Dubbing for Extras: High-Quality Neural Rendering for Data Sparse Visual Dubbing

Jack Saunders, Vinay Namboodiri
University of Bath
Claverton Down, Bath BA2 7AY
{jrs68, vpn22}@bath.ac.uk

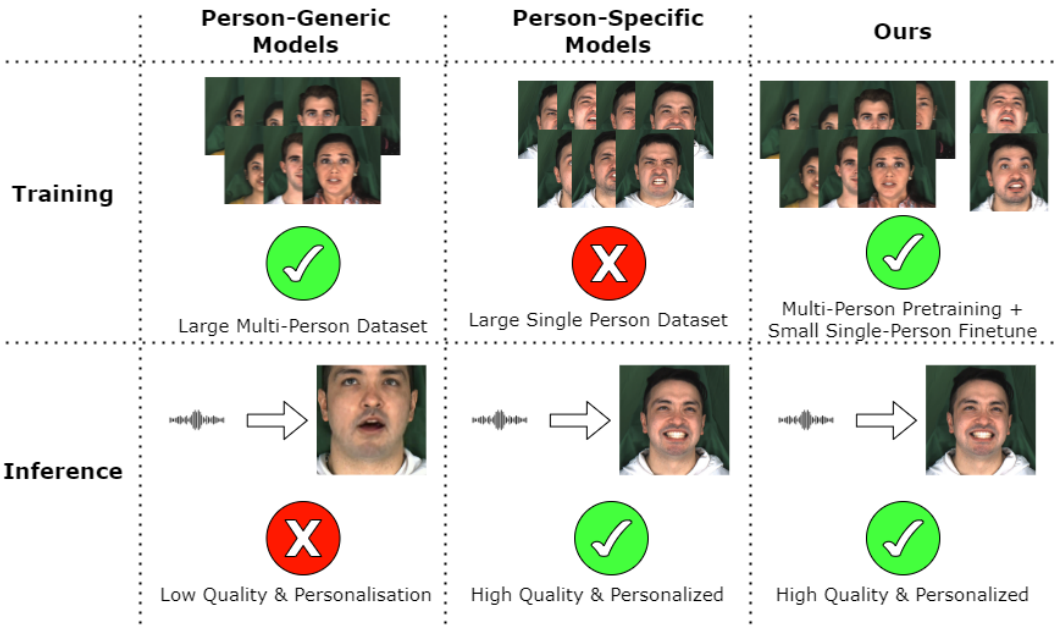

Figure 1: Our method, **Dubbing for Extras**, combines the best attributes of generalisable but low-quality person-generic models and high-quality but restrictive person-specific models. In doing so, we are able to create a model that achieves high visual quality and personalisation, that makes efficient use of all available data for a given actor.

Abstract

We present Dubbing for Extras, a novel approach to neural rendering for visual dubbing. At the core of our method is the separation of the deferred neural rendering pipeline into person-generic and person-specific components. Specifically, we train a person-generic image-to-image model and person-specific neural textures. When presented with a new identity, our model requires only the training of the neural texture, allowing for rapid training and high-quality models even with limited data. Our image-to-image network is modified in order to inject audio, improving mouth interior reconstruction. Initial, early-stage experiments show a significant reduction in the amount of data required, the amount of storage space and the training time.

1. Introduction

Visual dubbing is the process of altering the lip motion of a video to match dubbed audio. This allows for media content to cross linguistic and cultural boundaries and opens up otherwise inaccessible markets. While existing methods have been able to produce compelling results, as of yet no model has become viable as an alternative to conventional audio-only dubbing. We argue that any successful video dubbing method must be **high-quality**, **generalizable**, **scalable** and **recognizable**. It must be high-quality so that con-

Table 1: A comparison of person-generic methods (Generic), person-specific models (Specific) and Our method. We are able to get the visual quality and idiosyncrasies of person-specific models with far less data and storage.

	Generic	Specific	Ours
High Visual Quality	✗	✓	✓
Idiosyncrasies	✗	✓	✓
Data for new subject	1 frame	120s	3s
Storage for new subject	0B	100MB	1MB

sumers 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 the sense that all actors, from A-list stars to extras, should be dubbed effectively with as little as two seconds of dialogue. It must be scalable, a sufficiently large model may be high-quality and generalizable, but it will not be adopted if adding another actor requires retraining from scratch. Finally, 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.

Previous models are split between their focus on these criteria. Some produce incredible quality video for a single actor under controlled conditions [13, 12, 27, 19, 32, 24, 8], such methods are high-quality and recognizable but will work only on the actor they are trained on. Others produce low-quality, but generalizable video [18, 15, 21, 30, 20, 23]. These methods can be applied to any audio and any video, but the outputs are rarely of good visual quality and they do not capture the style of the actors, instead providing generic lips that synchronize well with the audio.

We propose Dubbing for Extras. Taking the best of both approaches, we create a model that meets our criteria, allowing the high-quality dubbing of all actors, including extras with short roles. We use networks that are trained across multiple actors and therefore can generalize, and also introduce actor-specific components that allow for our model to adapt to individuals and produce high-quality and recognizable dubbing. Our model is scalable, given a new identity, only the actor-specific components need training.

We find that the models with the highest visual quality build upon 3D Models. Specifically, Neural Textures [27, 13, 28] produce the best quality of all available methods. This serves as the basis of our method. However, we argue that training a deferred neural renderer for every actor is unnecessary, instead training a single rendering network, but having a unique neural texture per actor. Previous work [19] shows that introducing audio into the rendering process improves lip-synchronization, we build upon this in our approach, shifting the injection of audio from the neural texture to the image-to-image network. To summarise, the key novel contributions of this work are:

- An approach towards visual dubbing that aims to achieve high-quality and generalizable video while capturing person-specific details.
- A person-generic neural renderer that uses person-specific neural textures allowing for fast and accurate fine-tuning on limited data.
- A modification of an image-to-image network that allows for the injection of audio.

2. Related Works

2.1. Person Specific Models

By restricting models to be person specific, the ability to produce high-quality outputs is impressive. These models usually use intermediate 3D representations, introducing some prior information into the model and making the image formation process much easier. **3DMM:** The most popular approach to do this is to use a 3D Morphable Model (3DMM) for explicit 3D geometry [1, 3, 14]. Typically, such a model is fit to video data and neural rendering [28, 25] is used to invert the fitting process and achieve ultra-high-quality results. The intermediate 3DMM can then be used for visual dubbing, either through the use of a source actor to drive the lip motions [29, 13, 12], or else by learning a direct mapping from audio to model parameters [19, 27, 32, 24]. Our work builds upon this line of thinking, attaining the impressive visual quality possible with neural rendering. However, unlike previous work, we apply this neural rendering in a scalable way. We train a generic neural renderer, but person-specific neural textures, allowing our model to adapt to unseen people with minimal data. **Implicit Model based:** Recently, works have attempted to build upon the success of implicit neural models such the Neural Radiance Field [17]. These methods either attempt to augment a 3DMM-based approach [7, 33], allowing for fine-grained alterations of the low-quality geometry proxy, or else they control an implicit model directly using audio signal [5, 8].

2.2. Person Generic Models

Person generic models [11, 18, 15, 21, 30, 20, 23] have the advantage of working for any video and any audio. These models are able to produce lip synchronization by optimizing a generative model against a pre-trained lip-sync evaluation network. Early works [11, 18, 15, 21] follow a similar pattern, using reference frames and audio signals to fill in masked-out regions of target frames with GAN losses [6]. More recent work [20, 23] replaces the GAN as a generative model with diffusion-based approaches [22, 10]. The majority of these methods rely on a lip-sync network,

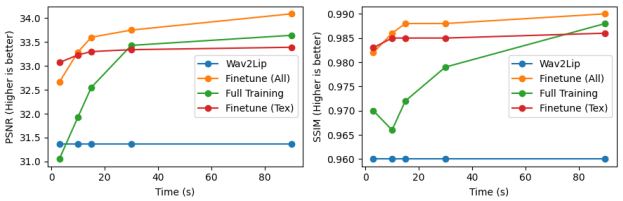

Figure 2: Results of our low data experiments. We report PSNR and SSIM scores vs the size of the dataset for several methods. Wav2Lip [18] is shown in blue, the naive method of training a deferred neural renderer from scratch is shown in green, the intensive process of fine-tuning the whole model is in orange and in red is the method of fine-tuning only the neural texture.

trained with contrastive learning, that can detect in or out-of-sync video and audio. This network is used as a loss to improve the sync in the generated videos.

The use of a reference frame in these methods allows the models to capture some amount of the target actor's idiosyncrasies. For this reason, we also use a reference frame in our method. However, these models do not fully capture the style of target actors as they rely on only a single frame of conditioning. In contrast, our method makes use of all available data for a given actor through fine-tuning, whether that is just a few seconds or multiple minutes. The consequence of this is that our model does capture person-specific speaker styles, while still being able to work in limited data scenarios.

3. Method

At the core of our method is a multi-stage training process that first creates a powerful person-generic model, and then allows for the fine-tuning of this large model to small person-specific datasets, using only small components. We first process our data in a way that is consistent across videos, and then perform monocular reconstruction to fit the FLAME model [14] to the videos (Section 3.1). Given the FLAME parameters, we then train a neural renderer (Section 3.2), inspired by deferred neural rendering [28], but differing in the fact that it is capable of working on many different identities. We are able to achieve this by splitting the renderer into components, some of which are person-generic and some person-specific. Finally, we describe a process (Section 3.3) that allows our person-generic renderer to adapt to new identities using minimal data and storage space.

3.1. Preprocessing

We warp and crop each frame so that it has a square shape (W, W). We do this by detecting 68 landmarks using mediapipe [16] and then finding for each video, the tightest possible bounding box that contains all the landmarks across all the frames. We then use EMOCA v2 [2, 4] to estimate the parameters of the FLAME [14] model independently for every frame. We note that the shape parameters for EMOCA vary slightly from frame to frame. This leads to noticeable jitter in the final video. We therefore replace the shape parameters for all frames, by setting this parameter to the mean of the per-frame estimations across all frames in a given video.

3.2. Person-Generic Deferred Neural Renderer

Given the FLAME parameters we have tracked using monocular reconstruction, we need a way of converting these back into photorealistic video. We use a deferred neural renderer based on neural texture [28] to do this. The neural texture approach combines a learnable, multi-channel UV texture with an image-to-image neural network. While this method has been shown to be very effective, it can only be applied in person-specific scenarios. This is because a new texture and rendering network must be trained for every subject.

We propose a generalized deferred neural renderer. We claim that only a single image-to-image network is necessary, as person and scene-specific detail can be represented in the neural textures. For N individuals, we use person-specific neural textures $\mathcal{T} = \{\mathbf{T}_i\}_{1 \leq i \leq N}$, combined with a person-generic deferred neural renderer $\mathcal{G}$. The deferred neural renderer has a UNET architecture, consisting of a CNN encoder and decoder with skip connections between.

To improve the generalization of the deferred neural renderer we also borrow from person generic models and we take a random frame from the same video to use as a reference frame. We concatenate this to the rasterised neural texture.

It has been shown [19] that including audio in the neural renderer improves the output. Previous works have conditioned the neural texture on audio. We do not consider this approach appropriate as we want to be able to learn textures for new identities with very limited data. We, therefore, shift the audio conditioning to the person-generic deferred neural renderer. To do this, we first encode the audio, represented as a MEL Spectrogram into a single vector of 32 dimensions. This representation is then repeated spatially and stacked with the input rasterization as in the input to the image-to-image network.

We use an ℓ_1 reconstruction loss $\mathcal{L}_1$, a VGG-based style loss $\mathcal{L}_{\mathbf{VGG}}$ and an adversarial loss $\mathcal{L}_{\mathbf{adv}}$ using the least-squares formulation. For the adversarial loss, we used the channel-stacked frames as in Wav2Lip [18], allowing the discriminator to see all T frames in the window. This helps improve temporal stability. We weigh each loss to get the final objective function:

$$\mathcal{L} = \lambda_1 \mathcal{L}_1 + \lambda_{\mathbf{VGG}} \mathcal{L}_{\mathbf{VGG}} + \lambda_{\mathbf{adv}} \mathcal{L}_{\mathbf{adv}} + \lambda_{\mathbf{sync}} \mathcal{L}_{\mathbf{sync}}$$

Another loss is applied at the rasterization level between the first three channels of the rendered neural texture and the target image. This encourages the first three channels to act like a classic diffuse texture.

3.3. Person-Specific Fine-Tuning

In previous work, given a new identity, a new model would need to be trained from scratch. In our method, however, this is not the case. As we use a person-generic image-to-image decoder, we need only train the person-specific neural texture. To do this, we randomly initialize a new neural texture and freeze the image-to-image decoder. We then train on the full objective function, optimizing the texture only. The pretrained nature of the image-to-image model allows the neural renderer to generalize to unseen expressions, even when there is limited data for the new individual. Furthermore, as the image-to-image model is the same for all subjects, only the neural texture need to be stored. This allows for an approximately 100 times reduction in the amount of storage space required.

4. Results

4.1. Experimental Setting

Dataset: We the MEAD dataset [31]. MEAD consists of 60 actors, captured across 6 camera angles. Each actor performs 30 sentences in 8 emotions and at three levels of intensity. We withhold three of these actors for our unseen identity tests and use only the front-facing cameras.

Our method does not claim any audio-to-expression generation as a novelty, so we test our methods using the tracked parameters unless otherwise specified.

4.2. Effect of Dataset Size

The most important claim of our work is that we are able to achieve the level of visual quality seen in person-specific models trained on several minutes of data, but using only a fraction of this. Recall that for our method, we require the fine-tuning of only the neural texture component of the model. To demonstrate this effect we compare this method to models that are trained from scratch (as is done in previous works [13, 28, 12, 27]), and models for which all layers are fine-tuned. To ensure a fair comparison and not bias the results by the choice of architecture, we use the same image-to-image network, including adding a reference frame and audio encoding. In addition, to show that our method surpasses the person-generic approach, we also compare our results to Wav2Lip [18].

We compare these methods using SSIM and PSNR. We run our tests using each of the three withheld MEAD subjects. The quantitative results can be seen in Figure 2. It can be clearly seen that the models trained from scratch, while working well with large datasets, struggle to perform on minimal data. Both our method of fine-tuning just the neural textures and the method of finetuning both textures and the image-to-image network perform well on smaller datasets. In particular, on very small (3-10 seconds) datasets, the difference between the two is minimal. Given that our method requires the training and storage of only approximately 1% of the number of parameters used in the image-to-image network, our method is useful where many identities are required.

5. Future Work

So far, we have demonstrated that our method works well in a limited context. In particular, we have shown that it surpasses person-generic models like Wav2Lip [18] in all attempted cases, and person-specific models [28] when the size of the training dataset is limited. In future work, we intend to repeat our experiments on a much larger dataset. We intend to train our model on a high-resolution, in-the-wild dataset, such as the 4K YouTube Faces dataset [9]. We hope to show that our model can then adapt to any new faces even in uncontrolled conditions. We will run the same experiment on this dataset, with comparisons to more models and the addition of user studies.

We also have some components that require an ablation study. In particular, the inclusion of a reference frame has not been studied in the context of 3DMM / Neural Rendering approaches. We intend to study this. Additionally, while it is known that audio improves video generation in the image-to-image network [19], we have not verified that this is still true with our architecture modifications. We will also compare to more state-of-the-art methods, particularly person-generic models.

We may also involve the use of an audio-to-expression network. While it is an important component of any 3DMM-based visual dubbing system, we had so far excluded it to focus on neural rendering. The addition of such a model should help show the robustness and actually applicability of our method. In particular, we will likely modify the Imitator [26] audio-to-expression network for our work.

6. Conclusion

We have presented **Dubbing for Extras** a novel approach for visual dubbing for actors with limited data. We have performed initial experiments to show the promise of our work. We hope that, with further development, our model will see adoption in both the entertainment industry and the creator economy.

References

[1] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '99, page 187–194, USA, 1999. ACM Press/Addison-Wesley Publishing Co. 2

[2] 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. 3

[3] 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. 2

[4] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. *ACM Transactions on Graphics (ToG), Proc. SIGGRAPH*, 40(4):88:1–88:13, August 2021. 3

[5] Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Niessner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8649–8658, June 2021. 2

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[7] 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. 2

[8] 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. 2

[9] 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. 4

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020. 2

[11] 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. 2

[12] 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. 2, 4

[13] 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. 2, 4

[14] 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. 2, 3

[15] Borong Liang, Yan Pan, Zhizhi Guo, Hang Zhou, Zhibin Hong, Xiaoguang Han, Junyu Han, Jingtuo Liu, Errui Ding, and Jingdong Wang. Expressive talking head generation with granular audio-visual control. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3377–3386, 2022. 2

[16] 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. 3

[17] 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. 2

[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. 2, 3, 4

[19] Jack Saunders and Vinay Namboodiri. Read avatars: Realistic emotion-controllable audio driven avatars, 2023. 2, 3, 4

[20] Shuai Shen, Wenliang Zhao, Zibin Meng, Wanhua Li, Zheng Zhu, Jie Zhou, and Jiwen Lu. Difftalk: Crafting diffusion models for generalized talking head synthesis. In *arxiv*, 2023. 2

[21] Hyoung-Kyu Song, Sang Hoon Woo, Junhyeok Lee, Seungmin Yang, Hyunjae Cho, Youseong Lee, Dongho Choi, and Kang-Wook Kim. Talking face generation with multilingual tts. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21393–21398, 2022. 2

[22] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv:2010.02502*, October 2020. 2

[23] Michał Stypułkowski, Konstantinos Vougioukas, Sen He, Maciej Zieba, Stavros Petridis, and Maja Pantic. Diffused Heads: Diffusion Models Beat GANs on Talking-Face Generation. In *https://arxiv.org/abs/2301.03396*, 2023. 2

[24] Anni Tang, Tianyu He, Xu Tan, Jun Ling, Runnan Li, Sheng Zhao, Li Song, and Jiang Bian. Memories are one-to-many mapping alleviators in talking face generation, 2022. 2

[25] A. Tewari, J. Thies, B. Mildenhall, P. Srinivasan, E. Tretschk, W. Yifan, C. Lassner, V. Sitzmann, R. Martin-Brualla, S. Lombardi, T. Simon, C. Theobalt, M. Nießner, J. T. Barron, G. Wetzstein, M. Zollhöfer, and V. Golyanik. Advances in neural rendering. *Computer Graphics Forum*, 41(2):703–735. 2

[26] Balamurugan Thambiraja, Ikhsanul Habibie, Sadegh Aliakbarian, Darren Cosker, Christian Theobalt, and Justus Thies. Imitator: Personalized speech-driven 3d facial animation, 2022. 4

[27] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural voice puppetry: Audio-driven facial reenactment. *ECCV 2020*, 2020. 2, 4

[28] 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. 2, 3, 4

[29] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. *Commun. ACM*, 62(1):96–104, dec 2018. 2

[30] 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, 2023. 2

[31] 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. 4

[32] 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, 2020. 2

[33] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Instant volumetric head avatars. 2023. 2