Face Translation between Images and Videos using Identity-aware CycleGAN

Zhiwu Huang[†], Bernhard Kratzwald[†], Danda Pani Paudel[†], Jiqing Wu[†], Luc Van Gool[†‡]

[†]Computer Vision Lab, ETH Zurich, Switzerland   [‡]VISICS, KU Leuven, Belgium

{zhiwu.huang, paudel, jwu, vangool}@vision.ee.ethz.ch, bkratzwald@ethz.ch

Abstract

This paper presents a new problem of unpaired face translation between images and videos, which can be applied to facial video prediction and enhancement. In this problem there exist two major technical challenges: 1) designing a robust translation model between static images and dynamic videos, and 2) preserving facial identity during image-video translation. To address such two problems, we generalize the state-of-the-art image-to-image translation network (Cycle-Consistent Adversarial Networks) to the image-to-video/video-to-image translation context by exploiting a image-video translation model and an identity preservation model. In particular, we apply the state-of-the-art Wasserstein GAN technique to the setting of image-video translation for better convergence, and we meanwhile introduce a face verificator to ensure the identity. Experiments on standard image/video face datasets demonstrate the effectiveness of the proposed model in both terms of qualitative and quantitative evaluations.

1. Introduction

In recent years, face translation between images has attracted a lot of attention. For example, [10, 14, 25, 15] translate facial images with identity-preserving. However, few works make efforts to translate faces between images and videos. To fill this gap, we propose the problem of bidirectional face translation between images and videos. Actually, image-video face translation has a broad range of applications. On one hand, given a static facial image, predicting its dynamic pattern in videos is very useful. On the other hand, it is also in high demand for video enhancement by translating low quality videos to high quality images.

Generally, image-video face translation suffers from more challenges than image-based face translation. One of the most challenging issues is treating the high heterogeneity between static images and dynamic videos. This challenge will become more complex when there are merely a limited number of paired image-video faces which belong to the same identity. Similar to the traditional image-to-

image translation task, it is also really difficult for us to collect paired images and videos as noted in [34]. As image-based face translation, the other major issue is that it is non-trivial to maintain the identity information especially when the involved faces are from two heterogeneous signals, that are 2D images and 3D videos.

With such motivation in mind, this paper focuses on the challenging problem of unpaired face translation between images and videos. In order to address the two crucial issues aforementioned, we aim to propose a bidirectional image-video face translation model in this paper. Inspired by the state-of-the-art image-to-image translation model (Cycle-Consistent Adversarial Networks, CycleGAN) [34], we design a cyclic image-video translation network with exploiting a image-to-video translation model, a video-to-image translation model, an image critic model, a video critic model and an identity preserving model. The image-to-video translation model consists of a pair of image encoder and video decoder, while the video-to-image translation model contains a pair of video encoder and image decoder. In addition to exploiting the identity mapping loss presented in [34], we also introduce a better identity preserving model that employs the state-of-the-art face verification network (FaceNet) [23] to ensure the faces are from the same identity. To train the whole translation network, we suggest an identity-aware cyclic GAN loss in an adversarial optimization on the translation between heterogeneous samples. In short, this paper brings two main innovations:

- To the best of our knowledge, this paper makes the first attempt to study the problem of unpaired face translation between images and videos.

- We propose bidirectional image-video translation models with exploiting different losses for the problem of image-to-video and video-to-image face translation.

2. Background

2.1. Image/Video Generation

As one of the most significant improvements on the research of image generation, Generative Adversarial Net-

works (GANs) [7, 21, 33, 4, 20] have drawn substantial attention from both the deep learning and computer vision society. The min-max two-player game involving in GANs provides a simple yet powerful way to estimate target distribution and generate novel image samples. Generally, the framework of GANs consists of a generative model G and a discriminative model D, one of which synthesizes highly-realistic images while the other of which distinguishes between synthesized images and real images. The state-of-the-art image GAN technique is Wasserstein GANs [2, 8] which empower the discriminative model with a critic C (replacing the discriminator D) of Earth-move distance (Wasserstein-1) to make the sample distribution close to the real data distribution:

$$\mathcal{L}_{\text{WGAN}} = \min_G \max_C \mathbb{E}_{\boldsymbol{x} \sim \mathbb{P}_r}[C(\boldsymbol{x})] - \mathbb{E}_{G(\boldsymbol{z}) \sim \mathbb{P}_g}[C(G(\boldsymbol{z}))]$$
$$+ \lambda \mathbb{E}_{\hat{\boldsymbol{x}} \sim \mathbb{P}_g}[(\|\nabla_{\hat{\boldsymbol{x}}} C(\hat{\boldsymbol{x}})\|_2 - 1)^2], \quad (1)$$

where $\boldsymbol{z}$ is random noise, $\hat{\boldsymbol{x}}$ is random samples following the distribution $\mathbb{P}_g$, G, C denotes the mapping functions of generator and critic respectively, and the hyperparameter λ controls the balance between the basic GAN loss and the gradient penalty regularization that aims to ensure the mapping function of the critic to be Lipschitz continuous.

As steady progress has been made toward better image generation, it is also important to study the video generation problem. Although the generated data has only one more dimension over time, the extension from generating images to generating videos typically turns out to be a much more challenging task for the following reasons. First, as a video sequence is a spatio-temporal recording of visual information of objects performing various actions, a generative model on video samples needs to learn the moving patterns of objects in addition to learning their appearance patterns. If the learned object constructs are incorrect, the generated videos typically contain objects performing physically implausible motion. Second, the temporal dimension brings in a huge amount of variation. This would be much more challenging when the background of videos is moving as well. Especially for video-based face generation, as the focus is on faces, one typically uses face detector to crop the involved faces while introducing more unstable changes in both foreground and background.

To our knowledge, there exist a limited number of video generation techniques [27, 26] through GANs. In particular, the state-of-the-art method [27] proposes two models: one employs one-stream video generation network, and the other contains two-stream generation networks, whose first stream serves to generate static background while the second stream produces dynamic foreground. [27] suggested to use the two-stream model that often outperforms the one-stream model. However, it is intuitive that the two-stream model works well only for videos with static backgrounds

with requiring an extra pre-processing step of background stabilization. In fact, the task of background stabilization is typically non-trivial and even may be impossible for videos in the wild. In this paper, we propose to apply the image-based Wasserstein GAN approach to video generation with using one-stream model to overcome the moving background issue that is often happened in video of faces.

2.2. Image-to-Image Translation

Among the first attempts, [9] suggested the problem of image-to-image translation, and employed a non-parametric texture model [6] from a single input-output training image pair. Recently, to address the problem, [13] built a pixel-to-pixel framework by utilizing a conditional generative adversarial network [7] to learn a mapping from input to output images. However, for many practical tasks, paired training data will not be available. Hence, several other works [16, 17, 22] attempted to tackle the unpaired setting, where the goal is to relate two data domains, $\boldsymbol{X}$ and $\boldsymbol{Y}$.

More recently, [34] proposed Cycle-Consistent Adversarial Netowrks (CycleGAN) to achieve state of the art for the problem of unpaired image-to-image translation. In addition to the standard GAN loss respectively for $\boldsymbol{X}$ and $\boldsymbol{Y}$, a pair of cycle consistency losses (forward and backward) was formulated using L1 reconstruction loss. For forward cycle consistency, given $\boldsymbol{x} \in \boldsymbol{X}$ the image translation cycle should reproduce $\boldsymbol{x}$. Backward cycle consistency is similar. Formally, the whole object of CycleGAN is

$$\mathcal{L}(G_X, G_Y, D_X, D_Y) = \min_{G_X, G_Y} \max_{D_X, D_Y} \mathcal{L}_{\text{GAN}}(G_X, D_X)$$
$$+ \mathcal{L}_{\text{GAN}}(G_Y, D_Y)$$
$$+ \gamma \mathcal{L}_{\text{cyc}}(G_X, G_Y), \quad (2)$$

where λ controls the relative importance between the GAN losses $\mathcal{L}_{\text{GAN}}(G_X, D_X), \mathcal{L}_{\text{GAN}}(G_Y, D_Y)$ and the cycle consistency loss $\mathcal{L}_{cyc}(G_X, G_Y)$:

$$\mathcal{L}_{\text{GAN}}(G_X, D_X) = \mathbb{E}_{\boldsymbol{x} \sim \mathbb{P}_{X_r}}[\log D_X(\boldsymbol{x})]$$
$$- \mathbb{E}_{G_X(\boldsymbol{y}) \sim \mathbb{P}_{X_g}}[\log D_X(G_X(\boldsymbol{y}))]. \quad (3)$$

$$\mathcal{L}_{\text{GAN}}(G_Y, D_Y) = \mathbb{E}_{\boldsymbol{y} \sim \mathbb{P}_{Y_r}}[\log D_Y(\boldsymbol{y})]$$
$$- \mathbb{E}_{G_Y(\boldsymbol{x}) \sim \mathbb{P}_{Y_g}}[\log D_Y(G_Y(\boldsymbol{x}))]. \quad (4)$$

$$\mathcal{L}_{\text{cyc}}(G_X, G_Y) = \mathbb{E}_{\boldsymbol{x} \sim \mathbb{P}_{X_g}}[\|G_X(G_Y(\boldsymbol{x})) - \boldsymbol{x}\|_1]$$
$$+ \mathbb{E}_{\boldsymbol{y} \sim \mathbb{P}_{Y_g}}[\|G_Y(G_X(\boldsymbol{y})) - \boldsymbol{y}\|_1]. \quad (5)$$

In this paper, our focus is on image-to-video and video-to-image translation by leveraging the GAN techniques. In contrast to the existing techniques for image-to-image translation, our proposed GAN approach allows hybrid image-based and video-based models to work together which proposes more challenges. In addition, as usual applications on faces, translation on faces is also quite different from translation on general objects.
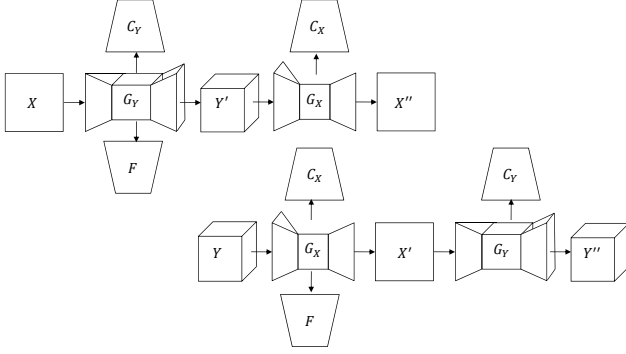
Figure 1. Conceptual illustration of the proposed Identity-aware CycleGAN (IdCycleGAN). Here X, Y, X', Y', X'', Y'' denotes input images, input videos, output images and output videos, reconstructed images and reconstructed videos. G_X, G_Y, C_X, C_Y, F represents image generator, video generator, image critic, video critic and face verificator respectively.

2.3. Identity-aware Face Translation

Face generation can be regarded as a kind of fine-grained object generations due to its special attributes. On one hand, in contrast to general objects, the global structures of faces are approximately the same such that each face can be aligned easier with a universal mean face template. Therefore, one typically can first detect faces in images, and then crop and align the faces for better generation. On the other hand, the variety of real faces are as much as general objects. Accordingly, it is also very challenging to generate identity-preserving faces from static images and even from dynamic videos.

Preserving facial identity has also been explored in synthesizing the corresponding frontal face image from a single side-view face image [10], where the identity preserving loss was defined based on the activations of the last two layers of the light CNN [29]. In multi-view image generation from a single view [32], a condition image (e.g. frontal view) was used to constrain the generated multiple views in their coarse-to-fine framework.[1] However, such works only designed generative models for image-based face generation rather than video-based generation which is much more challenging as studied before. Hence, to the best of our knowledge, this paper make the first attempt to generate identity-preserving faces in the context of image-to-video and video-to-image translation.

3. Our Model

The goal of our method is to learn mapping functions between image domain X and video domain Y given training

---

[1]There exist other works such as [15, 19] for identity-preserving face generation, most of which however have not been formally published to our best knowledge.
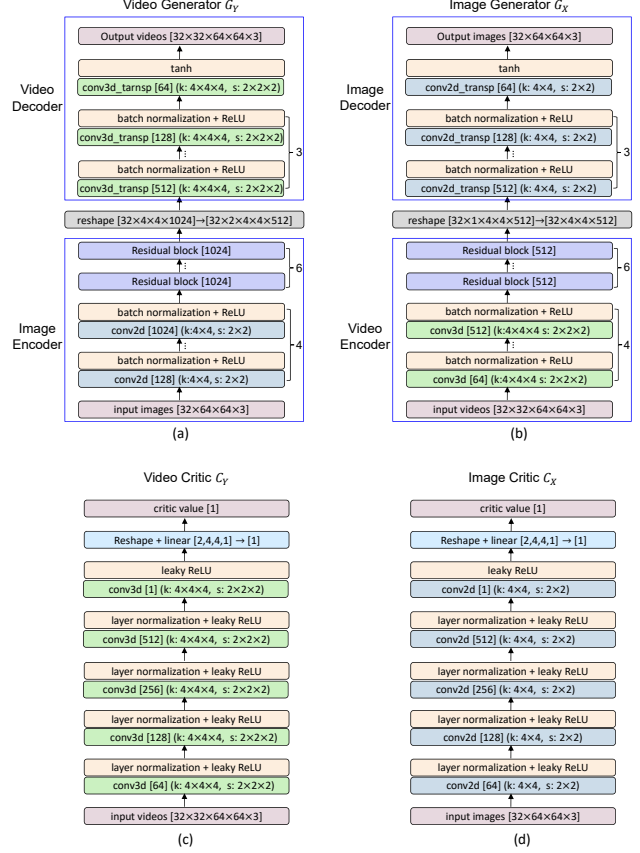

Figure 2. Architectures of the designed image-to-video translator (video generator), video-to-image translator (image generator), critic networks for images and videos respectively. The digits on the right of the big parantheses indicate the number of repeated components.

image samples $\{\boldsymbol{x}_i\}_{i=1}^N \in \boldsymbol{X}$ and video samples $\{\boldsymbol{y}_i\}_{j=1}^M \in \boldsymbol{Y}$. As illustrated in Fig.1, our model follows CycleGAN [34] to include two inverse mappings: $G_X : \boldsymbol{X} \to \boldsymbol{Y}$ and $G_Y : \boldsymbol{Y} \to \boldsymbol{X}$. Following the success of Wasserstein GANs [2, 8], we also introduce two critic model $C_{\boldsymbol{X}}$ and C_Y, where C_X aims to disambiguate between real images $\boldsymbol{x}$ and translated images $G_X(\boldsymbol{y})$, while C_Y aims to distinguish between real videos $\boldsymbol{y}$ and translated videos $G_Y(\boldsymbol{x})$. In addition, we also introduce a face identification network $F(\boldsymbol{x}, \boldsymbol{y})$ [23] to enable the translated images/videos preserve the identity of the input faces better from input images/videos.

3.1. Image-Video CycleGAN

For the image-to-video translation mapping $G_{\boldsymbol{Y}}$, our model designs a novel translator network that incorporates image decoder and video encoder; In the same way, for video-to-image translation $G_{\boldsymbol{X}}$, we devise a translator network that cascades video decoder with image generator. As

shown in Fig.2(a), the image-to-video translator receives a mini-batch of training image samples $\{x_i\}_{i=1}^B$ (B is the batch size) as input, then applies an image decoder to decode the images to lower-dimensional latent codes $\{z_i\}_{i=1}^B$, followed by a video decoder that finally produces video samples $\{y_i\}_{i=1}^B$ containing l frames of size $r \times r$. More specially, the image decoder contains four fractionally-strided 2D convolutional layers, each of which is followed one batch normalization layer and one ReLU layer. Following [34], we also insert three residual blocks into the the third and the fourth 2D convolutional layers respectively. After the image decoding process, we apply a video decoder to produce video samples from the encoded latent codes. The video decoder consists of four layers of fractionally-strided 3D convolutional layers [24], each of which works together with one batch normalization layer and one ReLU layer. The setting has proven to be an efficient approach to upsample while still maintaining spatial and temporal invariances [24]. Similarly, as illustrated in Fig.2 (b), the video-to-image translator contains a video encoder that encodes the input videos to latent codes, and a image decoder generates images from the resulting latent codes.

In order to generate realistic images and videos simultaneously, we apply the Wasserstein GAN framework to the setting of image-video translation. In particular, we design two different critic networks respectively for image generation and video generation. As shown in Fig.2 (c), the image critic network follows the architecture design of [2, 8], while the video critic model (see Fig.2 (d)) is specially designed to measure the Wasserstein distance between generated videos and real videos. In particular, the video critic network accepts a video as input, and maps it to a real valued input. It contains five convolutional layers as done in [2, 8] but is then followed by an additional linear layer. Like the image/video generator designed above, we also employ spatio-temporal convolutions with a bias to the output. The convolutions also include a certain of strides in order to allow efficient down-sampling of the high dimensional videos. Following [8], we replace regular batch normalization [12] layers with layer normalization [3] layers as batch normalization is not valid in the setting of penalizing the norm of the critcs gradient with respect to each input independently. In addition, all but the last convolutional layers use a leaky ReLU activation function. As [2, 8], we finally use a linear transformation layer to obtain the expectation of the input video samples such that the Wasserstein distance between real and fake samples can be calculated.

With designing such image-video translators and critic models for images and videos, we finally apply Wasserstein loss Eqn.1 to both the image-to-video mapping G_Y and video-to-image mapping G_X. So far, the network can only map the same set of input images/videos to any random permutation of videos/images in the target domain, where

any of the learned mappings can derive an output distribution that approximate the target distribution. Thus, inspired by [34], we additionally leverage the cycle-consistent loss, with which an individual input can be mapped to a desired output by further reducing the space of possible mapping functions. Hence, we express the objective of our image-to-video/video-to-image translation model as

$$
\begin{aligned}
\mathcal{L}(G_X, G_Y, C_X, C_Y) = \min_{G_X, G_Y} \max_{C_X, C_Y} & \mathcal{L}_{\text{WGAN}}(G_X, C_X) \\
& + \mathcal{L}_{\text{WGAN}}(G_Y, C_Y) \\
& + \gamma \mathcal{L}_{\text{cyc}}(G_X, G_Y),
\end{aligned}
\tag{6}
$$

where the involved Wasserestein loss $\mathcal{L}_{\text{WGAN}}(G_X, C_X)$ and $\mathcal{L}_{\text{WGAN}}(G_Y, C_Y)$ respects the form of Eqn.1, and the cycle consistency loss $\mathcal{L}_{\text{cyc}}(G_X, G_Y)$ can be achieved by Eqn.5. In experiments, we observe that the Wasserstein loss can make the whole training process more stable than using other traditional GAN losses.

3.2. Identity-aware CycleGAN

It is still very hard to make the input face and the resulting face being from the same identity. Therefore, we resort to one more constraint that can preserve identity during face translation. One of the most straightforward ways is to apply L2 reconstruction error between the translated samples and the real samples. As studied in [34], to encourage the mapping to preserve color composition between the input and output, it is helpful to introduce an additional loss, i.e., pixel-wise identity mapping loss:

$$
\begin{aligned}
\mathcal{L}_{\text{id}}(G_X, G_Y) = & \mathbb{E}_{x \sim \mathbb{P}_{X_g}}[\|(G_Y(x))_i - x\|_2] \\
& + \mathbb{E}_{y \sim \mathbb{P}_{Y_g}}[\|G_X(y) - y_i\|_2].
\end{aligned}
\tag{7}
$$

Intuitively, the approach can merely maintain the original appearance on the pixel level. As proved in [23], the faithful identity metric can be better achieved on in low-dimensional embedding space. Accordingly, we alternatively come up with using favorable identity metric loss to keep the identity of the original input faces when generating output faces. To this end, we additionally insert a verificator to verify whether the translated face and the input face belong to the same identity.

In this paper, we exploit the state-of-the-art face verification network (FaceNet) [23], which has surpassed human recognition accuracy on some standard face datasets. Essentially, FaceNet maps face images to an embedded space where the squared L2 distance directly correspond to the similarities of the involved faces. As done in the original FaceNet work [23] that treats video-to-video face verification, we use the average similarity of all pairs of the real (generated) image and the generated (real) frames within the same video clip. The lower score is, the more similar the facial image and video are. Formally, the additional

4

identity-preserving loss is expressed as

$$\mathcal{L}_{\text{id}}(G_X, G_Y) = \mathbb{E}_{\boldsymbol{x} \sim \mathbb{P}_{X_g}}[\|F((G_Y(\boldsymbol{x}))_i) - F(\boldsymbol{x})\|_2] \\ + \mathbb{E}_{\boldsymbol{y} \sim \mathbb{P}_{Y_g}}[\|F(G_X(\boldsymbol{y})) - F(\boldsymbol{y}_i)\|_2]. \quad (8)$$

Consequently, the final objective of the proposed model includes three terms: an adversarial Wasserstein loss $\mathcal{L}_{\text{WGAN}}(G_X, C_X), \mathcal{L}_{\text{WGAN}}(G_Y, C_Y)$ for matching the distribution of generated images/videos to the data distribution in the target images/videos; a cycle consistency loss $\mathcal{L}_{\text{cyc}}(G_X, G_Y)$ to prevent the learned mappings G_X and G_Y from contradicting each other; and a identity-preserving loss $\mathcal{L}_{\text{id}}(G_X, G_Y)$ with a hyperparameter ω to measure the identity metric between the generated images/videos and the input source images/videos.

4. Experiment

To evaluate our model for face translation between images and videos, we use standard face datasets: Celebrity Face (combining CelebA [18] with YouTubeFace [28]) and Point-and-Shoot Challenge (PaSC) [5]. To justify the effectiveness of our designed image-video translator architecture for robust video generation, we first compare our model against the state-of-the-art video generation technique VideoGAN [27]. Then we evaluate our proposed image-video translation model with using the Wasserstein setting of the baseline CycleGAN [34] loss Eqn.6 (denoted as *Model1*), adding pixel-wise identity loss Eqn.7 (denoted as *Model2*), or adding face identity loss Eqn.8 (denoted as *Model3*) on the two used face datasets for face translation between static images and dynamic videos.

4.1. Experimental Setting

The used Celebrity Face dataset (Celeb) consists of CelebA [18] and YouTubeFace [28], which contains facial images and videos respectively. The CelebA contains 202,599 images in total, and we use the frontal face detector with Haar cascades to get 156,254 frontal faces in total, which can constitute relatively high quality frontal facial images. The YouTubeFace has 3,425 videos of 1,595 different celebrities. An average of 2.15 videos are available for each subject, and the average length of a video clip is 181.3 frames. The video frames often exhibit a large variation of pose and illumination, as well as degradations such as compression effects. We employ the state-of-the-art face detector from [31] to detect faces from the YouTube videos, and finally obtain 625,988 facial frames of 626,146 in total. In contrast, the PaSC dataset [5] is designed to simulate more of a video surveillance system. The PaSC was collected for face recognition from stills and videos captured by point-and-shoot cameras. It includes 9,376 still images and 2,802 videos (303,222 frames) of 293 people in total.

The data has large variation in terms of distance to the camera, alternative sensors, frontal versus not-frontal views, and varying locations. As done for Celeb, we finally get 9,249 still facial images and 300,206 farcical video frames. Considering the balance of performance and computational cost of the evaluated models, all the faces are center cropped to the size of 64×64.

For our proposed IdCycleGAN models (Model1, Model2, Model3), following [8], we set the balance parameter λ of the gradient norm penalty regularization to 10. The parameter γ of the cycle consistency loss and is fixed as 1000, while the parameter ω of the identity-aware loss is set to 100. We follow [23] to employ the inception-ResNet-v1 model to achieve the verifcator model F of our IdCycleGAN, that actually uses the pretrained FaceNet. For the baseline video generation method VideoGAN [27], we set the trade-off parameter of using mask to combine the forground and background stream in the context of its two-stream model. For the baseline image-to-image translation model CycleGAN [34], we adapt it to the scenario of image-video face translation with using our proposed image-video translation model, more stable Wasserstein GAN loss and both forward and backward cycle consistency loss. The training of each deep model was performed on a GeForce GTX TITAN X GPU.

4.2. Qualitative Evaluation

Video Face Generation: We first degrade our model of image-video translation to video generation model with merely removing the image decoder out of the video generator and only using the Wasserstein GAN loss for generating videos. In the case, we can evaluate how robust our proposed model performs by comparing with the state-of-the-art video generation model VideoGAN [27] including its one-stream model and two-stream model. In Fig.3, we show some examples of the videos generated from the VideoGAN one-stream model, the VideoGAN two-stream model and our one-stream model. From the results on the Celeb and PaSC datasets, we can observe that the VideoGAN two-stream model performs badly on both the two datasets whereas the VideoGAN one-stream model works relatively normally. In contrast, our proposed model can perform the best with generating relatively sharper and more realistic faces. This not only verifies that our proposed approach with using one-stream model can perform well in the real-world setting, but also demonstrate that our generalized Wasserstein GAN framework in the context of videos can converge more well in comparison to the state-of-the-art VideoGAN that is under the traditional DCGAN [21] framework.

Image-Video Face Translation: Fig.4 and Fig.5 quantitatively exhibit the results of of the comparing our three models Model1, Model2, Model3 for face translation be-
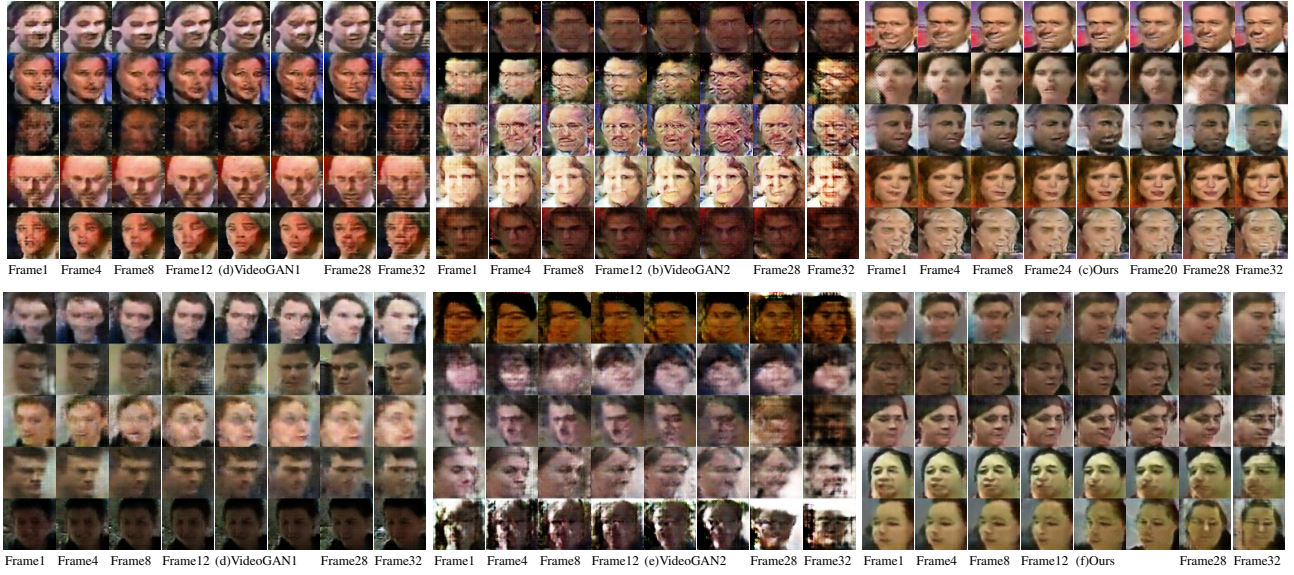
Figure 3. Video generation results of VideoGAN1 (one-stream) [27], VideoGAN2 (two-stream) [27] and our proposed model respectively on the Celeb (a)-(b)-(c) and the PaSC (d)-(e)-(f)
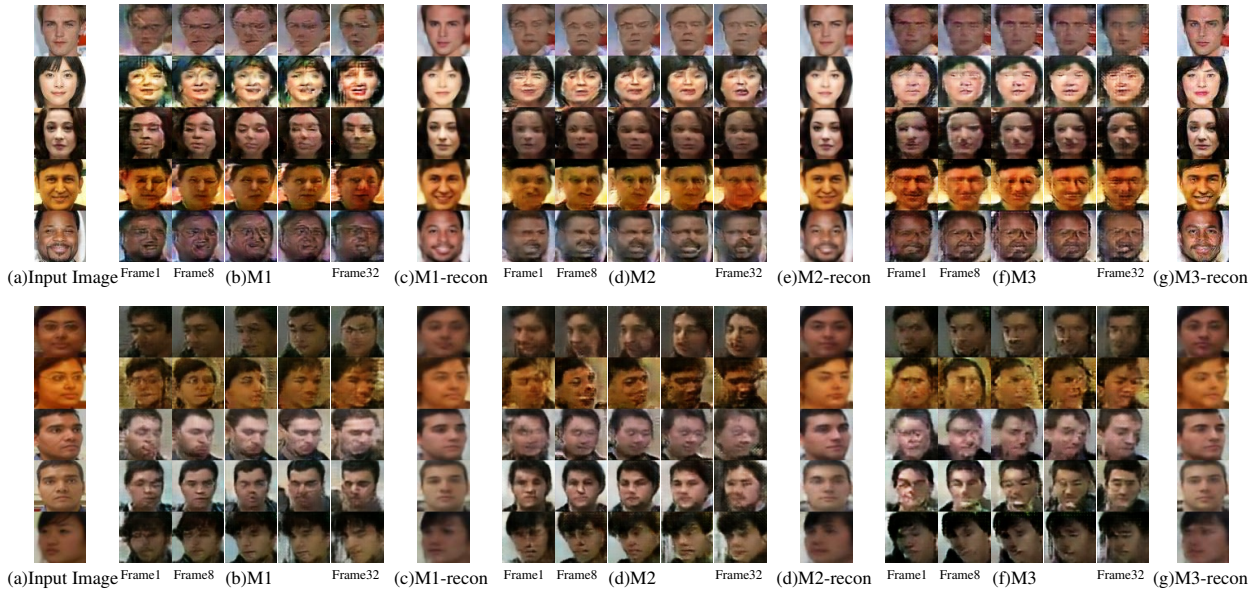

Figure 4. Image-to-Video translation results (b)Model1-(d)Model2-(f)Model3 and their reconstructed results (c)-(e)-(g) on the Celeb and PaSC databases

tween images and videos on the Celeb and the PaSC database. Note that we also show some more translated static and dynamic facial samples in Fig.7 which is best viewed with the Adobe reader.

In the qualitative evaluation for image-to-video translation Fig.4 (a)-(b)-(d)-(f), we can discover that all the three models can generate visually plausible and temporally continuous facial video frames from input facial images. In addition to using the stable Wasserstein GAN loss, we at-

tribute the good performance to using the proposed image-video generator network that enables the hybrid decoding and encoding process on involved images and videos to work well together. In addition, by comparing with Model1 and Model2, we can find that Model3 performs the best in terms of maintaining the identity of input faces from images/videos. This demonstrates that the proposed identity-aware loss can enhance the capability of preserving identity when translating faces between images and videos.

Frame1 Frame8 (a)Input Video Frame32 (b)M1 Frame1 Frame8 (c)M1-recon Frame32 (d)M2 Frame1 Frame8 (e)M2-recon Frame32 (f)M3 Frame1 Frame8 (g)M3-recon Frame32

Frame1 Frame8 (a)Input Video Frame32 (b)M1 Frame1 Frame8 (c)M1-recon Frame32 (d)M2 Frame1 Frame8 (e)M2-recon Frame32 (f)M3 Frame1 Frame8 (g)M3-recon Frame32

Figure 5. Video-to-Image translation results (b)Model1-(d)Model2-(f)Model3 and their reconstructed results (c)-(e)-(g) on the Celeb and PaSC databases

FaceNet Score	Image-to-Video		Video-to-Image	
	Celeb	PaSC	Celeb	PaSC
Model1	0.0611	0.0523	0.0675	0.0678
Model2	0.0493	0.0486	0.0621	0.0610
Model3	0.0428	0.0356	0.0538	0.0504

Table 1. FaceNet score (measured by L2 distance of the FaceNet [23] face embeddings) of image-to-video and video-to-image translation for the three proposed models on the Celeb and PaSC datasets.

In the video-to-image translation experiment (Fig.5 (a)-(b)-(d)-(f)), on the Celeb face dataset, Model1 can often translate plausible faces between images and videos, but the identity information is not kept during translation like the fourth and the fifth samples. While Model2 can address the problem to some extent, it still fails in some cases like the 5-th facial sample whose mouth is occluded by a melodica. Thanks to imposing the facial identity loss, Model3 can maintain the identity better in the process of face translation. Similar observation can be achieved on the PaSC face dataset by comparing the translated facial samples especially the first and the fourth samples that are shown in (a)-(b)-(d)-(f).

As we can see from some failure cases shown in Fig.7, the non-frontal and quick moving faces are the most difficult to translate for all of the three proposed models. In addition, the visual quality of the Model3's generated samples are not always the best which suggests the identity preserving term would be very strong to prevent the model from getting more visually plausible results.

4.3. Quantitative Evaluation

We use verification score of the state-of-the-art face verification network FaceNet [23] to measure the quantitative results of the generated 1000 samples using the three different translation model. As reported in Table.1, we can find that the proposed Model3 can achieve the lowest verification score which means it has the strongest ability to preserve the identity during the image-to-video and video-to-image face translation task.

In addition, we also study the identity score curve of the proposed Model3 that explicitly includes the identity preserving cost. As depicted in Fig.6, on the two used datasets, the identity-based cost can always decrease stably as the training iteration number increases towards 50K. This demonstrates that the proposed model can be trained to transfer samples with keeping identity better.

5. Conclusion and Future Work

This paper presented a new problem of unpaired face translation between static images and dynamic videos which can be applied to video face prediction and enhancement. To handle the problem, we proposed a cyclic image-video translation model to bidirectionally translate faces between images and videos in a unified GAN model. To better preserve the identity during translation, we additionally introduce a face verification model that ensures to maintain the facial identity of the original input faces in images or videos. Evaluations on two standard face database demonstrate the proposed model is able to perform image-

(a) Image-to-Video translation on Celeb

(b) Image-to-Video translation on PaSC

(c) Video-to-Image translation on Celeb

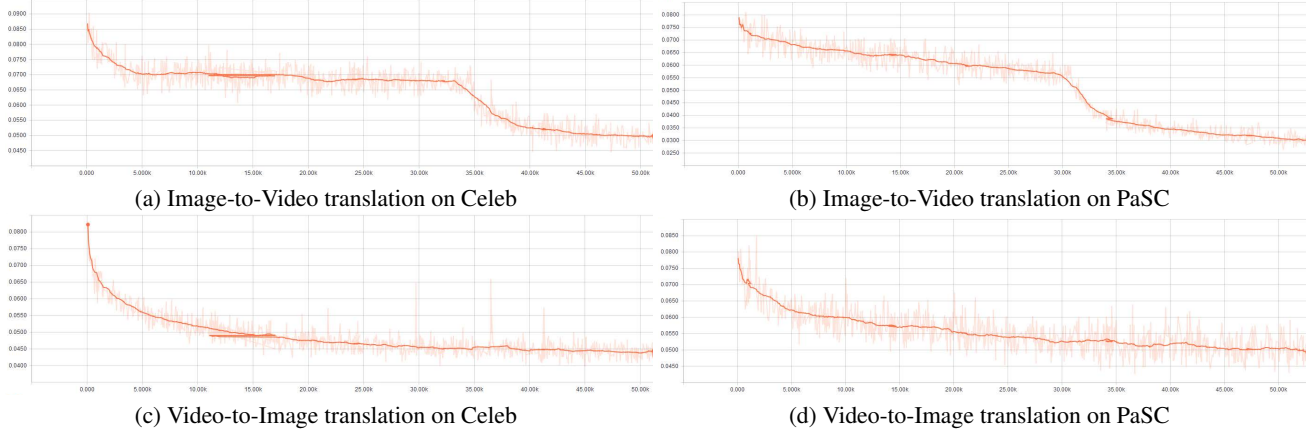(d) Video-to-Image translation on PaSC

Figure 6. Identity score curve of the proposed model during network training. Here, the values in x-axis and y-axis respectively denotes the training iteration number and the identity score, which is measured by L2 distance of the FaceNet [23] face embeddings.

(a) Input Image   (b) Model1   (c) Model2   (d) Model3   (e) Input Video   (f) Model1   (g) Model2   (h) Model3
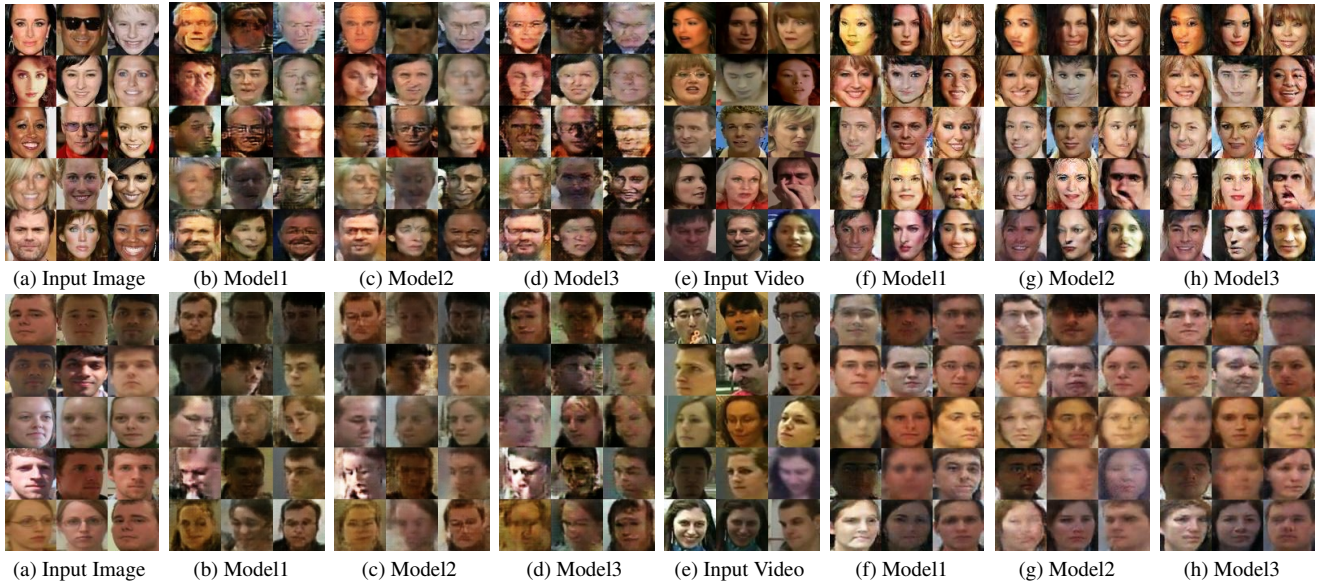
Figure 7. More image-to-video face translation results (b)-(c)-(d) and video-to-image face translation results (f)-(g)-(h) on the Celeb and PaSC databases. Please notice that the figure is best viewed via the Adobe Acrobat Reader. Click each involved video to play the dynamic video clip.

to-video/video-to-image face translation well in both terms of visual quality and identity maintenance.

The proposed model actually balances the three different terms: GAN loss, cycle consistency loss and identity preserving loss, which actually correspond to the designed critic model, generator model and face verification model. We empirically choose the default setting to make the balance among them. Hence, there may remain an interesting problem to balance such three functions better.

In addition, we actually freeze the network parameters of the pretrained FaceNet during the training process. Pursuing a better gradient feedback from the face model to better the generated results in both terms of visual appearance and

semantic maintenance still remains challenging. It would be also very interesting if we can simultaneously train the generator, discriminator (critic model) and face verification model.

As future work, for better balance between image quality and identity preservation, we will make some efforts to automatically control the hyperparameter between them. Besides, we will also borrow the successful insights of existing stacked GANs [11, 30, 1] to extend this work to generate higher-quality images and videos. Lastly, it would be also worth exploring the direction of generalizing the proposed model to translation on video-to-video and even 3D volume data like 3D brain image in medical imaging.

Acknowledgements

We would like to thank NVidia for donating the GPUs used in this work.

References

[1] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive Growing of GANs for Improved Quality, Stability, and Variation. *arXiv preprint arXiv:1710.10196*, 2017. 8

[2] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *ICML*, 2017. 2, 3, 4

[3] J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 4

[4] D. Berthelot, T. Schumm, and L. Metz. BEGAN: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017. 2

[5] J. R. Beveridge, J. Phillips, D. S. Bolme, B. Draper, G. H. Givens, Y. M. Lui, M. N. Teli, H. Zhang, W. T. Scruggs, K. W. Bowyer, P. J. Flynn, and S. Cheng. The challenge of face recognition from digital point-and-shoot cameras. In *BTAS*, 2013. 5

[6] A. A. Efros and T. K. Leung. Texture synthesis by non-parametric sampling. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pages 1033–1038. IEEE, 1999. 2

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014. 2

[8] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville. Improved training of Wasserstein GANs. *NIPS*, 2017. 2, 3, 4, 5

[9] A. Hertzmann, C. E. Jacobs, N. Oliver, B. Curless, and D. H. Salesin. Image analogies. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 327–340. ACM, 2001. 2

[10] R. Huang, S. Zhang, T. Li, and R. He. Beyond face rotation: Global and local perception GAN for photo-realistic and identity preserving frontal view synthesis. *ICCV*, 2017. 1, 3

[11] X. Huang, Y. Li, O. Poursaeed, J. Hopcroft, and S. Belongie. Stacked generative adversarial networks. *arXiv preprint arXiv:1612.04357*, 2016. 8

[12] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 4

[13] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint arXiv:1611.07004*, 2016. 2

[14] Y. Li, S. Liu, J. Yang, and M.-H. Yang. Generative face completion. *CVPR*, 2017. 1

[15] Z. Li and Y. Luo. Generate identity-preserving faces by generative adversarial networks. *arXiv preprint arXiv:1706.03227*, 2017. 1, 3

[16] M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. *arXiv preprint arXiv:1703.00848*, 2017. 2

[17] M.-Y. Liu and O. Tuzel. Coupled generative adversarial networks. In *NIPS*, 2016. 2

[18] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 5

[19] Y. Lu, Y.-W. Tai, and C.-K. Tang. Conditional CycleGAN for attribute guided face image generation. *arXiv preprint arXiv:1705.09966*, 2017. 3

[20] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smolley. Least squares generative adversarial networks. *ICCV*, 2017. 2

[21] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 2, 5

[22] R. Rosales, K. Achan, and B. J. Frey. Unsupervised image translation. In *ICCV*, 2003. 2

[23] F. Schroff, D. Kalenichenko, and J. Philbin. FaceNet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 1, 3, 4, 5, 7, 8

[24] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3D convolutional networks. In *ICCV*, 2015. 4

[25] L. Tran, X. Yin, and X. Liu. Disentangled representation learning GAN for pose-invariant face recognition. In *CVPR*, 2017. 1

[26] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz. MoCoGAN: Decomposing motion and content for video generation. *arXiv preprint arXiv:1707.04993*, 2017. 2

[27] C. Vondrick, H. Pirsiavash, and A. Torralba. Generating videos with scene dynamics. In *NIPS*, 2016. 2, 5, 6

[28] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *CVPR*, 2011. 5

[29] X. Wu, R. He, Z. Sun, and T. Tan. A light CNN for deep face representation with noisy labels. *arXiv preprint arXiv:1511.02683*, 2015. 3

[30] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*, 2017. 8

[31] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016. 5

[32] B. Zhao, X. Wu, Z.-Q. Cheng, H. Liu, and J. Feng. Multi-view image generation from a single-view. *arXiv preprint arXiv:1704.04886*, 2017. 3

[33] J. Zhao, M. Mathieu, and Y. LeCun. Energy-based generative adversarial network. *arXiv preprint arXiv:1609.03126*, 2016. 2

[34] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *ICCV*, 2017. 1, 2, 3, 4, 5