Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-18449-9

Simultaneous control of head pose and expressions in 3D facial keypoint-based GAN

Tomoyuki Hatakeyama[1] · **Ryosuke Furuta**[1] · **Yoichi Sato**[1]

Received: 10 October 2023 / Revised: 10 January 2024 / Accepted: 25 January 2024

Abstract
In this work, we present a novel method for simultaneously controlling the head pose and the facial expressions of a given input image using a 3D keypoint-based GAN. Existing methods for controlling head pose and expressions simultaneously are not suitable for real images, or they generate unnatural results because it is not trivial to capture head pose (large changes) and expressions (small changes) simultaneously. In this work, we achieve simultaneous control of head pose and facial expressions by introducing 3D facial keypoints for GAN-based facial image synthesis, unlike the existing 2D landmark-based approach. As a result, our method can handle both large variations due to different head poses and subtle variations due to changing facial expressions faithfully. Furthermore, our model takes audio input as an additional modality for further enhancing the quality of generated images. Our model was evaluated on the VoxCeleb2 dataset to demonstrate its state-of-the-art performance for both facial reenactment and facial image manipulation tasks, and our model tends not to be affected by the driving images.

Keywords Video generation · Facial attribute manipulation

1 Introduction

Many methods for manipulating facial images have been reported. Models used in domain-based image-to-image translation methods [1–6] learn a mapping between domains (i.e., image domains that have different facial attributes) and used to control facial attributes. For example, a facial image in the domain "sadness" can be converted into an image in another domain, e.g., "happiness." However, these methods cannot synthesize images for continuous

✉ Tomoyuki Hatakeyama
tsumlikt@gmail.com

✉ Ryosuke Furuta
furuta@iis.u-tokyo.ac.jp

✉ Yoichi Sato
ysato@iis.u-tokyo.ac.jp

[1] Institute of Industrial Science, The University of Tokyo, Tokyo, Japan

Published online: 01 March 2024

labels because they deal with only discrete domains. On the other hand, many generative models that solve this problem by enabling detailed adjustments have been reported [7–10]. For example, head pose redirection models [7, 8] control face direction, and expression manipulation models [9, 10] control the values of action units (AUs) [11], which describe the intensity of facial muscle movement and are used to control facial expressions. For example, AU12 corresponds to lip corner pulling, and AU45 corresponds to eyes blinking.

Although there are several models for manipulating head poses and expressions, only a few can be used to simultaneously control head poses and expressions. For example, several methods [12, 13] can control them simultaneously, but they deal with generated images using StyleGAN [14, 15] and cannot deal with real, user-provided images. Although FACEGAN [10] was targeted for use with real images, it generates unnatural results because 2D landmarks are used to describe the position of each facial part, and it is not a physically plausible method to achieve 3D rotations. More concretely, it is not trivial to capture head pose (large changes) and expressions (small changes) at the same time with its 2D landmark-based approach. One might think that a head pose-control method [8] can be simply combined with an expression-control method [9], but we experimentally demonstrated that this approach is impractical.

We propose a practical method for simultaneously controlling head poses and expressions that uses a 3D keypoint-based GAN. Simultaneous control is achieved by treating head pose changes and expression changes, respectively as rotations and deformations of 3D keypoint. Our method is more physically plausible compared to [10] in controlling rotations and deformations thanks to the 3D keypoints. In addition, to generate more accurate control, the proposed method handles multimodal inputs (images and audio). The audio is used to predict how the mouth will be deformed, which enables the mouth region of the generated images to be more accurate. We also propose to use a face parser to introduce segmentation loss, which enables the model to generate the eye and mouth areas more realistically. Figure 1 shows an overview of the proposed method and an example of the manipulation results. The experimental results show that the manipulation quality was much higher than that of the baseline methods. Besides, we show that our method can reconstruct images with quality comparable to that of state-of-the-art reconstruction-only methods. In addition, generated images using the proposed method do not have leakage from the driving (reference) images because of the model architecture, though generated images using the state-of-the-art methods are easily affected by the driving images.

Our main contributions are summarized as follows:

– Physically plausible head pose control and expression control can be achieved simultaneously using unsupervisedly learned 3D keypoints.

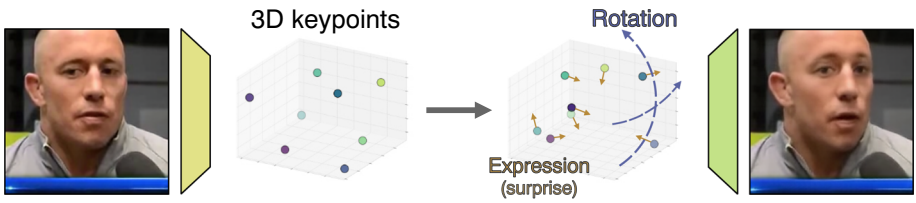


Fig. 1 Overview of the proposed method. In this example, head pose was rotated, and action units of "surprise" (AU1, AU2, AU5, AU26) were activated. Source image is decomposed into 3D keypoints, which are rotated and deformed to generate the target image. Head pose and expressions can be controlled simultaneously by treating head pose changes and expression changes respectively as rotations and 3D keypoint deformations

– The generated images are of comparable quality to those of state-of-the-art methods, and the facial images can be manipulated using parameters. In addition, the generated images are not affected by the driving images, unlike state-of-the-art methods.
– We use a face parser to introduce segmentation loss, which makes the eye and mouth regions of the generated images more realistic, and handle multimodal inputs (images and audio) to make the generated result more accurate.

2 Related works

Face reenactment In many methods [16–24], the source image is manipulated to have the pose and expressions of the driving image, preserving the source identity. Thies et al. proposed retrieving the mouth image and then transferring the source mouth region to an appropriate mouth region in the driving image [16]. However, their Face2Face method requires a sufficient number of frames containing the mouth region in the driving video. Averbuch et al. [17] proposed using 2D-warping from the source image to overcome this problem. However, both of these methods generate unnatural results because the hidden parts (e.g., inner mouth) are transferred from other images. Video-to-video synthesis methods [18, 19], which use sketch-to-face models with a spatiotemporal adversarial objective, require the preparation of other sketches in order to generate head poses and expressions. Zakharov et al. [20] proposed using a face-to-face model based on landmarks, but this method requires few-shot training before inference can be performed. In contrast, our approach requires only parameters rather than landmarks for inference. Furthermore, our approach does not require additional training. Tewari et al. [24] proposed using a first-order motion model (FOMM) to generate images on the basis of the motions of 2D keypoints learned in an unsupervised manner. This approach is aimed not only at face reenactment but also at animating objects such as the human body and cartoon animals. Therefore, facial attributes cannot be explicitly controlled.

Head pose and expression control The RaR [7] face rotation method uses a 3D morphable model (3DMM) and thus depends on the accuracy of the 3DMM feature extraction model. In contrast, in our approach, the feature extraction network is in an end-to-end trainable network. The approach most similar to ours is that used in FaceVid2Vid [8], in which 3D keypoints of the images are generated and used to control head pose, whereas FOMM uses 2D keypoints. Siarohin et al. [25] peoposed a novel view synthesis method by decomposing a RGB image into semantically meaningful parts such as depth and normal. While the aim of these works is to control only the head pose, our aim is also to control the expressions. GANimation [9] uses the cycle-consistency loss to generate manipulated images of continuous AU labels. While this method aims to control only the expressions, our method can also control the head pose. FACEGAN [10] uses the AU values to control the expressions. Although it can also control the head pose, 3D rotations cannot be applied theoretically because the model is based on 2D landmarks. In addition, the generated results are unsatisfactory because it is not easy to deal with changes in head pose (large rotations) and expressions (small deformations) simultaneously with a 2D landmark-based approach. These methods [12, 13] aim to control the facial attributes of the images generated by StyleGAN [14, 15], not the actual images (ie, real images) provided by the users. More concretely, their methods take latent codes as input and generate images. However, these methods do not aim at controlling the attributes of real images (StyleGAN output images), so their applications are limited ([12] showed the possibility of controlling the attributes of real images. However, to apply this to the problem

settings, inverse mapping from images to latent codes is non-trivial). Our method, on the contrary, aims to control the attributes of real facial images.

Audio-driven control Many recently reported methods [26–30] aim to manipulate facial images using audio. One [26] uses audio to deform landmarks and uses them to generate faces. Our approach does not deform landmarks but uses unsupervisedly learned keypoints, enabling the network to be trained without considering the accuracy of landmark detection. Another work [27] uses audio-based video editing as well as identity removal, and [30] can generate natural expression deformation with the audio. Unlike three of these methods [26, 28–30], our method can handle head pose and expressions simultaneously, where changes in expressions are predicted as deformations of 3D keypoints from both image and audio input.

3 Method

The model in our proposed method is trained by reconstructing the driving image x_d from the source image x_s using a two-step process: feature extraction and image generation (Fig. 5). In the first step, the identity feature I and the canonical keypoints are extracted from the source image. Then, the source and driving key points (K_s and K_d) are calculated by rotating and deforming the canonical keypoints. In the second step, the driving image is reconstructed using the extracted feature I and the calculated keypoints (K_s and K_d). Although these two steps are based on FaceVid2Vid [8], changes in head pose and expressions can be handled simultaneously, unlike with FaceVid2Vid [8]. The key idea is that head pose changes and expression changes are treated, respectively, as rotations and deformations of 3D keypoints. To achieve this, we introduce an **AU-driven deformation estimator** and an **audio-driven deformation estimator**, which predict deformations of 3D keypoints from AU and audio, respectively. In addition, to improve the quality of the generated image, we introduce a **segmentation loss**, which penalizes the difference between the face parsing results of the driving image and those of the generated image.

3.1 Feature extraction

As shown in Fig. 2, the 3D identity feature I is extracted, and the source and driving keypoints (K_s and K_d) are calculated. Two images are randomly selected from a speech video; one is treated as source image $x_s \in \mathbb{R}^{H\times W\times C}$ and the other is treated as driving image $x_d \in \mathbb{R}^{H\times W\times C}$. Figure 3 shows the network architectures used in the experiment. We explain the architectures in more detail.

The **identity extractor** extracts from x_s 3D identity feature $I \in \mathbb{R}^{H\times W\times D}$ for each channel (this feature represents hair color, eye shape, background, etc.). In addition, it is expected to compute the identity feature without dependence on camera parameters and perspective projection/distortion.

The **canonical detector** takes x_s as input and predicts canonical 3D keypoints $K_c \in \mathbb{R}^{k\times 3}$, where k is the number of the keypoints. These keypoints indicate the points in a neutral head pose (front face) and a neutral expression. Figure 4 shows the visualization of the canonical 3D keypoints. Note that these keypoints are learned without facial landmarks and optimized to reduce the loss.

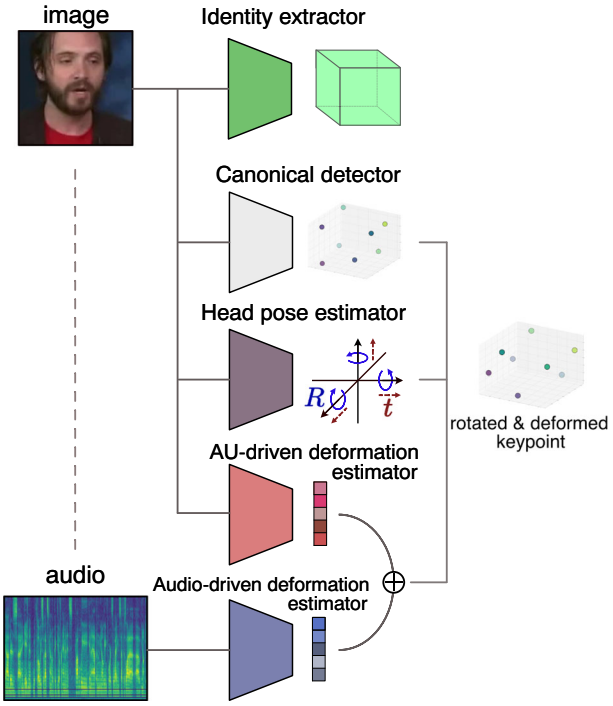


Fig. 2 Feature extraction procedure. Rotated and deformed keypoints are calculated from identity feature, canonical keypoints, head pose, AU-driven deformation, and audio-driven deformation. The first two (identity feature and canonical keypoints) are extracted from only the source image and not from the driving image

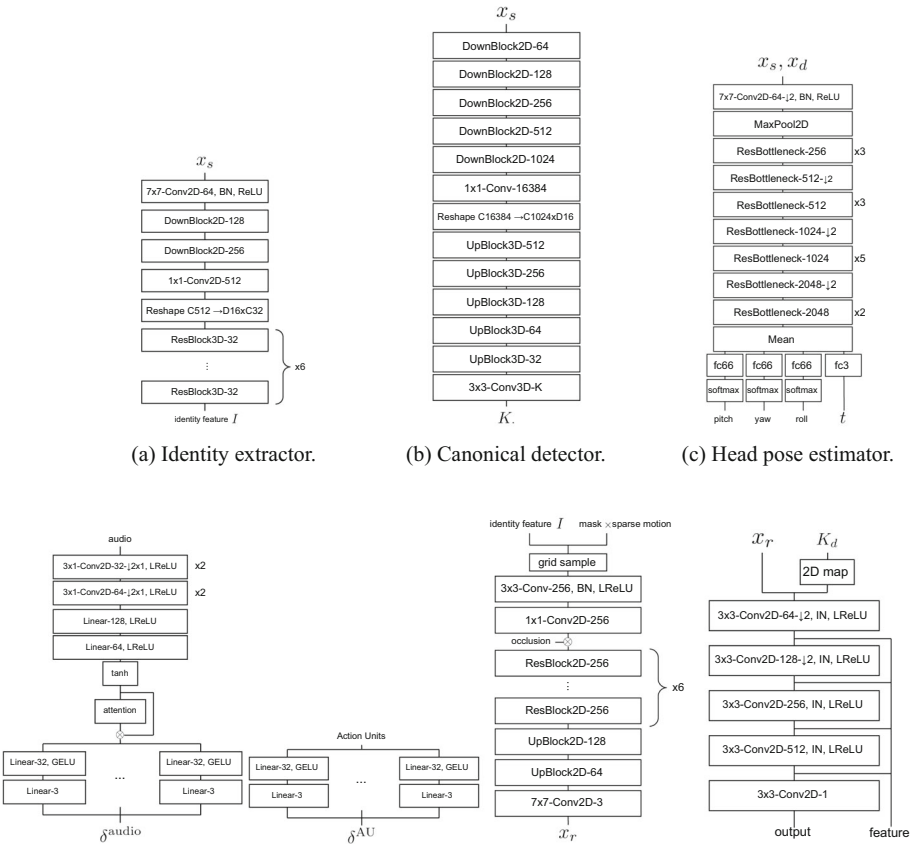
The **head pose estimator** outputs the angles $\in \mathbb{R}^3$ (pitch, yaw, and roll) to be used to estimate rotation matrices $R_* \in \mathbb{R}^{3\times3}$ from input image x_*, where $* \in \{s, d\}$. It also estimates translation vectors $t_* \in \mathbb{R}^3$, which represent the offset of the person's position.

The **AU-driven deformation estimator** takes AU values $\mathrm{AU}_* \in [0, 5]^{17}$ as input and predicts AU-driven deformation $\delta_*^{\mathrm{AU}} \in \mathbb{R}^{k\times3}$, which represents how the AUs affect the expressions.

In the network, $\mathrm{AU}_* \in [0, 5]^{17}$ is normalized to $[0, 1]^{17}$. Then, the deformation $\delta_{*,i}^{\mathrm{AU}} \in \mathbb{R}^3$ is extracted by two linear layers and activation of GELU for the keypoint i. The deformations are concatenated into the audio deformation

The **audio-driven deformation estimator** predicts audio-driven deformation $\delta_*^{\mathrm{audio}} \in \mathbb{R}^{k\times3}$, which enables the model to generate a more detailed mouth region. The audio of the corresponding frames is converted into a Mel spectrogram, and the speech content is analyzed using DeepSpeech [31] trained on Librispeech [32] to obtain audio feature $a_* \in \mathbb{R}^{n\times29}$, where n is the number of audio chunks. The audio feature is interpolated, as the number of chunks is the same as the number of video frames.

Each audio feature is combined with the adjacent features, and a per-frame feature $\in \mathbb{R}^{32}$ is computed using CNN and linear layers. The filtered audio expression $\in \mathbb{R}^{32}$ is calculated considering adjacent eight per-frame features through an self-attention architecture such as Audio2ExpressionNet [28]. The deformation $\delta_{*,i}^{\mathrm{audio}} \in \mathbb{R}^3$ is extracted by two linear layers and activation of GELU for the keypoint i. The deformations are concatenated into the audio deformation $\delta_*^{\mathrm{audio}} \in \mathbb{R}^{k\times3}$, where k is the number of keypoints.



(a) Identity extractor. (b) Canonical detector. (c) Head pose estimator.

(d) Audio-driven deformation estimator. (e) AU-driven deformation estimator. (f) Generator. (g) Discriminator.

Fig. 3 Network architectures. We determined hyperparameters based on [8], but some details differ to reduce computational resources. $\downarrow n$: stride of the convolution is n, **IN**: InstantNorm, **LReLU**: LeakyReLU

Finally, the rotated and deformed keypoints K_* are calculated as following:

$$K_* = R_*(K_c + \delta_*^{\text{AU}} + \delta_*^{\text{audio}}) + t_*, \tag{3.1}$$

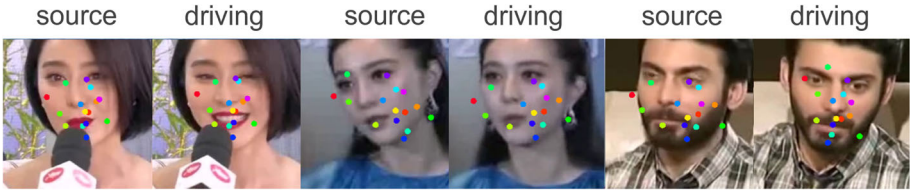


Fig. 4 Examples of the detected keypoints using the proposed method. In this model, for example, one keypoint (purple) is around the left eye, and another keypoint (yellow green) is located around the mouse

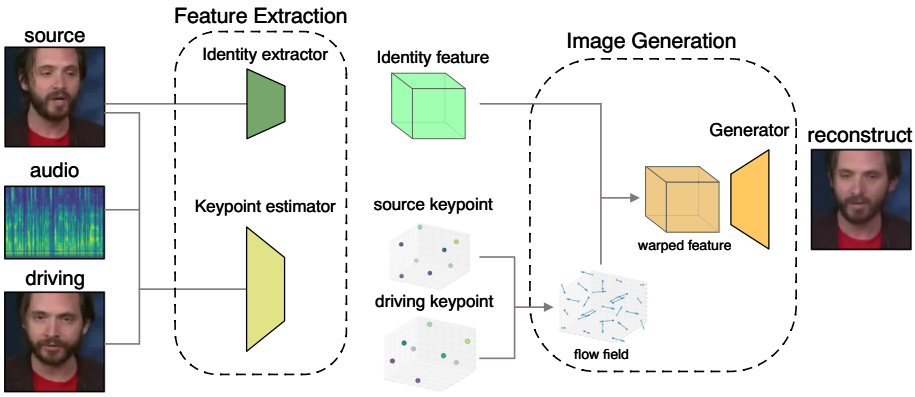


Fig. 5 Generator pipeline

3.2 Image generation

As shown in Fig. 5, driving image x_d is reconstructed using identity feature I and keypoints K_s and K_d, as described in more detail elsewhere [8, 24]. A heatmap representing the difference between the source and driving keypoints and the identity feature deformed by the Jacobian J_* are fed into the network and used to calculate the composite flow field w. Then, the warped identity feature $w(I)$ is fed into the image generator, and reconstruction results x_r are generated. Rotations are used as the Jacobian to reduce the computational resources required, as is done in FaceVid2Vid [8].

Driving and reconstructed images x_d and x_r, and driving keypoints K_d are used to train the discriminator, which distinguishes whether the given keypoints and images are real.

More details about the networks can be found in the supplemental material.

3.3 Losses

Our model is trained to minimize eight losses, which are described below. Total loss $\mathscr{L}$ is given by:

$$\begin{aligned}\mathscr{L} = \lambda_P \mathscr{L}_P + \lambda_G \mathscr{L}_G + \lambda_F \mathscr{L}_F + \lambda_E \mathscr{L}_E \\ + \lambda_K \mathscr{L}_K + \lambda_H \mathscr{L}_H + \lambda_\Delta \mathscr{L}_\Delta + \lambda_S \mathscr{L}_S,\end{aligned} \tag{3.2}$$

The four losses are aimed at making the generated image more realistic. **Perceptual loss** $\mathscr{L}_P$ is the L_1 distance of the features obtained by feeding the real and output images into a VGG19 model trained on the ImageNet dataset [33] and a VGG16 model trained on the VGGFace dataset [34]. **GAN loss** $\mathscr{L}_G$ makes the generated image more realistic by using a hinge loss term [35]. **Feature matching loss** $\mathscr{L}_F$ penalizes the difference in characteristics in the discriminator. **Equivariance loss** $\mathscr{L}_E$ is the L_1 distance between the driving keypoints and the transformed keypoints, which are inversed to match the driving keypoints.

Three losses are used to manage the dynamics of head poses and expressions. **Keypoint prior loss** $\mathscr{L}_K$ brings the key points closer. **Head pose loss** $\mathscr{L}_H$ penalizes the difference between the head pose predictions. The HopeNet model [36] trained on the 300W-LP dataset [37] is used to make the prediction. **Deformation prior loss** $\mathscr{L}_\Delta$ keeps deformation δ small, since the deformation is forced to be a slight change in expressions.

In addition to the above losses, which are used in FaceVid2Vid, we introduce **segmentation loss** $\mathscr{L}_S$, which penalizes the distance between the facial parts of the driving and generated images by using the face parser network. The BiSeNet model [38] trained on the CelebAMask-HQ dataset [39] is used as the face parser. Total loss consists of two losses: $\mathscr{L}_{\text{upper}}$ and $\mathscr{L}_{\text{lower}}$. $\mathscr{L}_{\text{upper}}$ is the L_1 distance between the eyes and eyebrow regions in the face parsing results obtained from the driving image and those from the generated image. Formally, let the face parser and the generated image be F and x_g, respectively. We denote the face parsing results as $F(x_g), F(x_d) \in [0, 1]^{h \times w \times c}$ where $c \in \{\text{eyes, eyebrows}\}$. $\mathscr{L}_{\text{upper}}$ is denoted as:

$$\mathscr{L}_{\text{upper}} = |F(x_g) - F(x_d)|_1, \tag{3.3}$$

On the other hand, $\mathscr{L}_{\text{lower}}$ is the L_1 distance of the lips and mouth regions and is denoted the same as in (3.3), but with $c \in \{\text{lips, mouth}\}$. Total loss is defined as the weighted sum of the losses:

$$\mathscr{L}_S = \mathscr{L}_{\text{upper}} + \lambda \mathscr{L}_{\text{lower}}, \tag{3.4}$$

where λ is a parameter indicating how much the facial reconstruction of the mouth part is emphasized.

3.4 Manipulation time

Reference-required control The head pose and expressions can be controlled to have the attributes as those of the driving image while maintaining the identity of the source image. Self-motion transfer is the setting in which the person in the source and driving images is the same. Cross-motion transfer is the setting in which the person in the source image and the one in the driving image differ.

Reference-free control If the driving image is the same as the source image, i.e., $x_d = x_s$, we can control the facial attributes without reference images. The head pose parameters can be perturbed using

$$\begin{cases} R_d \leftarrow R_u R_d \\ t_d \leftarrow t_u + t_d, \end{cases} \tag{3.5}$$

where $R_u \in \mathbb{R}^{3\times3}$ and $t_u \in \mathbb{R}^3$ are the user-defined rotation matrix and translation vector, respectively. Similarly, the expression parameters can be perturbed using

$$\begin{cases} \delta_d^{\text{AU}} \leftarrow \delta_u^{\text{AU}} \cdot \delta_d^{\text{AU}} \\ \delta_d^{\text{audio}} \leftarrow \delta_u^{\text{audio}} \cdot \delta_d^{\text{audio}} \end{cases}, \tag{3.6}$$

where $\delta_u^{\text{AU}} \in \mathbb{R}^k$ and $\delta_u^{\text{audio}} \in \mathbb{R}^k$ are the user-defined AU-driven deformation magnification vector and the audio-driven deformation magnification vector. Furthermore, AU-driven deformation can be controlled by directly changing AU values: $\text{AU}_d \leftarrow \text{AU}_d + \text{AU}_u$. For example, the lip corner is depressed when AU15 is large, and the eyes blink when AU45 is large.

4 Experiments

4.1 Training

In each of the three experiments, we trained the proposed method on three 32GB GPUs (NVIDIA V100 in DGX-1). The pipeline was implemented in PyTorch and optimized using

Table 1 Quantitative comparison of face reconstruction

	$L_1 \downarrow$	PSNR ↑	SSIM ↑	MS-SSIM ↑	CSIM ↑	FID ↓
Bilayer [44]	41.8	13.1	0.444	N/A	0.826	188.6
Bilayer [44] + segmentation	34.1	14.5	0.497	N/A	0.876	96.3
FOMM [24]	**13.8**	**22.2**	**0.749**	**0.815**	**0.975**	16.0
FaceVid2Vid [8]	21.4	18.6	0.703	0.730	0.956	18.2
FaceGANimation (a)	26.9	16.5	0.602	N/A	0.923	21.2
FaceGANimation (b)	26.0	16.8	0.589	N/A	0.927	44.3
Ours	17.8	19.9	0.717	0.740	0.959	**14.7**

(↑ means that larger is better; ↓ means that smaller is better)

the Adam [40] with custom settings ($\beta_1 = 0.5, \beta_2 = 0.999$) and a learning rate of 5.0×10^{-5}. Other settings and hyperparameters are shown in the supplemental material.

We trained the proposed and baseline methods on VoxCeleb2 [41], a large-scale video dataset containing more than 100K videos. In each video, one person is recorded talking for a few seconds, and their face is aligned to a size of 224×224. We resized them to 256×256 for comparison with other methods. Videos without audio data or for which AU detection using OpenFace [42, 43] failed were excluded. The data we used is openly available on the VoxCeleb2 website.[1]

4.2 Facial reenactment

For facial reenactment, the driving image was reconstructed from the source image and the driving attributes.

Baselines Three methods were used as baselines: Bilayer [44], FOMM [24], and FaceVid2Vid [8]. The Bilayer [44] model generates only the area of the head and body, so we complemented the background region of each image by reverting the background to the generated image. We also created a "FaceGANimation" baseline, which combines the FaceVid2Vid [8] and GANimation [9] models. FaceVid2Vid controls only head poses, and GANimation manipulates only expressions. This baseline had two variants: (a) FaceVid2Vid and GANimation were pre-trained independently on the dataset and then combined; (b) FaceVid2Vid and GANimation were trained after they were combined in the same model.

Evaluation metrics We evaluated the reconstruction results using six metrics. L_1 represents the average distance of the pixel values between the driving and generated images. The driving and generated images were compared using the peak signal-to-noise ratio (PSNR), structural similarity (SSIM) [45], multi-scale SSIM (MS-SSIM) [46], and cosine similarity (CSIM). The distributions of the real (driving) and generated images were compared using the Fréchet inception distance (FID) [47].

Self motion transfer As shown in Table 1, the proposed method obtained results comparable to those of the baseline facial reconstruction methods. The FOMM method obtained better results than the other methods, including ours. However, the difference between the FOMM results and our results could possibly be reduced by training our model with a larger batch size and using more computational resources (such as eight GPUs as were used in [8]). In

[1] https://www.robots.ox.ac.uk/~vgg/data/voxceleb/vox2.html

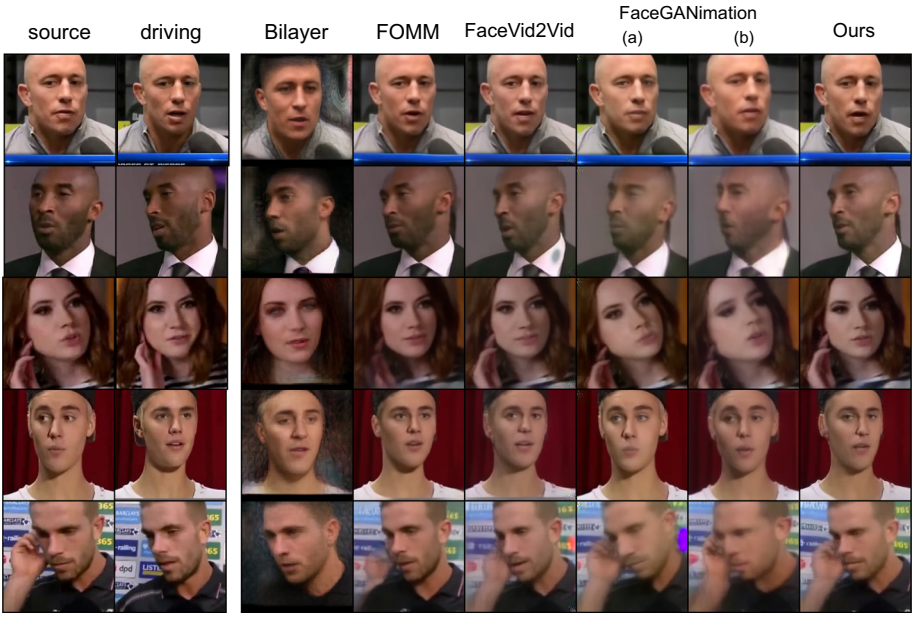


Fig. 6 Qualitative comparison of reconstruction results: Bilayer [44], FOMM [24], FaceVid2Vid [8], FaceGANimation, and proposed. FaceGANimation consists of FaceVid2Vid [8] and GANimation [9]: (a) independently trained; (b) simultaneously trained

addition, note that our objective is not to present a model that can reconstruct images with better quality. Instead, it is to present a model that can manipulate head pose and expressions simultaneously by using a 3D keypoint-based approach. The aim of the first experiment was simply to determine whether our model could obtain comparable quality if we reconstructed the driving images. As shown in Fig. 6, our method captured identity better than the other methods. For example, when there was a hand in the source image, the reconstruction images generated with our method showed the hand more clearly. Our method successfully captured the identity because the expressions are driven only by AUs and audio and are independent of identity.

Ablation study Table 2 shows the results of facial reconstruction after ablating the audio driven deformation and components of the proposed method, and Fig. 7 shows qualitative examples. We used our model with only the AU-driven deformation as a baseline. The audio-driven deformation made the images more accurate when the mouth in the driving image

Table 2 Quantitative comparison of face reconstruction after audio-driven deformation and segmentation loss were ablated

	$L_1 \downarrow$	PSNR ↑	SSIM ↑	MS-SSIM ↑	CSIM ↑	FID ↓
δ^{AU}	18.1	19.8	0.717	0.740	0.959	17.9
$\delta^{\mathrm{AU}} + \delta^{\mathrm{audio}}$	**17.6**	**20.0**	**0.720**	**0.744**	**0.960**	25.0
$\delta^{\mathrm{AU}} + \mathscr{L}_S$	18.1	19.8	0.711	0.730	0.958	**14.0**
$\delta^{\mathrm{AU}} + \delta^{\mathrm{audio}} + \mathscr{L}_S$	17.8	19.9	0.717	0.740	0.959	14.7

δ^{AU}: AU-driven deformation; δ^{audio}: audio-driven deformation; $\mathscr{L}_S$: segmentation loss

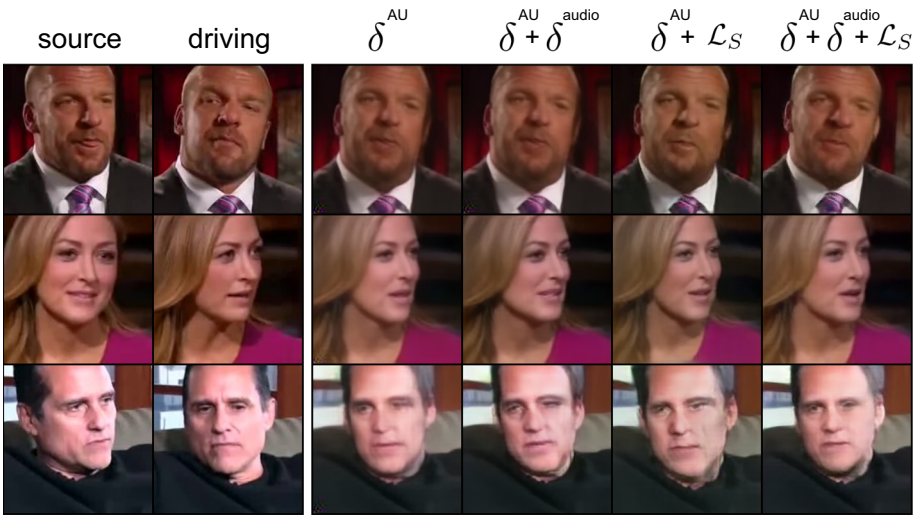


Fig. 7 Qualitative comparison of AU-driven deformation δ^{AU}, audio-driven deformation δ^{audio} and segmentation loss $\mathscr{L}_S$

was opened widely, as shown in the second row of the figure. However, the scores Table 2 did not improve significantly because only small regions, such as the mouth, were affected. Face parser (FP) helped the method generate images with a lower FID because it made the model focus on the eye and mouth regions, so the background noise tended to disappear.

As shown in Table 3, $\mathscr{L}_S$ makes better results with SSIM but worse with FID for FOMM [24]. This is because the accuracy for eyes and mouth was higher, though the generated images are relatively blurred.

Figure 8 shows the generated results with addition or removal of the deformations. AU-driven deformation δ^{AU} aims to make coarse-grained expressions, and audio-driven deformation δ^{Audio} aims to make expressions more detailed, especially for the mouth region.

Cross motion transfer Our method can also be used when the person in the source image differs from the one in the driving image. As shown in Fig. 9, FOMM had good reenactment results, but there tended to be blurred areas in the hair, body, and background. FaveVid2Vid tended to have a blurred facial outline (first and second rows). In contrast, our method struck the right balance between reflecting driving attributes and maintaining source identity. Furthermore, the images generated by FOMM and FaceVid2Vid tended to be affected by the driving identity. For example, the generated images have background artifacts (Fig. 9) and caption remnants (Fig. 9 second row). The artifacts appeared in the FOMM and FaceVid2Vid generated images because the models capture deformations from images. In contrast, our method extracts from only AUs and audio.

Table 3 Self motion transfer results of FOMM [24]; $\mathscr{L}_S$: segmentation loss

	L1 ↓	PSNR ↑	SSIM ↑	MS-SSIM ↑	CSIM ↑	FID ↓
FOMM	13.8	22.2	0.749	0.815	0.975	16.0
FOMM + $\mathscr{L}_S$	13.6	22.3	0.778	0.815	0.975	21.3

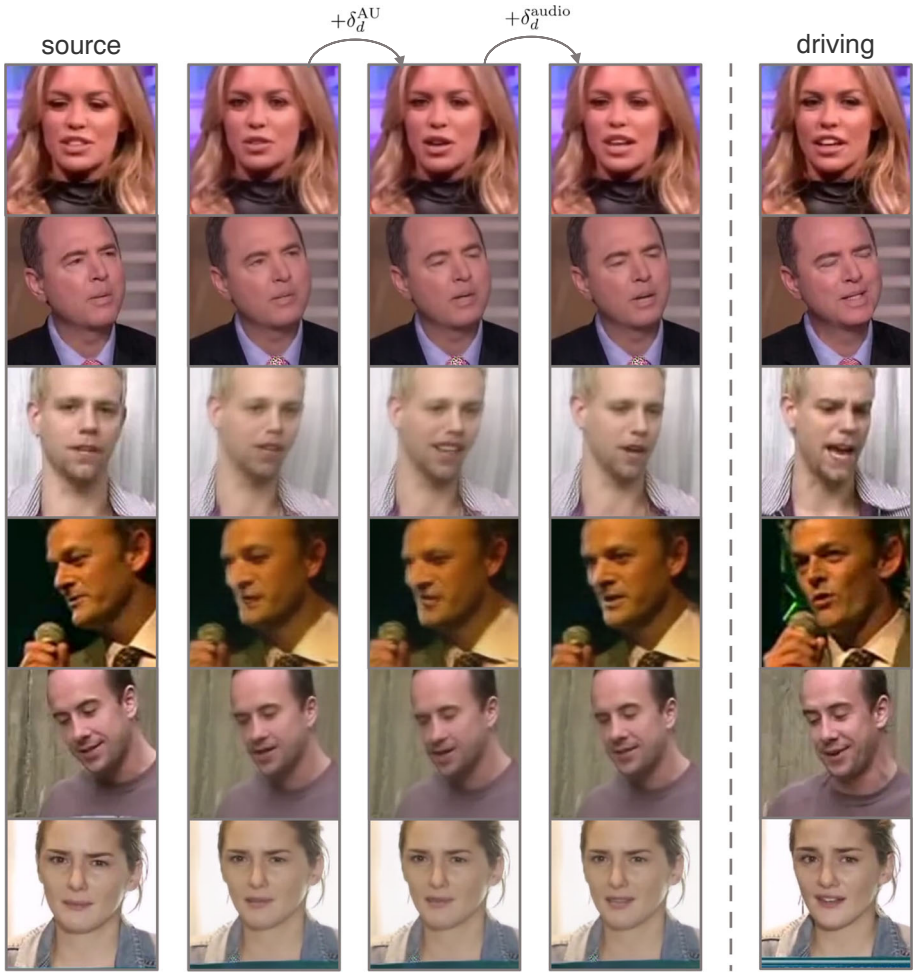


Fig. 8 The generated results with addition or removal of the deformations. By removing AU-driven and Audio-driven deformations for the source image, the expression of the source image gets to neutral. We can add the deformations corresponding to the AUs and the audio of the driving frame

4.3 Facial image manipulation

Our method can control head pose and expressions simultaneously, as illustrated in Fig. 10, which shows the results of the combined control. Both frontalization (head pose changes) and emotion control (expression changes) were achieved. In this experiment, each emotion was driven by activating the corresponding AUs. For example, the image was made "happy" by activating AU06 and AU12.

Figure 11 shows that our method can transform images for AUs and rotation (pitch/yaw/roll) with continuous labels. For example, the higher the intensity of AU12, the more smiling the generated face is. We can thus intuitively control facial images using parameters without reference images.

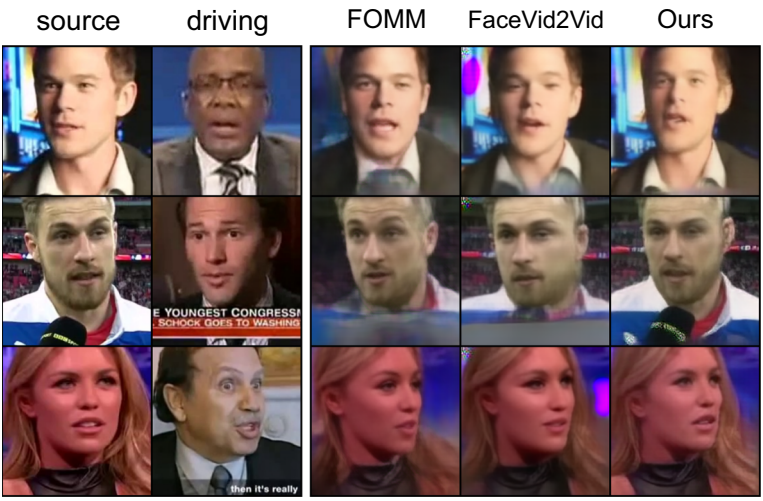


Fig. 9 Cross-motion transfer results for FOMM [24], FaceVid2Vid [8], and proposed methods

Table 4 shows the L_1 distance between the target attributes (head pose and AUs of the driving image) and those of the generated image. This experiment was conducted on cross motion transfer settings, and each attribute was predicted using OpenFace. As shown in the table, our method outperformed the baselines for manipulating all attributes. Additional results are shown in the supplemental material.

5 Limitations and discussions

Our method does have limitations. As exemplified by the failure cases illustrated in Fig. 12, we can control the AUs in Fig. 11, but the generated image can be affected by other AUs

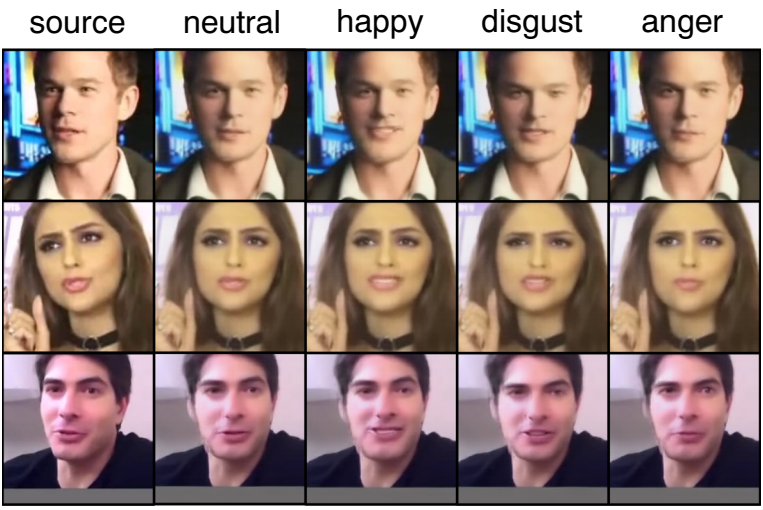


Fig. 10 Our method enables simultaneous frontalization and facial expression control

Fig. 11 Proposed method can control facial images continuously for rotations and action units. For AUs, the lower the image, the greater the intensity of activation. AU01: inner brow raiser; AU09: nose wrinkler; AU10: upper lip raiser; AU12: lip corner puller; AU14: dimpler; AU15: lip corner depressor; AU17: chin raiser; AU25: lips part; AU26: jaw drop; AU45: blinking

if the target AU tends to be activated along with the other AUs at the same time. That is, our model cannot control AUs independently. For example, AU09 (nose wrinkler) tends to be activated with mouth opening units such as AU10 (upper lip raiser), which affects the generated results (Fig. 12a left). We cannot control AUs that do not appear very often or intensely. For example, our model cannot control a face for AU05 (upper lid raiser), which does not appear clearly in many videos (Fig. 12a right). Furthermore, our model is based on OpenFace [42, 43] prediction, so the precision of manipulation depends on the prediction model.

Another limitation is related to head pose manipulation. Figure 12b shows examples of when our model failed to generate objects. Changes in head pose often distort objects, such as the caption and microphone. The generated images are thus blurred and unnatural when the head poses of the source and driving images significantly differ (Fig. 12c). This failure happens because the head pose does not change much in videos with an average duration of 8 seconds from which we randomly pick two frames to determine the source and driving images.

6 Conclusion

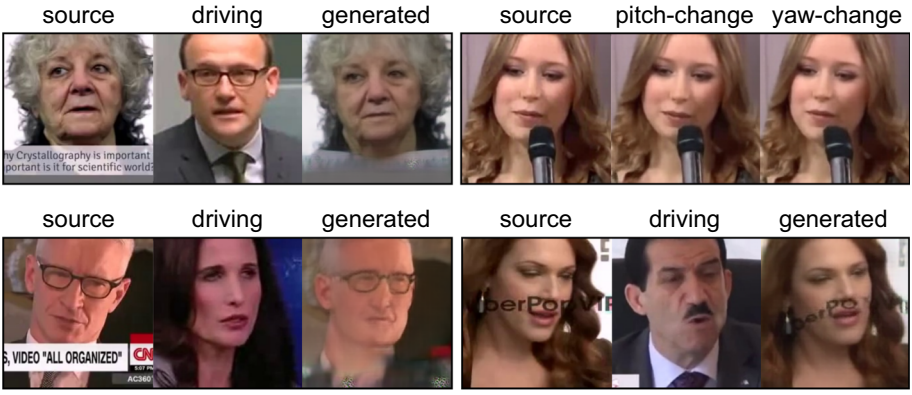
We have presented a physically plausible method based on unsupervisedly learned 3D keypoints for simultaneously controlling the head pose and expressions of a facial image. Existing methods that simultaneously control these attributes are not suitable for real images [12, 13]

Table 4 L_1 distance between the target attributes (head pose and AUs of the driving image) and the attributes of the generated image AU_{upper} : action units around eye, nose, and cheek

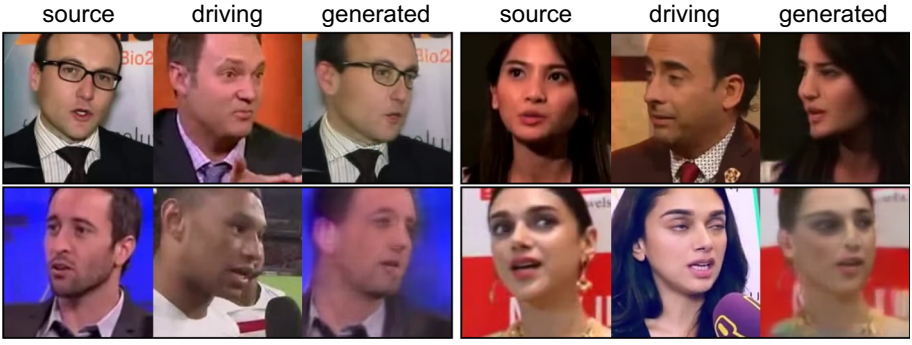
	Pitch	Yaw	Roll	AU_{upper}	AU_{lower}
FaceGANimation (a)	0.105	0.171	0.103	0.362	0.388
FaceGANimation (b)	0.120	0.175	0.096	0.593	0.426
Ours	**0.043**	**0.039**	**0.027**	**0.307**	**0.296**

AU_{lower}: action units around mouth and jaw. Pitch, yaw, and roll represent head rotations

(a) AU controls. Left: AU09 activation. Right: AU05 activation.



(b) Failure cases of generating object such as caption and microphone.



(c) Rotation failure cases.

Fig. 12 Example failure cases

or generate unnatural results due to using a 2D keypoint-based approach [10]. Suppose we combine the methods of the 3D keypoint-based rotation model [8] and the action unit-based animation model [9]. In that case, the generated images tend to collapse because it is difficult to capture the head pose (large changes) and the expressions (small changes) at the same time.

To alleviate this problem, the model used in our method handles head pose changes as rotations and expression changes as keypoint deformations. As a result, the model's outputs were comparable to those of state-of-the-art methods and of much higher quality than combined methods, and they were not affected by the driving images. Our experiments demonstrated that our model can simultaneously control the head pose and expression. Future work includes adding gaze control to our method.

Funding Open Access funding provided by The University of Tokyo.

Data Availibility Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Declarations

Conflict of Interest The authors declare that they have no conflict of interest.

References

1. Zhu JY, Park T, Isola P, Efros AA (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proc. of the IEEE/CVF international conference on computer vision (ICCV 2017), pp 2223–2232
2. Choi Y, Choi M, Kim M, Ha JW, Kim S, Choo J (2018) StarGAN: unified generative adversarial networks for multi-domain image-to-image translation. In: Proc. of the IEEE/CVF conference on computer vision and pattern recognition (CVPR 2018), pp 8789–8797
3. Li M, Zuo W, Zhang D (2016) Deep identity-aware transfer of facial attributes. In: arXiv preprint arXiv:1610.05586
4. Perarnau G, van de Weijer J, Raducanu B, Álvarez JM (2016) Invertible conditional GANs for image editing. In: Advances in neural information processing systems (NeurIPS 2016) workshop on adversarial training
5. Shen W, Liu R (2017) Learning residual images for face attribute manipulation. In: Proc of the IEEE/CVF conference on computer vision and pattern recognition (CVPR 2017), pp 4030–4038
6. Larsen ABL, Sønderby SK, Larochelle H, Winther O (2016) Autoencoding beyond pixels using a learned similarity metric. In: Proc of the machine learning research (PMLR 2016), pp 1558–1566
7. Zhou H, Liu J, Liu Z, Liu Y, Wang X (2020) Rotate-and-Render: unsupervised photorealistic face rotation from single-view images. In: Proc of the IEEE/CVF conference on computer vision and pattern recognition (CVPR 2020), pp 5910–5919
8. Wang TC, Mallya A, Liu MY (2021) One-shot free-view neural talking-head synthesis for video conferencing. In: Proc of the IEEE/CVF conference on computer vision and pattern recognition (CVPR 2021), pp 10039–10049
9. Pumarola A, Agudo A, Martinez AM, Sanfeliu A, Moreno-Noguer F (2018) GANimation: anatomically-aware facial animation from a single image. In: Proc of the European conference on computer vision (ECCV 2018), pp 835–851
10. Tripathy S, Kannala J, Rahtu E (2021) FACEGAN: facial Attribute controllable reenactment GAN. In: Proc of the IEEE/CVF winter conference on applications of computer vision (WACV 2021), pp 1329–1338
11. Ekman P, Friesen WV (1978) Facial action coding system: a technique for the measurement of facial movement. In: Palo Altom: consulting psychologists press
12. Deng Y, Yang J, Chen D, Wen F, Tong X (2020) Disentangled and controllable face image generation via 3D imitative-contrastive learning. In: Proc of the IEEE/CVF conference on computer vision and pattern recognition (CVPR 2020)
13. Tewari A, Elgharib M, Bharaj G, Bernard F, Seidel HP, Pérez P, Zöllhofer M, Theobalt C (2020) StyleRig: rigging StyleGAN for 3D control over portrait images. In: Proc of the IEEE/CVF conference on computer vision and pattern recognition (CVPR 2020), pp 6142–6151

14. Karras T, Laine S, Aila T (2019) A style-based generator architecture for generative adversarial networks. In: Proc of the IEEE/CVF conference on computer vision and pattern recognition (CVPR 2019), pp 4401–4410
15. Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J, Aila T (2020) Analyzing and improving the image quality of StyleGAN. In: Proc of the IEEE/CVF conference on computer vision and pattern recognition (CVPR 2020), pp 8110–8119
16. Thies J, Zollhöfer M, tamminger M, Theobalt C, Nießner M (2016) Face2Face: real-time face capture and reenactment of RGB videos. In: Proc of the IEEE/CVF conference on computer vision and pattern recognition (CVPR 2016) pp 2387–2395
17. Averbuch-Elor H, Cohen-Or D, Kopf J, Cohen MF (2017) Bringing portraits to life. ACM Trans Graph (TOG) 36(6):196:1-196:13
18. Wang TC, Liu MY, Zhu JY, Liu G, Tao A, Kautz J, Catanzaro B (2018) Video-to-video synthesis. In: Proc of the advances in neural information processing systems (NeurIPS 2018)
19. Wang TC, Liu MY, Tao A, Liu G, Kautz J, Catanzaro B (2019) Few-shot video-to-video synthesis. In: Proc of the advances in neural information processing systems (NeurIPS 2019)
20. Zakharov E, Shysheya A, Burkov E, Lempitsky V (2019) Few-shot adversarial learning of realistic neural talking head models. In: Proc of the IEEE/CVF International conference on computer vision (ICCV 2019), pp 9459–9468
21. Wang TC, Liu MY, Zhu JY, Tao A, Kautz J, Catanzaro B (2018) High-resolution image synthesis and semantic manipulation with conditional GANs. In: Proc of the IEEE/CVF conference on computer vision and pattern recognition (CVPR 2018), pp 8798–8807
22. Wiles O, Koepke AS, Zisserman A (2018) X2Face: A network for controlling face generation by using images, audio, and pose codes. In: Proc of the European conference on computer vision (ECCV 2018), pp 690–706
23. Nirkin Y, Keller Y, Hassner T (2019) FSGAN: subject agnostic face swapping and reenactment. In: Proc of the IEEE/CVF international conference on computer vision (ICCV 2019), pp 7184–7193
24. Siarohin A, Lathuilière S, Tulyakov S, Ricci E, Sebe N (2019) First order motion model for image animation. In: Proc of the advances in neural information processing systems (NeurIPS 2019)
25. Siarohin A, Menapace W, Skorokhodov I, Olszewski K, Ren J, Lee HY, Chai M, Tulyakov S (2023) Unsupervised volumetric animation. In: Proc of the IEEE/CVF conference on computer vision and pattern recognition (CVPR 2023)
26. Chen L, Maddox RK, Duan Z, Xu C (2019) Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In: Proc of the IEEE/CVF conference on computer vision and pattern recognition (CVPR 2019), pp 7824–7833
27. Song L, Wu W, Qian C, He R, Loy CC (2020) Everybody's Talkin': let me talk as you want. In: arXiv preprint arXiv:2001.05201
28. Thies J, Elgharib M, Tewari A, Theobalt C, Nießner M (2020) Neural voice puppetry: audio-driven facial reenactment. In: Proc of the European conference on computer vision (ECCV 2020)
29. Guo Y, Chen K, Liang S, Liu Y, Bao H, Zhang J (2021) AD-NeRF: audio driven neural radiance fields for talking head synthesis. In: Proc of the IEEE/CVF international conference on computer vision (ICCV 2021)
30. Stypulkowski M, Vougioukas K, He S, Zieba M, Petridis S, Pantic M (2024) Diffused heads: diffusion models beat GANs on talking-face generation. In: Proc of the IEEE/CVF winter conference on applications of computer vision (WACV 2024), pp 5091–5100
31. Amodei D, Anubhai R, Battenberg E, Case C, Casper J, Catanzaro B, JC et al (2016) Deep speech 2 : end-to-end speech recognition in English and Mandarin. In: Proc of the international conference on machine learning (ICML 2016), pp 173–182
32. Panayotov V, Chen G, Povey D, Khudanpur S (2015) LibriSpeech: an ASR corpus based on public domain audio books. In: Proc of the IEEE international conference on acoustics, speech and signal processing (ICASSP 2015), pp 5206–5210
33. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) ImageNet: a large-scale hierarchical image database. In: Proc of the IEEE/CVF conference on computer vision and pattern recognition (CVPR 2009), pp 248–255
34. Parkhi OM, Vedaldi A, Zisserman A (2015) Deep face recognition. In: Proc of the British machine vision conference (BMVC 2015), pp 41.1–41.12
35. Lim JH, Ye JC (2017) Geometric GAN. In: arXiv preprint arXiv:1705.02894
36. Ruiz N, Chong E, Rehg JM (2018) Fine-grained head pose estimation without keypoints. In: the IEEE conference on computer vision and pattern recognition (CVPR 2018) workshop, pp 2187–2196
37. Zhu X, Lei Z, Liu X, Shi H, Li SZ (2016) Face alignment across large poses: a 3D solution. In: Proc of the IEEE/CVF conference on computer vision and pattern recognition (CVPR 2016), pp 146–155

38. Yu C, Wang J, Peng C, Gao C, Yu G, Sang N (2018) BiSeNet: bilateral segmentation network for real-time semantic segmentation. In: Proc of the European conference on computer vision (ECCV 2018), pp 334–349
39. Lee CH, Liu Z, Wu L, Luo P (2020) MaskGAN: towards diverse and interactive facial image manipulation. In: Proc of the IEEE/CVF conference on computer vision and pattern recognition (CVPR 2020), pp 5549–5558
40. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. In: arXiv preprint arXiv:1412.6980
41. Chung JS, Nagrani A, Zisserman A (2018) VoxCeleb2: deep speaker recognition. In: Proc of the inter-speech, pp 1086–1090
42. Baltrusaitis T, Zadeh A, Lim YC, Morency LP (2018) OpenFace 2.0: facial behavior analysis toolkit. In: Proc of the IEEE international conference on automatic face gesture recognition (FG 2018), pp 59–66
43. Baltrusaitis T, Mahmoud M, Robinson P (2015) Cross-dataset learning and person-specific normalisation for automatic action unit detection. In: Proc of the IEEE international conference on automatic face gesture recognition (FG 2015), pp 1–6
44. Zakharov E, Ivakhnenko A, Shysheya A, Lempitsky V (2020) Fast bi-layer neural synthesis of one-shot realistic head avatars. In: Proc of the European conference on computer vision (ECCV 2020), pp 524–540
45. Wang Z, Bovik AC, Sheikh HR, Member S, Simoncelli EP, Member S (2004) Image quality assessment: from error visibility to structural similarity. In: Proc of the IEEE transactions on image processing (TIP 2004) pp 600–612
46. Wang Z, Simoncelli EP, Bovik AC (2003) Multi-scale structural similarity for image quality assessment. In: Proceedings of the IEEE asilomar conference on signals, systems, and computers (Asilomar), vol 2, pp 1398–1402
47. Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S (2017) GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In: Proc of the advances in neural information processing systems (NeurIPS 2017), pp 6629–6640