Dubbing for Everyone: Cost and Data-Efficient Visual Dubbing using Neural Rendering Priors

Anonymous authors
Paper under double-blind review

Abstract

Visual dubbing is the process of generating lip motions of an actor in a video to synchronize with given audio. Visual dubbing allows video-based media to reach global audiences. Recent advances have made progress towards realizing this goal. However, existing models are either zero-shot and, therefore, lack quality, or they are expensive methods requiring off-putting user enrollment with lengthy and costly model training. Our key insight is to train a large, multi-person prior network, which can then be adapted to new users. This method allows for **high-quality visual dubbing with just a few seconds of data**, that enables video dubbing for any actor - from A-list celebrities to background actors at a much lower cost. We show that we achieve state-of-the-art in terms of **visual quality** and **recognizability** both quantitatively and qualitatively through two user studies. Our prior learning and adaptation method **generalizes to limited data** better than existing person-specific models. Our experiments on real-world, limited data scenarios find that our model is preferred over all existing methodologies.

1 Introduction

Dubbing is the process of translating video content from one language to another. For the most part, dubbing is performed only on the audio tracks, leaving the video unchanged. This results in a poor visual experience. Visual dubbing involves reconstructing the lip and mouth movements of the actor in a video to match new audio in a different language. When done correctly, visual dubbing transforms how global audiences watch video content filmed in non-native languages, allowing content creators to reach more viewers worldwide.

We argue that any successful video dubbing method must be **high-quality**, **generalizable**, **recognizable** and **low-cost**. It must be **high-quality** so consumers are not distracted by the synthesized lips, avoiding the 'uncanny valley' effect. This requires good video quality and good lip sync. It must be **generalizable** in that all actors, from A-list stars to background actors, should be dubbed effectively with as little as a few seconds of dialogue. An actor's style should be **recognizable** in the dubbed video. For instance, the actor's lips and teeth should look the same in the dubbed video as in the real one. To be viable, a visual dubbing solution should also be low-cost.

Previous models are split between their focus on these criteria. Some produce excellent videos for a single actor under controlled conditions (e.g. Thies et al. (2020); Ye et al. (2023); Tang et al. (2022)). Such methods are high-quality and recognizable but will work only on the actor they are trained on. They also require significant training data, often at least 2 minutes. This makes them unsuitable for practical dubbing in scenes where actors may only be present for a few seconds. Other methods produce a low-quality but generalizable video (e.g. Wang et al. (2023a); Gupta et al. (2023); Guan et al. (2023). These methods can be applied to any audio and video cheaply, but the outputs are rarely of good visual quality and do not capture the style of the actors. For instance, they all produce overly generic teeth and mouth interiors; see Figure 6.

We propose Dubbing for Everyone. We create a model that meets ***all*** our criteria, allowing the high-quality, low-cost dubbing of all actors, including those with short roles. The critical insight of our work is the realisation that existing high-quality models can benefit from a large-scale pre-training of some of their components. We call this a prior network. Our prior network is trained across multiple actors and can

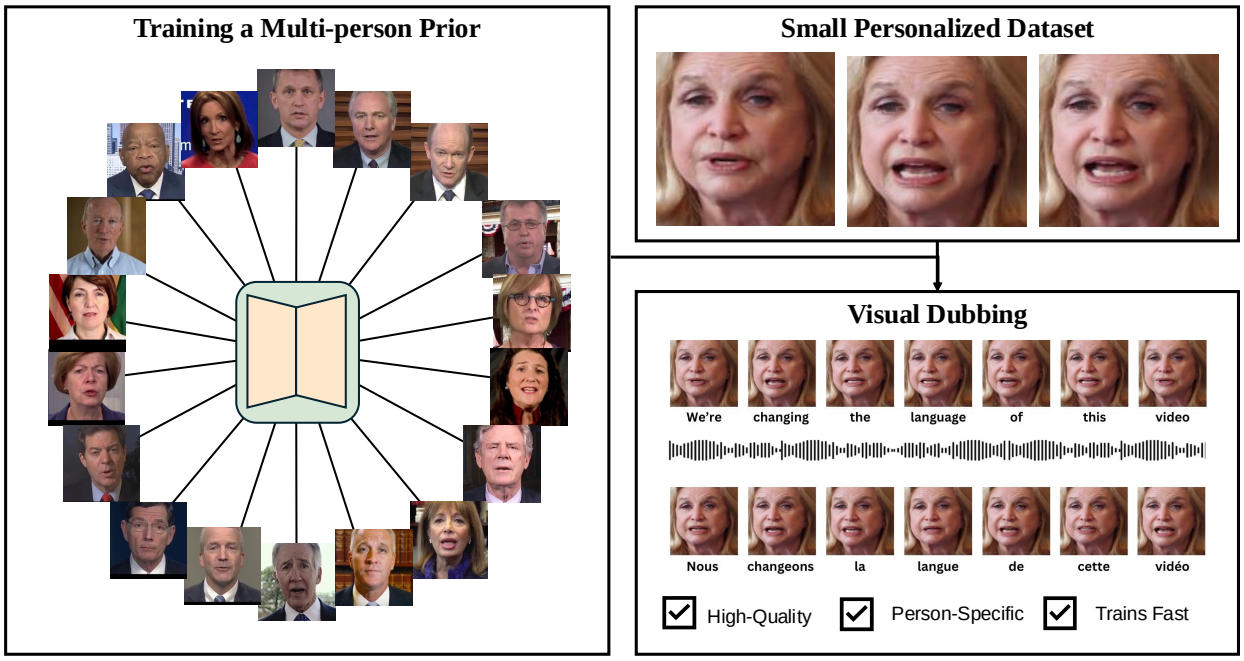

Figure 1: Our method, Dubbing for Everyone, allows for the reconstruction of lip movements when dubbing video from one language to another, using only a few seconds of training data. We do this by training a large-scale prior over many people, dramatically reducing the data requirements.

generalize across identities. We also maintain actor-specific components, designed to guide the prior, that allow our model to adapt to new individuals and capture actor-specific nuance. In particular, we adopt a multi-stage approach based on neural textures (Thies et al., 2020; 2019).

Our model is **generalisable** due to the person-generic prior network training on a large dataset. We demonstrate this in Table 1 by using as few as 4 seconds of actor-specific data and further discuss this quality in Section 5.6. It is **high-quality** due to the person-specific adaptation making effective use of all data, we achieve state-of-the-art in this respect as is shown in Table 1. We find an order-of-magnitude speedup compared to existing person-specific models (Section 5.4), leading to a similar order-of-magnitude cost reduction. By maintaining actor-specific components, or model is **recognisable** for any actor and does not appear overly generic. We validate this through a user study in Table 2. The novel contributions in this paper may be summarised as:

- We present **Dubbing for Everyone**, a visual dubbing model using person-generic and person-specific components, capable of producing **high-quality** and **idiosyncratic** results from just a few seconds of data.
- We train a prior deferred neural rendering network across many identities and learn actor-specific neural textures, allowing us to adapt our model to new identities. The training of the prior network allows for **data-efficent dubbing**, resulting in a significant reduction in data requirements compared to existing person-specific models.
- We propose a novel post-processing algorithm to remove artefacts in the border around the generated video. This improves perceived quality (Table 4).
- We perform an extensive evaluation to show that our method achieves state-of-the-art for **quality**(Table 1, Table 2, & Supplementary Video) and **recognisability**(Table 1), as well as being **an order of magnitude cheaper** (Section 5.4 & Table 3) and **robust in few-shot scenarios.** (Section 5.6 & Table 1)

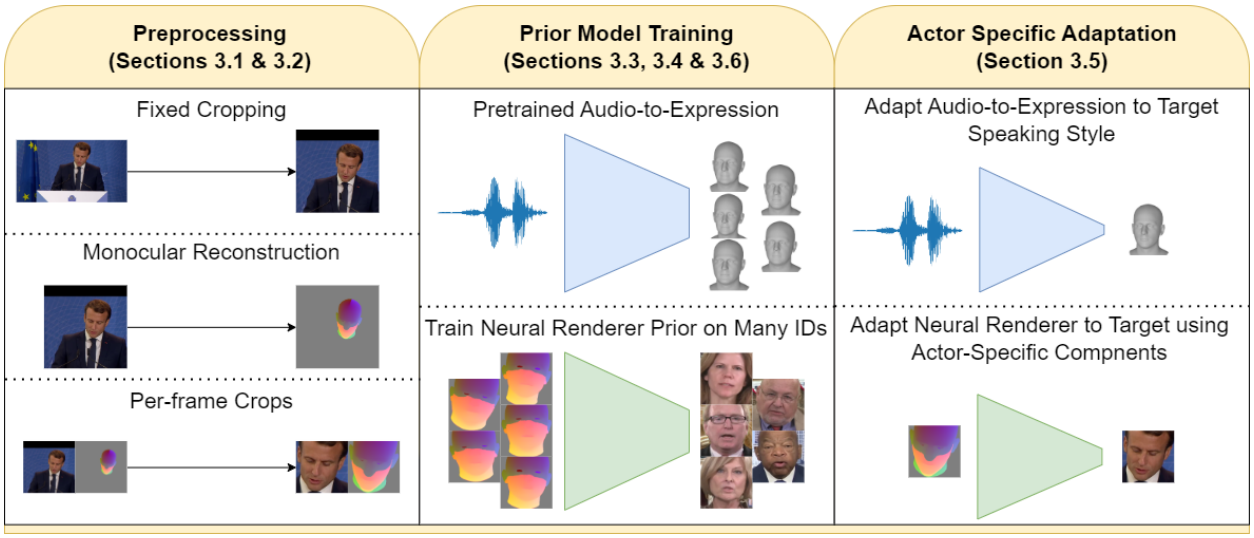

Figure 2: The pipeline of our method. We first apply preprocessing to our dataset (Section 3.2) to obtain 3D reconstructions, tightly and stably cropped to the face. We next obtain person-generic audio-to-expression and neural rendering models using multiple subjects (Section 3.4). Given a new subject, we then finetune both models for the given subject (Section 3.5).

2 Related Work

Early works in visual dubbing (e.g. Bregler et al. (1997); Ezzat et al. (2002)) use various methods. However, for this work, we consider post-deep learning models. Two separate classes of visual dubbing exist person-generic and person-specific models.

2.1 Person Generic Models

Person-generic models differ from person-specific models because they can work zero-shot for any new person and audio. These methods are typically 2D-based. One class of these models uses some form of expert discriminator to achieve lip sync. Early methods (Prajwal et al., 2020; K R et al., 2019) use an encoder-decoder model to predict frames from audio and use random reference frames of the same person. These methods use adversarial training combined with an expert Syncnet (Chung and Zisserman, 2016), which predicts if video and audio are in or out of sync. Most person-generic models build upon this framework but replace some of these components. Some seek to replace the encoder-decoder model with transformers (Wang et al., 2023b) or vector-quantised models (Gupta et al., 2023). Others change the type of expert discriminator (Wang et al., 2023a; Sun et al., 2022) or replace the adversarial loss with a diffusion process (Shen et al., 2023; Stypułkowski et al., 2023).

Another class of person-generic model works on the idea of flow. These models will estimate motion using either landmarks (Zhou et al., 2020; Ji et al., 2021; Wang et al., 2020; Gururani et al., 2023) or pixel-based flow (Siarohin et al., 2019). This motion is then used to warp a reference image.

In either case, these models use only a single reference frame to encode the identity. This is a significant issue as a single image cannot contain enough information about appearance or talking style. For example, if the mouth is closed in the reference frame, it is impossible to predict what the interior should look like. Our work, in contrast, can use all available frames of the target person for fine-tuning, enabling us to capture idiosyncratic qualities.

2.2 Person Specific Models

Person-specific models are trained per person, usually under controlled conditions. As a result of this, they are typically much higher quality than person-generic models but cannot produce results for anyone other than the person they were trained on. It is very common for person-specific models to use some form of 3D supervision in order to improve the results. By introducing 3D priors, certain characteristics, such as the face shape or pose, can be controlled for.

One line of work builds upon the 3D Morphable Model (Egger et al., 2020; Blanz and Vetter, 2023). The 3DMM allows pose, lighting, shape and texture to remain constant, only changing the expression of the face. Some works achieve dubbing by having one actor provide the lip motions for another (Kim et al., 2018; 2019), while others generate the lip motions from audio (Thies et al., 2020; Saunders and Namboodiri, 2023; Wen et al., 2020; Song et al., 2020; Ma et al., 2023).

Another, more recent, line of research builds upon implicit 3D geometry. In particular, Neural Radiance Fields (Mildenhall et al., 2020) (NeRFs) have been used to good effect for talking head generation. Audio-driven models (Guo et al., 2021; Shen et al., 2022; Tang et al., 2022; Ye et al., 2023) condition NeRFs on audio to produce free-viewpoint renderings. While promising, these results are often slow to train and render. Gaussian splatting-based models (Li et al., 2024; Cho et al., 2024) extend this line of work, achieving similar quality with much faster training and inference. However, they still require a lot of data and may suffer from artefacts. Several other works have also addressed the problem from a person-specific viewpoint and are less easily grouped. Models range from using a diffusion auto-encoder (Du et al., 2023) to person-specific landmark-based models (Lu et al., 2021; Suwajanakorn et al., 2017).

Person-specific models share some essential qualities. They all produce high-quality output but come with significant data requirements ranging from about 15 seconds Shen et al. (2022) to upwards of 10 minutes Du et al. (2023). In contrast, our method achieves similar quality using as little as 4 seconds of training data, thanks to our person-generic prior network training and person-specific adaptation.

2.3 Prior Learning for Faces

A few related methods have adopted a similar strategy of pretraining a prior model and fine-tuning to new identities. This work has been applied in face re-enactment (Burkov et al., 2020) and volumetric rendering for faces (Bühler et al., 2023). However, there has been little work using this method to visual dubbing. The most similar work to ours could be considered to be StyleSync (Guan et al., 2023). This method performs visual dubbing and has demonstrated an ability to adapt to new identities using fine-tuning. However, the model does not decouple person-specific and person-generic components, while ours does. This makes it less capable of capturing person-specific nuances (Table 1, Table 2). Another work, ROME (Khakhulin et al.), also uses a generalised version of the neural texture model for faces. However, it has no person-specific component, meaning it can only use one frame for personalisation. If, for example, it saw a frame with the mouth closed, it could never infer the open mouth. By contrast, our model effectively uses all available frames.

3 Method

Our method builds upon the deferred neural rendering approach of Thies et al. (2019). The key to our method is the training of a prior deferred neural rendering network (Section 3.4), which is person-generic, and the adaptation to new actors using neural textures (Section 3.5). This method (Figure 2) requires an order of magnitude less data than previous neural textures approaches. Before training the prior network, we run a preprocessing stage, which involves cropping the video frames to the face region and performing monocular reconstruction (Section 3.1) to get a parameterized 3D representation of the face.

Using an existing speech-driven animation model (Thambiraja et al., 2023), we are able to control the 3D model and, in turn, alter the lip motions of a given video (Section 3.6). Some artefacts are left during the video generation process, so we propose a postprocessing step to remove these (Section 3.7).

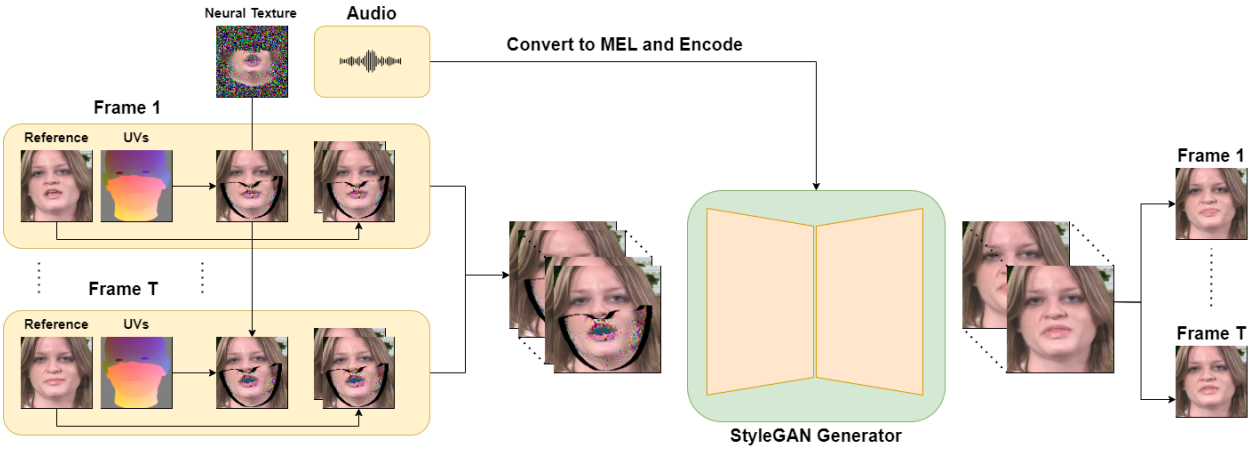

Figure 3: The architecture of our model. We take the UV rasterisations of each frame in a window of length T and sample a neural texture. We combine a mask with the real frame (Figure 4). These are concatenated with random reference frames of the same person and stacked across the channel dimension. A StyleGAN2-based (Guan et al., 2023; Karras et al., 2020) generator (see supplementary) is then used to convert these into T photorealistic frames.

3.1 Monocular Reconstruction

Following similar 3DMM-based neural texture approaches (Thies et al., 2020; 2019), we first fit a 3DMM to each frame of the ground truth videos using differentiable rendering. For the 3DMM, we use FLAME (Li et al., 2017). FLAME uses a combination of linear blendshapes and blend skinning to control a full-face rig with 5023 vertices. We use a three-stage process for this fitting. We cover more detail in the supplementary.

Stage 1: First, we estimate the shape parameters of the FLAME model using MICA (Zielonka et al., 2022). MICA predicts the shape parameters from a single frame and is shown to be very accurate. We then fix the shape parameters. **Stage 2:** Now, with the shape fixed, we optimise other non-varying parameters by jointly optimising a regularised photometric loss function over several frames simultaneously. We then fix the texture and camera parameters in addition to the shape. **Stage 3:** Finally, we optimise the variable parameters for each frame. These are expression, pose and lighting. We initialise parameters for frame t using the parameters for frame $t-1$.

3.2 Preprocessing

We next crop the face to 256×256 pixels. We find that pretrained face detectors ((Lugaresi et al., 2019; Deng et al., 2020)) suffer from two issues. The first is jitter between frames, and the second is the bounding box, which varies based on the jaw pose. We use our tracking data to generate bounding boxes to overcome these issues. We project the vertices of the meshes to obtain 2D landmarks using the parameters in Section 3.1. To prevent the jaw position from appearing in the box size, we project landmarks with the jaw in several positions and find the bounding box containing all these. We then add a small margin to get the final box. More details can be found in the supplementary material. We also use the tracked data to generate masks with the same multi-jaw approach. We rasterise a predefined mouth texture mask, which may be seen in Figure 4.

3.3 Architecture

In this section, we describe the architecture of our model. Inspired by previous work (Thies et al., 2019), we adopt a deferred neural rendering approach using neural textures. The model contains two components: learnable neural textures which are similar to standard RGB, uv-based, texture images with high-dimensional

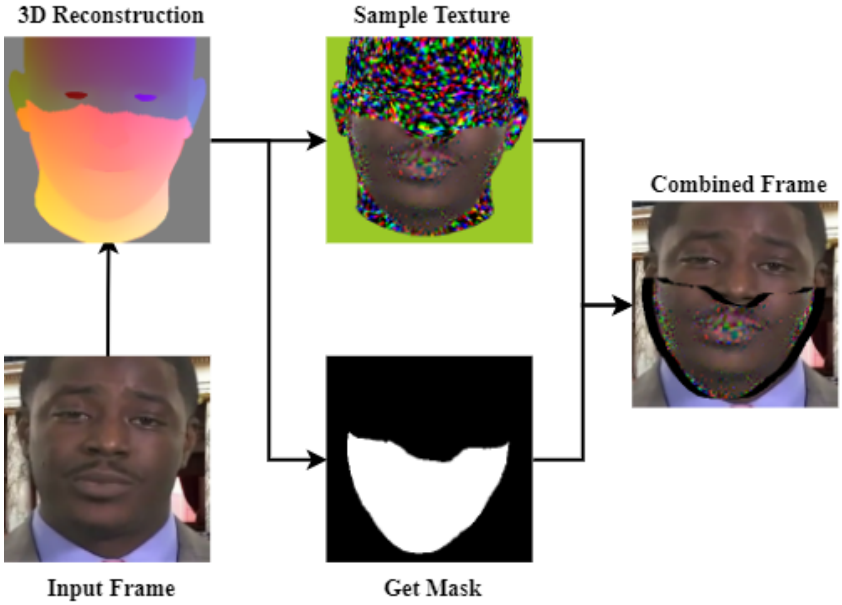

Figure 4: The input to the image-to-image network. We sample the neural texture onto the rasterised mesh and use the mesh to estimate a mask. The input is computed using this mask, the target frame, and the rasterised texture.

feature vectors; and a deferred neural renderer, an image-to-image network that takes these rasterised neural features and converts them into realistic video (Figure 3).

Existing models require several minutes of training data. Our work's primary novelty is adapting this method to work few-shot, using only a small dataset of a given actor. A key insight to obtaining this is to note that much of the person-specific information can be stored in the neural textures, allowing the image-to-image network to be generalised across multiple subjects. To help the image-to-image network generalise, we use a reference frame as is done in person-generic works (Prajwal et al., 2020; Gupta et al., 2023; Guan et al., 2023; Shen et al., 2023; Stypułkowski et al., 2023).

To improve the quality of the generations, we replace the UNET used in previous works with a modified version of the StyleGAN2 (Karras et al., 2020) architecture used in StyleSync (Guan et al., 2023). Instead of providing masked target frames as is done in StyleSync, we mask out the target frame using the rasterization of a predefined mask on the 3DMM (Section 3.2) and fill in the masked regions with the rasterized neural texels. This can be seen in Figure 4.

Using the generator architecture from StyleSync also allows us to condition the video generation on audio. To improve temporal stability, we provide the generator with access to a window of frames surrounding the target and predict the same window of the final video. The architecture is best understood by viewing Figure 3 and referring to the supplementary material and StyleSync paper (Guan et al., 2023). In short, however, we convert input audio into MEL-spectrograms and use a series of 2D, residual convolutional layers to get a latent representation of audio, which is used in place of style vectors in the StyleGAN-based generator (Karras et al., 2020).

3.4 Training the Prior Model

We use multiple identities to train the prior deferred neural rendering network model. The network weights are shared for all identities, but we have a different, randomly initialised neural texture for each. We jointly optimise the network and textures to minimise the following loss, as it has been shown to be effective in previous work (Thies et al., 2020):

$$\mathcal{L} = \lambda_1 \mathcal{L}_1 + \lambda_{\text{VGG}} \mathcal{L}_{\text{VGG}} + \lambda_{\text{reg}} \mathcal{L}_{\text{reg}} + \lambda_{\text{adv}} \mathcal{L}_{\text{adv}} \tag{1}$$

$\mathcal{L}_1$ is a simple ℓ_1 loss computed between the generated window of frames and the ground truth. To encourage the network to produce better results in the lower face and mouth region, we compute masks for these areas and weigh them higher. Specifically, we weigh the mouth region at ten times the background and the upper face and the lower face region at eight times the background.

$\mathcal{L}_{\text{VGG}}$ is a VGG-based (Johnson et al., 2016) style loss. This is computed using a pre-trained VGG network and is calculated for each frame in the window individually, taking the mean. This is a perceptual loss and is known to improve image quality.

$\mathcal{L}_{\text{adv}}$ is an adversarial loss. We jointly train the model with a discriminator and use an LSGAN (Mao et al., 2017) formulation for the adversarial loss. The discriminator is identical to the one used in StyleSync (Guan et al., 2023), but it is shown all frames in the window to encourage temporal consistency as demonstrated in previous works (Prajwal et al., 2020).

Finally, $\mathcal{L}_{\text{reg}}$ is a regularisation loss for the neural textures. It is computed as the ℓ_1 distance between the first three channels of the rasterised texture and the target frame. This encourages the first three channels of the texture to mimic a standard, diffuse RGB texture. λ values are given in section 4.

3.5 Adapting to New Identity

Given a new actor, we adapt our model to them. We first initialise a new random neural texture and use the deferred neural rendering prior network from section 3.4. We train the texture from scratch but use the prior network as initialisation for the deferred neural renderer. To help the model maintain its generalisation, we also include data from other identities from the training set of the generic model. Specifically, we include person-generic training data in the person-specific dataset at a ratio of 1:1. We refer to this as a **mixed training strategy**. The mixed training strategy allows the deferred neural rendering network to continue learning from a comprehensive data distribution, including, for example, poses that may not be in the actor-specific dataset. This effectiveness is shown in table 4.

3.6 Audio-to-Expression Model

Given our neural rendering model, we can convert rasterizations of the 3DMM to realistic video. To change the lip motions of the video, we need to alter the model's expression parameters. To do this, we make use of state-of-the-art speech-driven animation models. In particular, we use Imitator (Thambiraja et al., 2023). Imitator is a transformer-based model that allows for speaking style adaptation. We use the pre-trained Imitator model. To adapt to the speaking style of the new individual, we add a final layer to the network that independently applies a linear transformation for each expression and jaw pose parameter. We then train only this layer for each individual.

3.7 Post Processing

While our method produces high-quality results in the facial interior, it occasionally suffers from artefacts around the border between the face and background. Due to the strong bias introduced by the neural texture, pixels beyond the face region appear "stuck" to the face and follow its motion. This is best seen in video format (see the supplementary video). To reduce the effect of this artefact, we apply post-processing. We first apply a semantic segmentation network (Yu et al., 2021) to each generated and real frame. This separates the face and neck from the background. We can then replace the generated pixels with the real frame where both the generated and real frame agree the pixel is the background.

4 Implementation Details

To train our prior network, we used an Adam Optimiser with a learning rate of $1e-4$ and a batch size of 4. We set $\lambda_1 = 10.0$, $\lambda_{\text{adv}} = 1.0$, $\lambda_{\text{VGG}} = 10.0$, $\lambda_{\text{reg}} = 5.0$. We have one neural texture for each identity;

Method	HDTF 100 Frames						HDTF 1000 Frames						CelebV-HQ (Average 200 frames)					
	PSNR ↑	*SSIM* ↑	*FID* ↓	**Qual** ↑	**Lip**	↑ **ID** ↑	*PSNR* ↑	*SSIM* ↑	*FID* ↓	**Qual** ↑	**Lip** ↑	**ID** ↑	*PSNR* ↑	*SSIM* ↑	*FID* ↓	**Qual** ↑	**Lip** ↑	**ID** ↑
Wav2LipHQ Gupta et al. (2023)	27.70	0.895	6.78	3.53	3.67	3.10	27.70	0.895	6.78	3.53	3.67	3.10	24.81	0.832	15.84	3.31	2.38	4.08
StyleSync Guan et al. (2023)	29.26	0.899	7.07	3.77	3.50	3.20	29.26	0.899	7.07	3.77	3.50	3.20	29.14	0.895	9.26	3.23	2.77	3.38
TalkLip Wang et al. (2023a)	28.34	0.887	9.98	2.47	3.59	3.03	28.34	0.887	9.98	2.47	3.59	3.03	28.85	0.896	13.89	2.92	2.85	3.54
RAD-NeRF Tang et al. (2022)	22.63	0.732	30.63	2.17	1.67	2.75	26.23	0.835	22.94	2.33	1.67	3.00	20.37	0.65	32.12	1.92	2.15	3.23
GeneFace Ye et al. (2023)	21.87	0.720	38.50	2.00	1.92	2.77	24.43	0.797	26.38	2.67	2.75	3.08	21.04	0.665	45.29	1.85	2.00	3.00
Ours Baseline	27.92	0.888	10.52	2.90	3.33	3.13	29.00	0.899	8.61	3.06	3.23	3.10	26.61	0.870	16.29	2.31	3.23	3.23
Ours Full	29.10	0.899	5.76	3.57	3.77	3.83	29.30	0.904	5.44	3.90	3.80	4.03	30.17	0.912	5.35	3.38	3.69	4.46
Real	100.00	1.00	0.00	4.53	4.60	4.46	100.00	1.00	0.00	4.53	4.60	4.46	100.00	1.00	0.00	4.69	4.85	4.85

Table 1: Quantitative comparisons of our model with state-of-the-art. We compare in three settings, a very low data setting (100 frames) and a somewhat low data setting (1000 frames), both using HDTF and an unseen test set (CelebV-HQ). Our method is compared to person-specific models TalkLip (Wang et al., 2023a), Wav2LipHQ (Gupta et al., 2023) and StyleSync (Guan et al., 2023), and person-generic models including a baseline version of our model trained from scratch as well as GeneFace (Ye et al., 2023) and RAD-NeRF (Tang et al., 2022). We compare using quantitative (*italics*) and user ratings (**bold**). We highlight the Best and Second Best for each metric (excluding the ground truth).

these have a 256×256 resolution and 16 channels. The audio encoder and Generator architecture are taken from StyleSync (Guan et al., 2023), but the first convolutional layer is altered to reflect the difference in the image channels of the input (e.g. StyleSync uses 3-channel RGB and we use 16-channel neural texture features, see the appendix). The prior network is trained for seven days using an NVIDIA V100 GPU. The fine-tuning stage requires only 1-2 hours of training on the smaller L4 GPU.

5 Results

Data: We train our prior model using the HDTF Zhang et al. (2021) dataset. HDTF consists of around 400 videos, each in high-definition and several minutes long. The length of the videos is important as we want to run experiments with various video lengths, as is done in section 5.6. We select a random subset of 20 videos for finetuning and testing, and the rest is used for the generic pretraining stage. We manually inspect the dataset to ensure that the subjects in the training set are not also accidentally included in the test set. We resample each video to 25fps and use 1500 frames (1 minute). We use the last 10 seconds as testing data and subsets of the remaining 50s for fine-tuning. In addition, our model can generalise beyond the dataset on which it is trained. To show this, we also include results from a subset of CelebV-HQ Zhu et al. (2022) (note that the prior model does not see this dataset). CelebV-HQ videos are not as long, so we use the last 10 seconds for testing and all the remainder for fine-tuning. This is, on average, 200 frames.

Metrics: We look to evaluate our method on three criteria, **visual quality, lip sync** and **idiosyncracies**. Visual quality is measured using **FID**. As ground truth is available during re-enactment experiments, we also use **SSIM and PSNR**. While useful as proxies, these metrics are less important than how users perceive the method. For this reason, we also ask users to rate the three qualities: visual quality **(QUAL)**, idiosyncracies **(ID)** and lip-sync **(LIP)** out of 5. A total of 30 users provided ratings. Further details of this user study are provided in the supplementary material.

5.1 Comparisons to State-of-the-Art

We compare our model to the state-of-the-art. We separate these into person-specific and person-generic. To demonstrate the ability of our model to work for small and medium-sized datasets, we consider three scenarios: one with limited data (100 frames), one with significantly more data (1000 frames) and one using a different dataset (Zhu et al., 2022). We compare our model to three recent person-generic methods: Wav2LipHQ (Gupta et al., 2023), which uses a VQ-GAN to achieve ultra-high resolution outputs; TalkLip (Wang et al., 2023a), which uses a lip reading network for better lip-sync, and the StyleGAN2 (Karras et al., 2020) based StyleSync (Guan et al., 2023). We compare person-specific models to the NERF-based RAD-NeRF (Tang et al., 2022) and GeneFace (Ye et al., 2023). We also compare our work with a baseline model. For this, we train our model on each subject from scratch. We consider this a close re-implementation of similar pipelines (Thies et al., 2020; 2019). Therefore, we do not also compare our work to these models.

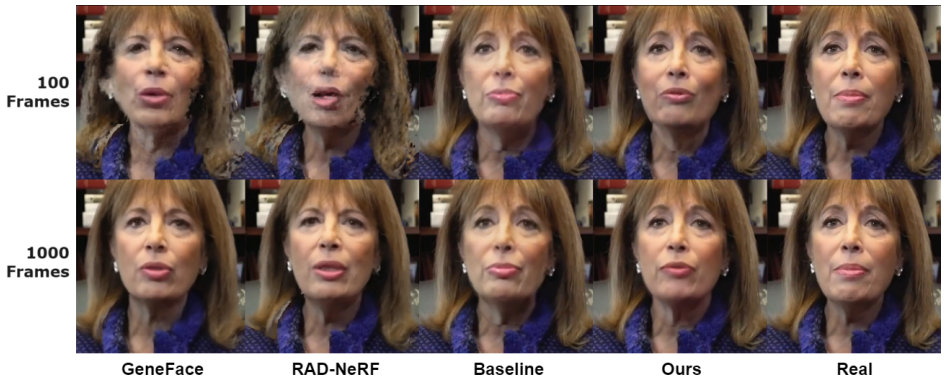

Figure 5: Qualitative comparisons to state-of-the-art person-specific models. We compare to GeneFace Ye et al. (2023), RAD-NeRF (Tang et al., 2022) and the baseline model trained from scratch.

Figure 6: Qualitative comparisons to the state-of-the-art person generic models. We compare to TalkLip (Wang et al., 2023a), Wav2LipHQ (Gupta et al., 2023) and StyleSync (Guan et al., 2023). We show two versions of our model, one fine-tuned on 100 frames of data and one with 1000 frames. We also show the ground truth frames for comparison. Closeups of the mouth region are included for more detail.

The results are shown in Table 1. **Our model outperforms the person-generic models in terms of visual quality as measured by FID**. This effect is more noticeable with additional data. User ratings for quality are slightly lower for our model with just 100 frames but higher with 1000. However, **our model can capture person-specific details that the generic models can not**. This is evidenced in Figure 6 and by the user ratings for idiosyncracies (ID). This may suggest that the generic models are producing visually appealing but generic-looking lips. Note that we do not use LSE metrics (Prajwal et al., 2020) for lip sync as some previous works do. We find that the person-generic models optimise this directly to the extent that they outperform the ground truth (9.34 for StyleSync vs 7.28 for real video), making this an unreliable metric. To further this point, users preferred the lip-sync of Wav2LipHQ (Rating=3.67) to StyleSync (3.50), but StyleSync outperforms it using LSE-C (9.34 vs 7.00), and this is more pronounced with real data (User Ratings = 4.6, LSE-C=7.28). Instead, we believe that the user experience of lip sync is the most important, and our model outperforms all others in this respect.

Compared to person-specific models, our method outperforms them across all metrics. **The difference is most prominent when trained on 4 seconds (100 Frames) of data, suggesting our model is using**

Row > Col % (95% CI)	Audio only	StyleSync	Baseline	Ours
Audio Only	-	**24** (16-33)	**23** (15-32)	**9** (4-16)
StyleSync	**76** (67-84)	-	**68** (58-77)	**38** (29-48)
Baseline	**77** (68-85)	**32** (23-42)	-	**9** (4-16)
Ours	**91** (84-96)	**62** (52-71)	**91** (84-96)	-

Table 2: User study performed on a translated section of an in-the-wild video clip. We show the percentage of 35 users who preferred the row to the column. We include 95% confidence intervals for each.

	Train Iterations to Reach PSNR					
Method	25	26	27	28	29	30
Baseline	2200	3000	6200	9800	18000	40000
Ours	200	300	300	400	700	1600
Speedup Factor	11	10	20.7	24.5	25.7	25

Table 3: Number of iterations (to the nearest 100) taken to reach a given PSNR value for our model and the baseline. Average of three runs with different identities.

the available data effectively. We further investigate this effect in Section 5.6. The NeRF-based models, in particular, fail with unseen poses when trained on just 100 frames. This can be seen easily in Figure 5.

5.2 User Study

To investigate our model in its intended context, altering the lip motion to match dubbed audio in a different language with limited data, we design a user study to replicate this. We take three videos of politicians speaking in their native language and use the automated (audio-only) dubbing provided by 11labs (Ele). These video clips are 15-20 seconds long, much shorter than those used in previous works (Thies et al., 2020; Wen et al., 2020; Ye et al., 2023). We compare our work to the highest-quality generic and specific models, measured by user ratings. We also consider audio-only dubbing (not altering the lips), which remains the industry standard. We perform a forced choice experiment. Users are given the same video dubbed using two random methods from our selection. The users are asked which they prefer. The results are in Table 2 and show that users prefer our method to all others within a 95% confidence interval. Further details of the user study are outlined in the supplementary material.

5.3 Ablations

We perform an ablation study of our model. We show that the post-processing step (Section 3.7) and the mixed training strategy for fine-tuning (Section 3.5) both improve the results of our model. We use the 100-frame setting for this experiment. The results can be seen in table 4. The mixed training strategy improves results across all metrics, showing that it helps the model generalise. The post-processing increases FID, suggesting worse visual quality. However, the user ratings show a preference for post-processing. This

	HDTF 100 Frames					
Method	PSNR ↑	SSIM ↑	FID ↓	Qual ↑	Lip ↑	ID ↑
Ours w/o post-processing	28.20	0.888	5.28	3.20	3.50	3.33
Ours w/o mixed data	28.95	0.897	5.88	3.17	3.50	3.33
Ours full	29.10	0.899	5.76	3.57	3.77	3.83

Table 4: Results of the ablation study. Including our post-processing step (Section 3.7) and mixed training strategy (Section 3.5) improves the results across many metrics. We highlight the best results.

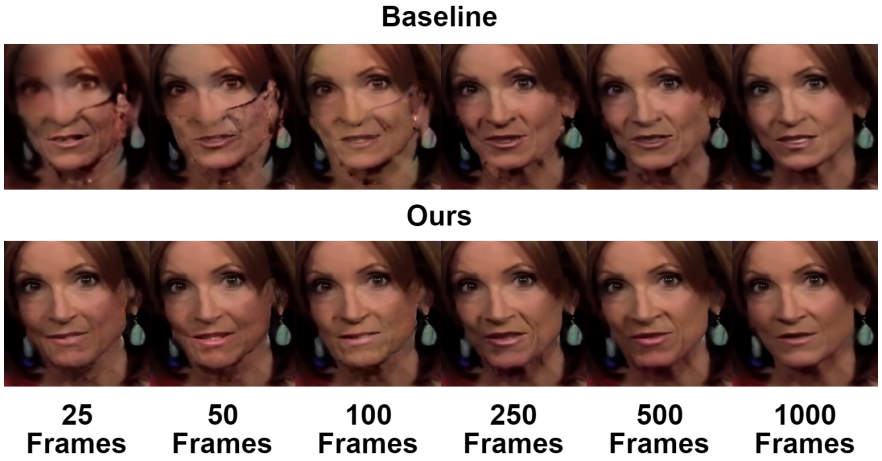

Figure 7: The effect training dataset size. The baseline model is trained from scratch and suffers when using limited data. Ours, by comparison, is far more robust.

may be because the post-processing noticeably removes the sharp boundary between the real and generated frame at the cost of some reconstruction accuracy.

5.4 Training Speed

Our model trains faster than existing person-specific models. This is due to the deferred neural rendering priors. As this network works across many identities, adapting to a new identity requires much less training. To demonstrate this effect, we compare our model to the baseline. This model is trained from scratch for each new identity, which does not use a prior network. We show that our model is faster to train by recording the number of training iterations required to reach a given value of PSNR on a withheld validation set. To show that this effect is not just due to limited data making the baseline weaker, we train on the 1000 frame setting. The results are shown in Table 3 and show clearly that **our model converges an order of magnitude faster than the baseline**. This takes 1-2 hours on an L4 GPU. For comparison, Geneface (Ye et al., 2023) takes approximately 80 hours, and RaD-NeRF (Tang et al., 2022) takes 7, meaning the costs of our model are 1-2 orders of magnitude lower than these state-of-the-art models.

5.5 Inference Speed

Our prior is not required at inference time, as it is only used to train the model. The additional steps we require for audio-to-expression generation and rasterization are very fast compared with the heavy image-to-image generator network used in these methods. This means that our model runs close to the same inference speed as image-to-image models (e.g. Thies et al. (2020); Guan et al. (2023); Gupta et al. (2023)) In practice, we can generate frames at around 5fps on an $L4$ GPU without any specific code optimisation.

5.6 Effect of Dataset Size

Our method allows dubbing actors with only a few seconds of data. To demonstrate this ability, we compare our prior network method to a baseline model, which trains both the deferred neural rendering network and neural texture from scratch. We evaluate the model using 10-second clips of one of our target actors, having trained both models on subsets of the training data of various sizes. The results are shown in Table 5, Figure 7 and in the supplementary video. It can be seen that **our model produces much higher quality results than the baseline on the small datasets**, but this effect reduces with dataset size.

N Frames	1	10	25	50	100	250	500	1000
Baseline	33.19	30.07	27.65	23.10	17.71	12.33	11.84	11.78
Ours	15.53	14.56	13.28	12.40	11.37	11.88	11.74	11.21

Table 5: Our model is robust even on very small datasets. We compare the number of frames of a given actor to the FID obtained by the model trained on this dataset.

6 Limitations and Future Work

While our method achieves state-of-the-art, is robust to small datasets and trains fast, it is not without limitations. First, there are noticeable artefacts around the border between the face and the background. The post-processing we introduce does mitigate this, but not completely. We think this could be addressed by first segmenting the background from the person, training the model on the foreground only and composing the result. We will address this in future work. Another significant issue is that the monocular reconstruction stage is very slow as it relies on optimisation through a differentiable renderer. Recent work has shown (Feng et al., 2021; Danecek et al., 2022; Filntisis et al., 2022) that regression-based reconstruction is possible, but it is still not temporally consistent enough for our purposes. We would like to investigate temporal regression models to this end, that could run in real-time.

7 Ethical Discussion

It is of paramount importance that the benefits and harms of the field of audio-driven visual dubbing are correctly weighted. This section discusses these and details our attempts to mitigate any harm.

Privacy: The HDTF dataset is released under the CC BY 4.0 License. We, therefore, have permission from the authors to use this dataset. To help protect the privacy of the individuals in this dataset and comply with GDPR, we will provide a contact form allowing any individuals to remove themselves from the dataset and model. As our model is two-part, consisting of neural textures and a generic rendering network, the model cannot reconstruct an individual without their neural texture. Simply deleting the texture will ensure that the individual is no longer represented in the model.

Associated Harms: The potential for misusing our technology is significant. ‘Deepfakes" refers broadly to the class of artificially generated videos of which our work may be considered. These models are known to cause harm through misinformation, defamation and non-consensual explicit material, although our work cannot be used for the latter. To help mitigate these harms, we will only provide access to the model to researchers at an accredited institution. Furthermore, we are investigating invisible watermarking (Navas et al., 2008; Bui et al., 2023) and deepfake detection methods (Rössler et al., 2019; Mirsky and Lee, 2021).

Associated Benefits: Our model enables media to cross language barriers. This helps promote diverse societies and allows for the spread of various cultures. In addition to this, the method has significant potential economic value. In these ways, the development of such models can benefit societies.

The exact weighting of the good and harms of developing “deepfake" models remains an open question. Still, we believe that visual dubbing, with its potential for spreading culture and the economic benefits, outweigh the potential risks when considering the mitigation we have put in place.

8 Conclusion

We have presented Dubbing for Everyone. Unlike person-specific models, which require several minutes of personalized data, or person-generic models, which cannot capture personalised appearances, our model is a hybrid person-generic, person-specific model. Using adaptation, our model is capable of high-quality, personalized visual dubbing using a few seconds of data for a given actor. Our experiments have shown that our model archives state-of-the-art across many metrics, including user ratings. We have also shown that our person-generic prior network training and adaptation strategy **trains faster, reaches higher quality and works on less data** than a similar model trained without priors.

References

Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural voice puppetry: Audio-driven facial reenactment. *ECCV 2020*, 2020.

Zhenhui Ye, Ziyue Jiang, Yi Ren, Jinglin Liu, Jinzheng He, and Zhou Zhao. Geneface: Generalized and high-fidelity audio-driven 3d talking face synthesis. *International Conference on Learning Representations*, 2023.

Jiaxiang Tang, Kaisiyuan Wang, Hang Zhou, Xiaokang Chen, Dongliang He, Tianshu Hu, Jingtuo Liu, Gang Zeng, and Jingdong Wang. Real-time neural radiance talking portrait synthesis via audio-spatial decomposition. *arXiv preprint arXiv:2211.12368*, 2022.

Jiadong Wang, Xinyuan Qian, Malu Zhang, Robby T Tan, and Haizhou Li. Seeing what you said: Talking face generation guided by a lip reading expert. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14653–14662, 2023a.

Anchit Gupta, Rudrabha Mukhopadhyay, Sindhu Balachandra, Faizan Farooq Khan, Vinay P. Namboodiri, and C. V. Jawahar. Towards generating ultra-high resolution talking-face videos with lip synchronization. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5198–5207, 2023. doi: 10.1109/WACV56688.2023.00518.

Jiazhi Guan, Zhanwang Zhang, Hang Zhou, Tianshu HU, Kaisiyuan Wang, Dongliang He, Haocheng Feng, Jingtuo Liu, Errui Ding, Ziwei Liu, and Jingdong Wang. Stylesync: High-fidelity generalized and personalized lip sync in style-based generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019.

Christoph Bregler, Michele Covell, and Malcolm Slaney. Video rewrite: Driving visual speech with audio. In *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '97, page 353–360, USA, 1997. ACM Press/Addison-Wesley Publishing Co. ISBN 0897918967. doi: 10.1145/258734.258880. URL https://doi.org/10.1145/258734.258880.

Tony Ezzat, Gadi Geiger, and Tomaso Poggio. Trainable videorealistic speech animation. *ACM Trans. Graph.*, 21(3): 388–398, jul 2002. ISSN 0730-0301. doi: 10.1145/566654.566594. URL https://doi.org/10.1145/566654.566594.

K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. 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*, MM '20, page 484–492, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379885.

Prajwal K R, Rudrabha Mukhopadhyay, Jerin Philip, Abhishek Jha, Vinay Namboodiri, and C V Jawahar. Towards automatic face-to-face translation. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, 2019. ISBN 978-1-4503-6889-6.

J. S. Chung and A. Zisserman. Out of time: automated lip sync in the wild. In *Workshop on Multi-view Lip-reading, ACCV*, 2016.

Jiayu Wang, Kang Zhao, Shiwei Zhang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Lipformer: High-fidelity and generalizable talking face generation with a pre-learned facial codebook. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13844–13853, June 2023b.

Yasheng Sun, Hang Zhou, Kaisiyuan Wang, Qianyi Wu, Zhibin Hong, Jingtuo Liu, Errui Ding, Jingdong Wang, Ziwei Liu, and Koike Hideki. Masked lip-sync prediction by audio-visual contextual exploitation in transformers. In *SIGGRAPH Asia 2022 Conference Papers*, SA '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450394703. doi: 10.1145/3550469.3555393. URL https://doi.org/10.1145/3550469.3555393.

Shuai Shen, Wenliang Zhao, Zibin Meng, Wanhua Li, Zheng Zhu, Jie Zhou, and Jiwen Lu. Difftalk: Crafting diffusion models for generalized audio-driven portraits animation. In *CVPR*, 2023.

Michał Stypułkowski, Konstantinos Vougioukas, Sen He, Maciej Zięba, Stavros Petridis, and Maja Pantic. Diffused Heads: Diffusion Models Beat GANs on Talking-Face Generation. In *https://arxiv.org/abs/2301.03396*, 2023.

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.

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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *ECCV*, 2020.

Siddharth Gururani, Arun Mallya, Ting-Chun Wang, Rafael Valle, and Ming-Yu Liu. Space: Speech-driven portrait animation with controllable expression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20914–20923, October 2023.

Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *Conference on Neural Information Processing Systems (NeurIPS)*, December 2019.

Bernhard Egger, William A. P. Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, Christian Theobalt, Volker Blanz, and Thomas Vetter. 3d morphable face models—past, present, and future. *ACM Trans. Graph.*, 2020.

Volker Blanz and Thomas Vetter. *A Morphable Model For The Synthesis Of 3D Faces*. Association for Computing Machinery, New York, NY, USA, 1 edition, 2023. ISBN 9798400708978. URL `https://doi.org/10.1145/3596711.3596730`.

Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Nießner, Patrick Pérez, Christian Richardt, Michael Zollöfer, and Christian Theobalt. Deep video portraits. *ACM Transactions on Graphics (TOG)*, 37(4):163, 2018.

Hyeongwoo Kim, Mohamed Elgharib, Hans-Peter Zollöfer, Michael Seidel, Thabo Beeler, Christian Richardt, and Christian Theobalt. Neural style-preserving visual dubbing. *ACM Transactions on Graphics (TOG)*, 38(6):178:1–13, 2019.

Jack Saunders and Vinay P. Namboodiri. Read avatars: Realistic emotion-controllable audio driven avatars. In *British Machine Vision Conference (BMVC)*, 2023.

Xin Wen, Miao Wang, Christian Richardt, Ze-Yin Chen, and Shi-Min Hu. Photorealistic audio-driven video portraits. *IEEE Transactions on Visualization and Computer Graphics*, 26(12):3457–3466, December 2020. doi: 10.1109/TVCG.2020.3023573. URL `https://richardt.name/audio-dvp/`.

Linsen Song, Wayne Wu, Chen Qian, Ran He, and Chen Change Loy. Everybody's talkin': Let me talk as you want. *arXiv preprint arxiv:2001.05201*, 2020.

Yifeng Ma, Suzhen Wang, Zhipeng Hu, Changjie Fan, Tangjie Lv, Yu Ding, Zhidong Deng, and XinYu. Styletalk: One-shot talking head generation with controllable speaking styles. In *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI)*, 2023.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

Yudong Guo, Keyu Chen, Sen Liang, Yongjin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Shuai Shen, Wanhua Li, Zheng Zhu, Yueqi Duan, Jie Zhou, and Jiwen Lu. Learning dynamic facial radiance fields for few-shot talking head synthesis. 2022.

Jiahe Li, Jiawei Zhang, Xiao Bai, Jin Zheng, Xin Ning, Jun Zhou, and Lin Gu. Talkinggaussian: Structure-persistent 3d talking head synthesis via gaussian splatting. *arXiv preprint arXiv:2404.15264*, 2024.

Kyusun Cho, Joungbin Lee, Heeji Yoon, Yeobin Hong, Jaehoon Ko, Sangjun Ahn, and Seungryong Kim. Gaussiantalker: Real-time high-fidelity talking head synthesis with audio-driven 3d gaussian splatting, 2024.

Chenpeng Du, Qi Chen, Tianyu He, Xu Tan, Xie Chen, Kai Yu, Sheng Zhao, and Jiang Bian. Dae-talker: High fidelity speech-driven talking face generation with diffusion autoencoder. *arXiv preprint arXiv:2303.17550*, 2023.

Yuanxun Lu, Jinxiang Chai, and Xun Cao. Live Speech Portraits: Real-time photorealistic talking-head animation. *ACM Transactions on Graphics*, 40(6), 2021. doi: 10.1145/3478513.3480484.

Supasorn Suwajanakorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: Learning lip sync from audio. *ACM Trans. Graph.*, 36(4), jul 2017. ISSN 0730-0301. doi: 10.1145/3072959.3073640. URL `https://doi.org/10.1145/3072959.3073640`.

Egor Burkov, Igor Pasechnik, Artur Grigorev, and Victor Lempitsky. Neural head reenactment with latent pose descriptors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Marcel C Bühler, Kripasindhu Sarkar, Tanmay Shah, Gengyan Li, Daoye Wang, Leonhard Helminger, Sergio Orts-Escolano, Dmitry Lagun, Otmar Hilliges, Thabo Beeler, et al. Preface: A data-driven volumetric prior for few-shot ultra high-resolution face synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3402–3413, 2023.

Taras Khakhulin, Vanessa Sklyarova, Victor Lempitsky, and Egor Zakharov.

Balamurugan Thambiraja, Ikhsanul Habibie, Sadegh Aliakbarian, Darren Cosker, Christian Theobalt, and Justus Thies. Imitator: Personalized speech-driven 3d facial animation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20621–20631, October 2023.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proc. CVPR*, 2020.

Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. URL `https://doi.org/10.1145/3130800.3130813`.

Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. In *European Conference on Computer Vision (ECCV)*. Springer International Publishing, October 2022.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. Mediapipe: A framework for perceiving and processing reality. In *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019*, 2019. URL `https://mixedreality.cs.cornell.edu/s/NewTitle_May1_MediaPipe_CVPR_CV4ARVR_Workshop_2019.pdf`.

Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *CVPR*, 2020.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016.

Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017.

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*, 129:3051–3068, 2021.

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 the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3661–3670, 2021.

Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. CelebV-HQ: A large-scale video facial attributes dataset. In *ECCV*, 2022.

Eleven labs dubbing. `https://elevenlabs.io/dubbing`. Accessed: 2023-06-11.

Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. volume 40, 2021. URL https://doi.org/10.1145/3450626.3459936.

Radek Danecek, Michael J. Black, and Timo Bolkart. EMOCA: Emotion driven monocular face capture and animation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20311–20322, 2022.

Panagiotis P. Filntisis, George Retsinas, Foivos Paraperas-Papantoniou, Athanasios Katsamanis, Anastasios Roussos, and Petros Maragos. Visual speech-aware perceptual 3d facial expression reconstruction from videos. *arXiv preprint arXiv:2207.11094*, 2022.

K. A. Navas, Mathews Cheriyan Ajay, M. Lekshmi, Tampy S. Archana, and M. Sasikumar. Dwt-dct-svd based watermarking. In *2008 3rd International Conference on Communication Systems Software and Middleware and Workshops (COMSWARE '08)*, pages 271–274, 2008. doi: 10.1109/COMSWA.2008.4554423.

Tu Bui, Shruti Agarwal, Ning Yu, and John Collomosse. Rosteals: Robust steganography using autoencoder latent space. In *Proc. CVPR WMF*, 2023.

Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *International Conference on Computer Vision (ICCV)*, 2019.

Yisroel Mirsky and Wenke Lee. The creation and detection of deepfakes: A survey. *ACM Computing Surveys (CSUR)*, 54(1):1–41, 2021.

Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016.

Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular rgb videos. *arXiv preprint arXiv:2112.01554*, 2021.

Youtube: European central bank. https://www.youtube.com/@ecbeuro/videos. Accessed: 2023-07-11.