Audio2Head: Audio-driven One-shot Talking-head Generation with Natural Head Motion

Suzhen Wang[1] , **Lincheng Li**[1] , **Yu Ding**[1*] , **Changjie Fan**[1] , **Xin Yu**[2]

[1] Virtual Human Group, Netease Fuxi AI Lab, China
[2]University of Technology Sydney
{wangsuzhen, lilincheng, dingyu01, fanchangjie}@corp.netease.com, xin.yu@uts.edu.au

Abstract

We propose an audio-driven talking-head method to generate photo-realistic talking-head videos from a single reference image. In this work, we tackle two key challenges: *(i)* producing natural head motions that match speech prosody, and *(ii)* maintaining the appearance of a speaker in a large head motion while stabilizing the non-face regions. We first design a head pose predictor by modeling rigid 6D head movements with a motion-aware recurrent neural network (RNN). In this way, the predicted head poses act as the low-frequency holistic movements of a talking head, thus allowing our latter network to focus on detailed facial movement generation. To depict the entire image motions arising from audio, we exploit a keypoint based dense motion field representation. Then, we develop a motion field generator to produce the dense motion fields from input audio, head poses, and a reference image. As this keypoint based representation models the motions of facial regions, head, and backgrounds integrally, our method can better constrain the spatial and temporal consistency of the generated videos. Finally, an image generation network is employed to render photo-realistic talking-head videos from the estimated keypoint based motion fields and the input reference image. Extensive experiments demonstrate that our method produces videos with plausible head motions, synchronized facial expressions, and stable backgrounds and outperforms the state-of-the-art.

1 Introduction

Delivering information in an audio-visual manner is more attractive to humans compared to an audio-only fashion. Given an audio clip and one image of an arbitrary speaker, authentic audio-visual content creation has received great attention recently and also has widespread applications, such as human-machine interaction and virtual reality. A large number of one-shot talking-head works [Chen *et al.*, 2019; Zhou *et al.*, 2019; Song *et al.*, 2019; Vougioukas *et al.*, 2019; Prajwal *et al.*, 2020; Zhu *et al.*, 2020] have been proposed to synchronize audio and lip movements. However, most of them neglect to infer head motions from audio, and a still head pose is less satisfactory for human observation. Natural and rhythmic head motions are also one of the key factors in generating authentic talking-head videos [Ding *et al.*, 2013; Chen *et al.*, 2020].

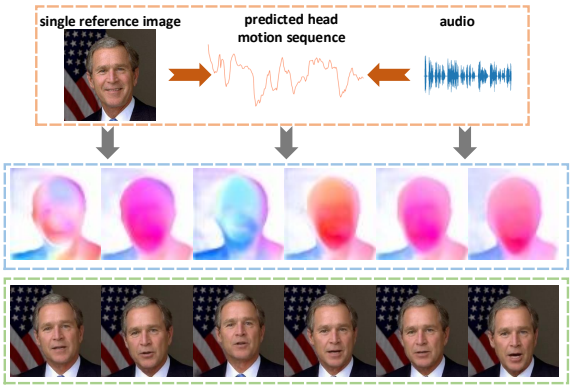

Figure 1: Illustration of the proposed audio-driven single image based talking-head video generation method. First row: the input reference image and audio, and the predicted head pose; middle row: generated motion fields from the audio and image; bottom row: synthesized talking-head frames.

Albeit recent works [Chen *et al.*, 2020; Zhou *et al.*, 2020] take head movements into consideration, they often suffer from the ambiguous correspondences between head motions and audio in the training dataset, and fail to produce realistic head movements. For example, given the same audio content, one performer may move the head from left to right while another may move the head from right to left. These examples would introduce ambiguity in training and thus a network will produce a still-alike head to minimize the head motion loss. Moreover, when generating head motions, reducing artifacts in background regions is critical as well. However, the methods [Chen *et al.*, 2020; Zhou *et al.*, 2020] only focus on the face regions and thus result in obvious distortions in non-face regions (*e.g.*, hair and background). Therefore, how to produce realistic head motions and artifact-free frames for one-shot talking-head generation is still challenging.

*Corresponding author.

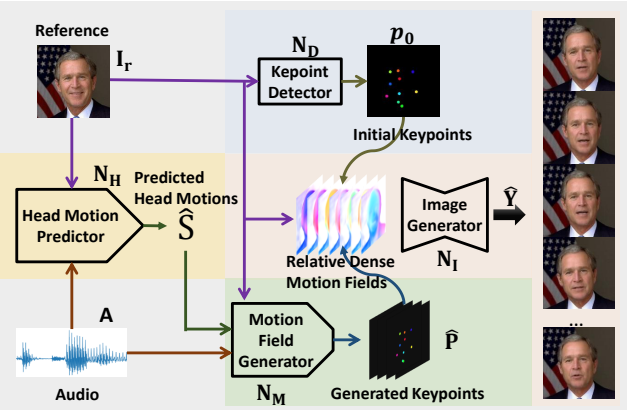

Figure 2: Pipeline of the proposed framework.

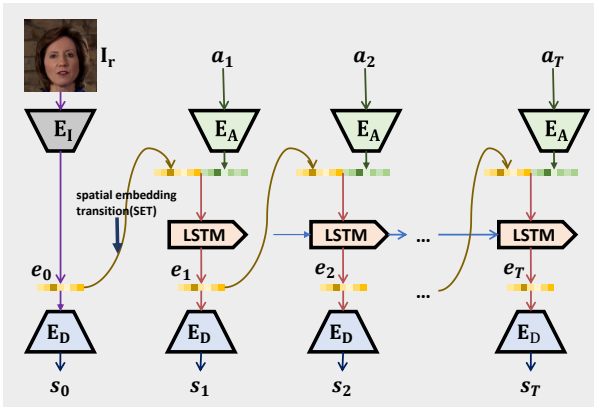

Figure 3: Architecture of the head motion predictor.

In this paper, a novel one-shot talking-head video generation framework is proposed to produce natural and rhythmic head motions while remarkably suppressing background artifacts caused by head motions. Firstly, we propose to disentangle head motions and expression changes since head motions might exhibit ambiguity as aforementioned. The head motion is modeled as a rigid 6 degrees of freedom (DoF) movement and it will act as the low-frequency holistic movement of talking head. We generate head motions via a motion aware recurrent neural network (RNN) (see Figure 1) and let another network focus on producing detailed facial movements.

Considering the ambiguous correspondences between head motions and audio, we opt to enforce the structural similarity between the generated and the ground truth head motion matrices making up of successive 6D motion vectors, thus achieving more diverse head motions.

After obtaining head motions, we resort to a keypoint based dense motion field representation [Siarohin *et al.*, 2019b] to depict the entire image content motions, including facial region, head and background movements. We develop an image motion field generator to produce keypoint based dense motion fields from the input audio, head poses and reference image, which will be used to synthesize new frames. The relative dense motion field is then described by the differences between the predicted keypoints and those of the reference image (see Figure 2). Compared to the 3D face model or landmark-based representations that are used in prior works, the keypoint based representation models the motions of the facial regions, head, and backgrounds integrally, allowing for better governing the spatial and temporal consistency in the generated videos. Finally, an image generation network [Yu *et al.*, 2018; Yu and Porikli, 2018; Yu *et al.*, 2019a; Yu *et al.*, 2019b] is employed to render photo-realistic talking-head videos from the estimated keypoint based motion fields and the input reference image.

Extensive comparisons with the state-of-the-art methods show that our method achieves superior visual quality and authentic head motions without introducing noticeable artifacts. In summary, we make the following technical contributions:

- We develop a new audio-driven talking-head generation framework that produces photo-realistic videos with natural head motions from a single image.
- We design a motion aware recurrent neural network to predict natural-looking head motions that match the input audio rhythm.
- We present an image motion field generator to produce keypoint based dense motion fields and thus are able to govern the spatial and temporal consistency of the generated videos.
- We achieve state-of-the-art results in terms of visual quality and rhythmic head motions.

2 Related Work

2.1 Speech-driven Talking-head Generation

Given the input audio, some approaches achieve great success to synthesize the talking head of a specific speaker [Suwajanakorn *et al.*, 2017; Fried *et al.*, 2019; Thies *et al.*, 2020; Li *et al.*, 2021]. However, they need to be trained on minutes to hours of videos of each speaker. Other works aim to reduce the speaker's reference information to a single image. Song *et al.* [2019] and Vougioukas *et al.* [2019] take the temporal dependency into account and generate the talking face video with GANs. Zhou *et al.* [2019] design an end-to-end framework to generate videos from the disentangled audio-visual representation. Prajwal *et al.* [2020] employ a carefully-designed lip-sync discriminator for better lip-sync accuracy. These works directly learn the mapping from audio to facial pixels, which tend to produce blurry results. For better facial details, some works utilize intermediate representations to bridge the variation of audio and pixels. Chen *et al.* [2020] and Zhang *et al.* [2021] predict the coefficients of a 3D face model from audio, which is used to guide the image generation. Chen *et al.* [2019] and Zhou *et al.* [2020] learn to generate facial landmarks from audio first, and then synthesize images from landmarks. Although these methods improve the visual quality of the inner face, none of the mediums models the non-face regions (e.g. hair and background). Hence, the texture and the temporal consistency of the outer face regions are not as good as that of the inner face.

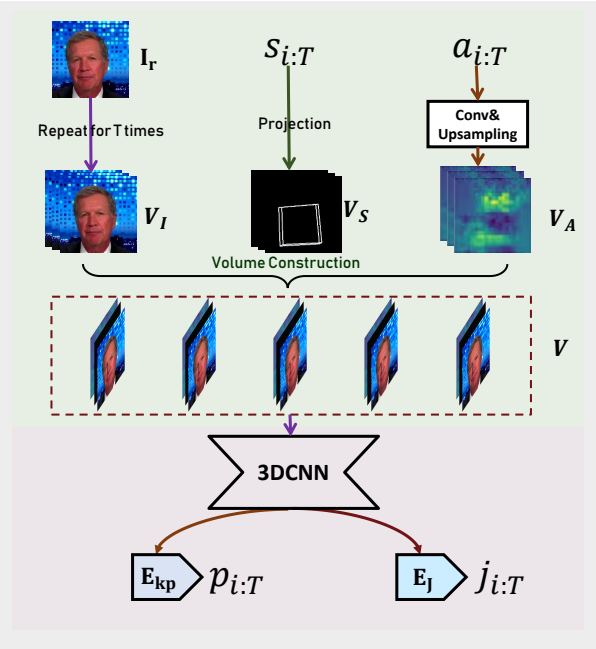

Figure 4: Architecture of the motion field generator.

Figure 5: Comparison with the state-of-the-art. Please see more dynamic demos in our supplementary materials.

2.2 Video-driven Talking-head Generation

Video-driven methods control the motions of a subject with a driving video. Subject-specific methods [Bansal *et al.*, 2018; Kim *et al.*, 2018] focus on a specific person by training their models on that person. Subject-independent approaches drive the reenactment using landmarks [Zhang *et al.*, 2020; Ha *et al.*, 2020; Nirkin *et al.*, 2019], latent embeddings [Burkov *et al.*, 2020] or feature warping [Wang *et al.*, 2019]. Recently, Siarohin *et al.* [2019b] represent the dense motion flow of the driving video as self-learned keypoints, and use the flow to warp the features of the reference image for reenactment. High visual quality is obtained due to the dense intermediate representation of the entire image.

2.3 Head Motion Prediction

Traditional head motion prediction methods are designed for 3D avatars, which only involve the 3D head rotation [Ding *et al.*, 2013; Greenwood *et al.*, 2017; Sadoughi and Busso, 2018]. Methods designed for 2D images, on the other hand, need to produce both head rotation and translation. Chen *et al.* [2020] constrain the mean value and the standard deviation of the predicted head pose sequence to be similar to that of the real sequence. However, the statistical constraints cannot model the local motion details. Zhou *et al.* [2020] constrain the predicted facial landmarks to be consistent with that of the ground truth. Due to the L2 loss term, their method suffers from the ambiguous correspondences between head motion and audio, and converge to the slightly swinging. Different from the above two works, our method constrains the full head motion sequence for more natural head motion patterns.

3 Proposed Method

3.1 Overview

Our method takes a reference image $\mathbf{I_r}$ and an audio clip $\mathbf{A}$ as input, and synthesizes video frames $\hat{\mathbf{Y}} = \hat{y}_{1:T}$ of the reference speaker synchronized with $\mathbf{A}$. As illustrated in Figure 2, the pipeline of our method consists of four components.

Head Motion Predictor $\mathbf{N_H}$. As the representation of low-frequency holistic movements, the head poses are predicted individually. From both $\mathbf{I_r}$ and $\mathbf{A}$, $\mathbf{N_H}$ produces the natural-looking and rhythmic head motion sequence $\hat{\mathbf{S}} = \hat{s}_{1:T}$.

Motion Field Generator $\mathbf{N_M}$. $\mathbf{N_M}$ produces the self-learned keypoint sequence $\hat{\mathbf{P}} = \hat{p}_{1:T}$ that controls the dense motion field. $\hat{\mathbf{P}}$ contains the information of synchronized facial expressions, head motions, and non-face region motions.

Keypoint Detector $\mathbf{N_D}$ and Image Generator $\mathbf{N_I}$. $\mathbf{N_D}$ detects the initial keypoints p_0 from $\mathbf{I_r}$. $\mathbf{N_I}$ renders the synthesized images from the relative dense motion between $\hat{\mathbf{P}}$ and p_0. The network architectures of $\mathbf{N_D}$ and $\mathbf{N_I}$ are adopted from [Siarohin *et al.*, 2019b].

3.2 Head Motion Predictor

As preprocessing, the input raw audio $\mathbf{A}$ is first converted to an acoustic feature sequence $a_{1:T}$. a_i refers to an acoustic feature frame. To be consistent with the frequency of the videos sampled at 25 fps, $a_i \in \mathbb{R}^{4\times 41}$ makes up of acoustic features from 4 successive sliding windows. In total, 41 acoustic features are extracted from each sliding window, including 13 Mel Frequency Cepstrum Coefficients (MFCC), 26 Mel-filterbank energy features (FBANK), pitch and voiceless. The sliding window has a window size of 25ms and a step size of 10ms.

		PSNR ↑	SSIM ↑	FID↓
LRW	Prajwal *et al*. [2020]	**20.17**	0.62	**42.41**
	Zhou *et al*. [2020]	19.20	0.59	48.13
	ours	19.53	**0.63**	42.55
GRID	Prajwal *et al*. [2020]	28.82	0.86	31.14
	Zhou *et al*. [2020]	24.55	0.78	34.57
	ours	**30.93**	**0.91**	**22.50**
VoxCeleb	Prajwal *et al*. [2020]	17.33	0.50	61.58
	Zhou *et al*. [2020]	17.45	0.50	53.95
	ours	**21.19**	**0.68**	**39.86**

Table 1: Quantitative comparison with the state-of-the-art.

In our work, $a_{1:T}$ is used to predict a head pose sequence describing head rotation and translation in the camera coordinate system. For reference images with different head scales and body positions, the natural head trajectories in pixel coordinates are also different. Specially, we employ an encoder $\mathbf{E_I}$ (ResNet-34) to extract the initial head and body state from $\mathbf{I_r}$. The extracted spatial embedding e_0 encodes the initial head rotation and translation. Afterwards, we use a two-layer LSTM to create natural head motion sequence that matches audio rhythm, as shown in Figure 3. At each time step i, we first extract the audio embedding with another ResNet-34 encoder $\mathbf{E_A}$ from a_i and concatenate it with the spacial embedding e_{i-1} at step $i-1$. Then, the LSTM takes the concatenated embeddings as input and outputs the current e_i. Such spatial embedding transition (SET) passes the previous e_{i-1} to the next time step, and therefore contribute to more natural head motions and better synchronization with audio. Finally, a decoder $\mathbf{E_D}$ is used to decode e_i to head pose $\hat{s}_i \in \mathbb{R}^6$ (3 for rotation and 3 for translation). Our head motion predictor supports an arbitrary length of audio input. The procedure is formulated as:

$$(h_i, e_i) = \mathbf{LSTM}(h_{i-1}, \mathbf{E_A}(a_i) \oplus e_{i-1}), \quad (1)$$

$$\hat{s_i} = \mathbf{E_D}(e_i), \quad (2)$$

where h_i is the hidden state of step i, $\oplus$ means concatenation.

Since the mapping from audio to head motion is one-to-many mapping, the widely-used L1 loss and L2 loss are not suitable choices. Instead, we treat $\hat{s}_{0:T} \in \mathbb{R}^{6\times T}$ as an image of size $6 \times T$, and impose the structural constraint on it using the Structural Similarity (SSIM)[Wang *et al.*, 2004] loss:

$$\mathcal{L}_{SSIM} = 1 - \frac{(2\mu\hat{\mu} + C_1)(2cov + C_2))}{(\mu^2 + \hat{\mu}^2 + C_1)(\sigma^2 + \hat{\sigma}^2 + C_2))}. \quad (3)$$

$\hat{\mu}$ and $\hat{\sigma}$ are the mean and standard deviation of $\hat{s}_{0:T}$, and μ and σ are that of the groundtruth head pose sequence extracted by OpenFace [Baltrusaitis *et al.*, 2018]. cov is the covariance. C_1 and C_2 are two small constants. To improve the fidelity and smoothness of the predicted head motions, we also employ a discriminator $\mathbf{D}$ based on the PatchGAN. Specifically, we adapt the original PatchGAN to perform 1D convolution operations on head motion sequence along temporal trunks. The total loss function for $\mathbf{N_H}$ is defined by:

$$\mathcal{L}_{\mathbf{N_H}} = arg \min_{\mathbf{N_H}} \max_{\mathbf{D}} (\mathbf{N_H}, \mathbf{D}) + \mathcal{L}_{SSIM}. \quad (4)$$

The GAN loss is calculated by LSGAN. When training $\mathbf{N_H}$, we set the window length T to 256.

3.3 Motion Field Generator

From the $\mathbf{I_r}$, $a_{1:T}$ and the predicted $s_{1:T}$, the motion field generator produces the keypoints sequence that controls the dense motion field. For the i-th frame, the keypoints include N positions $\hat{p}_i \in \mathbb{R}^{N\times 2}$, and the corresponding Jacobian $\hat{j}_i \in \mathbb{R}^{N\times 2\times 2}$. Each $(\hat{p}_i^k, \hat{j}_i^k)$ pair represents a local image affine transform. The N local affine transforms with adaptive masks constitute the dense motion field [Siarohin *et al.*, 2019b].

Figure 4 shows the structure of $\mathbf{N_M}$. The multimodal input $s_{1:T}$, $\mathbf{I_r}$ and $a_{1:T}$ are first converted to a unified structure. $\mathbf{I_r}$ is expected to provide the identity constraint for generated keypoints, we downsample $\mathbf{I_r}$ to a size of $[W, H]$ and repeat it T times as the tensor V_I of size $[W, H, 3, T]$. Instead of directly taking $s_{1:T}$ as input, we draw a 3D box in the camera coordinate system to represent the head pose and project it to the image. In this way, we render a binary image for each pose frame, and stack them to get the pose tensor V_S of size $[W, H, 1, T]$. As to $a_{1:T}$, we encode each a_i to the feature map of the shape of $[W, H, 2]$ using an encoder composed with conv and upsampling operations. The feature maps are also stacked as the tensor V_A with size $[W, H, 2, T]$. Finally, we construct V by concatenating V_I, V_S, V_A along the channel dimension. In our experiments, W and H are set to 64, T is set to 64. Afterwards, we employ the 3D Hourglass Network (Hourglass-3D) to deal with the temporal dependence and ensure the smoothness of motion field between consecutive frames. The Hourglass-3D takes V as input and outputs a group of continuous latent features, which are then decoded into $\hat{p}_{1:T}$ and $\hat{j}_{1:T}$ by two decoders $\mathbf{E_{kp}}$ and $\mathbf{E_J}$ separately.

The training process of $\mathbf{N_M}$ goes through two stages. In the first stage, we use the pretrained $\mathbf{N_D}$ as guidance. The loss function is defined as:

$$\mathcal{L}_{\mathbf{N_M^1}} = \frac{1}{T}\sum_{i=1}^{T}(\lambda_m \mathbf{L_1}(\hat{m}_i, m_i) + \lambda_p \mathbf{L_1}(\hat{p}_i, p_i) + \\ + \lambda_j \mathbf{L_1}(\hat{j}_i, j_i)). \quad (5)$$

$\mathbf{L_1}(\cdot,\cdot)$ denotes the L1 loss. p_i and j_i are the positions and Jacobians extracted by the pretrained $\mathbf{N_D}$ from the training video. $\hat{m}_i$ and m_i are heatmaps produced by $\mathbf{E_{kp}}$ and $\mathbf{N_D}$, the keypoint positions are estimated from these heatmaps as in [Siarohin *et al.*, 2019a]. The heatmap term helps the convergence in the beginning. λ_p and λ_j are both set to 10. λ_m is set to 1 and decays to zero after a certain time.

In the second stage, we import the pretrained $\mathbf{N_I}$ to help the fine-tuning of $\mathbf{N_M}$. Giving the predicted $\hat{p}_i$, $\mathbf{N_I}$ renders the reconstructed frame $\hat{I}_i$ for the reconstruction loss:

$$\mathcal{L}_{rec}(\hat{I}, I) = \sum_{i=1}^{l} \mathbf{L_1}(C_i(\hat{I}), C_i(I)), \quad (6)$$

where $C_i(\cdot)$ is the ith channel feature of a specific pretrained VGG-16 layer with l channels. We apply $\mathcal{L}_{rec}(\hat{I}, I)$ on the image pyramid of multiple resolutions, and sum them as $\mathcal{L}_{rec}^{mul}(\hat{I}, I)$. For more stable performance, we also adopt the equivariance constraint losses of [Siarohin *et al.*, 2019b], denoted as $\mathcal{L}_{eq}^{P}(\hat{p})$ and $\mathcal{L}_{eq}^{J}(\hat{j})$ respectively. Please refer to

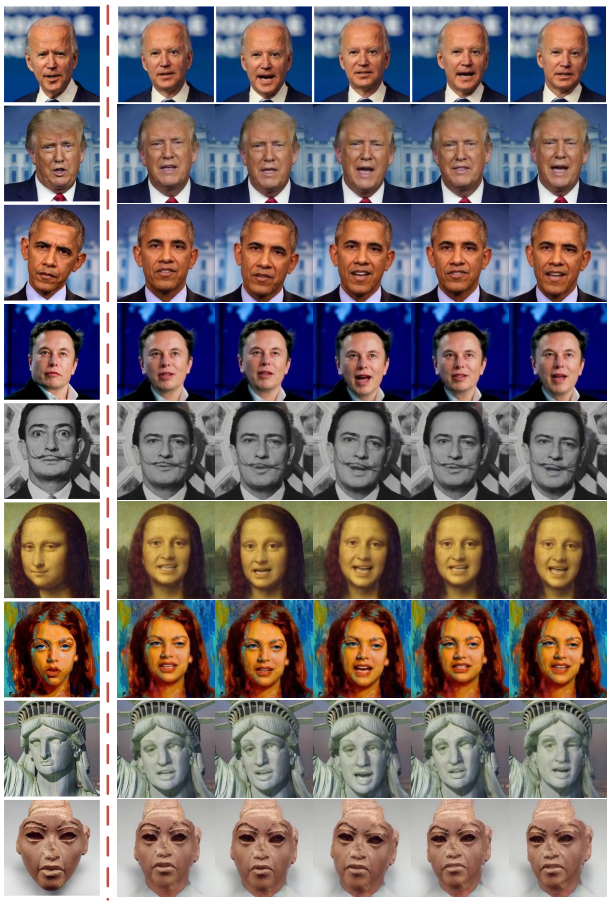

Figure 6: Samples generated with the same audio. Please zoom in for more details.

[Siarohin *et al.*, 2019b] for more details. The total loss function is defined as:

$$\mathcal{L}_{\mathbf{N_M^2}} = \frac{1}{T}\sum_{i=1}^{T}(\lambda'_p \mathbf{L_1}(\hat{p}_i, p_i) + \lambda_{rec}\mathcal{L}_{rec}^{mul}(\hat{I}_i, I_i) + \\ + \lambda_{eq}^{P}\mathcal{L}_{eq}^{P}(\hat{p}_i) + \lambda_{eq}^{J}\mathcal{L}_{eq}^{J}(\hat{j}_i). \quad (7)$$

λ'_p is set to 100, λ_{rec}, λ_{eq}^{P} and λ_{eq}^{J} are all set to 10.

4 Experiment Setup

4.1 Datasets

We use prevalent benchmark datasets **VoxCeleb** [Nagrani *et al.*, 2017], **GRID** [Cooke *et al.*, 2006] and **LRW** [Chung and Zisserman, 2016] to evaluate the proposed method. **VoxCeleb** consists of speech clips collected from YouTube. **GRID** contains video clips of 33 speakers in the experimental condition. **LRW** contains 500 different words spoken by hundreds of people in the wild. Specially, for **VoxCeleb** and **GRID**, we re-crop and resize the original videos to 256×256 as in [Siarohin *et al.*, 2019b], to get 54354 and 25788 shot video clips respectively. For **LRW**, we use its original format by aligning the face in the middle of each frame. We split each dataset into training and testing sets following the setting of previous works. All videos are sampled at 25 fps.

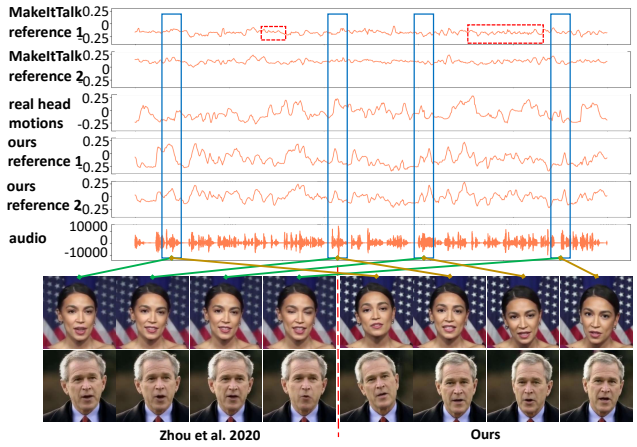

Figure 7: Comparison of head motion patterns on two reference images. Frames in the bottom are sampled from the blue boxes.

4.2 Implementation Details

All our networks are implemented using PyTorch. We adopt Adam optimizer during training, with an initial learning rate of 2e-4 and weight decay to 2e-6. $\mathbf{N_H}$ is trained on VoxCeleb for one day on one RTX 2080 Ti with batchsize 64. When training $\mathbf{N_D}$, $\mathbf{N_I}$ and $\mathbf{N_M}$, we separate **LRW** and the other two datasets into two groups due to the different cropping strategies. For each group, We first follow the training details of [Siarohin *et al.*, 2019b] to train $\mathbf{N_D}$ and $\mathbf{N_I}$, then we train $\mathbf{N_M}$ with frozen $\mathbf{N_D}$ and $\mathbf{N_I}$. The training of $\mathbf{N_D}$ and $\mathbf{N_I}$ takes 3 days with batchsize 28, and that of $\mathbf{N_M}$ takes one week with batchsize 4 on 4 RTX 2080 Ti. Specially, for **LRW**, the window length is set to 32 because of the short video clips.

5 Experiments Results

5.1 Evaluation of Visual Quality

We show the comparisons with the state-of-the-art methods in Figure 5, including Vougioukas *et al.* [2019], Chen *et al.* [2019], Prajwal *et al.* [2020] and Zhou *et al.* [2020]. The samples are generated with the same reference image and audio. Instead of only synthesizing a fixed head or cropped face, our method creates more realistic videos with head motions and full background. Compared to Zhou *et al.* [2020], our results produce more plausible head movements and a more stable background with fewer artifacts. Besides, our method holds the identity of the speaker well even after a large pose change, owing to the motion field representation. The last row of Figure 5 shows the video-driven result of Siarohin *et al.* [2019b]. Although our results are generated from audio instead of video, we show comparable results in visual quality. We present more results in Figure 6 for unseen identities, including non-realistic paintings and human-like statues. The results show the excellent generalization ability of our method.

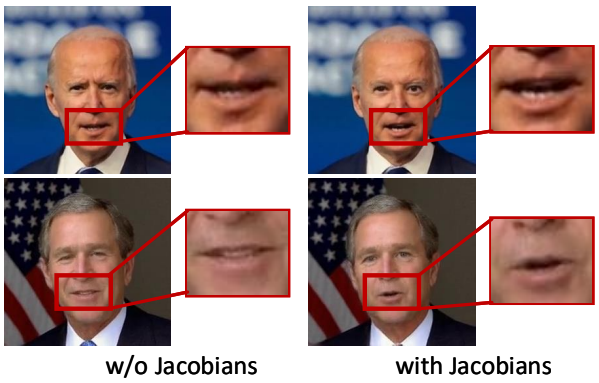

Figure 8: Results with and w/o Jacobians.

The quality of generated videos is evaluated using common reconstruction metrics SSIM, PSNR and Frechet Inception Distance (FID). We compare our approach with recent state-of-the-art methods including Zhou *et al*. [2020], and Prajwal *et al*. [2020], which also synthesize images with background rather than cropped face only. The quantitative results are shown in Table 1. Comparing with Zhou *et al*. [2020] and Prajwal *et al*. [2020], we achieve the highest PSNR/SSIM and the lowest FID score on **GRID** and **VoxCeleb**. Since videos of **LRW** are very short (about 1.16s), most of their head pose and background barely move. Prajwal *et al*. [2020] only edits the mouth region of the reference image, resulting in the best PSNR and FID in **LRW**.

5.2 Evaluation of Lip-sync

We employ SyncNet [Chung and Zisserman, 2016] to evaluate the audio-visual synchronization of the proposed method. The metric values (confidence/offset) of each method are listed on the right side of Figure 5. We obtain competitive lip-sync accuracy comparing with the state-of-the-art methods, even though we address a more challenging task.

5.3 Evaluation of Head Motion

Figures 6 shows that our head motion predictor creates natural and rhythmic head motions depending on both the audio and the identity. We further compare the head motion predictor with Zhou *et al*. [2020] (MakeItTalk) with the same audio and two reference images. The head movements of MakeItTalk are detected from the generated videos. For better visualization, we reduce the six dimensional head motions into one dimension by PCA, and show the sequential results in Figure 7. MakeItTalk hardly changes the head orientation and contains many repetitive behaviors, as shown in the red boxes. Their head motion patterns tend to be slightly swinging around the initial pose. In contrast, our head motions preserve the rhythm and synchronization with audio, and are much closer to the ground truth, as shown in the blue boxes. Furthermore, we produce corresponding motion sequences with different input identities.

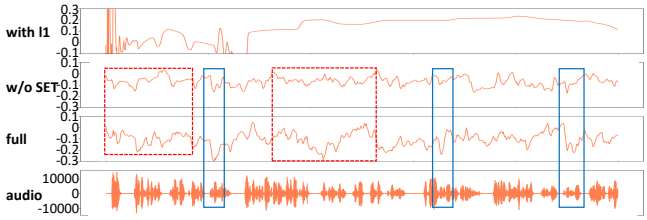

Figure 9: Ablation study on the head motion predictor.

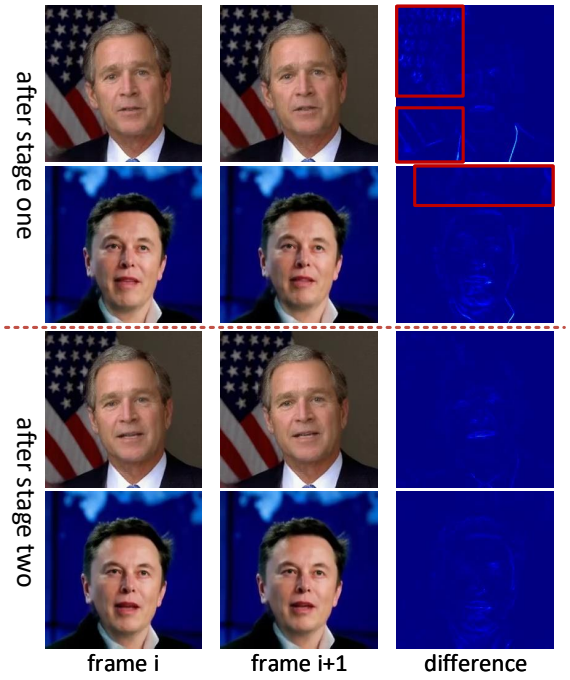

Figure 10: Results with one/two stage training. The red box shows the inconsistency between adjacent frames. Please zoom in for more details.

5.4 Ablation Study

We perform the quantitative evaluation of ablation study on VoxCeleb to illustrate the contribution of each component, the results are shown in Table 2 . We construct three variants including GT-keypoint (using GT keypoints), GT-head (using GT head movements) and our full model. Here, we also evaluate L1 distance between predicted and GT head movement vectors, marked as HE, L1 distance between generated and GT keypoints marked as KE. Although there exists numerical differences, the generated videos are still natural-looking.

We evaluate the effectiveness of Jacobians by removing $\mathbf{E_J}$ from $\mathbf{N_M}$. The generated result with and without Jacobians are shown in Figure 8. Without Jacobians, the lip seems to have only open-and-close patterns. It illustrates that the local affine transformations benefit to the lip shape details.

To show the effectiveness of the SSIM loss and SET in $\mathbf{N_H}$, we conduct two variants by replacing the SSIM loss with L1 loss (with L1), or removing SET from our model

	GT-keypoint	GT-head	full model
HE ↓	-	-	0.1910
KE ↓	-	0.0143	0.0630
PSNR ↑	26.40	25.37	21.19
SSIM ↑	0.82	0.79	0.68

Table 2: Quantitative results of ablation study on VoxCeleb.

	1	2	3	4	5	'natural'(4+5)
Chen *et al*. [2019]	8%	31%	39%	17%	5%	22.0%
Vougioukas *et al*. [2019]	41%	30%	14%	12%	2%	14.4%
Prajwal *et al*. [2020]	5%	20%	40%	30%	6%	35.6%
Zhou *et al*. [2020]	8%	35%	30%	21%	7%	28.0%
Ours	3%	8%	22%	39%	27%	**66.7%**

Table 3: Statistics of user study.

(w/o SET). The results are compared with our full method (full) in Figure 9. The model trained with L1 loss creates unnatural head motion sequence, because it suffers from the one-to-many mapping ambiguity. Results of the model without SET contains less dynamics and is not well synchronized with audio, especially in the red and blue boxes.

We then compare the results with or without the second training stage in Sec.3.3. We visualize the difference between two consecutive frames in Figure 10. Results without the refinement of the second stage contain texture inconsistency and slight jitters. The constraint on pixel level is helpful for improving the temporal coherence and the fidelity of videos.

5.5 User Study

We further conduct an online user study to compare the integrated quality of our method with state-of-the-arts. We create 4 videos for each method with the same input, to obtain $4 \times 5 = 20$ video clips. 33 participants are asked to rate "does the video look natural?" for each video from 1 to 5. The statistical results are shown in Table 3. Our method outperforms all compared methods significantly with the 66.7% of the cases that are judged as natural. It indicates that in addition to lip-sync, people are also quite sensitive to both frozen head pose and background artifacts. Besides, videos of only cropped faces [Chen *et al.*, 2019; Vougioukas *et al.*, 2019] are rated lower compared with others.

6 Conclusion and Discussion

In this paper, we propose a novel framework for one-shot talking-head generation from audio, which creates high fidelity videos with natural-looking and rhythmic head motions. We decouple the head motions from full-frame audio-dependent motions and predict the head motions individually in accordance with audio dynamics. Then, the motion field generator produces the keypoints that control the dense motion field from audio and head poses. Finally, an image rendering network synthesizes the videos using the dense motion field. Our method is evaluated qualitatively and quantitatively. The evaluation results show that our method predicts natural head motions, and produces few artifacts in non-face regions and between consecutive frames even though the head goes through a large pose change. Our method is proved to have a higher visual quality compared to the state-of-the-art.

Although our method outperforms previous works, our lip-sync accuracy drops on the bilabial and labiodental phonemes such as *p*, *f* and *m*. Compared to methods that focus on lip-sync, our framework trades a slight drop of lip-sync accuracy for much better head motion and visual quality. Such a trade-off is proved to be favored by most participants in our user study. We will be devoted to increasing the lip-sync accuracy without decreasing the current visual quality in future works. Besides, our method cannot capture the blink pattern, and fails on input reference images with extreme pose or expressions, which also needs to be addressed in the future.

Ethical Impact

With the convenience of creating photo-realistic videos for arbitrary identity and audio, our method has widespread positive applications, such as video conferencing and movie-dubbing. On the other hand, it may be misused by immoralists. To ensure proper use, we will release our code and models to promote the progress in detecting fake videos. Besides, we strongly require that any result created using our code and models must be marked as synthetic.

References

[Baltrusaitis *et al.*, 2018] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *IEEE International Conference on FG*, pages 59–66, 2018.

[Bansal *et al.*, 2018] Aayush Bansal, Shugao Ma, Deva Ramanan, and Yaser Sheikh. Recycle-gan: Unsupervised video retargeting. In *ECCV*, pages 119–135, 2018.

[Burkov *et al.*, 2020] Egor Burkov, Igor Pasechnik, Artur Grigorev, and Victor Lempitsky. Neural head reenactment with latent pose descriptors. In *CVPR*, pages 13786–13795, 2020.

[Chen *et al.*, 2019] 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*, pages 7832–7841, 2019.

[Chen *et al.*, 2020] Lele Chen, Guofeng Cui, Celong Liu, Zhong Li, Ziyi Kou, Yi Xu, and Chenliang Xu. Talking-head generation with rhythmic head motion. In *ECCV*, pages 35–51. Springer, 2020.

[Chung and Zisserman, 2016] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In *ACCV*, pages 87–103. Springer, 2016.

[Cooke *et al.*, 2006] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. An audio-visual corpus for speech perception and automatic speech recognition. *JASA*, 120(5):2421–2424, 2006.

[Ding *et al.*, 2013] Yu Ding, Catherine Pelachaud, and Thierry Artières. Modeling multimodal behaviors from

speech prosody. In *International Conference on Intelligent Virtual Agents*, pages 217–228, 2013.

[Fried *et al.*, 2019] Ohad Fried, Ayush Tewari, Michael Zollhöfer, Adam Finkelstein, Eli Shechtman, Dan B Goldman, Kyle Genova, Zeyu Jin, Christian Theobalt, and Maneesh Agrawala. Text-based editing of talking-head video. *ACM TOG*, 38(4):1–14, 2019.

[Greenwood *et al.*, 2017] David Greenwood, Stephen Laycock, and Iain Matthews. Predicting head pose from speech with a conditional variational autoencoder. *Proc. Interspeech 2017*, pages 3991–3995, 2017.

[Ha *et al.*, 2020] Sungjoo Ha, Martin Kersner, Beomsu Kim, Seokjun Seo, and Dongyoung Kim. Marionette: Few-shot face reenactment preserving identity of unseen targets. In *AAAI*, volume 34, pages 10893–10900, 2020.

[Kim *et al.*, 2018] 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 TOG*, 37(4):1–14, 2018.

[Li *et al.*, 2021] Lincheng Li, Suzhen Wang, Zhimeng Zhang, Yu Ding, Yixing Zheng, Xin Yu, and Changjie Fan. Write-a-speaker: Text-based emotional and rhythmic talking-head generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[Nagrani *et al.*, 2017] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: a large-scale speaker identification dataset. *arXiv*, 2017.

[Nirkin *et al.*, 2019] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *ICCV*, pages 7184–7193, 2019.

[Prajwal *et al.*, 2020] 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 *ACM MM*, pages 484–492, 2020.

[Sadoughi and Busso, 2018] Najmeh Sadoughi and Carlos Busso. Novel realizations of speech-driven head movements with generative adversarial networks. In *ICASSP*, pages 6169–6173. IEEE, 2018.

[Siarohin *et al.*, 2019a] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Animating arbitrary objects via deep motion transfer. In *CVPR*, pages 2377–2386, 2019.

[Siarohin *et al.*, 2019b] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *NIPS*, pages 7137–7147, 2019.

[Song *et al.*, 2019] Yang Song, Jingwen Zhu, Dawei Li, Andy Wang, and Hairong Qi. Talking face generation by conditional recurrent adversarial network. In *IJCAI*, pages 919–925, 2019.

[Suwajanakorn *et al.*, 2017] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017.

[Thies *et al.*, 2020] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural voice puppetry: Audio-driven facial reenactment. In *European Conference on Computer Vision*, pages 716–731. Springer, 2020.

[Vougioukas *et al.*, 2019] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Realistic speech-driven facial animation with gans. *International Journal of Computer Vision*, pages 1–16, 2019.

[Wang *et al.*, 2004] 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*, 13(4):600–612, 2004.

[Wang *et al.*, 2019] Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Bryan Catanzaro, and Jan Kautz. Few-shot video-to-video synthesis. *Advances in Neural Information Processing Systems*, 32:5013–5024, 2019.

[Yu and Porikli, 2018] Xin Yu and Fatih Porikli. Imagining the unimaginable faces by deconvolutional networks. *TIP*, 27(6):2747–2761, 2018.

[Yu *et al.*, 2018] Xin Yu, Basura Fernando, Bernard Ghanem, Fatih Porikli, and Richard Hartley. Face super-resolution guided by facial component heatmaps. In *ECCV*, pages 217–233, 2018.

[Yu *et al.*, 2019a] Xin Yu, Basura Fernando, Richard Hartley, and Fatih Porikli. Semantic face hallucination: Super-resolving very low-resolution face images with supplementary attributes. *IEEE transactions on pattern analysis and machine intelligence*, 42(11):2926–2943, 2019.

[Yu *et al.*, 2019b] Xin Yu, Fatemeh Shiri, Bernard Ghanem, and Fatih Porikli. Can we see more? joint frontalization and hallucination of unaligned tiny faces. *IEEE transactions on pattern analysis and machine intelligence*, 42(9):2148–2164, 2019.

[Zhang *et al.*, 2020] Jiangning Zhang, Xianfang Zeng, Mengmeng Wang, Yusu Pan, Liang Liu, Yong Liu, Yu Ding, and Changjie Fan. Freenet: Multi-identity face reenactment. In *CVPR*, pages 5325–5334, 2020.

[Zhang *et al.*, 2021] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

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

[Zhou *et al.*, 2020] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makelttalk: speaker-aware talking-head animation. *ACM TOG*, 39(6):1–15, 2020.

[Zhu *et al.*, 2020] Hao Zhu, Huaibo Huang, Yi Li, Aihua Zheng, and Ran He. Arbitrary talking face generation via attentional audio-visual coherence learning. In *IJCAI*, pages 2362–2368, 7 2020.