# TALKINGFLOW: TALKING FACIAL LANDMARK GENERATION WITH MULTI-SCALE NORMALIZING FLOW NETWORK

*Sen Liang*[⋆†] *Zhize Zhou*[⋆] *Rong Li*[†] *Juyong Zhang*[‡] *Hujun Bao*[⋆†]

[⋆] State Key Lab of CAD&CG , Zhejiang University
[†] Zhejiang Lab [‡] University of Science and Technology of China

## ABSTRACT

Deterministic models dominate the field of talking facial landmark generation by directly mapping speech signals to a certain lip-sync facial landmark sequence, which often suffer from regression to the mean face. In contrast, probability generative models are more beneficial to handle complex data space and generate diverse samples. In this work, we propose a flow-based probabilistic network named TalkingFlow to generate natural talking facial landmark with head movements from speech data. It is implemented by a weighted multi-scale architecture to improve model representation capability and a conditional Temporal Convolutional Network module to fuse speech data. Extensive experiments results show that it can effectively generate diverse and natural facial landmark from speech data. All code will be made publicly available online.

***Index Terms***— Talking Head, Facial Landmark, Generative Model, Motion Synthesis, Normalizing Flow

## 1. INTRODUCTION

Talking facial landmark generation task is aiming to synthesize facial landmark from speech signals, which has been found in the core of various applications, such as talking head [1, 2], avatar animation [3] and VR/AR applications [4]. Although there exists a huge amount of talking videos in the wild that can be used for training data-driven models, it remains a challenging task to generate natural talking facial landmark, as facial landmark is a high-level face representation including both speech-related information (*e.g.*, lips movement and expression) and speech-irrelevant information (*e.g.*, identity and head pose).

Recent proposed talking facial landmark generation methods are deterministic, which can only predict a certain sequence of facial landmark from one speech clip. Eskimez *et al.* firstly proposed a multi-layer LSTM network to generate talking facial landmark from speech data [5] with a data preprocessing of aligning landmarks to a fixed head pose and eye coordinates. Chen *et al.* proposed an ATNet model to transfer audio data to facial landmark represented by PCA components in [6] and further added an additional head-pose learner from speech in [1]. Ji *et al* extended the above work to an EVP model to generate landmarks with content encoding and emotional encoding of speech [2]. Although these methods can produce good sequences of lip-sync facial landmark, they are still limited to generate diverse samples and failed to capture natural facial landmark with head movement.

To address the above mentioned shortcomings of deterministic models, we introduce a probabilistic generative model to make talking facial landmark generation conditioned on speech signal, which is capable of modeling the whole space of speech-related and speech-irrelevant information of the facial landmark. Basically, there are three kinds of probabilistic methods for sequence generation. GANs are effective for data manifold modeling but often suffer from mode collapse and are failed to generalize properly [7]. VAEs often model sequence data with an auto-regression framework but are still subjected to posterior collapse [8]. Unlike the above two kinds, Normalizing Flows compose a chain of inverse functions to directly maximize the data likelihood, which avoid the above shortcomings and are convenient for training [9, 10]. To our knowledge, flow-based models have been explored in skeleton-based human motion synthesis [11], video generation [12], speech synthesis [13] *etc.*, but have not received comparatively much attention on talking facial landmark generation yet.

In this paper, we propose a model named TalkingFlow to extend flow-based generative models into the setting of talking facial landmark generation. The model is designed with a weighted multi-scale architecture to model the different variance of each scale data distribution, where the network of each scale includes multiple normalizing flow steps. Moreover, we introduce a conditional Temporal Convolutional Network (ConditionTCN) module in each flow step to fuse speech signal and latent variables time by time. We summarize our contributions as follows: (1) We propose an end-to-end flow-based model called TalkingFlow, which is the first probabilistic model for the task of talking facial landmark generation. (2) We propose a weighted multi-scale architecture and apply a ConditionTCN module to improve model representation capability. (3) Extensive qualitative and quantitative experimental results demonstrate the effectiveness of our method.

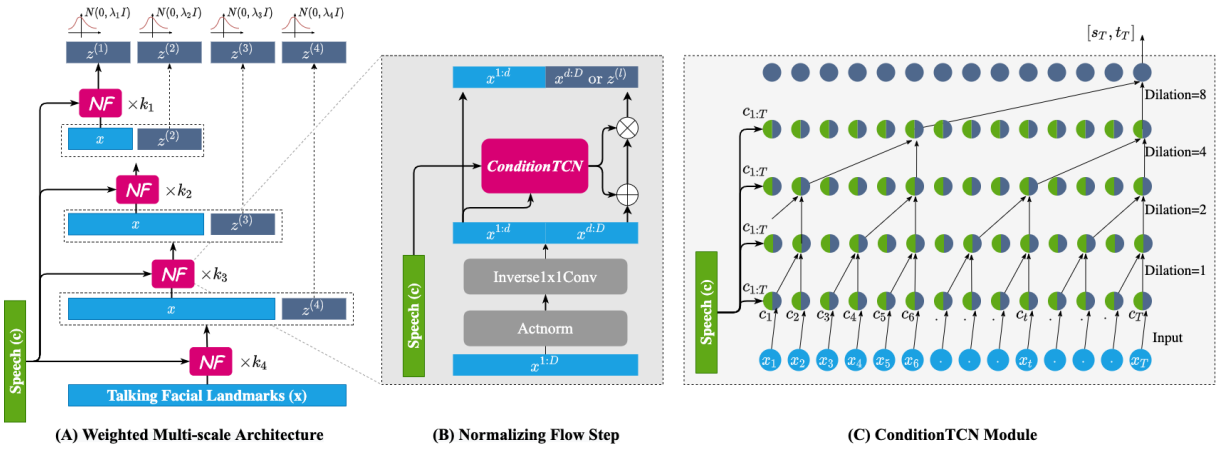


**Fig. 1**. Our proposed TalkingFlow model architecture.

## 2. TALKINGFLOW

The architecture of our proposed TalkingFlow model is illustrated in Fig. 1. A weighted multi-scale architecture containing a series of normalizing flow steps constitutes the main body of our model, where Glow network is adopted for each step [10]. Furthermore, a ConditionTCN module is integrated into each Glow step for fusing speech signal and latent variable. By combining Glow and ConditionTCN, a hierarchical bijection that maps multi-scale Gaussian Noise to our desired talking facial landmark is established in our model, which makes our model capable of generating lip-sync talking facial landmark with natural head movements from speech data.

Considering a speech sequence $c = \{c_1, c_2, ..., c_T\}$ with $T$ time-steps and its corresponding sequence of facial landmark $x = \{x_1, x_2, ..., x_T\}$, the probability of generating such a sequence of landmarks $x$ from $c$ can be factorized as a product of conditional probabilities as

$$p(x|c) = \prod_{t=1}^{T} p(x_t|x_{<t}, c_{\leq t}) \tag{1}$$

where the facial landmark $x_t$ at time $t$ are conditioned on all the previous landmarks $x_{<t}$ before $t$ and speech signals $c_{\leq t}$ before and at $t$.

We model this conditional probability density function $p(x|c)$ with a normalizing flow method that maps an unknown complex distribution to a simple distribution [14]. First, we suppose that there is a bijective map $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^d$ that can generate $x$ from a random variable $z \sim p(z)$ as $x = \phi(z, c)$ and recover $z$ by its inverse map $\phi^{-1}$ as $z = \phi^{-1}(x, c)$. Under this assumption, for each time $t$, $x_t|_{x_{<t}, c_{\leq t}} = \phi(z_t, c_{\leq t})$, and therefore we derive that

$$p(x_t|x_{<t}, c_{\leq t}) = p(z_t)|\det\frac{\partial\phi}{\partial z_t}|^{-1} \tag{2}$$

with the change of variable theory and assuming that the prior distribution $p(z_t)$ is independent of the speech signal.

In practice, $\phi$ is composed of a chain of $K$-step invertible transformations (*i.e.*, normalizing flow steps), the relation between $x$ and $z$ can then be represented as

$$x_t = \phi_K(\phi_{K-1}(...\phi_1(z_{t,0}, c_{\leq t}))) \tag{3}$$

$$z_{t,0} = \phi_1^{-1}(\phi_2^{-1}(...\phi_K^{-1}(x_{<t}, c_{\leq t}))) \tag{4}$$

where $z_{t,0}$ denotes the latent variabe at the beginning of the chain (*i.e.*, the 0-th step) at time $t$. Besides, we define that $z_{k+1} = \phi_k(z_k, c)$, and then $z_{t,0} = z_t$ and $z_{t,K} = x_t$. Equ. 2 can thus be extended to the following

$$\log p(x_t|x_{<t}, c_{\leq t}) = \log p(z_{t,0}) - \sum_{k=1}^{K} |\log\det\frac{\partial\phi_k}{\partial z_{t,k}}|, \tag{5}$$

### 2.1. Weighted Multi-Scale Architecture

The multi-scale architecture in our model is designed to realize the bijective map $\phi$, where the latent variable $z$ is constructed as a stack of variables in different scales as $z = cat[z^{(1)}, z^{(2)}, ..., z^{(L)}]$, and each variable $z^{(l)}$ is obtained by a network with $k_l$ normalizing flow steps as shown in Fig. 1(A). The variable $z^{(l)}$ at each scale $l$ encodes the distributions of part of talking facial landmark with speech condition, which is acquired from the last affine coupling layers of each scale network. The probability of $z^{(l)}$ at each scale can then be obtained from extending Equ. 1 as

$$p(z^{(l)}|c) = \prod_{t=1}^{T} p(z_t^{(l)}|z_{<t}^{(<l)}, c_{\leq t}) \tag{6}$$

where $z_{<t}^{(<l)}$ denotes the latent variable before time $t$ and scale $l$. Different from the models such as WaveGlow [13] and VideoFlow [12], our approach introduces a weighted variance for the distributions of each scale to improve its representation capability that $p(z^{(l)}) \sim N(0, \lambda_l I)$, where $\lambda_l$ is the weight for the variance of the latent data distribution at scale $l$.

For each normalizing flow step, Glow model [10] is adopted as illustrated in Fig. 1(B), which is convenient for incorporating conditional speech signals into the affine coupling layer [9] with a conditional neural network. In the affine coupling layer, the input $z_{t,k}^{(l)}$ with $D$ dimensions is split into $z_{t,k}^{1:d}$ and $z_{t,k}^{d:D}$ along the channel axis with $d = D/2$, and $z_{t,k+1}^{(l)}$ is updated by

$$[z_{t,k}^{1:d}, z_{t,k}^{d:D}] = split(z_{t,k}^{(l)}) \tag{7}$$

$$[s_{t,k}, t_{t,k}] = ConditionNet(z_{t,k}^{1:d}, c_t) \tag{8}$$

$$z_{t,k+1}^{(l)} = [z_{t,k}^{1:d}, \exp(s_{t,k}) \odot z_{t,k}^{d:D} + t_{t,k}] \tag{9}$$

where $\odot$ is Hadamard product, and $ConditionNet$ is the conditional network to fuse speech data into backbone, which will be described in detail in the next subsection. Specifically, in the last affine coupling layer of scale $l$, the output of Equ. 9 will be divided into a higher-scale variable $z_{t,k+1}^{(l+1)}$ as the input of next Glow step and a new high-level variable $z_t^{(l+1)}$ for weighted variance distribution modelling, which can be formulated as

$$[z_{t,k+1}^{(l+1)}, z_t^{(l+1)}] = [z_{t,k}^{1:d}, \exp(s_{t,k}) \odot z_{t,k}^{d:D} + t_{t,k}] \tag{10}$$

### 2.2. ConditionTCN Module

The mathematical property of the affine coupling layer in Glow model that $\log\det(\partial\phi_k/\partial z_k) = sum(\log(|s|))$ makes the $ConditionNet$ of Equ. 8 unnecessary to be invertible [9], which means that it can be any neural network. In this work, we implement a Conditional Temporal Convolutional Network (ConditonTCN) as the $ConditionNet$ to fuse conditional speech data and latent variables, which is illustrated in Fig. 1(C). This module stacks four 1-D dilated convolution layers along the time axis with 1, 2, 4, and 8 dilations respectively, and integrates causal convolutions to make sure the output $h_t$ is independent of any of the future landmarks and speech information at each time-step. In each layer, we merge latent variables with conditional speech data timestep-wisely by a gated activation unit similar to PixelCNN [15] and WaveGlow [13].

### 2.3. Loss Functions

The loss function of our model is to minimize the negative log-likelihood of talking facial landmark sequence by extending Equ. 5 to a multi-scale setting as

$$\min -\log p(x|c) = \tag{11}$$
$$-\frac{1}{L \times T}\sum_{t=1}^{T}\Big[\sum_{l=1}^{L}\log p(z_t^{(l)}) + \sum_{k=1}^{K}\sum_{l=1}^{L}|\log\det\frac{\partial\phi_k}{(l)}\partial z_{t,k}^{(l)}|\Big]$$

where $p(z_t^{(l)}) \sim N(0, \lambda_l I)$ with standard variance $\lambda_l$ is the distribution of the output at each scale $l$.

## 3. EXPERIMENTS

Our model is trained on the Obama dataset [16] which consists of 300 Obama weekly addresses videos. Each video is about 3 minutes on average, resulting in 17 hours in total. And we evaluated our model on both Obama dataset and GRID dataset [17]. In the data pre-processing stage, we first extract the 2D facial landmark of each video frame with an off-the-shelf method [18] and then normalize the coordinate values to $0 \sim 1$ for stable training. In the meantime, we split the synchronized audio signal from each video, and extract its mel-spectrogram feature as the condition speech data.

This algorithm is implemented with Pytorch library [19]. All experiments are evaluated on Ubuntu with one NVIDIA V100 GPU and CUDA 12.0 and takes about two days with about 100 epochs in training processing. The batch size is set to 128 with the learning rate of $10^{-4}$ by the Adam optimization method. By conducting hype-parameters grid search experiments, the number of scales $L$ of the multi-scale architecture is set to 4, and the weights $\lambda_1 \sim \lambda_4$ for the four scales are set to 0.25, 0.5, 0.75 and 1.0 respectively.

### 3.1. Quantitative Results

We quantitatively evaluate the effectiveness of our model by comparing with the following four state-of-the-art methods: (1) **Eskimez *et al.*** proposed multiple LSTM layer network which maps speech log-mel spectrogram feature to facial landmark [5]. (2) **Chen *et al.*** proposed AT-Net which transfers MFCC features to the PCA components of facial landmark [6]. (3) **Ji *et al.*** proposed EVP network which generates motional facial landmark with speech content and emotional latent code [2]. (4) **Lu *et al.*** presents a live system to generate personalized photo-realistic talking-head animation [20].

Three metrics are utilized to evaluate the performances of different methods: (1) **Mean Opinion Score (MOS)** is an investigation of people's opinions by constructing a user study to assess the lip-audio synchronicity and the quality of generated motions, where the designed rating score for each video is set as 1-bad, 2-poor, 3-fair, 4-good, 5-excellent. (2) **SyncNet Score**, which is modified from the SyncNet model proposed by Chung *et al.* [21], is an indicator to measure the facial lip-audio synchronicity by computing the similarity of lip-part facial landmark's embedding and its corresponding speech signal's embedding. (3) **Mouth Open Rate Distance ($L_{MOR}$)**, which is the distance between the inferred mouth open rate and GT in each frame [22].

The comparison result on Obama Dataset is shown in Table 1, which shows that our method achieves the best performance on all the three metrics. We further evaluate our method on the GRID dataset [17], which's results shown in Table 2. Our method is optimal in SyncNet Score and performs better than Ji *et al.* and Lu *et al.* in $L_{MOR}$. But ours is slightly worse than Eskimez *et al.* and Chen *et al.* in $L_{MOR}$.

| Methods | $MOS \uparrow$ | $SyncNet \downarrow$ | $L_{MOR} \downarrow$ |
|---|---|---|---|
| Eskimez *et al.* | 3.3184 | $0.8752 \pm 0.3388$ | 0.0852 |
| Chen *et al.* | 3.2275 | $1.1430 \pm 0.5814$ | 0.1798 |
| Ji *et al.* | 2.8377 | $1.3112 \pm 0.4850$ | 0.0891 |
| Lu *et al.* | 3.4572 | $0.6308 \pm 0.2314$ | 0.0857 |
| *Baseline* | – | $1.0183 \pm 0.27934$ | 0.1021 |
| + *ConditionTCN* | – | $0.3940 \pm 0.2754$ | 0.0916 |
| Our Full Model | **3.5921** | $\mathbf{0.2978 \pm 0.1586}$ | **0.0441** |
| *Ground Truth* | 3.9605 | $0.1785 \pm 0.1046$ | 0.0000 |

**Table 1**. Quantitative comparison result and ablation studies on Obama dataset.

The probable reason can be that the models of Eskimez *et al.* and Chen *et al.* were trained on GRID dataset, but ours was trained on Obama dataset without fine-tuning. This actually reflects a good generalization ability of our method.

### 3.2. Qualitative Results

The qualitative results of comparing our method with above-mentioned competing methods and ground truth (GT) are shown in Fig. 2, which shows that our method is able to generate high-quality talking landmarks with better lip synchronicity and natural head movement. The mouths of Ji *et al.* are all open. The lip movements of Eskimez *et al.* are much less than Chen *et al.* and ours. It is obvious that ours is closest to the ground truth. Specifically, only ours can reflect the pouting action as shown at time-step $t = 43$.


**Fig. 2**. Qualitative experiments result of comparison method.

### 3.3. Ablation Study

As our model is benefited from using both the weighted multi-scale architecture and the ConditionTCN module, a quantitative ablation study is conducted to explore their significance, and the result is shown in Table 1. The $Baseline$ model is without multi-scale architecture and ConditionTCN module, and the $+ConditionTCN$ model only includes ConditionTCN module. The result shows that even $Baseline$ has already got a better performance than Chen *et al.* and Ji *et al.* which proves the effectiveness of normalizing flow working on this task. Besides, the SyncNet score is significantly reduced by $+ConditionTCN$, which reveals that the ConditionTCN plays a more important role in our model.

| Methods | $SyncNet \downarrow$ | $L_{MOR} \downarrow$ |
|---|---|---|
| Eskimez *et al.* | $0.5879 \pm 0.2185$ | 0.0774 |
| Chen *et al.* | $0.5361 \pm 0.2429$ | **0.0747** |
| Ji *et al.* | $1.1486 \pm 0.1962$ | 0.1847 |
| Lu *et al.* | $0.9761 \pm 0.2404$ | 0.1548 |
| Our Model | $\mathbf{0.3591 \pm 0.1302}$ | 0.0918 |
| *Ground Truth* | $0.1876 \pm 0.1212$ | 0.0000 |

**Table 2**. Quantitative comparison result on GRID dataset.

## 4. DISCUSSION

Our work presented here is focused on talking facial landmark generation from speech signal, which is a core problem in the talking-head research field. Existing approaches solved it as a regression problem with supervised model, like the works mentioned above [5, 6, 2], which often suffer from regression to the mean face and cannot handle the speech-irrelevant information properly. In this work, we are the first to employ probabilistic generative model to address this problem with flow-based method, which has more advantages than GANs and VAEs in the training stage.

Multi-scale architecture is a common practice in mostly recent flow-based approaches, such as Real-NVP[9], WaveGlow [13] and VideoFlow [12]. Real-NVP lets half of the variables be directly modeled as Gaussians using a squeezing operation. WaveGlow implemented it with an early outputs operator to output some dimensions early to add information at multiple time scales. VideoFlow added a latent prior for each scale variable. These works model the variables at each scale as the normal distribution. Therefore, we add weight for each scale to increase the representation ability of the model. Although this operation brings extra hyper-parameters, we can optimize it with adaptive multi-scale weighted in future work. At last, we believe that our flow-based facial landmark generation model will help in the development of talking-head, 2D avatar, and AV/VR research area.

## 5. ACKNOWLEDGMENTS

This work is supported by the State Key Lab of CAD&CG in Zhejiang University and the National Natural Science Foundation of China (Grant No.61210007). Thank the support and encouragement of my wife Huanhuan Qu, more than anything in this world I want to spend the rest of my life with you.

## 6. REFERENCES


[1] Lele Chen, Guofeng Cui, Celong Liu, Zhong Li, Ziyi Kou, Yi Xu, and Chenliang Xu, "Talking-head generation with rhythmic head motion," in *European Conference on Computer Vision*. Springer, 2020, pp. 35–51.

[2] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Wayne Wu, Chen Change Loy, Xun Cao, and Feng Xu, "Audio-driven emotional video portraits," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14080–14089.

[3] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li, "Makelttalk: speaker-aware talking-head animation," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 6, pp. 1–15, 2020.

[4] Naima Otberdout, Claudio Ferrari, Mohamed Daoudi, Stefano Berretti, and Alberto Del Bimbo, "3d to 4d facial expressions generation guided by landmarks," *arXiv preprint arXiv:2105.07463*, 2021.

[5] Sefik Emre Eskimez, Ross K Maddox, Chenliang Xu, and Zhiyao Duan, "Generating talking face landmarks from speech," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2018, pp. 372–381.

[6] Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu, "Hierarchical cross-modal talking face generation with dynamic pixel-wise loss," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7832–7841.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[8] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[9] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio, "Density estimation using real nvp," *arXiv preprint arXiv:1605.08803*, 2016.

[10] Diederik P Kingma and Prafulla Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," *arXiv preprint arXiv:1807.03039*, 2018.

[11] Gustav Eje Henter, Simon Alexanderson, and Jonas Beskow, "Moglow: Probabilistic and controllable motion synthesis using normalising flows," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 6, pp. 1–14, 2020.

[12] Manoj Kumar, Mohammad Babaeizadeh, Dumitru Erhan, Chelsea Finn, Sergey Levine, Laurent Dinh, and Durk Kingma, "Videoflow: A flow-based generative model for video," *arXiv preprint arXiv:1903.01434*, vol. 2, no. 5, 2019.

[13] Ryan Prenger, Rafael Valle, and Bryan Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3617–3621.

[14] Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka, "Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows," *ACM Transactions on Graphics (TOG)*, vol. 40, no. 3, pp. 1–21, 2021.

[15] Aaron van den Oord, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, and Koray Kavukcuoglu, "Conditional image generation with pixelcnn decoders," *arXiv preprint arXiv:1606.05328*, 2016.

[16] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman, "Synthesizing obama: learning lip sync from audio," *ACM Transactions on Graphics (ToG)*, vol. 36, no. 4, pp. 1–13, 2017.

[17] 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*, vol. 120, no. 5, pp. 2421–2424, 2006.

[18] Xiaojie Guo, Siyuan Li, Jinke Yu, Jiawan Zhang, Jiayi Ma, Lin Ma, Wei Liu, and Haibin Ling, "Pfld: A practical facial landmark detector," *arXiv preprint arXiv:1902.10859*, 2019.

[19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al., "Pytorch: An imperative style, high-performance deep learning library," *arXiv preprint arXiv:1912.01703*, 2019.

[20] Y. Lu, J. Chai, and X. Cao, "Live Speech Portraits: Real-time photorealistic talking-head animation," *ACM Transactions on Graphics*, vol. 40, no. 6, 2021.

[21] Joon Son Chung and Andrew Zisserman, "Out of time: automated lip sync in the wild," in *Asian conference on computer vision*. Springer, 2016, pp. 251–263.

[22] S. Liang, Z. Zhou, Y. Guo, J. Zhang, and H. Bao, "Facial landmarks disentangled network with variational autoencoder," *Applied Mathematics-A Journal of Chinese Universities*, 2021 (accepted).