Perceptual Conversational Head Generation with Regularized Driver and Enhanced Renderer

Ailin Huang*
Megvii Research
Wuhan University
huangailin@megvii.com

Zhewei Huang*
Megvii Research
huangzhewei@megvii.com

Shuchang Zhou
Megvii Research
zsc@megvii.com

ABSTRACT

This paper reports our solution for ACM Multimedia ViCo 2022 Conversational Head Generation Challenge, which aims to generate vivid face-to-face conversation videos based on audio and reference images. Our solution focuses on training a generalized audio-to-head driver using regularization and assembling a high-visual quality renderer. We carefully tweak the audio-to-behavior model and post-process the generated video using our foreground-background fusion module. We get first place in the listening head generation track and second place in the talking head generation track on the official leaderboard. Our code is available at https://github.com/megvii-research/MM2022-ViCoPerceptualHeadGeneration.

CCS CONCEPTS

• **Computing methodologies** → **Reconstruction**.

KEYWORDS

Conversational Head Generation

ACM Reference Format:
Ailin Huang, Zhewei Huang, and Shuchang Zhou. 2022. Perceptual Conversational Head Generation with Regularized Driver and Enhanced Renderer. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3503161.3551577

1 INTRODUCTION

In face-to-face communication, people can observe real-time expressions and demeanor and more accurately capture their counterpart's feelings. It is interesting to understand this communication behavior and generate vivid talking head videos using computer vision technology. Proper responsive listening behavior is essential as well to effective communication, and also of critical importance to make digital humans more realistic during face-to-face human-computer interaction and animation production.

Generating conversational head videos is challenging because it involves not only the processing of serialized speech signals but also video synthesis [3]. We need to use the voice signal to infer the changes in the expressions and lip shapes of the people in the conversation. Then we also need to synthesize high-quality generated video frames based on the reference images of speakers and listeners. The prior art [24] has shown that realistic digital humans can be generated from a large number of videos of the same speaker. However, further research is needed to build a digital human system for any speaker with less available digital information.

In our paper, we focus on training an audio-to-head driver with limited data and assembling a powerful renderer to generate vivid videos. Our main techniques include:

- We apply several neural network training techniques to improve the performance of audio-to-head driver training on limited data, and further explore ensemble learning to make the model more robust.
- We assemble an enhanced renderer for producing visually better and more stable generated videos.

2 RELATED WORK

Audio-Driven Head Synthesis

With a reference head image and audio streams, many methods explore synthesizing a video based on a reference image, considering the head motion, expression, and lip shape changes. Some straightforward methods directly model the relationship between audio streams and reference images [14, 31]. To better correlate the video and audio factors, many recent works use a two-stage strategy to map reference images and audios firstly to feature representations (*e.g.* landmarks [4], 3D morphable models (3DMM) [27, 30], face parsing), and then render the final videos.

Recently, some very popular and advanced techniques have been applied to this task. StyleRig [26] describes a method to control StyleGAN [15] via a 3DMM. StyleGAN-based methods are very visually attractive, but their inference efficiency and identity retention performance still need improvement. The neural radiance field AD-NeRF [8] can show very promising results for a scenario where the data of the target person is relatively sufficient.

Neural Model and Regularization

The head generation task requires a special collection of processed data. Training effective models with limited data is an important topic of the ViCo challenge [1]. Recently, deep neural network with residual learning [9] has shown good performance in various tasks [7, 12, 18, 25]. Meanwhile, many regularization techniques have been proposed to increase the generalization of neural networks [28]. Among these techniques, Dropout [23] and Batch Normalization (BN) [13] are the most popular and powerful.

*Both authors contributed equally to this work.

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

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3551577

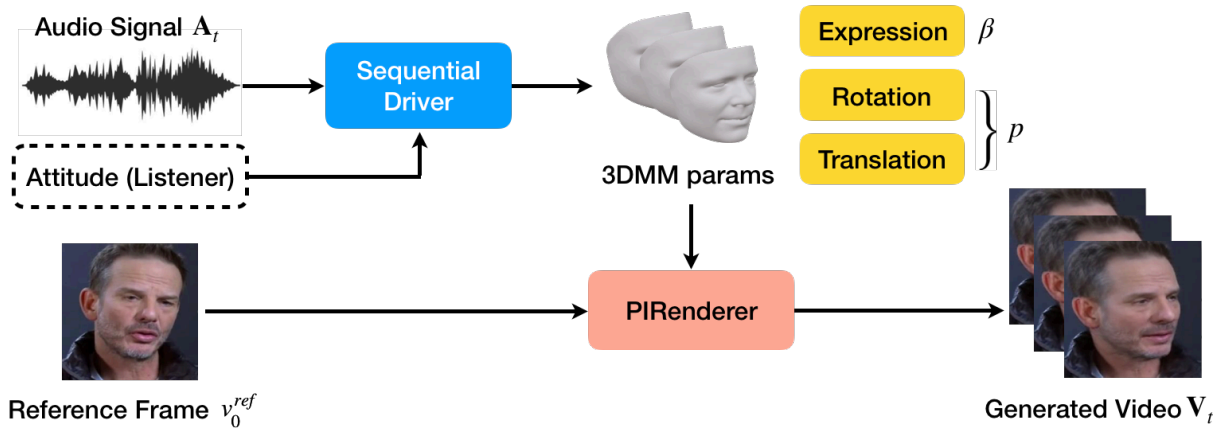

Figure 1: Overview of video-driven head generation pipeline. Given an audio signal sequence, a sequential driver approximates the 3DMM parameters for every video frame. Then a pre-trained PIRenderer [22] renders the final video based on these parameters and reference frame

Besides, ensemble learning [6] is a well-explored technique to improve the accuracy and generalization of the model. So it is usually used in various task, such as image classification [29], reinforcement learning [16] and combination optimization [2].

3 TASK OVERVIEW

3.1 Definition

Our work uses a unified framework to learn both following tasks. We briefly introduce the definitions of these two tasks. The main results and experiments of this paper are presented on the talking head generation task.

Vivid Talking Head Video Generation

Given the an input audio signal sequence $\mathrm{A}_t = a_1, ..., a_t$ of the speaker in time stamps ranging from $\{1, ..., t\}$ and a reference image v_0^{ref}, our goal is to generate a talking head video $\mathrm{V}_t^{Talking} = \{v_0, v_1, ..., v_t\}$.

Responsive Listening Head Video Generation

In addition to the input of the Talking Head Video Generation, the listening head generation additionally receives the input of the listener's attitude. Our goal is to generate a listening head video $\mathrm{V}_t^{Listening} = \{v_0, v_1, ..., v_t\}$.

3.2 Dataset

ViCo dataset [32] contains 483 video clips of 76 listeners responding to 67 speakers. The total length of these clips is approximately 95 minutes. Following previous work [5, 22], we extract 3DMM parameters for each frame. As ViCo baseline [32], relative dynamic and identity-independent features face can be represented parametrically using $\{\beta \in \mathbb{R}^{64}, p \in \mathbb{R}^6\}$ which denotes the expression and pose. Here, p represents rotations with $SO(3) \in \mathbb{R}^3$ and translations in $\mathbb{R}^3$. Additionally, for better modeling head movements, the baseline uses a "crop" parameter c of $\mathbb{R}^3$. This annotates where we will place and size the parametric 3D face in the original image.

Table 1: The official final leaderboard in ViCo challenge [1]. We intercept the results of the top few teams

Team (Talking)	PSNR↑	FID↓	LMD↓
sysu_hcp	**17.767**	29.709	**10.101**
Ours	17.179	**24.678**	10.646
iLearn	16.546	25.050	10.900
Avatar	17.696	34.571	11.643
Digital_Human	17.331	30.944	12.837
THU-Talking	16.390	45.361	12.707
Team (Listening)	PSNR↑	FID↓	ExpFD↓
Ours	**18.512**	**21.350**	**0.116**
iLearn	18.491	26.675	0.133
cheese	16.202	42.019	0.137
en_train	16.780	80.538	0.161
LIMMC	16.265	86.983	0.167

3.3 Evaluation

We consider evaluating our models in terms of image quality and semantics. Because the generated images and the real video are unlikely to be pixel-by-pixel aligned, the traditional metric (PSNR, SSIM) may not be reasonable. We mainly consider a metric at the feature level, Fréchet Inception Distance (FID) [10]. We further analyze the landmark distance (LMD) and expression feature distance (ExpFD) between generated faces and ground truth. The whole system can render 4 frames per second at 256×256 resolution. Comparing our model with methods from other teams shown in Table 1, our model mainly gains an advantage on the FID index.

4 METHOD

4.1 Framework

We illustrate the overall pipeline in Figure 1. We use a two-stage method to generate the head video with 3DMM parameters as the

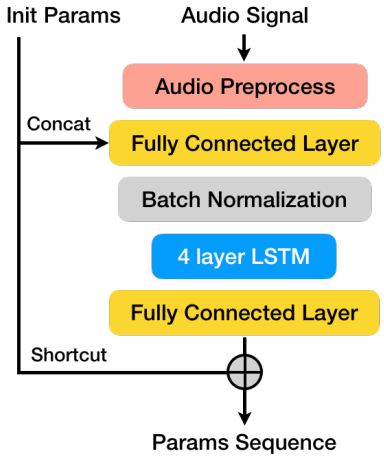

Figure 2: Architecture of the sequential driver model. For each frame, we approximate the residual relative to the initial parameters of the first frame (reference image)

mediator interface. We firstly approximate the parameters of every frame using a sequential driver model. We then use a pre-trained PIRenderer [22] to render the final video based on these parameters and the reference image. Furthermore, we enhance PIRenderer use an image boundary inpainting trick and a foreground-background fusion module.

In the ViCo challenge [1], participants are restricted to training their methods on ViCo training set. We observe large performance differences between the models on the training and validation sets. To overcome this over-fitting issue, we introduce several model regularization techniques, including residual learning [9], Dropout [23], and training using BN [13] with big batch size. The architecture of sequential driver containing a four-layer LSTM [11] model is illustrated in Figure 2.

4.2 Learning

Loss Function

For model optimization, we supervise each prediction using 3DMM parameters extracted from training videos. We randomly sample clips with $T = 90$ frames and calculate the loss function as:

$$\mathrm{L}_{gen} = \sum_{t=1}^{T} \|\beta_t - \widehat{\beta}_t\|_2 + \|c_t - \widehat{c}_t\|_2 + \|p_t - \widehat{p}_t\|_1, \tag{1}$$

where β, p, c denote the ground truth and $\widehat{\beta}, \widehat{p}, \widehat{c}$ represent the result of driver model. We experimentally search the choice of L_1 and L_2 loss functions. A head motion constraint loss L_{mot} is applied to encourage the inter-frame continuity:

$$\mathrm{L}_{mot} = \sum_{t=1}^{T} \|\mu(c_t) - \mu(\widehat{c}_t)\|_2, \tag{2}$$

where $\mu(\cdot)$ measures the inter-frame changes *i.e.* $\mu(c_t) = c_t - c_{t-1}$. The final loss function can be formulated as:

$$\mathrm{L}_{total} = \mathrm{L}_{gen} + \mathrm{L}_{mot}. \tag{3}$$

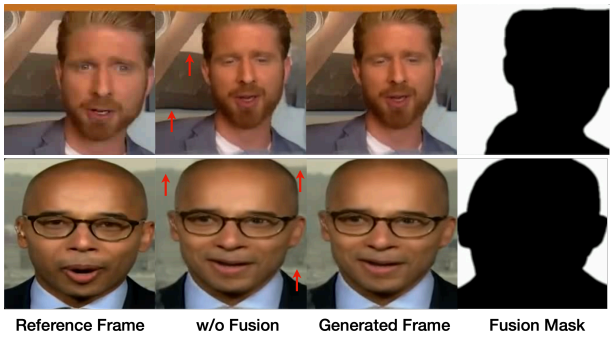

Figure 3: The effect of foreground-background fusion. The movement of the background is easily noticeable during video playback

Details

For audio preprocessing, we follow the ViCo baseline [32]. We extract the 14-dim Mel-frequency cepstral coefficients (MFCC) feature with the corresponding MFCC Delta and Delta-Delta (second-order difference) feature. The energy, loudness and zero-crossing rate (ZCR) are also embedded into audio features S_i for each audio frame clip A_i. The 45-dimensional audio feature extracted from A_s^t is denoted as $S_s^t = s_1, ..., s_t$.

Our driver model is trained on four TITAN 2080Ti GPUs for about three hours. We use AdamW [17, 19] optimizer with a weight decay of 0.05. Our training uses a batch size of 128. We gradually reduce the learning rate from 5×10^{-3} to 10^{-4} using cosine annealing during the whole training process.

4.3 Enhanced Renderer

We use a controllable portrait image generation model, PIRenderer [22], to convert 3DMM face parameters to video based on the reference image. PIRenderer employs a subset of 3DMM parameters as the head motion descriptor. Overall, PIRenderer works very well. However, two issues may degrade the image quality, including background distortion and image border artifacts.

Foreground-Background Fusion

Since we cannot estimate camera motion in the audio-driven head generation, we assume that the background is completely static. PIrenderer still needs background textures to complement some backgrounds revealed during head movement, which may not exist in the reference image. Due to the out-of-focus and complex texture of the background, the optical flow inferred by PIRenderer is not only on the surface of the head as wished. Therefore, we often observe unreasonable background distortions in synthetic videos, as shown in Figure 3.

To address this, We use a pre-trained foreground-background segmentation model, U^{2}Net [21], to segment the static background area. For the generated image $\widehat{I}_t^{gen}$ and reference image I^{ref}, we detect their background regions:

$$M_t, M^{ref} = seg(\widehat{I}_t^{gen}), seg(I^{ref}), \tag{4}$$

where $seg(\cdot)$ denotes the background region segmented from the image. To enhance inter-frame consistency, we calculate a median

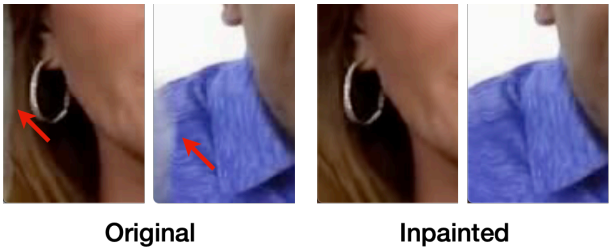

Figure 4: The effect of image boundary inpainting. When the features close to the boundary are warped, filling these holes with the boundary values of the original feature map will eliminate some artifacts

segmentation result, each pixel of which is the median of the results of the previous five frames. Formally,

$$M_t^{med}(x,y) = median\{M_t(x,y), M_{t-1}(x,y), ..., M_{t-4}(x,y)\}, \quad (5)$$

where we calculate the recent median for each pixel-wise location.

Then we paste the common background area from reference image I_{ref} to generated image $\widehat{I}_{gen}$:

$$M_t^{fusion} = Gaussian(M_t^{med} \cap M^{ref}), \quad (6)$$

$$\widehat{I}_t^{fusion} = (1 - M_t^{fusion}) \odot \widehat{I}_t^{gen} + M_t^{fusion} \odot I^{ref}, \quad (7)$$

where we use a 7×7 Gaussian filter to smooth the seam of the stitched image.

Image Boundary Inpainting

For a good visual effect, the human hair and upper body should move with the movement of the head. In this case, we need to inpaint some texture around the edges of the image. We can easily perform this technique by setting the padding mode to "border" in *grid_sample* function of PyTorch [20] when warping image features. The result is shown in Figure 4.

Combining these two modules, we observe about 0.4dB gain and 1.2 drop of FID on our validation set without any extra training overhead.

5 EXPERIMENTS

5.1 Ablation Studies

We do ablation studies on some designs, including regularization and model ensemble. The results are shown in Table 2.

Ablation on Model Design

Ablation experiments on model design are reported in Table 2. Our model learns the residual added to the initial 3DMM parameters. Removing this design will greatly reduce the performance of the model (1.46dB). We also observe that BN and Dropout can help alleviate the over-fitting issue.

We further increase the batch size for training, specifically 128. Our experiments show that large batch sizes have overall competitive performance while reducing training time exponentially.

Table 2: The ablation studies on model design and training batch size

Setting	PSNR↑	FID↓	LMD↓
w/o BN [13]	16.16	17.67	50.00
w/o residual learning [9]	14.79	19.20	65.18
w/o Dropout [23]	**16.38**	17.12	44.34
Final Model	16.25	**16.73**	**43.02**
batch size 8	16.52	18.13	43.07
batch size 32	**16.70**	17.18	**42.06**
batch size 128	16.25	**16.73**	43.02

Table 3: The ablation study on ensemble learning

Setting	PSNR↑	FID↓	LMD↓
w/o ensemble	16.25	16.73	43.02
self ensemble (3×)	16.34	16.31	42.65
cross-model ensemble (3×)	**16.56**	**15.81**	**41.73**

5.2 Model Ensemble

We observe whether ensemble trained models in the same training process (self ensemble) or models in different training processes (cross-model ensemble), all performance metrics can be improved. The experiment results are shown in Table 3. In the ViCo competition, we finally validate 10 models on our validation set and cross-ensemble the best 5 models. The most time spent in our system is in the PIRenderer part, so the ensemble on the driver model only slightly increases the inference overhead.

6 FUTURE WORK

We do not have enough time to fully explore the following relevant techniques in this short challenge, which might be very useful for this task. 1) Finetuning the renderer for current tasks and metrics is a reasonable point of improvement. The existing renderers produce results with inaccurate character identity retention and background disturbance. We could get a more targeted model to restrict the renderers to a specific application scenario, including rendering the same character and static background. 2) The lip shape generated by our talking head model is relatively conservative and insufficient to distinguish different syllables. The mapping of syllables to lips is a long-studied topic, and some experience from traditional methods may help to improve our model further. 3) There is also much room for exploration in more feature engineering techniques and stronger models.

7 CONCLUSION

In this paper, we introduce our solution for the conversational head generation competition. We propose a regularized driver and enhanced renderer to synthesize perceptually impressive videos. Hopefully, our discoveries and engineering practices can help future researchers.

REFERENCES

[1] JD Explore Academy. 2021. Conversational Head Generation Challenge. https://vico-challenge.github.io.

[2] Zixuan Cao, Yang Xu, Zhewei Huang, and Shuchang Zhou. 2022. ML4CO-KIDA: Knowledge Inheritance in Dataset Aggregation. *arXiv preprint arXiv:2201.10328* (2022).

[3] Lele Chen, Guofeng Cui, Celong Liu, Zhong Li, Ziyi Kou, Yi Xu, and Chenliang Xu. 2020. Talking-head generation with rhythmic head motion. In *European Conference on Computer Vision*. Springer, 35–51.

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

[5] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. 2019. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 0–0.

[6] Thomas G Dietterich. 2000. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*. Springer, 1–15.

[7] Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. 2021. Repvgg: Making vgg-style convnets great again. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13733–13742.

[8] Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. 2021. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5784–5794.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[10] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 30 (2017).

[11] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[12] Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. 2020. Rife: Real-time intermediate flow estimation for video frame interpolation. *arXiv preprint arXiv:2011.06294* (2020).

[13] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*. PMLR, 448–456.

[14] Amir Jamaludin, Joon Son Chung, and Andrew Zisserman. 2019. You said that?: Synthesising talking faces from audio. *International Journal of Computer Vision* 127, 11 (2019), 1767–1779.

[15] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4401–4410.

[16] Łukasz Kidziński, Sharada Prasanna Mohanty, Carmichael F Ong, Zhewei Huang, Shuchang Zhou, Anton Pechenko, Adam Stelmaszczyk, Piotr Jarosik, Mikhail Pavlov, Sergey Kolesnikov, et al. 2018. Learning to run challenge solutions: Adapting reinforcement learning methods for neuromusculoskeletal environments. In *The NIPS'17 Competition: Building Intelligent Systems*. Springer, 121–153.

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

[18] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. 2017. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 136–144.

[19] Ilya Loshchilov and Frank Hutter. 2018. Fixing weight decay regularization in adam. (2018).

[20] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. (2017).

[21] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar Zaiane, and Martin Jagersand. 2020. U2-Net: Going Deeper with Nested U-Structure for Salient Object Detection. *Pattern Recognition* 106, 107404.

[22] Yurui Ren, Ge Li, Yuanqi Chen, Thomas H Li, and Shan Liu. 2021. PIRenderer: Controllable Portrait Image Generation via Semantic Neural Rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13759–13768.

[23] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* 15, 1 (2014), 1929–1958.

[24] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. 2017. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (ToG)* 36, 4 (2017), 1–13.

[25] Zachary Teed and Jia Deng. 2020. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*. Springer, 402–419.

[26] Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. 2020. Stylerig: Rigging stylegan for 3d control over portrait images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6142–6151.

[27] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. 2020. Neural voice puppetry: Audio-driven facial reenactment. In *European conference on computer vision*. Springer, 716–731.

[28] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*. 1096–1103.

[29] Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. *arXiv preprint arXiv:2203.05482* (2022).

[30] Ran Yi, Zipeng Ye, Juyong Zhang, Hujun Bao, and Yong-Jin Liu. 2020. Audio-driven talking face video generation with learning-based personalized head pose. *arXiv preprint arXiv:2002.10137* (2020).

[31] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. 2019. Talking face generation by adversarially disentangled audio-visual representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 9299–9306.

[32] Mohan Zhou, Yalong Bai, Wei Zhang, Tiejun Zhao, and Tao Mei. 2021. Responsive Listening Head Generation: A Benchmark Dataset and Baseline. *arXiv preprint arXiv:2112.13548* (2021).