StyleFaceV: Face Video Generation via Decomposing and Recomposing Pretrained StyleGAN3

Haonan Qiu,¹ Yuming Jiang,¹ Hang Zhou,² Wayne Wu,³ Ziwei Liu¹

¹S-Lab, Nanyang Technological University, ²The Chinese University of Hong Kong, ³SenseTime Research

{haonan002, yuming002, ziwei.liu}@ntu.edu.sg, zhouhang@link.cuhk.edu.hk, wuwenyan0503@gmail.com

Abstract

Realistic generative face video synthesis has long been a pursuit in both computer vision and graphics community. However, existing face video generation methods tend to produce low-quality frames with drifted facial identities and unnatural movements. To tackle these challenges, we propose a principled framework named StyleFaceV, which produces high-fidelity identity-preserving face videos with vivid movements. Our core insight is to decompose appearance and pose information and recompose them in the latent space of StyleGAN3 to produce stable and dynamic results. Specifically, StyleGAN3 provides strong priors for high-fidelity facial image generation, but the latent space is intrinsically entangled. By carefully examining its latent properties, we propose our decomposition and recomposition designs which allow for the disentangled combination of facial appearance and movements. Moreover, a temporal-dependent model is built upon the decomposed latent features, and samples reasonable sequences of motions that are capable of generating realistic and temporally coherent face videos. Particularly, our pipeline is trained with a joint training strategy on both static images and high-quality video data, which is of higher data efficiency. Extensive experiments demonstrate that our framework achieves state-of-the-art face video generation results both qualitatively and quantitatively. Notably, StyleFaceV is capable of generating realistic $1024 \times 1024$ face videos even without high-resolution training videos.

Introduction

Face generation has been a long-standing research topic in vision and graphics. As humans perceive the world from a video-point of view, the exploration of video face generation is of great importance. Despite the great progress (Goodfellow et al. 2014; Karras, Laine, and Aila 2019; Karras et al. 2020b) in generating static face images, the task of face video generation is less explored due to the extra complexity introduced by the temporal dimension. Compared to a face image, we expect a generated face video to possess three desired properties: 1) High fidelity. The face of each frame should be photo-realistic. 2) Identity preservation. The same face identity should be kept among frames in the generated video. 3) Vivid dynamics. A natural face video is supposed
to contain a reasonable sequence of motions.

However, current face video generation methods still struggle to generate visually satisfying videos that meet all desired properties above. Most previous works design specific end-to-end generative models (Tian et al. 2021; Tulyakov et al. 2018; Clark, Donahue, and Simonyan 2019; Skorokhodov, Tulyakov, and Elhoseiny 2021; Yu et al. 2022) that target directly at video generation. Such designs are expected to achieve both spatial quality and temporal coherence in a single framework trained from scratch, imposing great burden to the network. In another fashion, some recent works (Fox et al. 2021; Tian et al. 2021) synthesize videos by employing a pre-trained StyleGAN and finding the corresponding sequence of latent codes. But those methods generate videos by manipulating the entangled latent codes thus making identity easily change with head movements.

To address these challenges, we propose a new framework, named StyleFaceV with three desired properties: photo-realism, identity preservation and vivid dynamics. In Fig. [1] we show some videos generated by our framework. The two videos in each row share the same face identity but with different facial movements. Our framework uses the StyleGAN3 (Karras et al. 2021) to synthesize a sequence of face frames to obtain the final generated video. With the powerful image generator, the key to generate a realistic video is: How can we generate a temporally coherent video with the same identity by traversing through the latent space of StyleGAN3? Our core insight is to decompose the latent space of the StyleGAN3 into the appearance and pose information via a decomposition and recomposition pipeline.

The decomposition module extracts pose and appearance information from the images synthesized by StyleGAN3. We design two extraction networks to extract pose features and appearance features, respectively. With the decomposed features, the recomposition module is employed to fuse them back to the latent codes as the inputs to StyleGAN3. The fused latent codes are then used to generate the facial images. With the decomposition and recomposition pipeline, we can easily sample a well disentangled sequence of latent codes with the same identity by sampling the pose information while keeping the appearance information fixed. The sampling of pose sequences is achieved by an additional LSTM (Hochreiter and Schmidhuber 1997) module.

To train the decomposition and recomposition pipeline, we need faces with the same identity but different poses. Such paired data can be easily obtained from face videos. However, large-scale and high-quality video datasets are lacking. To alleviate the problems caused by the lack of high-quality videos, we propose the joint training strategy, which utilizes the rich image priors captured in the well-pretrained StyleGAN3 model as well as the limited video data. Specifically, we do the self augmentation by synthesizing a new frame with a randomly translated and rotated face through the affine layer embedded in StyleGAN3 to simulate paired frames of face videos. The use of synthetic data eases the requirements on the video dataset. Notably, this joint training design enables our model to generate realistic 1024 × 1024 face videos even without accessing high-resolution training videos.

In summary, our main contributions include:

- We propose a novel framework StyleFaceV which decomposes and recomposes pre-trained StyleGAN3 to generate new face videos via sampling decomposed pose sequences.
- We design a joint training strategy which utilizes information from both the image domain and video domain.
- Our work exhibits that latent space of StyleGAN3 contains a sequence of latent codes which is able to generate a temporally coherent face video with vivid dynamics.

Related Work

StyleGAN Models. With the success of Generative Adversarial Networks (GAN) (Goodfellow et al. 2014), generative models have become a popular technique to generate natural images. After that, various GAN architectures (Karras et al. 2017; Miyato et al. 2018; Brock, Donahue, and Simonyan 2018) are proposed to stabilize the training process and the quality of generated images is greatly improved. After StyleGAN series (Karras, Laine, and Aila 2019; Karras et al. 2020b,a) have achieved high-resolution generation and style controlling, (Karras et al. 2021) propose StyleGAN3 to solve the “texture sticking” problem, paving the way for generative models better suited for video and animation. Besides generating images from scratch, those pre-trained image generators are also useful for performing several image manipulation and restoration tasks by utilizing captured natural image priors, such as image colorization (Pan et al. 2021; Wu et al. 2021), super-resolution (Chan et al. 2021; Wang et al. 2021; Menon et al. 2020) and facial image editing (Shen et al. 2020; Patashnik et al. 2021; Jiang et al. 2021). In our proposed StyleFaceV framework, we also utilized a pre-trained StyleGAN3 to render photorealistic face videos.

Video Generation. Different from image generation which only requires sampling at the spatial level, video generation involves an additional sampling at temporal level. Many works of video synthesis have achieved impressive results, e.g., video-to-video translation with given segmentation masks (Wang et al. 2018; 2019) or human poses (Chan et al. 2019), and face reenactment with given both human identities and driven motions (Stiho et al. 2019; Zhou et al. 2021). However, in the unconditional setting, generating videos from scratch remains still unresolved. Early works (Tulyakov et al. 2018; Saito, Matsumoto, and Saito 2017; Hyun, Kim, and Heo 2021; Aich et al. 2020; Munoz et al. 2021; Saito et al. 2020) proposed to use a single framework to synthesize low-resolution videos by sampling from some noise. (Tulyakov et al. 2018) decompose noises in the content domain and motion domain. The content noise controls the content of the synthesized video, while motion noises handle the temporal consistency. Recently, MoCoGAN-HD (Tian et al. 2021), and StyleVideoGAN (Fox et al. 2021) scaled up the resolution of synthesized videos by employing pre-trained StyleGAN. MoCoGAN-HD and StyleVideoGAN proposed to sample a sequence of latent codes, which were then fed into pre-trained StyleGAN to synthesize a sequence of images. The pose infor-
Figure 2: Overview of Our Proposed StyleFaceV. (a) Our proposed StyleFaceV decomposes the image into an appearance representation \( f_a \) and a pose representation \( f_p \). Then we re-compose both representations to get the intermediate embedding \( w \) as the input of the synthesis network \( G \). (b) Besides video training data, joint training strategy also samples generated images for training to utilize rich image priors captured in the well pretrained StyleGAN3 model by sampling initial content noise \( z \). To simulate paired frames with different poses but the same appearance, self-augmentation is used to randomly transform \( z \) \( \in I \) as the input of the synthesis network \( G \). (c) To generate videos from scratch, an LSTM-based motion sampling module is trained to sample a sequence of pose representations \( \{ f_p(1), ..., f_p(n) \} \) from motion noise \( \epsilon \), which is then re-composed with a randomly sampled appearance representation \( f_a \) to generate a sequence of frames.

Pre-Trained Image Generator

The recently proposed StyleGAN3 \cite{Karras2021} does not suffer from the "texture sticking" problem and thus paves the way for generating videos by pre-trained image generators. Among several variants of StyleGAN3, only StyleGAN3-R is equivalent to the rotation transformations. Considering face rotations are commonly observed in talking-face videos, we adopt the pre-trained StyleGAN3-R as our image generator. It contains a mapping network \( M \) that maps the content noise \( z \) into an intermediate latent space \( W \), and a synthesis network \( G \) that synthesizes image \( x \in I \) from the intermediate embedding \( w \in W \):

\[
w = M(z), x = G(w).
\]  

We slightly fine-tune the generator on video dataset so that it can generate images from the distribution of video dataset.

In the unconditional setting, face video generation task aims to generate a realistic face video sequence \( v \in V \) from scratch. Regarding the generated image as a single frame, a sequence of generated images can form the final synthesized video \( v = \{x_1, x_2, ..., x_T\} \). Therefore, the rest problem is how to find a sequence of latent codes from the latent space of StyleGAN3 to generate a temporally coherent sequence of face images with the same identity.

Decomposition and Recomposition

Our framework is built on top of powerful pre-trained image generator StyleGAN3 \cite{Karras2021}, which provides high quality face frames. As shown in Fig. 2, a decomposition and recombination pipeline is used to find some sequences of latent codes from the latent space of StyleGAN3 to generate a temporally coherent sequence of face images with the same identity. Finally, the motion sampler models the distribution of natural movements and samples a sequence of poses to drive the movements of faces.

Our Approach

Our framework utilizes image priors captured in the pre-trained StyleGAN3 and synthesizes facial videos with diverse identities.
changing the identity, we propose a decomposition and re-
composition pipeline as shown in Fig. 2(a). In this pipeline,
we decompose a face image into an appearance representa-
tion \( f_a = E_{\text{app}}(x) \) and a pose representation \( f_p = E_{\text{pose}}(x) \). Then we can simply re-compose \( f_a \) and \( f_p \) to get the inter-
mediate embedding \( w \) as the input of the synthesis network \( G \) by the re-composition network \( F_{\text{fuse}} \):
\[
w = F_{\text{fuse}}(f_a, f_p).
\] (2)

This pipeline allows us to drive the face by changing the pose representation \( f_p \), while the identity is kept by freezing the appearance representation \( f_a \).

**Pose Extraction.** Pose representations in face videos mainly consist of two levels of information: 1) overall head movement, 2) local facial movement. Both of these two kinds of movements will be reflected in landmarks. Therefore, we use a pre-trained landmark detection model (Wang, Bo, and Fuxin 2019) to supervise the pose estimation. However, all predicted landmarks also contain the identity information, like face shape. To purify the pose information without re-
serving the identity information, our pose extractor \( E_{\text{pose}} \) are only trained to reserve information of selected key face points \( p^{\text{target}} \).
\[
f_p = E_{\text{pose}}(x),
\] (3)
\[
L_{\text{pose}} = \|E_{\text{kp}}(f_p) - p^{\text{target}}\|_2^2,
\] (4)
where \( E_{\text{kp}} \) is key face points predictor and \( p^{\text{target}} \) presents targeted key face points selected from predicted landmarks from the pre-trained landmark detection model. As shown in Fig. 3, points \( p^{\text{target}} \) only trace the position of facial attributes and their status (e.g., open/close for mouth). The number of key face points is significantly less than that of landmarks. In addition, more detailed pose information which is not obtained by landmarks will be supervised during the recon-
struction stage.

**Appearance Extraction.** Different from the pose represent-
tations, it is hard to directly define appearance representa-
tion. In this face video generation task, we assume appearance is not changed. Based on this assumption, if we sample two frames \( x_i \) and \( x_j \) from the same video, we can get:
\[
x_i^* = G(F_{\text{fuse}}(f_{a(i)}, f_{p(j)}))
\]
\[
= G(F_{\text{fuse}}(E_{\text{app}}(x_i), E_{\text{pose}}(x_j))).
\] (5)

In theory, \( E_{\text{app}} \) can be learned by making \( x_i^* \) close to \( x_j \). However, we find that networks can hardly converge if we train the framework together from scratch. It is mainly be-
cause \( E_{\text{app}} \) and \( E_{\text{pose}} \) disentangle with each other, and can not provide effective appearance/pose information if we random-
ly initialize the network. Therefore, we first train \( E_{\text{pose}} \) with the single loss in Eq. 4. After \( E_{\text{pose}} \) is pretrained, we train the whole framework together.

**Objective Functions.** To use the video dataset, each time we sample \( n \) frames \( x = \{x_i, ..., x_i+k \times (n-1)\} \) with interval \( k \) and predict each frame \( x \) through Eq. 5. Here we simply set \( k = 3 \) to make sampled frames have obvious pose differences. By training to reconstruct frames with the same appearance, \( E_{\text{app}} \) is guided to ignore the pose information while \( E_{\text{pose}} \) learns to extract the differences among frames, which are pose representations in this work. The overall function for training with video dataset is:
\[
L_{\text{overall}} = \lambda_1 L_1 + \lambda_{\text{per}} L_{\text{per}} + \lambda_{\text{gan}} L_{\text{gan}} + \lambda_{\text{pose}} L_{\text{pose}}.
\] (6)
where \( L_1 = \|x^* - x\|_1 \) and \( L_{\text{per}} = \|VGG16(x^*) - VGG16(x)\|_1 \) are constraints to guarantee the reconstruction quality in both pixel domain and perceptual feature domain extracted by VGG16 (Simonyan and Zisserman 2014). \( L_{\text{gan}} \) is an adversarial loss (Goodfellow et al. 2014) to enhance local details. \( L_{\text{pose}} \) is a pose training loss introduced in Eq. 4. \( \lambda \) is the weight for different losses when using video data. When the pipeline is able to do roughly good images reconstruction with decomposition, \( \lambda \) for training with sampled images will decrease to make pose representations bring a finer control, which will be further analyzed in ablation study.

**Joint Training Strategy.** A well-trained StyleGAN3 model contains rich image priors captured in the large-scale and high-quality face image datasets. Utilizing this feature, we can also sample identities from distribution of original image datasets. However, there is no shared appearance among sampled images and each image is the individual frame. To prevent the networks from being lazy to do only reconstruction without decomposing, we design the joint training strategy as shown in Fig. 2(b). We firstly do the self augmentation through the affine layer embedded in StyleGAN3-R to produce a new frame with a randomly translated and rotated face while its face appearance is kept and background is still natural. Then we use the original version to provide the pose information and the transformed variants to provide the appearance information.

Specifically, instead of sampling two frames \( x_i \) and \( x_j \) from the same video in Eq. 5, we use the equation below to get \( x_i \) and \( x_j \) from the StyleGAN3-R:
\[
x_i = G(w; (r, t_x, t_y)), x_j = G(w; (0, 0, 0)),
\] (7)
where \( r \) is the random rotation angle, and \( t_x \), \( t_y \) are the random translation parameters.

In addition, we add a self-supervised embedding loss function \( L_w \) to help the convergence of training:
\[
L_w = \|w^* - w\|_1^2,
\] (8)
where \( w = M(z) \), and \( w^* = F_{\text{fuse}}(E_{\text{app}}(x_i), E_{\text{pose}}(x_j)) \). Other loss functions are the same as Eq. 6.

**Motion Sampling**
In order to generate videos from random motion noise \( \epsilon \), we design a motion sampling module to sample meaningful sequences of pose representations. Here we choose
LSTM \cite{hochreiter1997long} to generate later poses of $n-1$ frames $\{f_{p(2)}, \ldots, f_{p(n)}\}$ by giving the initial pose $f_{p(1)}$ and sampled motion noises $\{\epsilon_2, \ldots, \epsilon_n\}$:
\[
\{f_{p(2)}, \ldots, f_{p(n)}\} = \text{LSTM}(f_{p(1)}, \{\epsilon_2, \ldots, \epsilon_n\}).
\] (9)

Because we have no ground truth for generated results, we use a 2D adversarial loss $L^\text{2D}_{\text{gan}}$ to make the single generated pose representation fit the distribution of real videos and a 3D adversarial loss $L^\text{3D}_{\text{gan}}$ whose inputs are consecutive pose representations to constrain the temporal consistency:
\[
L^\text{3D}_{\text{gan}} = \mathbb{E} \left[ \log D_v(\{x_1, x_2, \ldots, x_n\}) \right] + \mathbb{E} \left[ \log (1 - D_v(G(F\text{fuse}(f_{a(1)}), \{f_{p(1)}, \ldots, f_{p(n)}\)))) \right],
\] (10)

where $D_v$ is the video discriminator to predict whether a video is real or fake according to the given sequence.

Additionally, we add a diversity loss $L_d$ to make the generated pose representations $p$ vary with the given motion noises $\epsilon$:
\[
L_d = -\frac{1}{n-1} \sum_{t=2}^{n} \left( \text{FC}(h_t)^T \epsilon_t / \|\text{FC}(h_t)\| \|\epsilon_t\| \right),
\] (11)

where $\text{FC}(\cdot)$ is a fully connected layer and $h$ is the hidden feature.

For the motion sampler of $256 \times 256$ resolution, we generate final videos with sampled pose sequence and apply supervisions on image/video representations because we find this operation brings more diverse motions. But for $1024 \times 1024$ resolution, all losses are added on pose representations only, which are much smaller than image space, saving a large amount of computation cost. This design allows the training process to supervise a longer sequence.

**Experiments**

**Joint Unaligned Dataset**

Previous face video generation methods \cite{tulyakov2018faceforensics++, saito2020towards, skorokhodov2021styleganv} usually use FaceForensics \cite{rossler2019faceforensics++} as the video dataset and are able to generate good videos whose identities are from FaceForensics. However, StyleGAN-V \cite{skorokhodov2021styleganv} points out the limitation that only 700 identities of FaceForensics can be sampled. To overcome this problem, we propose the joint training strategy. We choose unaligned FFHQ \cite{karras2019analyzing} as image dataset that provides tens of thousands of identities for sampling and RAVDESS \cite{livingstone2018dataset} as video dataset which provides 2451 high-quality vivid face motion sequences. Compared to FaceForensics, RAVDESS has more various motions and better video quality thus having a smaller gap with the image dataset FFHQ, helping the convergence of joint training.

In addition, many previous video generation meth-
Table 1: Quantitative Comparisons on FID and FVD Score. We compute the FID-RA VDESS between the generated frames and real frames from RA VDESS. FID-Mixture is calculated against a mixture dataset. FVD is computed between synthesized videos and videos in RA VDESS dataset.

| Methods      | FID-RA VDESS (↓) | FID-Mixture (↓) | FVD (↓) |
|--------------|-----------------|-----------------|---------|
| MCHD-pre     | 174.44          | 53.53           | 1425.77 |
| MCHD-re      | 146.08          | 73.73           | 1350.67 |
| StyleGAN-V   | 17.93           | 95.35           | 171.60  |
| DIGAN        | 12.02           | 77.01           | 142.08  |
| StyleFaceV   | 15.42           | 25.31           | 118.72  |

ods (Tulyakov et al. 2018, Saito et al. 2020) Skorokhodov, Tulyakov, and Elhoseiny (2021) stick the face in the center by aligning the face bounding box when preprocessing face video data. It erases natural head movement thus reducing the difficulty of face generation. However, this operation brings obvious face distortions and shaking among frames. In this paper, we use the setting (Starokin et al. 2019) and allow the face to freely moved in a fixed window. It is more consistent with real-world face videos because, in real-world face videos, the camera is usually fixed.

Evaluation Metrics. For evaluation, we report Frechet Inception Distance (FID) (Heusel et al. 2017) and Frechet Video Distance (FVD) (Unterthiner et al. 2018). In this paper, we use two versions of FID: FID-RA VDESS and FID-Mixture, where FID-RA VDESS is computed against RA VDESS dataset and FID-Mixture is calculated against a mixture dataset (RA VDESS and FFHQ). We also conduct a user study to evaluate the quality of generated videos. We mix our generated videos with those generated by baselines because paired comparison is not supported in unconditional generation. A total of 30 users were asked to give two scores (1-5, the best score is 5) for the naturalism of movements and the identity preservation throughout the movements.

Experimental Settings

We compare our proposed StyleFaceV with the following representative methods on the face video generation task. Compared to previous sticking setting, unaligned face generation is more challenging. Considering huge computational costs for DI-GAN, all results are fixed with 60 frames for the fair comparison.

MoCoGAN-HD (Tian et al. 2021) uses a pre-trained image generator and synthesizes videos by modifying the embedding in latent space. We compare with two versions of MoCoGAN-HD: 1) MCHD-pre uses the released pre-trained model which supports sampling identities from FFHQ domain. 2) MCHD-re retrain the models using StyleGAN3-R (Karras et al. 2021) with video dataset RA VDESS (Livingstone and Russo 2018) for a fair comparison, as the original version is built upon StyleGAN2 (Karras et al. 2020) which does not support various unaligned face poses.

DI-GAN (Yu et al. 2022) uses an INR-based video generator to improve the motion dynamics by manipulating the space and time coordinates separately. Because it only supports the video dataset, we retrain it on RA VDESS dataset. The number of training frames is set to 60 which is the same as the testing phase to avoid the collapse.

| Methods         | Movement Naturalism | Identity Preservation |
|-----------------|---------------------|-----------------------|
| MCHD-pre        | 1                   | 1                     |
| MCHD-re         | 2                   | 2                     |
| StyleGAN-V      | 3                   | 3                     |
| DIGAN           | 3                   | 3                     |
| StyleFaceV      | 4                   | 4                     |

Figure 5: User Study. We asked users to give 1 - 5 scores (the higher is better) for videos generated by different methods in terms of movement naturalism and identity preservation. Our method achieves the highest score.

StyleGAN-V (SG-V) (Skorokhodov, Tulyakov, and Elhoseiny 2021) builds the model on top of StyleGAN2, and redesigns its generator and discriminator networks for video synthesis. We also retrain it on RA VDESS dataset.

Qualitative Comparison

As presented in Fig. 4, MCHD-pre always generates faces located at the center of the image because it is based on StyleGAN2. In addition, generated results will crash with the time going as this model is pre-trained on 16 frames setting. Our retrained MCHD-pre improves those problems by using StyleGAN3. Meanwhile, video dataset RA VDESS makes MCHD-re learn the face motion. However, the man in the second video clip of MCHD-re gradually becomes a woman with the pose movement. The reason is that MCHD-re still moves along the entangled latent codes thus making identity change with other actions. Our method well decomposes the appearance representations and pose representations, preserving the identity throughout the pose movement.

Another state-of-the-art method, DI-GAN exactly produces videos without identity changes. However, some of frames will be blurred or even crash. The most recent work StyleGAN-V does not suffer from the crash problem. However, its generated faces are easy to deform and distort. The poor stability is mainly because both DI-GAN and StyleGAN-V use a single end-to-end trained framework to handle both the quality and coherence of all synthesized frames, which gives a huge burden to the network. Therefore, DI-GAN and StyleGAN-V cannot model the unaligned faces with movements. In contrast, our framework uses a pre-trained StyleGAN3-R which only focuses on the quality of every single frame and uses other modules to undertake the duty of coherence. Results shown in Fig. 4 exhibit that our StyleFaceV outperforms all other methods in both quality and identity keeping.

Quantitative Comparison

As shown in Table 1, our proposed StyleFaceV achieves the FID-RA VDESS score comparable to that of DIGAN. Since our method adopts a joint training strategy which utilizes information from both the image domain and video domain, apart from RA VDESS dataset, it is able to generate images from the FFHQ distribution. In other words, the synthesized frames of our method are more diverse than RA VDESS dataset. Therefore, it has a slightly higher FID-RA VDESS than DIGAN. To further evaluate if the synthesized data are truly from FFHQ distribution and RA VDESS distribution, we set up a mixture testing dataset, which is uniformly
mixed with FFHQ dataset and RAVDESS dataset. We compute the FID-Mixture with this mixture dataset, verifying that our method is capable of generating frames similar to both of the domains. FVD is adopted as the metric for evaluating the quality of synthesized videos, and it is computed against two sets of videos. Compared to the state-of-the-art methods, our method has the lowest FVD score, indicating that videos synthesized by our method are more realistic. Fig. 5 shows the results of the user study. We compare our proposed StyleFaceV with MCHD-pre, MCHD-re, and DI-GAN. We report the preference percentage for each method. Our approach achieves the highest scores for both the naturalism of movement and identity preservation, outperforming baseline methods by a large margin.

Ablation Study

Self-Supervised Embedding Loss. Unaligned images generated by StyleGAN3-R make models hardly converge. Therefore, the model trained without the self-supervised embedding loss generates blurry results. To help the model converge, we add the self-supervised embedding loss on \( w^* \). Since the loss is computed directly on the latent codes, the gradients are better preserved and propagated to \( F_{\text{fuse}} \) and \( E_{\text{app}} \) compared to the training with only a reconstruction loss at image level. As shown in Fig. 6 without the help of self-supervised embedding loss, the model is not able to recover more details of the original sampled image, thus produces severely blur results with both FID (FID-RA VDESS=102.97, FID-Mixture=67.91) and FVD (461.17). We also report arcface cosine similarity between the reconstructed images and the original images. The arcface similarity is averaged over 512 images. As shown in Table 2, the model without self-supervised embedding loss has an arcface similarity of 0.6676, while our final method achieves 0.7266 in this metric, further indicating the effectiveness of the self-supervised embedding loss.

Weight Re-Balance. At the beginning of the training of our decomposition and re-composition pipeline, the biggest challenge is the reconstruction of unaligned images with various poses. After that, the model is supposed to be capable of tracing finer details from pose representations, especially the actions of the mouth. However, the change of mouth actions is only present in video data. To make networks focus on it, we need to re-balance the weight and make all \( \lambda \)s for training with sampled images significantly smaller (\( \times \frac{1}{10} \)). Fig. 7 shows that the model with weight re-balance can successfully trace fine details from pose representations. As shown in Table 3, for the setting without Weight Re-Balance, it focus on FFHQ reconstruction (FID-RA VDESS=119.90) but performs poorly on RAVDESS reconstruction (FID-Mixture=21.09) and motion sampling (FVD=722.84).

Implementation Details

Network structure. \( F_{\text{app}} \), \( F_{\text{pure}} \), and \( F_{\text{fuse}} \) are formed by res-blocks. For 2D Discriminator, we just use NLayerDiscriminator [Isola et al. 2017]. 3D Discriminator is similar but uses conv3d instead.

Loss weights. Initially, \( \lambda_1 = 10, \lambda_{\text{per}} = 100, \lambda_{\text{gan}} = 0, \lambda_{\text{pose}} = 10, \lambda_w = 100 \) for both image data and video data. After the disentanglement pipeline roughly converges and is able to reconstruct identity, we do weight rebalance to make the pose extractor mainly focus on fine motion capture. Then

| Settings | Arcface Cosine Similarity (↑) |
|----------|-------------------------------|
| Without \( \mathcal{L}_w \) | 0.6676 |
| With \( \mathcal{L}_w \) | 0.7266 |

Table 2: Arcface Cosine Similarity Influenced by Self-Supervised Embedding Loss. The arcface similarity is averaged over 512 images. The model without self-supervised embedding loss has an arcface similarity of 0.6676, while our final method achieves 0.7266 in this metric.

| Methods | FID-RA VDESS (↓) | FID-Mixture (↓) | FVD (↓) |
|---------|-----------------|-----------------|--------|
| Without Reweight | 119.90 | 21.09 | 722.84 |
| Without \( \mathcal{L}_w \) | 102.97 | 67.91 | 461.17 |
| StyleFaceV | 15.42 | 25.31 | 118.72 |

Table 3: Quantitative Ablations on FID and FVD Score. We compute the FID-RA VDESS between the generated frames and real frames from RAVDESS. FID-Mixture is calculated against a mixture dataset. FVD is computed between synthesized videos and videos in RAVDESS dataset.
videos than state-of-the-art methods. Proposed StyleFaceV can generate more visually appealing video datasets. Extensive experiments demonstrate that our method is able to generate high-quality videos without large-scale training data for the domain. With the joint training strategy, our method is able to use information from both the image domain and video domain.

We also propose a joint-training strategy to make StyleFaceV synthesize videos by leveraging pre-trained image generator StyleGAN3. At the core of our framework, we decompose the inputs to the StyleGAN3 into appearance generator StyleGAN3 and pose information. In this way, we can generate videos from noise by sampling a sequence of pose representations while keeping the appearance information unchanged. We also propose a joint-training strategy to make use of information from both the image domain and video domain. With the joint training strategy, our method is able to generate high-quality videos without large-scale training video datasets. Extensive experiments demonstrate that our proposed StyleFaceV can generate more visually appealing videos than state-of-the-art methods.

**Results in High Resolution**

We also demonstrate the capability of StyleFaceV to generate 1024 × 1024 videos in Fig. 8. It should be noted that the models are trained without accessing any high-resolution training videos.

**Limitation**

Our main limitation is that the generated videos whose identity from the video domain have more vivid motions than those from the image domain. The reason lies in that we have direct temporal training objectives on the video domain while there is no temporal constraint for the image domain. This is a common issue for all baselines and our proposed method has already demonstrated plausible results.

**Conclusions**

In this work, we propose a novel framework named StyleFaceV to synthesize videos by leveraging pre-trained image generator StyleGAN3. At the core of our framework, we decompose the inputs to the StyleGAN3 into appearance and pose information. In this way, we can generate videos from noise by sampling a sequence of pose representations while keeping the appearance information unchanged. We also propose a joint-training strategy to make use of information from both the image domain and video domain. With the joint training strategy, our method is able to generate high-quality videos without large-scale training video datasets. Extensive experiments demonstrate that our proposed StyleFaceV can generate more visually appealing videos than state-of-the-art methods.

**References**

Aich, A.; Gupta, A.; Panda, R.; Hyder, R.; Asif, M. S.; and Roy-Chowdhury, A. K. 2020. Non-adversarial video synthesis with learned priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6090–6099.

Brock, A.; Donahue, J.; and Simonyan, K. 2018. Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*.

Chan, C.; Ginosar, S.; Zhou, T.; and Efros, A. A. 2019. Everybody dance now. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5933–5942.

Chan, K. C.; Wang, X.; Xu, X.; Gu, J.; and Loy, C. C. 2021. GLEAN: Generative latent bank for large-factor image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Clark, A.; Donahue, J.; and Simonyan, K. 2019. Adversarial video generation on complex datasets. *arXiv preprint arXiv:1907.06571*.

Fox, G.; Tewari, A.; Elgharib, M.; and Theobalt, C. 2021. Stylevideoagan: A temporal generative model using a pre-trained stylegan. *arXiv preprint arXiv:2107.07224*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation*, 9(8): 1735–1780.

Hyun, S.; Kim, J.; and Heo, J.-P. 2021. Self-Supervised Video GANs: Learning for Appearance Consistency and Motion Coherency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10826–10835.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-Image Translation with Conditional Adversarial Networks. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*.

Jiang, Y.; Huang, Z.; Pan, X.; Loy, C. C.; and Liu, Z. 2021. Talk-to-Edit: Fine-Grained Facial Editing via Dialog. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13799–13808.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2017. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*.

Karras, T.; Aittala, M.; Hellsten, J.; Laine, S.; Lehtinen, J.; and Aila, T. 2020a. Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems*, 33: 12104–12114.

Karras, T.; Aittala, M.; Laine, S.; Härkönen, E.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2021. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4401–4410.
Image Quality and Perks of StyleGAN2.

Skorokhodov, I.; Tulyakov, S.; and Elhoseiny, M. 2021. StyleGAN-V: A Continuous Video Generator with the Price. arXiv preprint arXiv:2112.14683.