Hierarchical Cross-Modal Talking Face Generation with Dynamic Pixel-Wise Loss

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

Abstract

We devise a cascade GAN approach to generate talking face video, which is robust to different face shapes, view angles, facial characteristics, and noisy audio conditions. Instead of learning a direct mapping from audio to video frames, we propose first to transfer audio to high-level structure, i.e., the facial landmarks, and then to generate video frames conditioned on the landmarks. Compared to a direct audio-to-image approach, our cascade approach avoids fitting spurious correlations between audiovisual signals that are irrelevant to the speech content. We, humans, are sensitive to temporal discontinuities and subtle artifacts in video. To avoid those pixel jittering problems and to enforce the network to focus on audiovisual-correlated regions, we propose a novel dynamically adjustable pixel-wise loss with an attention mechanism. Furthermore, to generate a sharper image with well-synchronized facial movements, we propose a novel regression-based discriminator structure, which considers sequence-level information along with frame-level information. Thoughtful experiments on several datasets and real-world samples demonstrate significantly better results obtained by our method than the state-of-the-art methods in both quantitative and qualitative comparisons.

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 [28, 3, 2] to text-to-video generation [23, 19] and to skeleton-to-image/video generation [21, 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 (see Fig. 7). 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.

Figure 1: Problem description. The model takes an arbitrary audio speech and one face image, and synthesizes a talking face saying the speech. The synthesized frames (last row) consist of synthesized attention (first row) and motion (second row), which demonstrate where and how the dynamics are synthesizing. For example, the face in the green box looks similar to the example face so that the attention map is almost dark; the face in the red box differs much from the example image, and hence the attention highlights the mouth region and the motion part hints white pixels for teeth.

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, 12, 17] tended to formulate video generation as a temporally independent image generation problem. For example, Chung et al. [3] proposed an encoder-decoder structure to generate one image from 0.35-second audio at each time. Song et al. [27] adopted a recurrent network to consider temporal dependency. They applied RNN in the feature extraction part, however, each frame was generated inde-

pendently in the generation stage. 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 [30], will result in low-quality videos. For example, in web videos [5, 24] (e.g., LRW and VoxCeleb datasets), speakers move significantly when they are talking. Nonetheless, all the recent photo-realistic talking face generation methods [3, 12, 27, 1, 28, 35] 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. It is worth to mention that in a recent audio-driven facial landmarks generation work [8], such irrelevant visual dynamics are removed in the training process by normalizing and identity-removing the facial landmarks. This has been shown to result in more natural synchronization between generated mouth shapes and speech audio.

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 [6], LRW [5], VoxCeleb [24] and TCD [13]). 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 contributions of our work can be summarized as follows: (1) We propose a novel cascade network structure to reduce the effects of the sound-irrelevant visual dynamics in the image space. Our model explicitly constructs high-level representation from the audio signal and guides video generation using the inferred representation. (2) We exploit a dynamically adjustable pixel-wise loss along with an attention mechanism which can alleviate temporal discontinuities and subtle artifacts in video generation. (3) We propose a novel regression-based discriminator to improve the audio-visual synchronization and to smooth the facial movement transition while generating realistic looking images. The code has been released at **https://github.com/lelechen63/ATVGnet**.

2. Related Work

In this section, we first briefly survey related work on the talking face generation task. Then we discuss the related work of each technique used in our model.

Talking Face Synthesizing The success of traditional approaches has been mainly limited to synthesizing a talking face from speech audio of a specific person [11, 9, 29]. For example, Suwajanakorn et al. [29] synthesized a taking face of President Obama with accurate lip synchronization, given his speech audio. The mechanism is to first retrieve the best-matched lip region image from a database through audiovisual feature correlation and then compose the retrieved lip region with the original face. However, this method requires a large amount of video footage of the target person. More recently, by combining the GAN/encoder-decoder structure and the data-driven training strategy, [27, 4, 1, 12] can generate arbitrary faces from arbitrary input audio.

High-Level Representations In recent years, high-level representations of images [31, 14, 34, 15] have been exploited in video generation tasks by using an encoder-decoder structure as the main approach. Given a condition, we can transfer it to high-level representations and feed them to a generative network to output a distribution over locations that a pixel is predicted to move. By adopting human body landmarks, Villegas et al. [31] proposed an encoder-decoder network which achieves long-term future prediction. Suwajanakorn et al. [28] transferred the audio signal to lip shapes and then synthesized the mouth texture based on the transferred lip shapes. These works have inspired us to use the facial landmarks to bridge audio with row pixel generation.

Attention Mechanism Attention mechanism is an emerging topic in natural language tasks [20] and image/video generation task [26, 37, 22, 36]. Pumarola et al. [26] generated facial expression conditioned on action units annotations. Instead of using a basic GAN structure, they exploited a generator that regresses an attention mask and a RGB color transformation over the entire image. The attention mask defines a per-pixel intensity specifying to what extend each pixel of the original image will contribute

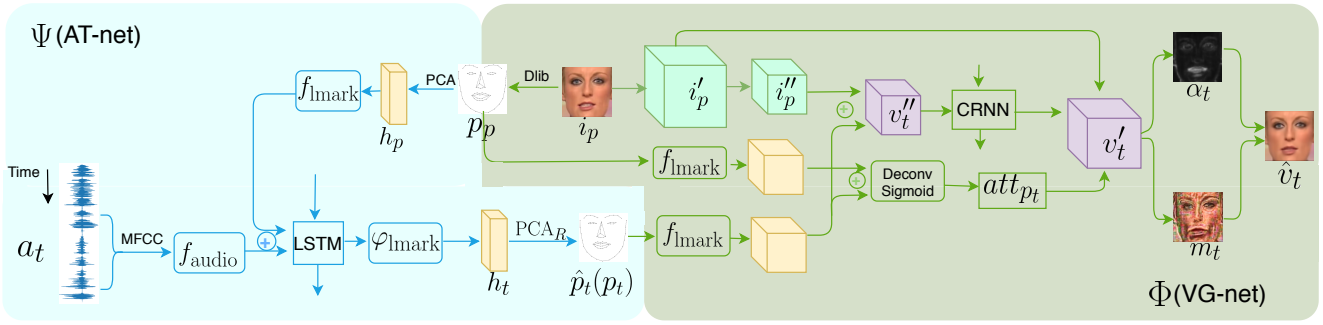

Figure 2: Overview of our network architecture. The blue part illustrates the AT-net, which transfers audio signal to low-dimensional landmarks representation and the green part illustrates the VG-net, which generates video frames conditioned on the landmark. During training, the input to VG-net are 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.

to the final rendered image. We adopt this attention mechanism to make our network robust to visual variations and noisy audio conditions. Feng et al. [10] observed that integrating a weighted mask into the loss function during training can improve the performance of the reconstruction network. Based on this observation, rather than using a fixed loss weights, we propose a dynamically adjustable loss by leveraging the attention mechanism to emphasize the audio-visual regions.

3. Architecture

This section describes the architecture of the proposed model. Fig. 2 shows the overall diagram, which is decoupled into two parts: audio transformation network (AT-net) and visual generation network (VG-net). First, we explain the overall architecture and the training strategy in Sec. 3.1. Then, we introduce two novel components: attention-based dynamic pixel-wise loss in Sec. 3.2 and a regression-based discriminator structure in Sec. 3.3 used in our VG-net.

3.1. Overview

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 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_{audio}, f_{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_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 [34, 31], 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 (size of $128 \times 8 \times 8$). 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. At the meanwhile, we can also compute an attention map (att_{p_t}) based on the dif-

Figure 3: The results of our baseline method. The synthesized frames with pixel jittering problem. The discontinuous problem and subtle artifacts will be amplified after composing into a video.

ference 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' (size of $128 \times 32 \times 32$) 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.

3.2. Attention-Based Dynamic Pixel-wise Loss

Recent works on video generation adopt either GAN-based methods [1, 32, 27] or Encoder-Decoder-based methods [3]. However, one common problem is the pixel jittering between adjacent frames (see Fig. 3). Pixel jittering is not obvious in single image generation but is a severe problem for video generation as humans are sensitive to any pixel jittering, e.g., temporal discontinuities and subtle artifacts in a video. The reason is that GAN loss or L1/L2 loss can barely generate perfect frames that all pixels are consistently changing in temporal domain, especially for audiovisual-non-correlated regions, e.g., background and head movements. In order to solve the pixel jittering problem, we propose a novel dynamic pixel-wise loss to enforce the generator to generate consistent pixels along temporal axis.

As mentioned in Sec. 2, Pumarola et al. [26] exploited a generator that regresses an attention mask and a RGB color transformation over the entire image. We adapt this attention mechanism in our VG-net to disentangle the motion part from audiovisual-non-correlated regions. Therefore, our final frame output is governed by the combination:

$$\hat{v}_t = \boldsymbol{\alpha}_t \odot m_t + (\mathbf{1} - \boldsymbol{\alpha}_t) \odot i_p \ , \tag{8}$$

where attention $\boldsymbol{\alpha}_t$ is obtained by applying convolution and Sigmoid activation operations on v_t', motion m_t is obtained by applying another convolution and hyperbolic tangent activation operations on v_t'. This step enforces the network to generate stable pixels in audiovisual-non-correlated regions while generating movements in audiovisual-correlated regions.

From Fig. 5, we can conclude that the pixels in audiovisual-non-correlated regions (e.g., hair, background etc.) usually attract less attention and are irrelevant to given condition (audio). In contrast, the network is mainly focusing on correlated regions (e.g., mouth, jaw, and cheek). 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 what 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:

$$\mathcal{L}_{\text{pix}} = \sum_{t=1}^{T} \|(v_t - \hat{v}_t) \odot (\overline{\boldsymbol{\alpha}}_t + \boldsymbol{\beta})\|_1) \ , \tag{9}$$

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.

3.3. Regression-Based Discriminator

Recently, people find that perceptual loss [16] is helpful for generating sharp images in GAN/VAE [27, 1]. Perceptual loss utilizes high-level features to compare generated images and ground-truth images resulting in better sharpness of the synthesized images. The key idea is that the weights of the perceptual network part are fixed, and the loss will only contribute to the generator/decoder part. Based on this intuition, we propose a novel discriminator structure (see Fig. 4). 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 regresses 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. 4) and sequence-level part D_s (red arrows in Fig. 4).

The D_p observes example landmarks and video frames, then regresses the landmarks sequence based on observed information. By yielding the facial landmark, it can evaluate the input image based on high-level representation in a

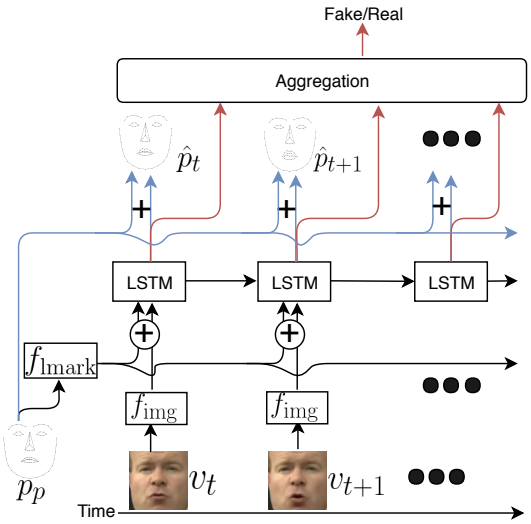

Figure 4: The overview of the regression-based discriminator. The $\oplus$ means concatenation. The $+$ means element-wise addition. The blue arrow and red arrow represent D_p and D_s, respectively.

frame-wise fashion. Specifically, the $\hat{p}_t$ is calculated by:

$$\begin{aligned}\hat{p}_t &= \mathrm{D}_p(p_p, v_t) \\ &= p_p + \mathrm{LSTM}(f_{\text{lmark}}(p_p) \oplus f_{\text{img}}(v_t)) \ , \end{aligned} \tag{10}$$

which observes ground truth image during discriminator training stage and observes synthesized image during generator training stage.

Besides D_p, the LSTM cell unit yields another branch D_s, which obtains vectors from each LSTM cell unit and aggregates them by average pooling. By passing through a Sigmoid activation function, D_s yields final discriminative score s for the overall input sequence. The score s can obtained by:

$$\begin{aligned} s &= \mathrm{D}_s(p_p, v_{1:T}) \\ &= \sigma(\frac{1}{T}\sum_{t=1}^{T}(\mathrm{LSTM}(f_{\text{lmark}}(p_p) \oplus f_{\text{img}}(v_t)))) \ . \end{aligned} \tag{11}$$

The D_p part is optimized to minimize the L2 loss between the predicted landmarks and the ground truth landmarks. Thus our GAN loss can be expressed as:

$$\begin{aligned} \mathcal{L}_{\text{gan}} =& \mathbb{E}_{p_p, v_{1:T}}[\log \mathrm{D}_s(p_p, v_{1:T})] + \\ & \mathbb{E}_{p_p, p_{1:T}, i_p}[\log(1 - \mathrm{D}_s(p_p, \mathrm{G}(p_p, p_{1:T}, i_p)))] + \\ & \|(\mathrm{D}_p(p_p, \mathrm{G}(p_p, p_{1:T}, i_p)) - p_{1:T}) \odot \mathrm{M}_p\|_2^2 + \\ & \|(\mathrm{D}_p(p_p, v_{1:T}) - p_{1:T}) \odot \mathrm{M}_p\|_2^2 \ , \end{aligned} \tag{12}$$

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.

3.4. Objective Function

By linearly combining all partial losses introduced in Sec. 3.2 and Sec. 3.3, the *full loss* function $\mathcal{L}$ can be expressed as:

$$\mathcal{L} = \mathcal{L}_{\text{gan}} + \lambda * \mathcal{L}_{\text{pix}} \ , \tag{13}$$

where λ is a hyper-parameter that controls the relative importance of different loss terms. We set $\lambda = 10.0$ in our experiments.

4. Experiments

In this section, we conduct thoughtful experiments to demonstrate the efficiency and effectiveness of the proposed architecture for video generation. Sec. 4.1 explains datasets and implementation in detail. Sec. 4.2 shows our results along with other state-of-the-art methods. We show user studies and ablation study in Sec.4.3 and Sec. 4.4 respectively.

4.1. Experimental Setup

Dataset We quantitatively and qualitatively evaluate our ATVGnet on LRW dataset [4] and GRID dataset [6]. The LRW dataset consists of 500 different words spoken by hundreds of different speakers in the wild. We follow the same train-test split as in [4]. In GRID dataset, there are 1000 short videos, each spoken by 33 different speakers in the experimental condition. For the image stream, all the talking faces in the videos are aligned based on key-points (eyes and nose) of the extracted landmarks using [18] 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. Similar to [3, 27], we remove the first coefficient from the original MFCC vector, and eventually yield a 28×12 MFCC feature for each audio chunk.

Implementation Details 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. During the training, the AT-net converges after 3 hours and the VG-net is stable after

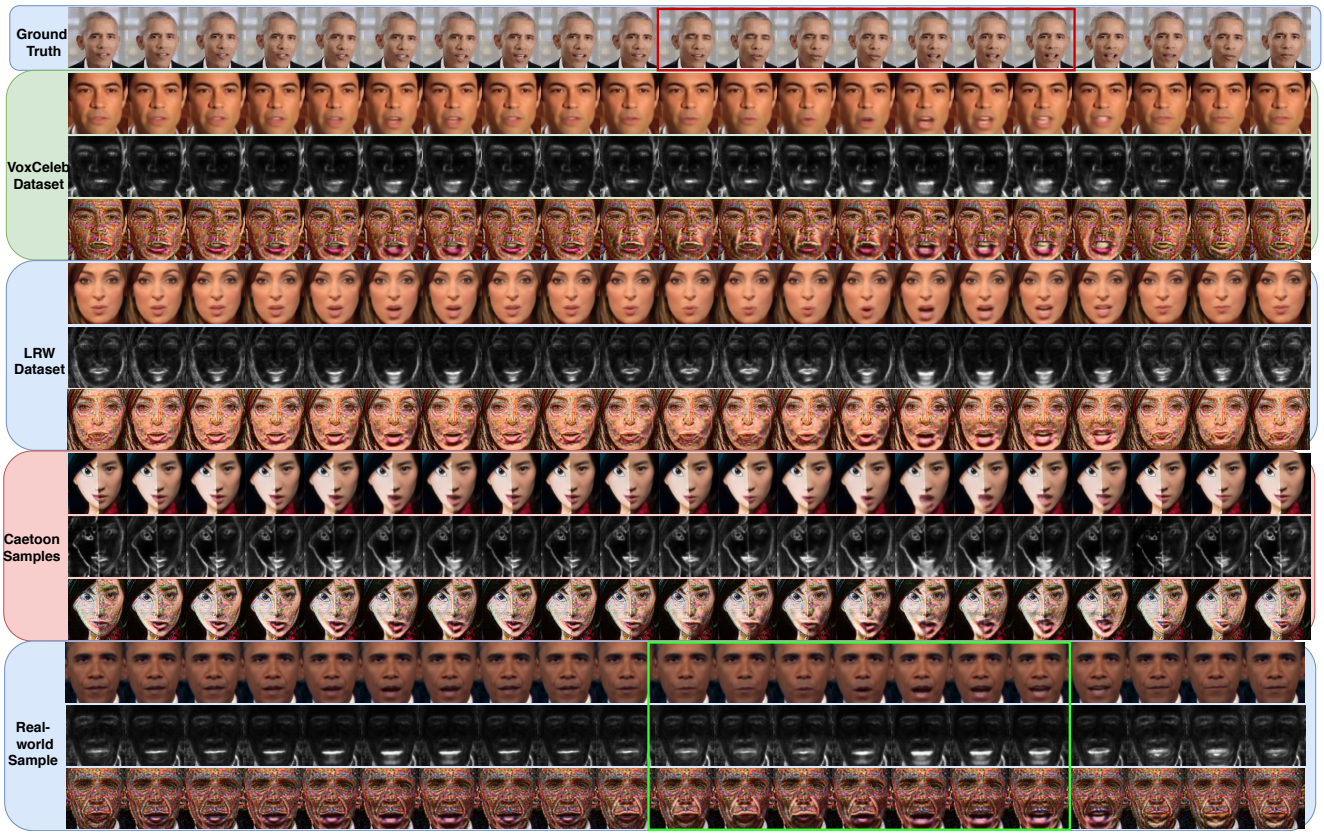

Figure 5: The outputs of ATVGnet. The inputs are one real-world audio sequence and different example identity images range from real-world people to cartoon characters. The first row is ground truth images paired with the given audio sequence. We mark the different sources of the identity image on the left side. From this figure, we can find that the lip movements of our synthesized frames (e.g., the green box in the last row) are well-synchronized with the ground truth (red box in first row). Meanwhile, the attention (middle row of the green box) accurately indicates where need to move and the motion (last row of the green box) indicates what the dynamics look like (e.g. white pixels for teeth and red pixels for lips).

Method	Real time	ATVGnet(our)	Chung et al.[3]	Zhou et al.[12]	Wiles et al.[35]
Inference time (FPS)	30	34.53	19.10	10.00	10.53

Table 1: The inference time of difference models. We use frame rate (FPS) to measure the time.

24 hours. Table 1 shows the generation time during inference stage. We can find that our inference time can achieve around 34.5 frames per second (FPS), which is much faster than [34, 12, 3] and slightly faster than real time (30 FPS).

4.2. Results

Image results are illustrated in Fig. 5 and Fig. 7. To evaluate the quality of the synthesized video frames, we compute PSNR and SSIM [33]. 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, 35]. All of them are trained on LRW dataset while Chung et al. [3] require extra VGG-M network pretrained on VGG Face dataset [25] and Wilels et al. [35] need extra MFCC feature extractor pretrained by [5]. The quantitative results are illustrated in Table 2. The Baseline model is a straightforward model without any features (e.g., DMA, MMCRNN, DAL and RD explained in Sec. 4.4) as mentioned in Sec. 3. The model ATVG-ND has the same network structure as ATVGnet. But it is trained end-to-end without the decoupled training strategy (see Sec. 3.1). We can find that our ATVGnet achieves the best results both in image quality (SSIM, PSNR) and the correctness of audio-visual synchronization (LMD).

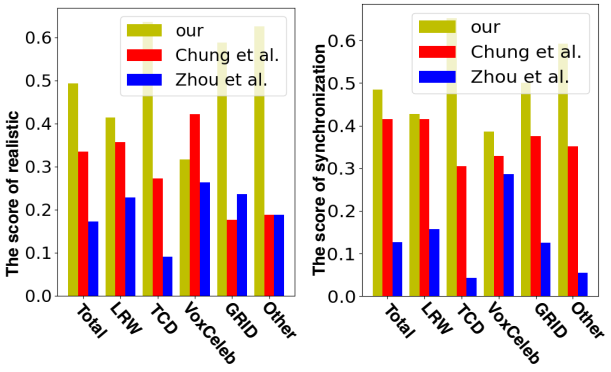

Figure 6: Statistics of user studies. The y-axis is the percentage of votes and the x-axis is different data sources (e.g., $total$ means all the video samples, $Other$ means sampled videos from YouTube.) The left histogram is the rating on authenticity. The right histogram is the rating on synchronization between facial movements and audio.

Method	LRW			GRID		
	LMD	SSIM	PSNR	LMD	SSIM	PSNR
Chen [1]	1.73	0.73	29.65	1.59	0.76	29.33
Wiles [35]	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
Baseline	1.71	0.72	28.95	1.82	0.77	28.78
ATVG-ND	**1.35**	0.78	30.27	1.34	0.79	30.51
ATVGnet	1.37	**0.81**	**30.91**	**1.29**	**0.83**	**32.15**

Table 2: Quantitative results of different methods on LRW dataset and GRID dataset. Our models mentioned in this table are trained from scratch. We bold each leading score.

4.3. User Studies

Our goal is to generate realistic videos based on the audio information. The evaluation in 4.2 can only evaluate the quality in a single frame style. To evaluate the performance in a video level, we conduct thoughtful user studies in this section. Human subjects evaluation (see Fig. 6) is conducted to investigate the visual qualities of our generated results compared with Chung et al. [3] and Zhou et al. [12]. The ground truth videos are selected from different sources: we randomly select samples from the testing set of LRW [5], VoxCeleb [24], TCD [13], GRID [6] and real-world samples from YouTube (in total 38 videos). Three methods are evaluated w.r.t. two different criteria: whether participants could regard the generated talking faces as realistic and whether the generated talking faces temporally sync with the corresponding audio. We shuffle all the sample videos and the participants are not aware of the mapping between videos to methods. They are asked to score the images on a scale of 0 (worst) to 10 (best). There are overall 10 participants involved, and the results are summed over persons and video time steps.

According to the ratings from Fig. 6, we can find that our method outperforms other two methods in terms of the extent of synchronization and authenticity. More specifically, our model achieves the best results on all datasets in terms of lip synchronization with audio input. As for image authenticity, our model achieves the highest score the on most of the datasets but slightly lower than Chung et al. [3] on the VoxCeleb testing set. We attribute this to the audio noise (e.g. background music) in the test samples.

Method	LRW			GRID		
	LMD	SSIM	PSNR	LMD	SSIM	PSNR
ATVGnet	**0.80**	**0.86**	**33.45**	**0.70**	**0.89**	**33.84**
w/o DMA	0.98	0.83	30.22	1.10	0.84	29.90
w/o MM-CRNN	1.03	0.80	30.61	0.81	0.86	32.68
w/o DAL	0.86	**0.86**	31.35	0.76	0.87	33.11
w/o RD	0.82	0.84	32.84	0.73	0.88	33.25
Baseline	1.27	0.81	29.55	1.17	0.80	29.45
ATVG-P	0.90	0.84	30.45	0.75	0.87	31.78

Table 3: Ablation studies on the LRW dataset and the GRID dataset. We remove each feature at a time. We bold the highest scores.

4.4. Ablation Studies

We conduct ablation experiments to study the contributions of the four components introduced in Sec. 3: Dynamic Motion & Attention (DMA), Multi-Modal-crnn (MMCRNN), Dynamically Adjustable Loss (DAL) and Recreational Discriminator (RD). The ablation studies are conducted on both LRW dataset and GRID dataset. Results are shown in Table 3. Here we follow the protocols mentioned in Sec. 4.1. We test each model using ground truth landmarks rather than fake landmarks generated by AT-net, so that we can eliminate the errors caused by uncorrelated noise and focus on each component.

As shown in Table 3, each component contributes to the full model. We can find that MMCRNN and DMA are critical to our full model. We attribute this to the better ability of generating smooth transactions between adjacent frames. The ATVG-P model has the same structure as ATVGnet but conditioned on the last fake frame $\hat{v}_{t-1}$ rather than the example frame i_p in Eq. 8 in Sec. 3.2. We suppose it could yield better performance. However, the error amplifies quickly through time until it overwhelms the visual information from example frame, which leads to a trivial solution that $\boldsymbol{\alpha}_t = 0_{n\times n}$ and decreases the performance.

We investigate the model performance w.r.t. the gen-

Figure 7: Qualitative results produced by ATVGnet, Chung et al. [3] and Zhou et al. [12] on samples from LRW and VoxCeleb dataset. We can observe from it that our mouth opening is closer to ground truth compared with the other two methods. It is worthwhile to mention that the second sample is recorded outside with loud background **noise**.

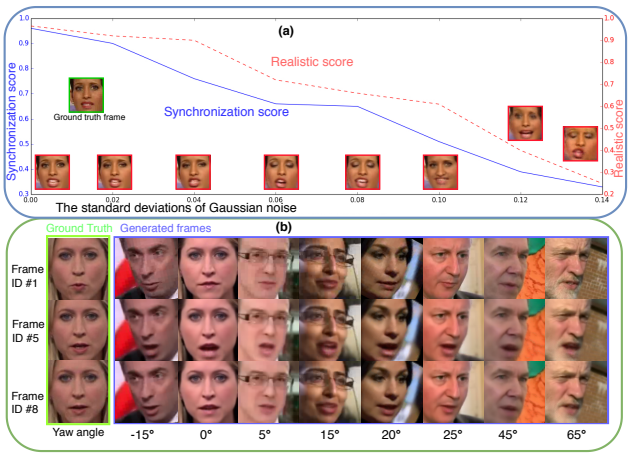

Figure 8: The trend of image quality w.r.t. (a) the landmarks (top) and (b) the poses (bottom). Please zoom in on a computer screen.

erated landmarks accuracy and different pose angles (see Fig. 8). 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. 8(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. 8(b) demonstrate the robustness of our method w.r.t. the different pose angles.

5. Conclusion and Discussion

In this paper, we present a cascade talking face video generation approach utilizing facial landmarks as intermediate high-level representations to bridge the gap between two different modalities. We propose a novel Multi-Modal Convolutional-RNN structure, which considers the correlation between adjacent frames in the generation stage. Meanwhile, we propose two novel components: dynamically adjustable loss and regression-based discriminator. In our perspective, these two techniques are general that could be adopted in other tasks (e.g., human body generation and facial expression generation) in the future. Our final model ATVGnet achieves the best performance on several popular datasets in both qualitative and quantitative comparisons. For future work, applying other techniques to enable our network to generate unconscious head movements/expressions could be an interesting topic, which has been bypassed in our current approach.

Acknowledgement. This work was supported in part by NSF IIS 1741472, IIS 1813709, and the University of Rochester AR/VR Pilot Award. This article solely reflects the opinions and conclusions of its authors and not the funding agents.

References

[1] L. Chen, Z. Li, R. K. Maddox, Z. Duan, and C. Xu. Lip movements generation at a glance. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII*, pages 538–553, 2018.

[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, Mountain View, CA, USA, October 23 - 27, 2017*, pages 349–357, 2017.

[3] J. S. Chung, A. Jamaludin, and A. Zisserman. You said that? In *British Machine Vision Conference 2017, BMVC 2017, London, UK, September 4-7, 2017*, 2017.

[4] J. S. Chung and A. Zisserman. Lip reading in the wild. In *Computer Vision - ACCV 2016 - 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II*, pages 87–103, 2016.

[5] J. S. Chung and A. Zisserman. Out of time: Automated lip sync in the wild. In *Computer Vision - ACCV 2016 Workshops - ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II*, pages 251–263, 2016.

[6] 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.

[7] X. Di, V. A. Sindagi, and V. M. Patel. Gp-gan: Gender preserving gan for synthesizing faces from landmarks. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 1079–1084, Aug 2018.

[8] S. E. Eskimez, R. K. Maddox, C. Xu, and Z. Duan. Generating talking face landmarks from speech. In *Latent Variable Analysis and Signal Separation - 14th International Conference, LVA/ICA 2018, Guildford, UK, July 2-5, 2018, Proceedings*, pages 372–381, 2018.

[9] B. Fan, L. Wang, F. K. Soong, and L. Xie. Photo-real talking head with deep bidirectional LSTM. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015*, pages 4884–4888, 2015.

[10] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIV*, pages 557–574, 2018.

[11] P. Garrido, L. Valgaerts, H. Sarmadi, I. Steiner, K. Varanasi, P. Pérez, and C. Theobalt. Vdub: Modifying face video of actors for plausible visual alignment to a dubbed audio track. *Comput. Graph. Forum*, 34(2):193–204, 2015.

[12] Z. L. P. L. X. W. Hang Zhou, Yu Liu. Talking face generation by adversarially disentangled audio-visual representation. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

[13] N. Harte and E. Gillen. TCD-TIMIT: an audio-visual corpus of continuous speech. *IEEE Trans. Multimedia*, 17(5):603–615, 2015.

[14] S. Hong, X. Yan, T. S. Huang, and H. Lee. Learning hierarchical semantic image manipulation through structured representations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 2708–2718. Curran Associates, Inc., 2018.

[15] S. Hong, D. Yang, J. Choi, and H. Lee. Inferring semantic layout for hierarchical text-to-image synthesis. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[16] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, 2016.

[17] T. Karras, T. Aila, S. Laine, A. Herva, and J. Lehtinen. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Trans. Graph.*, 36(4):94:1–94:12, 2017.

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

[19] Y. Li, M. R. Min, D. Shen, D. E. Carlson, and L. Carin. Video generation from text. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 7065–7072, 2018.

[20] T. Luong, H. Pham, and C. D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1412–1421, 2015.

[21] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. V. Gool. Pose guided person image generation. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 405–415, 2017.

[22] S. Ma, J. Fu, C. Wen Chen, and T. Mei. Da-gan: Instance-level image translation by deep attention generative adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[23] T. Marwah, G. Mittal, and V. N. Balasubramanian. Attentive semantic video generation using captions. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 1435–1443, 2017.

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

[25] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *Proceedings of the British Machine Vision Conference 2015, BMVC 2015, Swansea, UK, September 7-10, 2015*, pages 41.1–41.12, 2015.

[26] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part X*, pages 835–851, 2018.

[27] Y. Song, J. Zhu, X. Wang, and H. Qi. Talking face generation by conditional recurrent adversarial network. *CoRR*, abs/1804.04786, 2018.

[28] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Trans. Graph.*, 36(4):95:1–95:13, 2017.

[29] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Trans. Graph.*, 36(4):95:1–95:13, 2017.

[30] R. Villegas, J. Yang, S. Hong, X. Lin, and H. Lee. Decomposing motion and content for natural video sequence prediction. *ICLR*, 2017.

[31] R. Villegas, J. Yang, Y. Zou, S. Sohn, X. Lin, and H. Lee. Learning to generate long-term future via hierarchical prediction. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 3560–3569, 2017.

[32] C. Vondrick, H. Pirsiavash, and A. Torralba. Generating videos with scene dynamics. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 613–621, 2016.

[33] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Processing*, 2004.

[34] N. Wichers, R. Villegas, D. Erhan, and H. Lee. Hierarchical long-term video prediction without supervision. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 6033–6041, 2018.

[35] O. Wiles, A. S. Koepke, and A. Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIII*, pages 690–706, 2018.

[36] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[37] H. Zhang, I. J. Goodfellow, D. N. Metaxas, and A. Odena. Self-attention generative adversarial networks. *CoRR*, abs/1805.08318, 2018.