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

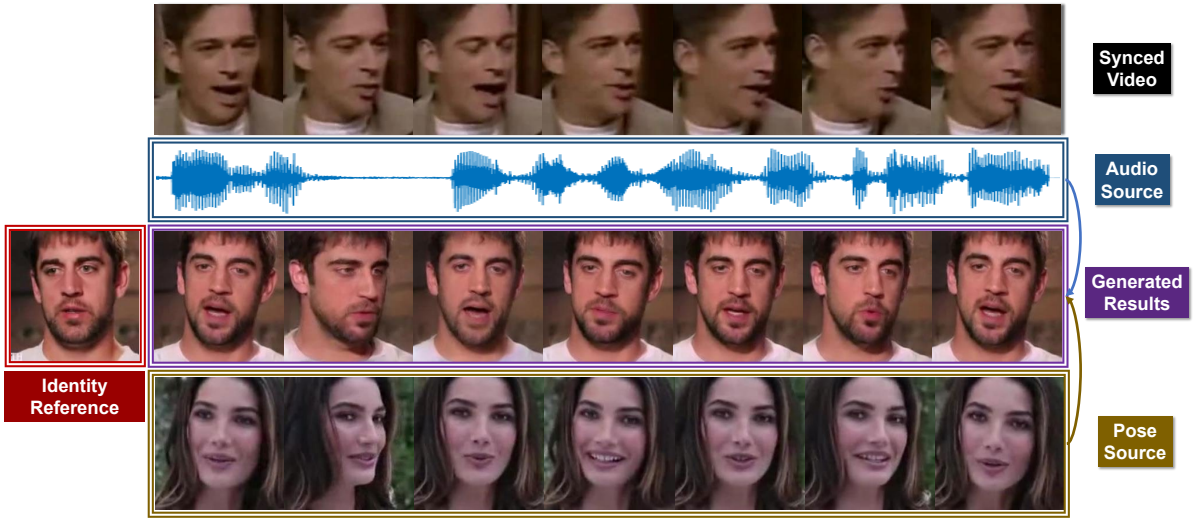


Figure 1: **Illustration of Pose-Controllable Audio-Visual System (PC-AVS).** Our approach takes one frame as identity reference and generates audio-driven talking faces with pose controlled by another *pose source* video. The mouth shapes of the generated frames are matched with the first row (synced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

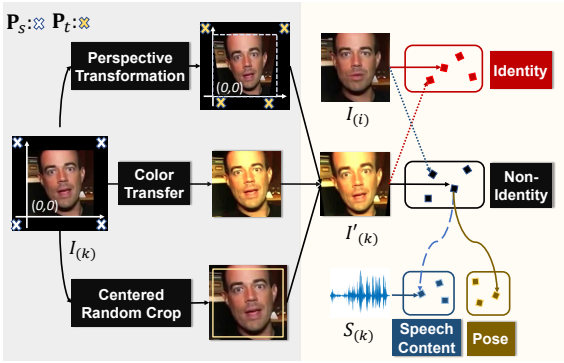
1. Introduction

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.

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. Very recently, a few works have addressed the problem of generating personalized rhythmic head movements from audios [3, 12]. However, they rely on a short clip of video to learn individual rhythms [3]. 2) On the other hand, all the above methods rely on 3D structural intermediate representations [10, 3, 12]. The pose information is inherently coupled with facial movements. Thus the most plausible way is to leverage 3D models [3] where the pose and expression parameters are explicitly disentangled [1]. Nevertheless, 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. 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. 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 is sufficient to represent a head pose. Thus we define a mapping from the non-identity space to a low dimension code which im-



(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. (2) The feature spaces that we target at learning.

plicitly stands for the pose. Then with additional identity supervision, the *modularization* of the whole talking face representations has been completed.

Our contributions are summarized as follows: **1)** We propose to modularize the representations of talking faces into three spaces, by devising a low-dimensional pose code inspired by 3D pose prior. **2)** The modularization is implicitly and complementarily learned in a StyleGAN2-based framework. **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. Our Approach

We present **Pose-Controllable Audio-Visual System (PC-AVS)** that achieves 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. 2.1), then we provide the *modularization* of audio-visual representations (Sec. 2.2). Finally, we introduce our generator and generating process (Sec. 2.3).

2.1. Identifying Non-Identity Feature Space

At first, we revisit the general setting of previous pure reconstruction-based methods. 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 for more compact information preservation. Previous studies [11, 8] 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, 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.

In this *non-identity space* lies the encoded features $\mathbf{F}_n = \{f_{n(1)}, \dots, f_{n(K)}\}$ from the augmented target frames $\boldsymbol{V}' = \{I'_{(1)}, \dots, I'_{(K)}\}$ by encoder E_n. Notably, data augmentation is also introduced in [2] for learning face reenactment.

2.2. Modularization of Representations

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 [11, 8]. 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 $\mathbf{F}_c^v = \mathrm{mp}_c(\mathbf{F}_n) = \{f^v_{c(1)}, \dots, f^v_{c(K)}\}$. Meanwhile, the audio inputs are encoded by the encoder E_c^a. Under our assumption, the audio features $\mathbf{F}_c^a = \mathrm{E}_c^a(\boldsymbol{A}) = \{f^a_{c(1)}, \dots, f^a_{c(K)}\}$ 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 [11] protocol to seek the synchronization between audio and visual features. 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:

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

The audio to visual synchronization loss $\mathcal{L}_c^{a2v}$ can also be achieved in a symmetric way as illustrated in the *speech*

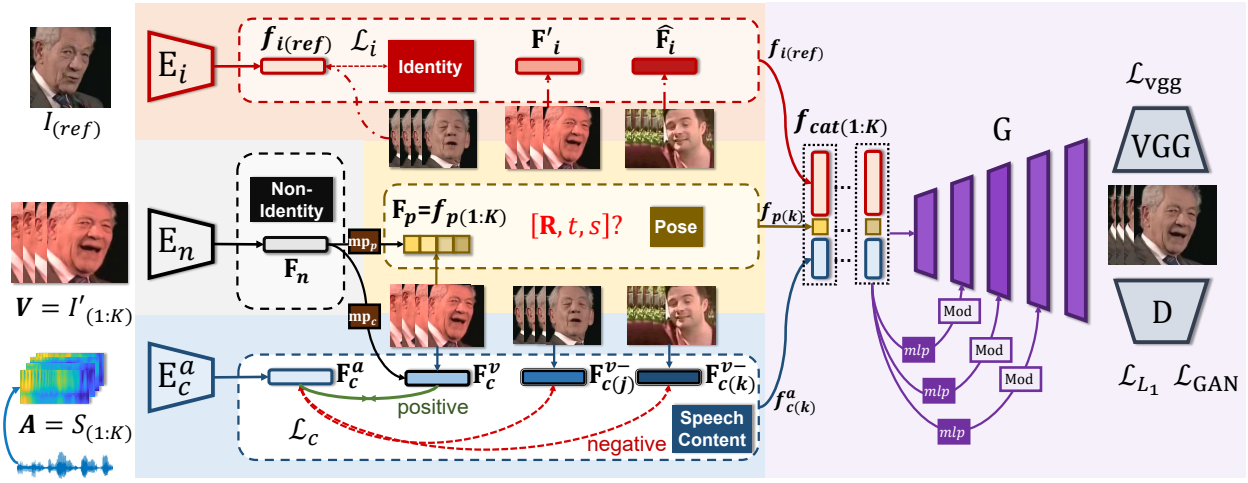


Figure 3: **The pipeline of our Pose-Controllable Audio-Visual System(PC-AVS) framework**. The identity reference $I_{(ref)}$ is encoded by E_i to the *identity space* (red). Encoder E_n encodes video clip $\boldsymbol{V}$ to $\mathbf{F}_n$ in the *non-identity space* (grey). Then it is mapped to $\mathbf{F}_c^v$ in the *speech content space* (blue), which it shares with $\mathbf{F}_c^a$ encoded by E_c^a from audio spectrograms $\boldsymbol{A}$. Specifically, we map $\mathbf{F}_n$ to pose features $\mathbf{F}_p = f_{p(1:k)}$ in the *pose space* (yellow). Finally, a pair of features $\{f_{i(i)}, f_{p(k)}, f_{c(k)}^a\}$ are assembled together and sent to generator G.

content space of Fig 3, which we omit here. The total loss for encoding this space is the sum of both:

$$\mathcal{L}_c = \mathcal{L}_c^{v2a} + \mathcal{L}_c^{a2v}. \tag{2}$$

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 [1]. 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. Differently, we only use the prior knowledge on the minimum dimension of data needed. A pose code with larger dimensions may contain additional information that is not desired.

Identity Space Encoding. The learning of identity space has been well addressed in previous studies [5, 11, 12]. 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 $\mathcal{L}_i$.

2.3. 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 $f_{i(ref)}$, $f_{c(k)}^a$ 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. With the recent development of generative model structures, style-based generator has achieved great success in the field of image generation [6]. Their expressive ability in recovering details and style manipulation is also a crucial component of our framework. In this paper, the concatenated features $f_{cat(k)} = \{f_{i(ref)}, f_{c(k)}^a, f_{p(k)}\}$ serve as latent codes to modulate the weights of the convolution kernels of a StyleGAN2 generator [6] as shown in Fig. 3.

Network Training. Finally, the feature space modularization and generator are trained jointly by image reconstruction. We directly borrow the same loss functions applied in [7]. 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 with $\mathcal{L}_{L_1}$, and adversarial generative learning $\mathcal{L}_{\mathrm{GAN}}$. The perceptual loss $\mathcal{L}_{\mathrm{vgg}}$ that relies on a pretrained VGG network with N_P layers is also used. All loss functions can refer to [9].

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:

$$\mathcal{L}_{total} = \mathcal{L}_{\mathrm{GAN}} + \lambda_1\mathcal{L}_{L_1} + \lambda_v\mathcal{L}_{\mathrm{vgg}} + \lambda_c\mathcal{L}_c + \lambda_i\mathcal{L}_i, \tag{3}$$

where the λs are balancing coefficients.

3. Experiments

Evaluation Metrics. We use **SSIM** and the cumulative probability blur detection (**CPBD**) to account for the genera-

Table 1: **The quantitative results on LRW and VoxCeleb2.** 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.

	LRW				VoxCeleb2			
Method	SSIM ↑	CPBD ↑	LMD ↓	Sync_{conf} ↑	SSIM ↑	CPBD ↑	LMD ↓	Sync_{conf} ↑
ATVG [4]	0.810	0.102	5.25	4.1	0.826	0.061	**6.49**	4.3
Wav2Lip [8]	**0.862**	0.152	5.73	**6.9**	0.846	0.078	12.26	4.5
MakeitTalk [12]	0.796	0.161	7.13	3.1	0.817	0.068	31.44	2.8
Rhythmic Head† [3]	-	-	-	-	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**

Figure 4: **Qualitative results**.

tion quality. Then we use both Landmarks Distance (**LMD**) around the mouths and the confidence score ($\textbf{Sync}_{conf}$) proposed in SyncNet to account for the accuracy of mouth shapes and lip sync.

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 Voxceleb2 and LRW datasets. On LRW, though Wav2Lip [8] outperforms our method given two metrics, the reason is that their method keeps most parts of the input unchanged while samples in LRW are mostly frontal faces. 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.

The qualitative results are shown in Fig. 4. Please refer to https://hangz-nju-cuhk.github.io/projects/PC-AVS for code, models and demo videos.

References

[1] Volker Blanz, Thomas Vetter, et al. A morphable model for the synthesis of 3d faces. In *SIGGRAPH*, 1999. 1, 3

[2] Egor Burkov, Igor Pasechnik, Artur Grigorev, and Victor Lempitsky. Neural head reenactment with latent pose descriptors. In *CVPR*, 2020. 2

[3] Lele Chen, Guofeng Cui, Celong Liu, Zhong Li, Ziyi Kou, Yi Xu, and Chenliang Xu. Talking-head generation with rhythmic head motion. *ECCV*, 2020. 1, 4

[4] Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *CVPR*, 2019. 4

[5] Joon Son Chung, Amir Jamaludin, and Andrew Zisserman. You said that? In *BMVC*, 2017. 3

[6] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 3

[7] 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. 3

[8] 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 *ACMMM*, 2020. 2, 4

[9] 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 *CVPR*, 2018. 3

[10] 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. 1

[11] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. Talking face generation by adversarially disentangled audio-visual representation. In *AAAI*, 2019. 2, 3

[12] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makeittalk: Speaker-aware talking head animation. *SIGGRAPH ASIA*, 2020. 1, 3, 4