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

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

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.

- **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.

- **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 [here](#).**

- **Acknowledgement:** This work was supported in part by NSF IIS 1714172, 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

- **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, $c_t$ 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 $c_t$ as a reference to represent to what extent each pixel contributes to the loss.

$$L_{pix} = \sum_t \sum_i \left((c_t - \hat{c}_t) \odot (\hat{x}_t + \beta)\right)$$

- **Regression-Based Discriminator** observes example landmarks $p$ and either ground truth video frames $T_f$ or synthesized video frames $\hat{T}_f$ then regresses landmarks shapes $p_{1:T_f}$ paired with the input frames, and additionally, gives a discriminative score $s$ for the entire sequence.

$$L_{Dc} = \sum_{t} \left(\frac{1}{2}s - \log(1 + \exp(-s))\right)$$

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.

- **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.

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