identities (i.e., many-to-many mapping) with a unified model, one would require the learning of ID-invariant representation from the observed facial landmarks. In other words, such a representation would only describe the information of pose and expression from the input face image, which allows the generator to produce images with preferred identities (i.e., images in the target domain). In this paper, we propose a novel model **CrossID-GAN** for such *many-to-many* face reenactment tasks. Our CrossID-GAN is able to transfer the pose and expression from source-domain face images to the target-domain one, while the extracted pose and expression features are identity independent. As illustrated in Fig. 1, our model is designed to derive ID-invariant motion information from the input source images/videos, while being able to transfer motion to face images of desirable identities with the guidance from target-domain image and the aforementioned facial landmarks. Table 1 highlights and compares recent works on face reenactment, showing the advantage of our CrossID-GAN. To the best of our knowledge, existing face reenactment approaches are not able to extract such ID-invariant features across multiple persons in a unified model (as ours does). The contributions of this paper are highlighted below.

- We propose CrossID-GAN for cross-identity face reenactment. With only identity labels observed (not the ground truth landmarks), our model extracts and transfers ID-invariant landmark information across face images of different identities.
- Our landmark encoder extracts ID-invariant facial landmark representation describing pose and expression information, while our content encoder extracts ID-preserving content information for identity recovery guarantees.
- Our CrossID-GAN can be trained in both supervised and unsupervised settings. The former requires pairwise training input data for encoding the aforementioned facial landmark and content representations, while the latter does not require such correspondence information.

2. Related Work

2.1. Human Face Reenactment

Previously, most existing works on face reenactment apply 3D parametric models to fit human faces. For example, Face2Face [17] utilizes a 3D parametric model with 269 parameters to fit pose, expression, illumination, and shape information of a human face. With detected and tracked face images, it would transfer the expression of source-domain faces to the target-domain one by adjusting the associated parameters. Since only the appearance of the inner face can be modeled, subtle details might not be described and

altered properly. With the development of deep learning models, methods like X2Face [22] show promising reenactment with improved warping techniques. X2Face consists of two sub-networks: an embedding network aiming to learn an embedded face representation across input target-domain images, and a driving network outputs a pixel sampler (i.e., optical flow map) which warps from this embedded face with the pose of interest. Since no facial landmark information is considered, pose and expression might not be transferred accurately. Also, due to the limitation of warping based methods, missing details in the embedded faces still make the produced outputs less realistic and thus suffer from distortion due to extreme poses.

GAN-based reenactment without facial landmarks. GAN-based methods have recently been utilized for face reenactment, due to its success in producing realistic face image outputs. And, with the structural similarity between face images, face reenactment can also be considered as a frame-by-frame style translation problem. CycleGAN [26] introduced a cycle consistency loss to find an appropriate mapping between two domains, but it sometimes encounters problems such as perceptual mode collapse or generated output spatially tied to input, which are mentioned in [1]. RecycleGAN applied predictors of two domains to further consider temporal information to deal with the problems and extended the type of domains to a wider range. However, both of CycleGAN and RecycleGAN are designed only for 1-to-1 mapping, namely transferring images between only two domains. This limited the potential of models in real-life applications. Another problem is that the training data hardly cover all the pose of human, this may cause the model to learn an inappropriate mapping between two domains.

GAN-based reenactment with facial landmarks. GAN-based models integrating facial landmarks have also been proposed for face reenactment. For example, GANnotation [13] takes in the concatenation of source landmarks heatmap and target identity image for face reenactment. However, it cannot properly transfer motion due to identity mismatch across domains. ReenactGAN [23] first trains a face boundary encoder to map the source frame onto an embedding space. Then, for every target identity, an ID-specific boundary transformer converts the encoded source face boundary to the shape matching the target identity. Finally, an ID-specific decoder decodes the transformed boundary to the reenacted image. While this work is able to perform reenactment, different face boundary transformers and decoders are required for each target identity. In other words, it cannot advance a unified framework for multi-ID face reenactment.

Recently, [25] utilize a pair of encoder and generator with meta-learning strategies and adaptive instance normalization mechanisms. With the observed source-domain landmarks, it synthesizes a series of images fitting the identity of the target-domain inputs. Nevertheless, since the extracted facial landmarks are not designed to adapt to the target-domain faces, the shape might be distorted which makes the transferred outputs less realistic (i.e., less consistent with the target-domain identity).

2.2. Feature Disentanglement

Aimed at learning interpretable data representation, feature disentanglement has drawn lots of attention in recent years. In the following, we mainly focus on the disentanglement technique applied in human face and video regime.

Feature Disentanglement for Human Faces. The goal of this research topic is to disentangle a representation of human face into two parts: an identity related (content) code and a pose or expression related (attribute) code. Given the label of pose view and correspond human face image, DR-GAN [19] utilizes an encoder-decoder network to learn a pose-invariant identity representation, which is preferable for face recognition with large pose discrepancy. To deal with the problem of learning incomplete representation in DR-GAN, CR-GAN [18] introduces a generation sideway to maintain the completeness of the learned embedding space. With extracted facial landmarks and correspond expression labels during training, GC-GAN [12] transfers the expression of the source face to the target person. While these models are able to transfer face attribute properly, they require pose/expression labels for every image during training. Moreover, these works are only designed for image-to-image translation but not videos, the discrete attributes labels are not able to cover the whole movement of human action.

Feature Disentanglement in Videos. As a representative work of video synthesis with disentanglement ability, MoCoGAN [20] divides video representations into content and motion features. By sampling a point in the content subspace and sampling different trajectories in the motion subspace, it allows the generation of videos of the same object performing different motions. However, this model is mainly designed for video generation, and cannot be easily extended to cross-identity human face reenactment.

3. Methods

Given a source-domain video $\boldsymbol{V_s} = \{v_s^1, ..., v_s^T\}$ with T frames and the corresponding facial landmarks $H_s = \{h_s^1, ..., h_s^T\}$, and a target-domain image I of identity x, our goal is to generate an output video $\tilde{\boldsymbol{V}}_{\boldsymbol{x}}$ while acting like $\boldsymbol{V_s}$.

We deal with the problem of face reenactment in a frame by frame manner with temporal continuity guarantees. To

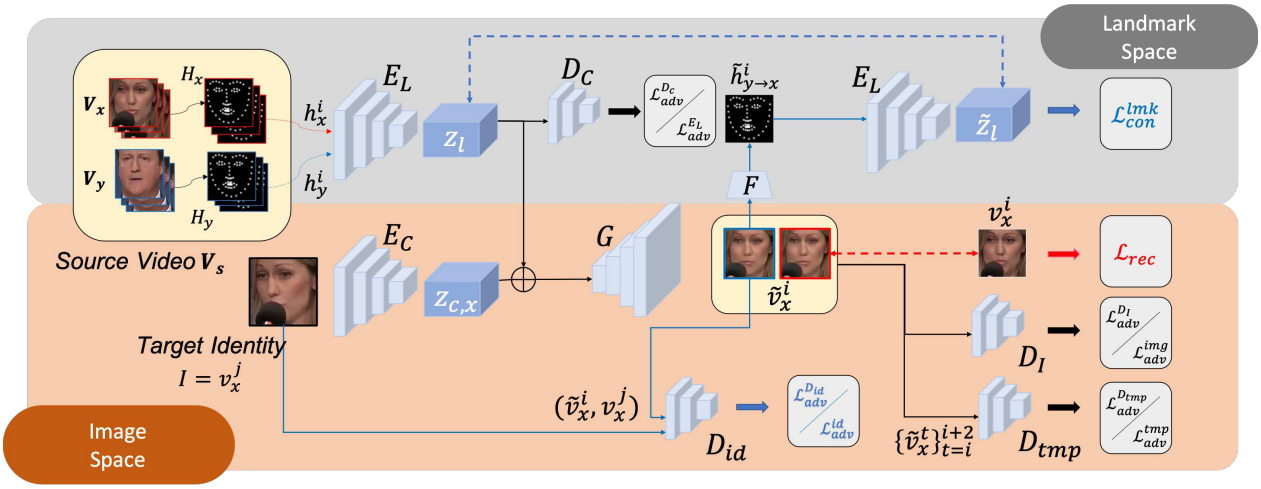

Figure 2. Our CrossID-GAN with learning ID-invariant facial landmarks for face reenactment. Since the observed source-domain landmarks can be from the same or distinct identity as that in the target domain during training, our model can be trained in supervised or unsupervised settings. (Note that arrows in red/blue indicate supervised/unsupervised training processes.)

be more specific, our model takes in the target-domain image $I = v_x$ and the facial landmarks frame $h_s^i \in H_s$ extracted from the source-domain input $v_s^i \in V_s$, and the produced output frame would be $\tilde{v}_x^i$. Note that the identity s can be the same or different from that of the target domain (i.e., $s = x$ or y) during training, while we consider $s = y$ in testing for reenactment purposes.

In our work, we propose a network model of *crossID-GAN* for motion-adapted yet ID-preserving face reenactment. As depicted in Fig. 2, our crossID-GAN contains three major components: an ID-invariant facial landmark encoder E_L, an ID-preserving content encoder E_C, and a conditional generator G. We have E_L aim at encoding observed landmarks h_s to ID-invariant motion latent code z_l, E_C maps the input I to an identity latent code $z_{c,I}$, and G take both latent codes for producing the video of interest. Additional discriminators are deployed for allowing the learning of the above network modules. The details of our proposed network architecture and learning strategies will be presented in the following subsections.

3.1. Supervised Learning of CrossID-GAN

We now first discuss how our CrossID-GAN can be trained in a supervised setting (i.e., both source-domain input video frames/landmarks h_x and the target-domain input image $I = v_x$ are of the same identity x).

As illustrated in Figure 2, our landmark encoder E_L encodes source landmarks h_x^i, resulting a landmark latent code z_l^i. On the other hand, we have the content encoder E_C to encode the target-domain input v_x^j from a different frame j and derive the content latent code $z_{c,x}$ for identity x. The generator G of our CrossID-GAN takes in the concatenation of both codes produce the output image $\tilde{v}_x^i$, with the goal to transfer the extracted landmark from h_x^i. Since the inputs from both domains are for the same identity, we expect v_x^i to serve as the ground truth for the generator output $\tilde{v}_x^i$, which is calculated by the L2-norm loss and perceptual loss [8] as the reconstruction loss $\mathcal{L}_{rec}$. Note that the perceptual loss enforces the features of $\tilde{v}_x^i$ to be visually similar to v_x^i, in which we apply a VGG-19 [15] for extracting visual features. Thus, we have $\mathcal{L}_{rec}$ defined as

$$\mathcal{L}_{rec} = ||\tilde{v}_x^i - v_x^i||^2 + \sum_f |\Phi_{VGG}^f(\tilde{v}_x^i) - \Phi_{VGG}^f(v_x^i)|. \tag{1}$$

where f denotes the layer index of VGG19.

In order to enhance the quality of the produced image outputs, we employ an image discriminator D_I which distinguishes between the synthesized image $\tilde{v}_x^i = G(z_{c,x}, z_l^i)$ and the real ones v_x^i, which is realized as follows

$$\begin{aligned} \mathcal{L}_{adv}^{D_I} &= \mathbb{E}[\log(D_I(\tilde{v}_x^i))] - \mathbb{E}[\log(1 - D_I(v_x^i)], \\ \mathcal{L}_{adv}^{img} &= -\mathbb{E}[\log(D_I(\tilde{v}_x^i))]. \end{aligned} \tag{2}$$

3.2. Unsupervised Learning of CrossID-GAN

It is worth noting that, the above supervised learning scheme only produces image outputs with landmarks extracted and adapted from visual data of the same identity x. To perform practical face reenactment, the source-domain inputs v_s^i and the target-domain image $I = v_x$ would belong to different identities (i.e., $s = y$). Moreover, the above training strategy does *not* guarantee that the landmark latent code extracted by E_L is *ID-invariant*, since the source-domain videos are of the same identity as that of the target-domain input image. To further allow the effectiveness of our reenactment network while not requiring landmark-level ground truth information, we extend our

training strategy to unsupervised learning, which simply requires the source-domain inputs are from a distinct identity y instead of x.

To address the above problem, we first utilize adversarial learning on the landmark latent code z_l by introducing a content/ID classifier D_C, which is jointly trained with E_L via reversal back propagation to ensure the invariance of z_l to identity. The associated loss functions are defined as follows:

$$\begin{aligned}\mathcal{L}_{adv}^{D_C} &= \mathbb{E}[\log P(l_r = r_s | E_L(h_s^i))] \\ \mathcal{L}_{adv}^{E_L} &= -\mathbb{E}[\log P(l_r = r_s | E_L(h_s^i))],\end{aligned} \tag{3}$$

where P is the probability distribution over domains l_r, and the identity representation r_s is a one-hot vector. Thus, our E_L is enforced to encode ID-invariant information (pose/expression) from h_s^i without face shape and contour information.

To further allow the source and target domain inputs to our CrossID-GAN are with different identities for practical reenactment purposes, our model needs to be learned in such unsupervised setting (i.e., no correspondence between training input data across domains). To achieve this goal, we additionally advance the consistency of ID-invariant landmark codes between the output $\tilde{v}_x^i$ and the *source-domain* input v_y^i, as well as the consistency of identity between $\tilde{v}_x^i$ and *target-domain* input v_x^j.

As shown in Fig. 2, we have F as a pretrained landmark extractor. As for ID-invariant landmark consistency, it is realized by minimizing the loss $\mathcal{L}_{con}^{lmk}$, which measures the L1 difference between the landmark latent codes of $\tilde{v}_x^i$ and v_y^i.

Thus, such a consistency loss $\mathcal{L}_{con}^{lmk}$ is calculated as

$$\mathcal{L}_{con}^{lmk} = |z_l^i - \tilde{z}_l^i|. \tag{4}$$

To ensure identity consistency between $\tilde{v}_x^i$ and v_x^j, we calculate their perceptual loss via a Siamese discriminator network D_{id}, which is trained to distinguish between the real image pair (v_x^n, v_x^m) sampled from the real videos and the synthesized one $(\tilde{v}_x^i, v_x^j)$. As a result, we calculate D_{id} by

$$\begin{aligned}\mathcal{L}_{adv}^{id} &= -\mathbb{E}[\log(D_{id}(\tilde{v}_x^i, v_x^j))] \\ \mathcal{L}_{adv}^{D_{id}} &= \mathbb{E}[\log(D_{id}(\tilde{v}_x^i, v_x^j))] - \mathbb{E}[\log(1 - D_I(v_x^n, v_x^m)].\end{aligned} \tag{5}$$

Similarly, we can apply existing pretrained face recognition models like VGGFace [10] or LightCNN [24] to extract visual features and to calculate the associated perceptual losses.

3.3. Ensuring Temporal Smoothness

While the learning and design details for each encoder, generator, and discriminator are described above, one typically consider an effective face reenactment model to produce video outputs instead of images. To better model temporal continuity across each output frame, we finally deploy a video discriminator [7] D_{tmp}, which consists of 3D-convolutional layers, to enforce temporal smoothness and thus improve the quality of output videos. More precisely, we take three consecutive frames as the input video sequence $\{\tilde{v}_x^t\}_{t=i}^{i+2}$ to D_{tmp} in each iteration during training, and D_{tmp} is trained to distinguish between the real sequence $\{v_x^t\}_{t=n}^{n+2}$ sampled from the source-domain video and the fake ones (i.e., $\{\tilde{v}_x^t\}_{t=i}^{i+2}$) generated by our model. The resulting temporal smoothness loss of D_{tmp} is then defined as follows:

$$\begin{aligned}\mathcal{L}_{adv}^{tmp} &= -\mathbb{E}[\log(D_{tmp}(\{\tilde{v}_x^t\}_{t=i}^{i+2}))], \\ \mathcal{L}_{adv}^{D_{tmp}} &= \mathbb{E}[\log(D_{tmp}(\{\tilde{v}_x^t\}_{t=i}^{i+2}))] \\ &\quad - \mathbb{E}[\log(1 - D_{tmp}(\{v_x^t\}_{t=n}^{n+2}))].\end{aligned} \tag{6}$$

3.4. Full Objectives

To summarize the training of our proposed CrossID-GAN, the full objective function as listed below:
For the supervised setting, we have

$$\mathcal{L}_{pair} = \mathcal{L}_{rec} + \mathcal{L}_{adv}^{img}. \tag{7}$$

As for unsupervised learning scheme, we observe

$$\mathcal{L}_{unpair} = \mathcal{L}_{con}^{lmk} + \mathcal{L}_{adv}^{E_L} + \mathcal{L}_{adv}^{img} + \mathcal{L}_{adv}^{id} + \mathcal{L}_{adv}^{tmp}. \tag{8}$$

And, the discriminators are trained by calculating

$$\mathcal{L}_{dis} = \mathcal{L}_{adv}^{D_I} + \mathcal{L}_{adv}^{D_C} + \mathcal{L}_{adv}^{D_{id}} + \mathcal{L}_{adv}^{D_{tmp}}. \tag{9}$$

It is worth noting that, both supervised and unsupervised learning schemes are jointly considered when training our CrossID-GAN. In the following section, we will present our qualitative and quantitative face reenactment results, and provide comparisons to baseline or recent reenactment models.

4. Experiments

In this section, we describe the dataset we use, the implementation details of our proposed method, and the experiments we do to verify the effectiveness of our CrossID-GAN. Qualitative and quantitative comparisons with state-of-the-art methods are performed to demonstrate the superiority of our model. Also, we provide ablation studies to explore the effectiveness of our proposed components. Finally, experiments show that our model owns the ability to generalize to source landmarks from unseen identities.

4.1. Dataset

300 Videos in the Wild. We adopt 300VW [14] video dataset in our experiments. 300VW contains 114 videos of

Figure 3. Qualitative comparison with state-of-the-art methods.Note that the results of X2Face suffer from image distortion when source-domain images are with extreme poses. Few-shot talking-head generates realistic images, but the identity of the reenacted images is not consistent with the target-domain one. Finally, our CrossID-GAN is able to transfer ID-independent motion information while preserving the identity of interest in the target domain. (More results and comparisons are available at supplementary materials.)

different people. Every frame of each video is annotated with 68 coordinate points representing 68 facial landmarks. We take the first 102 videos as training videos, with the rest serve as videos of unseen identities. We crop the face region of every video to shape 128×128. The facial landmarks are represented by a 68 channels heatmap, with every channel showing a unit Gaussian centered at its corresponding landmark location. Pairwise and unpaired frames are sampled from these videos to train our model. For pairwise data, we randomly sample 1000 pairs from every video in every epoch, which is 102000 pairs in total. For unpaired data, we randomly sample the same number as pairwise data (i.e. 102000 pairs) of different identity frame pairs from different videos.

4.2. Implementation Details

We implement our model with PyTorch [11]. For encoders, both E_L and E_C share the same structure, which consists of five convolutional layers and four residual blocks [4] to map input landmarks and image to latent code of dimension $8 \times 8 \times 256$. Our generator G contains five residual blocks followed by four transposed convolutional layers to decode the concatenation of latent codes to two intermediate products: a predicted mask M and a hallucinated output O. The final output image will be the linear combination of I and O using M, where $\tilde{v}_x^i = (1-M) \times I + M \times O$. For our discriminators, different architectures are applied. We adopt the multi-layer (five layers to be specific) discriminator from CycleGAN as D_I. D_C contains four convolutional layers followed by three fully-connected layers that map landmark code to $l \in R^n$ with n representing the number of identity in our training data. The feature extractor of AlexNet [9] is used to serve as the base model of Siamese network D_{id}, which is initialized with PyTorch pretrained weights. After extracting features from v_x^j and $\tilde{v}_x^i$, the subtraction of both features is forward to a fully-connected layer to predict the realism score. D_{tmp} contains four 3D-convolutional layers [7] and a fully-connected layer to predict the realism score of consecutive three inputs.
The pretrained landmark extractor F is trained following the state-of-the-art Pix2Pix [6] approach. We train F using our training dataset, with face image as input and landmarks heatmap as ground truth. The weights of F is fixed during the process of training our CrossID-GAN.

4.3. Qualitative Evaluation

We compare our method with two state-of-the-art works: X2Face [22] and Few-shot talking head [25] since *both methods utilize a unified network to output reenacted images* given a source-domain video and few target-domain images. We trained both models on our training dataset. For X2Face, since only the training code of their first stage is open source, we implement the second stage (stage with identity loss) by ourselves. For Few-shot talking head, since the official source code is not available, we use the code provided by the third party[1] which implemented the meta-learning stage of this work perfectly. Since both models can take multiple target-domain images as input, say K images, we set K to 3 to ensure that we demonstrate the full ability of these models.
Figure 3 shows the comparison results. While X2Face outputs sharp images, reenacted faces get distorted easily

[1]https://github.com/grey-eye/talking-heads

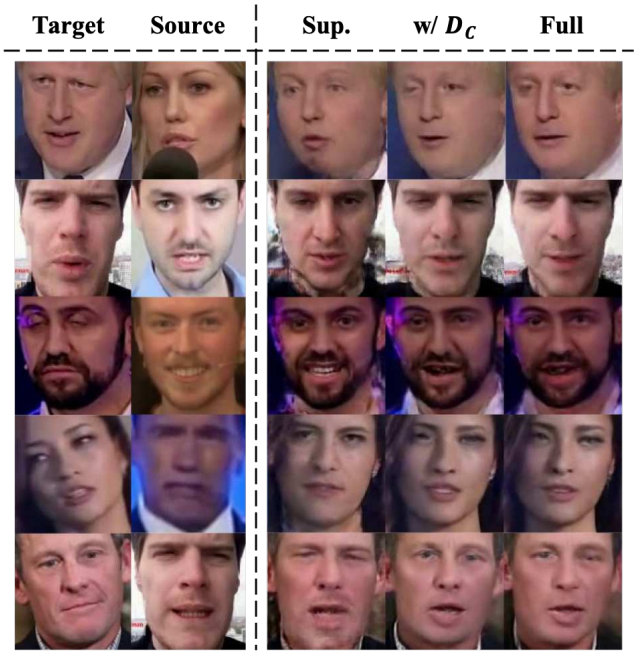

Figure 4. Ablation studies of our training strategy. Note that introduction of D_C allows learning of ID-invariant landmarks while preserving the target identity. The full version (trained with both supervised/unsupervised settings) produces the best results.

because of lacking clear guidance of pose (i.e., landmarks information). Also, it fails to give reasonable output if the source image is under extreme poses (column 3 and 4). This show the shortcoming of warping based methods which are not able to synthesis image under different illumination due to the pixels of output images are warped from input images.

On the other hand, with extra landmark information, the pose of output images from Few-shot talking head always corresponds to the source frame v_s^i . However, severe identity difference occurs between output images and I. This is mainly because they did not deal with the problem of cross-identity landmarks difference. Their model fails to preserve identity while doing face reenactment. The problem is especially obvious when doing face reenactment across genders.

The last row shows the results of our CrossID-GAN. Thanks to our unsupervised learning strategy, the pose and expression of our output images always correspond to the source frame while the identity is also well preserved.

4.4. Ablation Studies

In our ablation studies, we first show how D_C helps in landmark feature extraction. Then, we compare the results before and after the use of unpaired data.

As Fig. 4 shows, the target-domain images I and source-domain image v_s^i are listed in the first and second column

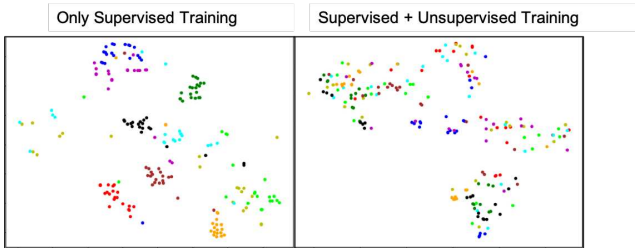

Figure 5. t-SNE visualization of landmark features z_l without/with our unsupervised learning strategy. Each color represents a different identity/video and each point represents a single frame. Note that with our unsupervised learning strategy described in Sect. 3.2, the derived z_l will be ID-invariant (not guided by person IDs).

respectively. The third column (**Sup.**) shows the results of our model after only supervised training. The shape of output face and mouth mainly depend on that of v_s^i, causing serious identity gap between I and output images. After adding D_C during training, as the fourth column (**w/ D_C**) shows, the contour of output faces is much more similar to I. This demonstrates the effectiveness of D_C to help E_L learning to extract ID-invariant code. However, while the distortion of reenacted images is much improved, missing details such as the difference in skin color or dissimilarity of nose and cheek still exist. The last column (**Full**) shows the results of our model after taking a fully supervised and unsupervised training process. We can tell that the problems are well tackled by the increase of identity similarity. This proves the effectiveness of our strategy of considering both pairwise and unpaired data during training.

4.5. Quantitative Evaluation

Besides qualitative comparisons, we further perform quantitative comparisons with X2Face and Few-shot talking head to evaluate the capability of identity preservation and the quality of synthesized images. For the evaluation regarding the capability of identity preservation, since there are no ground truth videos for reenacted outputs, it is intuitive to calculate the cosine similarity (CSIM) between embedding vectors of reenacted frames and the target-domain image, which are extracted through the state-of-the-art face recognition network [2]. However, since there are numerous pairs of source frames and target images, it is not practical for us to perform reenactment on all the combinations. As an alternative, we randomly sample 10 pairs of source-domain video and target-domain image and then calculate the mean CSIM score of all pairs. As Table 2 shows, our CrossID-GAN reaches the highest score of CSIM, which confirms that our model owns the best capability in preserving the identity of the target-domain image. To evaluate the quality of synthesized images, we follow the setting of Few-shot talking-head [25], which reconstructs the input video *when source-domain landmarks and target-*

Method	X2Face	Talking-Head	CrossID-GAN(Ours)
CSIM(↑)	0.646	0.607	**0.655**

Table 2. Quantitative comparisons results with state-of-the-art face reenactment methods in terms of visual/ID consistency between the target image and the produced video outputs.

Method	FID (↓)	SSIM (↑)	CSIM (↑)
X2Face	15.06	**0.83**	0.867
Talking-Head	22.95	0.80	**0.893**
Ours	**9.31**	0.80	0.883

Table 3. Quantitative comparisons with state-of-the-art methods of face reenactment in quality of synthesized images.

domain images are of the same identity. Three metrics, including FID [5], SSIM [21], and CSIM of features extracted from [2] between reconstructed images and the input (ground truth) frames are calculated. To be more specific, Frechet-inception distance (FID) is a measurement of similarity between two sets of images, which is often used to estimate the realism of generated images with real images by computing the Fréchet distance between the two Gaussian distributions fitted to feature representations of the Inception network. Structured similarity (SSIM) and cosine similarity (CSIM) aim to measure the low-level and the high-level identity preservation to the input (ground truth) frames respectively. As shown in Table 3, our model shows competitive results in synthesizing images. Despite these metric scores of different methods are close, our CrossID-GAN is capable of transferring motion across multiple identities well, while othes fail to present accurate face reenactment.

4.6. T-SNE Visualization of Encoded Landmarks Features

In this section, we demonstrate the capability of our E_L in extracting ID-invariant pose information. We apply t-SNE to visualize the distribution of embedded landmarks features z_l with/without our introduced unsupervised learning scheme, which confuses the ID labels in z_l with data recovery guarantees. We pick the first 10 videos from our dataset and randomly sample 20 frames from each video for visualization. As shown in Fig. 5, the left image shows the distribution of z_l with E_L only trained under the supervised setting. With each identity represented by a color, we observe that the distribution of z_l mainly depended on their identity, this indicates that the features encoded by E_L still contain identity information in this stage. On the other hand, the right image shows the distribution after our full training strategy. We can see that the encoded features are now mixed up and do not exhibit ID-oriented information. This confirms that our E_L extracts ID-invariant landmark information. With ID information is eliminated, pose information is preserved so that our model is able to perform face reenactment across multiple identities successfully.

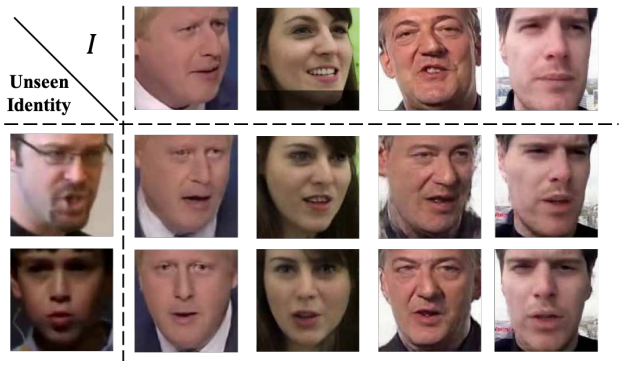

Figure 6. Examples of face reenactment with poses transferred from unseen identities. Note that methods like RecycleGAN cannot deal with source poses from unseen identity inputs.

4.7. Pose from Unseen Identity

Finally, we show that our CrossID-GAN owns the ability to generalize to unseen poses. Unseen pose means that the identity of source landmarks is not observed during training phase. Despite this circumstance, our CrossID-GAN is still able to perform face reenactment from these unseen poses. As shown in Fig. 6, the first column listed the input source-domain images, and the first row listed the target identity images. We can see that the motion from unseen poses is transferred accurately while the target identities are still well preserved. This shows that the range of source-domain is not limited to our training data but can be expanded to unseen identities, which makes our model more practical in real-world applications.

5. Conclusion

We proposed a unified deep learning model of CrossID-GAN, which extracts and transfers ID-invariant motion patterns across visual domains for multi-ID face reenactment. Different from most existing face reenactment methods, our model can be trained via supervised and unsupervised settings, while no ground truth landmarks are required. More importantly, our CrossID-GAN is designed to handle multiple ID in an unified framework. With qualitative and quantitative comparisons with state-of-the-art models, we confirmed the effectiveness and superiority of our method for cross-ID face reenactment.

Acknowledgement This work is supported in part by the Ministry of Science and Technology of Taiwan under grant MOST 109-2634-F-002-037.

References

[1] Aayush Bansal, Shugao Ma, Deva Ramanan, and Yaser Sheikh. Recycle-gan: Unsupervised video retargeting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 119–135, 2018. 1, 2, 3

[2] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 7, 8

[3] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 1

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[5] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017. 8

[6] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 6

[7] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2012. 5, 6

[8] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 4

[9] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 6

[10] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al. Deep face recognition. In *bmvc*, volume 1, page 6, 2015. 5

[11] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 6

[12] Fengchun Qiao, Naiming Yao, Zirui Jiao, Zhihao Li, Hui Chen, and Hongan Wang. Geometry-contrastive gan for facial expression transfer. *arXiv preprint arXiv:1802.01822*, 2018. 2, 3

[13] Enrique Sanchez and Michel Valstar. Triple consistency loss for pairing distributions in gan-based face synthesis. *arXiv preprint arXiv:1811.03492*, 2018. 2, 3

[14] Jie Shen, Stefanos Zafeiriou, Grigoris G Chrysos, Jean Kossaifi, Georgios Tzimiropoulos, and Maja Pantic. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 50–58, 2015. 5

[15] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4

[16] Justus Thies, Michael Zollhöfer, Matthias Nießner, Levi Valgaerts, Marc Stamminger, and Christian Theobalt. Real-time expression transfer for facial reenactment. *ACM Trans. Graph.*, 34(6):183–1, 2015. 1

[17] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2387–2395, 2016. 1, 2

[18] Yu Tian, Xi Peng, Long Zhao, Shaoting Zhang, and Dimitris N Metaxas. Cr-gan: learning complete representations for multi-view generation. *arXiv preprint arXiv:1806.11191*, 2018. 2, 3

[19] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1415–1424, 2017. 2, 3

[20] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1526–1535, 2018. 3

[21] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 8

[22] Olivia Wiles, A Sophia Koepke, and Andrew Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 670–686, 2018. 2, 3, 6

[23] Wayne Wu, Yunxuan Zhang, Cheng Li, Chen Qian, and Chen Change Loy. Reenactgan: Learning to reenact faces via boundary transfer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 603–619, 2018. 2, 3

[24] Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan. A light cnn for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, 13(11):2884–2896, 2018. 5

[25] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. *arXiv preprint arXiv:1905.08233*, 2019. 2, 3, 6, 7

[26] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 1, 2, 3