Fine-portraitist: Visualizing the Speaker's Face Portrait during Speech Listening

1st Jinting Wang*
Artificial Intelligence
HKUST (GZ)
Guangzhou, China
jwang644@connect.hkust-gz.edu.cn

2nd Li Liu†
Artificial Intelligence
HKUST (GZ)
Guangzhou, China
avrillliu@hkust-gz.edu.cn

3rd Jun Wang
Tencent AI Lab
Tencent
Shenzhen, China
junjunmin@gmail.com

***Abstract*—Speech-to-portrait generation (S2P) plays a crucial role in speech-driven, human-centered creative content generation, aiming to synthesize a speaker's face portrait with identity consistency from a given speech clip. However, existing S2P methods can typically only preserve attribute consistency, *e.g.,* gender and age, while failing to capture the more important part-appearance consistency due to the coarse speech-face correlation. In this work, we propose Fine-portraitist, a novel retrieval-augmented, easy-to-hard generation framework designed to tackle this problem. Specifically, Fine-portraitist enhances identity consistency in S2P through two key innovations: 1) We first explore the fine-grained speech-face correlation by decomposing the face portrait into speech-related and speech-unrelated parts. Based on this, we propose a two-stage, diffusion-based pipeline to progressively achieve S2P; 2) A retrieval prior is introduced, selected from a retrieval database based on speech feature similarity, providing supplementary external information for more accurate and realistic generation results. Extensive experiments on two datasets, *i.e.,* AVSpeech and VoxCeleb, demonstrate that Fine-portraitist significantly outperforms existing S2P methods.**

***Index Terms*—Speech-to-Portrait, Diffusion model, Retrieval augmentation generation, Cross-modal learning**

I. Introduction

Speech-to-portrait (S2P) generative models, including GAN-based [1], [2] and diffusion-based approaches [3], have undergone significant advancements in recent years. Given an audio speech, these methods endeavors to create the speaker's face portrait that is coherent with an audio speech. This technique attracts significant public interest in their potential applications, such as voice-based crimes.

A critical requirement for S2P is maintaining identity consistency, meaning the generated portrait must not only reflect the speaker's attributes but also preserve appearance consistency. However, as illustrated in Fig. 1, existing one-stage generators struggle to achieve accurate appearance consistency due to several challenges: **1)** these methods often rely on coarse semantic correlations, such as gender, age, and ethnicity [4], [5], without a deeper understanding of which specific facial features can be predicted by speech. This uncertainty increases model instability, resulting in generated portraits that lack accurate appearance consistency. **2)** The facial features available in short speech clips are often limited,

*Work done during the internship at Tencent AI Lab.
†Corresponding Author

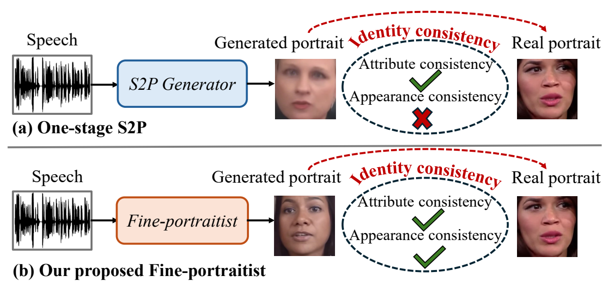

Fig. 1. Comparison with SOTA S2P methods. Our Fine-portraitist can not only achieve high attribute consistency, *i.e.,* gender, age, but also shows excellent performance in appearance consistency.

making it difficult to generate precise facial portraits based solely on speech, especially in real-world scenarios. To address these challenges, **firstly**, we conduct a fine-grained analysis of speech-face correlations, distinguishing between speech-related and speech-unrelated facial features. considering that, we propose a progressive generation pipeline for S2P, which first extracts speech-related facial features from the speech, followed by the synthesis of speech-unrelated features while ensuring overall facial coherence. **secondly**, to further enhance identity consistency, we incorporate a retrieval face prior as supplementary information, which helps to more effectively model facial features, particularly the speech-unrelated components.

In summary, the contributions of this work are threefold: **1)** We investigate the fine-grained correlation between speech and facial features and propose a two-stage method that formulates S2P in an easy-to-hard manner; **2)** We design a retrieval prior to guide S2P, enhancing identity consistency by leveraging knowledge from the retrieved samples; **3)** Extensive qualitative and quantitative experiments demonstrate that our Fine-portraitist framework surpasses state-of-the-art (SOTA) S2P methods in terms of identity consistency.

II. Related work

A. Face-Voice Correlation

The human voice reveals traits like gender [6], [7], age [8], [9], and emotion [10], [11], which are used in audio-visual tasks like identity verification [12], [13] and deepfake

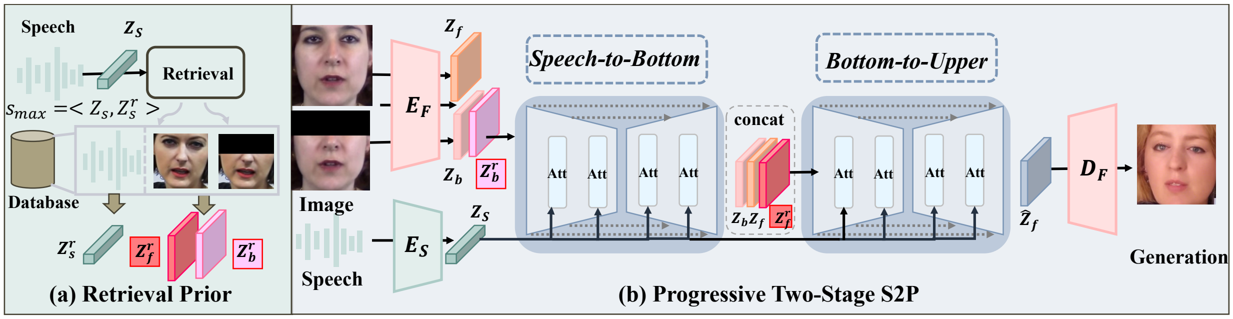

Fig. 2. Overview of Fine-portraitist. (a) The source speech are sent to calculate the similarity with the candidate speech in the retrieval database. And the paired face portraits of the most similar ones are fed into the S2P generation framework, serving as prior information for accurate portrait generation. (b) Our progressive two-stage S2P: The Speech-to-Bottom stage maps the audio speech to speech-related facial bottom part, and the Bottom-to-Upper stage synthesizes speech-unrelated facial up part with the bottom part and speech driven sources.

detection [14], [15] by analyzing lip-speech synchronization. Talking head generation [16], [17] also aligns lip movements with speech for realistic video creation. Prior studies [18], [19] explored the link between facial features and phonemes, aiding speech-synchronized 3D face generation. Unlike existing work focused on semantic or linguistic correlations, our research investigates the implicit connections between speech and facial structures in S2P.

B. Speech-to-Portrait Generation

S2P has garnered significant attention in recent years. Some existing methods [1], [20] leverage the rich facial information embedded in speech to design face generators using speech embeddings as input. To preserve shared speaker identity information across audio and visual modalities, certain approaches assign each speaker an identity label, using it to supervise model training [2], [21], [22]. For speech-to-face generation with random identities, self-supervised cross-modal identity matching [23] is used to exploit the shared identity information between audio and visual data. Motivated by the success of diffusion models in image generation [24], [25], recent works [3], [26] have applied latent diffusion model (LDM) to S2P, achieving higher-quality results compared to GAN-based methods [1], [2], [20]–[23]. In this work, we propose a diffusion-based framework designed for accurate identity-consistent generation, based on a fine-grained investigation of speech-face correlation in open scenarios.

III. Methods

A. Exploring Fine-grained Speech-Face Correlation

Human speech is produced by phonatory structures [27], which are likely crucial for generating facial portraits from speech. Although both speech and facial features convey speaker identity information, only certain facial regions directly involved in phonation, such as the jaw, mouth, and nose, are hypothesized to be predictable from speech, as hypothesized in our task design. To explicitly test the correlation between speech and specific facial features, we conduct a toy experiment. In this experiment, we assess the generation accuracy of various facial parts, including the eyes, eyebrows, nose, lips, and jaw, using their respective accuracy as a measure of correlation with the speech input. If a trained generator achieves higher accuracy on certain facial parts when utilizing speech input as opposed to without it, and the results are statistically significant, we infer that those facial features are speech-related, and vice versa. Using $N = 5000$ samples, we conducted a t-test on the generation results, setting the significance level at 95%, and subsequently referencing $t_{(0.95,4999)}$ from the t-distribution table. As shown in Table I, the probabilities for the jaw (t_{jaw}), mouth (t_{mouth}), and nose (t_{nose}) exceed $t_{(0.95,4999)}$, indicating statistical significance. Consequently, we confirm that the **bottom face**, comprising the nose and the following part, is speech-related, while the **upper face** is not.

TABLE I
The paired t-test results on facial parts.

$t_{(0.95,4999)}$	t_{jaw}	t_{mouth}	t_{nose}	t_{eyes}	$t_{eyebrows}$
1.96	2.85	4.94	2.26	0.47	0.32

B. Retrieval Prior-Guided Speech-to-Portrait Generation

Based on the investigation in Section III-A, facial structures can be divided into speech-related (bottom face) and speech-unrelated (upper face) parts. Therefore, the goal of this section is to generate a facial image from the input audio using a Speech-to-Bottom and Bottom-to-Upper pipeline, as illustrated in Fig. 2. To further improve the generation process, we incorporate a retrieval face prior as supplementary information.

Retrieval Prior. RAG is demonstrated effectiveness in multiple generation tasks [28]–[30], here, we employ RAG to provide face prior information for S2P enhancement. In detail, we first evaluate the feature similarities between the given speech clip and the candidates in the database. For each speech clip s, we extract $Z_s = E_S(s)$ as the speech query feature, where E_S is the pre-trained speech encoder [31]. Then, the speech features guide the retrieval process by selecting the sample with the highest similarity, calculated as: $s_{max} = Max < Z_s, Z_s^{ri} >$, where $\langle \cdot, \cdot \rangle$ denotes the cosine similarity between the two feature vectors, and Z_s^{ri} is the speech feature of the i_{th} sample in the retrieval database. The corresponding face portrait and bottom face portrait with the

TABLE II
COMPARISON RESULTS ON AVSEECH DATASET. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD. NOTE THAT ↓ INDICATES THAT A SMALLER VALUE IS PREFERABLE, WHILE ↑ INDICATES THAT A LARGER VALUE IS PREFERABLE.

Method	Year	Feature Similarity			Identity Preservation		Retrieval Performance		
		L1 ↓	L2 ↓	cos ↓	gender (%) ↑	age (%) ↑	$R@1$ ↑	$R@2$ ↑	$R@5$ ↑
Wav2Pix [1]	2019	144.72	24.32	82.51	67.4	41.3	2.46	6.72	14.26
Speech2Face [4]	2019	67.18	3.94	46.97	95.6	65.2	9.17	14.94	28.31
Choi *et al.* [32]	2019	60.26	3.57	35.89	95.8	69.6	10.84	17.37	32.91
SF2F [22]	2022	89.31	17.49	64.83	72.1	48.9	7.37	13.45	20.72
Kato *et al.* [3]	2023	46.35	2.73	21.96	96.7	81.3	18.44	28.31	49.24
Fine-portraitist (Ours)	-	**22.78**	**0.58**	**6.26**	**99.6**	**89.9**	**26.23**	**47.46**	**72.42**

TABLE III
COMPARISON RESULTS ON VOXCELEB DATASET. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Method	Year	Feature Similarity			Identity Preservation		Retrieval Performance		
		L1 ↓	L2 ↓	cos ↓	gender (%) ↑	age (%) ↑	$R@1$ ↑	$R@2$ ↑	$R@5$ ↑
Wav2Pix [1]	2019	137.58	22.19	79.36	74.5	49.6	4.81	9.56	12.94
Speech2Face [4]	2019	66.46	2.77	44.38	96.1	69.4	7.79	14.38	20.14
Wen *et al.* [2]	2019	59.82	2.41	42.54	97.4	72.5	8.26	15.62	23.51
Choi *et al.* [32]	2019	56.32	2.24	30.49	97.6	74.8	9.43	16.32	28.67
SF2F [22]	2022	78.45	13.31	58.79	79.3	57.6	9.25	17.17	22.53
Kato *et al.* [3]	2023	40.11	2.26	18.74	98.1	83.8	16.19	25.64	42.38
Fine-portraitist (Ours)	-	**18.86**	**0.64**	**5.35**	**99.8**	**94.7**	**28.87**	**49.72**	**78.94**

highest similarity are fed into the pre-trained face encoder E_F to obtain the retrieval priors *e.g.,* the retrieved full face features Z_f^r and the bottom face features Z_b^r. To construct the retrieval database, we simply use all the training data as entities. The retrieved face priors are concatenated with a noise vector as input for the S2P generation pipeline.

Speech-to-Bottom Generation. Given a source audio clip, our goal in this stage is to train a model for bottom face portrait generation while preserving the identity information conveyed in the speech condition. We employ a pre-trained audio extractor E_S and face encoder E_F to derive speech representations Z_s and bottom face features Z_b, respectively. In this setup, the speech features Z_s serve as a basic condition, while the retrieved bottom face prior Z_b^r is introduced as an additional condition to guide the denoising process of LDM. The objective function is defined as:

$$L_{LDM}^b := \mathbb{E}_{Z_b^t, Z_s, Z_b^r, \epsilon, t}[\|\epsilon - \mathcal{M}(Z_b^t, Z_s, Z_b^r, t)\|_2^2],$$

where ϵ represents Gaussian noise, Z_b^t is the noised version of Z_b during the diffusion process, and t denotes the time steps.

Bottom-to-Upper Generation. Given that the bottom face is closely related to speech and there are physiological and anatomical connections between the bottom and upper face, we propose a bottom-augmented approach for upper face generation. Specifically, for a given speech clip, we first utilize the speech-to-bottom generation module to synthesize the speech-related bottom face. The features from this generated bottom face are then used as additional conditions to guide the learning process for upper face generation. Formally, we use a pre-trained face encoder E_F to extract both the full face features Z_f and the bottom face features Z_b. And then the full face features Z_f are corrupted into Z_f^t by sequentially injecting Gaussian noise ϵ at t time steps. The noised features are concatenated with the bottom-face features Z_b and the retrieved face prior Z_f^r along the channel dimension. This concatenated result is then fed into the LDM to learn the upper face generation conditioned on the speech input. The objective function can be formulated as:

$$L_{LDM}^u := \mathbb{E}_{Z_s, Z_f^t, Z_f^r, \epsilon, t}\left[\|\epsilon - \mathcal{M}(Z_s, Z_f^t, Z_f^r, Z_b, t)\|_2^2\right].$$

Finally, the face decoder D_F recovers the generated face latent $\hat{Z}_f$ into portrait image.

IV. EXPERIMENTS

A. Datasets

Following existing S2P methods, we conduct our experiments using two datasets: the AVSpeech dataset [33] and VoxCeleb [34].

B. Implementation Details

We extract 6-second speech segments from video clips and convert them into spectrograms using the Short-Time Fourier Transform. The face images are first cropped and then resized to 256×256 pixels. We perform collaborative pre-training on the audio extractor, face encoder, and face decoder to ensure audio-visual alignment and accurate face reconstruction. The learning rate for the face encoder and decoder is set to 0.0001, while the speech encoder is trained with a learning rate of 0.001. In the two-stage generation pipeline, the face encoder, face decoder, and speech encoder are frozen. The Speech-to-Bottom and Bottom-to-Upper modules are independently trained using the Adam optimizer.

C. Evaluation Metrics

Feature Similarity. Following [4], we measure cosine, L1, and L2 distances between the features of the ground truth face image and the generated face image, both extracted using VGGFace [35]. **Identity Preservation.** We employ the Face++[1] commercial API for face attribute recognition to

[1]https://www.faceplusplus.com/attributes.

TABLE IV
ABLATION RESULTS ON AVSPEECH DATASET. "ONE STAGE" MEANS DIRECTLY GENERATE FACE PORTRAIT ONLY WITH SPEECH CONDITION. "TWO STAGE*" MEANS WITHOUT RETRIEVAL PRIOR IN THE TWO STAGE GENERATION PIPELINE.

Method	Feature Similarity			Identity Preservation		Retrieval Performance		
	L1 ↓	L2 ↓	cos ↓	gender (%) ↑	age (%) ↑	$R@1$ ↑	$R@2$ ↑	$R@5$ ↑
One-stage	44.27	2.38	20.41	96.4	80.3	18.97	29.32	49.96
Two-stage*	35.31	1.46	14.29	97.3	83.1	20.94	27.17	54.82
Fine-portraitist (Ours)	**22.78**	**0.58**	**6.26**	**99.6**	**89.9**	**26.23**	**47.46**	**72.42**

evaluate attributes such as age and gender. Age classification is considered accurate if the age difference between the generated face image and the ground truth is within 10 years. **Retrieval Performance.** Image retrieval involves analyzing the visual content of a large image database to find images that match the query image in terms of semantics or similarity [36]. We report retrieval performance using the Recall@K metric, including $R@1$, $R@2$, and $R@5$, which indicates whether the top K retrieved images contain a true match [37].

D. Comparisons with SOTAs

We compare our proposed method with six SOTA S2P methods, categorized into three groups: 1) **CNN-based** methods, such as Speech2Face [4] and SF2F [22]; 2) **GAN-based** methods, such as Wav2Pix [1], Wen *et al.* [2], and Choi *et al.* [32]; 3) **LDM-based** method, Kato *et al.* [3]. We perform experiments using the default settings and official implementations for Wav2Pix [1], Wen *et al.* [2], Choi *et al.* [32], SF2F [22], and Kato *et al.* [3]. However, as the code for Speech2Face [4], Choi *et al.* [32], and Kato *et al.* [3] is not available, we reproduce them based on the descriptions provided in their papers. Additionally, we only compare with Wen *et al.* on the VoxCeleb dataset, as the identity information of speakers is lacking in the AVSpeech dataset.

Quantitative Comparison. The comparison results on AVSpeech and VoxCeleb datasets are reported in Table II and Table III, respectively. Our method outperforms all the competitors in all metrics. Specifically, the cosine distance of our method achieves 7.26 on the AVSpeech test set and 6.35 on the VoxCeleb test set. The gender recognition accuracy achieves 99.4 and 99.8 on the two datasets. These results verify the effectiveness of our approach in producing identity-preserving portraits.

Qualitative Comparison. The qualitative comparisons shown in Fig. 3 and Fig. 4 underscore the effectiveness of our approach in generating realistic outputs that closely align with the speaker's attributes. This success is largely due to our two-stage generation pipeline, which divides the process into voice-related and voice-unrelated components. By employing this easy-to-hard strategy, our model achieves superior performance compared to prior methods, resulting in synthesized portraits that more accurately resemble the speakers.

E. Ablation studies

We conducted ablation studies on the AVSpeech dataset to validate the effectiveness of various components. The comparison results for different model versions are presented in Table IV. By comparing the one-stage approach with the two-stage* approach, we observe a performance gain attributed to the use of the easy-to-hard paradigm. Furthermore, the comparison between the two-stage* approach and Fine-portraitist shows that the retrieval prior offers valuable references, leading to more accurate portrait generation.

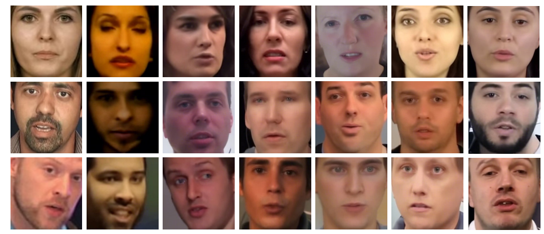

Fig. 3. Qualitative comparison between our model and previous SOTA methods on the AVSpeech dataset.

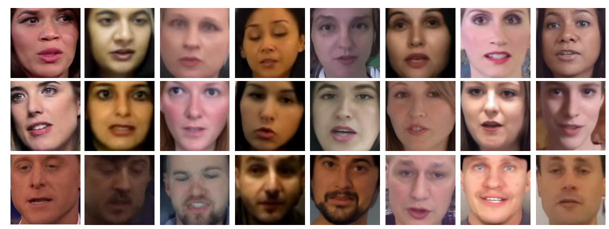

Fig. 4. Qualitative comparison between our model and previous SOTA methods on the Voxceleb dataset.

V. CONCLUSION

In this work, a novel retrieval-prior-guided generation framework (Fine-portraitist) is designed to improve the identity consistency of S2P. We investigate the fine-grained correlation between speech and facial features, which informs the design of our progressive two-stage generation process. By incorporating a retrieval face prior, Fine-portraitist achieves significant improvements in overall performance. Comparisons with state-of-the-art models across multiple performance metrics demonstrate that Fine-portraitist excels in generating identity-consistent face portraits.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (No. 62471420 and 62101351), the Tencent AI Lab (RBFR2023014), the Guangzhou-HKUST (GZ) Joint Funding Program (Grant No. 2023A03J0008), and the Education Bureau of Guangzhou Municipality.

References

[1] Amanda Cardoso Duarte, Francisco Roldan, Miquel Tubau, Janna Escur, Santiago Pascual, Amaia Salvador, Eva Mohedano, Kevin McGuinness, Jordi Torres, and Xavier Giro-i Nieto, "Wav2pix: Speech-conditioned face generation using generative adversarial networks.," in *ICASSP*, 2019, pp. 8633–8637.

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

[3] Shuhei Kato and Taiichi Hashimoto, "Speech-to-face conversion using denoising diffusion probabilistic models," .

[4] Tae-Hyun Oh, Tali Dekel, Changil Kim, Inbar Mosseri, William T Freeman, Michael Rubinstein, and Wojciech Matusik, "Speech2face: Learning the face behind a voice," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 7539–7548.

[5] Jianrong Wang, Jinyu Liu, Xuewei Li, Mei Yu, Jie Gao, Qiang Fang, and Li Liu, "Two-stream joint-training for speaker independent acoustic-to-articulatory inversion," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[6] Jianrong Wang, Zixuan Wang, Xiaosheng Hu, Xuewei Li, Qiang Fang, and Li Liu, "Residual-guided personalized speech synthesis based on face image," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4743–4747.

[7] Zhong-Qiu Wang and Ivan Tashev, "Learning utterance-level representations for speech emotion and age/gender recognition using deep neural networks," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 5150–5154.

[8] Joanna Grzybowska and Stanislaw Kacprzak, "Speaker age classification and regression using i-vectors.," in *INTERSPEECH*, 2016, pp. 1402–1406.

[9] Rita Singh, Joseph Keshet, Deniz Gencaga, and Bhiksha Raj, "The relationship of voice onset time and voice offset time to physical age," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5390–5394.

[10] Yan Rong and Li Liu, "Seeing your speech style: A novel zero-shot identity-disentanglement face-based voice conversion," *arXiv preprint arXiv:2409.00700*, 2024.

[11] Licai Sun, Bin Liu, Jianhua Tao, and Zheng Lian, "Multimodal cross- and self-attention network for speech emotion recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 4275–4279.

[12] Leda Sarı, Kritika Singh, Jiatong Zhou, Lorenzo Torresani, Nayan Singhal, and Yatharth Saraf, "A multi-view approach to audio-visual speaker verification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6194–6198.

[13] Ruijie Tao, Rohan Kumar Das, and Haizhou Li, "Audio-visual speaker recognition with a cross-modal discriminative network," *preprint arXiv:2008.03894*, 2020.

[14] Yipin Zhou and Ser-Nam Lim, "Joint audio-visual deepfake detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14800–14809.

[15] Wenyuan Yang, Xiaoyu Zhou, Zhikai Chen, Bofei Guo, Zhongjie Ba, Zhihua Xia, Xiaochun Cao, and Kui Ren, "Avoid-df: Audio-visual joint learning for detecting deepfake," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 2015–2029, 2023.

[16] Weichuang Li, Longhao Zhang, Dong Wang, Bin Zhao, Zhigang Wang, Mulin Chen, Bang Zhang, Zhongjian Wang, Liefeng Bo, and Xuelong Li, "One-shot high-fidelity talking-head synthesis with deformable neural radiance field," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17969–17978.

[17] 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*, 2023, pp. 14653–14662.

[18] Liao Qu, Xianwei Zou, Xiang Li, Yandong Wen, Rita Singh, and Bhiksha Raj, "The hidden dance of phonemes and visage: Unveiling the enigmatic link between phonemes and facial features," *preprint arXiv:2307.13953*, 2023.

[19] Xiang Li, Yandong Wen, Muqiao Yang, Jinglu Wang, Rita Singh, and Bhiksha Raj, "Rethinking voice-face correlation: A geometry view," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 2458–2467.

[20] Jingyi Mao, Yuchen Zhou, Junyu Li, Ziqing Liu, and Fanliang Bu, "Audio-face generating using squeeze-and-excitation enhanced generative adversarial network," in *2023 IEEE 5th International Conference on Power, Intelligent Computing and Systems (ICPICS)*. IEEE, 2023, pp. 966–970.

[21] Zheng Fang, Zhen Liu, Tingting Liu, Chih-Chieh Hung, Jiangjian Xiao, and Guangjin Feng, "Facial expression gan for voice-driven face generation," *The Visual Computer*, pp. 1–14, 2022.

[22] Yeqi Bai, Tao Ma, Lipo Wang, and Zhenjie Zhang, "Speech fusion to face: Bridging the gap between human's vocal characteristics and facial imaging," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 2042–2050.

[23] Hyeong-Seok Choi, Changdae Park, and Kyogu Lee, "From inference to generation: End-to-end fully self-supervised generation of human face from speech," *preprint arXiv:2004.05830*, 2020.

[24] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10684–10695.

[25] Prafulla Dhariwal and Alexander Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[26] Jinting Wang, Li Liu, Jun Wang, and Hei Victor Cheng, "Realistic speech-to-face generation with speech-conditioned latent diffusion model with face prior," *preprint arXiv:2310.03363*, 2023.

[27] Silvana Bommarito, Patricia Barbarini Takaki, Angélica da Veiga Said, Marcela Dinalli Gomes Barbosa, Gisele da Silva Dalben, Eduardo Kazuo Sannomiya, and Marilena Manno Vieira, "Correlation between voice, speech, body and facial types in young adults," *Global Journal of Otolaryngology*, vol. 20, no. 4, 2019.

[28] Deng Cai, Yan Wang, Lemao Liu, and Shuming Shi, "Recent advances in retrieval-augmented text generation," in *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, 2022, pp. 3417–3419.

[29] Xin Cheng, Di Luo, Xiuying Chen, Lemao Liu, Dongyan Zhao, and Rui Yan, "Lift yourself up: Retrieval-augmented text generation with self-memory," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[30] Yi Yuan, Haohe Liu, Xubo Liu, Qiushi Huang, Mark D Plumbley, and Wenwu Wang, "Retrieval-augmented text-to-audio generation," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 581–585.

[31] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel, "Audioclip: Extending clip to image, text and audio," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 976–980.

[32] Hyeong-Seok Choi, Changdae Park, and Kyogu Lee, "From inference to generation: End-to-end fully self-supervised generation of human face from speech," in *International Conference on Learning Representations*, 2019.

[33] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *preprint arXiv:1804.03619*, 2018.

[34] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "Voxceleb: a large-scale speaker identification dataset," *preprint arXiv:1706.08612*, 2017.

[35] Zakariya Qawaqneh, Arafat Abu Mallouh, and Buket D Barkana, "Deep convolutional neural network for age estimation based on vgg-face model," *preprint arXiv:1709.01664*, 2017.

[36] Mehwish Rehman, Muhammad Iqbal, Muhammad Sharif, and Mudassar Raza, "Content based image retrieval: survey," *World Applied Sciences Journal*, vol. 19, no. 3, pp. 404–412, 2012.

[37] Yihan Wu, Hongyang Zhang, and Heng Huang, "Retrievalguard: Provably robust 1-nearest neighbor image retrieval," in *International Conference on Machine Learning*. PMLR, 2022, pp. 24266–24279.