Sound to Visual: Hierarchical Cross-Modal Talking Face Video Generation

Lele Chen Haitian Zheng Ross K. Maddox Zhiyao Duan Chenliang Xu
University of Rochester, USA
{lchen63, hzheng15, rmaddox}@ur.rochester.edu, {zhiyao.duan, chenliang.xu}@rochester.edu

1 Introduction

Modeling the dynamics of a moving human face/body conditioned on another modality is a fundamental problem in computer vision, where applications are ranging from audio-to-video generation [3] to text-to-video generation and to skeleton-to-image/video generation [7]. This paper considers such a task: given a target face image and an arbitrary speech audio recording, generating a photo-realistic talking face of the target subject saying that speech with natural lip synchronization while maintaining a smooth transition of facial images over time (see Fig. 1). Note that the model should have a robust generalization capability to different types of faces (e.g., cartoon faces, animal faces) and to noisy speech conditions. Solving this task is crucial to enabling many applications, e.g., lip-reading from over-the-phone audio for hearing-impaired people, generating virtual characters with synchronized facial movements to speech audio for movies and games.

The main difference between still image generation and video generation is temporal-dependency modeling. There are two main reasons why it imposes additional challenges: people are sensitive to any pixel jittering (e.g., temporal discontinuities and subtle artifacts) in a video; they are also sensitive to slight misalignment between facial movements and speech audio. However, recent researchers [3, 2] tended to formulate video generation as a temporally independent image generation problem. In this paper, we propose a novel temporal GAN structure, which consists of a multi-modal convolutional-RNN-based (MMCRNN) generator and a novel regression-based discriminator structure. By modeling temporal dependencies, our MMCRNN-based generator yields smoother transactions between adjacent frames. Our regression-based discriminator structure combines sequence-level (temporal) information and frame-level (pixel variations) information to evaluate the generated video.

Another challenge of the talking face generation is to handle various visual dynamics (e.g., camera angles, head movements) that are not relevant to and hence cannot be inferred from speech audio. Those complicated dynamics, if modeled in the pixel space, will result in low-quality videos. For example, in web videos (e.g., LRW and VoxCeleb datasets), speakers move significantly when they are talking. Nonetheless, all the recent photo-realistic talking face generation methods [3, 9] failed to consider this problem. In this paper, we propose a hierarchical structure that utilizes a high-level facial landmarks representation to bridge the audio signal with the pixel image. Concretely, our algorithm first estimates facial landmarks from the input audio signal and then generates pixel variations in image space conditioned on generated landmarks. Besides leveraging intermediate landmarks for avoiding directly correlating speech audio with irrelevant visual dynamics, we also propose a novel dynamically adjustable loss along with an attention mechanism to enforce the network to focus on audiovisual-correlated regions.

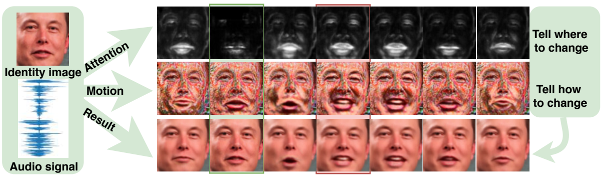

Figure 1: Problem description. The model takes an arbitrary audio speech and one face image, and synthesizes a talking face saying the speech.

Combining the above features, which are designed to overcome limitations of existing methods, our final model can capture informative audiovisual cues such as the lip movements and cheek movements while generating robust talking faces under significant head movements and noisy audio conditions. We evaluate our model along with state-of-the-art methods on several popular datasets (e.g., GRID [5], LRW [4], VoxCeleb [8] and TCD [6]). Experimental results show that our model outperforms all compared methods and all the proposed features contribute effectively to our final model. Furthermore, we also show additional novel examples of synthesized facial movements of the human/cartoon characters who are not in any dataset to demonstrate the robustness of our approach. The code has been released at **https://github.com/lelechen63/ATVGnet**.

2 Overview of Proposed Approach

Cascade Structure and Training Strategy We tackle the task of talking face video generation in a cascade perspective. Given the input audio sequence $a_{1:T}$, one example frame i_p and its landmarks p_p, our model generates facial landmarks sequence $\hat{p}_{1:T}$ and subsequently generates frames $\hat{v}_{1:T}$. To solve this problem, we come up with a

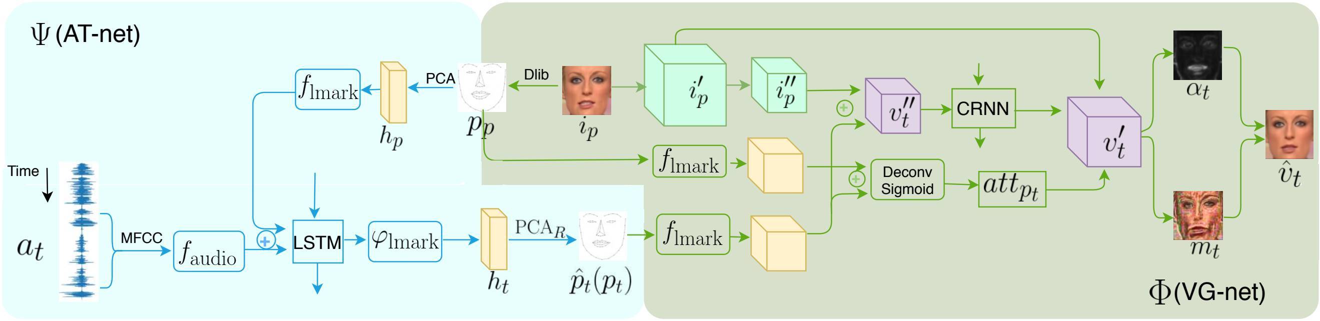

Figure 2: Overview of our network architecture. The blue part illustrates the AT-net, which transfers audio signal to low-dimensional landmarks representation and the green part illustrates the VG-net, which generates video frames conditioned on the landmark. During training, the input to VG-net are ground truth landmarks ($p_{1:T}$). During inference, the input to VG-net are fake landmarks ($\hat{p}_{1:T}$) generated by AT-net. The AT-net and VG-net are trained separately to avoid error accumulation.

novel cascade network structure:

$$\hat{p}_{1:T} = \Psi(a_{1:T}, p_p) \ , \tag{1}$$

$$\hat{v}_{1:T} = \Phi(\hat{p}_{1:T}, i_p, p_p) \ , \tag{2}$$

where the AT-net Ψ (see Fig. 2 blue part) is a conditional LSTM encoder-decoder and the VG-net Φ (see Fig. 2 green part) is a multi-modal convolutional recurrent network. During inference, the AT-net Ψ (see Eq. 1) observes audio sequence $a_{1:T}$ and example landmarks p_p and then predicts low-dimensional facial landmarks $\hat{p}_{1:T}$. By passing $\hat{p}_{1:T}$ into VG-net Φ (see Eq. 2) along with example image i_p and p_p, we subsequently get synthesized video frames $\hat{v}_{1:T}$. Ψ and Φ are trained in a decoupled way so that Φ can be trained with teacher forcing strategy. To avoid the error accumulation caused by $\hat{p}_{1:T}$, Φ is conditioned on ground truth landmarks $p_{1:T}$ during training.

Audio Transformation Network (AT-net) Specifically, the AT-net (Ψ) is formulated as:

$$[h_t, c_t] = \varphi_{\text{lmark}}(\text{LSTM}(f_{\text{audio}}(a_t), f_{\text{lmark}}(h_p), c_{t-1})), \tag{3}$$

$$\hat{p}_t = \text{PCA}_\text{R}(h_t) = h_t \odot \omega * \text{U}^T + \text{M} \ . \tag{4}$$

Here, the AT-net observes the audio MFCC a_t and landmarks PCA components h_p of the target identity and outputs PCA components h_t that are paired with the input audio MFCC. The $f_{\text{audio}}, f_{\text{lmark}}$ and φ_{lmark} indicate audio encoder, landmarks encoder and landmarks decoder. The c_{t-1} and c_t are outputs from cell units. PCA$_\text{R}$ is PCA reconstruction and ω is a boost matrix to enhance the PCA feature. The U corresponds to the largest eigenvalues and M is the mean shape of landmarks in the training set. In our empirical study, we observe that PCA can decrease the effect of none-audio-correlated factors (e.g., head movements) for training the AT-net.

Visual Generation Network (VG-net) Intuitively, similar to [1], we assume that the distance between current landmarks p_t and example landmarks p_p in feature space can represent the distance between current image frame and example image in image feature space. Based on this assumption (see Eq. 5), we can obtain current frame feature v''_t. Different from their methods, we replace element-wise addition with channel-wise concatenation in Eq. 5, which better preserves original frame information in our empirical study. In the meanwhile, we can also compute an attention map (att_{p_t}) based on the difference between p_t and p_p (see Eq. 6). By feeding the computed v''_t and att_{p_t} along with example image feature i'_p into the MMCRNN part, we obtain the current image feature v'_t (see Eq. 7). The resultant image feature v'_t will be used to generate video frames as detailed in the next section. Specifically, the VG-net is performed by:

$$v''_t = f_{\text{img}(i_p)} \oplus (f_{\text{lmark}}(p_t) - f_{\text{lmark}}(p_p)) \ , \tag{5}$$

$$att_{p_t} = \sigma(f_{\text{lmark}}(p_t) \oplus f_{\text{lmark}}(p_p)) \ , \tag{6}$$

$$v'_t = (\text{CRNN}(v''_t)) \odot \text{att}_{p_t} + i'_p \odot (\mathbf{1} - \text{att}_{p_t}) \ , \tag{7}$$

where $\oplus$ and $\odot$ are concatenation operation and element-wise multiplication, respectively. The CRNN part consists of Conv-RNN, residual block and deconvolution layers. i'_p is the middle layer output of $f_{\text{img}}(i_p)$, and σ is Sigmoid activation function. We omit some convolution operations in equations for better understanding.

Attention-Based Dynamic Pixel-wise Loss In order to solve the pixel jittering problem discussed in Sec. 1, we propose a novel dynamic pixel-wise loss to enforce the generator to generate consistent pixels along temporal axis. Intuitively, $\mathbf{0} \leq \boldsymbol{\alpha}_t \leq \mathbf{1}$ can be viewed as a spatial mask that indicates which pixels of given face image i_p need to move at time step t. We can also regard $\boldsymbol{\alpha}_t$ as a reference to represent to which extend each pixel contributes to the loss. The audiovisual-non-correlated regions should contribute less to the loss compared with the correlated regions. Thus, we propose a novel dynamic adjustable pixel-wise loss by leveraging the power of $\boldsymbol{\alpha}_t$, which is defined as:

$$\mathcal{L}_{\text{pix}} = \sum_{t=1}^{T} \|(v_t - \hat{v}_t) \odot (\overline{\boldsymbol{\alpha}}_t + \boldsymbol{\beta})\|_1) \ \ , \tag{8}$$

where $\overline{\boldsymbol{\alpha}}_t$ is the same as $\boldsymbol{\alpha}_t$ but without gradient. It represents the weight of each pixel dynamically that eases the generation. We remove the gradient of $\boldsymbol{\alpha}_t$ when back-propagating the loss to the network to prevent trivial solutions (lower loss but no discriminative ability). We also give base weights $\boldsymbol{\beta}$ to all pixels to make sure all pixels will be optimized. Here, we manually tune the hyper-parameter $\boldsymbol{\beta}$ and set $\boldsymbol{\beta} = \mathbf{0.5}$ in all of our experiments.

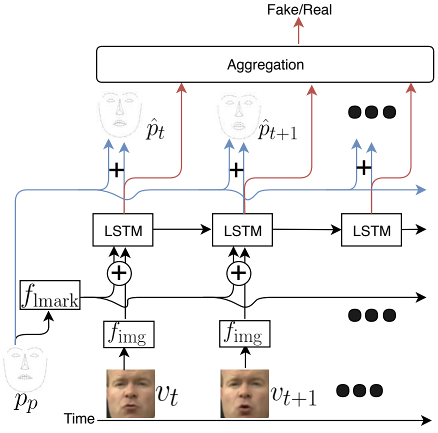

Figure 3: The overview of the regression-based discriminator. The $\oplus$ means concatenation. The $+$ means element-wise addition. The blue arrow and red arrow represent D_p and D_s, respectively.

Regression-Based Discriminator Perceptual loss utilizes high-level features to compare generated images and ground-truth images resulting in better sharpness of the synthesized images. Based on the perceptual loss, we propose a novel discriminator structure (see Fig. 3). The discriminator observes example landmarks p_p and either ground truth video frames $v_{1:T}$ or synthesized video frames $\hat{v}_{1:T}$, then regesses landmarks shapes $\hat{p}_{1:T}$ paired with the input frames, and additionally, gives a discriminative score s for the entire sequence. Specifically, we formulate discriminator into frame-wise part D_p (blue arrows in Fig. 3) and sequence-level part D_s (red arrows in Fig. 3). Thus our GAN loss can be expressed as:

$$\begin{aligned}\mathcal{L}_{\text{gan}} = & \mathbb{E}_{p_p, v_{1:T}}[\log \text{D}_s(p_p, v_{1:T})] + \\ & \mathbb{E}_{p_p, p_{1:T}, i_p}[\log(1 - \text{D}_s(p_p, \text{G}(p_p, p_{1:T}, i_p))] + \\ & \|(\text{D}_p(p_p, \text{G}(p_p, p_{1:T}, i_p)) - p_{1:T}) \odot \text{M}_p\|_2^2 + \\ & \|(\text{D}_p(p_p, v_{1:T}) - p_{1:T}) \odot \text{M}_p\|_2^2 \ \ ,\end{aligned} \tag{9}$$

where M_p is a pre-defined weight mask hyper-parameter which can penalize more on lip regions. By updating the parameters based on the regression loss when training the discriminator, the D_p can learn to extract low-dimensional representations from raw image data. When we train the generator, we will fix the weights of discriminator including D_s and D_p so that D_p will not compromise to generator. The loss back-propagated from D_p will enforce generator to generate accurate face shapes (e.g., cheek shape, lip shape etc.) and the loss back-propagated from D_s will enforce the network to generate high-quality images.

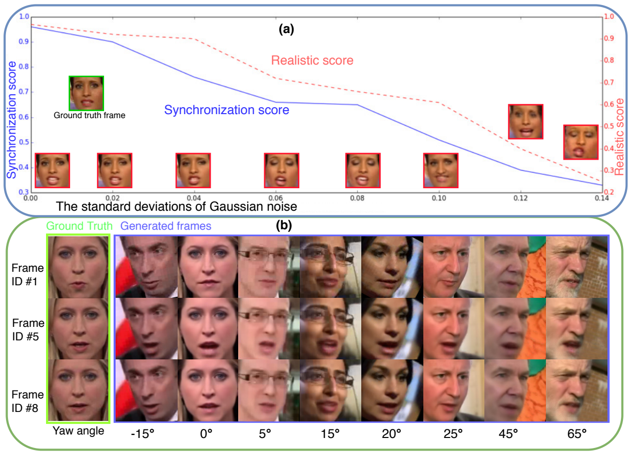

Figure 4: The trend of image quality w.r.t. (a) the landmarks (top) and (b) the poses (bottom). Please zoom in on a computer screen.

3 Experiments

Dataset and Implementation Details We quantitatively evaluate our ATVGnet on LRW [4], VoxCele [8] and TCD [6] datasets. For the image stream, all the talking faces in the videos are aligned based on key-points (eyes and nose) of the extracted landmarks at the sampling rate of 25FPS, and then resize to 128 × 128. As for audio data, each audio segment corresponds to 280ms audio. We extract MFCC at the window size of 10ms and use center image frame as the paired image data. Our network is implemented using Pytorch 0.4 library. We adopt Adam optimizer during training with the fixed learning rate of 2×10^{-4}. We initialize all network layers using random normalization with mean=0.0, std=0.2. All models are trained and tested on a single NVIDIA GTX 1080Ti. our inference time can achieve around 34.5 frames per second (FPS), which is slightly faster than real time (30 FPS).

Results To evaluate whether the synthesized video contains accurate lip movements that correspond to the input audio, we adopt the evaluation matrix Landmarks Distance (LMD) proposed in [1]. We compare our model with other three state-of-the-art methods [1, 3, 10]. The quantitative results are illustrated in Table 1. We can find that our

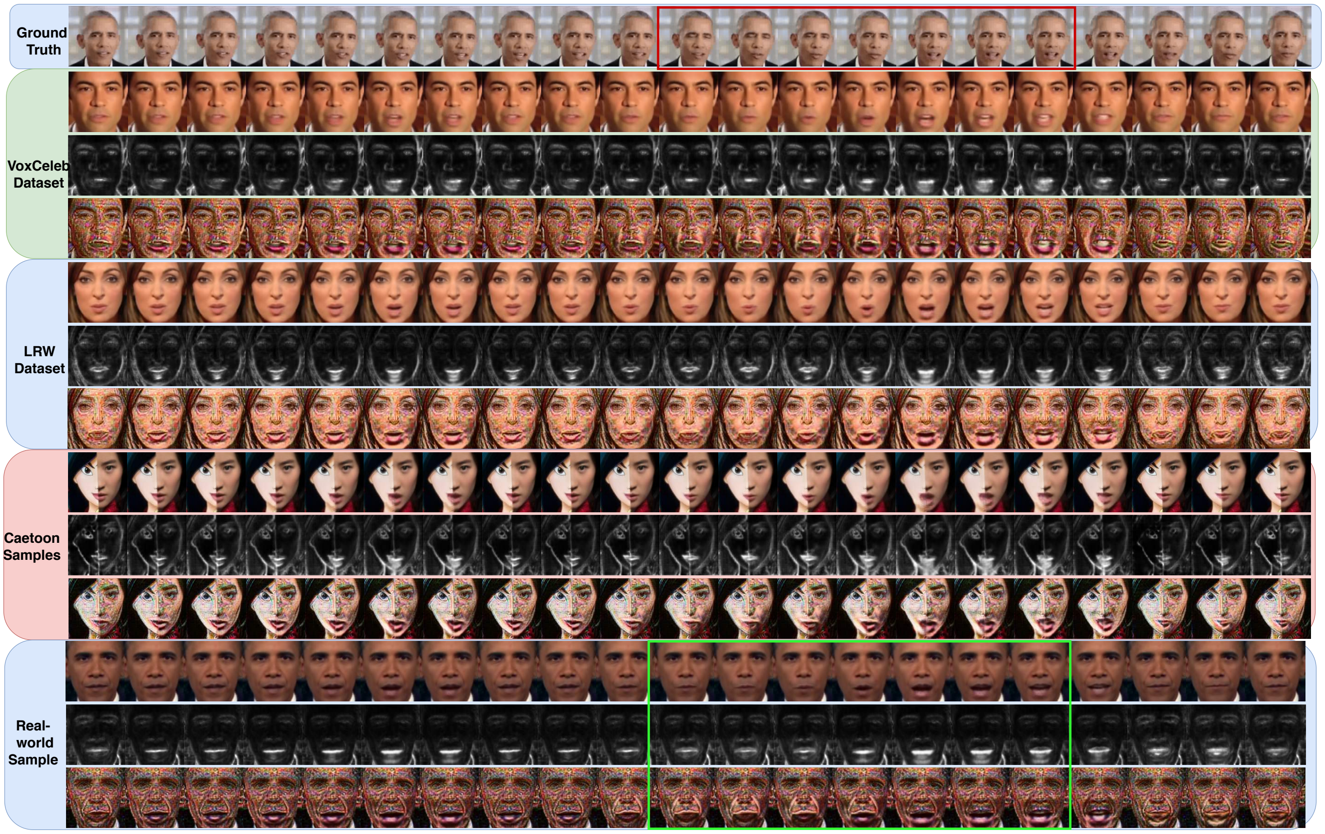

Figure 5: The outputs of ATVGnet. The first row is ground truth images paired with the given audio sequence. We mark the different sources of the identity image on the left side.

Method	LRW			GRID		
	LMD	SSIM	PSNR	LMD	SSIM	PSNR
Chen [1]	1.73	0.73	29.65	1.59	0.76	29.33
Wiles [10]	1.60	0.75	29.82	1.48	0.80	29.39
Chung [3]	1.63	0.77	29.91	1.44	0.79	29.87
ATVGnet	**1.37**	**0.81**	**30.91**	**1.29**	**0.83**	**32.15**

Table 1: Quantitative results of different methods on LRW dataset and GRID dataset.

ATVGnet achieves the best results both in image quality (SSIM, PSNR) and the correctness of audiovisual synchronization (LMD). In Fig. 4, we investigate the model performance w.r.t. the generated landmarks accuracy and different pose angles. We add Gaussian noises with different standard deviations to the generated landmarks during inference and conduct user study on the generated videos. The image quality drops (see Fig. 4(a)) if we increase the standard deviation. This phenomenon also indicates that our AT-net can output promising intermediate landmarks. To investigate the pose effects, we test different example images (different pose angles) with the same audio. The results in Fig. 4(b) demonstrate the robustness of our method w.r.t. the different pose angles. We also show our visual results in Fig. 5 to demonstrate the generalizability of our model in different datasets.

References

[1] L. Chen, Z. Li, R. K. Maddox, Z. Duan, and C. Xu. Lip movements generation at a glance. In *ECCV, 2018*, 2018. 2, 3, 4

[2] L. Chen, S. Srivastava, Z. Duan, and C. Xu. Deep cross-modal audio-visual generation. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*. 1

[3] J. S. Chung, A. Jamaludin, and A. Zisserman. You said that? In *BMVC*, 2017. 1, 3, 4

[4] J. S. Chung and A. Zisserman. Lip reading in the wild. In *ACCV 2016*. 1, 3

[5] M. Cooke, J. Barker, S. Cunningham, and X. Shao. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 2006. 1

[6] N. Harte and E. Gillen. TCD-TIMIT: an audio-visual corpus of continuous speech. *IEEE Trans. Multimedia*. 1, 3

[7] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. V. Gool. Pose guided person image generation. In *NeurIPS*, 2017. 1

[8] A. Nagrani, J. S. Chung, and A. Zisserman. Voxceleb: a large-scale speaker identification dataset. In *INTERSPEECH*, 2017. 1, 3

[9] S. Suwajanakorn, S. M. Seitz, and K. Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Trans. Graph.* 1

[10] O. Wiles, A. S. Koepke, and A. Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In *ECCV 2018*, 2018. 3, 4