Coherence and Identity Learning for Arbitrary-length Face Video Generation

Shuquan Ye
Department of Computer Science
City University of Hong Kong
Email: shuquanye2-c@my.cityu.edu.hk

Chu Han
Department of Radiology
Guangdong Provincial People's Hospital
Guangdong Academy of Medical Sciences
Email: zq1992@gmail.com

Jiaying Lin
Department of Computer Science
City University of Hong Kong
Email: jiayinlin5-c@my.cityu.edu.hk

Guoqiang Han
School of Computer Science and Engineering
South China University of Technology
Email: csgqhan@scut.edu.cn

Shengfeng He*
School of Computer Science and Engineering
South China University of Technology
Email: hesfe@scut.edu.cn

***Abstract*—Face synthesis is an interesting yet challenging task in computer vision. It is even much harder to generate a portrait video than a single image. In this paper, we propose a novel video generation framework for synthesizing arbitrary-length face videos without any face exemplar or landmark. To overcome the synthesis ambiguity of face video, we propose a divide-and-conquer strategy to separately address the video face synthesis problem from two aspects, face identity synthesis and rearrangement. To this end, we design a cascaded network which contains three components, Identity-aware GAN (IA-GAN), Face Coherence Network, and Interpolation Network. IA-GAN is proposed to synthesize photorealistic faces with the same identity from a set of noises. Face Coherence Network is designed to re-arrange the faces generated by IA-GAN while keeping the inter-frame coherence. Interpolation Network is introduced to eliminate the discontinuity between two adjacent frames and improve the smoothness of the face video. Experimental results demonstrate that our proposed network is able to generate face video with high visual quality while preserving the identity. Statistics show that our method outperforms state-of-the-art unconditional face video generative models in multiple challenging datasets.**

I. Introduction

Human face synthesis draws great attention during the last decade with the evolution of image generation techniques, especially Generative Adversarial Network. It is challenging to generate a high resolution and artifact-free face image from noises without any other condition. Recently, progressive growing GANs [1] and StyleGAN [2] have breached the bottleneck of face image synthesis based on latent codes with massive annotated data and large computational power. However, when generating a face video, it is much more challenging because of the temporal coherence of the face identity. Existing conditional face generation techniques cannot be simply extended to video due to the lack of exemplars or face landmarks.

Recent GAN-based video generation approaches [3], [4], [5] have made significant progress to conditionally predict a limited number of frames of motion or body poses. However, these methods can only produce reasonable results for a few consecutive frames, and the requirement of a given first frame prevents them from unconstrained video generation.

In this paper, we propose a face video synthesis framework given only noises without any other condition. The goal of our model is to generate realistic and identity-preserving human face videos in arbitrary length. However, generating a coherence face video has two main problems, face synthesis and inter-frame consistency. It is difficult to simultaneously handle these two problems. Therefore, instead of using an end-to-end network to solve the problem in one shot, we follow a divide-and-conquer strategy by separating the problem into two parts, face identity synthesis and rearrangement. We propose a cascaded network which contains three independent components. We first propose an identity-aware GAN (IA-GAN) to learn both the identity and appearance features of human faces. To force the IA-GAN learning face identity, we propose a face verifier to measure the similarity of the identities of two synthetic faces. Then a Face Coherence Network is proposed to rearrange the face candidates generated by IA-GAN. Interpolation Network is introduced to eliminate discontinuity between two adjacent frames and make the video smoother (example results are shown in Fig. 1). Extensive experiments have been conducted to prove the effectiveness of our proposed method on face video generation with arbitrary length. The major contributions of this paper are summarized as follows:

- We makes the first attempt to synthesize high-quality and diverse face videos directly from noises without any exemplar or landmark of faces. The whole face video generation framework is lightweight and can generate arbitrary length face video.
- We propose an Identity-aware GAN to generate faces while the identities can be controlled. A pre-trained face verifier is proposed to learn the mapping between latent

*Corresponding author.

Fig. 1. Consecutive frames of face videos generated by our method.

code and the face identities.

- We construct a face video by proposing a face coherence network. It solves face video synthesis in a divide-and-conquer manner, and provides an alternative for general video synthesis problem.

II. Literature Reviews

A. Face Generation

Face synthesis is one of the main focus in image synthesis. [6] focused on the generation of non-frontal views from a single portrait using CNN. [7] propose a CNN model to reconstruct the face geometry directly from a 2D facial image. Some researchers propose some GAN variations to synthesize fake faces, which are almost indistinguishable against real faces. [8] disentangle the attributes and identity to recombine them for identity preserving face synthesis. The progressive growing of GANs [1] is able to generate high resolution images with high quality by progressively training the network.

However, these existing methods impose certain restrictions on the training data [7] or need prior knowledge of the face distribution [9], such as 3D templates. Furthermore, the face coherence and identity are also the major challenges for face video generation other than single face image generation. In contrast, we propose an identity-preserving framework without face masking or using template, and does not require any attribute annotations.

B. Video Generation

It is almost impossible for a single end-to-end model to generate videos with realistic faces while keeping vivid and coherent motion dynamics. To tackle this problem, researchers introduces prior knowledge or decompose the video generation into several sub-tasks. A Multi-stage Dynamic GAN [5] is proposed to generate realistic time-lapse videos by dividing the video generating process into content and motion modeling. However, in previous works [10], [5], the spatio-temporal consistency of the generated videos may be limitedly verified by the discriminator which could not extend in time. This is because concatenating two images or using 3D convolutions of fixed size could hardly model video sequences in variable length, nor generate long video sequences. Recently, [11] aims to directly generate face video sequences with a single image and 3D face landmark model. However, it requires several constraints such as a face exemplar and high quality face landmarks for face motion generation. On the contrary, our proposed method can generate identity preserved human face videos in arbitrary length without any exemplar and landmark, and it costs less GPU memory than using 3D convolutional operators.

III. Our Approach

In this paper, we propose a face video synthesis framework, as shown in Fig. 2. Our proposed network consists of three cascaded components. The Identity-aware GAN (IA-GAN) in Sec. III-A synthesizes realistic faces given a set of noises. The Face Coherence Network introduced in Sec. III-B rearranges the face results which are synthesized from IA-GAN while keeping the inter-frame coherence. The Interpolation Network is introduced to make the video smoother. The face video synthesis pipeline is demonstrated in Sec. III-D.

A. Identity-Aware GAN

We propose an Identity-aware GAN (IA-GAN) to generate a set of face candidates for a face video by only giving the network a set of noise vectors $\mathcal{Z} = \{Z_1, Z_2, ..., Z_n\}$. Each noise vector Z_i is concatenated by two vectors, an identity vector z_i^{id}, which is used to control the face identity, and an appearance vector z_i^{app}, which represents the face appearance. IA-GAN takes the noises $\mathcal{Z}$ as inputs and generates a set of faces. As shown in Fig. 3, IA-GAN consists of three parts, a generator G for face generation, a discriminator D for distinguishing real images from the fake, and a face verifier V for measuring the similarity of the face identities.

Inspired by the Energy-based GAN [12] and Boundary Equilibrium GAN [13], the generator $G : \mathbb{R}^{N_Z} \rightarrow \mathbb{R}^{N_x}$ is designed as a decoder since it aims to generate a face x from a latent code Z. The discriminator $D : \mathbb{R}^{N_x} \rightarrow \mathbb{R}^{N_Z} \rightarrow \mathbb{R}^{N_x}$ is modeled as an auto-encoder. The generator G shares the same network structure with the decoder part in the discriminator D. The Boundary Equilibrium GAN [13] objective is defined as follows:

$$\begin{cases} \mathcal{L}_D = \mathcal{L}(x;\theta_D) - k_t\mathcal{L}(G(Z_i;\theta_G);\theta_D) \\ \mathcal{L}_G = \mathcal{L}(G(Z_i;\theta_G);\theta_D) \\ k_{t+1} = k_t + \lambda_k(\gamma\mathcal{L}(x;\theta_D) - \mathcal{L}(G(Z_i);\theta_G);\theta_D), \end{cases} \tag{1}$$

where θ_D and θ_G are the parameters in the discriminator D and the generator G. $\mathcal{L}$ is the auto-encoder loss. The variable $k_t \in [0, 1]$ is used to control the importance of $\mathcal{L}(G(Z_i;\theta_G);\theta_D)$ during gradient descent in training step t. For easier training of GAN, $\gamma \in [0, 1]$ is an equilibrium term to balance two competing goals of discriminator, auto-encodes real images and discriminate the real face from the fake one. Here, lower γ value leads to lower image diversity, because the discriminator pays more attention to the auto-encoding of

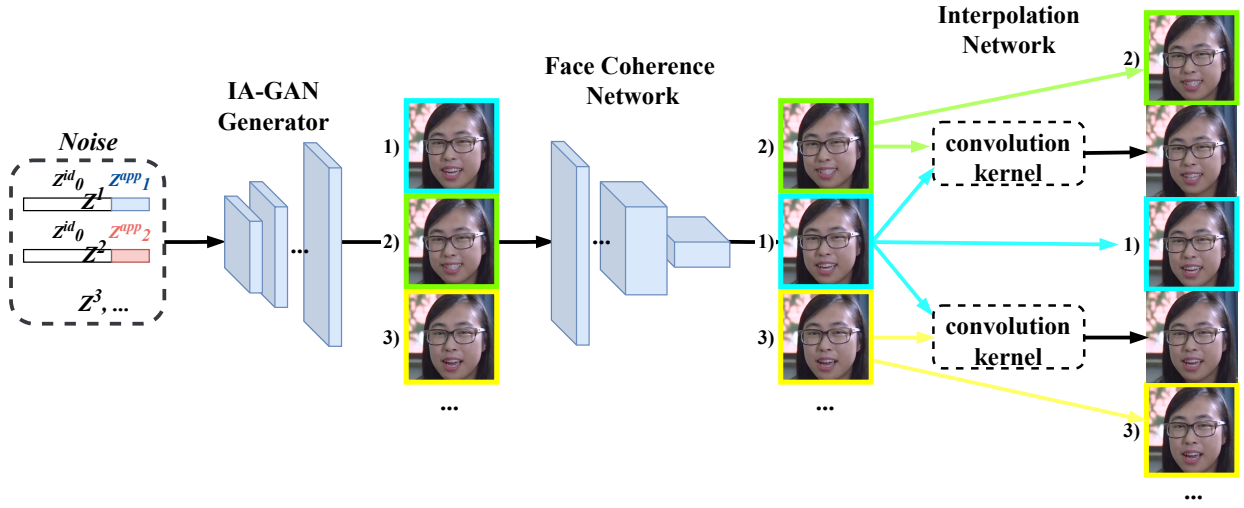

Fig. 2. System overview. Our proposed network takes a sequence of noises $Z\_list = \{Z_G^1, Z_G^2, ..., Z_G^n\}$ as input, then generates realistic face video. The whole network consists of two components which serve for specific purposes. The Identity-aware GAN (IA-GAN) synthesizes realistic faces. The Face Coherence Network re-arranges image sequences with regard to the motion dynamics of the adjacent frames. The Video Interpolation Network serves for frame interpolation.

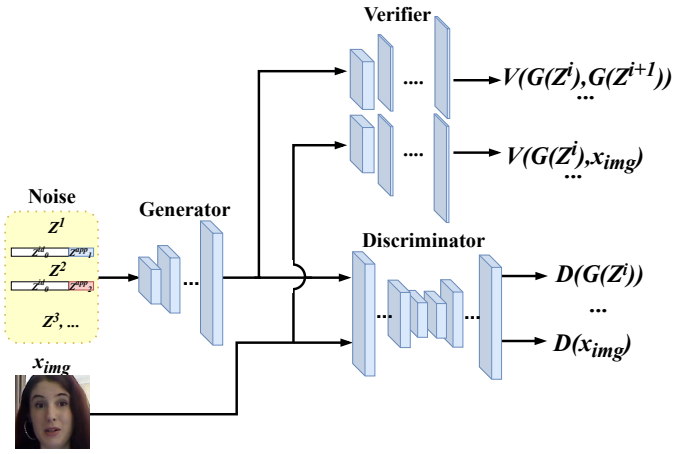

Fig. 3. The structure for IA-GAN.

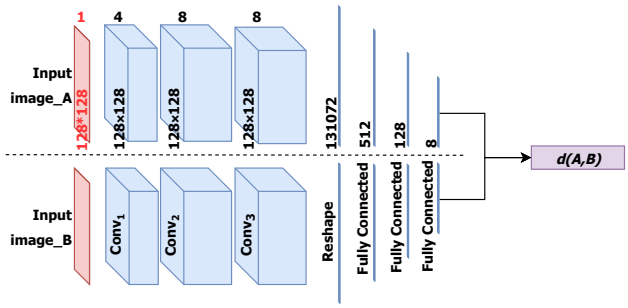

Fig. 4. Network architecture for the face verifier V. The Siamese networks is designed to measure the similarity of two faces. These two network branches share both the same architecture and weights.

the real images. We initialize $k_0 = 0$ and γ_k is the proportional gain for k.

To further improve the image quality and suppress the artifacts, perceptual loss is introduced in IA-GAN:

$$\mathcal{L}_{\text{vgg}} = \sum_{i=1}^{N} \frac{1}{M_i} ||f_i(x) - f_i(G(Z_i))||_1, \tag{2}$$

where N is the total number of layers, f_i and M_i denotes the layer i and the number of elements in f_i.

While keeping the diversity of the face appearances, we have to guarantee the temporal coherence of the face identity at the same time. Therefore, we embed an additional face verifier V into IA-GAN in order to measure the similarity of the identities between two faces. Such that it can force the IA-GAN to learn a better mapping of the face identity in the latent space. The face verifier V is constructed as a Siamese Network [14], as shown in Fig. 4. It is utilized to guarantee the identity of two faces to be the same, no matter the faces are real or fake. We pre-trained the face verifier by minimizing the contrastive loss function which is defined as follows:

$$\mathcal{L}_V = (Y - \frac{1}{2})(d(f_V(x_A), f_V(x_B))^2, \tag{3}$$

where $f_V(\cdot)$ denotes the identity features from an single branch of the Siamese Network V. Y denotes a label whether the two faces x_A and x_B are of the same identity. $d(\cdot, \cdot)$ is the pairwise distance, which is L_2 norm in our paper.

With the pre-trained face verifier V, we proposed an identity loss $\mathcal{L}_{\text{id}}$ for IA-GAN to force the generator G to generate faces which have similar feature representations in network V. We measure the identities between fake and real faces as well as two adjacent synthetic faces, shown as follows:

$$\begin{aligned}\mathcal{L}_{\text{id}} =& ||f_V(x) - f_V(G(Z_i))||_2 + \\ & \omega ||f_V(G(Z_{i-1})) - f_V(G(Z_i))||_2,\end{aligned} \tag{4}$$

IA-GAN Training Procedure: The network training includes two stages. In the first stage, we pre-trained IA-GAN and V separately in order to obtain a robust face generator (IA-GAN) and a more precise face verifier V. IA-GAN is trained by the following objective function:

$$\mathcal{L}_{\mathrm{IA-GAN}} = \mathcal{L}_{\mathrm{GAN}} + \lambda_{\mathrm{vgg}}\mathcal{L}_{\mathrm{vgg}}, \tag{5}$$

where λ_{vgg} is the weight of $\mathcal{L}_{\mathrm{vgg}}$. By minimizing the GAN loss $\mathcal{L}_{\mathrm{GAN}}$ and the perceptual loss $\mathcal{L}_{\mathrm{vgg}}$, IA-GAN is able to synthesize realistic faces with high quality and low artifacts. Then we pre-trained face verifier V by minimizing the contrastive loss $\mathcal{L}_V$.

In the second stage, we integrate face verifier V into the IA-GAN as an additional identity constraint to make sure same identity vectors can generate faces with same identity. So we fine-tune the IA-GAN by the following objective function:

$$\mathcal{L}_{\mathrm{IA-GAN}} = \mathcal{L}_{\mathrm{GAN}} + \lambda_{\mathrm{vgg}}\mathcal{L}_{\mathrm{vgg}} + \lambda_{\mathrm{id}}\mathcal{L}_{\mathrm{id}}. \tag{6}$$

Inspired by the progressively growing GANs [1], we introduce a progressive training strategy in the second stage by keeping increasing the number of input noise vectors n, starting from $n = 2$. As n increases, the generator G is imposed to establish the mapping between the face identities and the identity vectors. So that our model is expected to generate more diverse but identity-preserved faces. As a result, our video synthesis will be more robust on identity preservation.

B. Face Coherence Network

Given a set of synthetic faces from IA-GAN, to generate a face video, we have to guarantee the consistency of the face identity in temporal domain. We propose a Face Coherence Network to show the confidence of two frames to be consecutive frames according to their motion dynamics. In specific, we use a random pair of continuous frames as positive samples (x_A, x_{A+1}) and a random pair of discontinuous frames as negative samples (x_A, x_B). Considering the negative samples may have more than one scenario, we further extend the negative cases to three categories, two discontinuous frames (x_A, x_B), two frames with difference identities (x_A, x'_A), and two faces generated by distinct random noises $(G(Z_i), G(Z_{i'}))$ with different identity vector z^{id}. We assign weights to balance the contribution of each negative case.

$$\begin{aligned}\min_{\Theta_{\mathcal{S}}} \mathcal{L}_{\mathrm{coherence}} = &- \mathcal{S}(x_A, x_{A+1}) + \lambda_1\mathcal{S}(x_A, x_A{}') \\ &+ \lambda_2\mathcal{S}(x_A, x_B) + \lambda_3\mathcal{S}(G(Z_i), G(Z_{i'})).\end{aligned} \tag{7}$$

By minimizing $\mathcal{L}_{\mathrm{coherence}}$, the Face Coherence Network is trained to predict the confidence score of whether two frames can be consecutive frames. Given the first frame of the video, we greedily select the next frames with the highest score.

The architecture of Face Coherence Network is shown in Fig. 5. We concatenate two face images and feed it into the network. The network keeps downsampling and finally output a confidence score. A Convolutional Gate Recurrent Unit (Conv-GRU) layer is introduced to this network to continuously maintain the dynamic fidelity over time compared with the fixed size 3D convolutional layers and save memory usage.

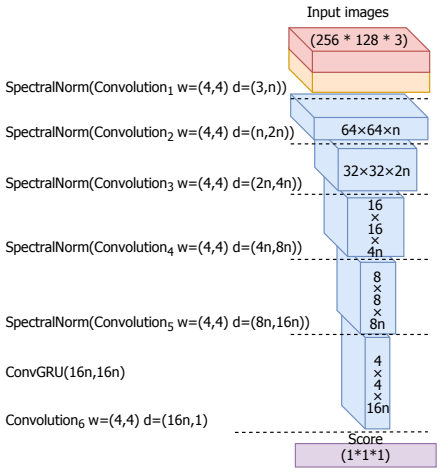

Fig. 5. Network architecture of our Face Coherence Network C. We concatenate two images and feed them into the network. n is set to 64 in our implementation.

C. Video Interpolation Network

One issue of our previous structure is that all the frames generated by IA-GAN are from random noises. Even if we proposed a Face Coherence Network to re-arrange the frames, there may still be some discontinuity between adjacent frames. Therefore, we apply a video frame interpolation method [15] to obtain a smoother face video result. In this method, a CNN is used to estimate the spatially-adaptive convolution kernel of each pixel. These kernels are used to capture motion as well as interpolation coefficients, and convolve directly with the input image to synthesize intermediate video frames.

D. Face Video Synthesis Pipeline

For face video synthesis, we initialize the input noises $\mathcal{Z} = \{Z_1, Z_2, ..., Z_n\}$ with the same identity vector z^{id} but different appearance vectors z^{app} and feed into the generator G of IA-GAN one by one, to synthesize a list of candidate face images $\mathcal{F} = \{f_1, f_2, ..., f_n\}$. We randomly choose one image from $\mathcal{F}$ as the first frame of the video, say, f_i. Then, feed the rest of $n-1$ images into the Face Coherence Network C for the confidence score. The one with the highest confidence score is the second frame. Then we keep choosing the best next frame until we get enough frames $\{\hat{f}_1, \hat{f}_2, \hat{f}_3, ..., \hat{f}_m\}$ for the video. We set m is ten times smaller than n since we want more candidate faces with more diversity. After that, the interpolation network is applied to obtain a longer and smoother face video.

IV. Experiments

A. Datasets

We use five datasets to train our network. **CelebA face image dataset** [16] and **Deep-funneled face image dataset**

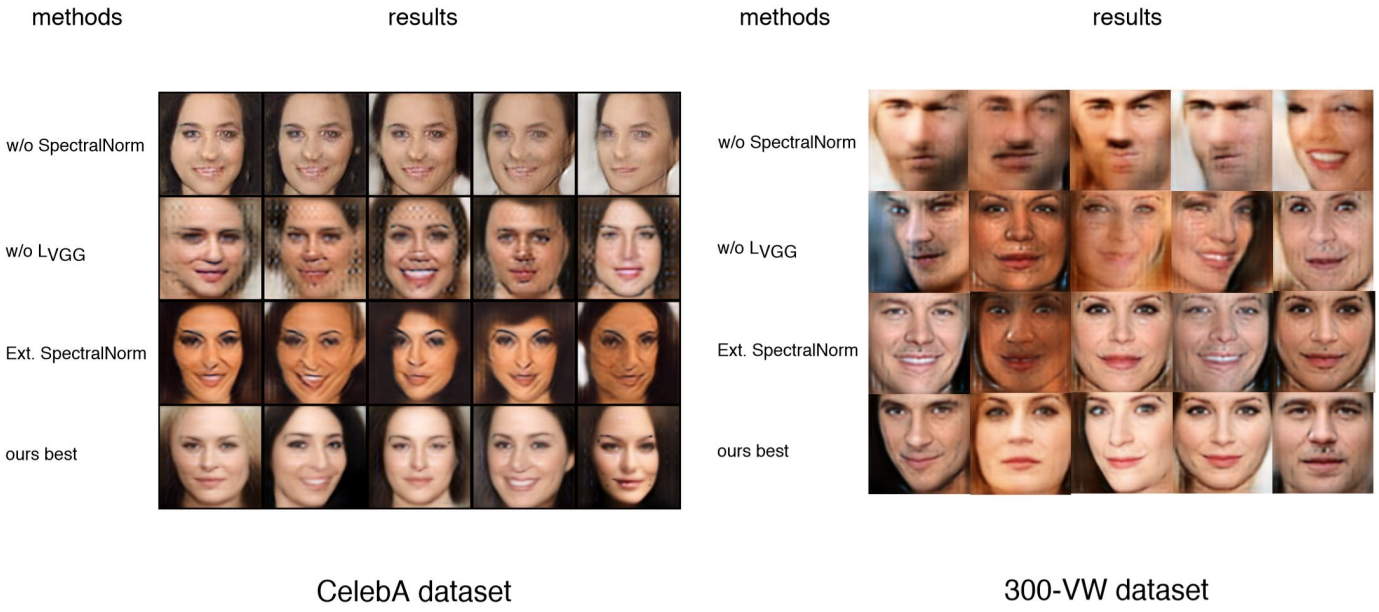

Fig. 6. Qualitative comparison between generated images from network G using different training strategies and methods. The generators are trained with two different datasets with $n_f = 1$

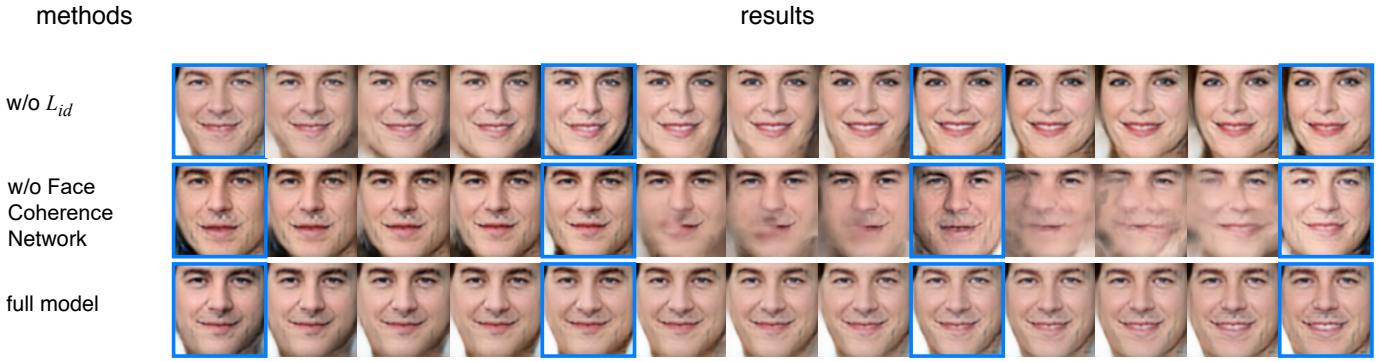

Fig. 7. Qualitative comparison of generated video on difference models. The faces in the blue box are the ones generated by IA-GAN. The others are the interpolation results. We interpolated three frames between every two consecutive frames.

LFW [17] are used separately and independently for training IA-GAN as well as for testing the effectiveness and robustness of our improvement. **Aligned FaceScrub dataset** [18] is with over a million faces of 530 people. We use this dataset to train the face verifier and the IA-GAN, to learn identity information. **In-the-wild video dataset 300-VW** [19] includes 114 videos with a total of 218,595 frames. The duration of each video is around 1 min at 25-30 fps. **Talking Face video dataset** [20] consists of 5000 frames of a person engaged in conversation. The two datasets above are used to train Face Coherence Network.

We have built a relatively large time-lapse video dataset from the Internet. The last two datasets are filtered and clipped, and the appropriate number of frames is skipped when intercepting, so that the inter-frame changes will not be too subtle. At the same time, we collected over 20 different people's daily conversation and speech videos, detected the facial region with a face recognition library [21], cropped and aligned them. And then, the dataset for training Face Coherence Network is obtained by merging the above three data sources through manual screening. Finally, we split the videos into segments. There is no overlap between two consecutive segments. The total number of frames per segment is more than 45 and less than 100, in order to effectively reduce mode collapse during training. We collect 684 training video segments, and totally 43,261 frames, each containing 63.25 frames on average. Before providing the segments to the Face Coherence Network, we normalize the color values to $[-1, 1]$. A random affine transformation is carried out as preprocessing.

B. Ablation Studies

We conduct the ablation studies to evaluate the effectiveness of our proposed method. This experiment is divided into two parts, one for evaluating the single image quality, another for

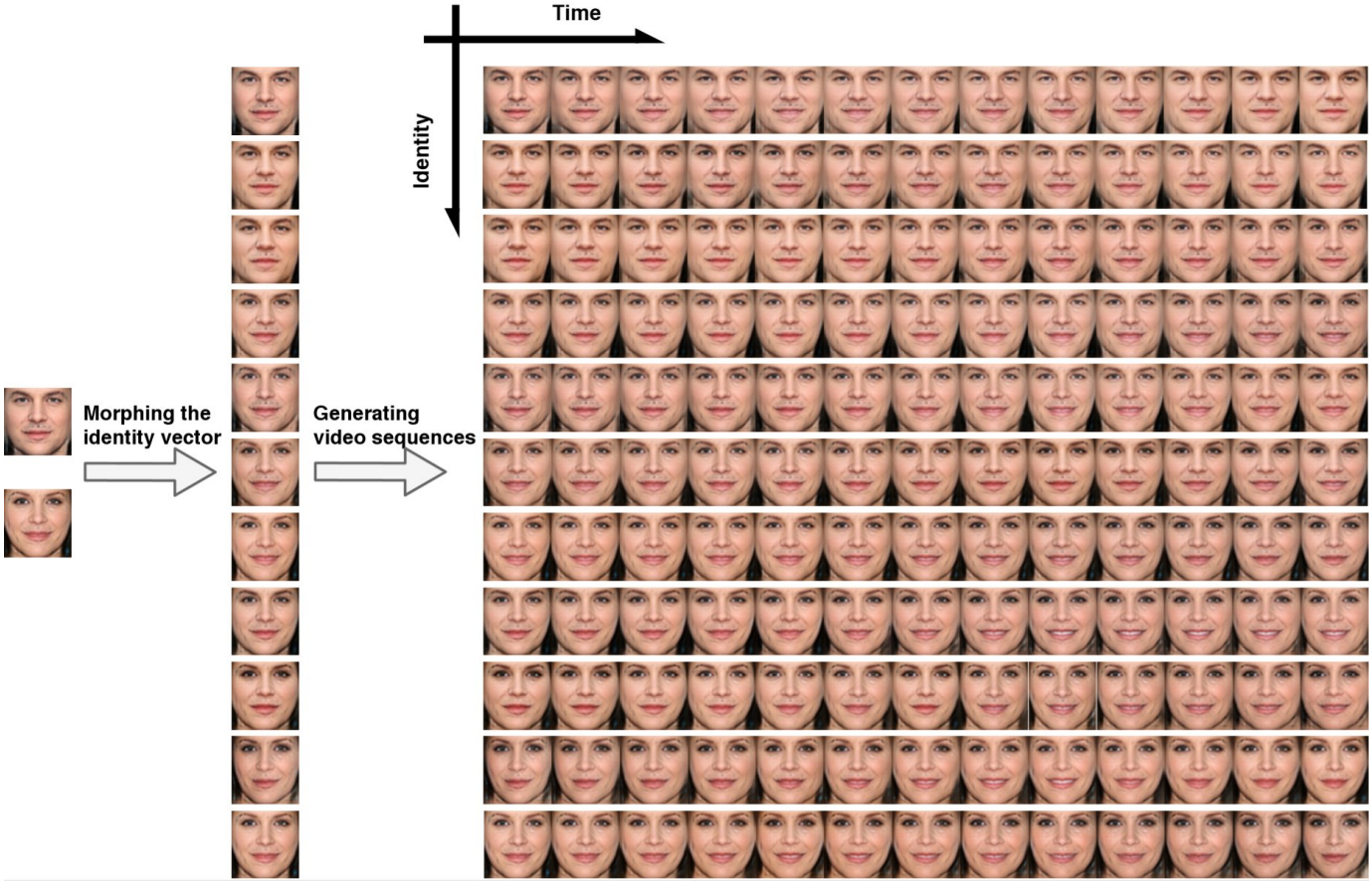

Fig. 8. By interpolating between the identity vectors of two faces (a man and a woman), we synthesize a series of images that continuously changing the facial features from a male subject with wider faces to a female with bigger eyes and thinner eyebrows. By modifying the expression vectors, we generate video sequences of each morphed identity.

the video quality. We compared the image quality of our complete model against three variants: 1) without SpectralNorm layers in discriminator D; 2) IA-GAN without perceptual loss $\mathcal{L}_{\text{vgg}}$; 3) extra SpectralNorm layers in decoder part of D. We compared the video quality of our complete model against two variants: 1) without the face verifier V and $\mathcal{L}_{\text{id}}$; 2) without Face Coherence Network. Note that, all the other training strategy and experiment setup remain the same for each variant.

To compare the image quality, we carry out two experiments on 300-VW dataset and CelebA dataset, respectively. Figure. 6 shows the qualitative comparison of the image quality. We can observe that removing SpectralNorm layer during the training process causes mode collapse and poor learning of facial features and details. The Removal of perceptual loss introduces large among of artifacts and the boundaries are not sharp enough. If a SpectralNorm layer is added into the decoder redundantly, the generated samples will be distorted and the facial area will be corrupted. Our complete model yields the best results, which can generate more realistic images with the least artifacts.

We qualitatively compare the face videos on 300-VW dataset, as shown in Fig. 7. We can observe that without the face verifier during the training process will make it difficult to maintain the face identity. For example, as shown in the first row of Fig. 7, the face identity gradually changes from a man's facial feature to a woman's. Without Face Coherence Network, temporal coherence of the face are no longer maintained, such as the results in the blue box in Fig. 7. In addiction, since the difference between two consecutive frames is large without Face Coherence Network. It also drastically increase the difficulties of the interpolation. Our complete generation pipeline achieved the best video results with the best face coherence.

C. Visualization of the Identity Latent Space in IA-GAN

We conduct an experiment to visualize latent space of the face identities by morphing one face identities vector to another. First, we randomly generate two identity vectors z_1^{id} and z_2^{id}, together with two lists of appearance vectors. Then we get two set of noises $\mathcal{Z}^1 = [Z_0^1, Z_1^1, Z_2^1, ...Z_N^1]$ and $\mathcal{Z}^2 = [Z_0^2, Z_1^2, Z_2^2, ...Z_N^2]$. Second, we obtain a series of identity vectors by the linear interpolation, i.e., $z_i = w_i z_1^{\text{id}} + (1 - w_i) z_2^{\text{id}}$, $w \in [0, 1]$. Note that, when interpolating continuously between the identity vectors, appearance vectors are also interpolated. Then we obtain a series of noise vectors

TABLE I
QUANTITATIVE EVALUATION OF FACE VERIFIER

Threshold	0.6	0.7	0.8	0.9
Accuracy	94.1%	95.8%	97.1%	97.3%

$\{\mathcal{Z}^1, \mathcal{Z}^2, ..., \mathcal{Z}^M\}$ Fig. 8 displays the results generated by the above noises. The faces from the top row to the bottom row are generated from $\mathcal{Z}^1$ to $\mathcal{Z}^M$. Each row is the results generated by our model.

When we scan through the faces vertically, we can observe that the face is gradually change from the top to the bottom. When we change the identity of the synthetic videos, identity of each video is well preserved. For example, the smaller face and the gradual disappearance of the beard are believable, showing good generalization. These smooth semantic changes indicate that the model has learned essential identity representations for face synthesis. Also, similar to [22], the results of walking on the latent space should be sufficient to indicate that there is hardly any hierarchically collapse.

D. Qualitative Evaluation of Face Verifier

To obtain a good face video result, we have to guarantee the face identity while generating the faces. Therefore, we conduct an experiment to qualitatively evaluate the effectiveness of our proposed face verifier on FaceScrub dataset. We randomly select two faces from the dataset, no matter they are from the same person or not. We let the face verifier to tell whether these two faces are from the same person. Since we design the face verifier as a Siamese Network. So the network can only output the similarity score of two faces. The similarity is defined as the pairwise distance between model outputs. To calculate pairwise distance, we apply $\mathcal{L}_2$ norm. A smaller value of the distance d indicates more similar of two faces. Then, the pairwise distance is compared to a threshold. If the distance is smaller than the threshold, they are predicted to have the same identity, and vice versa. To evaluate the performance of our model, we gradually increase the threshold $[0.6, 0.7, 0.8, 0.9]$ to loose the similarity baseline. The accuracy of the face verifier is reported in Table. I. We can observe that when the threshold looses/increases, the face verifier accuracy keep increasing. When the threshold equals 0.6, the face verifier can still show high precision at 94.1%. It demonstrates the effectiveness of the face verifier on distinguishing the face identity.

E. Comparison with Other Methods

We are the first one trying to generate face video from noises. There is no existing work to directly compare with. Therefore, we conduct an quantitative experiment to compare our method with a combination of BEGAN [13] and facemorpher [23], named BEGAN-morpher. For face video generation using BEGAN-morpher, we first utilize BEGAN to generate a set of face candidates. To be fair, we also apply our proposed Face Coherence Network on the face candidates generated by BEGAN to maintain the temporal coherence. After that, we apply facemorpher to interpolate frames for every two consecutive frames and generate a face video. The reason why we do not apply the same interpolation method of ours is because BEGAN is not able to control the identity of the synthetic faces. To directly apply the method of Niklaus et al. [15] may introduce a large among of artifacts. Besides BEGAN-morpher, we also compare our method with our model without progressive training strategy (ours w/o PT) and the real face videos.

TABLE II
QUANTITATIVE COMPARISON ON IDENTITY-PRESERVING ABILITY

Method	$n_f = 30$	$n_f = 45$	$n_f = 60$	$n_f = 75$
Ours (with PT)	**98.0%**	**97.8%**	**97.5%**	**97.4%**
Ours (w/o PT)	94.1%	93.5%	93.2%	93.0%
BEGAN-morpher	92.1%	91.0%	90.9%	86.8%
Real Videos	100.0%	100.0%	100.0%	100.0%

TABLE III
USER STUDY RESULTS OF DIFFERENT METHODS

Standards	Methods			
	Real	Ours	MoCo-GAN	BEGAN-morpher
Face Identity	4.677	3.442	**4.012**	2.942
Face Coherence	4.498	**3.634**	3.401	2.909
Face Angle Diversity	3.566	**3.457**	2.459	2.781
Video Quality	4.421	**3.357**	2.862	2.992
Overall Preference	4.392	**3.469**	3.124	2.842

We conduct this quantitative experiment on 300-VW dataset, where the faces are cropped and resized to 128×128. We then measure the identity-preserving ability by comparing the identity of the frames using a face recognition technique [24]. The comparison is carried out over 100 generated videos for each method. For each video, we collect the face pairs by randomly selecting two frames with a fix interval n_f. The score is the percentage of the correctness that these two frames are successfully recognized as the same person. The final scores are as shown in Table. II. The identity-preservation scores show that our proposed method achieves better performance than BEGAN-morpher and ours (w/o PT). The face identity of BEGAN-morpher cannot be retained with the increase of frame interval. That is because they do not have a specific identity vector to control the face identity. Without progressive training strategy, the performance decrease drastically. Our complete model preserves identity well for arbitrary-length video generation.

F. User Study

We also conduct a user study on Amazon Mechanical Turk (AMT) to further evaluate the visual preference of our results. We compare our method with BEGAN-morpher, MoCoGAN [25] (a state-of-the-art video generation network) and the real face videos. Note that, since we cannot train a better model of MoCoGAN on the face dataset we used. The

results of MoCoGAN for comparison are directly from their papers and official implementation. The resolution of images generated by MoCoGAN is 96×96 while others are 128×128. Moreover, the video length of MoCoGAN is 48 frames per video, while others generate more than 100 frames on average.

We randomly select 20 generated video clips for each model in random order. For each group of results, users were asked to evaluate the results in five aspects: face identity, face coherence, face angle diversity, video quality and the overall preference. The score range is $[1, 5]$. Total 22 participants joint this test and the user study results are shown in Table. III. The user study results show that our method outperforms other models most of the times, except face identity compared with MoCoGAN. That is because MoCoGAN focuses on the facial expression and the direction of the faces and illumination are almost unchanged, while other methods aim at extracting a longer and variable-length video including more information closer to real videos.

V. Conclusion

In this paper, we propose a face video generation framework, which can synthesize arbitrary-length face videos without any exemplar or landmark of faces. We address face video generation with a divide-and-conquer strategy, solving the synthesis and rearrangement sub-task individually. High quality and coherent face videos can be generated with our lightweight network. We make the first attempt to generate face video from noises, and experimental evaluations demonstrate the proposed method outperforms the direct extension of state-of-the-art face image generators.

References

[1] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," CoRR, vol. abs/1710.10196, 2017.

[2] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," arXiv preprint arXiv:1812.04948, 2018.

[3] W. J. Baddar, G. Gu, S. Lee, and Y. M. Ro, "Dynamics transfer gan: Generating video by transferring arbitrary temporal dynamics from a source video to a single target image," CoRR, vol. abs/1712.03534, 2017.

[4] B. Kratzwald, Z. Huang, D. P. Paudel, A. Dinesh, and L. V. Gool, "Improving video generation for multi-functional applications," arXiv preprint arXiv:1711.11453, 2017.

[5] W. Xiong, W. Luo, L. Ma, W. Liu, and J. Luo, "Learning to generate time-lapse videos using multi-stage dynamic generative adversarial networks," in CVPR, 2018.

[6] X. Peng, X. Yu, K. Sohn, D. N. Metaxas, and M. K. Chandraker, "Reconstruction for feature disentanglement in pose-invariant face recognition," CoRR, vol. abs/1702.03041, 2017.

[7] E. Richardson, M. Sela, and R. Kimmel, "3d face reconstruction by learning from synthetic data," 2016 Fourth International Conference on 3D Vision (3DV), pp. 460–469, 2016.

[8] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua, "Towards open-set identity preserving face synthesis," in CVPR, 2018.

[9] Y. Shen, P. Luo, J. Yan, X. Wang, and X. Tang, "Faceid-gan: Learning a symmetry three-player gan for identity-preserving face synthesis," in CVPR, 2018.

[10] M. Saito, E. Matsumoto, and S. Saito, "Temporal generative adversarial nets with singular value clipping," ICCV, pp. 2849–2858, 2017.

[11] K. Songsri-in and S. Zafeiriou, "Face video generation from a single image and landmarks," arXiv preprint arXiv:1904.11521, 2019.

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

[13] D. Berthelot, T. Schumm, and L. Metz, "Began: Boundary equilibrium generative adversarial networks," CoRR, vol. abs/1703.10717, 2017.

[14] G. R. Koch, "Siamese neural networks for one-shot image recognition," in ICML Deep Learning Workshop, vol. 2, 2015.

[15] S. Niklaus, L. Mai, and F. Liu, "Video frame interpolation via adaptive separable convolution," ICCV, pp. 261–270, 2017.

[16] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in ICCV, 2015.

[17] G. B. Huang, M. Mattar, H. Lee, and E. Learned-Miller, "Learning to align from scratch," in NIPS, 2012.

[18] H. Ng and S. Winkler, "A data-driven approach to cleaning large face datasets," in ICIP, Oct 2014, pp. 343–347.

[19] G. G. Chrysos, E. Antonakos, S. Zafeiriou, and P. Snape, "Offline deformable face tracking in arbitrary videos," in ICCVW, Dec 2015, pp. 954–962.

[20] T. F. Cootes, "Talking face video," http://personalpages.manchester.ac.uk/staff/timothy.f.cootes/data/talking_face/talking_face.html, Online; accessed June 2, 2016.

[21] A. Geitgey, "Face recognition," https://github.com/ageitgey/face_recognition/, Online; accessed June 2, 2018.

[22] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," Computer Science, 2015.

[23] A. Quek, "Face morpher," https://github.com/alyssaq/face_morpher, Online; accessed June 2, 2018.

[24] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015.

[25] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz, "MoCoGAN: Decomposing motion and content for video generation," in CVPR, 2018, pp. 1526–1535.