Pose-Controllable Talking Face Generation by Implicitly Modularized Audio-Visual Representation

Hang Zhou[1], Yasheng Sun[2,3], Wayne Wu[2,4], Chen Change Loy[4], Xiaogang Wang[1], Ziwei Liu[4] ✉
[1]CUHK - SenseTime Joint Lab, The Chinese University of Hong Kong [2]SenseTime Research
[3]Tokyo Institute of Technology [4]S-Lab, Nanyang Technological University
{zhouhang@link,xgwang@ee}.cuhk.edu.hk, wuwenyan@sensetime.com, {ccloy,ziwei.liu}@ntu.edu.sg

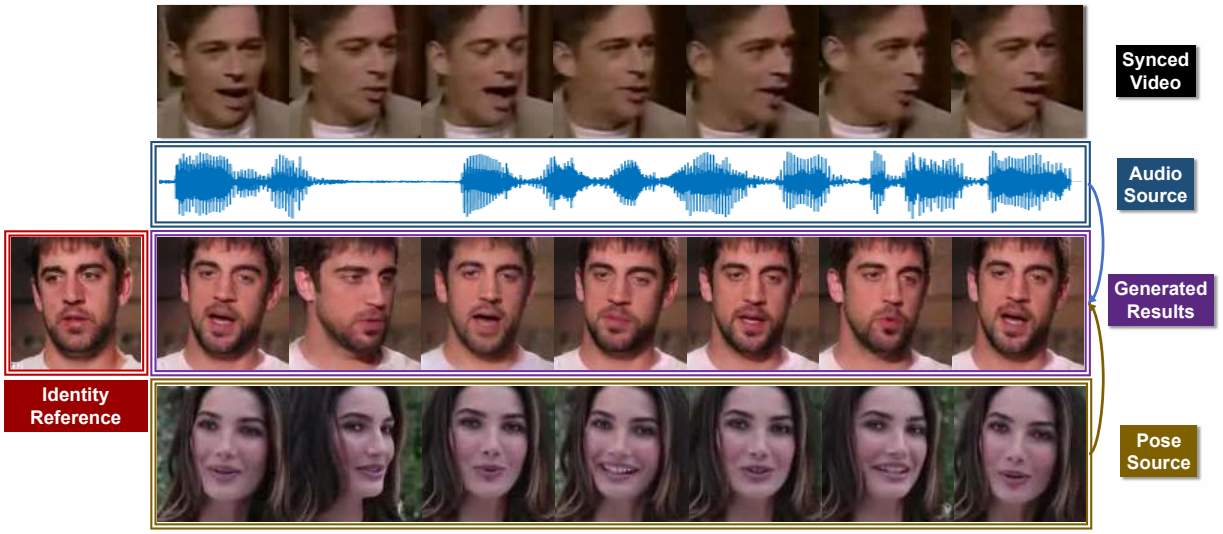


Figure 1: **Illustration of Pose-Controllable Audio-Visual System (PC-AVS).** Our approach takes one frame as identity reference and generates audio-driven talking faces with pose controlled by another *pose source* video. The mouth shapes of the generated frames are matched with the first row (synced video with audio) while the pose is matched with the bottom row (pose source).

Abstract

While accurate lip synchronization has been achieved for arbitrary-subject audio-driven talking face generation, the problem of how to efficiently drive the head pose remains. Previous methods rely on pre-estimated structural information such as landmarks and 3D parameters, aiming to generate personalized rhythmic movements. However, the inaccuracy of such estimated information under extreme conditions would lead to degradation problems. In this paper, we propose a clean yet effective framework to generate pose-controllable talking faces. We operate on non-aligned raw face images, using only a single photo as an identity reference. The key is to modularize audio-visual representations by devising an implicit low-dimension pose code. Substantially, both speech content and head pose information lie in a joint non-identity embedding space. While speech content information can be defined by learning the intrinsic synchronization between audio-visual modalities, we identify that a pose code will be complementarily learned in a modulated convolution-based reconstruction framework.

Extensive experiments show that our method generates accurately lip-synced talking faces whose poses are controllable by other videos. Moreover, our model has multiple advanced capabilities including extreme view robustness and talking face frontalization.[1]

1. Introduction

Driving a static portrait with audio is of great importance to a variety of applications in the field of entertainment, such as digital human animation, visual dubbing in movies, and fast creation of short videos. Armed with deep learning, previous researchers take two different paths towards analyzing audio-driven talking human

[1]Code, models, and demo videos are available at https://hangz-nju-cuhk.github.io/projects/PC-AVS.

faces: 1) through pure latent feature learning and image reconstruction [14, 76, 9, 71, 50, 54, 44], and 2) to borrow the help of structural intermediate representations such as 2D landmarks [51, 10, 18] or 3D representations [1, 52, 49, 8, 74, 46, 28]. Though great progress has been made in generating accurate mouth movements, most previous methods fail to model head pose, one of the key factors for talking faces to look natural.

It is very challenging to control head poses while generating lip-synced videos with audios. 1) On the one hand, pose information can rarely be inferred from audios. While most previous works choose to keep the original pose in a video, very recently, a few works have addressed the problem of generating personalized rhythmic head movements from audios [8, 74, 65]. However, they rely on a short clip of video to learn individual rhythms [65, 8], which might be inaccessible for a variety of scenes. 2) On the other hand, all the above methods rely on 3D structural intermediate representations [65, 8, 74]. The pose information is inherently coupled with facial movements, which affects both reconstruction-based [14, 71] and 2D landmark-based methods [10, 18]. Thus the most plausible way is to leverage 3D models [33, 52, 8, 70] where the pose parameters and expression parameters are explicitly defined [2]. Nevertheless, long-term videos are normally needed in order to learn person-specific renderers for 3D models. More importantly, such representations would be inaccurate under extreme cases such as large pose or low-light conditions.

In this work, we propose **Pose-Controllable Audio-Visual System (PC-AVS)**, which achieves free pose control when driving arbitrary talking faces with audios. Instead of learning pose motions from audios, we leverage another *pose source* video to compensate only for head motions as illustrated in Fig. 1. Specifically, *no structural information* is needed in our work. The key is to *devise an implicit low-dimension pose code that is free of mouth shape or identity information*. In this way, audio-visual representations are *modularized* into spaces of three key factors: speech content, head pose, and identity information.

In particular, we identify the existence of a non-identity latent embedding from the visual domain through data augmentation. Intuitively, the complementary *speech content* and *pose* information should originate from it. Extracting the shared information between visual and audio representations could lead to the *speech content space* by synchronizing both the modalities, which is also proven to be beneficial for various downstream audio-visual tasks [17, 40, 72, 36, 25]. However, there is no explicit way to model pose without precisely recognized structural information. Here we leverage the *prior knowledge* of 3D pose parameters, that a mere vector of 12 dimensions, including a 3D rotation matrix, a 2D positional shifting bias, and a scale factor, is sufficient to represent a head pose. Thus we define a mapping from the non-identity space to a low dimension code which implicitly stands for the pose. Notably, we do not use other 3D priors to model the transformation between different poses. Then with additional identity supervision, the *modularization* of the whole talking face representations has been completed.

The last key is the cross-frame reconstruction between video frames, where all representations are complementarily learned. A generator whose convolutional kernels are modulated by the embedded features is also designed. Specifically, we assemble the features from the *modularized* spaces and use them to scale the weights of the kernels as proposed in [31]. The expressive ability of the weight modulation also enforces our *modularization* to audio-visual representations in an implicit manner, *i.e.*, in order to ensure low reconstruction loss, the low-dimensional code automatically controls pose while the speech content embedding takes care of the mouth. During inference, we can drive an arbitrary face by a clip of audio with head movements controlled by another video. Multiple advanced properties can also be achieved such as extreme view robustness and frontalizing talking faces.

Our contributions are summarized as follows: **1)** We propose to modularize the representations of talking human faces into the spaces of speech content, head pose, and identity respectively, by devising a low-dimensional pose code inspired by 3D pose prior in talking faces. **2)** The modularization is implicitly and complementarily learned in a construction-based framework with modulated convolution. **3)** Our model generates pose-controllable talking faces with accurate lip synchronization. **4)** As no structural intermediate information is used in our system, our model requires little pre-processing and is robust to input views.

2. Related Work

Audio-Driven Talking Face Generation. Driving talking faces with audio input [7] has long been important research interest in both computer vision and graphics, where structural information and stitching techniques play a crucial role [4, 3, 56, 55, 75]. For example, Bregler *et al.* [4] rewrite the mouth contours. With the development of deep learning, different methods have been proposed to generate landmarks through time modeling [20, 21, 51]. However, previous methods are mostly speaker-specific. Recently, with the development of end-to-end cross audio-visual generation methods [77, 11, 69, 24, 68, 63, 73, 23, 64, 22], researchers have explored the speaker-independent setting [14, 50, 45, 71, 74], which seeks a universal model that handles all identities with one or few frame references. After Chung *et al.* [14] firstly propose an end-to-end reconstruction-based network, Zhou *et al.* [71] further disentangle identity from speech content by adversarial representation learning. The key idea for reconstruction-based methods is to determine the synchronization between audio and videos [17, 71, 45, 44] where

Prajwal *et al.* [44] sync mouth with audio for inpainting-based reconstruction. However, these reconstruction-based works normally neglect head movements due to the difficulty in decoupling head-poses from facial movements.

As more compact and easy-to-learn targets, structural information is leveraged as intermediate representations within GAN-based reconstruction pipelines [10, 18, 52, 49, 74, 65]. Chen *et al.* [10] and Das *et al.* [18] both use two-stage networks to predict 2D landmarks first and generate faces. But the pose and mouth are also entangled within 2D landmarks. On the other hand, 3D tools [2, 19, 5, 29] serve as strong intermediate representations. Zhou *et al.* particularly model 3D talking landmarks with personalized movements, which generates the currently most natural results. However, free pose control is not achieved in their model. Chen *et al.* and Yi *et al.* both leverage 3D model to learn natural pose. However, their methods cannot render good quality under the "one-shot" condition. Moreover, the 3D fitting accuracy would drop significantly under extreme conditions. Here, we propose to achieve free pose control for one reference frame without relying on any intermediate structural information.

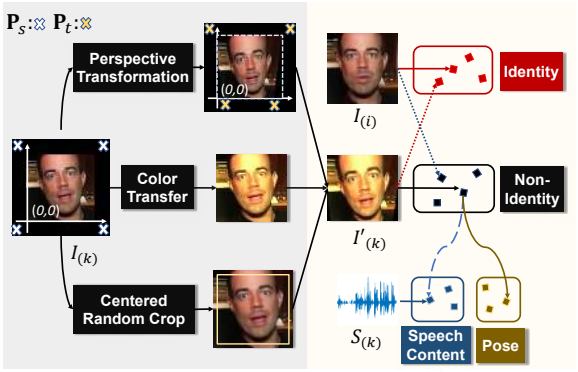
Face Reenactment. Our work also relates to visually driven talking faces, which is studied in the realm of face reenactment. Similar to their audio-driven counterparts, most face reenactment works depend on structural information such as landmark [61, 26, 67, 66], segmentation map [12, 6] and 3D models [53, 33, 32]. Specifically, certain works [60, 47, 58] learn the warping between pixels without defined structural information. While quite a number of papers [66, 6, 18] adopt a meta-learning procedure on a few frames for a new identity, our method does not require this procedure.

3. Our Approach

We present **Pose-Controllable Audio-Visual System (PC-AVS)** that aims to achieve free pose control while driving static photos to speak with audio. The whole pipeline is depicted in Fig 3. In this section, we first explore an efficient feature learning formulation by identifying the non-identity space (Sec. 3.1), then we provide the *modularization* of audio-visual representations (Sec. 3.2). Finally, we introduce our generator and generating process (Sec. 3.3).

3.1. Identifying Non-Identity Feature Space

At first, we revisit the general setting of previous pure reconstruction-based methods. The problem is formulated in an image-to-image translation manner [27, 57], where massively-available talking face videos provide natural self-supervision. Given a K-frame video clip $\boldsymbol{V} = \{I_{(1)}, \dots, I_{(K)}\}$, the natural training goal is to generate any *target* frame $I_{(k)}$ conditioned on one frame of identity reference $I_{(ref)}$ ($ref \in [1, \dots, K]$) and the accompanied audio inputs. The raw audios are processed into spectrograms $\boldsymbol{A} = \{S_{(1)}, \dots, S_{(K)}\}$ as 2D time-frequency representations



(1) Target Frame Augmentation (2) Feature Space Encoding

Figure 2: We **identify a non-identity space** through augmenting the (target) frames corresponding to the conditional audio. (1) Three data augmentation procedures are used to account for texture, facial deformation and subtle scale perturbation, which are irrelevant to learning pose and speech content. (2) The feature spaces that we target at learning.

for more compact information preservation. Previous studies [9, 76, 71, 45, 44] have verified that learning the mutual and synchronized *speech content* formation within both audio and visual modalities is effective for driving lips with audios. However, as no absolute pose information can be inferred from audios [35], methods formulated in this way mostly keep the original pose unchanged.

In order to encode additional pose information, we first point out the existence of a general *non-identity space* for representing all identity-repelling information including poses and facial movements. As depicted in Fig. 2, the encoding of such a space is through careful data augmentation on the target frame $I_{(k)}$. To account for two major aspects, namely texture and facial structure information, we apply two types of data augmentation to the target frames: *color transfer* and *perspective transformation*. Additionally, a *centered random crop* is also applied to alleviate the influence of facial scale changes in face detectors.

The *color transfer* is made by simply altering RGB channels randomly. As for the *perspective transformation*, we set four source points $\mathbf{P}_s = [[-r_s, -r_s],\ [\mathrm{W} + r_s, -r_s],\ [-r_s, \mathrm{H} + r_s],\ [\mathrm{W} + r_s, \mathrm{H} + r_s]]^{\mathrm{T}}$ outside of the original images by a random margin of r_s. Then through symmetrically moving the points along the x axis, we set target points $\mathbf{P}_t = [[-r_s + r_t, -r_s],\ [\mathrm{W} + r_s, -r_s],\ [-r_s, \mathrm{H} + r_s],\ [\mathrm{W} + r_s + r_t, \mathrm{H} + r_s]]^{\mathrm{T}}$ with another random step r_t (see Fig. 2 (1) for details). The transformation can be learned by solving:

$$[\mathbf{P}_t, \mathbf{e}] = [\mathbf{P}_s, \mathbf{e}] * \mathbf{M}, \tag{1}$$

where $\mathbf{M}$ is a 3×3 matrix, $*$ denotes the matrix multiplication and $\mathbf{e} = [1, 1, 1, 1]^{\mathrm{T}}$.

In this *non-identity space* lies the encoded features $\mathbf{F}_n = \{f_{n(1)}, \dots, f_{n(K)}\}$ from the augmented target frames

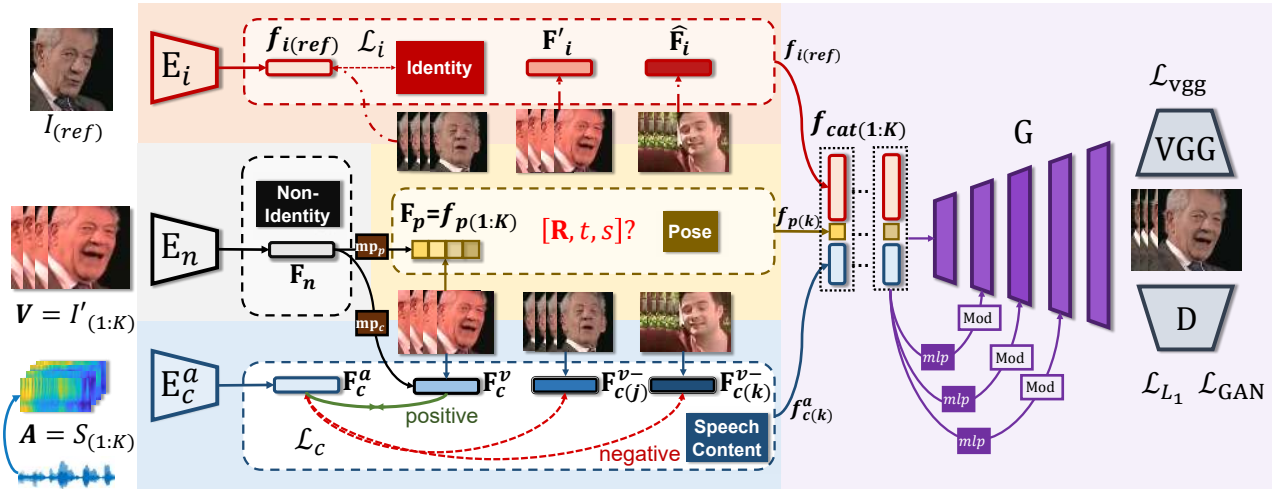


Figure 3: **The overall pipeline of our Pose-Controllable Audio-Visual System(PC-AVS) framework**. The identity reference $I_{(ref)}$ is encoded by E_i to the *identity space* (red).Encoder E_n encodes video clip $\boldsymbol{V}$ to $\mathbf{F}_n$ in the *non-identity space* (grey). Then it is mapped to $\mathbf{F}_c^v$ in the *speech content space* (blue), which it shares with $\mathbf{F}_c^a$ encoded by E_c^a from audio spectrograms $\boldsymbol{A}$. We also draw encodings $\mathrm{F}_{c(j)}^{v-}$, $\mathrm{F}_{c(k)}^{v-}$ from two negative examples. Their features in *pose* and *identity* spaces are also shown. Specifically, we map $\mathbf{F}_n$ to pose features $\mathbf{F}_p = f_{p(1:k)}$ in the *pose space* (yellow). Though motivated by 3D priors, the pose features are not supervised by or necessarily represent the traditional $[\mathbf{R}, t, s]$ 3D parameters. Finally, a pair of features $\{f_{i(i)}, f_{p(k)}, f_{c(k)}^a\}$ are assembled together and sent to generator G.

$\boldsymbol{V}' = \{I'_{(1)}, \ldots, I'_{(K)}\}$ by encoder E_n. Notably, data augmentation is also introduced in [6] for learning face reenactment. Different from their goal, our derivation of this space is to assist better feature learning and further representation modularization.

3.2. Modularization of Representations

Supported with such a *non-identity space*, we then modularize audio-visual information into three feature spaces namely the *speech content space*, the *head pose space* and *identity space*.

Learning Speech Content Space. It has been verified that learning the natural synchronization between visual mouth movements and auditory utterances is valuable for driving images to speak [71, 44]. Thus embedding space that contains synchronized audio-visual features as the *speech content space*.

Specifically, we first define a mapping of fully connected layers from non-identity features $\mathbf{F}_n$ to the visual speech content features $\mathbf{F}_c^v = \mathrm{mp}_c(\mathbf{F}_n) = \{f_{c(1)}^v, \ldots, f_{c(K)}^v\}$. Meanwhile, the audio inputs are encoded by the encoder E_c^a. Under our assumption, the audio features $\mathbf{F}_c^a = \mathrm{E}_c^a(\boldsymbol{A}) = \{f_{c(1)}^a, \ldots, f_{c(K)}^a\}$ share the **same space** with $\mathbf{F}_c^v$. Thus the feature distance between timely aligned audio-visual pairs should be lower that non-aligned pairs.

We adopt the contrastive learning [17, 36, 71] protocol to seek the synchronization between audio and visual features. Different from the L_2 contrastive loss mostly leveraged before, we take the advantage of the more stable form of InfoNCE [39]. Concretely, for visual to audio synchronization, we regard the ensemble of timely aligned features $\mathbf{F}_c^v \in \mathbb{R}^{l_c}$ and $\mathbf{F}_c^a \in \mathbb{R}^{l_c}$ as positive pairs and sample N^- negative audio features $\mathbf{F}_c^{a-} \in \mathbb{R}^{N^- \times l_c}$. The negative audio clips could be sampled from other videos or from the same recording with a time-shift. For feature distances measurement, we adopt the cosine distance $\mathcal{D}(\mathbf{F}_1, \mathbf{F}_2) = \frac{\mathbf{F}_1^{\mathbf{T}} * \mathbf{F}_2}{|\mathbf{F}_1| \cdot |\mathbf{F}_2|}$, where closer features render larger scores. In this way, the contrastive learning can be formulated to a classification problem with $(N^- + 1)$ classes:

$$\mathcal{L}_c^{v2a} = -\log\left[\frac{\exp(\mathcal{D}(\mathbf{F}_c^v, \mathbf{F}_c^a))}{\exp(\mathcal{D}(\mathbf{F}_c^v, \mathbf{F}_c^a)) + \sum_{j=1}^{N^-} \exp(\mathcal{D}(\mathbf{F}_c^v, \mathbf{F}_{c(j)}^{a-}))}\right]. \tag{2}$$

The audio to visual synchronization loss $\mathcal{L}_c^{a2v}$ can also be achieved in a symmetric way as illustrated in the *speech content space* of Fig 3, which we omit here. The total loss for encoding this space is the sum of both:

$$\mathcal{L}_c = \mathcal{L}_c^{v2a} + \mathcal{L}_c^{a2v}. \tag{3}$$

Devising Pose Code. Without relying on any precisely recognized structural information, such as pre-defined 3D parameters, it is difficult to explicitly model a pose. Here, we propose to devise an implicit pose code using only subtle prior knowledge of 3D pose parameters [2]. Concretely, the 3D head pose information can be expressed by a mere of 12 dimensions with a rotation matrix $\mathbf{R} \in \mathbb{R}^{3\times3}$, a positional translation vector $\mathbf{t} \in \mathbb{R}^2$ and a scale scalar s. Thus we define another fully connected mapping from the *non-identity space* to a low-dimensional feature with exactly the size of 12: $\mathbf{F}_p = \mathrm{mp}_p(\mathbf{F}_n) = \{f_{p(1)}, \ldots, f_{p(K)}\}$.

The idea has some similarity with papers that use 3D priors for unsupervised 3D representation learning [38, 48]. Differently, we only use the prior knowledge on the minimum dimension of data needed. Though the implicitly learned feature could not possibly possess the same value of real 3D pose parameters, this intuition is important to our design. A pose code with larger dimensions may contain additional information that is not desired. The reason for the defined pose code to work is described in Sec. 3.3.

Identity Space Encoding. The learning of identity space has been well addressed in previous studies [14, 71, 74]. As we train our networks on the videos of celebrities [15], the identity labels naturally exist. Our identity space $\mathbf{F}_i = \mathrm{E}_i(\boldsymbol{V}) = \{f_{i(1)}, \ldots, f_{i(K)}\}$ can be learned on identity classification with softmax cross-entropy loss $\mathcal{L}_i$.

3.3. Talking Face Generation

The features embedded in the three modularized spaces are composed for the final reconstruction of target frames $\boldsymbol{V}$. For a specific case, we concatenate $f_{i(ref)}$, $f^a_{c(k)}$ and $f_{p(k)}$ which are encoded from $I_{(ref)}$, $S_{(k)}$ and $I'_{(k)}$ respectively, and target to generate $I_{(k)}$ through a generator G.

Generator Design. A number of previous studies [14, 71, 45, 44] leverage skip connections for better input identity preserving. However, such a paradigm further restricts the possibility of altering poses, as the low-level information preserved through the connection greatly affects the generation results. Different from their structures, we directly encode the spatial dimension of the features to be one. With the recent development of generative model structures, style-based generator has achieved great success in the field of image generation [30, 31]. Their expressive ability in recovering details and style manipulation is also a crucial component of our framework.

In this paper, we empirically propose to generate faces through modulated convolution, which has been proven to be effective in image-to-image translation [41]. Detailedly, the concatenated features $f_{cat(k)} = \{f_{i(ref)}, f^a_{c(k)}, f_{p(k)}\}$ serve as latent codes to modulate the weights of the convolution kernels of the generator. At each convolutional block, a multi-layer perceptron is learned to map a $f_{cat(k)}$ to a modulation vector $\mathcal{M}$ which has the same dimension as the input feature's channels. For each value w_{xyz} in the convolution kernel weight w, where x is its position on the input feature channels, y is related to output channel numbers and z represents the spatial location, it is modulated and normalized given the x's value of $\mathcal{M}$ as:

$$w^m_{xyz} = \frac{\mathcal{M}_x \cdot w_{xyz}}{\sqrt{\sum_{x,z}(\mathcal{M}_x \cdot w_{xyz})^2 + \epsilon}}, \tag{4}$$

where ϵ is a small constant for avoiding numerical errors.

Network Training. Finally, the feature space modularization and generator are trained jointly by image reconstruction. We directly borrow the same loss functions applied in [42]. The generated and ground truth images are sent to a multi-scale discriminator D with N_D layers. The discriminator is utilized for both computing feature map L_1 distances within its layers, and adversarial generative learning. The perceptual loss that relies on a pretrained VGG network with N_P layers is also used. All loss functions can be briefly denoted as:

$$\begin{aligned}\mathcal{L}_{\text{GAN}} = \min_{\mathrm{G}}\max_{\mathrm{D}} \sum_{n=1}^{N_D} (&\mathbb{E}_{I_{(k)}}[\log \mathrm{D}_n(I_{(k)})] \\ &+ \mathbb{E}_{f_{cat(k)}}[\log(1 - \mathrm{D}_n(\mathrm{G}(f_{cat(k)}))]),\end{aligned} \tag{5}$$

$$\mathcal{L}_{L_1} = \sum_{i=1}^{N_D} \|\mathrm{D}_n(\mathrm{I}_{(k)}) - \mathrm{D}_n(\mathrm{G}(f_{cat(k)}))\|_1, \tag{6}$$

$$\mathcal{L}_{\text{vgg}} = \sum_{i=1}^{N_P} \|\mathrm{VGG}_n(I_{(k)}) - \mathrm{VGG}_n(\mathrm{G}(f_{cat(k)}))\|_1. \tag{7}$$

This reconstruction training not only maps features to the image space but also implicitly ensures the representation modularization in a complementary manner. While features in the content space $\mathbf{F}_c$ are synced audio-visual representations without pose information, in order to suppress the reconstruction loss, the low-dimension f_p automatically compensates for pose information. Moreover, in most previous methods, the poses between generated images and their supervisions are not matched, which harms the learning process. Differently in our setting, as the generated pose is aligned with ground truth through our pose code f_p, the learning of the speech content feature can further be benefited from the reconstruction loss. This leads to more accurate lip synchronization.

The overall learning objective for the whole system is formulated as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{\text{GAN}} + \lambda_1\mathcal{L}_{L_1} + \lambda_v\mathcal{L}_{\text{vgg}} + \lambda_c\mathcal{L}_c + \lambda_i\mathcal{L}_i, \tag{8}$$

where the λs are balancing coefficients.

4. Experiments

4.1. Experimental Settings

Datasets. We leverage two in-the-wild audio-visual datasets which are popularly used in a great number of previous studies, VoxCeleb2 [15] and LRW [16]. Both datasets provide detected and cropped faces.

- **VoxCeleb2** [15]. There are a total of 6,112 celebrities in VoxCeleb2 covering over 1 million utterances. While 5,994 speakers lie in the training set, 118 are in the test set. The qualities of the videos differ largely. There are extreme cases with large head pose movements, low-light conditions, and different extents of blurry. No test identity has been seen during training.

Table 1: **The quantitative results on LRW [16] and VoxCeleb2 [15].** All methods are compared under the four metrics. For LMD the lower the better, and the higher the better for other metrics. †Note that we directly evaluate the authors' generated samples on VoxCeleb2 under their setting. They have not provided examples on LRW.

	LRW [16]				VoxCeleb2 [15]			
Method	SSIM ↑	CPBD ↑	LMD ↓	Sync_{conf} ↑	SSIM ↑	CPBD ↑	LMD ↓	Sync_{conf} ↑
ATVG [10]	0.810	0.102	5.25	4.1	0.826	0.061	**6.49**	4.3
Wav2Lip [44]	**0.862**	0.152	5.73	**6.9**	0.846	0.078	12.26	4.5
MakeitTalk [74]	0.796	0.161	7.13	3.1	0.817	0.068	31.44	2.8
Rhythmic Head† [8]	-	-	-	-	0.779	0.802	14.76	3.8
Ground Truth	1.000	0.173	0.00	6.5	1.000	0.090	0.00	5.9
Ours-Fix Pose	0.815	0.180	6.14	6.3	0.820	**0.084**	7.68	5.8
PC-AVS (Ours)	**0.861**	**0.185**	**3.93**	6.4	**0.886**	**0.083**	6.88	**5.9**

- **Lip Reading in the Wild (LRW)** [16]. This dataset is originally proposed for lip reading. It contains over 1000 utterances of 500 different words within each 1-second video. Compared to VoxCeleb2, the videos in this dataset are mostly clean with high-quality and near-frontal faces of BBC news. Thus most of the utterances and identities in the test set are seen during training.

Implementation Details. The structure of E_i is a ResNeXt50 [62]. E_n is borrowed from [71] and E_c^a is a ResNetSE34 borrowed from [13]. The generator consists of 6 blocks of modulated convolutions. Different from certain previous works [9, 71, 8], we do not align facial key points for each frame. All images are of size 224×224. The audios are pre-processed to 16kHz, then converted to mel-spectrograms with FFT window size 1280, hop length 160 and 80 Mel filter-banks. For each frame, 0.2s mel-spectrogram with the target frame time-step in the middle are sampled as condition. The λs are empirically set to 1. Our models are implemented on PyTorch [43] with eight 16 GB Tesla V100 GPUs. Note that our identity encoder is pre-trained on the labels provided in the Voxceleb2 [15] dataset with $\mathcal{L}_i$. The speech content space is also pretrained first with loss $\mathcal{L}_c$. Then these models are loaded to the overall framework for generator and pose space learning.

Comparison Methods. We compare our methods with the best models currently available that support arbitrary-subject talking face generation. They are: **AVTG** [10], the representative of 2D landmark-based method; **Wav2Lip** [44], a reconstruction-based method that claims state-of-the-art lip sync results. **MakeitTalk** [74] which leverages 3D landmarks and generates personalized head movements according to the driving audios. **Rhythmic Head** [8] which generates rhythmic head motion under a different setting. Note that its code is not run-able until this paper's final version, thus we only show two of its one-shot results in Fig. 4, which is generated with the help of the authors. The numerical comparisons in Table 1 are conducted on the authors' provided VoxCeleb2 samples under their setting for reference. Specifically, we also show the evaluation directly on the **Ground Truth**.

4.2. Quantitative Evaluation

Evaluation Metrics. We conduct quantitative evaluations on metrics that have previously been involved in the field of talking face generation. We use **SSIM** [59] to account for the generation quality, and the cumulative probability blur detection (**CPBD**) measure [37] is adopted from [54] to evaluate the sharpness of the results. Then we use both Landmarks Distance (**LMD**) around the mouths [10] and the confidence score ($\textbf{Sync}_{conf}$) proposed in SyncNet [16] to account for the accuracy of mouth shapes and lip sync. Though we do not use landmarks for video generation, we specifically detect landmarks for evaluation. Results on less informative metrics such as PSNR for image quality and CSIM [8, 66] for identity preserving are shown in supplementary material.
Image Generation Paradigm. We conduct the experiments under the self-driven setting that, the first image of each video within the test sets is selected as the identity reference. Then the audios are used as driving conditions to generate their accompanying whole videos. The evaluation is conducted between all generated frames and the ground truths. When the pose code is not given for our method, we can fix it with the same angle as the input, which keeps the head still. We refer results generated under this setting as **Ours-Fix Pose**. As our method is pose-controllable by another *pose source* video, we seek to leverage the pose information under a fair setting. Specifically, we use the target frames as pose sources to drive another reference image together with a different audio. In this way, an extra video with supposedly the same pose but different identities and mouth shapes is generated, serving as the pose source for our method.
Evaluation Results. The results are shown in Table 1. It can be seen that our method reaches the best under most of the metrics on both datasets. On LRW, though Wav2Lip [44] outperforms our method given two metrics, the reason is

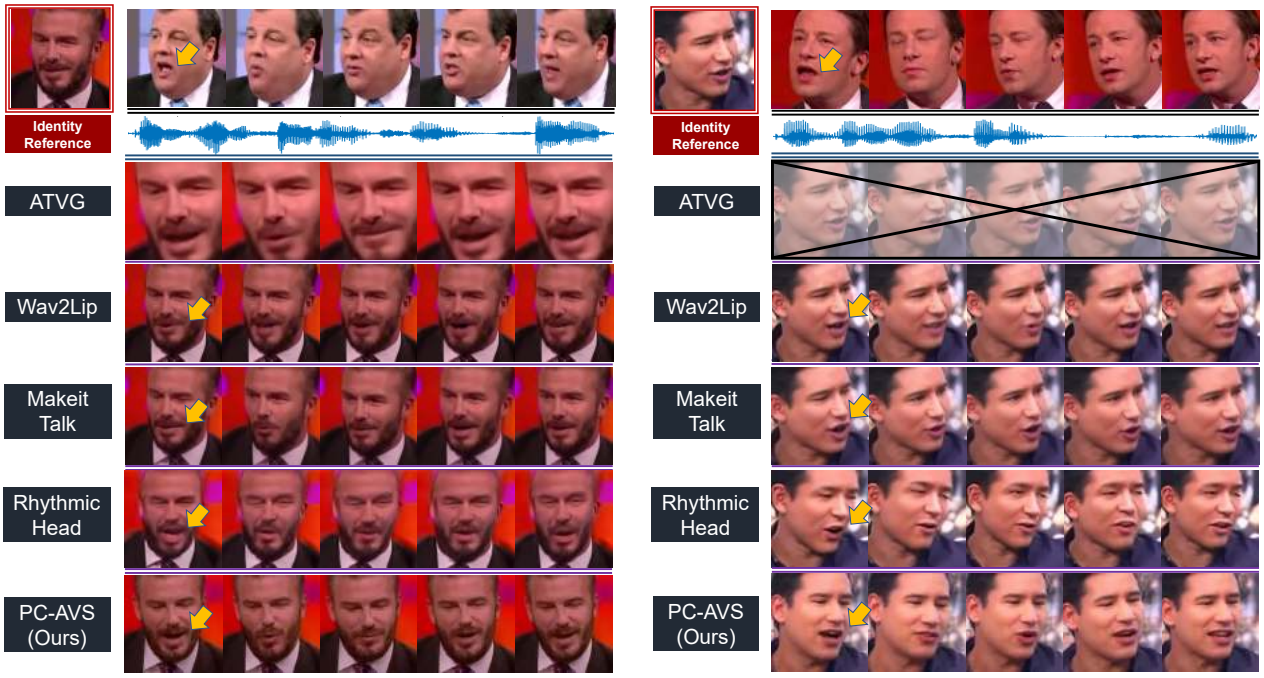


Figure 4: **Qualitative results**. In the top row are the audio-synced videos. ATVG [10] are accurate on the left. But with cropped faces, results seem non-real. Moreover, its detector fails on the right case. The mouth openings of Wav2Lip [44] are basically correct, but they generate static results. While MakeitTalk [74] generates subtle head motions, the mouth shapes of their model are not accurate. Both our method and Rhythmic Head [8] create diverse head motions, but their identify-preserving is much worse.

Table 2: **User study measured by Mean Opinion Scores.** Larger is higher, with the maximum value to be 5.

MOS on \ Approach	ATVG [10]	Wav2Lip [44]	MakeitTalk [71]	**PC-AVS (Ours)**
Lip Sync Quality	2.87	**3.98**	2.33	**3.98**
Head Movement Naturalness	1.26	1.60	2.93	**4.20**
Video Realness	1.60	2.84	3.22	**4.07**

that their method keeps most parts of the input unchanged while samples in LRW are mostly frontal faces. Notably, their SyncNet confidence score outperforms that of ground truth's. But this only proves that their lip-sync results are nearly comparable to the ground truth. Our model performs better than theirs on the LMD metric. Moreover, the SyncNet confidence score of our results is also close to the ground truth on the more complicated VoxCeleb2 dataset, meaning that we can generate accurate lip-sync videos robustly.

4.3. Qualitative Evaluation

Subject evaluation is crucial for determining the quality of the results[2] for generative tasks. Here we show the comparison of our methods against previous state-of-the-arts (listed in Sec. 4.1) on two cases in Fig. 4. For our method, we sample one random *pose source* video whose first frame's pose feature distance is closer to the input's. It can be seen that our method generates more diverse head motions and more accurate lip shapes. For the left case, only results of ATVG [10] and Rhythmic Head [8] match our mouth shapes. Notably, the relied landmark detector [34] of ATVG fails on the right case, thus no results can be produced. Similarly, the lip sync quality of MakeitTalk [74] is better on the left side. The Face under the large pose on the right degrades the 3D landmarks' prediction. Both these facts verify the non-robustness of structural information-based methods. While producing dynamic head motions, the generated quality of Rhythmic Head's is much worse given only one identity reference.

User Study. We conduct a user study of 15 participants for their opinions on 30 videos generated by ours and three competing methods. Twenty videos are generated from fifteen reference images in the test set of VoxCeleb and ten from LRW. The driving audios are also arbitrarily chosen from the test set in a cross-driving setting. Notably, we cannot generate Rhythmic Head [8] cases without the help of the authors, this method is not involved. We adopt the widely used Mean Opinion Scores (MOS) rating protocol. The users are required to give their ratings (1-5) on the following three aspects for each video: (1) Lip sync qualities; (2) naturalness

[2]Please refer to https://hangz-nju-cuhk.github.io/projects/PC-AVS for demo videos and comparisons.

Table 3: **Ablation study with quantitative comparisons on VoxCeleb2 [15].** The results are shown when we vary the loss function. pose feature length and generator structure.

Method	SSIM ↑	Mouth LMD ↓	Pose LMD ↓	Sync_{conf} ↑
w/o $\mathcal{L}_c$	0.836	13.52	16.51	4.7
Pose-dim 36	0.860	9.17	9.40	5.5
AdaIN G	0.750	10.58	8.78	5.5
Ours	**0.886**	**6.88**	**5,9**	**7.62**

Figure 5: **Ablation study with visual results.** The mouth shapes are same among results but not synced with pose source.

of head movements; and (3) the realness of results, whether they can tell the videos are fake or not.

The results are listed in Table 2. As both ATVG [10] and Wav2Lip [44] generate near-stationary results, their scores on head movements and video realness are reasonably low. However, the lip sync score of Wav2Lip [44] is the same as ours, outperforming MakeitTalk [74]. While the latter is famous for the realness of their generated image, the users prefer our results on the head motion naturalness and video realness, demonstrating the effectiveness of our method.

4.4. Further Analysis

Ablation Studies. We conduct ablation studies given three important aspects of our method. The contrastive loss for audio-visual synchronization;The pose code length which is empirically set to 12 and design of the generator. Thus we conduct experiments on our model (1) w/o contrastive loss; (2) with different pose feature lengths and (3) change generator to AdaIN-based [30] form. Note that the traditional skip-connection form does not work in our case. Except for the previous metrics, we propose an additional metric namely *Pose LMD*, which computes only the landmark distances between facial contours to represent the pose.

The numerical results on VoxCeleb2 are shown in Table 3 and the visualizations are shown in Fig. 5. Without the contrastive loss, the audios cannot be synced perfectly with the speech content, thus the whole modularization learning procedure would break, leading to the failure of pose code. Under the same training time, we observe drops on both metrics for larger pose code, possibly due to the growth of

Figure 6: **Results under extreme condition.** We can drive faces under large poses, and even frontalize them by setting the pose code to all zeros.

training difficulties along with the information in the pose code. Note that we also try learning without the target frame data augmentation (Sec. 3.1), similar to Ours w/o $\mathcal{L}_c$, the learning of the pose would fail.

Extreme View Robustness and Face Frontalization. Here we show the ability of our model to handle extreme views and achieving talking face frontalization. As ATVG fails again on the reference input of Fig. 6, Wav2Lip would create artifacts on the mouth. Our model, on the other hand, not only can generate accurate lip motion but also can frontalize the faces while preserving the identity when setting the values of the pose code to zero.

5. Conclusion

In this paper, we propose **Pose-Controllable Audio-Visual System (PC-AVS)**, which generates accurately lip-synced talking faces with free pose control from other videos. We emphasize several appealing properties of our framework: 1) Without using any structural intermediate information, we implicitly devise a pose code and modularize audio-visual representations into the latent identity, speech content, and pose space. 2) The complementary learning procedure ensures more accurate lip sync results than previous works. 3) The pose of our talking faces can be freely controlled by another pose source video, which can hardly be achieved before. 4) Our model shows great robustness under extreme conditions, such as large poses and viewpoints.

Acknowledgements. We would like to thank Lele Chen for his generous help with the comparison, and Siwei Tang for his voice included in our video. Hang Zhou would also like to thank his grandmother, Shuming Wang, for her love throughout her life. This research was conducted in collaboration with SenseTime. It is supported in part by the General Research Fund through the Research Grants Council of Hong Kong under Grants (Nos. 14202217, 14203118, 14208619), in part by Research Impact Fund Grant No. R5001-18, and in part by NTU NAP and A*STAR through the Industry Alignment Fund - Industry Collaboration Projects Grant.

References

[1] Robert Anderson, Björn Stenger, Vincent Wan, and Roberto Cipolla. An expressive text-driven 3d talking head. In *SIGGRAPH*, 2013. 2

[2] Volker Blanz, Thomas Vetter, et al. A morphable model for the synthesis of 3d faces. In *SIGGRAPH*, 1999. 2, 3, 4

[3] Matthew Brand. Voice puppetry. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 1999. 2

[4] Christoph Bregler, Michele Covell, and Malcolm Slaney. Video rewrite: Driving visual speech with audio. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, 1997. 2

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

[6] Egor Burkov, Igor Pasechnik, Artur Grigorev, and Victor Lempitsky. Neural head reenactment with latent pose descriptors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3, 4

[7] Lele Chen, Guofeng Cui, Ziyi Kou, Haitian Zheng, and Chenliang Xu. What comprises a good talking-head video generation?: A survey and benchmark. *arXiv preprint arXiv:2005.03201*, 2020. 2

[8] Lele Chen, Guofeng Cui, Celong Liu, Zhong Li, Ziyi Kou, Yi Xu, and Chenliang Xu. Talking-head generation with rhythmic head motion. *European Conference on Computer Vision (ECCV)*, 2020. 2, 6, 7

[9] Lele Chen, Zhiheng Li, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Lip movements generation at a glance. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2, 3, 6

[10] Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 3, 6, 7, 8

[11] Lele Chen, Sudhanshu Srivastava, Zhiyao Duan, and Chenliang Xu. Deep cross-modal audio-visual generation. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, pages 349–357, 2017. 2

[12] Zhuo Chen, Chaoyue Wang, Bo Yuan, and Dacheng Tao. Puppeteergan: Arbitrary portrait animation with semantic-aware appearance transformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[13] Joon Son Chung, Jaesung Huh, Seongkyu Mun, Minjae Lee, Hee Soo Heo, Soyeon Choe, Chiheon Ham, Sunghwan Jung, Bong-Jin Lee, and Icksang Han. In defence of metric learning for speaker recognition. In *Interspeech*, 2020. 6

[14] Joon Son Chung, Amir Jamaludin, and Andrew Zisserman. You said that? In *BMVC*, 2017. 2, 5

[15] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*, 2018. 5, 6, 8

[16] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In *ACCV*, 2016. 5, 6

[17] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *ACCV*, 2016. 2, 4

[18] Dipanjan Das, Sandika Biswas, Sanjana Sinha, and Brojeshwar Bhowmick. Speech-driven facial animation using cascaded gans for learning of motion and texture. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 3

[19] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of IEEE Computer Vision and Pattern Recognition Workshop on Analysis and Modeling of Faces and Gestures*, 2019. 3

[20] Bo Fan, Lijuan Wang, Frank K Soong, and Lei Xie. Photo-real talking head with deep bidirectional lstm. In *ICASSP*, 2015. 2

[21] Bo Fan, Lei Xie, Shan Yang, Lijuan Wang, and Frank K Soong. A deep bidirectional lstm approach for video-realistic talking head. *Multimedia Tools and Applications*, 2016. 2

[22] Chuang Gan, Deng Huang, Peihao Chen, Joshua B Tenenbaum, and Antonio Torralba. Foley music: Learning to generate music from videos. *Proceedings of the European conference on computer vision (ECCV)*, 2020. 2

[23] Chuang Gan, Deng Huang, Hang Zhao, Joshua B Tenenbaum, and Antonio Torralba. Music gesture for visual sound separation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[24] Ruohan Gao and Kristen Grauman. 2.5 d visual sound. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[25] Ruohan Gao and Kristen Grauman. Visualvoice: Audio-visual speech separation with cross-modal consistency. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[26] Po-Hsiang Huang, Fu-En Yang, and Yu-Chiang Frank Wang. Learning identity-invariant motion representations for cross-id face reenactment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[27] 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)*, 2017. 3

[28] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Wayne Wu, Chan Change Loy, Xun Cao, and Feng Xu. Audio-driven emotional video portraits. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[29] Zi-Hang Jiang, Qianyi Wu, Keyu Chen, and Juyong Zhang. Disentangled representation learning for 3d face shape. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[30] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5, 8

[31] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 5

[32] Hyeongwoo Kim, Mohamed Elgharib, Michael Zollhöfer, Hans-Peter Seidel, Thabo Beeler, Christian Richardt, and Christian Theobalt. Neural style-preserving visual dubbing. *ACM Transactions on Graphics (TOG)*, 2019. 3

[33] Hyeongwoo Kim, Pablo Garrido, 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)*, 2018. 2, 3

[34] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 2009. 7

[35] Givi Meishvili, Simon Jenni, and Paolo Favaro. Learning to have an ear for face super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[36] Arsha Nagrani, Joon Son Chung, Samuel Albanie, and Andrew Zisserman. Disentangled speech embeddings using cross-modal self-supervision. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020. 2, 4

[37] Niranjan D Narvekar and Lina J Karam. A no-reference perceptual image sharpness metric based on a cumulative probability of blur detection. In *2009 International Workshop on Quality of Multimedia Experience*, 2009. 6

[38] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 5

[39] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4

[40] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. *European Conference on Computer Vision (ECCV)*, 2018. 2

[41] Taesung Park, Alexei A. Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European Conference on Computer Vision (ECCV)*, 2020. 5

[42] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5

[43] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems (NeurIPS)*, 2019. 6

[44] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia (ACMMM)*, 2020. 2, 3, 4, 5, 6, 7, 8

[45] K R Prajwal, Rudrabha Mukhopadhyay, Jerin Philip, Abhishek Jha, Vinay Namboodiri, and CV Jawahar. Towards automatic face-to-face translation. In *Proceedings of the 27th ACM International Conference on Multimedia (ACMMM)*, 2019. 2, 3, 5

[46] Alexander Richard, Colin Lea, Shugao Ma, Jurgen Gall, Fernando de la Torre, and Yaser Sheikh. Audio-and gaze-driven facial animation of codec avatars. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021. 2

[47] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 3

[48] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 5

[49] Linsen Song, Wayne Wu, Chen Qian, Ran He, and Chen Change Loy. Everybody's talkin': Let me talk as you want. *arXiv preprint arXiv:2001.05201*, 2020. 2, 3

[50] Yang Song, Jingwen Zhu, Dawei Li, Xiaolong Wang, and Hairong Qi. Talking face generation by conditional recurrent adversarial network. *IJCAI*, 2019. 2

[51] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (TOG)*, 2017. 2

[52] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural voice puppetry: Audio-driven facial reenactment. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2, 3

[53] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[54] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Realistic speech-driven facial animation with gans. *International Journal of Computer Vision*, 2019. 2, 6

[55] Lijuan Wang, Wei Han, and Frank K Soong. High quality lip-sync animation for 3d photo-realistic talking head. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012. 2

[56] Lijuan Wang, Xiaojun Qian, Wei Han, and Frank K Soong. Synthesizing photo-real talking head via trajectory-guided sample selection. In *Eleventh Annual Conference of the International Speech Communication Association*, 2010. 2

[57] 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)*, 2018. 3

[58] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[59] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 2004. 6

[60] O. Wiles, A.S. Koepke, and A. Zisserman. X2face: A network for controlling face generation by using images, audio, and pose codes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 3

[61] Wayne Wu, Yunxuan Zhang, Cheng Li, Chen Qian, and Chen Change Loy. Reenactgan: Learning to reenact faces via boundary transfer. In *European Conference on Computer Vision (ECCV)*, 2018. 3

[62] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 6

[63] Xudong Xu, Bo Dai, and Dahua Lin. Recursive visual sound separation using minus-plus net. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 2

[64] Xudong Xu, Hang Zhou, Ziwei Liu, Bo Dai, Xiaogang Wang, and Dahua Lin. Visually informed binaural audio generation without binaural audios. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[65] Ran Yi, Zipeng Ye, Juyong Zhang, Hujun Bao, and Yong-Jin Liu. Audio-driven talking face video generation with natural head pose. *arXiv preprint arXiv:2002.10137*, 2020. 2, 3

[66] 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)*, 2019. 3, 6

[67] Jiangning Zhang, Xianfang Zeng, Mengmeng Wang, Yusu Pan, Liang Liu, Yong Liu, Yu Ding, and Changjie Fan. Freenet: Multi-identity face reenactment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[68] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 2

[69] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *Proceedings of the European conference on computer vision (ECCV)*, 2018. 2

[70] Hang Zhou, Jihao Liu, Ziwei Liu, Yu Liu, and Xiaogang Wang. Rotate-and-render: Unsupervised photorealistic face rotation from single-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[71] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. Talking face generation by adversarially disentangled audio-visual representation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2019. 2, 3, 4, 5, 6, 7

[72] Hang Zhou, Ziwei Liu, Xudong Xu, Ping Luo, and Xiaogang Wang. Vision-infused deep audio inpainting. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 2

[73] Hang Zhou, Xudong Xu, Dahua Lin, Xiaogang Wang, and Ziwei Liu. Sep-stereo: Visually guided stereophonic audio generation by associating source separation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[74] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makeittalk: Speaker-aware talking head animation. *SIGGRAPH ASIA*, 2020. 2, 3, 5, 6, 7, 8

[75] Yang Zhou, Zhan Xu, Chris Landreth, Evangelos Kalogerakis, Subhransu Maji, and Karan Singh. Visemenet: Audio-driven animator-centric speech animation. *ACM Transactions on Graphics (TOG)*, 2018. 2

[76] Hao Zhu, Huaibo Huang, Yi Li, Aihua Zheng, and Ran He. Arbitrary talking face generation via attentional audio-visual coherence learning. *IJCAI*, 2020. 2, 3

[77] Hao Zhu, Mandi Luo, Rui Wang, Aihua Zheng, and Ran He. Deep audio-visual learning: A survey. *arXiv preprint arXiv:2001.04758*, 2020. 2