Supplementary Materials: Audio-Driven Identity Manipulation for Face Inpainting

Anonymous Authors

1 DATASET PROPROCESSING

Our method aims to generate faces of exceptional fidelity using audio references. While previous datasets such as VoxCeleb [4] and VoxCeleb2 [1] have successfully gathered a large number of videos, the majority of the collected faces suffer from issues like blurriness, low resolution (128×128), and noise, which ultimately result in subpar quality outcomes. Consequently, our foremost challenge lies in curating a dataset of audio face pairs that exhibits superior quality and fidelity.

Celeb-ID, introduced by Dolhansky et al. [3], serves as a benchmark dataset for eye inpainting tasks. It comprises approximately 17,000 identities and more than 100,000 images, boasting a significantly higher resolution of 300×300. Although Celeb-ID lacks audio recordings, we fortuitously discover that there is a partial overlap in identities between Celeb-ID and VoxCeleb. Leveraging this intersection, we assemble a top-tier dataset named VoxCeleb-ID, which combines audio from VoxCeleb with face images from Celeb-ID, resulting in a collection of high-quality audio face pairs. VoxCeleb-ID encompasses 953 unique identities, 7,080 face images, and an impressive 117,551 audio recordings.

In addition to VoxCeleb-ID, we also process two exceptional talking video datasets: FaceForensics++ [5] and HDTF [8]. These datasets are primarily employed for tasks such as fake face detection and talking face generation. We download a total of 761 and 358 videos, respectively, from the provided URLs, ensuring that they maintain a resolution of either 720p or 1080p. To prepare the data, we utilize the FaceXlib open-source tool to detect, align, and crop the faces in each video frame. All faces are subsequently resized to a standardized resolution of 256×256.

For audio recordings in the three datasets, we set the sampling rate to 16 kHz, the channel number to one, and cut them into 6-second audio segments. If the audio is not long enough, we repeat it to ensure a minimum duration of 6 seconds. Following the established methodology outlined in [7], we presently remove silence regions of each segment using a voice activity detector and extract 64-dimensional log mel-spectrograms using a Hann window of 255mm, 100ms hop, and 1024 FFT frequency bands. Additionally, we perform mean and variance normalization of each mel-frequency bin. Given that the number of audio segments significantly exceeds the number of corresponding face images, we randomly sample at most 20 audio segments for each identity. Conversely, in the cases of FaceForensics++ and HDTF, where facial appearances remain largely consistent across different frames, we randomly select 10 faces for each video to maintain diversity and prevent redundancy.

Moreover, in the implementation of our method, we disregard the audio content and prioritize the visual aspect. Hence, we presently overlook the changes in lip movements in the FaceForensics++ and HDTF datasets. Instead, we manually choose standard faces that exhibit frontal views and are free from noticeable lip changes. For each selected standard face, we compute a VGG-Face feature using a ResNet-50 model pre-trained on the VGGFace2 dataset. These VGG-Face features serve as the face identity labels for our dataset.

2 TRAINING DETAILS

We commence the training process by training the audio face decoder on the FaceForensics++ and HDTF datasets. Throughout the training, we randomly select a face and an audio segment from the same identity. The audio segment is then passed through a pre-trained speaker recognition network to extract an audio embedding. It is important to note that the speaker recognition network remains fixed during the training process. To enhance robustness, we normalize the embedding and introduce Gaussian noise as permutations. Subsequently, we feed the modified audio embedding into the audio face decoder, which reconstructs the corresponding faces. The maximum number of training iterations is set to 100k.

Moving forward, we train our complete method using the standardized faces from all three datasets. During training, we randomly resize and flip the faces to augment the training data. The batch size is set to 4, and the learning rate is set to 2e-4. We train the model for a maximum of 600,000 iterations. To calculate the loss within the mask region, we multiply the generated result with the corresponding mask. Furthermore, we introduce a discriminator to compute the GAN loss, with the learning rate set to 2e-4.

3 RESULTS

We present additional visual results in Figure 1 and report the CosFace [6] distance below the corresponding images. Our audio faces successfully learn fundamental identity attributes such as gender, age, and facial shape from the audio inputs, which effectively guide our method to generate high-quality faces while preserving the identity. In comparison to other methods, our approach achieves the smallest CosFace score, indicating superior performance in identity preservation.

Furthermore, we showcase retrieval results in Figure 2. The top 5 results retrieved by our outputs from the image gallery consistently correspond to the same identity as the ground truth, thus demonstrating the effectiveness of our method in preserving identity accurately.

4 ABLATION STUDY

In Table 1, we present the results of our ablation studies conducted on the three datasets. The configuration labeled as *Base + AudioEmb* denotes the inclusion of an audio embedding network, where the high-dimensional audio embedding is fused with the intermediate face feature in the face branch. However, we observe a decrease in performance with this configuration. We attribute this decline to the limited capacity of a single face decoder to effectively decode such a complex feature. The introduction of the audio face decoder leads to a slight improvement in performance, particularly

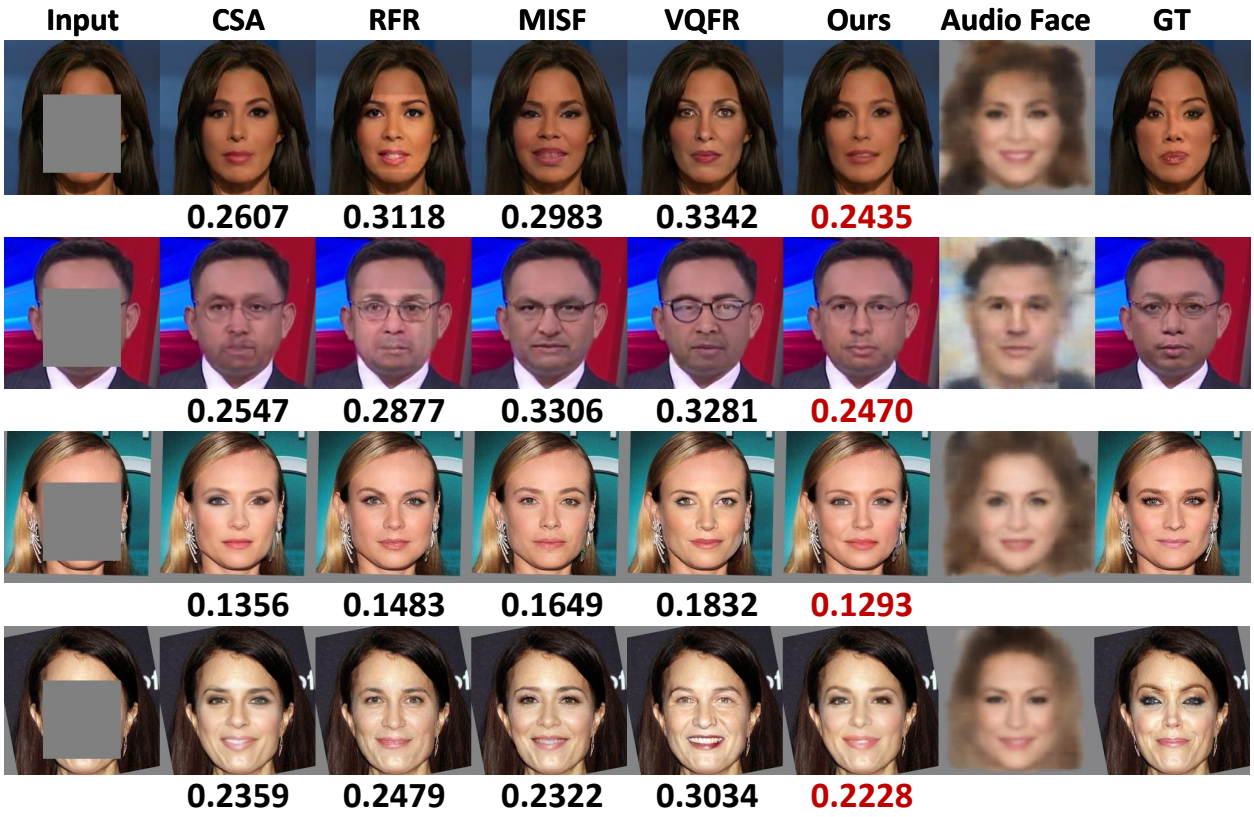

Figure 1: Visual results comparsion with the state-of-the-arts. We show the best result in red.

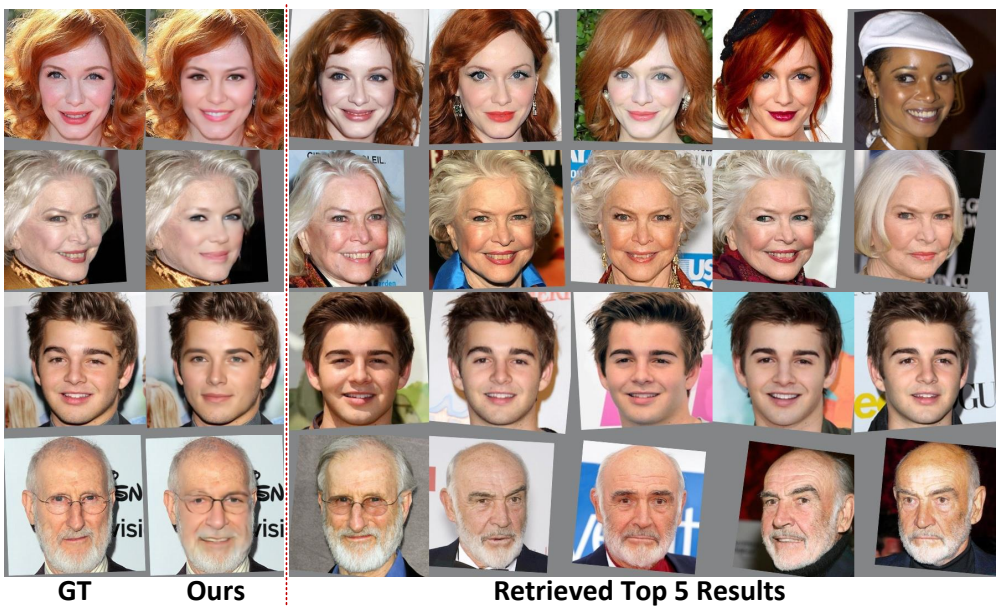

Figure 2: Top 5 image retrieval results.

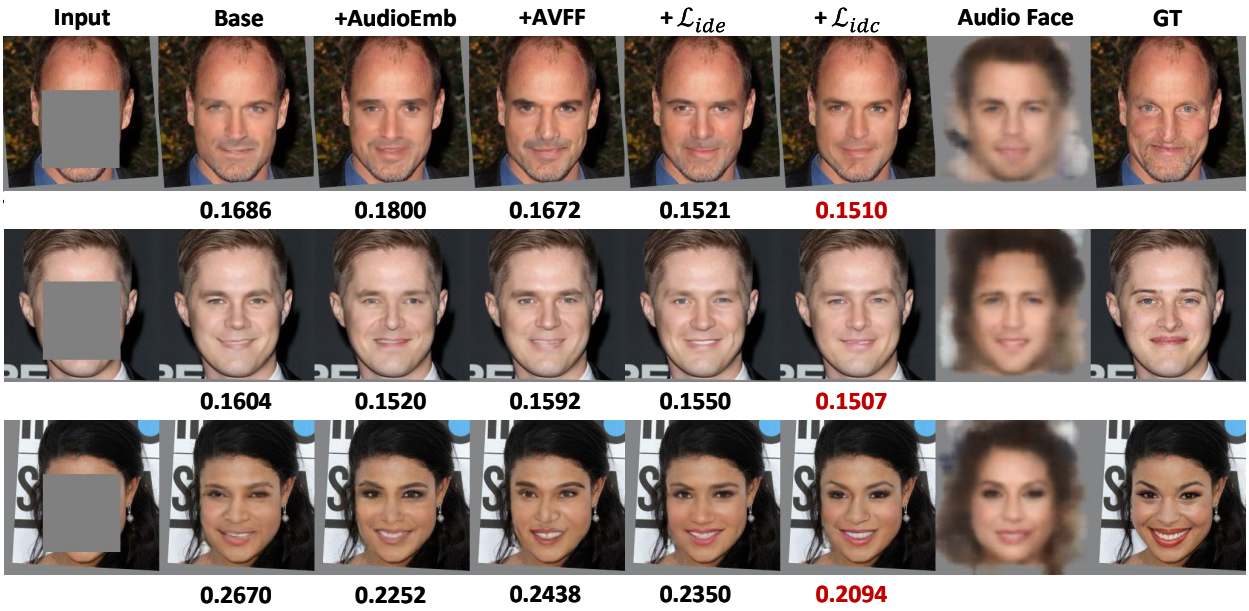

Figure 3: Visual results of ablation study. CosFace[6] distance are blow the image.

Methods	FaceForensics++				
	↑ PSNR	↑ SSIM	↓ LPIPS	↓ CosFace	↓ ArcFace
Base	28.1665	0.9154	0.0495	0.2329	3.8796
Base + AudioEmb	28.1596	0.9155	0.0497	0.2332	3.9130
Base + AudioDec + Concat	28.1401	0.9150	0.0492	0.2350	3.9080
Base + AudioDec + AVFF	28.1833	0.9164	0.0466	0.2299	3.8547
Base + AudioDec + AVFF + $\mathcal{L}_{ide}$	28.1886	0.9151	0.0492	0.2311	3.8441
Base + AudioDec + AVFF + $\mathcal{L}_{ide}$ + $\mathcal{L}_{idc}$	**28.2354**	**0.9166**	**0.0463**	**0.2278**	**3.7988**
Methods	HDTF				
	↑ PSNR	↑ SSIM	↓ LPIPS	↓ CosFace	↓ ArcFace
Base	28.4092	0.9083	0.0564	0.2289	3.8604
Base + AudioEmb	28.3743	0.9078	0.0552	0.2270	3.8627
Base + AudioDec + Concat	28.4239	0.9085	0.0549	0.2268	3.8309
Base + AudioDec + AVFF	28.3729	0.9086	0.0526	0.2266	3.8199
Base + AudioDec + AVFF + $\mathcal{L}_{ide}$	28.6179	0.9119	0.0543	0.2245	3.7319
Base + AudioDec + AVFF + $\mathcal{L}_{ide}$ + $\mathcal{L}_{idc}$	**28.7509**	**0.9156**	**0.0514**	**0.2184**	**3.6453**
Methods	VoxCeleb-ID				
	↑ PSNR	↑ SSIM	↓ LPIPS	↓ CosFace	↓ ArcFace
Base	25.9169	0.8942	0.0592	0.2176	3.6722
Base + AudioEmb	25.8960	0.8947	0.0574	0.2164	3.6477
Base + AudioDec + Concat	25.8909	0.8944	0.0577	0.2167	3.6461
Base + AudioDec + AVFF	25.8292	0.8938	0.0548	0.2156	3.6446
Base + AudioDec + AVFF + $\mathcal{L}_{ide}$	25.9368	0.8956	0.0568	0.2137	3.6291
Base + AudioDec + AVFF + $\mathcal{L}_{ide}$ + $\mathcal{L}_{idc}$	**25.9633**	**0.8963**	**0.0544**	**0.2122**	**3.6084**

Table 1: Quantitative results of ablation study on three datasets.

in the HDTF dataset. Nevertheless, we identify the misalignment between the features from the face decoder and the audio face decoder as a bottleneck hindering further performance improvement.

After applying the Audio-Visual Feature Fusion (AVFF) module, we observe a significant decrease in the face identity loss, as measured by the CosFace [6] and ArcFace [2] metrics. However, the intermediate features of the audio decoder encompass not only identity information but also noises, resulting in a decline in PSNR and SSIM scores. To address this issue, we introduce an identity embedding loss and an identity consistency loss. The identity embedding loss constrains the completed identity embedding to align with the ground truth, while the identity consistency loss ensures consistency between the final results and the audio faces. To provide further insights into the performance, we present additional visual results in Figure 3.

REFERENCES

[1] J. S. Chung, A. Nagrani, and A. Zisserman. 2018. VoxCeleb2: Deep Speaker Recognition. In *INTERSPEECH*.

[2] Jiankang Deng, J. Guo, and Stefanos Zafeiriou. 2019. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), 4685–4694.

[3] Brian Dolhansky and Cristian Canton-Ferrer. 2018. Eye In-painting with Exemplar Generative Adversarial Networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), 7902–7911.

[4] Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman. 2020. Voxceleb: Large-scale speaker verification in the wild. *Computer Speech & Language* 60 (2020), 101027.

[5] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. FaceForensics++: Learning to Detect Manipulated Facial Images. In *International Conference on Computer Vision (ICCV)*.

[6] H. Wang, Yitong Wang, Zheng Zhou, Xing Ji, Zhifeng Li, Dihong Gong, Jin Zhou, and Wei Liu. 2018. CosFace: Large Margin Cosine Loss for Deep Face Recognition. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), 5265–5274.

[7] Yandong Wen, Bhiksha Raj, and Rita Singh. 2019. Face reconstruction from voice using generative adversarial networks. *Advances in neural information processing systems* 32 (2019).

[8] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. 2021. 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*. 3661–3670.