One-Shot Identity-Preserving Portrait Reenactment

Sitao Xiang[1], Yuming Gu[1] *, Pengda Xiang[1] *, Mingming He[1], Koki Nagano[1],
Haiwei Chen[1], and Hao Li[1,2]

[1] University of Southern California / USC Institute for Creative Technologies
[2] Pinscreen

Fig. 1. Semantic image synthesis results produced by our method. Our method can not only synthesize images using only semantic segmentation masks as input, but also supports controllable synthesis via a reference style image.

Abstract. We present a deep learning-based framework for portrait reenactment from a single picture of a target (one-shot) and a video of a driving subject. Existing facial reenactment methods suffer from identity mismatch and produce inconsistent identities when a target and a driving subject are different (cross-subject), especially in one-shot settings. In this work, we aim to address identity preservation in cross-subject portrait reenactment from a single picture. We introduce a novel technique that can disentangle identity from expressions and poses, allowing identity preserving portrait reenactment even when the driver's identity is very different from that of the target. This is achieved by a novel landmark disentanglement network (LD-Net), which predicts personalized facial landmarks that combine the identity of the target with expressions and poses from a different subject. To handle portrait reenactment from unseen subjects, we also introduce a feature dictionary-based generative adversarial network (FD-GAN), which locally translates 2D landmarks into a personalized portrait, enabling one-shot portrait reenactment under large pose and expression variations. We validate the effectiveness of our identity disentangling capabilities via an extensive ablation study, and our method produces consistent identities for cross-subject portrait

---

* indicates joint second authors.

reenactment. Our comprehensive experiments show that our method significantly outperforms the state-of-the-art single-image facial reenactment methods. We will release our code and models for academic use.

Keywords: One-shot, landmark disentanglement, cross-subject, portrait reenactment, identity-preserving, generative adversarial network.

1   Introduction

Synthesizing facial animation from portrait images has a wide range of creative applications, including visual effects, multimedia messaging apps, and visual dubbing. Using someone's facial expressions and head pose in videos to drive the face of another person in a single image (known as portrait reenactment) is particularly popular due to its intuitive control and accessibility.

Traditionally, producing a realistically animated face is achieved by rendering a carefully digitized 3D head model with texture maps. Although many digital humans are still being produced this way for high-end visual effects and video games [50], this approach involves a tedious production effort including large teams of digital artists and often relies on complex 3D capture equipment [23].

More recently, deep learning-based methods have gained significant attention due to their success in producing realistic face reenactment. In particular, Deep-Fakes (*e.g.* [15]) have become a widely used approach for end-to-end video-based face-swapping. However, it cannot generalize well to unseen subjects, requiring thousands of internet photos and often days of training for each subject.

Currently, some advanced deep learning techniques for face image manipulation combine conditional GANs [31] with facial geometry information, such as 3D facial models [40,35,34] or 2D landmarks [64,42], to provide both better control and generalization capabilities w.r.t. arbitrary identities. These 3D model-based methods only work under specific conditions. They mostly rely on statistical face models and are often limited to certain face regions and linear shape variations. Furthermore, they lose accuracy for non-frontal portraits. However, our goal is to synthesize highly complex head poses, facial expressions and facial appearances (facial hair, stylized content, complex lighting conditions) as well as the image regions surrounding the face such as hair, background, etc. On the other hand, the 2D landmark-based methods are unable to properly preserve the accurate identity and complex facial expressions with the lack of appropriate landmark adaptation.

In this work, we wish to achieve a one-shot portrait reenactment of novel subjects (with no subject-specific training) for the cross-subject setting (meaning the ability to accommodate any driver). In particular, we aim to address the problem of portrait reenactment from a single image of someone (the "target") using a sequence of 2D facial landmarks from a video of another person (the "source").

Our goal is to improve the preservation of identities within a cross-subject setting. Several challenges need to be addressed for identity-preserving face reenactment, particularly when a target and a source are different subjects. First,

it is a non-trivial task to properly extract facial expressions/poses from person-specific facial features encoded using 2D facial landmark coordinates. *For example, how can we determine whether a person has narrow eyes or is squinting and thus properly transfer identity-invariant motion to the target?* The second challenge lies in synthesizing photorealistic and recognizable results with arbitrary expression and novel views. The third challenge is to achieve the above from a single reference image of the target subject without relying on any person-specific training or fine-tuning [64].

We achieve this by introducing two novel sub-networks. The first, *Landmark Disentanglement Network* (LD-Net), learns to disentangle the identity from the head poses and expressions, predicting facial landmarks that preserve the identity of the target while combining expressions and poses from another driving subject. The second, *Feature Dictionary-based Generative Adversarial Network* (FD-GAN), learns to transform the landmark positions into a personalized video portrait of the subject depicted in a single target image, which allows subject-agnostic reenactment of a portrait that preserves the target's identity and can be applied to unseen subjects without subject-specific training. To summarize, we make the following technical contributions: (1) We introduce a novel one-shot learning method that enables portrait reenactment using the identity from a single image and expressions and poses from videos of another subject. (2) We present a novel network that disentangles the identity from 2D face landmarks for cross-subject portrait reenactment. (3) We also propose a feature dictionary-based generative network to synthesize a high-fidelity face image, which is applicable to new subjects. We evaluate each sub-network as well as the full method extensively via quantitative measurements and qualitative comparisons with the state-of-the-art methods, and demonstrate our method's ability to preserve the target subject's identity and to generalize to unseen subjects for cross-subject face reenactment.

2   Related Work

3D Graphics-based Methods. Research in 3D facial animation and rendering dates back several decades. The seminal work of [5] showed that a morphable principal component model is effective in modeling human facial geometry and appearance. Over the years, a number of technical advances [17] have been made in capturing high-resolution textures and geometric details, as well as subject-specific expression deformations that are necessary to create realistic renderings of human face animations. To capture facial textures, recent work uses deep learning-based analysis to infer photorealistic skin albedo maps from a single image [48,20]. Olszewski et al. [44] showed that realistic face puppeteering is possible from a single picture by rendering a sequence of dynamic textures that are synthesized using generative adversarial networks (GANs). For face geometry, other recent approaches employ a non-linear morphable model [57] to improve the fidelity, use a local regressor to enhance high-resolution details [7],

or use a pixel-to-pixel translation network to learn mesoscopic geometry [30] or comprehensive skin reflectance [63,49].

The reconstructed 3D mesh can be used to re-render animated faces using single-view face tracking for video dubbing [14,18], realistic facial reenactment [54,56], facial replacement [43], or lip-syncing [53]. Previous work also explored the techniques for animating a photorealistic avatar by building person-specific skin deformation and texture models from RGB-D [11] or RGB scans [9]. In 3D-based methods, 3D face modeling and retargeting from a single-view input [6,8,19,27,28,38,47,61,37] can be performed to properly decompose person-specific facial shapes from expressions and pose. Alternatively, 2D facial landmarks and image-space detail transfer can be used to animate still portraits [2].

Deep Learning-based Methods. For still portrait synthesis, GANs [24] have been extensively studied for synthesizing a high-resolution human face, including person-specific details such as pores and facial hair [32,33]. A conditional GAN [31] has been introduced for pixel-to-pixel translation applications to manipulate a human face image from edge drawings [60] or image sequences [59]. For facial expression editing, Choi et al. [12] extended previous work to multi-domain pixel translations, enabling facial expression editing using discrete expression labels. Pumarola et al. [45] showed that a conditional GAN with a cycle consistency loss [65] can be used for unsupervised learning of continuous facial animation editing from a single image. However, the controls provided are too coarse to capture the fine-scale nature of human facial expressions. For face swapping applications, "DeepFake" frameworks (*e.g.* [15]) employ an encoder and decoder architecture to achieve a video-based face swapping of a pair of subjects. However, the framework cannot handle arbitrary pairs of subjects without additional subject-specific training. For many-to-many subject face swapping, Bao et al. [4] proposed a GAN-based training framework to decompose facial identity from other attributes such as expression, pose and illumination, allowing an end-to-end face swapping for unseen subjects.

An alternative approach for decomposing facial identity from other attributes is to explicitly model it as facial geometry such as 2D landmarks or 3D face models. For 2D geometry-guided methods, Natsume et al. [41] proposed a framework to achieve single-image face swapping between unseen identities conditioned on 2D landmarks. Nirkin et al. [42] proposed a recursive approach for improved identity preservation for a subject-agnostic face image synthesis. Siarohin et al. [51] introduced a first order motion model which can animate an image of a variety of categories via keypoints and local affine transformations including a human face portrait. However, it is challenging to properly separate person-specific identity, facial expressions and pose from coarse 2D landmarks (typically 68 fiducial points), and thus the previous work can still suffer from noticeable artifacts and identity mismatches. Zakharov et al. [64] relaxed the requirement for one-shot learning, thereby showing that few-shot learning could be employed to improve the identity preservation for portrait reenactment. However, it cannot adapt the landmarks when the source and target subjects are different and does not address cross-subject face reenactment. Wang et al. [58] proposed a
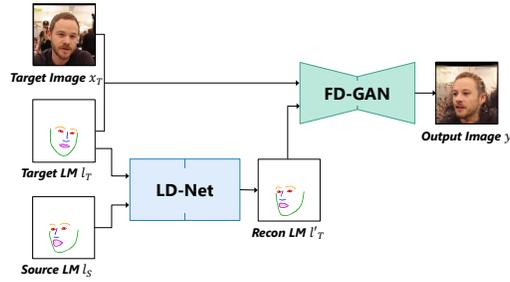
Fig. 2. Overview of our method, which consists of two sub-networks: LD-Net and FD-GAN. Given the landmarks of a pair of face images l_T and l_S, LD-Net first generates a landmark l'_T combining the target's identity with the source's pose/expression. Then, taking l_T, l'_T and the target image as input, FD-GAN generates a new face image y.

few-shot framework for general video-to-video translation applications and applied it to animating portraits. Unlike any of the above methods, our method only requires a single image of the target, it does not require subject-specific training, and it can handle cross-subject reenactment of unseen subjects (*i.e.*, it is subject-agnostic).

For 3D geometry-based methods, Kim et al. proposed a hybrid approach combining 3D morphable models and an image translation network to translate 3D rendering of the target face to a synthetic video for video portrait reenactment [35] or visual dubbing [34] between pairs of subjects. Nagano et al. [40] proposed a generalized solution that can synthesize arbitrary expressions of an unseen identity from a single picture, but only operates in the face region. Previous work [21,22] also addressed identity-agnostic face image synthesis using 3D face fitting and deep neural nets, but also addressed full portrait manipulation including the background using background warping [21] or blending [22].

3   Method

Our goal is to transfer the head pose and facial expression from a source video of one person to a target image of another subject while preserving the target's identity. Based on the observation that 2D facial landmarks contain information about the pose and expression as well as person-specific identity features (*e.g.* the size, shape, proportion, and layout of the facial features), we propose to disentangle the identity and pose/expression from the landmarks and use them for landmark synthesis. As shown in Fig. 2, our method consists of two sub-networks. The *Landmark Disentanglement Network* (LD-Net) first synthesizes new landmarks with the target's identity and the source's pose/expression. Then the *Feature Dictionary-based Generative Adversarial Network* (FD-GAN) takes both target and synthetic landmarks as input and translates them into a new face image.
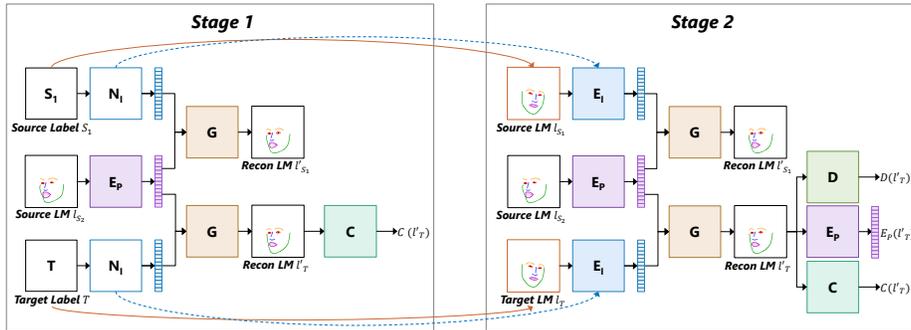
Fig. 3. The training procedure of LD-Net, consisting of two stages.

3.1   Landmark Disentanglement Network (LD-Net)

Disentangling landmarks into identity and pose/expression is difficult due to the lack of accurate numerical labeling for pose/expression. Inspired by [1], which can disentangle two complementary factors of variations with only one of them labeled, we propose a landmark disentanglement network (LD-Net) to disentangle identity and pose/expression using data with only the subject's identity labeled. More importantly, our network generalizes well to novel identities (*i.e.*, those unseen during training), unlike previous works (*e.g.* [1]).

Given 2D facial landmarks from a pair of face images, LD-Net first disentangles the landmarks into a pose/expression latent code and an identity latent code, then combines the target's identity code with the source's pose/expression code to synthesize new landmarks. As shown in Fig. 3, the training procedure of LD-Net is divided into two stages. *Stage 1* aims to train a stable pose/expression encoder and *Stage 2* generalizes to predict an identity code from landmarks instead of using identity labels so as to handle unseen identities.

Stage 1. In *Stage 1*, similar to [1], the network consists of four modules: (1) a pose/expression encoder E_P that computes a code from the input landmarks l that encodes only pose/expression without information about identity; (2) a one-layer network N_I that maps the input one-hot identity label k to an identity code; (3) a generator G that combines a pose/expression code and an identity code to reconstruct landmarks $l' = G(E_P(l), N_I(k))$; and (4) a classifier C that tries to classify the generated landmarks based on their identity.

As shown in Fig. 3, *Stage 1* is trained with two branches for each iteration, which share the same generator G but are associated with their own input and output. In the first branch, the input $k = S_1$ (identity label) and $l = l_{S_2}$ (landmark locations) are from the same subject so the reconstructed landmarks $l' = l'_{S_1}$ should be as same as the input l, which can be used to define a reconstruction loss. In the second branch, the input $k = T$ and $l = l_{S_2}$ are from different subjects and the reconstructed landmarks $l' = l'_T$ should contain no information about the identity of S_1.

To achieve this, the classifier C tries to classify $l' = l'_T$ as being a landmark of S_1 while the pose/expression encoder E_P, generator G and identity encoder N_I tries to prevent the classifier from doing so. C is trained with the classification loss of the form:

$$\mathcal{L}_C = \mathbb{E}[-\log P(S_1|l')]. \tag{1}$$

Meanwhile, E_P, G and N_I jointly optimize the reconstruction loss minus the identity classification loss as in Eq. 2. The reconstruction loss is defined as per-point squared Euclidean distance using landmark coordinates.

$$\mathcal{L}_G = \lambda_{\rm rec}\mathbb{E}[||l - l'||_2^2] + \lambda_C\mathbb{E}[\log P(S_1|l')], \tag{2}$$

where $\lambda_{\rm rec} = 1000$ and $\lambda_C = 0.1$.

All expectations are taken over $l \sim p(l), k \sim p(k)$ where $p(l)$ and $p(k)$ are training distributions of landmarks and identities where l and k may be from different subjects. Different from the network architecture in [1], all convolutional networks are replaced with Multi-layer Perceptrons (MLP) in LD-Net, since instead of images we operate on landmark coordinates.

Stage 2. To generalize to novel subjects, we introduce an identity encoder to *Stage 2*. A notable deficiency of [1] is that it does not include any encoder for the labeled data (i.e., identity in our task) and thus it is limited to generating new samples only for the labeled classes in the training data. In *Stage 2*, we replace the one-layer network N_I with a full-fledged identity encoder E_I that accepts landmarks as input and encodes them into an identity code.

The full training network of *Stage 2* is shown in Fig. 3, also involving two branches similar to *Stage 1*, that is, l_{S_1} and l_{S_2} are from the same subject while l_{S_2} and l_T belong to different subjects.

For the second input l_{S_2} and l_T in the second branch, due to unavailability of ground truth, we train a discriminator D and a classifier C to constrain the reconstructed landmarks l'_T. We use least square loss for the discriminator D following [39] to minimize:

$$\mathcal{L}_D = \mathbb{E}[(D(l_{S_2}) - 1)^2 + (D(l'_T) + 1)^2]. \tag{3}$$

The classifier C is trained with an adversarial loss on both input landmarks l_T and generated landmarks:

$$\mathcal{L}_C = \mathbb{E}[-\log P(k|C(l_T))] + \mathbb{E}[-\log(1 - P(k|C(l'_T))], \tag{4}$$

with expectations taken over $l_{S_2}, l_T \sim p(l)$, $k \sim p(k)$. k is the identity label of l_T. In addition, a content consistency loss term is defined between the generated pose/expression code and its ground truth $E_P(l_{S_2})$:

$$\mathcal{L}_{\rm cont} = \mathbb{E}[||E_P(l_{S_2}) - E_P(l'_T)||_2^2]. \tag{5}$$

For the first input l_{S_1} and l_{S_2}, the reconstructed landmarks l'_{S_2} should have the same pose/expression and identity as l_{S_2}. Thus, a reconstruction loss is defined to discourage E_I to encode pose/expression:

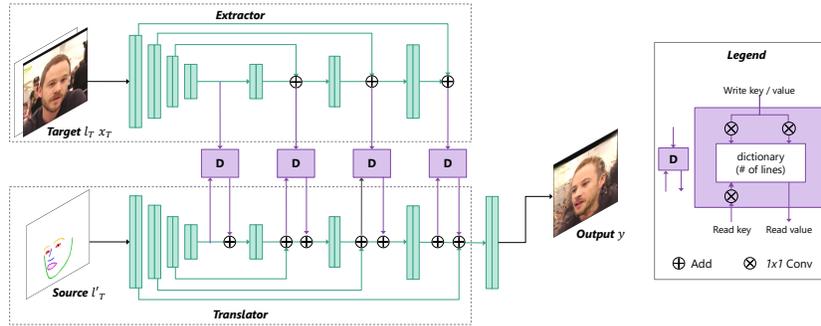$$\mathcal{L}_{\rm rec} = \mathbb{E}[||l_{S_2} - l'_{S_2}||_2^2], \tag{6}$$

Fig. 4. The network architecture of the second sub-network FD-GAN.

with the expectation over l_{S_1}, $l_{S_2} \sim p(l|k)$, and $k \sim p(k)$. l_1 and l_2 are landmarks from the same subject k.

Thus, the identity encoder E_I and generator G are jointly optimized to minimize a weighted sum of the above losses:

$$
\begin{aligned}
\mathcal{L}_G = {} & \lambda_{\text{rec}} \mathcal{L}_{\text{rec}} \\
& + \lambda_{\text{cont}} \mathbb{E}[\|E(l_{S_2}) - E(l'_{S_1})\|_2^2] \\
& + \lambda_D \mathbb{E}[D(l'_T)^2] \\
& + \lambda_C \mathbb{E}[-\log P(k|C(l'_T))],
\end{aligned}
\tag{7}
$$

where $\lambda_{\text{rec}} = 1000$, $\lambda_{\text{cont}} = 0.01$, $\lambda_D = 0.1$, and $\lambda_C = 0.1$.

3.2 Feature Dictionary-based Generative Adversarial Network (FD-GAN)

With the predicted landmarks rasterized into a landmark image, our next goal is to translate it to a photorealistic face image. We can think of the translation procedure as follows: a local patch around each location in the landmark image indicates "which facial part should be here", and for each location, we want to translate this into "how it should appear". Thus, we propose a novel feature dictionary-based generative adversarial network (FD-GAN) to achieve these intuitive objectives.

The architecture of FD-GAN is illustrated in Fig. 4, which consists of an extractor and a translator. Given a target image x_T and its corresponding landmark image l_T, we train an extractor that constructs a "feature dictionary" in the module D, which is essentially a mapping from an annotation in the landmark image to its appearance in the target image. Concurrently, given another landmark image l'_T and the feature dictionary, we train a translator that retrieves relevant facial features from the dictionary based on the landmarks and composes a face image.

The dictionary mapping is realized with a mechanism similar to the memory bank in the Neural Turing Machine (NTM) [25]: the feature dictionary is a

memory matrix, with each memory row conceptually corresponding to some facial component and the value stored in that row corresponding to how that component should appear on a specific subject's face. The construction of such a feature dictionary corresponds to the write operation in NTM and the lookup step during translation corresponds to the read operation in NTM. Precisely, a feature dictionary consists of n rows, and each row i is associated with a write tag $\mathbf{t}^{(i)}$ and read tag $\mathbf{u}^{(i)}$, both of which are vectors of length m_T and are learnable parameters of the network, and a stored value $\mathbf{v}^{(i)}$, which is a vector of length m_V computed by the network.

During the writing phase in the extractor, the stored values are computed as follows: in the form of two convolutional feature maps, the extractor generates for each spatial location j a write key $\mathbf{p}^{(j)}$ and a write value $\mathbf{x}^{(j)}$. Then, the value of $\mathbf{v}^{(i)}$ is computed as:

$$\mathbf{v}^{(i)} = \frac{\sum_{n \in N} \mathbf{x}^{(n)} \cdot e^{(\mathbf{t}^{(i)} \cdot \mathbf{p}^{(n)})}}{\sum_{n \in N} e^{(\mathbf{t}^{(i)} \cdot \mathbf{p}^{(n)})}}, \tag{8}$$

where N is the set of spatial locations. That is, the value of row i in the feature dictionary is a weighted sum of the write values at each location, with weight being the softmax of $\mathbf{t}^{(i)} \cdot \mathbf{p}^{(n)}$ for location n.

Similarly, during the reading phase, the translator, as a convolutional feature map, generates for each spatial location j a read key $\mathbf{q}^{(j)}$. The return value $\mathbf{y}^{(j)}$ for each location of lookup operation is computed as:

$$\mathbf{y}^{(j)} = \frac{\sum_{k=1}^{n} \mathbf{v}^{(k)} \cdot e^{\mathbf{u}^{(k)} \cdot \mathbf{q}^{(j)}}}{\sum_{k=1}^{n} e^{\mathbf{u}^{(k)} \cdot \mathbf{q}^{(j)}}}, \tag{9}$$

which means the value read by each location j is a weighted sum of all the rows in the feature dictionary, with weight being the softmax of $\mathbf{u}^{(k)} \cdot \mathbf{q}^{(j)}$ for each row k. The translator then continues network operations on this returned convolutional feature map.

The extractor and translator are both fully convolutional, with U-Net skip connections as shown in Fig. 4. To train such a joint extractor-translator, we employ a combination of reconstruction loss, GAN loss, and an adversarial classifier loss. The discriminator and classifier are both patch-based with their loss averaged across spatial locations. In the following equations, T is the extractor-translator, and to avoid excessive notation, we use the same letters D and C for the discriminator and classifier as in Sec. 3.1. For simplicity, we omit the range over which the expectations are taken: x_T and x'_T are two frames from the same video clip, l_T and l'_T are their respective landmark images, and k is the identity label of x_T and x'_T.

The discriminator D minimizes:

$$\mathcal{L}_D = \mathbb{E}[(D(x_T) - 1)^2 + (D(T(x_T, l_T, l'_T)) + 1)^2]. \tag{10}$$

The classifier C minimizes:

$$\mathcal{L}_C = \mathbb{E}[-\log P(k|C(x_T))] \tag{11}$$
$$+ \mathbb{E}[-\log(1 - P(k|C(T(x_T, l_T, l'_T))))].$$

The loss of extractor-translator in FD-GAN is a weighted sum of adversarial discriminator loss, adversarial classifier loss and reconstruction:

$$\mathcal{L}_T = \lambda_{\mathrm{rec}}\mathbb{E}[||T(x_T, l_T, l'_T) - x'_T||_2^2] \tag{12}$$
$$+ \lambda_D \mathbb{E}[D(T(x_T, l_T, l'_T))^2]$$
$$+ \lambda_C \mathbb{E}[-\log P(k|C(T(x_T, l_T, l'_T)))],$$

where $\lambda_{\mathrm{rec}} = 50$, $\lambda_D = 1$, and $\lambda_C = 1$.

4   Experiments

We first conduct an evaluation and ablation study in Sec. 4.1 on the performance of LD-Net and FD-GAN independently, followed by comparisons of our full method with the state-of-the-art methods on cross-subject face reenactment in Sec. 4.2. For more results tested on unconstrained portrait images, please refer to the supplemental material.

Implementation details. For FD-GAN, the extractor and translator are based on U-Nets, with both networks joined together by dictionary writer/reader modules inserted into the up-convolution modules. The discriminator and classifier for FD-GAN are patch-based and have the same structure as the down-convolution part of the U-Nets. Please refer to the supplemental material for more details concerning the network structures and training strategies.

Performance. Our method takes approximately 0.08s for FD-GAN to generate one image and 0.02s for LD-Net to perform landmark disentanglement on a single NVIDIA TITAN X GPU.

Training datasets. The training dataset is built from VoxCeleb video training data [13] which is processed by dlib [36] to crop a 256×256 face image at 25fps and to extract its landmarks. In total, it contains 52,112 videos for 1,000 randomly selected subjects.

Testing datasets. We use three datasets to evaluate our method:

- LMTest: a landmark dataset which has 200,000 landmarks (100 subjects × 2000 frames of varying poses and expressions) with ground truth labels for both identity and poses/expressions. Using a video of one person performing and the first 100 neutral expression photos from the Compound Facial Expressions Database [16], we used single-view 3D face fitting [56] to retarget facial expressions and poses from the video subject to each subject's 3D face model and project 3D vertex positions to obtain ground truth 2D landmarks. This dataset is used to evaluate the effect of LD-Net.

- SelfTest: a video dataset of 8,000 frames for 80 subjects from Voxceleb testing data (100 frames per video at 25fps). It is only used for the ablation study when testing self-reenactment, where the ground truth is known.
- CrossTest: a video dataset of 8,000 frames for 80 pairwise subjects (100 frames per video at 25fps) randomly sampled from the Voxceleb testing data, used to compare our method with the baselines in one-shot cross-subject face reenactment.

Metrics. We use the following metrics for quantitative evaluation of generated images.

- Identity Similarity (ISIM): computes cosine similarity between embedding vectors of the face recognition network VGGFace2 [10] for identity matching.
- Pose Similarity (PSIM): computes cosine similarity between head rotation in radians around the X, Y, and Z axes estimated by OpenFace [3].
- Expression Distance (ED): computes L2 distance of intensities of corresponding facial action units detected by OpenFace [3] between the generated images and the driving images.
- Fréchet-Inception Distance (FID) [26]: measures the distance between the distributions of real data and generated data to quantify the result fidelity.
- Structured Similarity (SSIM): measures low-level similarity to ground truth images in the self-reenactment setting.

4.1   Evaluation

Evaluation of LD-Net. To validate the accuracy in disentangling identity and pose/expression, we test the LD-Net in isolation using the LMTest dataset. From the 200,000 landmarks, we sample pairs of landmarks in 3 different patterns: the same identity but different pose/expression, the same pose/expression but different identity, and both differing identity and pose/expression. In each case, we randomly sample one million pairs of landmarks and compute their distances in the latent space of identity encoder E_I and pose/expression encoder E_P respectively. We first use PCA to reduce the dimensionality of E_I's and E_P's latent codes to 8 before computing the mean Euclidean distance for each pair. For E_I's latent space, pairs of landmarks from the same subject should give a smaller mean distance than pairs from different identities, and similarly for landmarks with the same pose/expression in the latent space of E_P. Table 1 gives the mean distances for each case, which shows that the identity code and pose/expression code do control the respective aspect of the generated landmarks, no matter what kind of input is provided.

Ablation analysis of LD-Net. In Table 2, We show the effect of LD-Net on the generated images in terms of identity, expression and pose preservation in two settings: self-reenactment and cross-subject. We use three metrics: ISIM, PSIM, and ED to measure matching accuracy of identity, pose and expression, respectively. In self-reenactment, we compare results generated using ground truth

Sample by	In E_I's space ↓	In E_P's space ↓
Same identity	**2.0431**	3.5171
Same pose/exp	3.1793	**1.2730**
Both different	3.8274	3.6744

Table 1. Mean Euclidean distance in the latent spaces of the identity encoder E_I and the pose/expression encoder E_P in different cases.

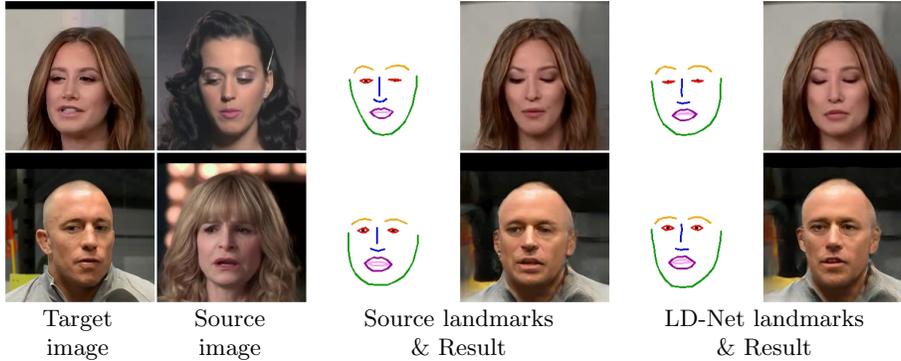|			
Target image	Source image	Source landmarks & Result	LD-Net landmarks & Result

Fig. 5. Qualitative comparison between the results generated using source landmarks and synthetic landmarks by LD-Net. Synthesized landmarks better preserve the target identity.

landmarks and synthetic landmarks by LD-Net. In cross-subject reenactment, we do a similar comparison between the results using the source subject landmarks as-is and LD-Net landmarks. Since PSIM and ED compare the poses/expressions with the source images, the results using the source landmarks (ground truth) always lead to better matching accuracy. However, the landmarks generated by LD-Net are very close to the ground truth landmarks when comparing in self-reenactment. Moreover, the cross-subject setting shows the importance of predicting personalized landmarks with higher identity accuracy in the cross-subject reenactment in Table. 2 and better visual quality in Fig. 5.

	Self-reenactment				Cross-subject		
LM from	ISIM ↑	PSIM ↑	ED ↓	LM from	ISIM ↑	PSIM ↑	ED ↓
Ground truth	**0.7986**	**0.9134**	**0.1296**	Source	0.7145	**0.8615**	**0.2080**
LD-Net	0.7984	0.8950	0.1655	LD-Net	**0.7726**	0.8398	0.2430

Table 2. Quantitative comparison between the results generated using ground truth landmarks vs landmarks predicted by LD-Net in self-reenactment, and using landmarks from the source subject as-is vs landmarks by LD-Net in cross-subject reenactment.

Ablation analysis of FD-GAN. We construct three baselines to evaluate the performance of FD-GAN: Pix2PixHD [60], FD-GAN-1, and AdaIN. The first, Pix2PixHD [60], is an advanced image-to-image translation network which can synthesize photo-realistic images from landmark images. In FD-GAN-1, we re-

duce the number of rows in the feature dictionary to 1 with its value being the mean of the convolution features. The final baseline, AdaIN, also uses a one-row dictionary but uses the written value to generate parameters for adaptive instance normalization (AdaIN) [29].

We compare FD-GAN with the three baselines in both the self-reenactment and cross-subject settings. To evaluate FD-GAN alone, we utilize two important image quality metrics, SSIM (unavailable in cross-subject reenactment) and FD, in addition to ISIM. As shown in Table. 3, the quantitative comparison in both settings demonstrates that our FD-GAN best preserves low-level image features, image fidelity, and identity information. Since our generative network learns a local mapping between the target image and landmarks to the final image, it is flexible enough to generalize to unseen subjects. But existing image-to-image approaches such as Pix2PixHD [60] lack the domain generalization capability needed to synthesize unseen subjects without any subject-specific learning.

Method	Cross-subject			Self-reectment		
	ISIM ↑	SSIM ↑	FID ↓	ISIM ↑	SSIM ↑	FID ↓
Pix2PixHD [60]	0.514	N/A	99.34	0.41	0.49	98.36
AdaIN	0.650	N/A	86.69	0.51	0.56	90.70
FD-GAN-1	0.6232	N/A	71.11	0.49	0.56	70.34
FD-GAN (Ours)	**0.7726**	N/A	**67.68**	**0.63**	**0.63**	**55.19**

Table 3. Quantitative comparison of FD-GAN with three baselines in both self-reenactment and cross-subject reenactment.

4.2   Comparison

Comparison with one-shot methods. We first show quantitative comparisons with two state-of-the-art one-shot face reenactment baselines, X2Face [62], and First-order-model [51], using their pre-trained models on the Voxceleb training dataset. We evaluate the models in the same setting without any fine-tuning on the CrossTest dataset. Both X2Face and First-order-model are warping-based methods which can well generalize to unseen subjects in the one-shot setting. In X2Face, the generated frame inherits the object proportions of the driving source video, and the quality of their results is very sensitive to the cropping region and face alignment as shown in Fig. 6 (we also test the algorithm with a different crop size in the supplemental material). From the quantitative comparison in Table 4 and the qualitative comparison in Fig. 6, we can see that the results from the First-order-model demonstrate the best image fidelity, since it uses a warping formulation to generate the deformed faces. However, its warping formulation, which is based on keypoints and local affine transformations, can hardly provide as accurate local control as our synthesized landmarks which better preserve the source pose/expression, especially when handling very different head poses and facial expressions. Therefore, compared to these methods, our model can generalize to unseen subjects with better identity preservation and more consistent quality under a large variety of poses/expressions.

Method	ISIM ↑	PSIM ↑	ED ↓	FID ↓
X2face [62]	0.6347	0.302	0.448	101.72
First-order-model [52]	0.7699	0.822	0.274	**55.94**
Ours	**0.7762**	**0.840**	**0.243**	67.68

Table 4. Quantitative comparison of methods for cross-subject reenactment on the CrossTest dataset between our method and [62], [60], and [52].

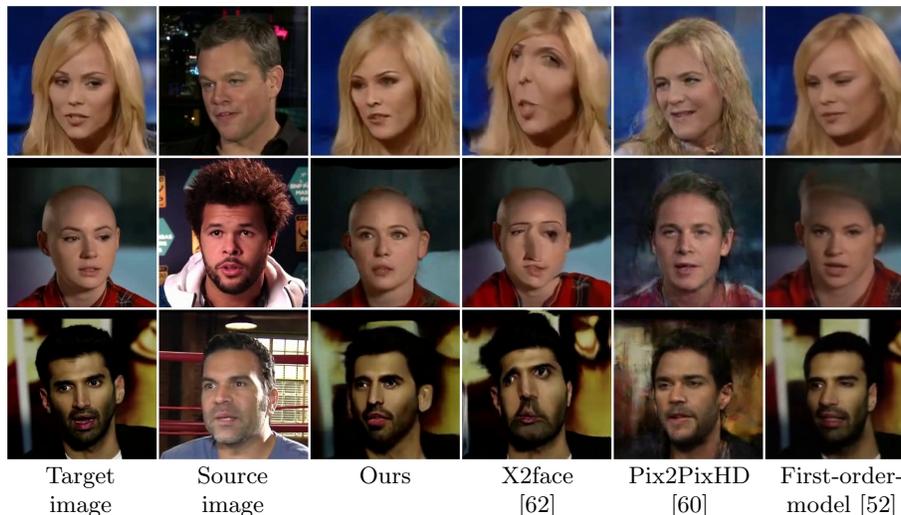| Target image | Source image | Ours | X2face [62] | Pix2PixHD [60] | First-order-model [52] |

Fig. 6. Qualitative comparison between our method and the baselines in cross-subject reenactment: X2face [62], Pix2PixHD [60], and First-order-model [52].

Qualitative comparison with 3D-based methods. Fig. 7 shows qualitative comparisons on the FaceForensics++ test dataset [46] with two state-of-the-art 3D-based methods (Face2Face [56] and NeuralTexture [55]). Compared to their methods which require 3D face fitting to maintain the target identity and cannot change head poses, our method can synthesize personalized faces with arbitrary head poses using only 2D landmarks.

5   Discussion and Future Work

We have demonstrated a technique for portrait reenactment that only requires a single target picture and 2D landmarks of the target and the driver. The resulting portrait is not only photorealistic but also preserves recognizable facial features of the target. Our comparison shows significantly improved results compared to state-of-the-art single-image portrait manipulation methods. Our extensive evaluations confirm that identity disentanglement of 2D landmarks is effective in preserving the identity when synthesizing a reenacted face. We have shown that our method can handle a wide variety of challenging facial expressions and poses of unseen identities without subject-specific training. This is made possible

| Target image | Source image | Ours | Face2Face [56] | Neural-Texture [55] |

Fig. 7. Qualitative comparison between our method and two 3D-based methods: Face2Face [56] and NeuralTexture [55].

thanks to our generator, which uses a feature dictionary to translate landmark features into a photorealistic portrait.

A limitation of our method is that the resulting portrait has only a resolution of 256×256, and it is still difficult to capture high-resolution person-specific details such as stubble hair. It could also suffer from some artifacts for non-facial parts and the background region, since we rely on the landmarks to transfer facial appearance but the landmarks contain no structural information about the hair or background. We believe such a limitation could be further addressed by incorporating dense pixel-wise conditioning [40] and segmentation. While our method can produce reasonably stable portrait reenactment results from a frame of target and 2D landmarks, the temporal consistency could be further improved by taking into account temporal information from the entire video.

Acknowledgment

References

1. Anonymous: Disentangling style and content in anime illustrations. In: Submitted to International Conference on Learning Representations (2020), under review. `https://openreview.net/forum?id=BJe4V1HFPr`
2. Averbuch-Elor, H., Cohen-Or, D., Kopf, J., Cohen, M.F.: Bringing portraits to life. ACM Trans. Graph. **36**(4), to appear (2017)
3. 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)
4. Bao, J., Chen, D., Wen, F., Li, H., Hua, G.: Towards open-set identity preserving face synthesis. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
5. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques. pp. 187–194. SIGGRAPH '99 (1999)
6. Bouaziz, S., Wang, Y., Pauly, M.: Online modeling for realtime facial animation. ACM Trans. Graph. **32**(4), 40:1–40:10 (Jul 2013)
7. Cao, C., Bradley, D., Zhou, K., Beeler, T.: Real-time high-fidelity facial performance capture. ACM Trans. Graph. **34**(4), 46 (2015)
8. Cao, C., Weng, Y., Zhou, S., Tong, Y., Zhou, K.: Facewarehouse: A 3d facial expression database for visual computing. IEEE TVCG **20**(3), 413–425 (2014)
9. Cao, C., Wu, H., Weng, Y., Shao, T., Zhou, K.: Real-time facial animation with image-based dynamic avatars. ACM Trans. Graph. **35**(4), 126 (2016)
10. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). pp. 67–74. IEEE (2018)
11. Casas, D., Feng, A., Alexander, O., Fyffe, G., Debevec, P., Ichikari, R., Li, H., Olszewski, K., Suma, E., Shapiro, A.: Rapid photorealistic blendshape modeling from rgb-d sensors. In: Proceedings of the 29th International Conference on Computer Animation and Social Agents. pp. 121–129. ACM (2016)
12. 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: IEEE CVPR (June 2018)
13. Chung, J.S., Nagrani, A., Zisserman, A.: Voxceleb2: Deep speaker recognition. arXiv preprint arXiv:1806.05622 (2018)
14. Dale, K., Sunkavalli, K., Johnson, M.K., Vlasic, D., Matusik, W., Pfister, H.: Video face replacement. ACM Trans. Graph. **30**(6), 130:1–130:10 (Dec 2011)
15. deepfakes: faceswap (2019), `https://github.com/deepfakes/faceswap`
16. Du, S., Tao, Y., Martinez, A.M.: Compound facial expressions of emotion. Proceedings of the National Academy of Sciences **111**(15), E1454–E1462 (2014). https://doi.org/10.1073/pnas.1322355111
17. Egger, B., Smith, W.A.P., Tewari, A., Wuhrer, S., Zollhoefer, M., Beeler, T., Bernard, F., Bolkart, T., Kortylewski, A., Romdhani, S., Theobalt, C., Blanz, V., Vetter, T.: 3D Morphable Face Models – Past, Present and Future. arXiv e-prints (Sep 2019)
18. Garrido, P., Valgaerts, L., Sarmadi, H., Steiner, I., Varanasi, K., Perez, P., Theobalt, C.: Vdub: Modifying face video of actors for plausible visual alignment to a dubbed audio track **34**(2), 193–204 (2015)

19. Garrido, P., Zollhöfer, M., Casas, D., Valgaerts, L., Varanasi, K., Pérez, P., Theobalt, C.: Reconstruction of personalized 3d face rigs from monocular video. ACM Trans. Graph. **35**(3), 28 (2016)
20. Gecer, B., Ploumpis, S., Kotsia, I., Zafeiriou, S.: Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
21. Geng, J., Shao, T., Zheng, Y., Weng, Y., Zhou, K.: Warp-guided gans for single-photo facial animation. ACM Trans. Graph. **37**(6), 231:1–231:12 (Dec 2018)
22. Geng, Z., Cao, C., Tulyakov, S.: 3d guided fine-grained face manipulation. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 9821–9830 (2019)
23. Ghosh, A., Fyffe, G., Tunwattanapong, B., Busch, J., Yu, X., Debevec, P.: Multiview face capture using polarized spherical gradient illumination. ACM Trans. Graph. **30**(6), 129:1–129:10 (2011)
24. 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)
25. Graves, A., Wayne, G., Danihelka, I.: Neural turing machines. arXiv preprint arXiv:1410.5401 (2014)
26. 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)
27. Hsieh, P.L., Ma, C., Yu, J., Li, H.: Unconstrained realtime facial performance capture. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1675–1683 (2015)
28. Hu, L., Saito, S., Wei, L., Nagano, K., Seo, J., Fursund, J., Sadeghi, I., Sun, C., Chen, Y.C., Li, H.: Avatar digitization from a single image for real-time rendering. ACM Trans. Graph. **36**(6) (2017)
29. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: ICCV (2017)
30. Huynh, L., Chen, W., Saito, S., Xing, J., Nagano, K., Jones, A., Debevec, P., Li, H.: Mesoscopic facial geometry inference using deep neural networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
31. Isola, P., Zhu, J., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: IEEE CVPR. pp. 5967–5976 (July 2017)
32. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196 (2017)
33. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. CoRR **abs/1812.04948** (2018), `http://arxiv.org/abs/1812.04948`
34. Kim, H., Elgharib, M., Zollhöfer, M., Seidel, H.P., Beeler, T., Richardt, C., Theobalt, C.: Neural style-preserving visual dubbing. ACM Transactions on Graphics (TOG) (2019)
35. Kim, H., Carrido, P., Tewari, A., Xu, W., Thies, J., Niessner, M., Pérez, P., Richardt, C., Zollhöfer, M., Theobalt, C.: Deep video portraits. ACM Trans. Graph. **37**(4), 163:1–163:14 (Jul 2018)
36. King, D.E.: Dlib-ml: A machine learning toolkit. Journal of Machine Learning Research **10**, 1755–1758 (2009)
37. Li, H., Weise, T., Pauly, M.: Example-based facial rigging. ACM Trans. Graph. **29**(3) (July 2010)

38. Li, H., Yu, J., Ye, Y., Bregler, C.: Realtime facial animation with on-the-fly correctives. ACM Trans. Graph. **32**(4) (July 2013)
39. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Smolley, S.P.: Least squares generative adversarial networks. In: Computer Vision (ICCV), 2017 IEEE International Conference on. pp. 2813–2821. IEEE (2017)
40. 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 Trans. Graph. **37**(6), 258:1–258:12 (Dec 2018)
41. Natsume, R., Yatagawa, T., Morishima, S.: Fsnet: An identity-aware generative model for image-based face swapping. In: Proc. of Asian Conference on Computer Vision (ACCV). Springer (Dec 2018)
42. Nirkin, Y., Keller, Y., Hassner, T.: Fsgan: Subject agnostic face swapping and reenactment. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 7184–7193 (2019)
43. Nirkin, Y., Masi, I., an Trãn, A.T., Hassner, T., Medioni, G.: On face segmentation, face swapping, and face perception. arXiv preprint arXiv:1704.06729 (April 2017)
44. Olszewski, K., Li, Z., Yang, C., Zhou, Y., Yu, R., Huang, Z., Xiang, S., Saito, S., Kohli, P., Li, H.: Realistic dynamic facial textures from a single image using gans
45. Pumarola, A., Agudo, A., Martinez, A., Sanfeliu, A., Moreno-Noguer, F.: Ganimation: One-shot anatomically consistent facial animation (2019)
46. Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: FaceForensics++: Learning to detect manipulated facial images. In: International Conference on Computer Vision (ICCV) (2019)
47. Saito, S., Li, T., Li, H.: Real-time facial segmentation and performance capture from rgb input. In: ECCV (2016)
48. Saito, S., Wei, L., Hu, L., Nagano, K., Li, H.: Photorealistic facial texture inference using deep neural networks. In: IEEE CVPR (2017)
49. Sengupta, S., Kanazawa, A., Castillo, C.D., Jacobs, D.W.: Sfsnet: Learning shape, refectance and illuminance of faces in the wild. In: Computer Vision and Pattern Regognition (CVPR) (2018)
50. Seymour, M., Evans, C., Libreri, K.: Meet mike: Epic avatars. In: ACM SIGGRAPH 2017 VR Village. pp. 12:1–12:2. SIGGRAPH '17, ACM, New York, NY, USA (2017). https://doi.org/10.1145/3089269.3089276, `http://doi.acm.org/10.1145/3089269.3089276`
51. Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., Sebe, N.: First order motion model for image animation. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32, pp. 7137–7147. Curran Associates, Inc. (2019), `http://papers.nips.cc/paper/8935-first-order-motion-model-for-image-animation.pdf`
52. Siarohin, A., Lathuilire, S., Tulyakov, S., Ricci, E., Sebe, N.: First order motion model for image animation. In: Conference on Neural Information Processing Systems (NeurIPS) (December 2019)
53. Suwajanakorn, S., Seitz, S.M., Kemelmacher-Shlizerman, I.: Synthesizing obama: learning lip sync from audio. ACM Trans. Graph. **36**(4), 95 (2017)
54. Thies, J., Zollhöfer, M., Nießner, M., Valgaerts, L., Stamminger, M., Theobalt, C.: Real-time expression transfer for facial reenactment. ACM Trans. Graph. **34**(6) (2015)
55. Thies, J., Zollhöfer, M., Nießner, M.: Deferred neural rendering: Image synthesis using neural textures. arXiv preprint arXiv:1904.12356 (2019)

56. Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., Nießner, M.: Face2face: Real-time face capture and reenactment of rgb videos. In: IEEE CVPR. pp. 2387–2395 (2016)
57. Tran, L., Liu, F., Liu, X.: Towards high-fidelity nonlinear 3d face morphable model. In: In Proceeding of IEEE Computer Vision and Pattern Recognition. Long Beach, CA (June 2019)
58. Wang, T.C., Liu, M.Y., Tao, A., Liu, G., Kautz, J., Catanzaro, B.: Few-shot video-to-video synthesis. In: Advances in Neural Information Processing Systems (NeurIPS) (2019)
59. Wang, T.C., Liu, M.Y., Zhu, J.Y., Liu, G., Tao, A., Kautz, J., Catanzaro, B.: Video-to-video synthesis. In: Advances in Neural Information Processing Systems (NeurIPS) (2018)
60. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: IEEE CVPR (2018)
61. Weise, T., Bouaziz, S., Li, H., Pauly, M.: Realtime performance-based facial animation. ACM Trans. Graph. **30**(4) (July 2011)
62. Wiles, O., Sophia Koepke, A., Zisserman, A.: X2face: A network for controlling face generation using images, audio, and pose codes. In: The European Conference on Computer Vision (ECCV) (September 2018)
63. Yamaguchi, S., Saito, S., Nagano, K., Zhao, Y., Chen, W., Olszewski, K., Morishima, S., Li, H.: High-fidelity facial reflectance and geometry inference from an unconstrained image. ACM Trans. Graph. **37**(4), 162:1–162:14 (Jul 2018)
64. Zakharov, E., Shysheya, A., Burkov, E., Lempitsky, V.: Few-shot adversarial learning of realistic neural talking head models. arXiv preprint arXiv:1905.08233 (2019)
65. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. arXiv preprint arXiv:1703.10593 (2017)

Appendix

In this supplementary material, we first explain details of the implementation, training strategy and performance of our method. We then provide additional results for evaluations and qualitative comparisons between our method and other one-shot face reenactment baselines on different datasets. Finally, we demonstrate the strong capability of our method by testing on in-the-wild portrait images from the Internet. More video results can be found in the supplementary video.

6   Implementation Details

All networks are MLPs in LD-Net, having 10 hidden layers with 512 features each. The length of the pose/expression code is 64 and the length of the identity code is 128.

In FD-GAN, the extractor and translator are based on U-Nets, with both networks joined together by dictionary writer/reader modules inserted in the up-convolution modules of the network, as shown in Fig. 3. in the paper.

In the down-convolution module, each level consists of a stride-2 convolution with 4×4 kernel, followed by a flat convolution with a 3×3 kernel. In the up-convolution module of the extractor, each level consists of a dictionary writer module, followed by a flat convolution with a 3×3 kernel and then a stride-2 convolution with 4×4 kernel. The up-convolution module of the translator is similar, with dictionary writers replaced with readers.

In the writer modules, write keys and write values are each computed from the input with a 1×1 convolution. In the reader modules, read keys are computed from the input with a 1×1 convolution and the values read from the dictionary are added back to the input feature, as in a residual block.

The discriminator and classifier for the generation part are patch-based and have the same structure as the down-convolution module of the U-Nets. For all networks, from the lowest level to the highest, the number of convolutions features, as well as the length of rows in the feature dictionary, are (32, 64, 128, 256). The number of rows in the dictionary are (512, 256, 128, 64), and the length of the read/write tags is 32 for all dictionaries.

7   Training Strategy

Although our FD-GAN implementation operates on 256×256 images, the LD-Net part should in principle be independent of image size. For LD-Net we normalize the landmark coordinates such that the square bounding box of all points span the range [-1, 1]. For FD-GAN, pixel values are normalized to [-1, 1].

The training configuration is given in table 5. Training time is in number of iterations.

Stage	Algorithm	LR	Batch	Time
LD-Net Stage 1	Adam	10^{-4}	32	4×10^5
LD-Net Stage 2	Adam	5×10^{-5}	32	10^6
FD-GAN	RMSprop	2×10^{-5}	4	10^6

Table 5. Training configuration of both sub-networks.

8   Performance

For one target image and its corresponding landmarks, the identity code in LD-Net and the feature dictionary in the FD-GAN can be reused for multiple source images. For each target, we measure the running time using all 100 frames in a corresponding test video. Landmark detection is performed separately in advance and is not included in the running time. It takes approximately 0.08s for FD-GAN to generate one image and 0.02s for LD-Net to do landmark disentangling on a single NVIDIA TITANX GPU.

9   Additional Qualitative Results

9.1   Ablation study

Comparison between with and without LD-Net. In Fig. 8 and Fig. 9, we show additional qualitative evaluations for self-reenactment and cross-subject face reenactment using the SelfTest dataset and CrossTest dataset, respectively. In Fig. 8, we compare results generated using ground truth landmarks (from the source video) and results using landmarks generated by LD-Net. As can be seen in the figure, our method can predict landmarks and synthesize high-quality images that are both close to the ground truth. For the cross-subject evaluation in Fig. 9, our results using landmarks by LD-Net not only have better identity preservation but also more precise poses/expressions (*e.g.* in the first row).

Comparison FD-GAN with baselines. Fig. 10 shows additional qualitative comparisons for cross-subject face reenactment on the CrossTest dataset with the three baselines, including the advanced image-to-image translation network Pix2PixHD [60], as well as two variants of FD-GAN (AdaIN and FD-GAN-1). AdaIN (colum 4) and FD-GAN-1 (column 5) use the full feature dictionary with AdaIN [29] or a one-line feature dictionary, respectively.

9.2   Comparison with one-shot methods

Additional results for one-shot cross-subject face reenactment on the CrossTest dataset and FaceForensics++ dataset [46] are shown in Fig. 11 and Fig. 12 respectively, with comparisons between our method and X2face [62] (column 3), X2face-aligned [62] (column 4), and First-order-model [51] (column 5). Note that in X2face [62], the generated frames inherit the object proportions of the driving

Fig. 8. Qualitative comparison on self-reenactment between the results using ground truth landmarks and the results with synthetic landmarks by LD-Net (LMs is short for Landmarks).

| Source | Target | Source LMs | Result1 | LD-Net LMs | Result2 |

Fig. 9. Qualitative comparison on cross-subject face reenactment between the results using source landmarks and the results with synthetic landmarks by LD-Net.
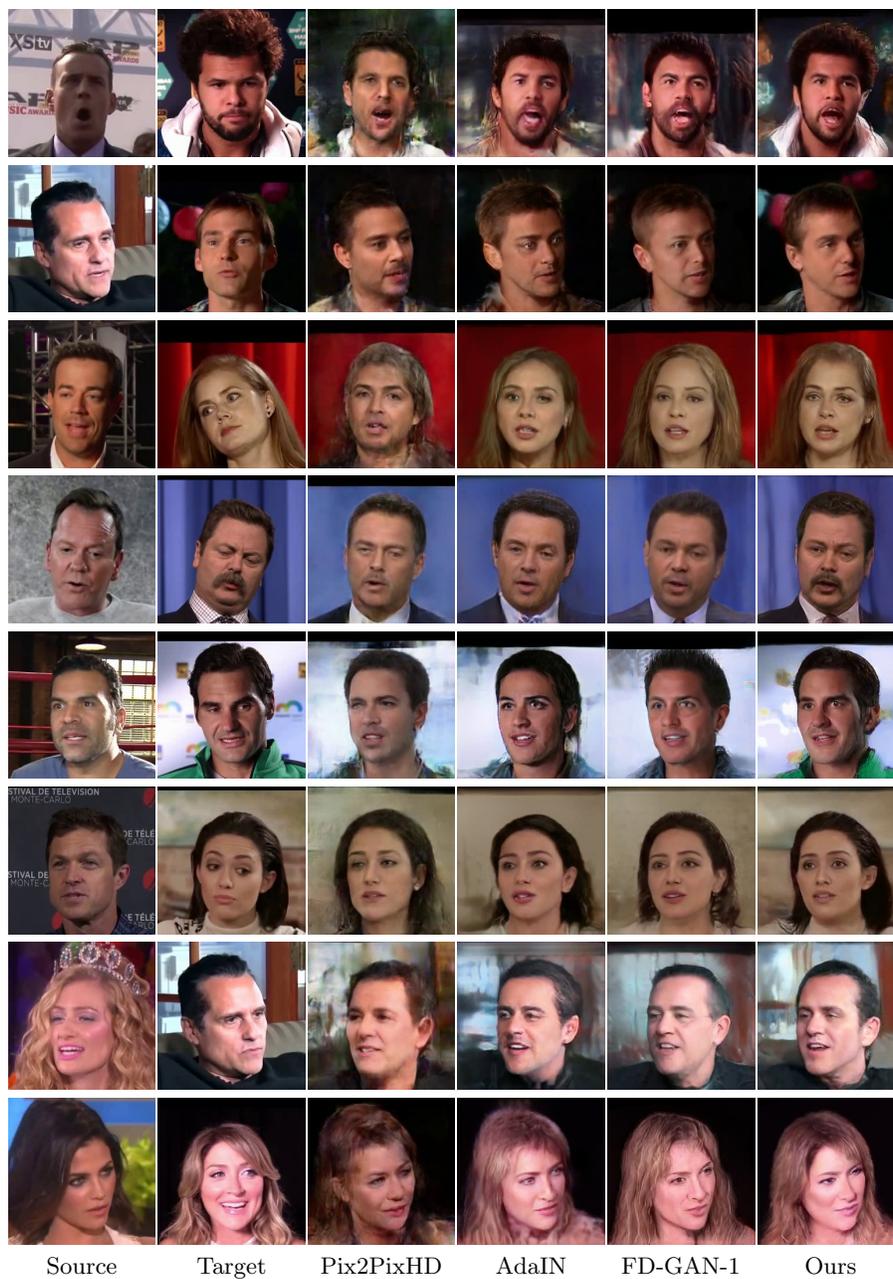
| Source | Target | Pix2PixHD | AdaIN | FD-GAN-1 | Ours |

Fig. 10. Qualitative comparison on cross-subject face reenactment between our FD-GAN and the baselines: Pix2PixHD [60], AdaIN and FD-GAN-1.

source video by transferring absolute coordinates, and thus it is very sensitive to face alignment. In addition to testing X2face [62] using exactly the same input configurations as other methods, we also take a smaller face region with a tighter bounding box as input to minimize the misalignment between source and target images for X2face [62] and obtain the results as shown as X2face-aligned. We also compute the metrics of the results by X2face-aligned on the CrossTest dataset, which are ISIM=0.7855, PSIM=0.7207, ED=0.344 and FID=62.14. Although the results are better than using non-aligned face images, our results are both quantitatively and qualitatively superior to theirs.

9.3   Comparison with 3D-based methods

Additional results for cross-subject face reenactment on the FaceForensics++ dataset [46] are shown in Fig. 13, with comparisons to two state-of-the-art 3D-based face reenactment methods, Face2Face [56] and NeuralTexture [55].

9.4   More Results

To demonstrate the capacity and generalization of our method, we test it on in-the-wild face images with the diverse appearance and challenging poses/expressions, including 2D paintings, historical photographs, as well as some celebrity portraits, as shown in Fig. 14. In the accompanying video, we provide more video examples for reference.

Source   Target   X2face   X2face   First-order   Ours
                                    -aligned   -model

Fig. 11. Qualitative comparison on cross-subject face reenactment using the CrossTest dataset between our method and one-shot methods: X2face [62], X2face-aligned [62] and First-order-model [51].
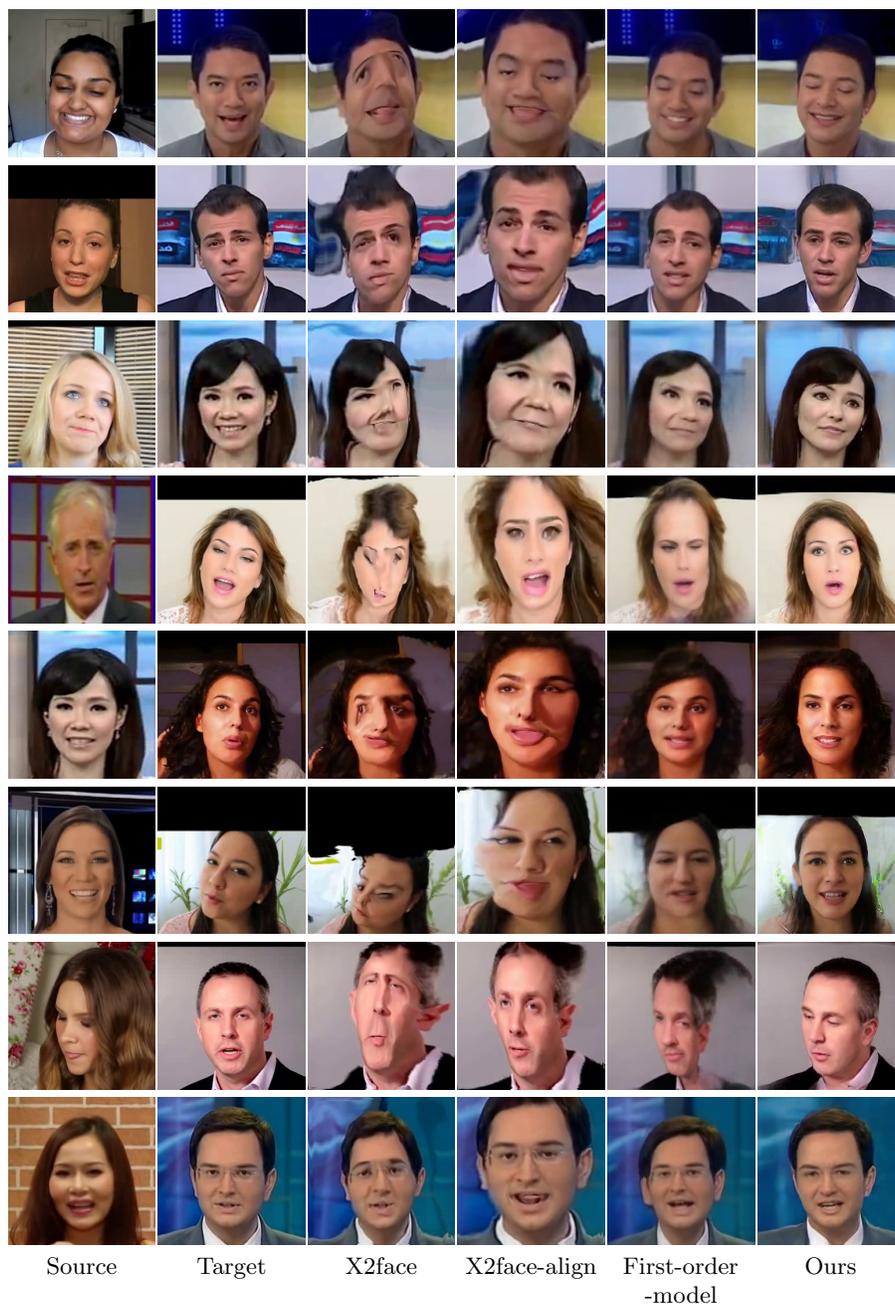
| Source | Target | X2face | X2face-align | First-order -model | Ours |

Fig. 12. Qualitative comparison on cross-subject face reenactment using FaceForensics++ dataset [46] between our method and one-shot methods: X2face [62], X2face-aligned [62] and First-order-model [51].
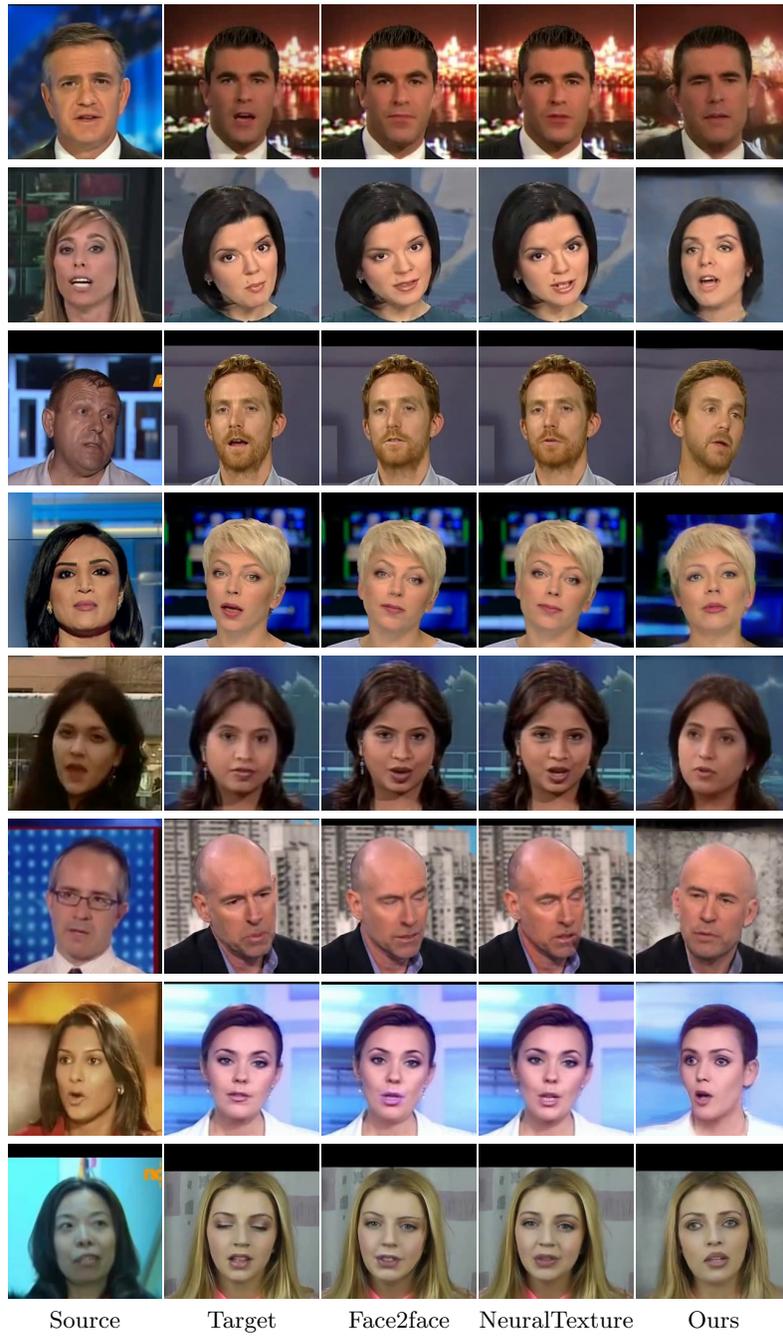
Source   Target   Face2face   NeuralTexture   Ours

Fig. 13. Qualitative comparison on cross-subject face reenactment with Face2face [56] and NeuralTexture [55] using input from FaceForensics++ [46]. Note that their methods control facial expressions only while ours can handle both the head pose and facial expressions.
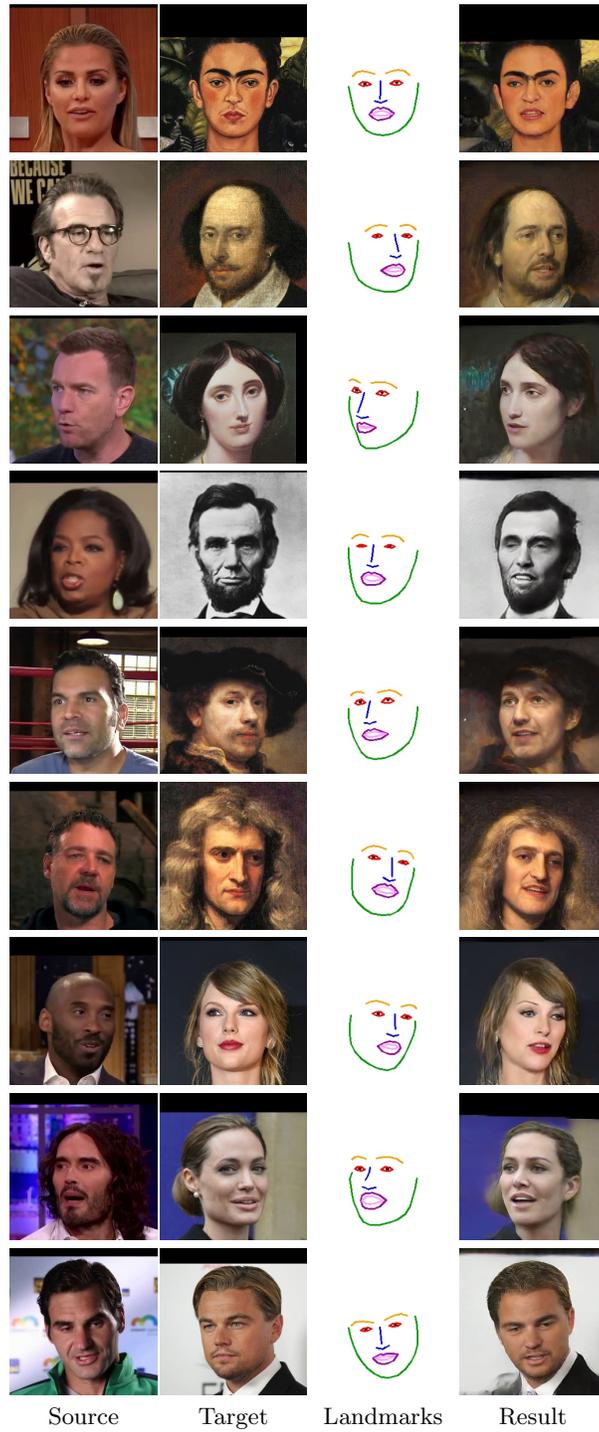
| Source | Target | Landmarks | Result |

Fig. 14. More results.