Talking Head from Speech Audio using a Pre-trained Image Generator

Mohammed M. Alghamdi
University of Leeds
Taif University
scmmalg@leeds.ac.uk

He Wang
University of Leeds
h.e.wang@leeds.ac.uk

Andrew J. Bulpitt
University of Leeds
a.j.bulpitt@leeds.ac.uk

David C. Hogg
University of Leeds
d.c.hogg@leeds.ac.uk

Figure 1: Given a single identity image and speech audio, our model generates high-resolution talking-head video of the identity lip-synced with the audio.

ABSTRACT

We propose a novel method for generating high-resolution videos of talking-heads from speech audio and a single ‘identity’ image. Our method is based on a convolutional neural network model that incorporates a pre-trained StyleGAN generator. We model each frame as a point in the latent space of StyleGAN so that a video corresponds to a trajectory through the latent space. Training the network is in two stages. The first stage is to model trajectories in the latent space conditioned on speech utterances. To do this, we use an existing encoder to invert the generator, mapping from each video frame into the latent space. We train a recurrent neural network to map from speech utterances to displacements in the latent space of the image generator. These displacements are relative to the back-projection into the latent space of an identity image chosen from the individuals depicted in the training dataset. In the second stage, we improve the visual quality of the generated videos by tuning the image generator on a single image or a short video of any chosen identity. We evaluate our model on standard measures (PSNR, SSIM, FID and LMD) and show that it significantly outperforms recent state-of-the-art methods on one of two commonly used datasets and gives comparable performance on the other. Finally, we report on ablation experiments that validate the components of the model. The code and videos from experiments can be found at https://mohammedalghamdi.github.io/talking-heads-acm-mm/

CCS CONCEPTS
• Computing methodologies → Computer vision; Animation; Rendering; Image-based rendering.
KEYWORDS
talking head generation, video generation, audio-driven synthesis

ACM Reference Format:
Mohammed M. Alghamdi, He Wang, Andrew J. Bulpitt, and David C. Hogg. 2022. Talking Head from Speech Audio using a Pre-trained Image Generator. In Proceedings of the 30th ACM International Conference on Multimedia (MM ’22), October 10–14, 2022, Lisboa, Portugal. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3503161.3548101

1 INTRODUCTION

Synthesising video of talking heads from speech audio has many potential applications, such as video conferencing, video animation production and virtual assistants. Although there has been considerable prior work on this task, the quality of generated videos is typically limited in terms of overall realism and resolution. In this paper, we propose an audio-driven model that synthesises high-resolution talking-head videos (1024 x 1024 in our experiments) from a single identity image.

Many previous models generate low-resolution video [31, 32] or cropped faces [3, 4]. Low resolution video is generally not suitable for deployment in many real world applications, such as a virtual assistant. A common approach has been to use intermediate features such as facial landmarks to map from audio to output video [3, 7, 41]. Another approach has been to edit an existing talking-head video to blend in a new mouth region synthesised from audio [28, 29].

Recent advances in image synthesis have been successful at generating high-resolution images from noise [12–14]. Karras et al. [14] propose a style-based generator StyleGAN that synthesises high quality images that are largely indistinguishable from real ones. Some works have studied the latent space of StyleGAN [1, 9, 22, 24] and discovered meaningful semantics for manipulating images. Recent work has leveraged the richness of a pre-trained StyleGAN generator [15] to generate high-resolution videos from noise by decomposing (disentangling) the motion and content in the latent space [8, 26, 30]. Tian et al. [30] discover motion trajectories in the latent space to render high-resolution videos while image and motion generators are trained on different domain datasets.

Inspired by these advances, we propose a novel method for generating high-resolution videos conditioned on speech audio by constructing trajectories in the latent space of a pre-trained image generator [15]. Our framework uses a pre-trained image encoder [21] to find the latent code of a given identity image in the latent space of the generator. We then train a recurrent audio encoder along with a latent decoder to predict a sequence of latent displacements to the encoded identity image. In this stage, we show our approach can generate talking-head videos with accurate mouth movements conditioned on speech audio. To improve the visual quality of the generated videos further, we tune the generator on a single image or short video of a target subject using the PTI [22] method. We compare our approach with other state of the art approaches qualitatively and quantitatively using benchmark measures: LMD, SSIM, PSNR and FID. We show that it achieves performance at least as good as the state of the art on two commonly used datasets.

Our principal contributions are:
- A method for generating high-resolution videos from speech audio by constructing motion trajectories in the latent space of a pre-trained image generator;
- A comparative evaluation, including a user study, demonstrating the performance of the method on both quantitative and qualitative criteria.

2 RELATED WORK

2.1 Audio-driven talking-head generation

Various methods have been proposed to generate videos of talking heads. Given audio, the task is to lip-sync the head to the audio. The audio itself may be generated automatically from text or be extracted from a video clip of someone speaking. Some of these methods are generic, and can generate videos of any identity given one or more images of that identity [3, 4, 20, 31, 32, 39–41]. However, they can only generate low-resolution videos. Other methods that generate high-resolution videos from speech audio [11, 17, 18, 27–29, 34, 36, 38] can only work on a single subject. These approaches require retraining the models for each new subject.

Chung et al. [4] propose a model to generate videos using multiple images of the target face and an audio speech sample. The model consists of an audio encoder and identity encoder that learn a joint embedding of the face and audio, and a decoder that generates a frame that best represents the audio sample for the target identity. Prajwal et al. [20] adopt a similar approach, except that a pretrained lip-sync discriminator, and a visual discriminator are used in addition to $L_1$ loss. Vougioukas et al. [31] expand the approach by training a recurrent-based decoder with a noise generator to model spontaneous facial expressions (e.g. blinks). Other models rely on 2D intermediate features (e.g. facial landmarks) to learn the mapping between audio input and video output [3, 7, 41]. Chen et al. propose a cascade approach that generates a talking face video given an image and audio. The method first transfers the given audio signal to facial landmarks and then generates video frames conditioned on the landmarks. Zhou et al. [41] adopt a similar approach that first disentangles the content and speaker information from the input audio. Then, these two components are mapped to content and speaker facial landmark spaces using a recurrent model on each.

Other methods rely on 3d intermediate features (e.g. through monocular reconstruction) to synthesise high-quality videos of a single subject. Some attempt to generate only the mouth region and blend it to a target video [17, 27–29]. Thies et al. [29] propose an approach that predicts the coefficients that drive a person-specific expression blendshape basis using audio features. A neural texture rendering network is then used to generate the mouth region. In addition, others modify the facial expressions, geometry or pose of a target video conditioned on the audio [11, 36]. In contrast, our
approach generates a full-frame video of a talking-head without editing a target video or relying on intermediate features.

### 2.2 Unconditional video generation using StyleGAN

Recently, there have been several works that use a pre-trained image generator (StyleGAN) to generate videos \cite{8, 26, 30}. They all share the idea of discovering motion trajectories in the latent space of a StyleGAN generator without conditioning on any driving source. Tian et al. \cite{30} propose a MOCOGAN-HD model that uses a motion generator to predict residual latent codes from an initial latent code sampled from the latent space of StyleGAN. The model is trained in the image space with a multi-scale video discriminator as well as contrastive image discriminator. Similarly, Fox et al. \cite{8} reduce the training cost by training a Wasserstein GAN model in the latent space instead of image space. Although their model is trained on a single subject dataset, it can transfer the learned motion to a new subject using an offset trick. Skorokhodov et al. \cite{26} modify the StyleGAN network to learn a continuous latent trajectory using a neural representation based approach. Our work differs from these methods by learning the motion trajectories in the latent space of StyleGAN conditioned on speech audio.

### 3 THE METHOD

Our method consists of four components: image encoder $E_I$, audio encoder $E_A$, latent decoder $D$ and image generator $G$. Given an identity image $I$ and speech audio $a$, the goal is to synthesise a video of the identity lip-synced with the audio. We first partition the audio clip into a sequence of $T$ fixed-duration audio segments $\{a_1, a_2, ..., a_T\}$. From this audio segment sequence, we target a video clip consisting of a sequence of $T$ video frames $\{x_1, x_2, ..., x_T\}$. There is therefore a one-to-one correspondence between input audio segments and output video frames.

The inference pipeline is as follows. We take the identity image $I$ and finds its latent code $\omega_I$ in the latent space $\mathbb{W}^+$ of the generator $G$ using the image encoder $E_I$. Next, we encode an audio segment $a_t$ using the audio encoder $E_A$ and extract an audio embedding vector $e_t = E_A(a_t)$. We then feed in the identity latent code $\omega_I$ and the audio vector $e_t$ jointly to the latent decoder $D$ to predict a latent displacement $d_t = D(\omega_I, e_t)$. We then calculate the displaced latent vector $\tilde{\omega}_t = \omega_I + d_t$ and generates the corresponding video frame $\tilde{x}_t$. An overview of the model can be seen in figure 2. In the following section, we describe each component in detail.

#### 3.1 Architecture

##### 3.1.1 Image generator

For our image generator $G$, we use the pre-trained StyleGAN \cite{15} trained on the FFHQ dataset \cite{14} to synthesise static images of faces. The StyleGAN architecture consists of mapping and synthesis networks. The mapping network is a non-linear 8-layer MLP which maps a latent code $z$ sampled from a latent space $\mathbb{Z}$ to an intermediate space $\mathbb{W}^+$. The produced $w$ controls the synthesis network through an adaptive instance
Video

Inverted video

Our model is trained in two stages. In stage one, we are only interested in learning trajectories in the latent space $W^+$ conditioned on the speech audio. The model predicts latent displacements to trajectories are learned by training the model using a talking-head. Disentangles the mouth motion and the image content. The motion is tested in learning trajectories in the latent space $W^+$. Our model is trained in two stages. In stage one, we are only inter-

3.2 Training

We represent each audio segment $a_t$ using Mel-frequency cepstral coefficients (MFCC). The audio encoder $E_A$ is a recurrent model that takes a sequence of audio segments $\{a_1, a_2, ..., a_T\}$ and produces a sequence of encoded audio segments $\{e_1, e_2, ..., e_T\}$. The encoder network $E_A$ consists of multiple convolutional layers followed by three LSTM layers.

3.1.4 Latent decoder. The latent decoder $D$ takes as input the identity latent vector $w_I$ and an encoded audio segment $e_t$ to predict a displacement $d_t$ to the identity vector in the latent space $W^+$. To reduce the high dimensionality of the latent space $W^+$, we first conduct principal component analysis (PCA) on the FFHQ dataset [14] mapped into $W^+$ using the image decoder. We obtain a subspace from the components with the largest eigenvalues, giving a basis $V$. We project the identity latent $w_I$ input into the subspace defined by $V$ and concatenate it with the encoded audio segment $e_t$. This provides the input to the decoder. We map the decoder’s output $h_t$ from the subspace to $W^+$ to get the displacement vector $d_t$. Thus, we obtain $\tilde{w}_t$ as follows:

$$\tilde{w}_t = w_I + d_t = w_I + h_t \cdot V, \quad t = 1, 2, 3, \ldots, T,$$  

where $\phi$ is based on the VGG neural network [25]. Thus, the overall loss for learning to predict latent displacements driven by audio is a weighted sum of the two losses:

$$L_{\text{stage1}} = \lambda_{\text{l2}}L_{\text{l2}} + \lambda_{\text{lpips}}L_{\text{lpips}}$$

where $\lambda_{\text{l2}}$ and $\lambda_{\text{lpips}}$ are hyperparameters.

3.2.2 Stage two. One could tune both the image encoder $E_I$ and the image generator $G$ on the target speaker using an auto-encoder. However, this would transform the latent space $W^+$ and the learned model at stage one would consequently fail to generate correct mouth movements. To improve the visual quality of the generated video, we use the PTI method [22] to tune only the generator on a single image or video of a target speaker. In experiments, we implement stage two on short videos, and as a limiting case, on a single image.

Given a video $x$ of a speaker, we tune the image generator $G$ on the video frames $x_t$. For this task, we use the pre-trained encoder $E_I$ to encode the frames $x_t$ to the latent space $W^+$ to get $\bar{w}_I$. Given $\bar{x}_t = G(\bar{w}_I; \theta^*)$, we tune the weights of the generator while keeping the encoder fixed. We use the same objective loss used in PTI [22]:

$$L_{\text{stage2}} = L_{\text{lpips}} + L_{\text{l2}}$$

Figure 3: This figure illustrates how the image encoder $E_I$ can encode a video (top row) into the latent space of StyleGAN $W^+$ while preserving its mouth movements and other facial expressions. Bottom row shows the inverted video.
4 EXPERIMENTS

4.1 Datasets
We evaluate our approach using two widely used datasets for synthesizing talking-head videos: GRID [5] and TCD-TIMIT [10]. The GRID dataset has 33 speakers uttering 1000 short sentences each containing 6 words. The TCD-TIMIT has 59 speakers each uttering 100 sentences. We hold-out ten speakers from each dataset for testing and use the remaining for training. The videos are resampled to 25 fps. To align the video frames, we use the same face alignment method used in preprocessing the FFHQ dataset [14] for training the original StyleGAN [15]. The input to the audio encoder $E_A$ is an audio segment of length 0.2 seconds which corresponds to a window of five frames. However, we choose only the middle frame as the ground truth frame. The identity image is a randomly chosen frame out of this window. We represent the audio speech using MFCC values extracted from the raw values. Each audio segment is a window of size 12 x 28, where the columns represent MFCC features for each time step.

4.2 Implementation details
We perform our experiments using PyTorch [19]. For the image generator, we use an unofficial implementation of StyleGAN1. For the pre-trained image encoder, we use the official implementation of p2p [21]. For training the audio encoder $E_A$ and the decoder $D$ in stage one, we use an Adam optimiser [16] with a learning rate of 0.0002. In Eqn. 4, we set $\lambda_{\text{latent}} = 250$ and $\lambda_{\text{LPIPS}} = 1$. We tune the generator with a learning rate of 0.0003. The tuning process takes less than two minutes for a single identity image. All experiments use an NVIDIA V100 GPU with 32 GB of memory.

4.3 Results
In this section, we evaluate our model after tuning the generator. Figure 4 shows the quality of the generated videos from tuning the generator on a single frame (middle row) and on a short video (bottom row). The figure shows that tuning the generator on multiple frames has resulted in better visual quality. This can be seen in the mouth appearance highlighted in red.

To evaluate the quality of the generated videos, we use two common reconstruction measures: The peak signal-to-noise ratio (PSNR) and the structural similarity (SSIM) [35]. For these measures, a larger score is better. We use a landmarks distance metric (LMD) [2] to evaluate the synchronisation between the mouth movement and the speech audio. This metric computes the Euclidean distance between mouth landmarks of each generated frame and its corresponding true frame. It then averages the score on the number of frames and number of mouth landmark points. We also use a Fréchet Inception Distance (FID) to quantitatively evaluate generated videos. For LMD and FID measures, a lower score is better.

We compare our work against four state of the art models [3, 20, 31, 41] and use the official available codes of these models to generate the videos and compute the evaluation measures. Table 1 shows that our approach outperforms other state of the art models on the TCD-TIMIT dataset [10]. On the GRID dataset [5], the model achieves better scores on the PSNR and FID measures. Figure 5 shows the visual quality of generated videos by our model in comparison with other models. The highlighted frames show that our model generates photo-realistic videos largely indistinguishable from the ground truth.

We compare our work against four state of the art models [3, 20, 31, 41] and use the official available codes of these models to generate the videos and compute the evaluation measures. Table 1 shows that our approach outperforms other state of the art models on the TCD-TIMIT dataset [10]. On the GRID dataset [5], the model achieves better scores on the PSNR and FID measures. Figure 5 shows the visual quality of generated videos by our model in comparison with other models. The highlighted frames show that our model generates photo-realistic videos largely indistinguishable from the ground truth.

Figure 6 shows a visual comparison of our model with others on a challenging mouth movement associated with the phoneme /p/ in the word "place". It can be seen that our model and Vougioukas et al. produce a closed-mouth shape (highlighted in green) in sync...
Figure 5: Qualitative comparisons. The videos are generated by the methods of Vougioukas et al. [31], Chen et al. [3], Zhou et al. [41], Prajwal et al. [20] and Ours on audio samples from TCDTIMIT (left) and GRID (right). It can be seen that our model synthesises a lip-movement that is closer to the ground truth than the other methods.

Table 1: We conducted quantitative comparisons in two benchmark datasets.

| Method             | TCD-TIMIT |             |          |          | GRID |             |          |          |
|--------------------|-----------|-------------|----------|----------|------|-------------|----------|----------|
|                    | PSNR↑     | SSIM↑       | FID↓     | LMD↓     | PSNR↑    | SSIM↑       | FID↓     | LMD↓     |
| Vougioukas, et al. [31] | 17.24     | 0.60        | 16.05    | 3.42     | 16.72    | 0.62        | 13.58    | 3.08     |
| Chen, et al. [3]     | 15.31     | 0.58        | 11.79    | 3.66     | 16.80    | 0.69        | 13.27    | 3.74     |
| Zhou, et al. [41]   | 18.10     | 0.58        | 18.02    | 2.59     | 18.53    | 0.61        | 11.87    | 2.64     |
| Prajwal, et al. [20] | 18.26     | 0.64        | 15.24    | 2.19     | 17.83    | 0.69        | 11.11    | 2.05     |
| Ours                | 20.55     | 0.65        | 8.11     | 2.18     | 20.33    | 0.65        | 5.30     | 2.18     |

with the ground truth while Prajwal et al. (highlighted in yellow) is out of sync with the ground truth. In addition, Chen et al. and Zhou et al. (highlighted in red) fail to produce the required mouth shape.

Table 2 shows the number of trainable parameters, inference time and output frame size for the five methods. We ran all experiments on a V100 Nvidia GPU and report the achieved frame rate (FPS) as a measure of inference time. The source videos are sampled at 25 FPS. We can see that the method of Vougioukas, et al. [31] and ours are faster than real-time. In addition, our method generates much higher resolution videos compared to others.

Table 2: Comparisons between our method and others in terms of number of parameters, inference time and output size.

| Method             | Number of parameters | Inference time | Output size |
|--------------------|----------------------|----------------|-------------|
| Vougioukas, et al. [31] | 55.28 M       | 441 FPS        | 96x128       |
| Chen et al. [3]     | 88.43 M       | 15.57 FPS      | 128x128      |
| Zhou, et al. [41]   | 36.40 M       | 10.32 FPS      | 526x256      |
| Prajwal, et al. [20] | 36.30 M       | 16 FPS         | 526x256      |
| Ours                | 29.68 M       | 35 FPS         | 1024x1024    |
4.4 User Study

We conducted a user study to compare our approach with related works using Amazon Mechanical Turk services. We evaluate both the audio-visual synchronisation and the visual quality of the state-of-the-art methods. We show participants a pair of videos: one generated by our method and the other generated by either Chen et al. [3] or Zhou et al. [41]. For each pair, we either ask which video looks more photo-realistic or which video has more accurate lip-sync with the audio. The choices for each question are ”right”, ”left”, ”none”, and ”both”. We randomly choose the order of videos in each pair. We obtained 80 answers from 20 participants for each question. Figure 7 shows the results of the user study. It can be seen our model achieves a better result in terms of both the audio-visual sync and the visual quality.

4.5 Ablation analysis

We analyse the effect of each loss in Eqn 4 on the performance of the model in generating talking-head videos. We train the model in stage one without $L_{\text{latent}}$ and $L_{\text{LPIPS}}$ separately and test it after tuning the generator in stage two. We observe that the choices of these losses do not affect the visual quality of the generated video but affect the lip-synchronisation accuracy. This is indicated in table 3 on the LMD column. The model trained on both losses outperforms the model trained on either of the losses alone.

Figure 7: User study results for evaluating our approach with state of the art methods in terms of the audio-visual synchronisation (top) and the visual quality (bottom)
We also compare the performance of the model at stage one without tuning the generator and after tuning the generator. Figure 8 shows a sample of generated video using the model trained in stage one. It can be seen that the model generates correct mouth positions but the visual quality inherits some distortion caused by the image generator. In stage two, we tune the generator on a single frame or short video of the target speaker. We can see from Table 4 that SSIM and PSNR are higher after tuning the generator, indicating that stage two is important to improve the quality of the generated video.

5 ETHICAL CONSIDERATION

Our framework can synthesise high-quality videos from speech audio. This is perfect for video production animation, a talking-head avatar and video-dubbing. Creative people may use our work to edit content in movies or generate new videos. However, the model can be misused to spread misinformation or manipulate existing data. Generated videos using deep learning are becoming hard to distinguish from the real thing, although there have been promising advances in forensics on the detection of "deepfake" videos [6, 23]. We will share generated videos of our framework with the community to help detecting manipulated videos.

6 CONCLUSION AND FUTURE WORKS

We propose a novel method for synthesising high-resolution videos from speech audio. The model can generate videos of a target speaker given a short video (or single image) of the speaker. Our model is built on top of a pre-trained image generator. We first learn to generate talking-head videos by constructing motion trajectories conditioned on speech audio. We then improve the image generator by tuning it on a short video of a target speaker.

We show that the method significantly outperforms recent state-of-the-art methods on TCD-TIMIT in quantitative experiments and gives performance comparable to the state-of-the-art on GRID. The method also performs best in the user study.

The generated faces depict only mouth movements because the training datasets (TCD-TIMIT and GRID) are neutral and expressionless. We anticipate our approach could in principle generate other facial expressions where these are present in the dataset (e.g. [33]), but have not yet demonstrated that this is the case.

ACKNOWLEDGMENTS

We are grateful to Rebecca Stone and Jose Martinez for their comments on this paper. This work was undertaken on ARC4, part of the High Performance Computing facilities at the University of Leeds. MA is supported by a PhD scholarship from Taif University.

REFERENCES

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. 2019. Image2stylegan: How to embed images into the stylegan latent space?. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 4432–4441.

[2] Lele Chen, Zhiheng Li, Ross K. Maddox, Zhiyao Duan, and Chenliang Xu. 2018. Lip Movements Generation at a Glance. CoRR abs/1803.10404 (2018). arXiv:1803.10404. http://arxiv.org/abs/1803.10404

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

[4] Joon Son Chung, Amit Jamaludin, and Andrew Zisserman. 2017. You said that? in British Machine Vision Conference.

[5] Martin Cooke, Jon Barker, Stuart P. Cunningham, and Xu Shao. 2006. An audio-visual corpus for speech perception and automatic speech recognition. The Journal of the Acoustical Society of America 120 5 Pt 1 (2006). 2421–4.

[6] Davide Cozzolino, Justus Thies, Andreas Rössler, Christian Riess, Matthias Nießner, and Luisa Verdoliva. 2018. Forensictransfer: Weakly-supervised domain adaptation for forgery detection. arXiv preprint arXiv:1812.02510 (2018).

[7] Dipanjan Das, S. Biwasi, Sanjana Sinha, and Rajaerdwaj Bhownick. 2020. Speech-Driven Facial Animation Using Cascaded GANs for Learning of Motion and Texture. In ECCV.

[8] Gereon Fox, Ayush Tewari, Mohamed Elgharib, and Christian Theobalt. 2021. StyleVideoGAN: A Temporal Generative Model using a Pretrained StyleGAN. arXiv:1710.10196 (2017).

[9] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. 2020. GanSpace: Discovering interpretable gan controls. Advances in Neural Information Processing Systems 33 (2020). 9841–9850.

[10] Naomi Harte and Eoin Gillen. 2015. TCD-TIMIT: An Audio-Visual Corpus of visual corpus for speech perception and automatic speech recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 7832–7841.

[11] Xinxi Ji, Hang Zhou, Kainiyaan Wang, Wayne Wu, Chen Change Loy, Xun Cao, and Feng Xu. 2021. Audio-Driven Emotional Video Portraits. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[12] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2017. Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196 (2017).
Talking Head from Speech Audio using a Pre-trained Image Generator

MM ’22, October 10–14, 2022, Lisboa, Portugal

[13] Tero Karras, Miksa Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2021. Alias-Free Generative Adversarial Networks. In Proc. NeurIPS.

[14] Tero Karras, Samuli Laine, and Timo Aila. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[15] Tero Karras, Samuli Laine, Miksa Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of stylegan. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 8110–8119.

[16] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014).

[17] Avisek Lahiri, Vivek Kwatra, Christian Frueh, John Lewis, and Chris Bregler. 2021. StyleGAN-Y. arXiv preprint arXiv:2104.15069 (2021).

[18] Moustafa Meshry, Saksham Suri, Larry S. Davis, and Abhinav Shrivastava. 2021. Learned Spatial Representations for Few-Shot Talking-Head Synthesis. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 13829–13838.

[19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Advances in Neural Information Processing Systems 32, H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alch˙e-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[20] K R Praywal, Radhika Muthopadhyay, Vimy P. Namboodiri, and C V Jawahar. 2020. A Lip Sync Expert Is All You Need for Speech to Lip Generation In The Wild. In Proceedings of the 28th ACM International Conference on Multimedia (Seattle, WA, USA) (MM ’20). Association for Computing Machinery, New York, NY, USA, 484–492. https://doi.org/10.1145/3394171.3413532

[21] Elad Richardson, Yoval Alaluf, Oh Pontuhaslik, Yotam Nitzan, Yaniv Azar, Shapira, and Daniel Cohen-Or. 2021. Encoding in Style: a StyleGAN Encoder for Image-to-Image Translation. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[22] Daniel Roach, Ron Mokady, Amit H Berman, and Daniel Cohen-Or. 2021. Pivotal tuning for latent-based editing of real images. arXiv preprint arXiv:2106.05744 (2021).

[23] Andreas Rossler, Davide Cozzolino, Luisa Verdecli, Christian Ries, Justus Thies, and Matthias Nießner. 2019. Faceforensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 1–11.

[24] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. 2020. Interpreting the Latent Space of GANs for Semantic Face Editing. In CVPR.

[25] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In International Conference on Learning Representations.

[26] Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elleuchy. 2021. StyleGAN-V: A Continuous Video Generator with the Price. Image Quality and Perks of StyleGAN2. arXiv:2112.14683 [cs] (Dec. 2021). http://arxiv.org/abs/2112.14683 arXiv: 2112.14683

[27] Linsen Song, Wayne Wu, Chen Qian, Ran He, and Chen Change Loy. 2020. Everybody’s Talking. Let Me Talk as You Want. arXiv preprint arXiv:2012.05744 (2020).

[28] Supasorn Suwajanakorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. 2017. Synthesizing Obama: Learning Lip Sync from Audio. ACM Trans. on Graph. (Proceedings of SIGGRAPH ’16), 4, Article 95 (July 2017), 13 pages.

[29] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. 2020. Neural Voice Puppetry: Audio-driven Facial Reenactment. ECCV (2020).

[30] Yu Tian, Jian Ren, Menglei Chai, Kyle Olczewski, Xi Peng, Dimitris N. Metaxas, and Sergey Tulyakov. 2021. A Good Image Generator Is What You Need for High-Resolution Video Synthesis. arXiv:2104.15069 [cs] (April 2021). http://arxiv.org/abs/2104.15069

[31] Konstantinos Vougioukas, Stavros Petrakis, and Maja Pantic. 2018. End-to-End Speech-Driven Facial Animation with Temporal GANs. In BMVC.

[32] Konstantinos Vougioukas, Stavros Petrakis, and Maja Pantic. 2019. Realistic Speech-Driven Facial Animation with GANs. International Journal of Computer Vision (IJCV) (13 Oct 2019).

[33] Kaistuynan Wang, Quanyu Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. 2020. MEAD: A Large-scale Audas-vision Dataset for Emotional Talking-face Generation. In ECCV.

[34] Suzhen Wang, Linsen Li, Yu Ding, Changjie Fan, and Xiao Yu. 2021. Audio2Head: Audas-driven One-shot Talking-head Generation with Natural Head Motion. In Proceedings of the Third International Joint Conference on Artificial Intelligence, IJCAI-21, The Hua Zhou (Ed.) International Joint Conferences on Artificial Intelligence Organization, 1098–1105. https://doi.org/10.24963/ijcai.2021/152 Main Track

[35] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing 13, 4 (2004), 600–612. https://doi.org/10.1109/TIP.2003.819861

[36] Chenxu Zhang, Yifan Zhao, Yifei Huang, Ming Zeng, Saifeng Ni, Madhukar Budagavi, and Xiaohu Guo. 2021. FACIAL: Synthesizing Dynamic Talking Face With Implicit Attribute Learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 3867–3876.

[37] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In CVPR.

[38] Zhiming Zhang, Lincheng Li, Yu Ding, and Changjie Fan. 2021. Flow-Guided One-Shot Talking-Head Generation With a High-Resolution Audio-Visual Dataset. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 3661–3670.

[39] Hang Zhou, Yu Tian, Zhiwei Li, and Xiaohu Guo. 2019. Talking Face Generation by Adversarially Disentangled Audio-Visual Representation. In AAAI Conference on Artificial Intelligence (AAAI).

[40] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, and Ziwei Liu. 2019. Pose-Controllable Talking Face Generation by Implicitly Modulated Audio-Visual Representation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[41] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingyu Li. 2020. MakeItTalk: Speaker-Aware Talking-Head Animation. ACM Transactions on Graphics 39, 6 (2020).