# **OPT: ONE-SHOT POSE-CONTROLLABLE TALKING HEAD GENERATION**

Jin Liu<sup>1,2</sup>, Xi Wang<sup>1\*</sup>, Xiaomeng Fu<sup>1,2</sup>, Yesheng Chai<sup>1</sup>, Cai Yu<sup>1,2</sup>, Jiao Dai<sup>1\*</sup>, Jizhong Han<sup>1</sup>

<sup>1</sup>Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China <sup>2</sup>School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China

## ABSTRACT

One-shot talking head generation produces lip-sync talking heads based on arbitrary audio and one source face. To guarantee the naturalness and realness, recent methods propose to achieve free pose control instead of simply editing mouth areas. However, existing methods do not preserve accurate identity of source face when generating head motions. To solve the identity mismatch problem and achieve high-quality free pose control, we present One-shot Pose-controllable Talking head generation network (OPT). Specifically, the Audio Feature Disentanglement Module separates content features from audios, eliminating the influence of speakerspecific information contained in arbitrary driving audios. Later, the mouth expression feature is extracted from the content feature and source face, during which the landmark loss is designed to enhance the accuracy of facial structure and identity preserving quality. Finally, to achieve free pose control, controllable head pose features from reference videos are fed into the Video Generator along with the expression feature and source face to generate new talking heads. Extensive quantitative and qualitative experimental results verify that OPT generates high-quality pose-controllable talking heads with no identity mismatch problem, outperforming previous SOTA methods.

*Index Terms*— Talking head generation, Generative Model, Audio driven animation

#### 1. INTRODUCTION

Talking head generation aims to drive the source face image with the audio signal and produces a lip-sync talking head video, which is significant to various practical multimedia applications, such as film making, virtual education, video conferencing, digital human animation and short video creation.

Talking head generation can be divided into two categories: speaker-specific methods and speaker-independent methods. The speaker-specific methods [1, 2, 3] only generate talking heads of fixed subject and requires large amount of person-specific high-quality videos, which limits the application and generalization.

The speaker-independent methods are designed to animate video portraits given one unseen source face and driving audio. Some one-shot works [4, 5] simply edit the mouth region and keep the other areas of source face unchanged. Their generated talking head videos are unnatural with the fixed facial contour, blending traces around the mouth and no head motion changes. Therefore, current one-shot speakerindependent works focus on full-frame generation [6, 18, 7], which produce the whole head areas, together with neck parts and background.

To improve the naturalness and realness, some methods propose to add natural head poses into talking heads. PC-AVS [6] modularizes audio-visual representations by devising an implicit low-dimension pose code. Audio2Head [18] utilizes a motion-aware recurrent neural network to predict head motions from audio. However, in talking heads of above methods with new poses, the source identity is not well preserved due to the facial structure change, as shown in Fig. 2.

The identity mismatch problem means the inability of generated talking heads to preserve the identity of source faces. Previous image driven face reenactment works [8] focus on solving the identity mismatch problem in visual modality, which caused by the inconsistent facial contour between driving subject and source person. When it comes to audio-driven paradigm, the gap between audio and visual modality becomes even larger. All the information contained in audio signal affects the driving representation extraction, among which the content feature is the most important since it directly relates to the mouth shape. Given the fact that different speakers' audios with the same content are different, we believe that it is necessary to disentangle identity and content features from audio signal. Furthermore, it is important to extract accurate mouth expression features since the facial structure changes when performing different head poses.

Specifically, we present the One-shot Pose-controllable Talking head generation network (OPT). The *Audio Feature Disentanglement Module* separates identity and content features explicitly from audio signals. Later, the facial expression feature is extracted from content feature and source face, during which the landmark loss is designed to enhance the accuracy of facial structure and identity preserving quality.

<sup>\*</sup>Corresponding authors.

This research is supported in part by the National Key Research and Development Program of China (No. 2020AAA0140000), and the National Natural Science Foundation of China (No. 61702502).

Finally, the head pose feature reconstructed from other pose videos using 3DMM [9] is fed into the Video Generator along with expression feature and source face to generate new talking heads. Extensive experimental results demonstrate the superior performance of OPT and the effectiveness of several modules. Our contributions are summarized as follows:

- The proposed OPT is the first to simultaneously perform one-shot identity-independent pose-controllable talking head generation with almost no *identity mismatch problem*.
- To solve the identity mismatch problem, the *Audio Feature Disentanglement Module* is proposed to successfully decompose intrinsic identity features and content features over audio signals.
- The landmark loss is designed to enhance the accuracy of facial shape and the identity preserving quality. The explicit head pose feature is also utilized to guide the free pose control.

### 2. OUR METHOD

Fig. 1 summarizes the pipeline of our proposed method. OPT takes driving audio  $A_{dri}$ , source image  $I_{src}$  and pose image  $I_{pose}$  as inputs to generate  $I_G$ , indicating  $I_{src}$  speaking the corpus of  $A_{dri}$  with the head pose of  $I_{pose}$ . The Audio Feature Disentanglement Module separates content feature  $F_{con}$  and identity feature  $F_{id}$  from  $A_{dri}$ . Then  $F_{con}$  and source feature  $F_{src}$  from  $I_{src}$  are fed into the Audio-to-Expression Module to produce expression feature  $F_e$ . Finally, the Video Generator takes  $I_{src}$ ,  $F_e$  and head pose feature  $F_{hp}$  from  $I_{pose}$  as inputs to generate  $I_G$ . Each module will be introduced detailedly in the following sections.

#### 2.1. Audio Feature Disentanglement Module

Given that audios of the same content but of different speakers are diverse, we believe the inherently entangled identity and content features need to be independently extracted from audio signals, to achieve audio-based identity control for talking head generation. During inference, the identity of  $A_{dri}$  and  $I_{src}$  are usually different, since the driving audio is chosen arbitrarily. Unlike previous methods [5, 6] who merely extract entangled features from audio signals, we propose the Audio Feature Disentanglement Module (AFDM) to map audios into two separate latent audio spaces: a content-agnostic encoding space of the identity and a content-dependent encoding space of the corpus corresponding to audio.

Specifically, AFDM contains two encoders  $E_{con}$  and  $E_{id}$  to individually extracts corresponding features from  $A_{dri}$ . The content loss  $\mathcal{L}_{con}$  is as follows:

$$\mathcal{L}_{con} = \left\| E_{con} \left( A \right) - E_{con} \left( \tilde{A} \right) \right\|_{1}, \tag{1}$$



**Fig. 1**. Overview of OPT. The Audio Feature Disentanglement Module separates content feature from driving audio. Then the Audio-to-Expression Module extracts the expression feature given conent feature and soure face. Finally, the head pose feature from other video and expression feature are fed into the Video Generator to generate new talking heads.

where audio A and  $\tilde{A}$  shares the same content but spoken by different subjects. The identity loss  $\mathcal{L}_{id}$  is used to train  $E_{id}$ :

$$\mathcal{L}_{cls} = -\sum_{i=1}^{N} \left( p * \log q \right), \tag{2}$$

where N denotes the total amount of speaker identities, p means the ground truth identity probability and q represents the  $E_{id}$  prediction probability. In this way, the pure content feature  $F_{con}$  can be accurately separated from  $A_{dri}$  through AFDM to guide the following extraction process of expression feature.

### 2.2. Audio-to-Expression Module

The Audio-to-Expression Module (AEM) predicts expression feature  $F_e$  from features  $F_{con}$  and  $F_{src}$ . Considering the good reconstruction performance of 3DMM [9], we use part of 3DMM parameters to represent the expression feature. With a 3DMM, the 3D shape **S** of a face is parameterized as follows:

$$\mathbf{S} = \overline{\mathbf{S}} + \alpha \mathbf{B}_{id} + \beta \mathbf{B}_{exp},\tag{3}$$

where  $\overline{\mathbf{S}}$  is the average face shape,  $\mathbf{B}_{id}$  and  $\mathbf{B}_{exp}$  are basis of identity and expression via PCA. The coefficients  $\alpha \in \mathbb{R}^{80}$ and  $\beta \in \mathbb{R}^{64}$  describe the facial shape and expression respectively. In our method, we choose  $\beta$  as expression features  $f_e$ . The AEM contains multiple convolutional layers and linear layers. The L1 loss  $\mathcal{L}_{L1}$  is imposed on  $\tilde{F}_e$  and the ground truth  $F_e$ . In order to further improve the lip-sync quality and keep the accurate facial contour structure to alleviate the identity mismatch problem, we further design the facial landmark loss  $\mathcal{L}_{ldmk}$  using 3DMM:

$$\mathcal{L}_{ldmk} = \frac{1}{N} \sum_{n=1}^{N} \omega_n \left\| \tilde{\mathbf{l}}_n - \mathbf{l}_n \right\|^2$$
(4)

where  $\{l_n\}$  is facial landmarks of ground truth driving face image,  $\{\tilde{l}_n\}$  is the 3D landmark vertices projection of reconstructed shape using  $f_e$  onto the image plane. N denotes the number of landmarks. The weight  $\omega_n$  id set to 20 for inner mouth and nose points and others to 1.

## 2.3. Video Generator

To achieve pose-controllable talking head generation, the additional driving pose face image  $I_{pose}$  is need to provide head pose feature. During inference,  $I_{pose}$  could come from real talking head videos to offer auxiliary pose feature sequences. The head pose feature  $F_{hp}$  is denoted by rotation  $R \in SO(3)$ and translation  $T \in \mathbb{R}^3$ , which could also be obtained during the process of 3DMM reconstruction [9]. Finally, given  $I_{src}$ ,  $F_{hp}$  and  $F_e$ , the Video Generator(VG) produces new talking head $I_G$ . Detailedly, VG contains Conv and TransConv layers with residual connection. The two features are fed by AdaIn [10] opeation. The discriminator utilized Patch-GAN [11] and VG minimizes the reconstruction loss  $\mathcal{L}_{rec}$ between ground truth image I and generated head image  $I_G$ .

$$\mathcal{L}_{rec} = \|I - I_G\|_2. \tag{5}$$

The input to the discriminator D is the ground truth image I and  $I_G$ . The GAN loss is as follows:

$$\mathcal{L}_{GAN} = \log D(I) + \log(1 - D(I_G)). \tag{6}$$

Furthermore, the perceptual loss  $\mathcal{L}_{per}$  is also adapted to calculate the distance between activation maps of the pre-trained VGG-19 network :

$$\mathcal{L}_{per} = \sum_{i} \left\| \phi_i(\tilde{I}_g) - \phi_i(I) \right\|_1, \tag{7}$$

where  $\phi_i$  is the activation map of the i-th layer of the VGG-19 network [12]. During training, each module in OPT are trained separately utilizing corresponding loss combination.

#### 3. EXPERIMENTS

### **3.1.** Experimental settings

**Datasets** The AFDM is trained on audio-visual dataset MEAD [13] which contains annotations of speaking corpus scripts and identity information. Other modules of OPT are trained and tested on LRW [14] and LRS2 [15] datasets. The LRW dataset contains over 1000 utterances of 500 different words while the LRS2 dataset includes over 140,000 utterances of different sentences. Both videos are from BBC television in the wild.

**Implementation Details** Face frames are cropped to  $256 \times 256$  size at 25 FPS and audio to mel-spectrogram of size  $16 \times 16$  per frame. Mel-spectrograms are constructed from 16kHZ audio, window size 800, and hop size 200. Both encoders in the AFDM share the SE-ResNet [16] architecture.  $F_{src}$  is extracted by pre-trained ArcFace [17] model. OPT is trained in stages on 4 Tesla 32G V100 GPUs. The ADAM optimizer



Fig. 2. Qualitative comparison results with other state-of-theart methods on LRS2 dataset.

Methods	SSIM $\uparrow$	$\text{CSIM} \uparrow$	$LMD\downarrow$	LSE-C $\uparrow$
ATVG	0.781	0.76	5.32	4.165
Wav2Lip	0.792	0.81	5.73	7.237
Audio2Head	0.743	0.72	7.34	2.135
PC-AVS	<u>0.815</u>	0.74	6.14	6.420
Ours-Fix Pose	0.795	<u>0.83</u>	<u>5.25</u>	6.432
OPT (Ours)	0.823	0.88	3.78	<u>6.619</u>

**Table 1.** Quantitative comparison results on LRW dataset.The **bold** and <u>underlined</u> indicate the top-2 results.

is adopted with an initial leaning rage as  $10^{-4}$ . The learning rate is decreased to  $2 \times 10^{-5}$  after 300k iterations.

Comparing Methods The following SOTA one shot talking head generation methods are compared. ATVG [4] generates frames based on facial landmarks using the attention mechanism. Wav2Lip [5] utilizes a pre-trained lip-sync discriminator to focus on editing the mouth shape. Audio2Head [18] infers unique head pose sequences from audio and utilizes flow-based generator to produce talking heads. PC-AVS [6] extracts modularized audio-visual representations of identity, pose and speech content, generating pose-controllable talking heads. Besides, when  $I_{pose}$  is not given for OPT, we can fix it with the same pose as  $I_{src}$ , keeping the head still. We refer results under this setting as **Ours-Fix Pose**. Same as the generation paradigm in PC-AVS [6], an extra video with supposedly the same pose as driving frames corresponding to  $I_{dri}$  but different identities and mouth shapes is generated to serve as the  $I_{pose}$  in our method.



**Fig. 3**. Qualitative results of OPT. Four generated clips of the same word "we" driven by different pose videos are shown.

Methods	SSIM $\uparrow$	$\mathbf{CSIM}\uparrow$	$LMD\downarrow$	LSE-C $\uparrow$
Baseline	0.721	0.69	7.18	3.125
Ours w/o AFDM	0.753	0.75	5.78	4.259
Ours w/o $\mathcal{L}_{ldmk}$	0.762	0.74	6.14	3.842
Ours	0.823	0.88	3.78	6.619

**Table 2.** Ablation study on LRW dataset. The evaluated parts include the *Audio Feature Disentanglement Module* and the landmark loss utilized in the Audio-to-Expression Module.

## **3.2.** Quantitative Results

We evaluate the performance on image quality, identity preserving and lip-sync quality. The SSIM [19] scores are utilized to judge the talking head image quality. CSIM indicates the cosine similarity between face recognition features [17] of generated and ground truth talking heads. For the lip-sync quality, the Landmark Distance(LMD) and Lip-Sync Error-Confidence(LSE-C) [5] are applied.

Table 1 shows the quantitative comparison results using LRW datasets. It shows that OPT achieves leading SSIM, LMD and CSIM scores. As mentioned in [6], the leading LSE-C score only means that Wav2Lip is comparable to the ground truth, not better. Our leading LMD score indicates that we preserve better facial structure and accurate mouth shape. We also achieve high-level identity preserving quality, proved by the leading CSIM score.

## 3.3. Qualitative Results

We compare OPT with other methods, as displayed in Fig. 2. It shows that OPT generates high quality talking head videos with accurate mouth shape and facial contour that best match the ground truth. Concretely, ATVG merely focuses on cropped facial region. Wav2Lip generates faces with fixed head motions and blurry mouth shape. Audio2Head fails to keep the lip synchronization and accurate facial contour. PC-

Method	Visual	Lip-Sync	Identity
ATVG	2.39	3.69	2.81
Wav2Lip	3.57	4.10	3.92
Audio2Head	3.31	2.45	3.63
PC-AVS	3.92	3.87	3.14
Ours	4.16	4.31	4.27

Table 3. User study results by mean opinion scores.

AVS causes the identity mismatch problem and cannot preserve the accurate facial contour. We further demonstrate the performance of OPT in Fig. 3. It indicates that the generated talking heads can achieve free pose control and meanwhile maintain accurate lip synchronization and no identity mismatch problem.

### 3.4. Ablation Study

Table 2 shows our ablation study to prove the effect of each proposed component. The baseline method directly extracts expression feature from audio signal and utilizes merely reconstruction loss in the Audio Feature Disentanglement Module and Audio-to-Expression Module. The result indicates that AFDM helps solve the identity mismatch problem according to the obvious increase on CSIM score. Besides,  $\mathcal{L}_{ldmk}$  contributes a lot to the image quality and lip-sync quality since it extracts accurate facial structure representation during the training process.

## 3.5. User Study

We further conduct the user study to evaluate OPT and other state-of-the-art methods. For OPT and comparing methods, 15 video clips using randomly selected source faces and audios from LRW and LRS2 datasets are generated. We adopt the widely used Mean Opinion Scores (MOS) rating protocol. 10 participants are required to give their ratings (1-5) on the following three aspects for each generated talking head video: visual quality, lip-sync quality and identity preserving quality. As Table 3 shows, OPT achieves the best results on all aspects, especially the identity preserving quality.

### 4. CONCLUSION

In this paper, we propose a new method called OPT to generate pose-controllable identity-preserving talking head videos. Given one source face image and arbitrary driving audio, OPT generates lip-sync talking heads that preserve the source identity and can achieve pose control guided by auxiliary pose video. The proposed Audio Feature Disentanglement Module separates content features from audio signal and the landmark loss is adopted in the Audio-to-Expression Module, both contributing a lot to the generation of talking heads. In the future, we will focus on real time and high resolution generation to enhance the generalization.

### 5. REFERENCES

- [1] Chenxu Zhang, Yifan Zhao, Yifei Huang, Ming Zeng, Saifeng Ni, Madhukar Budagavi, and Xiaohu Guo, "Facial: Synthesizing dynamic talking face with implicit attribute learning," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 3867–3876.
- [2] Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang, "Ad-nerf: Audio driven neural radiance fields for talking head synthesis," in *Proceedings of the IEEE/CVF International Conference* on Computer Vision, 2021, pp. 5784–5794.
- [3] Yuhan Zhang, Weihua He, Minglei Li, Kun Tian, Ziyang Zhang, Jie Cheng, Yaoyuan Wang, and Jianxing Liao, "Meta talk: Learning to data-efficiently generate audiodriven lip-synchronized talking face with high definition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*). IEEE, 2022, pp. 4848–4852.
- [4] 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/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 7832–7841.
- [5] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV 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*, 2020, pp. 484–492.
- [6] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu, "Pose-controllable talking face generation by implicitly modularized audiovisual representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4176–4186.
- [7] Borong Liang, Yan Pan, Zhizhi Guo, Hang Zhou, Zhibin Hong, Xiaoguang Han, Junyu Han, Jingtuo Liu, Errui Ding, and Jingdong Wang, "Expressive talking head generation with granular audio-visual control," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3387–3396.
- [8] Jin Liu, Peng Chen, Tao Liang, Zhaoxing Li, Cai Yu, Shuqiao Zou, Jiao Dai, and Jizhong Han, "Li-net: Large-pose identity-preserving face reenactment network," in 2021 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2021, pp. 1–6.
- [9] 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 the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

- [10] Xun Huang and Serge Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1501–1510.
- [11] 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*, 2017, pp. 1125–1134.
- [12] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.
- [13] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy, "Mead: A large-scale audio-visual dataset for emotional talking-face generation," in *European Conference on Computer Vision*. Springer, 2020, pp. 700–717.
- [14] Joon Son Chung and Andrew Zisserman, "Lip reading in the wild," in Asian conference on computer vision. Springer, 2016, pp. 87–103.
- [15] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman, "Deep audiovisual speech recognition," *IEEE transactions on pattern analysis and machine intelligence*, 2018.
- [16] Jie Hu, Li Shen, and Gang Sun, "Squeeze-andexcitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.
- [17] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.
- [18] Suzhen Wang, Lincheng Li, Yu Ding, Changjie Fan, and Xin Yu, "Audio2head: Audio-driven one-shot talkinghead generation with natural head motion," arXiv preprint arXiv:2107.09293, 2021.
- [19] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.