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

Lele Chen Ross K. Maddox Zhiyao Duan Chenliang Xu
University of Rochester

CVPR LONG BEACH CALIFORNIA June 16-20, 2019

Introduction

- **Motivation:** The main difference between image generation and the talking face 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, see Fig. 1) in a video; they are also sensitive to slight misalignments between facial movements and speech audio. However, recent researchers tended to formulate video generation as a temporally independent image generation problem.

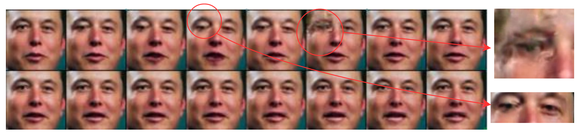

Figure 1: Pixel jittering problem in video generation

- **Objective:** Our method (see Fig. 2) takes an arbitrary audio speech and one face image, and synthesizes a talking face saying the speech **in real time**. The synthesized frames (last row) consist of synthesized attention (first row) and motion (second row), which demonstrate where and how the dynamics are synthesizing.

Figure 2: Problem description

- **Contributions**
  - 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.
  - We exploit a dynamically adjustable pixel-wise loss along with an attention mechanism which can alleviate temporal discontinuities and subtle artifacts in video generation.
  - 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.
  - **Code**, pretrained model and demo video are available at .

- **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.

Method

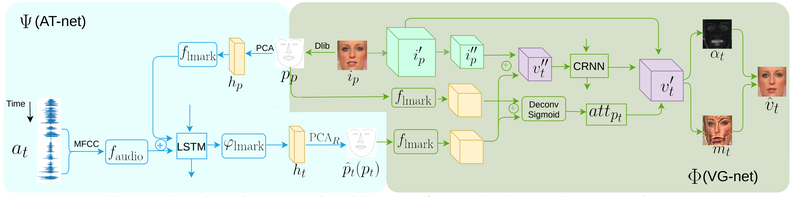

Figure 3: Overview of our network architecture. Blue part is AT-net and green part is VG-net.

- **AT-net** observes the audio MFCC and landmark PCA components of the target identity and outputs PCA components that are paired with the input audio MFCC.
- **VG-net** generates video frames conditioned on the landmarks and one single face image. It encodes image feature and landmark feature separately and then fuses the features in MMCRNN module.
- **Attention-Based Dynamic Pixel-wise Loss** is designed to avoid pixel jittering problems and to enforce the VG-net to focus on audiovisual-correlated regions. Intuitively, $\boldsymbol{\alpha}_t$ can be viewed as a spatial mask that indicates which pixels of given face image 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.

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

- **Regression-Based 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.

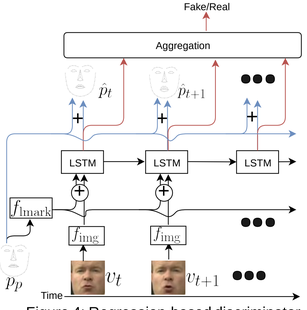

Figure 4: Regression-based discriminator

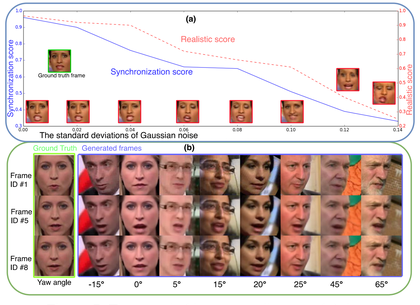

Figure 5: The trend of image quality w.r.t. (a) the landmarks (top) and (b) the poses (bottom).

Experiments

- **Qualitative results** are shown in Fig. 6, which are conditioned on one real-world audio sequence and different example identity images range from real-world people to cartoon characters.

Figure 6: Quantitative results

- **Comparison with SOTA methods** are illustrated in Fig. 7. The input samples are randomly selected from LRW and VoxCeleb dataset. Please refer to the paper for more results.

Figure 7: Qualitative results produced by ATVGnet, Chung et al. and Zhou et al.

- **Ablation studies** are shown in Fig. 5. Please refer to the paper for more details.