# Synthesizing Photorealistic Virtual Humans Through Cross-modal Disentanglement

Siddarth Ravichandran,   Ondřej Texler,   Dimitar Dinev,   Hyun Jae Kang
NEON, Samsung Research America
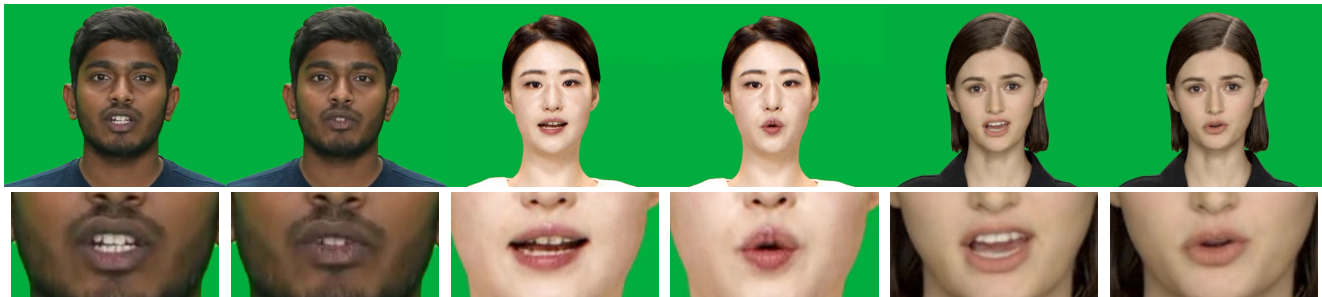{siddarth.r, o.texler, dimitar.d, hyunjae.k}@samsung.com

Figure 1. Three sequences of talking faces generated using the proposed framework. Our method produces high-texture quality, sharp images, and correct lips shapes that are inline with the spoken audio. See the zoom-in patches, our method faithfully reproduces fine textural and content details such as teeth.

## Abstract

*Over the last few decades, many aspects of human life have been enhanced with virtual domains, from the advent of digital assistants such as Amazon's Alexa and Apple's Siri to the latest metaverse efforts of the rebranded Meta. These trends underscore the importance of generating photorealistic visual depictions of humans. This has led to the rapid growth of so-called deepfake and talking-head generation methods in recent years. Despite their impressive results and popularity, they usually lack certain qualitative aspects such as texture quality, lips synchronization, or resolution, and practical aspects such as the ability to run in real-time. To allow for virtual human avatars to be used in practical scenarios, we propose an end-to-end framework for synthesizing high-quality virtual human faces capable of speaking with accurate lip motion with a special emphasis on performance. We introduce a novel network utilizing visemes as an intermediate audio representation and a novel data augmentation strategy employing a hierarchical image synthesis approach that allows disentanglement of the different modalities used to control the global head motion. Our method runs in real-time, and is able to deliver superior results compared to the current state-of-the-art.*

## 1. Introduction

With the metaverse gaining traction in the media and the rapid *virtualization* of many aspects of life, virtual humans are becoming increasingly popular. Propped up by technological advances in hardware, GPU computing, and deep learning, digitization of humans has become a very active research topic.

One challenging problem is creating virtual avatars that attempt to faithfully recreate a specific human indistinguishable from their real counterparts. This entails a perfect visual recreation, with high quality textures that recreate the minutia that make a human visually realistic, e.g., hair that appears to be composed of individual strands, clothes that appear made out of fabric, and pores or vellus hair on the skin. Recent advances in neural rendering have achieved considerable success in creating very convincing images that are able to visually recreate such fine details. Additionally, many use-cases involving virtual humans require them to appear on large screens, e.g., human-size displays in airport halls, hotels, and billboards. For such use-cases, it is vital for the virtual humans to be rendered in a relatively high resolution.

Unfortunately, rendering quality alone is not sufficient to recreate a convincing virtual human that can be interacted with. Just as important as the visual fidelity is the fidelity of the motion. Specifically, for speech-driven avatars, the

way the avatar talks, opens its mouth, and moves its lips is crucial. The human perception system is very sensitive to even the slightest mistake when it comes to the face, specifically the appearance of mouth and lips during speech. In order for a virtual avatar to be believable, its mouth and lip motion must match the speech a human listener hears.

In this work, we introduce a novel training regime that utilizes 1-dimensional audio features such as visemes [9] or wav2vec 2.0 [1]. These are higher-level features compared to traditional features such as MFCC [23] which require more complex encoders. This enables us to efficiently synthesize talking heads with high-quality rendering and lip motion. However, this alone is insufficient for creating an avatar suitable for the aforementioned use-cases. In order to have more control over the avatar, it needs to be conditioned on multiple constraints in addition to speech, such as head pose or body shape. With speech represented as a 1-dimensional vector of visemes, head pose represented as rotation angles, and body shape represented as a 2-dimensional outline, we quickly run into a problem with different modalities among our data representation.

Training deep neural networks on multimodal data is a difficult problem, especially when the different modalities are innately correlated. The network then tends to overfit on the modality that is easier to learn from, and this kind of overfitting leads to very poor performance at testing time. While there exist many techniques to mitigate traditional overfitting such as reducing number of parameters, regularization, etc., they are less effective when the overfitting is caused by multiple modalities. To break the correlations between these modalities, we introduce a novel data augmentation strategy: for every given lip shape, we use a generative oracle network to synthesize photorealistic images of the same lip shapes in a variety of head poses. This requires a generative network capable of producing images that are as high-resolution as our training data to preserve sharpness in important areas such as teeth.

Even with modern hardware and the latest research, generative neural networks still struggle with generating high-resolution images, and the resolution usually comes at the cost of quality. When increasing the resolution and the complexity of what the network is asked to generate, the generative network often loses the semantic quality or other qualitative aspects of the result images. Increasing the network's capacity alone by adding more learnable parameters does not necessarily help and can introduce noise. To alleviate this, we propose a hierarchical approach to generating a high-quality synthetic image of a talking face that preserves the quality in the mouth region via high-resolution supervision.

In order to reproduce a believable interaction with a virtual avatar, inference speed is paramount. Our paper presents an efficient framework for creating high-quality virtual artificial humans in real-time where each of the aforementioned concerns is addressed. The contribution of our works is:

- A data augmentation method to disentangle audio and visual modalities so that the whole framework can be trained end-to-end.

- A hierarchical *outpainting* approach which allows for generation of high-resolution synthetic data.

- An end-to-end framework that utilizes 2-encoder-2-decoder neural network architecture and leverages synthetic data.

## 2. Related Work

**Talking faces.** In the last several decades, many methods to generate videos of talking faces have been developed, and most of the recent ones heavily leverage generative adversarial networks–GANs [12]. Having the photo-realism as a main objective, various methods, colloquially known as "deepfakes" [2,7,20,32,34,36,45], have been proposed and successfully used to synthesize images of human faces. Talking face animation can be categorized into face re-enactment approaches [15,24,25,40,42,47] and speech driven approaches [3,26,27,31,46,49,51]. Speech driven models have used a variety of intermediate representations such as 2D keypoints [3,32], meshes [6,27,33], which themselves can be derived from various audio features such as MFCC, spectral coefficients, or DeepSpeech features [13]. Some methods directly produce talking faces from audio features without additional geometric representations [46] and some [26,49] use an additional identity image or video as guiding signal. A common issue these methods face is the reduced high-frequency motion, such as lip motion during speech. Unrealistic lip movement is easy to notice, and disrupts the overall experience. This is partially addressed in the work of *Prajwal et al.* [26], they utilize SyncNet [5] to supervise the audio and lips synchronization, which results in a dynamic mouth movement. However, they do not consider the photorealistic appearance. As the mandate of our work is to create virtual humans that are indistinguishable from real people, we need to secure both realistic lip motion and photorealistic textures.

**Hierarchical image generation.** Synthesizing high-resolution images while retaining various fine details and overall sharpness is a difficult task. Researchers have been approaching this problem from various angles, and one of widely used solutions is to tackle this challenge in a hierarchical fashion. InsetGAN [11] propose to combine multiple pre-trained GANs, where one network generates the coarse body image, and other networks refine certain regions. The work of *Li et al.* [22] proposes to have several
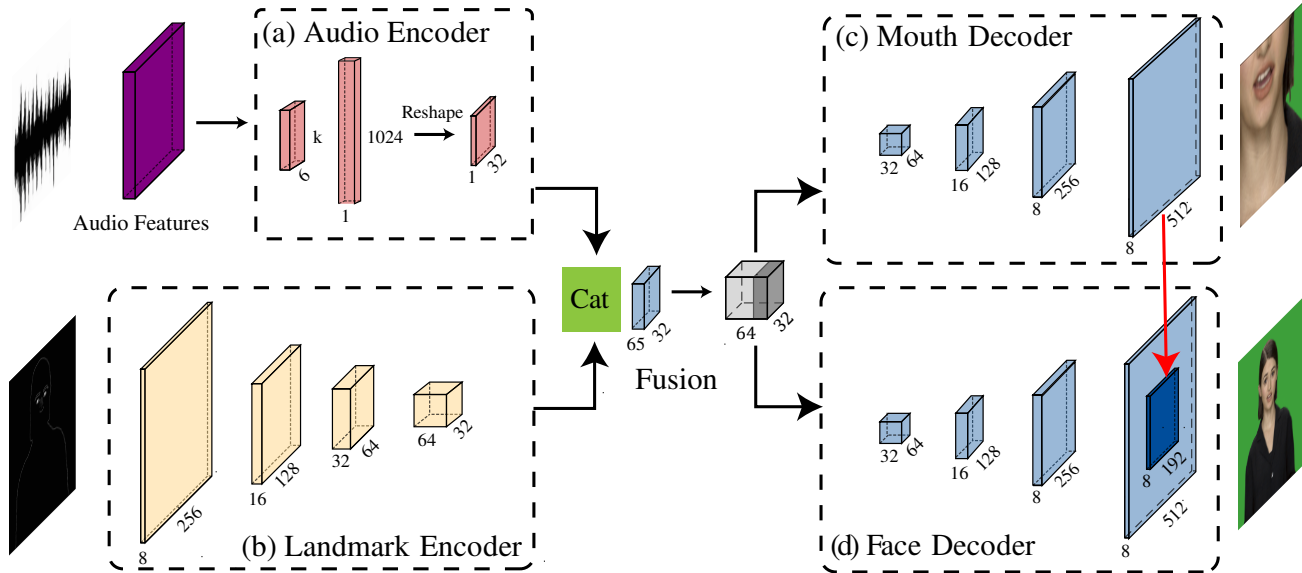
Figure 2. Overview of our multi-modal neural rendering pipeline. We generate $k$ audio features of the provided audio, and sample a sequence of keypoint drawings and contours which are then mapped to a shared latent space using Audio Encoder (a) and the Landmark Encoder (b) respectively; which is then passed through the mouth (c) and face (d) decoders to produce the mouth and face images respectively. The mouth feature maps are pasted into the head feature map as denoted by the red arrow. During inference, only the output of the head decoder is used.

class-specific generators that are used to refine certain semantic regions of the previously synthesized image. Tile-GAN [10] presents a way to combine small-resolution synthesized images in order to produce a single high-resolution texture. *Xu and Xu* [43] hierarchically compose the image from smaller patches, and *Iizuka et al.* [17] solve inpainting task using two discriminators, one for the whole image, and the other for a small area centered at the completed region. Although those approaches are not directly applicable in our talking virtual human scenario, our pipeline also leverages the hierarchical composition.

**Multi-modal training.** To puppet a virtual human, multiple different inputs are usually needed, e.g., speech to control the lips and head-pose to specify head position and rotation. Those inputs–modalities are usually strongly correlated, which translates in the need of a multi-modal training [4, 37, 48, 50]. NeuralVoicePuppetry [33] trains a generalized expression network and an identity targeted neural renderer method decoupling the modalities using an intermediate mesh representation. MakeItTalk [51] propose to disentangle the modalities by learning displacements from a reference facial landmark using a separate module and perform rendering from the intermediate 2-dimensional landmark representations. Similarly, *L. Chen et al.* [3] use a learned high-level hierarchical structure to bridge audio signal with the pixel image. MeshTalk [27] deals with the modality disentanglement in a lower dimensional space by employing the cross-modality loss. Many of these methods require an intermediate representation for the face in the

form of drawings or mesh, hindering their ability to produce diverse lip shapes. We address these limitations in our work by using a mix of 1-dimensional and 2-dimensional representations, requiring modality disentanglement in order to synthesize speech-driven talking faces.

## 3. Our Approach

We present a framework to generate photorealistic virtual human avatars that can be streamed in real-time to enable interactive applications. We train our system on a single identity and during the inference we drive it using audio features and partial face keypoints, as shown in Fig. 2. Due to bias in the way humans speak, there is a strong correlation between head motion and speech which causes the network to model undesirable relations. To combat that, we propose a novel data augmentation strategy employing keypoint mashing and a high-resolution generative oracle network to achieve disentanglement of lip motion from upper face motion. This allows synthesizing videos with arbitrary upper face keypoints and audio.

### 3.1. Input Features

**Audio features.** One of the crucial requirements for our method is the ability to faithfully reproduce motion and shapes of mouth and lips. Thus, it is vital to have an intermediate audio representation that is capable of capturing those geometric characteristics. Visemes [9] are lip shape categories that describe the basic lip shapes used to make

sounds common in human speech. They can be thought of as the geometric analog to phonemes, and have been utilized to create a geometric basis for head models used in commercial 3D graphics products to parameterize and animate speech [52]. Despite the aforementioned theoretical advantages, visemes have certain practical difficulties, often requiring commercial tools such as JALI [8]. Our framework can also handle different frame-level audio feature representations such as wav2vec 2.0 [1].

Commonly used intermediate representations, such as keypoints [39] can be inaccurate, especially for difficult-to-detect regions such as lips, We propose to forego the use of intermediate geometric representation and instead render the image directly given visemes.

**Facial features.** To render the rest of the head, landmarks and contours are strong guiding signals for the network to produce semantically correct images. We use these features of the non-mouth areas of the face as additional inputs for the network, enabling control of non-speech-related behavior such as head motion.

### 3.2. Multi-Modal Neural Rendering

We start by processing the audio to obtain $k$ features per frame $\mathbf{v} \in \mathbb{R}^{k \times 1}$. We take a window of 6 consecutive frames centered around the corresponding image frame. This results in a vector of size $\mathbf{v} \in \mathbb{R}^{k \times 6}$, shown in Fig. 2a. For visemes, $k = 34$.

The second modality is the upper face keypoints and the contours drawn on a canvas of size $512 \times 512$ px, in Fig. 2b. Rather than passing the keypoints as vectors, we draw them as a 2D image which helps the model to learn spatial correspondences, resulting in better generalization to translations and motion in the keypoints.

**Model architecture details.** Our model consists of two encoders, one for each modality. The audio encoder consists of a 1D convolutional neural network followed by a 2-layer perceptron network. The resulting latent vector is reshaped to a 2D representation and concatenated with the latent feature map of the keypoint encoder which consist of 4-layer strided convolution blocks with a residual block [14] in between consecutive strided convolutions. Taking the inspiration from Pix2PixHD [39], we adopted a two image decoder scheme. One decoder is trained to generate a full face (Fig. 2d), and the second decoder (Fig. 2c) is trained to generate a fixed crop region of the mouth at a larger resolution. Both decoders produce different images from the same latent space. The model is trained end to end in a GAN setup, using two patch discriminators [18]: one for the head and the other for the mouth. We use the VGG [29] feature matching loss and smooth L1 loss alongside the GAN loss to train the generators. To obtain sharper mouth images, especially for the teeth, we resize and place the mouth region feature map into the head feature map at the appro-
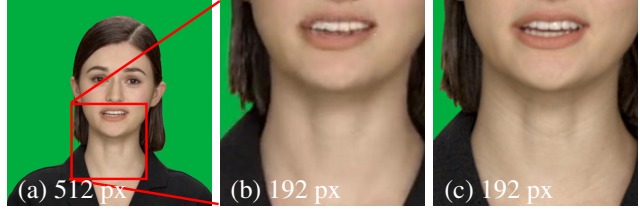


Figure 3. Training a full-head model at 512 px resolution (a) produces lower details in the 192 px mouth region (b) than training a mouth-only model on 512 px, then resampled to 192px (c).

priate crop position in the penultimate layer of the head decoder, as shown by the red arrow in Fig. 2. We do this in the penultimate layer rather than the final image to avoid any blending artifacts that could arise.

### 3.3. Disentanglement of Audio and Head-Pose

As mentioned in Section 3.2, we use two inputs to the network: a $k \times 6$ vector of audio features, and a $512 \times 512$ line drawing of the face. Since our training dataset only contains a single subject speaking one language, this leads to considerable entanglement between audio features and the 2D contour drawing, making the model learn unwanted relationships between the line drawing and lip motion. This manifests in, for example, lip motion being generated from just head motion in the contour drawing, even without providing any audio.

To remedy this, we propose a novel data augmentation strategy for disentangling the audio features from the head pose. One way to enforce the correlation between audio features and lip shape is to show the network samples of the subject uttering the same phrases with different head poses. While this can be done during data capture, this significantly lengthens and complicates the capture process. To remedy this, we propose training an oracle network that converts 2D contour drawings into photoreal renders. Then, we propose a method for transferring the mouth shape from one frame to the head pose of another. Combining these two methods allows us to augment the capture dataset with synthetic images containing a greater variety of lip shapes and head poses, forcing the network to break the correlation between head pose and lip shape.

**Oracle Network.** The oracle generates photorealistic images of the subject given the 2D line drawings from in Fig. 2. This can be accomplished by employing a network based on Pix2Pix [18,39]. While the images produced by this network appear photorealistic, Pix2Pix can struggle with blurriness, especially in regions where there is significant motion and occlusions, such as the teeth. This blurriness pollutes our training data, degrading the visual quality of our method. One observation is that if we train a higher-resolution network that focuses on the mouth then the teeth
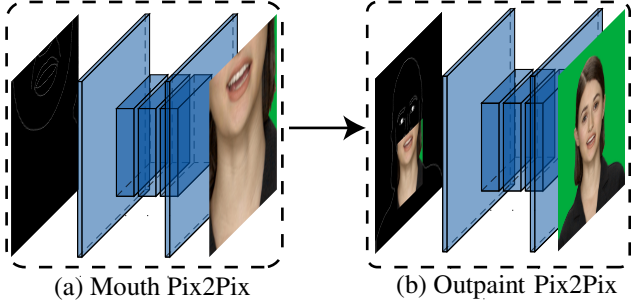
(a) Mouth Pix2Pix       (b) Outpaint Pix2Pix

Figure 4. Our oracle network works in a hierarchical fashion, starting with the network to generate mouth (a) in a high-resolution. The result is then fed into the subsequent network (b) to generate the rest of the head. The head image and the high resolution mouth image are used for supervision of our multi-modal renderer.



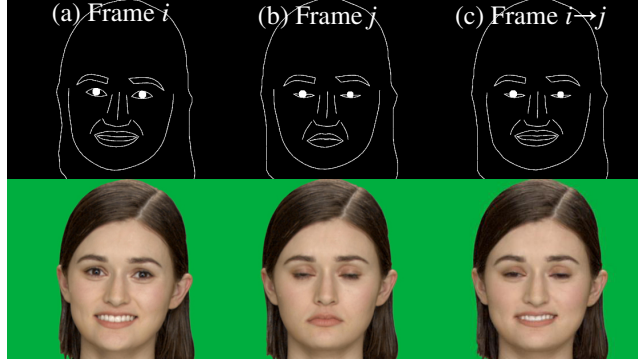(a) Frame $i$     (b) Frame $j$     (c) Frame $i \rightarrow j$

Figure 5. Using keypoint mashing and the proposed oracle network, we are able to generate a synthetic image (c) by combining the mouth keypoints of (a) with the head keypoints of (b).

will come out clearer, even if the result is then downsampled as shown in Fig. 3. Here, we trained two versions of the same network: one on the full-head $512 \times 512$ px data Fig. 3a and the other on high-resolution $512 \times 512$ px mouth crop, then downscaled to $192 \times 192$px Fig. 3c. If we extract the same $192 \times 192$ region from the full head model (Fig. 3b), we can see that there is a degradation in quality in the teeth region.

To produce high-resolution renderings, we propose to render the image in a hierarchical manner starting with the region where the details and quality is most important for the given task – the mouth region in our case. Our solution can be viewed as a *chaining* of neural network models where the *next* network is conditioned by the result of the *previous* network, shown in Fig. 4. First, we train a network that generates a high-quality mouth image( Fig. 4a). This image is downsampled and placed in a line drawing of the full head, similar to the process outlined in Section 3.2. Then, it is used as input to the full-head network( Fig. 4b), which learns to render the rest of the head. We refer to this hierarchical approach as *outpaint* since it resembles the task of outpainting.

**Keypoint mashing.** The next missing piece for generating novel head pose and mouth shape combinations is the ability to combine the mouth keypoints from one frame with the head keypoints in another frame. However, the two frames will have different head poses, meaning we need to remove the global pose information before the mouth keypoints can be replaced.

For any given frame, we extracted $n$ 2D facial keypoints $\mathbf{K} \in \mathbb{R}^{n \times 2}$ and a pose matrix $\mathbf{P} \in \mathbb{R}^{4 \times 4}$ based on a 3D canonical space for the keypoints. The position of the keypoints can be expressed as:

$$\mathbf{K} = proj(\mathbf{P} \mathbf{K^f}) \tag{1}$$

where $\mathbf{K^f} \in \mathbb{R}^{n \times 4}$ are the un-posed keypoints with depth in homogeneous coordinates, and $proj()$ is a projection func-

tion (e.g. perspective projection) that converts keypoints from 3D space to 2D screen space. Using Eq. 1, we can compute the un-posed keypoints:

$$\mathbf{K^f} = \mathbf{P^{-1}} proj^{-1}(\mathbf{K}) \tag{2}$$

Where $proj^{-1}$ is the inverse projection function. For an orthographic projection, this would just be an inverse of the projection matrix, provided the keypoint data is 3D. However, many keypoints detectors only provide 2D information. In order to fill in the missing depth, we obtain depth values by solving the forward problem Eq. 1 for the canonical keypoints and concatenating the resulting depth value to the extracted 2D keypoints. This same process can be used to recover the perspective divide if $proj$ is a perspective transformation. While this is only an approximation for the true depth values, most of the depth variation in the face comes from the underlying facial bone structure and geometry captured in the canonical keypoints, making this a viable approximation.

**Generating synthetic data.** With keypoint mashing and a generative oracle in place, we now define a method for data augmentation. We select two arbitrary frames $i$ and $j$, with corresponding keypoints and poses $\mathbf{K}_i, \mathbf{P}_i$ (Fig. 5a, top) and $\mathbf{K}_j, \mathbf{P}_j$ (Fig. 5b, top), we compute the un-posed keypoints $\mathbf{K}_i^f$. Then, we apply $\mathbf{P}_j$ resulting in a new keypoints set $\mathbf{K}_{i \rightarrow j}$: the keypoints from frame $i$ in the head pose of frame $j$. Since the two keypoint sets $\mathbf{K}_{i \rightarrow j}$ and $\mathbf{K}_j$ are now in the same space, we can trivially replace the mouth keypoints: $\mathbf{K}_j^{mouth} = \mathbf{K}_{i \rightarrow j}^{mouth}$ resulting in the mouth position from $i$ transferred over to frame $j$, $\mathbf{K}_j'$. Now we have audio features for frame $i$, but a new keypoint set $\mathbf{K}_j'$ (Fig. 5c, top).

Using our oracle network, we can now generate synthetic images (Fig. 5c, bottom) from $\mathbf{K}_j'$ to be used as the "ground truth" for this new head pose and mouth shape combination. This makes the network to see different head poses

(a) MakeItTalk     (b) Wav2lip     (c) NVP+Pix2Pix     (d) Ours     (e) Ground Truth
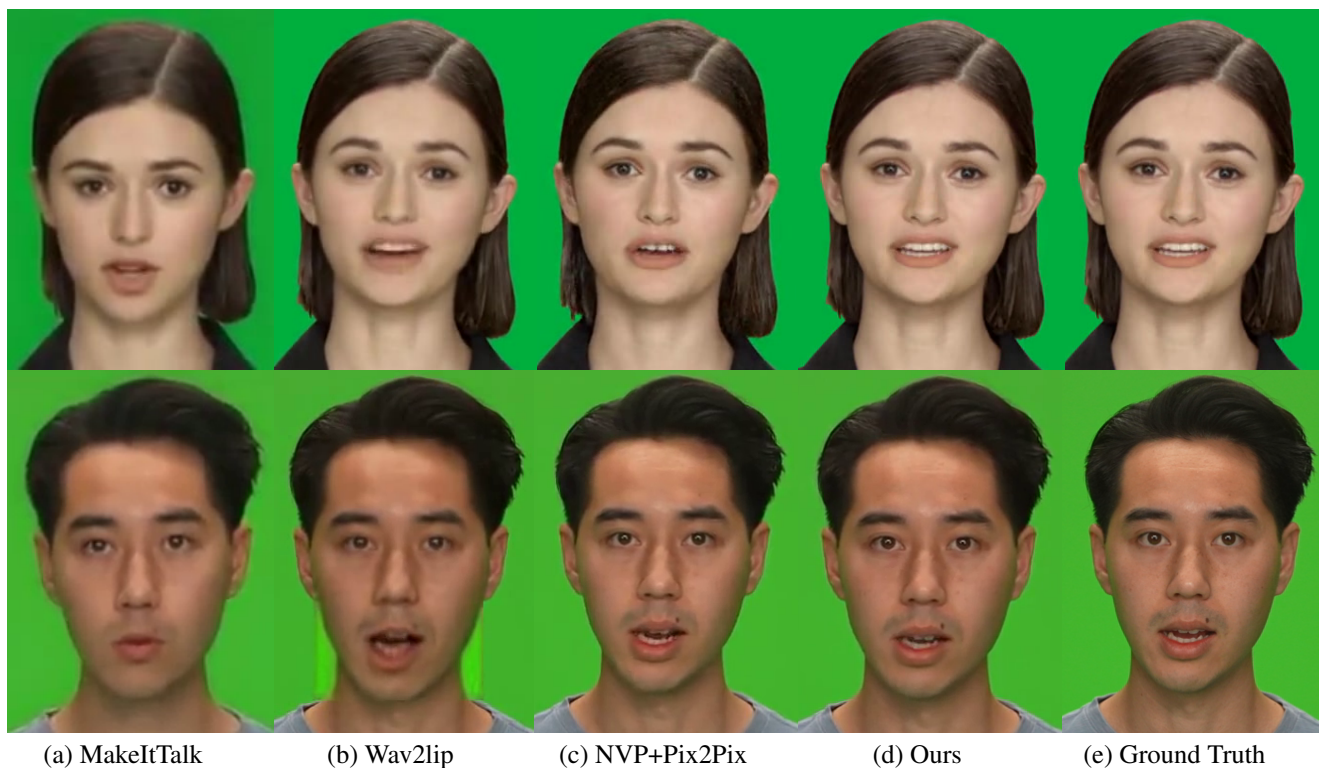
Figure 6. Comparison with state-of-the-art techniques. MakeItTalk [51] (a), Wav2Lip [26] (b), NVP [33]+Pix2pix [38] (c), our result (d), and ground truth image (e). The avatar says 'ee'. Our results have significantly better lip shapes as compared to (a, c), and better texture quality as compared to (a, b).

for every audio feature frame, preventing it from learning an erroneous correlation between the head motion and lip shape.

## 4. Results

First we present comparisons with the current state-of-the-art techniques. Then, we elaborate on data collection and implementation details, discuss hyper-parameters, followed by further analysis and discussion. Our method is single identity, and each subject shown has their own model trained on their own dataset. For more results, please, see our supplementary pdf and video.

### 4.1. Comparisons

**Qualitative comparisons.** We compare our method with two recent multi-identity approaches, Fig. 6a-b: MakeItTalk [51] and Wav2Lip [26]. These methods are trained on a large corpus of data containing many identities, and extract the identity of the subject at inference time. While it is indeed practical that MakeItTalk only requires a single image to synthesize a new identity, the resulting visual and lip sync quality is markedly worse than our method. Wav2Lip, which infers identity from a source video, arguably produces better lip sync than our method,

but at the expense of visual quality: Wav2Lip only supports $96 \times 96$, of which only $48 \times 96$ are used for the mouth, and attempting higher resolutions fails. Notice that both of these above methods produce black blobs in the mouth region without any teeth textures. We then compared our method to TalkingFace [30], which is also trained on a single identity using data which is similar to our approach making the comparison straightforward. Our method produces superior lip motion: Fig. 7 shows how the two methods compare when producing the word "we". Our method is able to correctly capture the lip transition from the more rounded "w" sound to the "e" sound. Another single-identity method for generating talking heads is outlined in Neural Voice Puppetry [33]. We implemented a variation using Pix2Pix [38] and trained it on our data; our method produces superior lip motion Fig. 6c-d. Overall, our method produces higher quality videos with sufficiently accurate lip motion than the competing methods. Please refer to the supplementary video for more detailed comparisons.

**Quantitative comparisons.** In Table 1, we provide quantitative comparisons between our method and the above competing methods. To evaluate image quality, we compute the PSNR and SSIM [41] scores for 500 inferred images against their ground truth counterparts, and show
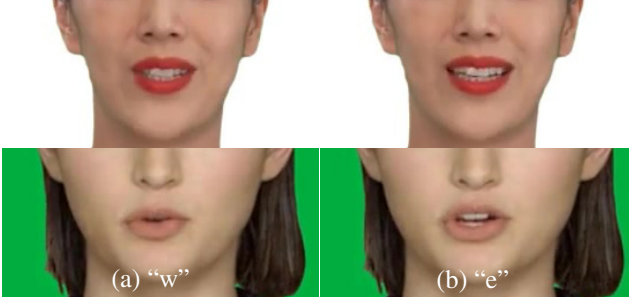
Figure 7. TalkingFace [30] (top) and our model (bottom) were both asked to produce the word "we". Our model captures the transition from "w" (a) to "e" (b) more accurately.

| Method | FPS ↑ | SSIM ↑ | PSNR ↑ | Sync ↑ | $\mathbf{D}_{lip}$ ↓ |
|---|---|---|---|---|---|
| Wav2Lip | 49 | 0.88 | 33.16 | **9.51** | **8.65** |
| TalkingFace | 63 | na | na | 6.45 | na |
| NVP+p2p | 47 | 0.82 | 27.19 | 4.83 | 18.34 |
| **Ours** | **110** | **0.94** | **35.71** | 7.41 | 12.81 |

Table 1. Quantitative comparisons with Wav2Lip [26], Talking-Face [30] and NVP [33]+Pix2Pix [38]. Sync confidence score **Sync** and Lip Distance metric $\mathbf{D}_{lip}$ are used to determine accuracy of lip synchronization. Structure similarity score **SSIM** and peak signal to noise ration **PSNR** are used to judge the image quality. Our methods outperform other methods by a large margin on image quality and trails close behind Wav2Lip in lip-sync accuracy while achieving the highest inference speed **FPS** by a significant margin.

that we produce the best results. We also provide two metrics to evaluate the quality of the lip sync: the score provided by SyncNet [5] (**Sync**) and Lip Marker Distance ($\mathbf{D}_{lip}$) in Table 1. In terms of lip-sync, our results are superior to all of the other comparisons except Wav2Lip. This is because Wav2Lip was trained to maximize **Sync**, which also translates into better $\mathbf{D}_{lip}$. Our method comes close to Wav2Lip in terms of lip sync quality, but with a higher visual quality. Since we do not have the ground truth videos for Talking Face, we are unable to compute SSIM, PSNR and $\mathbf{D}_{lip}$ metrics, which rely on ground truth frames.

**Performance.** All results shown in the Table 1 were run on a single RTX 3080 Ti with a Intel(R) Core(TM) i9-9900K CPU @ 3.60GHz. Our method is close to twice as fast as the second fastest method, TalkingFace. This is because by using a vector of visemes as input, we are able to greatly simplify our network. Additionally, since our architecture does not use any custom layers, our models can be easily converted to TensorRT. This further increases our inference speeds up to 200 frames per second, making it very suitable for production-level interactive applications.
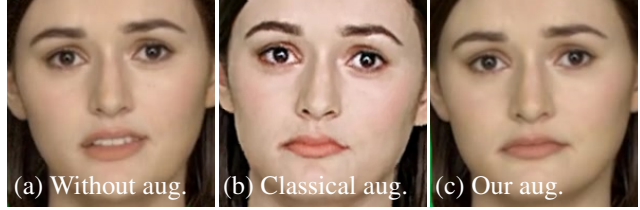


Figure 8. The proposed augmentation is essential for the model to correctly respect different inputs. Without the augmentation (a), the virtual human erroneously opens her mouth even when the audio is silence; while (b) classical data augmentation methods might slightly mitigate this issue the model still produces wrong lip shapes; this is fixed by our augmentation technique (c).

## 4.2. Discussion and Analysis

**Data Collection and Capture Setup.** We recorded subjects using cameras capable of capturing 6K footage at 30 frames per second. From the full-sized video, we crop out the head region and a high-resolution mouth image that is used to supervise the training. We then run facial landmark detection, pose estimation, and contour extraction on the head crops, which provides us with the information we need to produce line-drawings such as the ones in Fig. 2. For audio data, we extract visemes by providing the audio waveforms and a transcript to JALI [8].

**Implementation Details.** We detect facial keypoints and head pose using a commercial product AlgoFace (www.algoface.ai). To obtain contours, we first generate a head segmentation using a BiSeNetV2 [44] and then we draw a line on the background-foreground boundary. Finally, we use JALI [8] to extract visemes. The viseme window size is a hyper-parameter; we empirically determined that using a window of size 6 is optimal. Larger window sizes lead to muffled mouth motion and very small window sizes produces unstable results. For the generator and both of the discriminators, we use Adam optimizer with a learning rate of 0.0002 and learning rate scheduling. We train the system on a 20 minute video of the subject speaking and evaluate on out of domain TTS audio paired with arbitrary sequences of keypoints. The oracle and multi-modal renderer each take roughly 6-8 hours on two 3080 GPUs with a batch size of 1.

**Disentanglement.** Our network takes two types of input: speech data represented as visemes and head motion data represented as keypoint drawings. These two modalities are inherently connected: humans move their heads rhythmically while talking in a way that matches their speech, e.g., nodding up and down subtly when saying "yes". Without any special handling, our network learns a strong correlation between the head motion and word being spoken. This results in the model producing mouth openings just from head motion, depicted in Fig. 8a. Generating synthetic data

using classical data augmentation techniques such as rotation, scale, translation, cropping, or adding noise would not be effective in our case. While it is true that it might slightly help with the overfitting as it introduces randomness to the data and it might have effect similar to regularization, this alone does not break the correlation between head motion and audio, see Fig. 8b and our supplementary video, the lips still move and the mouth is not in the neutral resting pose. Our data augmentation successfully breaks the correlation by introducing a variety of head pose and viseme combinations. Fig. 8c shows that our model properly keeps its mouth closed during silence, regardless of head motion. We empirically observed that at least 80% of the training data have to be synthetic data for the disentanglement to have desired effect.



(a) Oracle result    (b) Multi-modal result

Figure 9. Even though the oracle (a) can sometimes produce artifacts in our synthetic training data, our multimodal renderer (b) is robust to them.

Since the keypoint mashing Section 3.3 can produce unnatural combinations that are substantially different from the oracle's training data, the synthetic data produced can contain noticeable artifacts (Fig. 9a). However, such artifacts only occur when the two frames have vastly different head poses; this is rare and results in only a small percentage of the training data being polluted. Our multi-modal network is robust to such artifacts, shown in Fig. 9b.

**Different backbones and performance optimization.** MobileNetV2 [28] is usually used as compressed architectures primarily for image recognition tasks. We tried to directly apply the MobileNetV2 backbone for our image synthesis tasks and observed that both *large* and *small* variants of the MobileNetV2 were slower than our current approach. Using hard swish activation as proposed in MobileNetV2 led to rendering artifacts across multiple runs, and we chose to use the original leaky relu activation functions. Finally, replacing all convolutions in our network with depth wise separable convolutions did improve the performance, but also introduced some quality degradation in the teeth and produced grid like artifacts near high contrast texture regions like the hair.

## 5. Limitations and Future Work

While our work significantly advances the state-of-the-art in generating realistic virtual human avatars, there are still certain limitations and a fair amount of potential future work that needs to be done before the virtual humans are truly indistinguishable from real people.

One of the major limitations of our framework is that it does not perform well on large motions, head rotations, and extreme head poses. To alleviate this in a future work, we envision involving 3D geometry and other aspects from the 3D rendering domain, in particular, using a mesh as an intermediate representation instead of 2D images, which would allow for better occlusion and collision handling and for using 3D neural rendering techniques, such as deferred neural rendering [35] that have been proven very successful.

Furthermore, we also observed certain texture-sticking artifacts between frames when the motion is large, this is due to the fully convolutional nature of our network; this is a known problem discussed in, for example, StyleGAN3 [19], and might be mitigated by adopting vision transfomers [21]. Also, as a part of further work, we encourage exploration into how multiple modalities can be used to target the same part of the face, enabling us to selectively learn from each of the modalities [16]. In the context of this work, this could mean, for example, having visemes as one modality and a signal to controls smiling as another modality; both modalities used to synthesize lips.

## 6. Conclusion

While the problem of creating virtual humans that perfectly mimic the appearance and behavior of real people is still far from being solved, our work considerably pushes the state-of-the-art boundaries. We presented a robust and efficient framework for generating photorealistic talking face animations from audio in real-time. Thanks to the proposed alternative data representation, a training system to prevent modality entanglement, and supervision from high resolution around the mouth area, we are able to produce superior face rendering quality with better lip synchronization compared to recent approaches, all while maintaining real-time inference.

### Acknowledgments

### References

[1] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised

learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460, 2020. 2, 4

[2] Christoph Bregler, Michele Covell, and Malcolm Slaney. Video rewrite: Driving visual speech with audio. *ACM Transactions on Graphics*, 31:353–360, 01 1997. 2

[3] Lele Chen, Ross K. Maddox, Zhiyao Duan, and Chenliang Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7824–7833, 2019. 2, 3

[4] Sen Chen, Zhilei Liu, Jiaxing Liu, Zhengxiang Yan, and Longbiao Wang. Talking head generation with audio and speech related facial action units. *arXiv preprint arxiv:2110.09951*, 2021. 3

[5] Joon Son Chung and Andrew Zisserman. Out of time: Automated lip sync in the wild. In *ACCV Workshops*, 2016. 2, 7

[6] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J Black. Capture, learning, and synthesis of 3d speaking styles. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 10101–10111, 2019. 2

[7] Nikita Drobyshev, Jenya Chelishev, Taras Khakhulin, Aleksei Ivakhnenko, Victor Lempitsky, and Egor Zakharov. Megaportraits: One-shot megapixel neural head avatars. *Proceedings of the 30th ACM International Conference on Multimedia*, 2022. 2

[8] Pif Edwards, Chris Landreth, Eugene Fiume, and Karan Singh. Jali: An animator-centric viseme model for expressive lip synchronization. *ACM Transactions on Graphics*, 35:1–11, 07 2016. 4, 7

[9] Cletus G Fisher. Confusions among visually perceived consonants. *Journal of speech and hearing research*, 11(4):796–804, 1968. 2, 3

[10] Anna Frühstück, Ibraheem Alhashim, and Peter Wonka. TileGAN: Synthesis of large-scale non-homogeneous textures. *ACM Transactions on Graphics*, 38(4):58:1–58:11, 2019. 3

[11] Anna Frühstück, Krishna Kumar Singh, Eli Shechtman, Niloy J Mitra, Peter Wonka, and Jingwan Lu. Insetgan for full-body image generation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 7723–7732, 2022. 2

[12] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014. 2

[13] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014. 2

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 4

[15] Fa-Ting Hong, Longhao Zhang, Li Shen, and Dan Xu. Depth-aware generative adversarial network for talking head video generation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 3397–3406, 2022. 2

[16] Xun Huang, Arun Mallya, Ting-Chun Wang, and Ming-Yu Liu. Multimodal conditional image synthesis with product-of-experts GANs. In *ECCV*, 2022. 8

[17] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and Locally Consistent Image Completion. *ACM Transactions on Graphics*, 36(4):107:1–107:14, 2017. 3

[18] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 5967–5976, 2017. 4

[19] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 852–863, 2021. 8

[20] Taras Khakhulin, Vanessa Sklyarova, Victor Lempitsky, and Egor Zakharov. Realistic one-shot mesh-based head avatars. In *Proceedings of European Conference on Computer Vision*, 2022. 2

[21] Kwonjoon Lee, Huiwen Chang, Lu Jiang, Han Zhang, Zhuowen Tu, and Ce Liu. ViTGAN: Training GANs with vision transformers. In *International Conference on Learning Representations*, 2022. 8

[22] Yuheng Li, Yijun Li, Jingwan Lu, Eli Shechtman, Yong Jae Lee, and Krishna Kumar Singh. Collaging class-specific gans for semantic image synthesis. In *Proceedings of IEEE International Conference on Computer Vision*, pages 14398–14407, 2021. 2

[23] Paul Mermelstein. Distance measures for speech recognition, psychological and instrumental. *Pattern recognition and artificial intelligence*, 116:374–388, 1976. 2

[24] Yuval Nirkin, Yosi Keller, and Tal Hassner. FSGAN: Subject agnostic face swapping and reenactment. In *Proceedings of IEEE International Conference on Computer Vision*, pages 7184–7193, 2019. 2

[25] Yuval Nirkin, Yosi Keller, and Tal Hassner. FSGANv2: Improved subject agnostic face swapping and reenactment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2

[26] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. 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*, pages 484–492, 2020. 2, 6, 7

[27] Alexander Richard, Michael Zollhöfer, Yandong Wen, Fernando de la Torre, and Yaser Sheikh. Meshtalk: 3d face animation from speech using cross-modality disentanglement. In *Proceedings of IEEE International Conference on Computer Vision*, pages 1173–1182, October 2021. 2, 3

[28] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted

residuals and linear bottlenecks. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018. 8

[29] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 4

[30] Hyoung-Kyu Song, Sang Hoon Woo, Junhyeok Lee, Seungmin Yang, Hyunjae Cho, Youseong Lee, Dongho Choi, and Kang-wook Kim. Talking face generation with multilingual tts. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 21425–21430, June 2022. 6, 7

[31] Yang Song, Jingwen Zhu, Dawei Li, Xiaolong Wang, and Hairong Qi. Talking face generation by conditional recurrent adversarial network. *arXiv preprint arXiv:1804.04786*, 2018. 2

[32] Supasorn Suwajanakorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: Learning lip sync from audio. *ACM Transactions on Graphics*, 36(4), 2017. 2

[33] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural voice puppetry: Audio-driven facial reenactment. In *Proceedings of European Conference on Computer Vision*, pages 716–731, 2020. 2, 3, 6, 7

[34] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 2387–2395, 2016. 2

[35] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics*, 38:1–12, 07 2019. 8

[36] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. End-to-end speech-driven realistic facial animation with temporal gans. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Workshops*, June 2019. 2

[37] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Realistic speech-driven facial animation with gans. *International Journal of Computer Vision*, 2020. 3

[38] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In *Advances in Neural Information Processing Systems*, pages 1144–1156, 2018. 6, 7

[39] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 8798–8807, 2018. 4

[40] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2

[41] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 6

[42] Wayne Wu, Yunxuan Zhang, Cheng Li, Chen Qian, and Chen Change Loy. Reenactgan: Learning to reenact faces via boundary transfer. In *Proceedings of European Conference on Computer Vision*, 2018. 2

[43] Xiaogang Xu and Ning Xu. Hierarchical image generation via transformer-based sequential patch selection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(3):2938–2945, 2022. 3

[44] Changqian Yu, Changxin Gao, Jingbo Wang, Gang Yu, Chunhua Shen, and Nong Sang. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *International Journal of Computer Vision*, 129:1–18, 11 2021. 7

[45] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor S. Lempitsky. Few-shot adversarial learning of realistic neural talking head models. *Proceedings of IEEE International Conference on Computer Vision*, pages 9458–9467, 2019. 2

[46] Sibo Zhang, Jiahong Yuan, Miao Liao, and Liangjun Zhang. Text2video: Text-driven talking-head video synthesis with personalized phoneme-pose dictionary. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2659–2663, 2022. 2

[47] Yunxuan Zhang, Siwei Zhang, Yue He, Cheng Li, Chen Change Loy, and Ziwei Liu. One-shot face reenactment. In *British Machine Vision Conference*, 2019. 2

[48] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 3661–3670, June 2021. 3

[49] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. Talking face generation by adversarially disentangled audio-visual representation. In *AAAI Conference on Artificial Intelligence*, 2019. 2

[50] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 3

[51] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makeittalk: Speaker-aware talking-head animation. *ACM Transactions on Graphics*, 39(6), 2020. 2, 3, 6

[52] Yang Zhou, Zhan Xu, Chris Landreth, Evangelos Kalogerakis, Subhransu Maji, and Karan Singh. Visemenet: Audio-driven animator-centric speech animation. *ACM Transactions on Graphics*, 37(4):1–10, 2018. 4