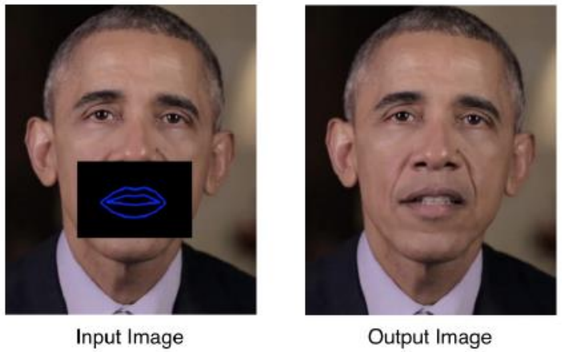
Paper 1445

# A Neural Lip-Sync Framework for Synthesizing Photorealistic Virtual News Anchors

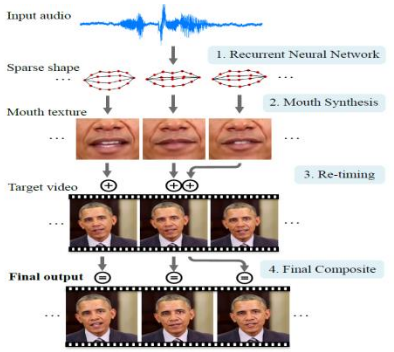
Ruobing Zheng, Zhou Zhu, Bo Song, Changjiang Ji

**Moviebook Technology**

**Lip Sync** “rewrite” the lip motions on a target video clip based on the given speech content.




[1]




[2]

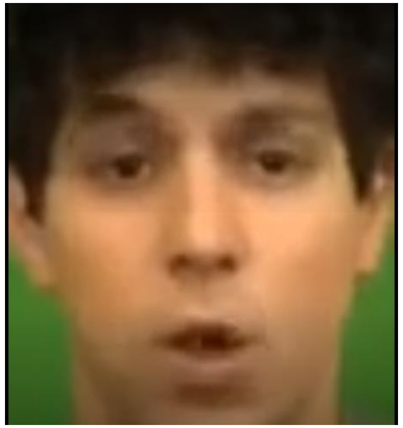
**Related work:**

[1] “Obamanet: Photo-realistic lip-sync from text” R. Kumar
[2] “Synthesizing obama: learning lip sync from audio” S. Suwajanakorn
......

## Two main problems

- **Quality**: Resolution, Visual consistency, Natural appearance
- **Efficiency**: Training, Inference


Obamanet 2017

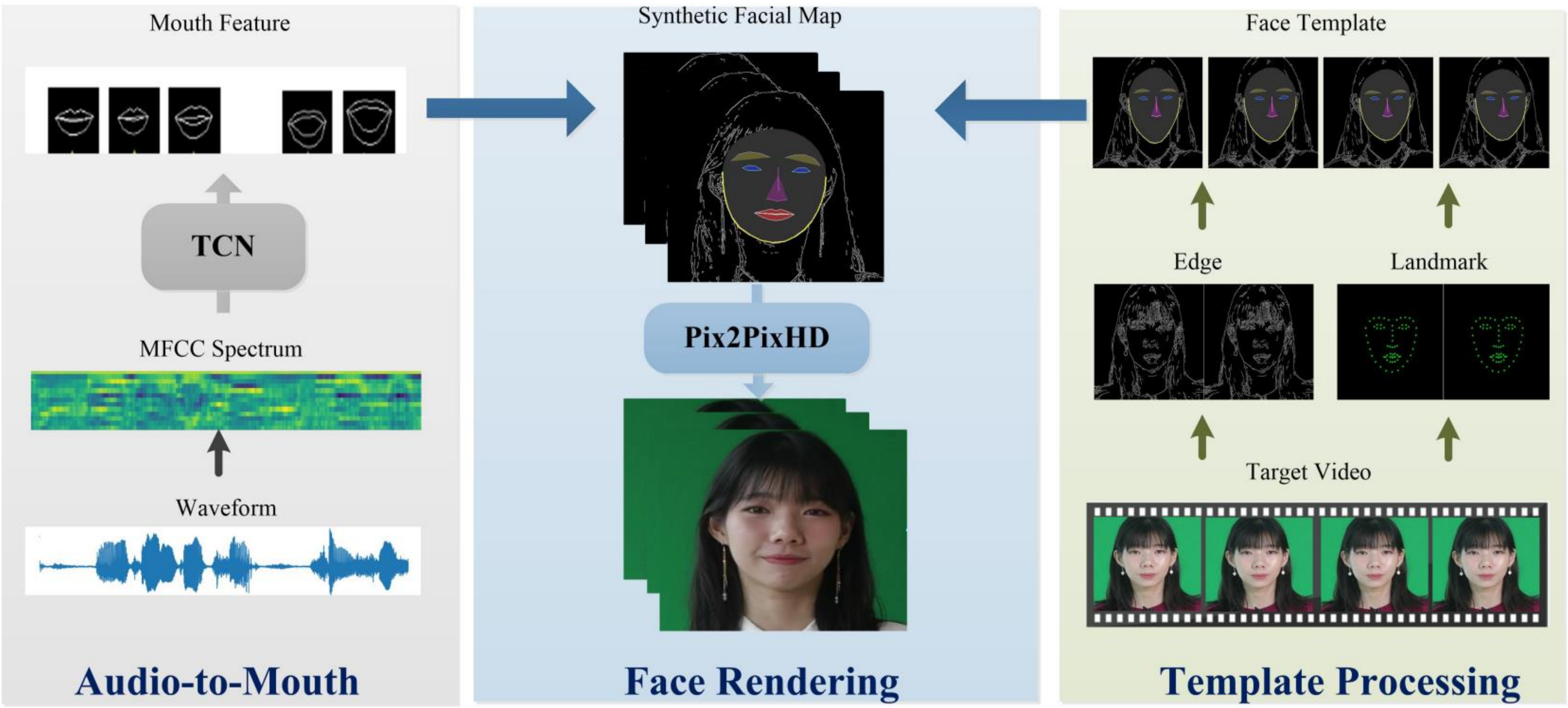


Realistic Speech-Driven Facial Animation with GANs 2019

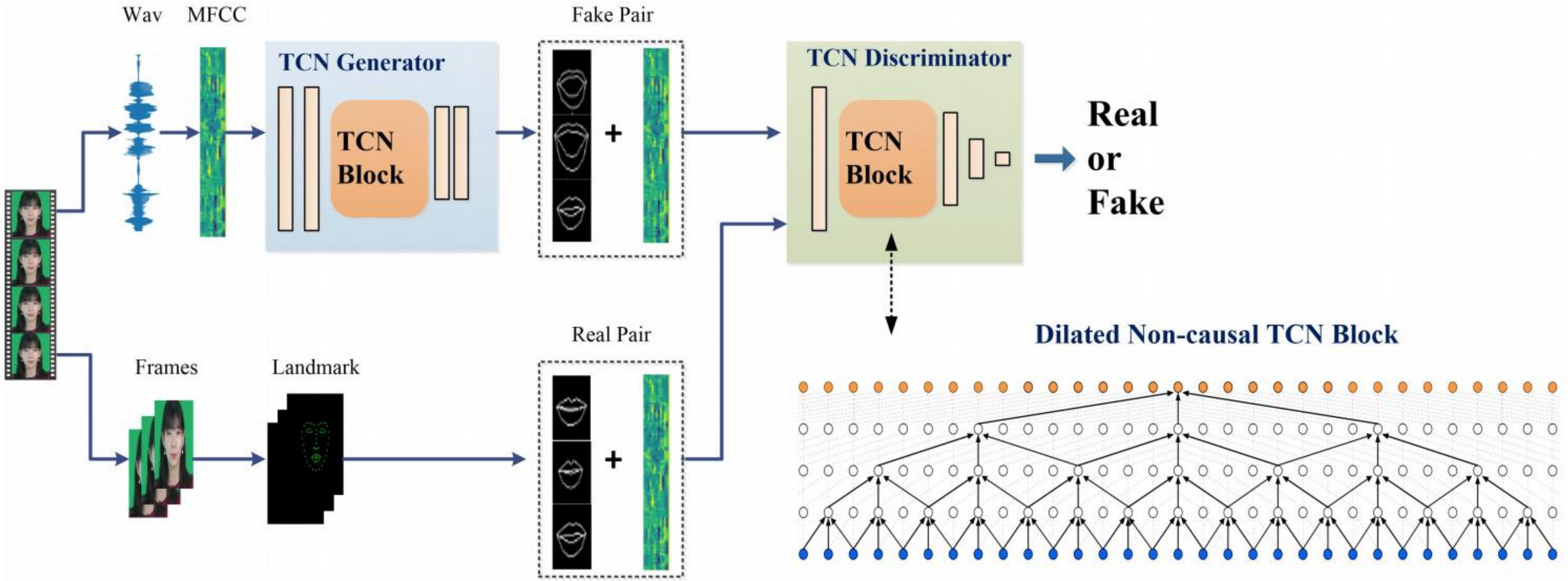
## Our Solution

1. A pair of **Temporal Convolutional Networks(TCN)** learning the seq-to-seq mapping from audio signals to lip motion

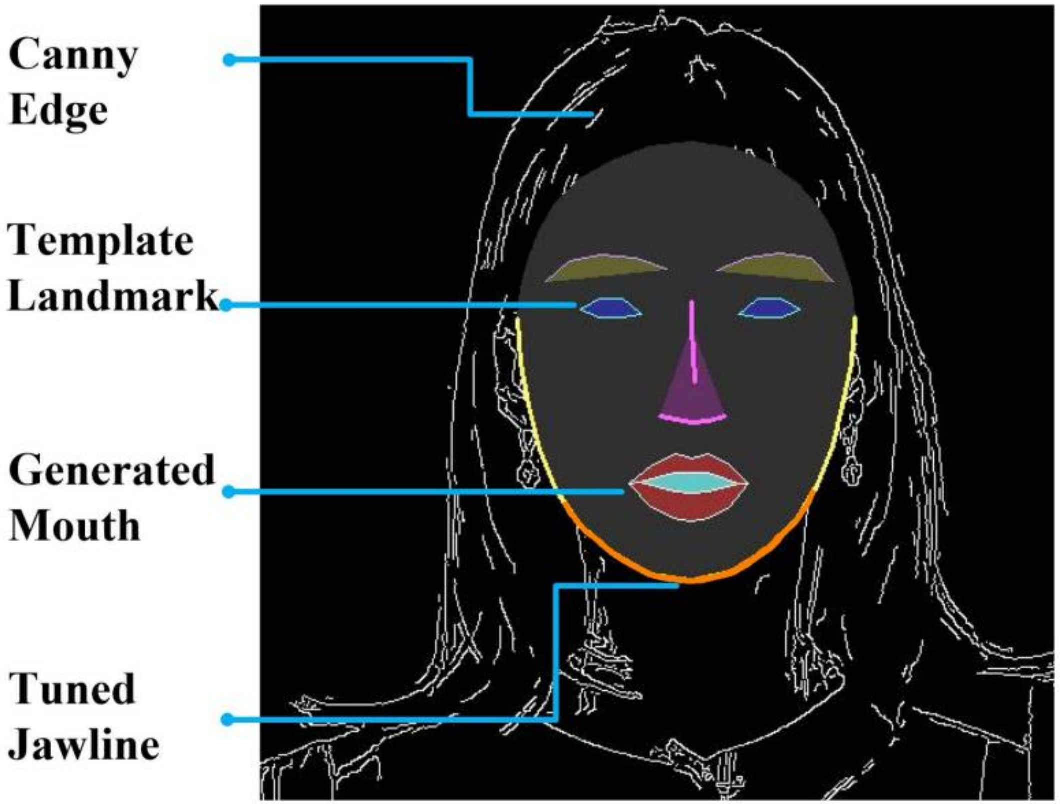
2. An image-to-image translation-based **neural rendering model** converts synthetic face maps to high-resolution and photorealistic video frames

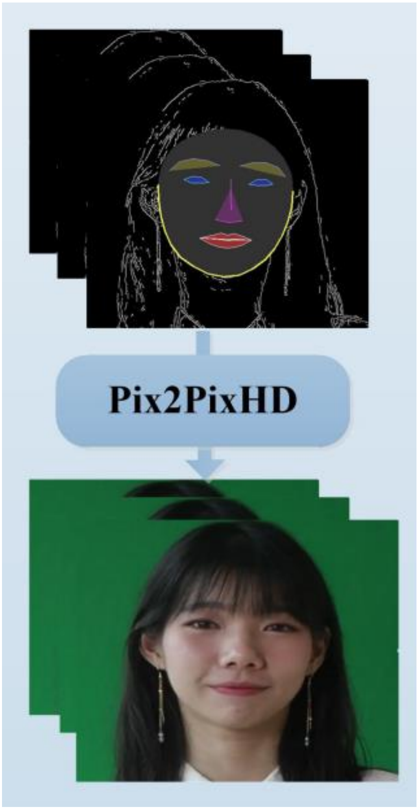

Mouth Feature
TCN
MFCC Spectrum
Waveform
Audio-to-Mouth
Synthetic Facial Map
Pix2PixHD
Face Rendering
Face Template
Edge
Landmark
Target Video
Template Processing

1. Learning Audio-to-Mouth Mapping



2. Neural Rendering



# **Experiments:** Audio-to-Mouth stage

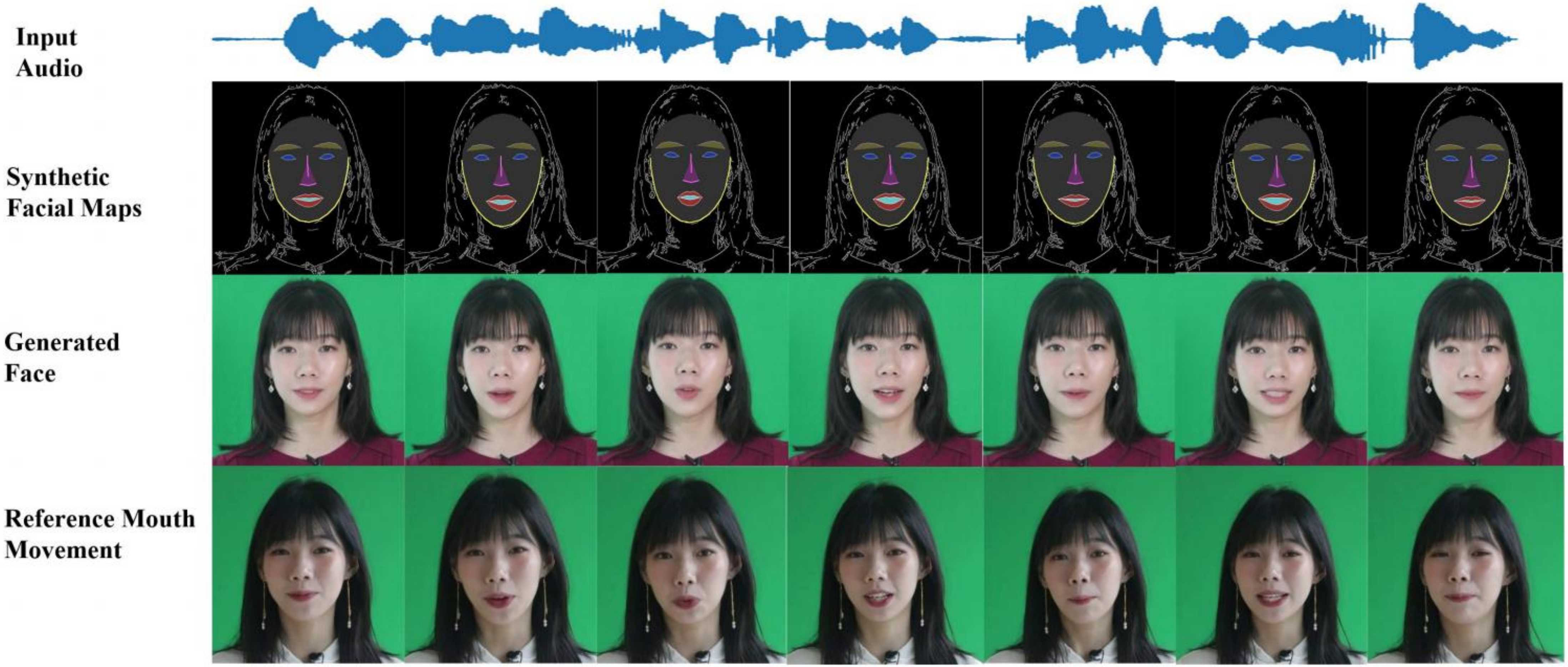
TABLE I
COMPARING THE PERFORMANCE OF AUDIO-TO-MOUTH MAPPING BETWEEN THE PROPOSED MODEL AND BASELINES.

| Model | MSE | MAE | Int-MSE |
|---|---|---|---|
| Time-delayed LSTM | 0.00366 | 0.0465 | 0.00735 |
| Bi-LSTM | 0.00357 | 0.0458 | 0.00712 |
| Non-Causal TCN | 0.00155 | 0.0278 | **0.00122** |
| Adversarial TCN (our) | **0.00141** | **0.0261** | 0.00132 |

TABLE II
COMPARING THE TRAINING AND INFERENCE TIME (1-MIN AUDIO) BETWEEN LSTM, BIDIRECTIONAL LSTM, AND TCN.

| Models | Batch training (s) | Total training (min) | Inference time (s) |
|---|---|---|---|
| LSTM | $0.069 \pm 0.005$ | $67.43 \pm 5.62$ | $2.272 \pm 0.269$ |
| Bi-LSTM | $0.124 \pm 0.007$ | $114.58 \pm 3.76$ | $3.376 \pm 0.201$ |
| TCN | $\mathbf{0.068 \pm 0.005}$ | $\mathbf{35.82 \pm 2.62}$ | $\mathbf{0.011 \pm 0.005}$ |

# **Experiments:** Rendering stage



# Thanks for watching our presentation!

If you are interested in our work, please contact

zrb915@gmail.com