# DEAD: Data-Efficient Audiovisual Dubbing using Neural Rendering Priors

Jack R. Saunders
https://jsaunders909.github.io/
Vinay P. Namboodiri
https://vinaypn.github.io/

University of Bath
Bath, UK

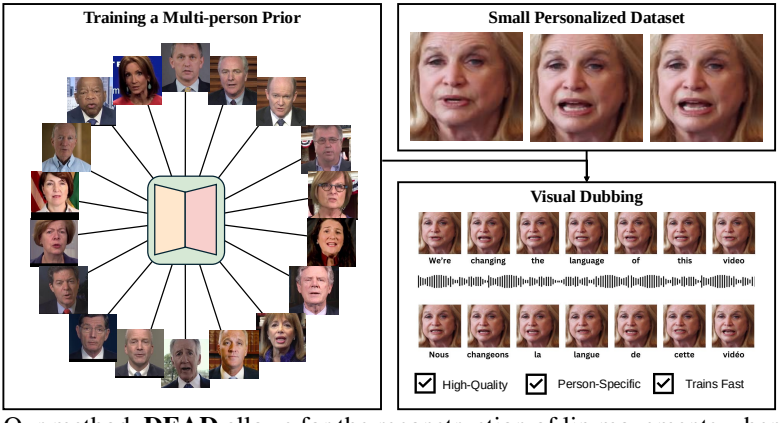


Figure 1: Our method, **DEAD** allows for the reconstruction of lip movements when dubbing video from one language to another, using only a few seconds of training data. We do this by training a large-scale prior over many people, dramatically reducing the data requirements.

**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 person-specific models see only one frame of the actor and, therefore, lack the ability to capture identity in the form of characteristic motion and related idiosyncracies, or they are expensive methods requiring off-putting large data requirements 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. 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 **is able to adapt to small datasets better than baselines**. Our experiments on real-world, limited data scenarios find that our model is preferred over baseline models.

# 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 [7].

We argue that any successful video dubbing method must be **high-quality**, **generalizable**, **recognisable** 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 who only appear for a few seconds of high-quality dialogue, should be dubbed effectively with as little as a few seconds of dialogue. An actor's idiosyncratic style should be **recognisable** 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. [24, 27, 32]). Such methods are high-quality and recognisable 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 generalisable video (e.g. [9, 10, 28]. 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 5.

We propose **DEAD** (Data Efficient Audiovisual Dubbing). 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 that Neural Rendering models can be de-coupled into person-generic and person-specific components and that they benefit from a large-scale pre-training of generic ones. We call this a prior network. Our prior network is trained across multiple actors and can generalise 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 [26, 27]. The novel contributions in this paper may be summarised as:

- We present **DEAD**, 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 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 than existing person-specific models** (supplementary) and **robust in few-shot scenarios.** (Section 4.4 & Table 1)

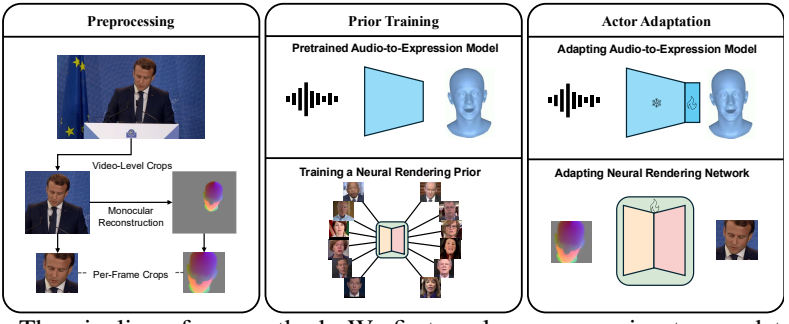


Figure 2: The pipeline of our method. We first apply preprocessing to our dataset. We next obtain person-generic audio-to-expression and neural rendering models using multiple subjects (Section 3.2). Given a new subject, we then finetune both models (Section 3.3).

# 2 Related Work

## 2.1 Person Generic Models

Person-generic models differ from person-specific models as they do not require any data from a given subject for training or fine-tuning. They only see a single reference frame given as input. One class of these models uses some form of expert discriminator to achieve lip sync. Early methods [12, 18] 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 [4]. Most person-generic models build upon this framework but replace some of these components. Some seek to replace the encoder-decoder model with transformers [29] or vector-quantised models [10]. Others change the type of expert discriminator [23, 28] or replace the adversarial loss with a diffusion process [16, 20, 22, 33].

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, requiring 2-5 minutes of data [27]. As a result, they are typically higher quality than person-generic models but cannot produce results for anyone other than the person they were trained on. In contrast, our method achieves similar quality using as little as 4 seconds of training data and can be adapted. Many works build upon the 3D Morphable Model [3, 6]. This 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 [14, 15], while others generate the lip motions from audio [19, 21, 27, 30]. While some of these models train a prior for the lip motion, none do so for the appearance. The most similar work to ours is StyleSync [9]. 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).

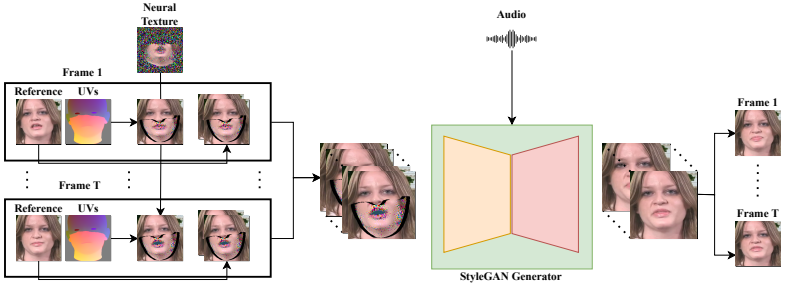


Figure 4: The architecture of our model. We take the UV rasterisations of each frame in a window of length T and sample a neural texture. We combine a mask with the real frame (Figure 3). These are concatenated with random reference frames of the same person and stacked across the channel dimension. A StyleGAN2-based [9, 13] generator (see supplementary) is then used to convert these into T photorealistic frames.

# 3 Method

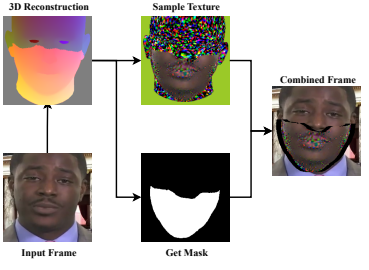


Figure 3: The input to the image-to-image network. We sample the neural texture onto the rasterised mesh and compute a mask. The input is computed using this mask, the target frame, and the rasterised texture.

Our method builds upon the deferred neural rendering approach of Thies et al. [26]. The key to our method is the training of a prior deferred neural rendering network (Section 3.2), which is person-generic, and the adaptation to new actors using neural textures (Section 3.3). This method (Figure 2) requires an order of magnitude less data than previous neural textures approaches. Using an existing speech-driven animation model [25], we are able to control the 3D model and, in turn, alter the lip motions of a given video (Section 3.4). We also include pre and post processing steps in the supplementary material.

## 3.1 Architecture

In this section, we describe the architecture of our model. Inspired by previous work [26], 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 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 4).

Existing ***person-specific*** models require several minutes of ***subject-specific*** 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 [9, 10, 18, 20, 22].

To improve the quality of the generations, we replace the UNet used in previous works with a modified version of the StyleGAN2 [13] architecture shown to be superior in StyleSync

[9]. 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 and fill in the masked regions with the rasterized neural texels. This can be seen in Figure 3.

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, as is shown effective in Wav2Lip [18]. The architecture is best understood by viewing Figure 4 and referring to the supplementary material and StyleSync paper [9]. 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 [13].

## 3.2 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 [27]:

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

$\mathcal{L}_1$ is a simple $\ell_1$ 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. This mouth weighting has been shown to be effective in previous work [31]. $\mathcal{L}_{\text{VGG}}$ is a VGG-based [11] 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 [17] formulation for the adversarial loss. The discriminator is identical to the one used in StyleSync [9], but it is shown all frames in the window to encourage temporal consistency as demonstrated in previous works [18]. Finally, $\mathcal{L}_{\text{reg}}$ is a regularisation loss for the neural textures. It is computed as the $\ell_1$ distance between the first three channels of the rasterised texture and the target frame.

## 3.3 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 3.2. 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 3.

| Method | HDTF 100 Frames | | | | | | | HDTF 1000 Frames | | | | | | | CelebV-HQ (Average 200 frames) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *PSNR* ↑ | *SSIM* ↑ | *FID* ↓ | *FVD*↓ | **Qual** ↑ | **Lip** | ↑**ID** ↑ | *PSNR* ↑ | *SSIM* ↑ | *FID* ↓ | *FVD* ↓ | **Qual** ↑ | **Lip** ↑ | **ID** ↑ | *PSNR* ↑ | *SSIM* ↑ | *FID* ↓ | *FVD* ↓ |
| Wav2LipHQ [34] | 27.70 | 0.895 | 6.78 | 100.52 | 3.53 | 3.67 | 3.10 | 27.70 | 0.895 | 6.78 | 100.52 | 3.53 | 3.67 | 3.10 | 24.81 | 0.832 | 15.84 | 270.96 |
| StyleSync [8] | 29.26 | 0.899 | 7.07 | 99.10 | 3.77 | 3.50 | 3.20 | 29.26 | 0.899 | 7.07 | 99.10 | 3.77 | 3.50 | 3.20 | 29.14 | 0.895 | 9.26 | 285.75 |
| TalkLip [25] | 28.34 | 0.887 | 9.98 | 128.53 | 2.47 | 3.59 | 3.03 | 28.34 | 0.887 | 9.98 | 128.53 | 2.47 | 3.59 | 3.03 | 28.85 | 0.896 | 13.89 | 346.01 |
| DiffDub [24] | 26.64 | 0.855 | 7.87 | 178.73 | 3.09 | 3.11 | 3.80 | 26.64 | 0.855 | 7.87 | 178.73 | 3.09 | 3.11 | 3.80 | 23.81 | 0.789 | 16.52 | 186.05 |
| MuseTalk [23] | 28.74 | 0.899 | 5.95 | 119.78 | 2.80 | 3.03 | 3.57 | 28.74 | 0.899 | 5.95 | 119.78 | 2.80 | 3.03 | 3.57 | 25.01 | 0.848 | 14.90 | 202.69 |
| TalkLip-FT [25] | 29.19 | 0.887 | 28.97 | 1218.89 | 2.57 | 3.06 | 3.60 | 27.95 | 0.864 | 10.21 | 126.95 | 2.69 | 2.86 | 3.57 | 28.42 | 0.870 | 25.97 | 254.04 |
| DiffDub-FT [24] | 26.71 | 0.857 | 7.69 | 154.12 | 3.17 | 3.31 | 3.69 | 26.77 | 0.858 | 7.73 | 166.04 | 3.29 | 3.23 | 3.80 | 23.84 | 0.799 | 18.120 | 192.86 |
| Ours Baseline | 27.92 | 0.888 | 10.52 | 138.06 | 2.90 | 3.33 | 3.13 | 29.00 | 0.899 | 8.61 | 126.45 | 3.06 | 3.23 | 3.10 | 26.61 | 0.870 | 16.29 | 187.20 |
| Ours Full | 29.10 | 0.899 | 5.76 | 99.10 | 3.57 | 3.77 | 3.83 | 29.30 | 0.904 | 5.44 | 94.161 | 3.90 | 3.80 | 4.03 | 30.17 | 0.912 | 5.35 | 86.65 |
| Real | 100.00 | 1.00 | 0.00 | 0.00 | 4.53 | 4.60 | 4.46 | 100.00 | 1.00 | 0.00 | 0.00 | 4.53 | 4.60 | 4.46 | 100.00 | 1.00 | 0.00 | 0.00 |

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). We compare using quantitative (*italics*) and user ratings (**bold**). We highlight the Best and Second Best for each metric (excluding the ground truth).

### 3.4 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 [25]. 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.

## 4 Results

**Data:** We train our prior model using the HDTF [34] 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 4.4. 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 [35] (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. We randomly select a further 20 videos from this dataset with the 'talking' annotation.

**Metrics:** We look to evaluate our method on three criteria, **visual quality, lip sync** and **idiosyncracies**. Visual quality is measured using **FID** at the image level and FVD at the video level. 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.

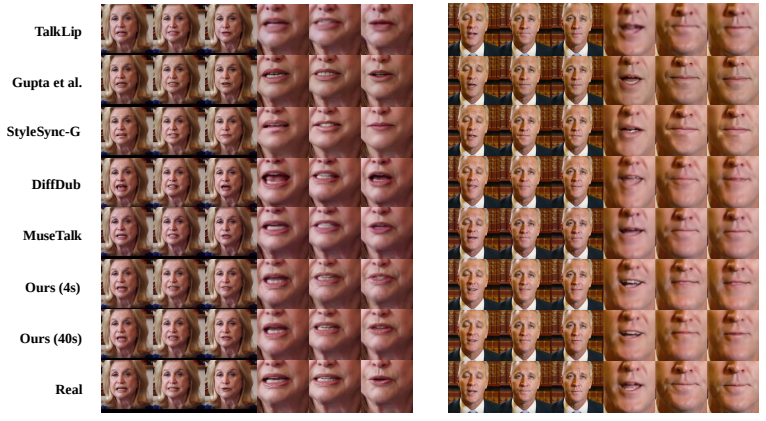


Figure 5: Qualitative comparisons to the state-of-the-art person generic models. We compare to TalkLip [28], Wav2LipHQ [10] and StyleSync [9]. We show two versions of our model, one fine-tuned on 100 frames of data and one with 1000 frames. We also show the ground truth frames for comparison. Closeups of the mouth region are included for more detail.

## 4.1 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 [35]. We compare our model to three recent person-generic methods: that of Gupta et al. [10], which uses a VQ-GAN to achieve ultra-high resolution outputs and we refer to as Wav2LipHQ; TalkLip [28], which uses a lip reading network for better lip-sync the StyleGAN2 [13] based StyleSync [9] as well as the diffusion-based DiffDub [16] and MuseTalk [33]. For more details on the baselines, please refer to the supplementary. We also compare our work with a baseline model where we train our model on each subject from scratch.

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. **Our model can capture person-specific details that the generic models can not**. This is evidenced in Figure 5 and by the user ratings for idiosyncracies (ID). This may suggest that the generic models are producing visually appealing but generic-looking lips.

| 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 in brackets.

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 the available data effectively**. We further investigate this effect in Section 4.4. The NeRF-based models, in particular, fail with unseen poses when trained on just 100 frames. This can be seen easily in Figure 6.


Figure 6: Qualitative comparisons to state-of-the-art person-specific models. We compare to fine-tuned versions of TalkLip [28] and DiffDub [16], as well as a baseline model trained from scratch.

## 4.2 User Study

We design a 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 take three videos of politicians speaking in their native language and use the automated (audio-only) dubbing [1]. These video clips are 15-20 seconds long, much shorter than those used in previous works. We compare our work to the highest-quality baseline model, 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.

## 4.3 Ablations

| Method | HDTF 100 Frames | | | | | | |
|---|---|---|---|---|---|---|---|
| | PSNR ↑ | SSIM ↑ | FID ↓ | *FVD* ↓ | **Qual** ↑ | **Lip** ↑ | **ID** ↑ |
| Ours w/o post-processing | 28.20 | 0.888 | 5.28 | 98.10 | 3.20 | 3.50 | 3.33 |
| Ours w/o mixed data | 28.95 | 0.897 | 5.88 | 107.30 | 3.17 | 3.50 | 3.33 |
| **Ours full** | 29.10 | 0.899 | 5.76 | 99.10 | 3.57 | 3.77 | 3.83 |

Table 3: Results of the ablation study, including our post-processing step (supplementary) and mixed training strategy (Section 3.3). We highlight the best results.

We perform an ablation study of our model. We show that the post-processing step and the mixed training strategy for fine-tuning both improve the results of our model (Table 3). We use the 100-frame setting for this experiment. For further details, see the supplementary.

## 4.4 Effect of Dataset Size

Our method allows dubbing actors with only a few seconds of data. To demonstrate this, 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 a target actor, having trained both models on subsets of the training data of various sizes. The results are shown in Table 4, 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 is reduced on large ones.

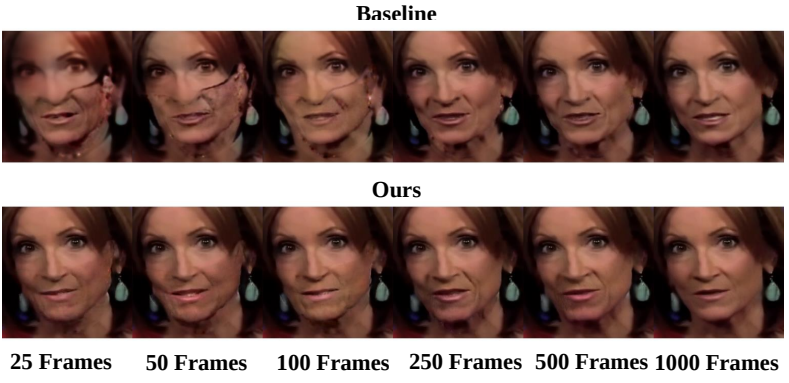


Figure 7: The effect training dataset size. The baseline model is trained from scratch and suffers when using limited data. Ours, by comparison, is far more robust.

## 5 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 and training the model in the foreground only. Another issue is that the monocular reconstruction stage is slow as it relies on optimisation through a differentiable renderer. Recent work has shown [5, 7, 8] 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 that could run in real time. Also, our work requires at least a few seconds of data to produce good results and performs poorly with only a single frame.

| N Frames | 100 | 250 | 500 | 1000 |
|---|---|---|---|---|
| Baseline | 17.71 | 12.33 | 11.84 | 11.78 |
| Ours | 11.37 | 11.88 | 11.74 | 11.21 |

Table 4: Our model is robust even on datasets as small as four seconds (100 frames). We compare the number of frames of a given actor to the FID obtained by the model trained on this dataset.

## 6 Conclusion

We have presented **DEAD**. Unlike person-specific models, which require several minutes of personalised 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, personalised 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 baseline model trained without priors.

# 7 Acknowledgments

This research was carried out as a collaboration between the University of Bath and DeepReel under the Innovate UK project "Feasibility study on improving video production process using data-efficient Generative-AI models." We would like to thank the DeepReel team for their assistance with this work. Jack Saunders is supported by the UKRI grant with code EP/S023437/1.

# References

[1] Eleven labs dubbing. https://elevenlabs.io/dubbing. Accessed: 2025-01-17.

[2] Flawless. https://braintitan.medium.com/the-worlds-first-movie-fully-translated-and-dubbed-by-ai-is-a Accessed: 2025-01-17.

[3] 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.

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] Tao Liu, Chenpeng Du, Shuai Fan, Feilong Chen, and Kai Yu. Diffdub: Person-generic visual dubbing using inpainting renderer with diffusion auto-encoder. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3630–3634. IEEE, 2024.

[17] 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.

[18] 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.

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

[20] 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.

[21] Linsen Song, Wayne Wu, Chen Qian, Ran He, and Chen Change Loy. Everybody's talkin': Let me talk as you want. *IEEE Transactions on Information Forensics and Security*, 17:585–598, 2022.

[22] Michal Stypulkowski, Konstantinos Vougioukas, Sen He, Maciej Zikeba, Stavros Petridis, and Maja Pantic. Diffused heads: Diffusion models beat gans on talking-face generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5091–5100, 2024.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

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

[28] 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, 2023.

[29] 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 2023.

[30] 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/.

[31] Jun Xiang, Xuan Gao, Yudong Guo, and Juyong Zhang. Flashavatar: High-fidelity head avatar with efficient gaussian embedding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[32] 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.

[33] Yue Zhang, Minhao Liu, Zhaokang Chen, Bin Wu, Yubin Zeng, Chao Zhan, Yingjie He, Junxin Huang, and Wenjiang Zhou. Musetalk: Real-time high quality lip synchorization with latent space inpainting. *arxiv*, 2024.

[34] 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.

[35] 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.