D^{2}Animator: Dual Distillation of StyleGAN For High-Resolution Face Animation

Zhuo Chen*
Tsinghua Shenzhen International Graduate School, Tsinghua University
Shenzhen, China
z-chen17@mails.tsinghua.edu.cn

Chaoyue Wang†
JD Explore Academy
Beijing, China
wangchaoyue9@jd.com

Haimei Zhao
The University of Sydney
Sydney, Australia
hzha7798@uni.sydney.edu.au

Bo Yuan
Qianyuan Institute of Sciences
Hangzhou, China
boyuan@ieee.org

Xiu Li
Tsinghua Shenzhen International Graduate School, Tsinghua University
Shenzhen, China
li.xiu@sz.tsinghua.edu.cn

Figure 1: High-resolution face animation from source (left top) to target (right top).

ABSTRACT

The style-based generator architectures (*e.g.* StyleGAN v1, v2) largely promote the controllability and explainability of Generative Adversarial Networks (GANs). Many researchers have applied the pretrained style-based generators to image manipulation and video editing by exploring the correlation between linear interpolation in the latent space and semantic transformation in the synthesized image manifold. However, most previous studies focused on manipulating separate discrete attributes, which is insufficient to animate a still image to generate videos with complex and diverse poses and expressions. In this work, we devise a dual distillation strategy (D^{2}Animator) for generating animated high-resolution face videos conditioned on identities and poses from different images. Specifically, we first introduce a Clustering-based Distiller (CluDistiller) to distill diverse interpolation directions in the latent space, and synthesize identity-consistent faces with various poses and expressions, such as blinking, frowning, looking up/down, *etc.* Then we propose an Augmentation-based Distiller (AugDistiller) that learns

* This work was performed when Zhuo Chen was visiting JD Explore Academy as a research intern.
†Corresponding author.

MM '22, October 10–14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3548002

to encode arbitrary face deformation into a combination of interpolation directions via training on augmentation samples synthesized by CluDistiller. Through assembling the two distillation methods, D^{2}Animator can generate high-resolution face animation videos without training on video sequences. Extensive experiments on self-driving, cross-identity and sequence-driving tasks demonstrate the superiority of the proposed D^{2}Animator over existing StyleGAN manipulation and face animation methods in both generation quality and animation fidelity.

CCS CONCEPTS

• **Computing methodologies → Image manipulation**.

KEYWORDS

GANs; High-resolution image generation; Face Animation

ACM Reference Format:

Zhuo Chen*, Chaoyue Wang†, Haimei Zhao, Bo Yuan, and Xiu Li. 2022. D^{2}Animator: Dual Distillation of StyleGAN For High-Resolution Face Animation. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), Oct. 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3503161.3548002

1 INTRODUCTION

Face animation technology aims to reenact a given portrait (*i.e.* source image) to different poses and expressions provided by targets while preserving the identity. By applying a sequence of frames as targets, videos of imitative motions can be generated. Seeing as such technology has a wide application prospect in teleconference, movie industry and artistic creation, plenty of works have been done to enhance the performance of face animation [61, 40, 49, 54, 14, 48, 70, 59, 70].

Traditional face animation methods [6, 5, 16] usually modify the source image by warping pixels, where the transforming of head pose and the hallucination of unseen parts are unachievable. With the great progress of deep learning [30, 17, 68, 23, 67, 22, 73, 21], generation-based methods [61, 71, 49, 59, 64, 18, 14, 70] attempt to synthesize a new image conditioned on the identity of the source and the pose extracted from the target frames instead of warping. The generation-based methods learn to model the face appearance implicitly via large-scale training on video sequences, therefore can handle more drastic deformation during inference. However, the training of generative models requires abundant video sequences and computing resources. Collecting and storing high-resolution videos are not only expensive but also risk the infringement of portrait rights. Not to mention the consumption of computing resources rises exponentially with video resolution. Therefore, most generation-based models suffer from low resolution and reality.

Recently, the style-based generator architectures (*e.g.* StyleGANs) [28, 29] made a breakthrough in unconditional image generation. They can synthesize realistic images of resolution up to 1024×1024. Moreover, some papers utilize pretrained StyleGANs to realize face manipulation through exploring the correlation between linear interpolation in the latent space and semantic transformation in the synthesized image manifold. However, most previous methods manipulate discrete attributes (*e.g.* smile or not) and can barely accomplish high-resolution video generation with complex motions.

In this work, we propose a dual distillation strategy (D^{2}Animator) that can generate high-resolution face animation videos with a pretrained StyleGAN while needing no extra training on videos. D^{2}Animator consists of a Clustering-based distiller (CluDistiller) and an Augmentation-based distiller (AugDistiller). In the CluDistiller phase, we distill the latent space by clustering the synthetic images, and the distillation contributes to discovering the interpolation directions that are in control of facial actions. Unlike previous latent space analysis methods, CluDistiller can search for the direction which controls specific facial parts, and then collecting a comprehensive set of directions that correspond to various facial actions. Then the AugDistiller learns to encode arbitrary facial actions into the combination of interpolation directions via training on augmentation samples synthesized by CluDistiller. Different from the previous face animation methods, the AugDistiller models the facial actions by distilling a pretrained image generator instead of training on additional video sequences. Through assembling the proposed two distillers, D^{2}Animator can generate high-resolution videos of face animation. In summary, the contributions of this work are as follows:

• We introduce a Clustering-based distillation module (CluDistiller) to find a comprehensive set of interpolation directions that controls diverse facial actions. The interpolation directions can be used for both flexible face manipulation and data augmentation.

• We present an Augmentation-based distillation module (AugDistiller) to learn facial actions from a pretrained image generator instead of video sequences.

• The proposed D^{2}Animator can be a new paradigm of applying a pretrained GAN in downstream tasks by employing a generator as a neural renderer and a data producer simultaneously.

• Experiments show that the proposed D^{2}Animator outperforms the previous start-of-the-art face animation methods in qualitative and quantitative, especially for high-resolution portrait animation.

2 RELATED WORK

Portrait Animation. Most previous works deal with portrait animation following conditional generation methods [58, 57, 56] where the generated frames are conditioned on the identity of the source image and the poses of the target frames. A few-shot method [71] uses the landmark image of the target frame as the input and injects the identity of the source by AdaIN [24] during the feed-forward process. FSGAN [40] warps the source face following a query Euler angle of target frames, and inpaints the warped face using Pix2PixHD network [58]. Some face swapping methods can also be considered portrait animation. FaceShifter [35] takes the identity feature of the source as input and injects the attribute features of the target frame to generate the face swapping result. In contrast, the input of SimSwap [9] is the target image, and the identity is controlled by AdaIN [24] blocks in the middle of the network. Besides AdaIN [24], SPADE [41] is another efficient condition controlling approach. For example, HeadGAN [14] and PuppeteerGAN [10] both employ SPADE [41] to introduce pose and texture information into the network and guide the generation.

X2Face [61] estimates a dense flow to warp the source image to an intermediate frontal state and then to the target pose. MonkeyNet [48] learns a key-point detector and a dense motion predictor to estimate the dense flow that warps the feature of the source

image to the pose of the target. Then the generator of MonkeyNet [48] takes the warped feature to generate the final output image. FOMM [11] detects key points and the one-order Jacobi matrix of each key point to achieve a more accurate motion transformation. Work [59] extends the 2D key-point detection to 3D for boosting the reality of the generated video. Disentanglement-based methods such as MoCoGAN [54] disentangle the input to content and motion vectors, and portrait animation is performed by combining the content vector of the source with the motion vector of the target. Siarohin [50] proposed a novel image animation method that can manipulate various articulated objects. LLA [60] trains an autoencoder to encode the relation between image warping and latent space navigation in a self-supervised manner. StyleHeat [66] employs a pretrained StyleGAN for face animation by warping the intermediate features.

StyleGAN Semantic Analysis. Style-based generators show surprising semantic interpretability not only in the latent W^+ space but also in the feature space. Labels4Free [3] distinguishes the foreground and background of the generated image easily by taking the intermediate features of StyleGANs as input. Xu *et al.* [62] reveal the potential of StyleGANs feature by transforming the concatenated feature to segmentation linearly supervised by only a few semantic labels. Endo *et al.* [15] carry out a image-to-image transfer method with the StyleGAN generator and the supervision of pseudo labels extracted from StyleGAN features. PSP [44] proposes a GAN-inversion method which can also generate new images conditioned on segmentations and sketches. DatasetGAN [72] extends segmentation generation based on a pretrained StyleGAN from the face field to various data. SemanticGAN [34] proposes a different way to train image and segmentation generation branches together instead of using a pretrained image generator. EditGAN [36] displays a precise image manipulation method by optimizing the latent code of the target while constraining the semantic label.

StyleGAN-based Image Manipulation. StyleGANs [28, 29, 27] make great progress in unconditional image generation. Benefiting from the excellent performance and disentanglement ability of Style-generators, various applications have been developed based on pretrained StyleGAN. Shen *et al.* [46] show linear interpolation in the W^+ [1, 2] space can manipulate the generated image with the help of the corresponding attribute classifier. Some methods attempt to search for the meaningful manipulation vector in w^+ space in an unsupervised/self-supervised manner. GANSpace [20] identify important manipulation vectors as the principal components (PCA) of the latent space. Plumerault *et al.* [43] introduce a method to find meaningful directions by self-augmentation such as translation, zoom or color variations. Closed-form factorization [47] between the latent W^+ space and the generated image space can also be used to find meaningful interpolatioin directions. Compared with supervised methods, unsupervised methods need no pretrained classifiers, which enables them to find more fantastic manipulation vectors, especially for attributes without specific datasets to train classifiers. However, a part of vector directions found by unsupervised searching in the latent space is incorrect or meaningless, and the meaning of the other part needs to be tagged manually.

Other methods use networks to process the latent code of StyleGAN in order to edit the generated image. Lu *et al.* [37] extract the pair of manipulation vector and attribute by a deformation directed by classification consistency and centroid constant. IALS [19] finds semantic directions for disentangled attributes in a step-by-step manner to keep the instance unaffected. StyleFusion [26] is able to extract different attributes from a set of inputs to generate harmonized style code to edit multiple attributes of the generated image directly. Yao *et al.* [65] employ a dedicated latent transformation network which uses a single layer of linear transformation to compute the difference value of latent code for each attribute editing. StyleFlow [4] formulates the conditional exploration as an instance of conditional continuous normalizing flows in the GAN latent space conditioned by attribute attributes.

More applications such as 3D shape manipulation [52], layout editing [63], style transfer [31], virtual try-on [33], face swapping [38, 74, 12] and text-driving editing [42] show more ways of image manipulation based on pretrained StyleGAN. MoCoGAN-HD [53] is the first work that tries to generate videos with pretrained StyleGAN. However, extra video datasets are needed to train the trajectory encoder and controllable pose retargeting is still needed.

3 METHOD

Human face animation aims to transfer the pose and expression of the target image t to the source image s. Employing a pretrained style-based generator G, our method realizes face animation by predicting the variation Δw in the latent space which is equivalent to the facial action between s and t. The proposed D^{2}Animator learns to model the relation between the latent space and facial action through a dual-distillation strategy. D^{2}Animator consists of two distillation modules Clustering-based Distiller (CluDistiller) and Augmentation-based Distiller (AugDistiller). The procedure of the CluDistiller searching the latent space for the interpolation directions that are in control of pose and expression is illustrated in Section 3.1. Based on CluDistiller, we present AugDistiller, an Augmentation-based distillation method that encodes the facial action to interpolation directions which is presented in Section 3.2.

3.1 Clustering-based Distiller

As aforementioned, StyleGANs provide a prior for face manipulation in the way of latent space interpolation. Clustering-based Distiller (CluDistiller) targets to distill the StyleGANs to find the interpolation directions that alter the pose and expression without affecting the identity. Specifically, CluDistiller can not only search for the interpolation directions that are equivalent to a designated facial part but also explore the latent space to collect a comprehensive set of interpolation directions to complete diverse pose and expression manipulations. As shown in Figure 2, CluDistiller takes four steps, *i.e.* sampling, detecting, clustering and collecting to find the interpolation directions for face animation. We explain the four steps in detail as follows.

Sampling: In the beginning, we generate a synthetic dataset with a pretrained StyleGAN by intensively sampling from the latent space, *i.e.* sampling from the learned image distribution. We randomly sample $z \in \mathcal{N}(0, 1)$, and then map them into the latent space W^+ [1, 2] (noted as W for short below) using the mapping network. We store the latent codes $\{w\}$ and the images $\{\mathcal{I}\}$ generated by G as two items of the synthetic dataset.

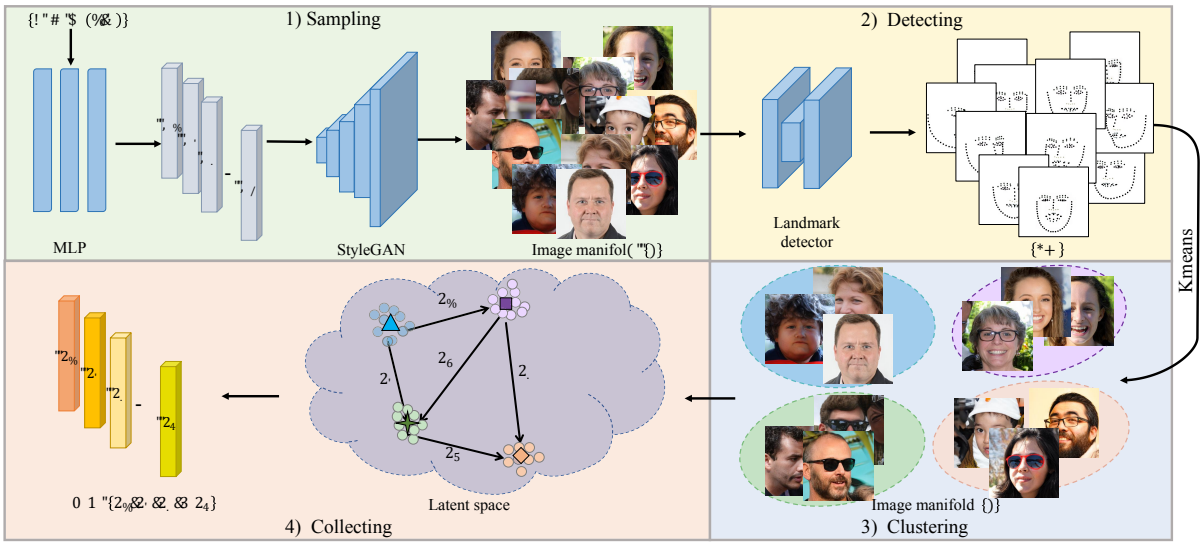

Figure 2: Illustration of CluDistiller framework, consisting of four steps, *i.e.* Sampling, Detecting, Clustering and Collecting.

Figure 3: Face manipulation and augmentation with the interpolation directions collected by CluDistiller.

Detecting: Facial landmark is an effective representation to separate the information of facial action from the appearance. Besides, the small data volume of detected landmarks makes it easier for subsequent processing. Therefore, we detect the facial landmarks of the synthesized images $\{\mathcal{I}\}$ as the representation of facial action. In practice, we employ one of the start-of-the-art facial landmark detectors [69] to predict 98 key points for each generated image.

Clustering: Previous methods [46] presented the classification-based distillation methods to discover interpolation directions for discrete attributes. However, diverse facial actions cannot be fully separated by multiple classifiers. In order to divide the samples into several contiguous clusters in accordance with pose and expression, we separate the synthetic dataset by clustering. Taking the 98 facial landmarks as a 196-dimensional representation of pose and expression for each sample, we utilize Kmeans [51] to cluster these samples. Furthermore, we can select the landmark points that are relative to certain facial parts to distinguish them more precisely. In addition, the number of clusters would affect the scale of discovered facial actions. By selecting different facial landmark points and setting diverse clustering numbers, we divide synthetic samples in different manners to collect various interpolation directions. The diagram of the clustering stage (third stage) shown in Figure 2 illustrates the effect of clustering points. Clustering the dataset according to the mouth landmark points can divide the samples into the blue and purple clusters, and the green and orange clusters are divided by cheek landmark points.

Collecting: Finally, we aim to collect the interpolation directions based on the clustering results. For each cluster, we compute the mean $\bar{w}$ of the samples as the centroid. Then we calculate the interpolation direction between each pair of centroids. Based on the observation that only the first 8 layers of w affect the pose and expression, we omit the last 10 layers during calculating. In order to eliminate the insignificant interpolation directions, we use Independent Component Analysis (ICA) [25] to choose the most significant components when the number of clusters is over 6.

By setting different numbers of clustering and facial landmark pints, we can collect a comprehensive set of interpolation direction $\mathcal{V} = \{v_0, v_1, v_2, ..., v_n\}$, which are in control of continuous facial actions. We show examples of face manipulation using some of the collected interpolation directions in Figure 3 (left). The image at the center is a randomly generated source image and the others are the manipulated results. In the first row, we show that the collected interpolation directions can alter the head pose of the source with large-scale motion. The two manipulated images in the second row display that we can edit subtle expressions such as blinking (left) and gazing (right). The last row presents the results of manipulating expression in different ways including excitement, happiness and surprise. It proves that CluDistiller can collect diverse interpolation directions that correspond to different facial actions.

3.2 Augmentation-based Distiller

Augmentation-based Distiller (AugDistiller) is a distillation framework that learns to model the facial actions with the variations in the latent space. Based on the interpolation directions $\mathcal{V}$ =

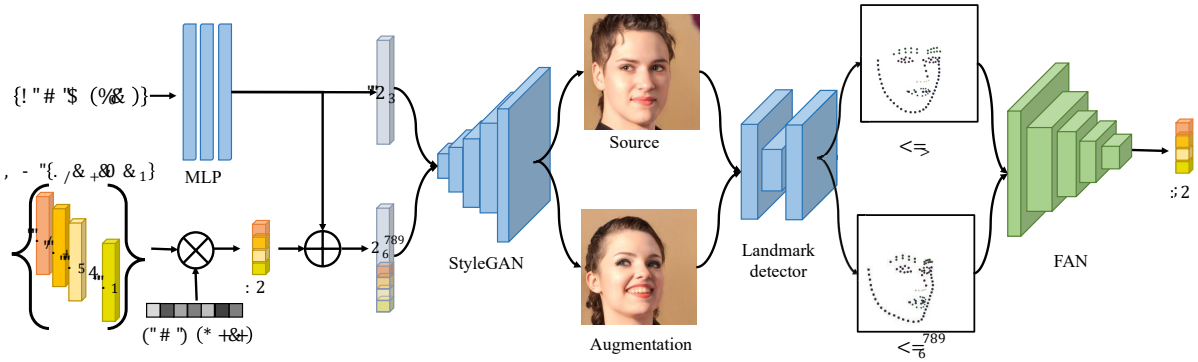

Figure 4: Illustration of AugDistiller framework which trains the FAN network in a self-supervised manner.

$\{v_0, v_1, v_2, ..., v_n\}$ provided by CluDistiller, we first develop an augmentation method to generate high-volume manipulated images by sampling on the explored directions in a flexible manner. Then the AugDistiller employs a facial action network (FAN) to predict the combination of the interpolation directions from the difference of the pose and expression between source and target landmarks.

Augmentation. The interpolation directions $\mathcal{V} = \{v_0, v_1, v_2, ..., v_n\}$ collected by CluDistiller can be adopted as a set of bases that conduct diverse manipulations. Next, we can achieve more complex manipulations by estimating the weighted sum of these bases. However, setting too large weights may lead to artifact and facial expression distortion, while using too small weights limits the diversity of augmentation. Our augmentation method solves this problem by adjusting the norm of augmented variation layer-by-layer. We first sum the interpolation directions with weights r sampled from a uniform distribution $\mathcal{U}(-1, 1)$. Then, for each layer of the summed vector, we re-norm it by restricting the euclidean distance between the altered w and the $\bar{w}$ of the latent space by a fixed threshold. Figure 3 (right) shows the images augmented by our method. Compared with the manipulation results in Figure 3 (left), the augmentation method can generate images with more various poses and expressions while remaining the quality and identity of the source. For simplicity, we still note the augmentation as sampling from a uniform distribution $\mathcal{U}(-1, 1)$ in the following sections.

Training. As shown in Figure 4, AugDistiller trains the FAN by distilling the StyleGAN through augmentation. We employ a pretrained StyleGAN G and a facial landmark detector $\mathcal{F}$ during the training, where neither video sequences nor real images are required. We first generate the w_s and I_s of source from randomly sampled $z \in \mathcal{N}(0, 1)$. Then we alter the pose and expression of the I_s by adding a variation Δw on the w_s,

$$w_t^{aug} = w_s + \Delta w, \tag{1}$$

such that we can input the augmented Δw to the generator and produce the augmented target image I_t^{aug}.

Next, we detect the facial landmarks of I_s and I_t^{aug}, denoted as:

$$\begin{aligned} kp_s &= \mathcal{F}(G(w_s)), \\ kp_t^{aug} &= \mathcal{F}((G(w_t^{aug})). \end{aligned} \tag{2}$$

The input of the FAN is the concatenated heatmaps of the landmarks, which represents the location of each landmark as a unimodal Gaussian distribution. The FAN predicts the variation as:

$$\Delta\widetilde{w} = FAN(kp_s, kp_t^{aug}). \tag{3}$$

In order to show the distilling process, we expand Equation 3 as:

$$\begin{aligned} \Delta\widetilde{w} &= FAN(\mathcal{F}(G(w_s)), \mathcal{F}(G(w_s + r \cdot \mathcal{V}))), \\ w_s &= MLP(z), \end{aligned} \tag{4}$$

where $z \in \mathcal{N}(0, 1)$ and $r \in \mathcal{U}(-1, 1)$. Since we already know the ground truth of the variation, losses on image content are not needed anymore. We train the FAN by minimizing the difference between the predicted $\Delta\widetilde{w}$ and the augment Δw, denoted as $\mathcal{L}$:

$$\mathcal{L} = \log(\cosh(\Delta\widetilde{w} - \Delta w)) + \lambda \cdot ||\Delta\widetilde{w} - \Delta w||_2, \tag{5}$$

where the hyper-parameter λ is usually set as 0.1.

In the training phase, the data and augmented ground truth are all sampled from the image manifold and latent space of a pretrained StyleGAN. After training, the FAN is able to predict the variation which is equivalent to the facial action between two images. Thus, the trained FAN can be used to animate a source image to target by providing the desired variation in the latent space.

Fine-tuning. We can further improve the accuracy of pose and expression retargeting by fine-tuning the FAN on specific source image without additional data. Given a source image I_s, we first inverse the image to the latent space as $\hat{w}_s$ by a GAN-inversion method [55]. Then we fix the $\hat{w}_s$ as the source instead of sampling randomly, with remained procedures the same as the training phase.

Inference. During inference on real images, the inversion of the source image is also needed. For each target pose, we input the facial landmarks of the source and target to the FAN, *i.e.*

$$\Delta\widetilde{w} = FAN(\mathcal{F}(G(\hat{w}_s)), \mathcal{F}(G(w_t)). \tag{6}$$

Then we add the predicted variation Δw back to $\hat{W}_s$ and synthesize the predicted image with StyleGAN:

$$\widetilde{I} = G(\hat{w}_s + \Delta\widetilde{w}). \tag{7}$$

4 EXPERIMENTS

Experimental settings. We implement the experiments in three situations including self-driving, cross-identity and sequence-driving animation. The cross-identity experiment is performed on high-resolution image datasets CelebA-HQ [32] and FFHQ [28]. For each comparison, we randomly select 1000 images as the sources and

Figure 5: Qualitative comparison pf cross-identity experiment on CelebA-HQ [32] and FFHQ [28].

Table 1: Quantitative experiments on CelebA-HQ [32] and FFHQ [28].

Methods	Cross-identity								Sequence-driving							
	CelebA-HQ [32]				FFHQ [28]				CelebA-HQ [32]				FFHQ [28]			
	FID(↓)	IS(↓)	NME(↓)	CSIM(↑)	FID(↓)	IS(↓)	NME(↓)	CSIM(↑)	FID(↓)	IS(↓)	NME(↓)	CSIM(↑)	FID(↓)	IS(↓)	NME(↓)	CSIM(↑)
FOMM[49]	100.60	0.157	3.60	0.46	126.41	0.168	4.31	0.45	114.29	0.187	**3.79**	0.44	144.93	0.193	**3.80**	0.42
Bi-layer[70]	200.55	0.202	5.08	0.43	204.48	0.270	5.47	0.34	175.83	0.327	6.57	0.40	217.95	0.416	6.76	0.36
PSP[44]	47.70	0.063	**2.34**	0.28	75.25	0.113	**3.70**	0.22	91.27	0.135	4.42	0.25	98.31	0.147	4.41	0.16
StyleFusion[26]	53.05	0.057	4.84	0.55	69.22	0.093	5.77	0.49	83.03	0.109	5.97	0.54	93.81	0.154	5.96	0.48
D^2Animator	**44.62**	**0.054**	3.60	**0.57**	**53.99**	**0.085**	5.05	**0.58**	**69.26**	**0.096**	4.71	**0.54**	**79.27**	**0.114**	5.52	**0.56**

another 1000 images as the targets. In the sequence-driving experiment, we pick 100 sequences of different people from the VoxCeleb1 [39] dataset and sample 4 frames in each sequence as the targets. The target frames are aligned to the layout of FFHQ [28]. The source images are 50 high-resolution images from CelebA-HQ [32] and FFHQ [28]. Due to the limited space, we display the self-driving experimental results in the supplementary material.

Comparison Methods. We compared D^2Animator with both methods trained on real video sequences and StyleGAN-based inversion methods. We take two popular generation-based face animation methods FOMM [49] and Bi-layer [70] for comparison. Both of them are trained on real video sequences, whose performance is usually restricted by the resolution and quality of training videos. PSP [44] is the state-of-the-art inversion and conditional generation method based on StyleGAN. We use PSP [44] for animation by reproducing the source image conditioned on the segmentation of the target. StyleFusion [26] manipulates images through swapping facial parts. The images animated by StyleFusion [26] usually remain the face and background of the source, while the pose, eye and mouth may not be the same as the target.

Evaluation Metrics. In the experiment, we use four metrics to quantify the reality, identity-preserving and facial action accuracy of the animated images. We use Frechet Inception Distance (FID) [7] and Inception Score (IS) [45] to describe the reality and quality of the generated images. To measure the identity proximity of the source and the animated face, we computed the cosine similarity (CSIM) between extracted ArcFace [13] features. We evaluate the facial action accuracy by the normalized mean error (NME) [8] between the facial landmarks of the animated and target images.

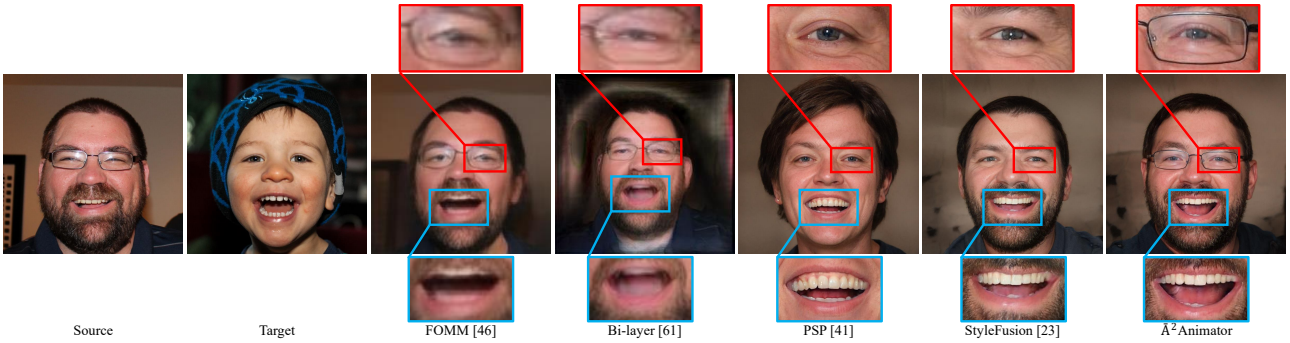

Figure 6: Details (eye and mouth) of animated faces in cross-identity experiment on FFHQ [28] .

Table 2: Quantitative results of ablation study.

Methods	NME(↓)	CSIM(↑)
w/o CluDistiller	4.2022	0.5681
w/o fine-tuning	3.7514	0.5663
D^{2}Animator	**3.6027**	**0.5689**

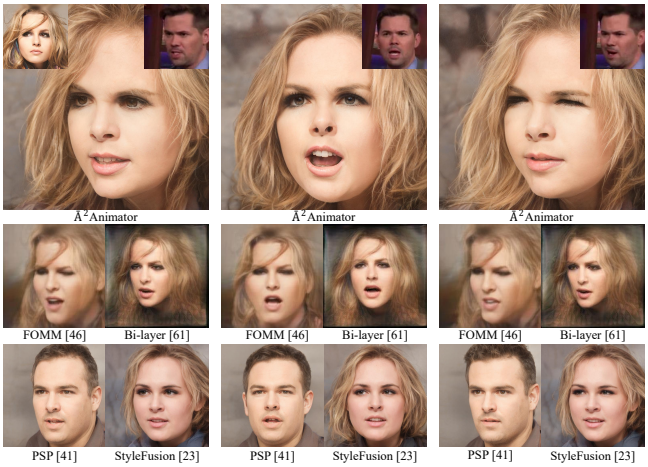

Figure 7: Sequence-driving animation results on CelebA-HQ.

Figure 8: Qualitative comparison of ablation study.

4.1 Cross-identity Animation

Cross-identity animation experiment compare D^{2}Animator with the other four methods on CelebA-HQ [32] and FFHQ [28], respectively. The qualitative comparisons of cross-identity situation are shown in Figure 5. Compared with other methods, ours can produce high-resolution results while keeping the identity better.

The quantitative results in Table 1 show that D^{2}Animator improves the reality of animated images by lower FID and IS scores. D^{2}Animator also outperforms the other methods in identity preservation as shown by the CSIM score. PSP [44] achieves the best NME score while failing to preserve the identity. Our method gains a comparable NME score with FOMM [49], which proves the effectiveness of learning facial action by distilling StyleGAN. Figure 6 displays an example animated face in detail. As shown in the red rectangle, our method can preserve the glasses during the animation while PSP [44] and StyleFusion [26] lose the detail. FOMM [49] and Bi-layer [70] can also generate the glasses, but the glass frames are distorted. The blue rectangles magnify the mouth, where our method produces the most similar teeth to the source.

4.2 Sequence-driving Experiment

The quantitative results of the sequence-driving experiment are reported in Table 1. Compared with generation-based methods FOMM [49] and Bi-layer [70], our method promotes reality and identity fidelity largely as shown by the FID, IS and CSIM scores. Figure 7 compares the visual quality of different methods. The images animated by D^{2}Animator are far more realistic than the generation-based methods FOMM [49] and Bi-layer [70]. D^{2}Animator outperforms inversion-based method PSP [44] in identity-preserving. Compared with StyleFusion [26], the pose and expression of the face animated by our method are much closer to the target.

4.3 Ablation Study

The proposed D^{2}Animator consists of two modules, *i.e.* CluDistiller and AugDistiller, and AugDistiller includes a training phase and a fine-tuning phase. As the training of AugDistiller is indispensable for face animation, we further show the efficiency of CluDistiller and fine-tuning in the ablation.

Compared with former methods, the CluDistiller can discover more various interpolation directions for face manipulation, which also improves the controllability of face animation. To verify the

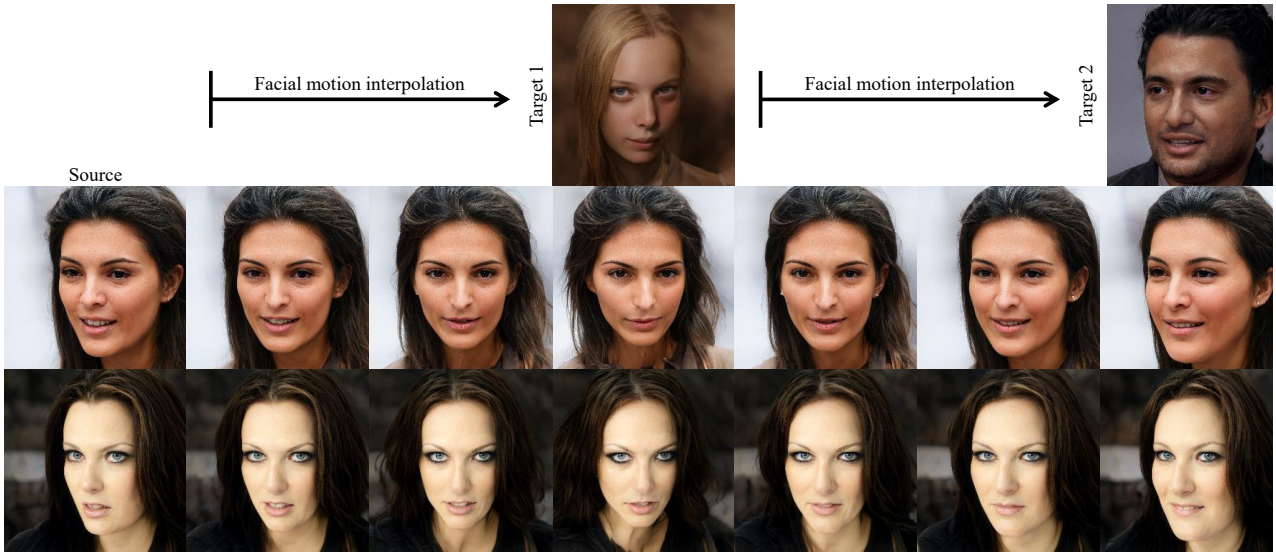

Figure 9: Facial action interpolation between the source and the targets.

Figure 10: Facial action disentanglement illustration.

effectiveness of our CluDistiller, we replace the interpolation directions collected by CluDistiller with the directions discovered by expression classifiers following [46]. Then we use the directions in the same way with the D^{2}Animator. The quantitative results in the Table 2 show that CluDistiller can provide more various manipulations which improves the accuracy and identity fidelity of face animation.

Fine-tuning the FAN of AugDistiller on the given source after training can further improve the identity fidelity and precision of face animation. We test the effect of fine-tuning by comparing the model trained on randomly augmented data with those fine-tuned on specific sources. As shown in Figure 8, the models fine-tuned on the specific sources can generate images with more similar poses and expressions to the target. The CSIM score and NME score reported in Table 2 also confirm the validity of fine-tuning for identity-preserving and accurate pose-retargeting.

5 APPLICATIONS

D^{2}Animator provides a powerful tool for various face animation tasks such as portrait frontalization, teleconference anonymous and video editing. Besides them, we show two novel applications of the proposed D^{2}Animator, *i.e.* facial action disentanglement and facial action interpolation.

Facial action disentanglement. D^{2}Animator enables us to retarget the pose and expression of the source at different targets respectively. We divide the interpolation directions collected by AugDistiller into two groups, *i.e.* pose and expression. Then we train AugDistiller with two groups of interpolation directions separately. We show the disentanglement of facial action in Figure 10 by animating sources to the pose shown on the left top and the expression provided on the left bottom.

Facial action interpolation. Our method can also be used to complement the intermediate frames of facial animation. Since D^{2}Animator realize face animation by interpolating in the latent space, our method can generate intermediate frames with more semantically realistic facial actions. As shown in Figure 9, D^{2}Animator can generate sequences of continuous facial actions between the source and target to smooth the animation video.

6 CONCLUSION

In this work, we propose D^{2}Animator, a dual distillation framework for high-resolution face animation. D^{2}Animator takes a Clustering-based Distiller to collect the interpolation directions that correspond to diverse facial actions in the latent space. Then D^{2}Animator uses an Augmentation-based Distiller to inverse facial actions into latent space interpolations. By assembling the two distillers, D^{2}Animator learns to generate high-resolution face animation videos without any real video sequences for training. D^{2}Animator outperforms the prior face animation methods in both reality and identity fidelity. The results shown in the experiments and applications confirm the superiority and versatility of D^{2}Animator.

ACKNOWLEDGMENTS

This work is supported by Science and Technology Innovation 2030 –"Brain Science and Brain-like Research" Major Project (No. 2021ZD0201405).

REFERENCES

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. 2019. Image2stylegan: how to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4432–4441.

[2] Rameen Abdal, Yipeng Qin, and Peter Wonka. 2020. Image2stylegan++: how to edit the embedded images? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8296–8305.

[3] Rameen Abdal, Peihao Zhu, Niloy Mitra, and Peter Wonka. 2021. Labels4free: unsupervised segmentation using stylegan. *arXiv preprint arXiv:2103.14968*.

[4] Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. 2021. Styleflow: attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics (TOG)*, 40, 3, 1–21.

[5] Hadar Averbuch-Elor, Daniel Cohen-Or, Johannes Kopf, and Michael F Cohen. 2017. Bringing portraits to life. *ACM Transactions on Graphics (TOG)*, 36, 6, 196.

[6] Dmitri Bitouk, Neeraj Kumar, Samreen Dhillon, Peter Belhumeur, and Shree K Nayar. 2008. Face swapping: automatically replacing faces in photographs. In *ACM Transactions on Graphics (TOG)* number 3. Vol. 27. ACM, 39.

[7] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. 2017. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 3722–3731.

[8] Adrian Bulat and Georgios Tzimiropoulos. 2017. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *International Conference on Computer Vision*.

[9] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. 2020. Simswap: an efficient framework for high fidelity face swapping. In *Proceedings of the 28th ACM International Conference on Multimedia*, 2003–2011.

[10] Zhuo Chen, Chaoyue Wang, Bo Yuan, and Dacheng Tao. 2020. Puppeteergan: arbitrary portrait animation with semantic-aware appearance transformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13518–13527.

[11] Mengyu Chu, You Xie, Jonas Mayer, Laura Leal-Taixé, and Nils Thuerey. 2020. Learning temporal coherence via self-supervision for gan-based video generation. *ACM Transactions on Graphics (TOG)*, 39, 4, 75–1.

[12] Edo Collins, Raja Bala, Bob Price, and Sabine Susstrunk. 2020. Editing in style: uncovering the local semantics of gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5771–5780.

[13] Jiankang Deng, Jia Guo, Xue Niannan, and Stefanos Zafeiriou. 2019. Arcface: additive angular margin loss for deep face recognition. In *CVPR*.

[14] Michail Christos Doukas, Stefanos Zafeiriou, and Viktoriia Sharmanska. 2021. Headgan: one-shot neural head synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14398–14407.

[15] Yuki Endo and Yoshihiro Kanamori. 2021. Few-shot semantic image synthesis using stylegan prior. *arXiv preprint arXiv:2103.14877*.

[16] Pablo Garrido, Levi Valgaerts, Ole Rehmsen, Thorsten Thormahlen, Patrick Perez, and Christian Theobalt. 2014. Automatic face reenactment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4217–4224.

[17] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *NIPS*.

[18] Hao Guan, Chaoyue Wang, and Dacheng Tao. 2021. Mri-based alzheimer's disease prediction via distilling the knowledge in multi-modal data. *NeuroImage*, 244, 118586.

[19] Yuxuan Han, Jiaolong Yang, and Ying Fu. 2021. Disentangled face attribute editing via instance-aware latent space search. *arXiv preprint arXiv:2105.12660*.

[20] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. 2020. Ganspace: discovering interpretable gan controls. *arXiv preprint arXiv:2004.02546*.

[21] Fengxiang He and Dacheng Tao. 2020. Recent advances in deep learning theory. *arXiv preprint arXiv:2012.10931*.

[22] Fengxiang He, Bohan Wang, and Dacheng Tao. 2020. Piecewise linear activations substantially shape the loss surfaces of neural networks. In *International Conference on Learning Representations*.

[23] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

[24] Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, 1501–1510.

[25] Aapo Hyvärinen and Erkki Oja. 2000. Independent component analysis: algorithms and applications. *Neural networks : the official journal of the International Neural Network Society*, 13 4-5, 411–30.

[26] Omer Kafri, Or Patashnik, Yuval Alaluf, and Daniel Cohen-Or. 2021. Stylefusion: a generative model for disentangling spatial segments. *arXiv preprint arXiv:2107.07437*.

[27] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. 2020. Training generative adversarial networks with limited data. *arXiv preprint arXiv:2006.06676*.

[28] Tero Karras, Samuli Laine, and Timo Aila. 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.

[29] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8110–8119.

[30] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60, 84–90.

[31] Sam Kwong, Jialu Huang, and Jing Liao. 2021. Unsupervised image-to-image translation via pre-trained stylegan2 network. *IEEE Transactions on Multimedia*.

[32] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. 2020. Maskgan: towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[33] Kathleen M Lewis, Srivatsan Varadharajan, and Ira Kemelmacher-Shlizerman. 2021. Vogue: try-on by stylegan interpolation optimization. *arXiv preprint arXiv:2101.02285*.

[34] Daiqing Li, Junlin Yang, Karsten Kreis, Antonio Torralba, and Sanja Fidler. 2021. Semantic segmentation with generative models: semi-supervised learning and strong out-of-domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8300–8311.

[35] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. 2020. Advancing high fidelity identity swapping for forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5074–5083.

[36] Huan Ling, Karsten Kreis, Daiqing Li, Seung Wook Kim, Antonio Torralba, and Sanja Fidler. 2021. Editgan: high-precision semantic image editing. In *Thirty-Fifth Conference on Neural Information Processing Systems*.

[37] Yu-Ding Lu, Hsin-Ying Lee, Hung-Yu Tseng, and Ming-Hsuan Yang. 2020. Unsupervised discovery of disentangled manifolds in gans. *arXiv preprint arXiv:2011.11842*.

[38] Tianxiang Ma, Dongze Li, Wei Wang, and Jing Dong. 2021. Face anonymization by manipulating decoupled identity representation. *arXiv preprint arXiv:2105.11137*.

[39] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. Voxceleb: a large-scale speaker identification dataset. *Telephony*, 3, 33–039.

[40] Yuval Nirkin, Yosi Keller, and Tal Hassner. 2019. Fsgan: subject agnostic face swapping and reenactment. In *Proceedings of the IEEE/CVF international conference on computer vision*, 7184–7193.

[41] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. 2019. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2337–2346.

[42] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. 2021. Styleclip: text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2085–2094.

[43] Antoine Plumerault, Hervé Le Borgne, and Céline Hudelot. 2020. Controlling generative models with continuous factors of variations. *arXiv preprint arXiv:2001.10238*.

[44] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. 2021. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2287–2296.

[45] Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. In *NIPS*, 2226–2234. http://papers.nips.cc/paper/6125-improved-techniques-for-training-gans.

[46] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. 2020. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9243–9252.

[47] Yujun Shen and Bolei Zhou. 2021. Closed-form factorization of latent semantics in gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1532–1540.

[48] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019. Animating arbitrary objects via deep motion transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2377–2386.

[49] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019. First order motion model for image animation. *Advances in Neural Information Processing Systems*, 32, 7137–7147.

[50] Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. 2021. Motion representations for articulated animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13653–13662.

[51] Robert R. Sokal and Peter H. A. Sneath. 1961. *Principles of Numerical Taxonomy*. W. H. Freeman.

[52] Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. 2020. Stylerig: rigging stylegan for 3d control over portrait images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6142–6151.

[53] Yu Tian, Jian Ren, Menglei Chai, Kyle Olszewski, Xi Peng, Dimitris N Metaxas, and Sergey Tulyakov. 2021. A good image generator is what you need for high-resolution video synthesis. *arXiv preprint arXiv:2104.15069*.

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

[55] Tengfei Wang, Yong Zhang, Yanbo Fan, Jue Wang, and Qifeng Chen. 2022. High-fidelity gan inversion for image attribute editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[56] Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Jan Kautz, and Bryan Catanzaro. 2019. Few-shot video-to-video synthesis. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 5013–5024.

[57] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018. Video-to-video synthesis. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 1152–1164.

[58] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8798–8807.

[59] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. 2021. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10039–10049.

[60] Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva. 2021. Latent image animator: learning to animate images via latent space navigation. In *International Conference on Learning Representations*.

[61] Olivia Wiles, A Koepke, and Andrew Zisserman. 2018. X2face: a network for controlling face generation using images, audio, and pose codes. In *Proceedings of the European conference on computer vision (ECCV)*, 670–686.

[62] Jianjin Xu and Changxi Zheng. 2021. Linear semantics in generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9351–9360.

[63] Ceyuan Yang, Yujun Shen, and Bolei Zhou. 2021. Semantic hierarchy emerges in deep generative representations for scene synthesis. *International Journal of Computer Vision*, 129, 5, 1451–1466.

[64] Zuopeng Yang, Daqing Liu, Chaoyue Wang, Jie Yang, and Dacheng Tao. 2022. Modeling image composition for complex scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7764–7773.

[65] Xu Yao, Alasdair Newson, Yann Gousseau, and Pierre Hellier. 2021. A latent transformer for disentangled face editing in images and videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13789–13798.

[66] Fei Yin et al. 2022. Styleheat: one-shot high-resolution editable talking face generation via pretrained stylegan. *arXiv preprint arXiv:2203.04036*.

[67] Shan You, Tao Huang, Mingmin Yang, Fei Wang, Chen Qian, and Changshui Zhang. 2020. Greedynas: towards fast one-shot nas with greedy supernet. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1999–2008.

[68] Shan You, Chang Xu, Chao Xu, and Dacheng Tao. 2017. Learning from multiple teacher networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1285–1294.

[69] Baosheng Yu and Dacheng Tao. 2021. Heatmap regression via randomized rounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[70] Egor Zakharov, Aleksei Ivakhnenko, Aliaksandra Shysheya, and Victor Lempitsky. 2020. Fast bi-layer neural synthesis of one-shot realistic head avatars. In *European Conference of Computer vision (ECCV)*. (Aug. 2020).

[71] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. 2019. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9459–9468.

[72] Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. 2021. Datasetgan: efficient labeled data factory with minimal human effort. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10145–10155.

[73] Haimei Zhao, Wei Bian, Bo Yuan, and Dacheng Tao. 2020. Collaborative learning of depth estimation, visual odometry and camera relocalization from monocular videos. In *IJCAI*, 488–494.

[74] Yuhao Zhu, Qi Li, Jian Wang, Cheng-Zhong Xu, and Zhenan Sun. 2021. One shot face swapping on megapixels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4834–4844.