PuppeteerGAN: Arbitrary Portrait Animation with Semantic-aware Appearance Transformation

Zhuo Chen[1,2] Chaoyue Wang[2] Bo Yuan[1] Dacheng Tao[2]
[1]Shenzhen International Graduate School, Tsinghua University
[2]UBTECH Sydney AI Centre, School of Computer Science, Faculty of Engineering,
The University of Sydney, Darlington, NSW 2008, Australia
z-chen17@mails.tsinghua.edu.cn,chaoyue.wang@sydney.edu.au,
yuanb@sz.tsinghua.edu.cn,dacheng.tao@sydney.edu.au

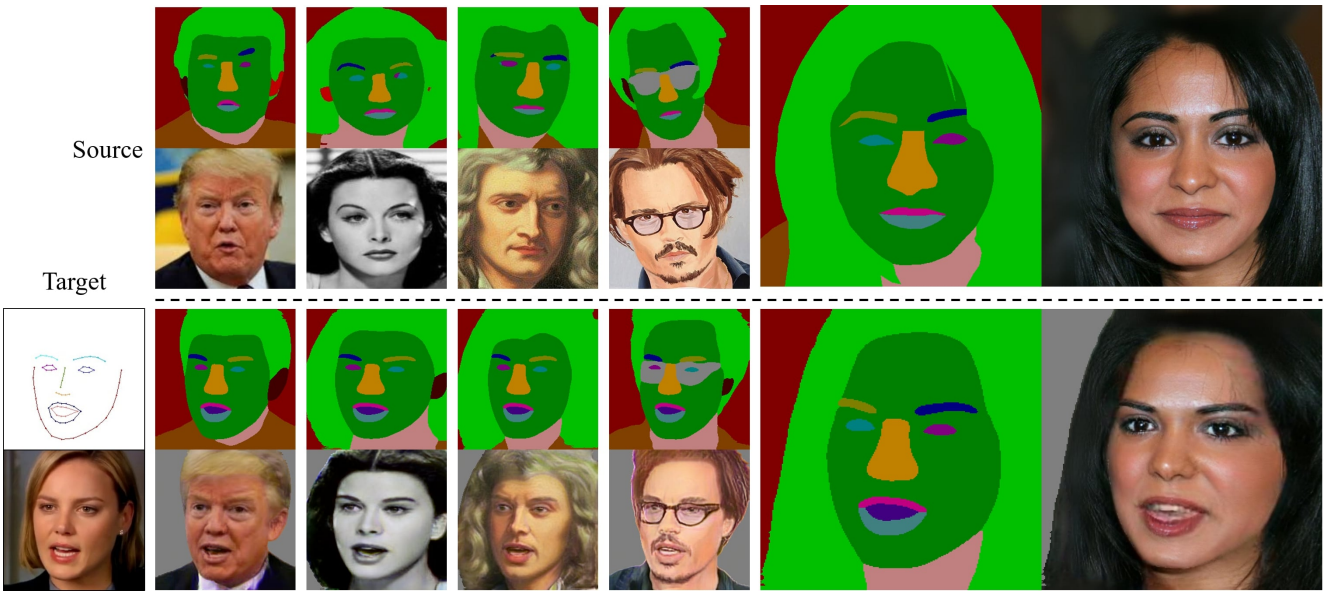

Figure 1: **Examples of animated portraits generated by the proposed PuppeteerGAN.** The results are at the same pose as the target frame (left column) while keeping the same appearance of the source (top row). As shown in the source images, our method can be applied to various portraits including color photos, black-and-white photos, paintings, cartoon characters and high-resolution images.

Abstract

Portrait animation, which aims to animate a still portrait to life using poses extracted from target frames, is an important technique for many real-world entertainment applications. Although recent works have achieved highly realistic results on synthesizing or controlling human head images, the puppeteering of arbitrary portraits is still confronted by the following challenges: 1) identity/personality mismatch; 2) training data/domain limitations; and 3) low-efficiency in training/fine-tuning. In this paper, we devised a novel two-stage framework called PuppeteerGAN for solving these challenges. Specifically, we first learn identity-preserved semantic segmentation animation which executes pose retargeting between any portraits. As a general representation, the semantic segmentation results could be adapted to different datasets, environmental conditions or appearance domains. Furthermore, the synthesized semantic segmentation is filled with the appearance of the source portrait. To this end, an appearance transformation network is presented to produce fidelity output by jointly considering the wrapping of semantic features and conditional generation. After training, the

1. Zhuo Chen has been a visiting PhD student at UBTECH Sydney AI Centre, School of Computer Science, Faculty of Engineering, The University of Sydney, since January 2019.

two networks can directly perform end-to-end inference on unseen subjects without any retraining or fine-tuning. Extensive experiments on cross-identity/domain/resolution situations demonstrate the superiority of the proposed PuppetterGAN over existing portrait animation methods in both generation quality and inference speed.

1. Introduction

Portraits including paintings, photographs or other artistic representations of human beings have always been one of the most important research objects in computer vision and graphics. In this work, we consider the task of portrait animation, which aims to animate a given portrait using poses provided by driven frames. This kind of technique has attracted great attention throughout the community because of its potentially wide usage in the film industry [7], art-making [47] and personalized media generation [35].

In recent years, with the rapid development of deep learning algorithms, some deep generative models are proposed for solving portrait animation tasks. Among them, one kind of solution is trained to decompose the disentangled appearance and pose representations from the input portrait image [36, 25, 46]. Ideally, by recombining the appearance feature of the source portrait with the pose feature of the target frames, the generator/decoder is supposed to generate the desired outputs. Though promising progress has been made, the challenges remain on extracting the desired appearance and pose representations from unseen images (or videos) during inference.

Facial landmarks, which can be conveniently detected by recent techniques [38, 18, 31], are regarded as one kind of replacement of the disentangled pose representation. In [47, 33, 10], given the landmarks of the target frame as input, the network is trained to reconstruct the target frame by conditioning on appearance information extracted from other frames of the same video/person. Compare to directly disentangling pose representations from images, methods using facial landmarks are more robust and usually with better generation results. However, during inference, it may encounter a misalignment between the driven landmarks and the source portrait (*e.g*. different face shape), which will result in poor results. In addition, considering the appearance representation may fail to be extracted from unseen portraits that following different distribution with training data [4, 22], the few-shot fine-tuning strategy is employed to learn accurate appearance information [18, 45]. By utilizing a few images of the same person to fine-tune the pre-trained model, these methods will achieve better results for each specific identity. However, in real-world applications, the fine-tuning would cost much more time and computation resources than simply feed-forward inference.

Therefore, although recent works have achieved convincing results on synthesizing or controlling portraits, we argue that the puppeteering of arbitrary portraits is still troubled by the following challenges: 1) identity/personality mismatch between the animated source portrait with the provided driven frames; 2) training data/domain limitations which lead the pre-trained model failing to understand unseen portraits from other identities, domains or resolutions; 3) low-efficiency retraining/fine-tuning which may cost a considerable amount of time and computation resource in real-world applications.

In this work, we proposed a novel two-stage generation framework called PuppetterGAN for arbitrary portrait animation tasks. Different from existing pose/appearance decomposition strategies, we separate portrait animation into two stages: pose retargeting and appearance transformation. In the first stage, we aim to perform the identity-preserved pose retargeting between the semantic segmentation of any portraits. Specifically, a sketching network is trained to synthesize the animated segmentation masks and landmarks that keep characteristic details (*e.g*. facial shape, hairstyle) of the source portrait yet with the same pose as the driven frame. Since the generated results could act as a general representation of different kinds of portraits, we can simply train the sketching network on a specific talking-head video dataset but perform inference on arbitrary portraits. In the second stage, we devised an appearance transformation network to fill in the animated semantic segmentation mask with the appearance of the source portrait. This network consists of an appearance encoder, a segmentation mask conditioned decoder and the proposed Warp-based semantic-Aware SkIp-Connections (WASIC). The coloring network makes full use of the texture information extracted by the shallow layers of the encoder through the proposed WASIC, thus, it avoids the fine-tuning/retraining step during the inference.

Finally, we show the generated portraits of our method on various experiments including self-driven, cross-identity/domain/resolution cases and compare the proposed method with five different methods [41, 45, 1, 43, 40]. Experiment results demonstrate our method is superior to the existing work in fidelity, reality, generalization ability and inference speed.

2. Related Works

Deformation based methods. Traditional deformation based algorithms compute a transformation from the source portrait to the target pose based on facial landmarks or optical flow. Since the large scale deformation are prone to distortion, some methods [20, 3, 11] builds a face image database of the source person to retrieve the most similar expression to the target as the basis of deformation. Although these methods succeed in face mimic for the specific source person, collecting and pre-processing such a large dataset

for each source person is a high cost in practice.

Averbuch *et al*. [1] detects additional key points around the face on the source and target image in order to control the deformation of the whole head. By copying the mouth region from the target, this method can manipulate facial expression without any training or extra database on a still portrait. X2face [43] learns a pair 2D pixel-wise deformation from the source pose to the frontal and the target pose which can be extracted from multiple optional media including video, audio and pose angle. Overall, Deformation based methods are efficient in transferring facial expression in both identity fidelity and face reality with a relatively small computational cost. However, a large motion of the head or the synthesizing of the hidden part will lead to distortion of the generated portraits.

Video-to-video. Training a specific network for every single person can boost the quality of the generated portraits significantly. Vid2Vid [40] can generate temporally coherent videos conditioned on segmentation masks, sketches, and poses after being trained on the specific video sequences of the same appearance. Its few-shot adaptive variation [39] learns to synthesize videos of previously unseen subjects or scenes by leveraging few example images of the target at test time. DeepFake [7] builds an autoencoder framework consist of an pose encoder and an appearance decoder. In the inference phase, they employ the encoder of the driven identity to extract pose and the decoder of the source person to recover appearance. Because the re-implementation and training process is easy to follow, DeepFake becomes popular for face swapping applications. ReenactGAN [44] generates mimic videos guided by the face boundary transformed into the source identity. While the encoder can be generic, the decoder and transformer are specific to each source identity. Moreover, 3D morphable model based methods also require individual training for each given source person. For example, Hyeongwoo *et al*. [17] trains a rendering-to-video network to render the 3D model with the target pose and expression conditioned by the illustration and identity of the source. [34] aims to learn accurate lip motions driven by the video, while numerous hours of the target person's speech footage is vital for training.

Although video-to-video methods are able to generate realistic portrait even with large motion, the identity specific training prevents most common users from using this technology since the difficulty of collecting training data, high computational cost and long inference time.

Conditional image synthesis. Benefit from the progress in (conditional) generative adversarial networks (GANs) [48, 6, 37, 28, 21], some synthesis based methods [45, 26, 43] are proposed to generate fake portraits using the identity of the source person and the pose (or facial expression) of the target image.

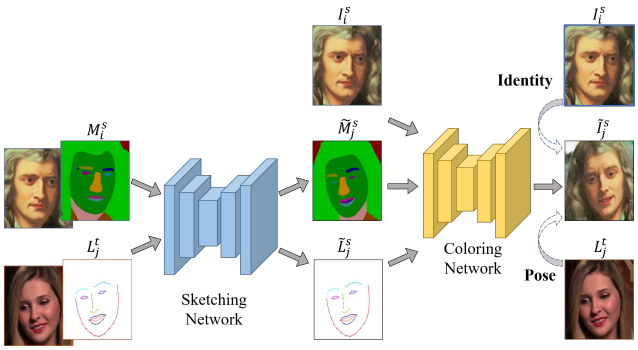

Figure 2: The complete framework of the proposed method. PuppeteerGAN performs portrait animation in two stages: 1) pose retargeting by the sketching network and 2) appearance transformation by the coloring network.

Various representations are introduced as the conditional signal for portrait animation. Landmarks [45, 33, 10] detected from the driven frame are the most frequently used description of the target pose. Nirkin *et al*. [26] and Geng *et al*. [12] warp the input portrait to the target pose based on the landmarks and further fills in the missing part with an inpainting algorithm. Pumarola *et al*. [29] utilizes the action unit in a similar way. Although these methods perform well in some cases, as mentioned before, using the landmarks detected from the driven frame may result in the identity mismatch issue when there is a large gap between the source and driven portrait. paGAN [23] generates videos of photo-real faces conditioned on the deformation of 3D expression meshes which is only applicable to facial area rather than the whole portrait. Moreover, some works [27, 2, 32] attempt to disentangle the shape and appearance features from the portrait through adversarial learning. Although the learning of identity and pose feature can mitigate the personality mismatch problem, the joint learning of pose retargeting and appearance is a challenging task and may limit by the training data/domain.

3. Method

Given a source portrait I_i^s and a target/driven frame I_j^t, the proposed PuppeteerGAN aims to generate an output image $\tilde{I}_j^s$, whose identity is the same as I_i^s, meanwhile, being consistent with I_j^t in pose and expression. In addition, the facial landmarks L_i^s / L_j^t and semantic segmentation mask M_i^s are detected for assisting our portrait animation process. In the following, we introduce our pipeline in Section 3.1. Section 3.2 and Section 3.3 demonstrate the sketching and coloring network, respectively.

3.1. Model framework

As aforementioned, some generation based methods synthesize the animated portrait with the facial landmark detected on the driven frame as the condition signal. Com-

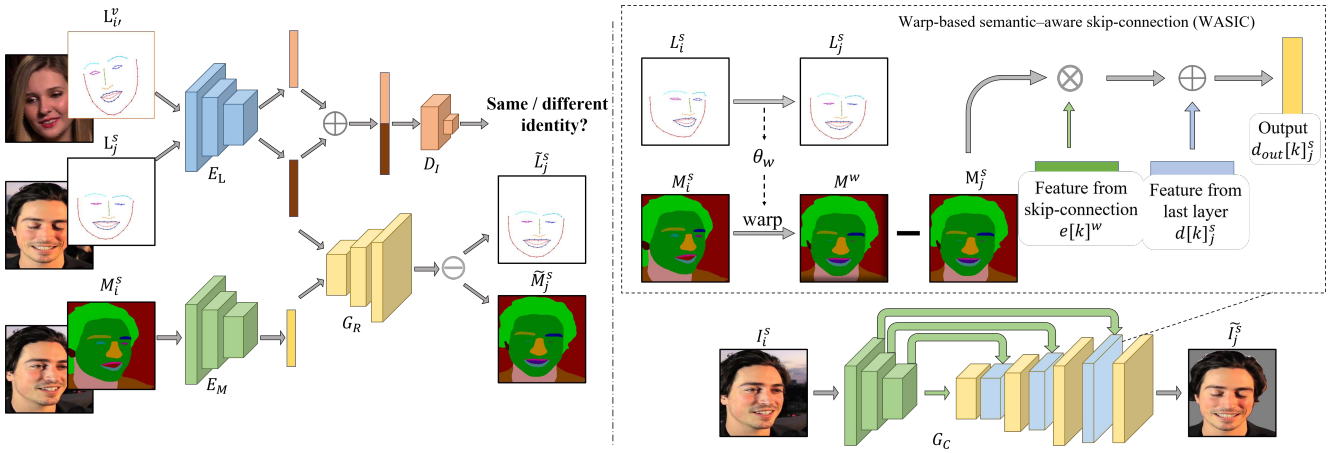

Figure 3: **The training process of the sketching network and the coloring network.** The sketching network (left) consists of two encoders (for segmentation mask and landmark), a generator and a discriminator. Similar to U-net [30], the coloring network (right bottom) includes an encoder, a generator and several proposed warp-based semantic-aware skip-connections (WASIC). We illustrate the structure of WASIC in the dotted box (right top).

pared with facial landmark, the segmentation mask is a pixel-wise semantic classification of the portrait, which marks the external region of the face such as the ears, the hair and the neck as well. Therefore, we take advantage of the segmentation mask and use facial landmark as auxiliary to generate realistic images.

We proposed a two-stage generation pipeline (Fig. 2) including sketching and coloring network. First, identity-preserved pose retargeting is solved in the first sketching stage. Specifically, We take the segmentation mask M_i^s of the source portrait and landmark L_j^t of the driven frame as input, intending to generate the target segmentation mask $\tilde{M}_j^s$ and landmark $\tilde{L}_j^s$. Through introducing an identity discriminator, the generated $\tilde{M}_j^s$ and $\tilde{L}_j^s$ are learned to be consistent with the identity of M_i^s and the pose of L_j^t. Since the segmentation mask and facial landmark could be applied to any portraits, this module is naturally adaptive to images of different domains.

In the second stage, our coloring network performs appearance transformation conditioned on the generated segmentation mask $\tilde{M}_j^s$ and the source portrait I_i^s. Benefit from the generated segmentation mask, this stage generates realistic outputs based on the precise geometry guidance and does not need to pay attention to the consistency with the target pose. In order to make full use of the texture features extracted by the shallow layers of the encoder network, we devised the Warp-based semantic-Aware SkIp-Connections (WASIC).

3.2. Sketching network

During training, the input of the sketching network is a group of landmarks and segmentation masks denoted as $[L_i^s, L_j^s, L_{i'}^v, M_i^s, M_j^s]$. L_i^s and L_j^s are facial landmarks detected from two frames in the same video sequence. $L_{i'}^v$ is the landmark of another person. M_i^s and M_j^s denote the segmentation masks of the source portrait at pose i and j.

Figure 3 - left illustrates the structure of the sketching network. The network consists of four components, the landmark encoder E_L, the segmentation encoder E_M , the identity discriminator D_I and the joint generator G_R. The generation process can be formulated as:

$$(\tilde{L}_j^s, \tilde{M}_j^s) = G_R\big(E_L(L_j^s), E_M(M_i^s)\big). \tag{1}$$

The generated segmentation mask and landmark should be the same as M_j^s and L_j^s, which is detected from frame j of the video/identity s. The reconstruction losses are:

$$\mathcal{L}_{ld}(E_L, E_M, G_R) = ||\tilde{L}_j^s - L_j^s||, \tag{2}$$

$$\mathcal{L}_{seg}(E_L, E_M, G_R) = ||\tilde{M}_j^s - M_j^s||. \tag{3}$$

Identity-preserved pose retargeting. In order to avoid identity/personality mismatch, we introduce an identity discriminator D_I [9]. The input of D_I is a pair of features extracted from the facial landmarks by E_L. The discriminator is required to determine whether the identities of the landmarks are the same or not. For each step, a real pair $(E_L(L_i^s), E_L(L_j^s)))$ and a fake pair $(E_L(L_{i'}^v), E_L(L_j^s)))$ are used for training. Meanwhile, we train E_L to fool the discriminator. Finally, the outputs of landmark encoder E_L portray the pose and expression implicitly while being identity indistinguishable. The adversarial training process can be defined as:

$$\begin{aligned}\mathcal{L}_{idt}(E_L, D_M) = \min_{E_L}\max_{D_I} & D_I(E_L(L_i^s), E_L(L_j^s)) \\ & - D_I(E_L(L_{i'}^v), E_L(L_j^s)).\end{aligned} \tag{4}$$

Finally, the total loss for training the network is:

$$\begin{aligned}&\mathcal{L}_{seg}(E_L, E_M, G_R, D_I) = \mathcal{L}_{ld}(E_L, E_M, G_R) \\ &+\lambda_1 \cdot \mathcal{L}_{seg}(E_L, E_M, G_R) + \lambda_2 \cdot \mathcal{L}_{idt}(E_L, D_I).\end{aligned} \tag{5}$$

Since the sketching network deals with the pose retargeting on semantic guidance (segmentation mask and landmark) instead of the real portraits, this module can be potentially applied to different kinds of portraits regardless of the domain gaps. Through the disentanglement and recombination of the identity and pose feature, the sketching phase completes the pose retargeting task precisely in our framework. The outputs of the sketching network will be used as the geometry conditional signals in the next coloring network.

3.3. Coloring network

The coloring network aims to synthesize the target portraits based on the source images and the geometry guidance generated in the former sketching phase. The inputs are [I_i^s, I_j^s, L_i^s, L_j^s, M_i^s, M_j^s], standing for the portraits, landmarks, and segmentation masks of two frames in the same sequence.

Here, we utilize the segmentation mask M_j^s and source image I_i^s as input, and aim to synthesize the portrait image $\tilde{I}_j^s$. Since the pose retargeting problem is solved in the sketching stage, we directly utilize the segmentation mask M_j^s extracted from the I_j^s as conditional input, and attempt to synthesize the portrait $\tilde{I}_j^s$ based on the source image I_i^s in the training phase. In this stage, the challenge remains on the appearance transformation between different frames of the same person. First, we observe that, for the generated image $\tilde{I}_j^s$, most of its appearance information could be directly found in the input image I_i^s. Inspired by deformation based methods, we devised the Warp-based semantic-aware skip-connection(WASIC) for transforming these appearances. However, for unseen parts (*e.g.* open mouth), we hope that the coloring network could work as a conditional generation network, which is able to imagine these parts based on the input images.

As shown in Fig. 3 - right, the structure of the coloring network is based on U-net [30]. We replace the single convolution layers with residual blocks [13] in both encoder and decoder. In order to constrain the generation with the segmentation mask, we use spatially-adaptive normalization (SPADE) [28] instead of the batch normalization [15] in the decoder. We attach the shallow layers of encoder to the corresponding layers of decoder through the proposed WASIC.

Warp-based semantic-Aware SkIp-Connection (WASIC). Skip-connection used in U-net [30] between the corresponding layers of the encoder and decoder can improve the generated result by bridging the features extracted by the encoder to the decoder. However, straightforward skip-connection between the encoder and decoder is unhelpful for our problem because of the geometry misalignment between the source and the target frames. Therefore we proposed the WASIC to warp and transpose the appearance features from the encoder to the decoder. Specifically, given an input group, we first compute a formulated transformation parameter θ_w from the source landmark L_i^s to the target L_j^s. For the k^{th} layer of the network, the intermediate output of the encoder and the corresponding output of the decoder are denoted as $e[k]_i^s$ and $d[k]_j^s$, and the two features are of the same size. Both M_i^s and M_j^s are resized to the equal geometry scale as well. Then, $M[k]_i^s$ and $e[k]_i^s$ are warped to $M[k]^w$ and $e[k]^w$ according to θ_w. Finally, the warped feature $e[k]^w$ and the generated feature $d[k]_j^s$ are weighted and added up to be the output of this layer directed by the semantic segmentation masks. This step is shown in Fig. 3 - right and formulated as:

$$\begin{aligned} d_{out}[k]_j^s =& \mu \cdot M[k]^f \cdot e[k]^w + (1-\mu) \cdot d[k]_j^s \\ &+ \mu \cdot (1 - M[k]^f) \cdot d[k]_j^s, \end{aligned} \tag{6}$$

where $M[k]^f$ is a mask of the same part of $M[k]^w$ and $M[k]_j^s$ and μ is a learned weight parameter.

Geometry dropout strategy. We further expand the available data for training the coloring network from video dataset to image dataset through the proposed geometry dropout strategy. Since we accomplished pose retargeting in the former sketching stage, our coloring phase can be regarded as an image-to-image transformation task. Furthermore, the proposed skip-connection allows us to change the training process by modifying the segmentation masks. Based on these, we adopt a simple but useful geometry dropout strategy to train the network on the image dataset. For training on one image, we randomly zero one or more parts of the source portrait I_i^s and segmentation mask M_i^s to form I^w, e^w and M^w. As the geometry dropout is almost the same as the gap caused by deformation for the generation network, the image dataset can play the same role as the video dataset for the training of the coloring network.

Training loss for the coloring network is a combination of several widely used loss functions for image generation,

$$\begin{aligned} \mathcal{L}_{col}(G_C, D_C) = \mathcal{L}_{rec}(G_C) &+ \gamma_1 \cdot L_{perc}(G_C) \\ &+ \gamma_2 \cdot \mathcal{L}_{GAN}(G_C, D_C) + \gamma_3 \cdot \mathcal{L}_{feat}(G_C, D_C). \end{aligned} \tag{7}$$

The first term is an $L1$ reconstruction loss and the second term measures the perceptual distance by a VGG16 network pretrained on VGGFace2 [5]. Moreover, we adapt a multi-scale patch discriminator similar to PatchGAN [16] and add feature matching loss to all the discriminators.

Profited from the proposed WASIC, the generalization ablility of our coloring network can be improved a lot, and no fine-tuning is required for any specific input portrait. Therefore, our network can cost much less time and computation resource in the inference phase.

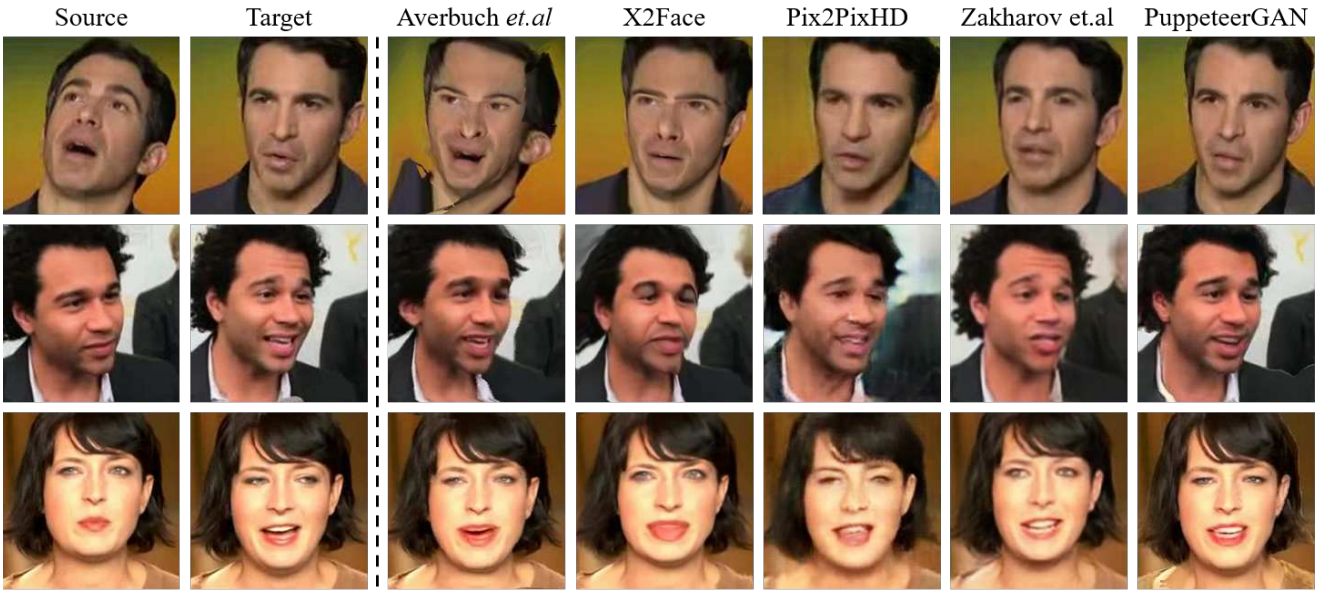

Figure 4: **Self-driven portrait animation results.** The animated portrait was driven by another frame in the same video. In each row, the result of Averbuch *et al*. [1], X2Face [43], Pix2PixHD [41], Zakharov *et al*. [45] and PuppeteerGAN are shown.

4. Experiment

In this section, we evaluated the proposed PuppeteerGAN in terms of generation quality, generalization capability and extensibility. First, we evaluated the framework on self-driven sequence in both quality and quantity. Then we tested the generalization capability of the proposed framework in cross-identity and cross-domain experiment. Finally, we trained the coloring network on image dataset to animate high-resolution images.

Comparison methods. We compared the proposed method with five previous methods: two general conditional generation based methods Pix2PixHD [41] and Vid2Vid [40], a pixel-wise deformation method X2Face [43], a geometry warp method Averbuch *et al*. [1] and the latest generation based method Zakharov *et al*. [45]. For X2Face [43], we adapted the official code and pretrained model. We reimplemented the methods proposed by Averbuch *et al*. [1] and Zakharov *et al*. [45]. Except Averbuch *et al*. [1], the other four compared methods require fine-tuning/retraining for each source portrait to generate comparable result.

Datasets. In comparison experiments, our model was trained on a subset of VoxCeleb1 [24] with video sequences of 672 person for training and 50 for test. We trained the Pix2PixHD model with the same settings. For comparison, we tested the methods on the test split of VoxCeleb1 [24] and the CelebMask-HQ [19] with resolution reduction. In the cross-domain experiments, we collected 300 portraits of different appearance domains on the Internet. For the cross-resolution experiment, we trained and tested the coloring network on CelebMask-HQ [19].

Methods	Vox [24]				
	SSIM↑	FID↓	PSNR↑	CSIM↑	MSE↓
Averbuch *et al*. [1]	0.6733	73.1115	**31.4702**	0.7600	0.1502
Pix2PixHD [41]	0.5500	70.3599	29.2691	0.4145	0.1876
PuppeteerGAN	0.7255	33.6119	31.3506	**0.8178**	**0.1033**
Zakharov *et al*. [45] (1)	0.6700	43.0000	-	-	-
X2Face [43] (1)	0.6800	45.8000	-	-	-
Pix2PixHD [41] (1)	0.5727	67.4887	29.6024	0.4789	0.1689
Zakharov *et al*. [41] (8)	0.7100	38.0000	-	-	-
X2Face [43] (8)	0.7300	51.5000	-	-	-
Pix2PixHD [41] (8)	0.5854	66.6279	29.8199	0.5117	0.1567
Zakharov *et al*. [41] (32)	0.7400	**29.5000**	-	-	-
X2Face [43] (32)	**0.7500**	56.5000	-	-	-
Pix2PixHD [41] (32)	0.6072	64.6087	30.2076	0.6082	0.1463
Vid2Vid [40]	0.6744	51.2171	31.4291	0.7715	0.1265

Table 1: Quantitative results. The measurement of the statistic distances between the generated portraits and GT.

4.1. Self-driven experiment

Following [45], we chose the source portrait and the driven frame from the same video. We compared the methods on the test split of VoxCeleb1 [24], which contains 312 videos sequences of 50 different people. All the identities were unseen in the training phase. We selected 16 pairs of frames for each identity randomly.

As shown in Fig. 4, Averbuch *et al*. [1] can change the facial expression well while fails to mimic large scale actions. X2face [43] preserves the texture of the source portrait and performs better on large deformations, but it is hard to generate the unseen parts such as an open mouth. The generation based methods, Pix2PixHD [41] and Zakharov *et al*. [45] both suffer from the artifact and blur in the results.

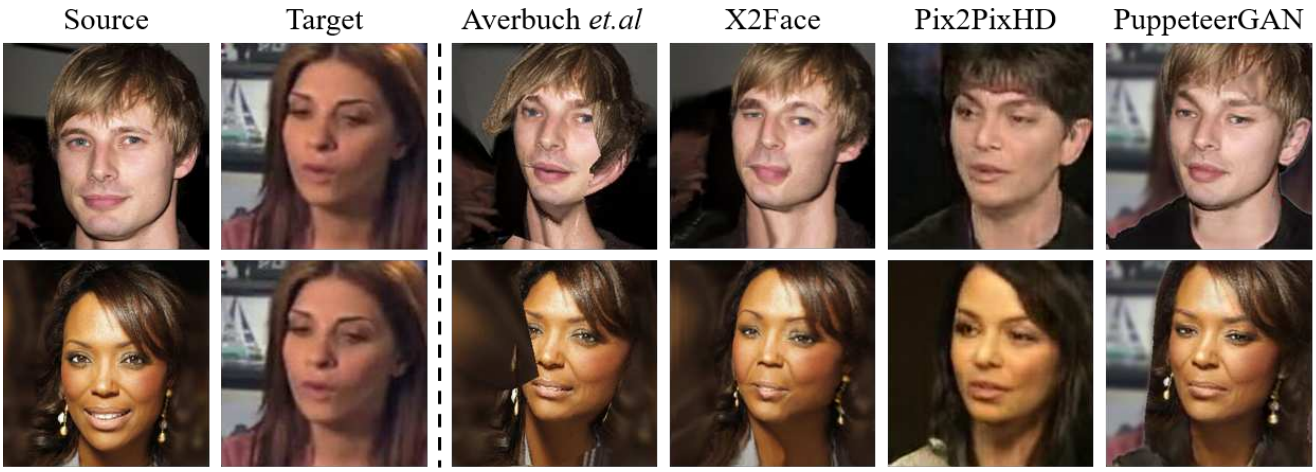

Figure 5: **Cross-identity portrait animation results.** We animated each source portrait by a target frame of a different person. We compared the proposed method with Averbuch *et al*. [1], X2face [43] and Pix2PixHD [41].

Methods	Time Cost (s)		
	Fine-tuning	Inference	Sum
Averbuch *et al*. [1]	-	0.9396	0.9396
X2Face [43]	4.3800	0.2257	4.6056
Pix2PixHD [41]	28.0415	0.2559	28.2974
Zakharov *et al*. [45]	48.7117	1.0220	49.7334
PuppeteerGAN	-	0.6117	**0.6117**

Table 2: Comparisons of the average time cost per frame.

Our method outperforms the previous methods in both reality and fidelity. Due to the segmentation mask guided generation, our method is not troubled by the problem of large scale motion. Profited from WASIC, our method can preserve the details and imagine the hidden parts of the source.

We also evaluated the proposed method in quantity as displayed in Table 1. We used Mean squared error (MSE), Peak signal-to-noise ratio (PSNR) and Structured similarity (SSIM) [42] to assess the statistical error of the generated image according to the target image. Moreover, We took Frechet-inception distance (FID) [14] and Cosine similarity (CSIM) [8] to quantify the reality and fidelity of the result respectively. Since the results of our re-implementation did not reach the expecting performance, we copied the scores from Zakharov *et al*. [45].

The quantitative results also demonstrates the efficiency of our method. Especially, our method outperforms the other methods largely in the zero-shot inference, and still surpasses their fine-tuned results based on several metrics.

The time cost experiment was conducted to compare the efficiency of different methods. We compared the inference speed by measuring the average time cost of animating a given portrait to target pose for each method except Vid2Vid [40] because the time cost of retraining is much more than fine-tuning. As our method requires no fine-tuning, it is far more efficient than the other methods as shown in Table 2.

4.2. Cross-identity experiment

We evaluated the proposed PuppeteerGAN through animating a source portrait by a driven frame of another person, which demonstrates our method can alleviate the problem of identity mismatch between the source and target portrait. Since the two frames were extracted from different videos, it increases the difficulty of portrait animation, which may result in low fidelity of the generated portraits.

We compared PuppeteerGAN with three different methods as illustrated in Fig. 5. Because of the large gap between the landmark of the source and target, the results of warp based methods, Averbuch *et al*. [1] and X2face [43] were distorted. The generation based method Pix2PixHD [41] was able to generate realistic portraits, but failed to preserve the identity of the source portraits. The results of PuppeteerGAN shown in the last column of Fig. 5 are superior to the other methods in both reality and fidelity. By animating two source portraits to the same target pose, we displayed the identity preserving ability of our method in both geometry shape (*e.g*. face shape and hair style) and appearance(*e.g*. the texture and color of skin and eyes).

4.3. Cross-domain experiment

The proposed PuppetterGAN aims to animate arbitrary source portraits to mimic the target frame. Our cross-domain experiments demonstrate that our method is applicable for diverse kinds of portraits without fine-tuning. As shown in Fig. 6, firstly, our sketching network transforms the shape from source to target based on segmentation mask and landmark, which is domain-adaptive to different kinds of portraits. Then our coloring network synthesizes realistic portraits based on the generated segmentation and the source image. To evaluate the generalization ability of our method, we animated source images of various appearance domains. The first row of Fig. 6 shows the reenactment

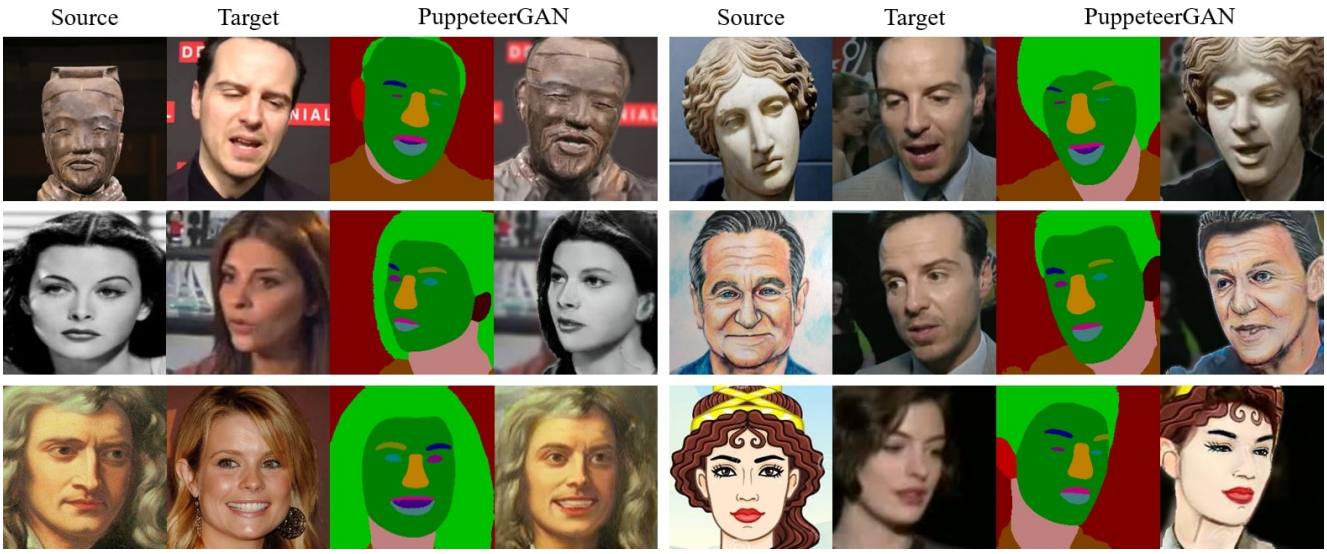

Figure 6: **Cross-domain portrait animation results.** Six animated portraits of different appearance domains. In each pair, we display the source portrait, the target frame, the generated segmentation mask and portrait in a row.

Figure 7: **Cross-resolution portrait animation results.** We animated the portraits with the resolution of 512. The input is the source portrait and the driven frame which is at the right top corner. The generated segmentation masks are shown in the same position of the animated portraits.

of a traditional Chinese sculpture Terracotta Warriors and the Greek sculpture. We also animated a black-and-white photo, a comic character, a painting and a cartoon character as shown in the last two lines. The results of our method are the same as the source portraits in identity, style and texture, except the poses are alike the targets.

4.4. Cross-resolution experiment

Benefit from the proposed two-stage framework, our coloring network can be trained on image datasets independently, which could improve the extensibility of our framework as the image data is more abundant and accessible than videos. In this section, we trained our coloring network on VoxCeleb [24] and CelebAMask-HQ [19], then tested well-trained model on images with the resolution of 512 (Fig. 7). The realistic generated portraits demonstrate that our method is fitting for high-resolution portrait puppeteering even without a video dataset of corresponding resolution for training. More details can be found in the Supplementary Materials.

5. Conclusion

In this work, we proposed PuppeteerGAN, a portrait animation framework that can animate arbitrary types of portrait using the target poses. Through separating portrait animation into pose retargeting and appearance transformation, PuppeteerGAN preserves the identity and appearance of the source portrait without any fine-tuning. Meanwhile, it can be extended to animating high-resolution (HQ) portraits by simply training the coloring network on the HQ image dataset. Experiments demonstrate that PuppeteerGAN outperforms previous methods in terms of usability, generalization capability, and extensibility. In future work, besides the appearance transformation, we will work towards synthesizing more realistic details, *e.g.* eye gaze and wrinkle.

References

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

[2] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. Towards open-set identity preserving face synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6713–6722, 2018.

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

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

[5] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face and Gesture Recognition*, 2018.

[6] Xinyuan Chen, Chang Xu, Xiaokang Yang, and Dacheng Tao. Attention-gan for object transfiguration in wild images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 164–180, 2018.

[7] DeepFakes. Faceswap. https://github.com/deepfakes/faceswap/. 2019-02-06.

[8] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4690–4699, 2019.

[9] Emily L Denton et al. Unsupervised learning of disentangled representations from video. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4414–4423, 2017.

[10] Jiali Duan, Xiaoyuan Guo, Yuhang Song, Chao Yang, and C-C Jay Kuo. Portraitgan for flexible portrait manipulation. *arXiv preprint arXiv:1807.01826*, 2018.

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

[12] Jiahao Geng, Tianjia Shao, Youyi Zheng, Yanlin Weng, and Kun Zhou. Warp-guided gans for single-photo facial animation. In *SIGGRAPH Asia 2018 Technical Papers*, page 231. ACM, 2018.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[14] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6626–6637, 2017.

[15] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning (ICML)*, pages 448–456, 2015.

[16] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1125–1134, 2017.

[17] Hyeongwoo Kim, Pablo Carrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. Deep video portraits. *ACM Transactions on Graphics (TOG)*, 37(4):163, 2018.

[18] Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis. Fast face-swap using convolutional neural networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3677–3685, 2017.

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

[20] Tommer Leyvand, Daniel Cohen-Or, Gideon Dror, and Dani Lischinski. Data-driven enhancement of facial attractiveness. In *ACM Transactions on Graphics (TOG)*, volume 27, page 38. ACM, 2008.

[21] Jingyuan Li, Fengxiang He, Lefei Zhang, Bo Du, and Dacheng Tao. Progressive reconstruction of visual structure for image inpainting. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5962–5971, 2019.

[22] Feng Liu, Jie Lu, Bo Han, Gang Niu, Guangquan Zhang, and Masashi Sugiyama. Butterfly: A panacea for all difficulties in wildly unsupervised domain adaptation. In *NeurIPS LTS Workshop*, 2019.

[23] Koki Nagano, Jaewoo Seo, Jun Xing, Lingyu Wei, Zimo Li, Shunsuke Saito, Aviral Agarwal, Jens Fursund, Hao Li, Richard Roberts, et al. pagan: real-time avatars using dynamic textures. *ACM Trans. Graph.*, 37(6):258–1, 2018.

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

[25] Ryota Natsume, Tatsuya Yatagawa, and Shigeo Morishima. Rsgan: face swapping and editing using face and hair representation in latent spaces. *arXiv preprint arXiv:1804.03447*, 2018.

[26] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 7184–7193, 2019.

[27] Tae-Hyun Oh, Tali Dekel, Changil Kim, Inbar Mosseri, William T Freeman, Michael Rubinstein, and Wojciech Matusik. Speech2face: Learning the face behind a voice. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7539–7548, 2019.

[28] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive nor-

malization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2337–2346, 2019.

[29] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 818–833, 2018.

[30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention (MICCAI)*, pages 234–241. Springer, 2015.

[31] Enrique Sanchez and Michel Valstar. Triple consistency loss for pairing distributions in gan-based face synthesis. *arXiv preprint arXiv:1811.03492*, 2018.

[32] Yujun Shen, Ping Luo, Junjie Yan, Xiaogang Wang, and Xiaoou Tang. Faceid-gan: Learning a symmetry three-player gan for identity-preserving face synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 821–830, 2018.

[33] Kritaphat Songsri-in and Stefanos Zafeiriou. Face video generation from a single image and landmarks. *arXiv preprint arXiv:1904.11521*, 2019.

[34] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (TOG)*, 36(4):95, 2017.

[35] Justus Thies, Michael Zollhöfer, Christian Theobalt, Marc Stamminger, and Matthias Nießner. Headon: Real-time reenactment of human portrait videos. *ACM Transactions on Graphics (TOG)*, 37(4):164, 2018.

[36] Chaoyue Wang, Chaohui Wang, Chang Xu, and Dacheng Tao. Tag disentangled generative adversarial network for object image re-rendering. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2901–2907, 2017.

[37] Chaoyue Wang, Chang Xu, Xin Yao, and Dacheng Tao. Evolutionary generative adversarial networks. *IEEE Transactions on Evolutionary Computation*, 23(6):921–934, 2019.

[38] Miao Wang, Guo-Ye Yang, Ruilong Li, Run-Ze Liang, Song-Hai Zhang, Peter M Hall, and Shi-Min Hu. Example-guided style-consistent image synthesis from semantic labeling. pages 1495–1504, 2019.

[39] Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Jan Kautz, and Bryan Catanzaro. Few-shot video-to-video synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[40] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1144–1156, 2018.

[41] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8798–8807, 2018.

[42] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[43] Olivia Wiles, A Sophia Koepke, and Andrew Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 670–686, 2018.

[44] Wayne Wu, Yunxuan Zhang, Cheng Li, Chen Qian, and Chen Change Loy. Reenactgan: Learning to reenact faces via boundary transfer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 603–619, 2018.

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

[46] Yunxuan Zhang, Siwei Zhang, Yue He, Cheng Li, Chen Change Loy, and Ziwei Liu. One-shot face reenactment. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2019.

[47] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. Talking face generation by adversarially disentangled audio-visual representation. In *AAAI*, volume 33, pages 9299–9306, 2019.

[48] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2223–2232, 2017.