INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Speech-to-Face Conversion Using Denoising Diffusion Probabilistic Models

Shuhei Kato[1], *Taiichi Hashimoto*[1]

[1]RevComm Inc., Japan
{kato,taiichi.hashimoto}@revcomm.co.jp

Abstract

Speech-to-face conversion is the task of generating face images from speech signals. Many studies have been conducted to address this task, and achieved good performances. In this paper, we introduce denoising diffusion probabilistic models (DDPMs) to generate face images instead of generative adversarial networks (GANs) or autoencoders, which are used in most of the prior studies. Moreover, unlike prior studies, several components of our system are designed to use high-resolution face image datasets instead of audio-visual paired data. As a result, our system can generate high-resolution face images from speech signals with an architecture that is simpler and more flexible than the ones used in prior studies. In addition, introducing DDPMs enables us to utilize techniques that control outputs of DDPMs or improve performance of them in succeeding studies.

Index Terms: speech-to-face, denoising diffusion probabilistic model, classifier-free guidance

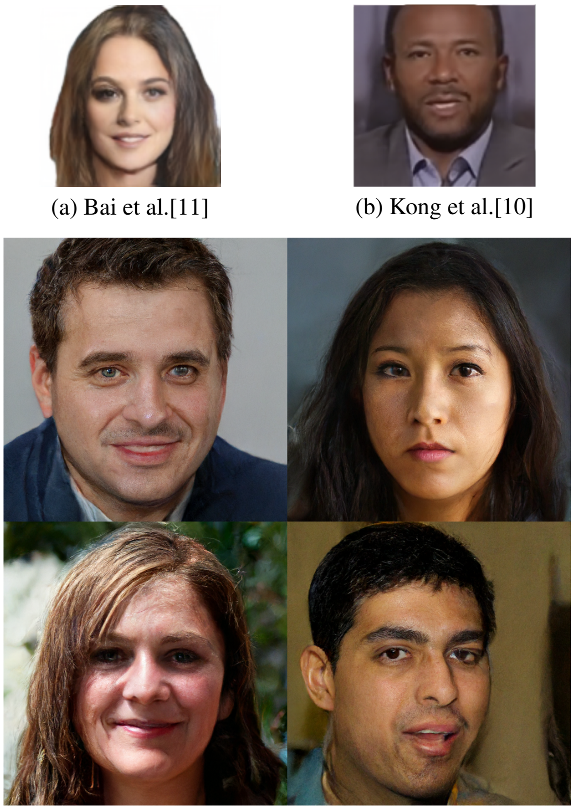
1. Introduction

Humans have the ability to imagine how a speaker looks when they hear a voice but cannot see the person [1, 2]. This relationship between speech and appearance is partially obvious, considering a speech signal conveys various speaker attributes such as age and gender [3]. Recent developments in machine learning have enabled researchers to take on the challenge in generating a human's face solely from his or her speech while keeping the speaker's attributes [4, 5, 6, 7, 8, 9, 10, 11]. Such a *speech-to-face* conversion technique can be applied, for example, to generating virtual face images reflecting speakers' attributes on the phone apps, or generating avatars for entertainment purposes or privacy protection.

Most of prior studies use generative adversarial networks (GANs) [12] or autoencoders, and have achieved good performance. Despite their success, because audio-visual paired data are needed to train the whole system, the output image resolution is limited.

In the field of image generation, denoising diffusion probabilistic models (DDPMs) [13, 14] have achieved state-of-the-art performance in recent years [15, 16, 17, 18, 19]. DDPMs tend to have the ability to generate samples that are more diverse than those of GANs and autoencoders while maintaining or improving quality, and can be stably trained with stationary loss functions. In addition, we can use simple and flexible methods to train conditional models [17, 20].

In this paper, we propose a novel method of speech-to-face conversion using DDPMs and investigated how well generated face images reflect the speakers' attributes. Generated image samples by our system and prior studies are shown in Fig. 1. The key contributions of this paper are as follows:



(a) Bai et al.[11]

(b) Kong et al.[10]

(c) Ours

Figure 1: *Samples of generated face images generated from speech. (a): Generated by a GAN-based decoder* (128×128)*. (b): Generated by an autoencoder-based decoder* (128×128)*. (c): Ours (DDPM-based, each image is* 512×512*). Samples (a) and (b) are referenced from original papers.*

- We introduce DDPMs to speech-to-face conversion to replace GANs or autoencoders. DDPMs enable our system to generate high-resolution face images with a simpler and more flexible architecture. Introducing DDPMs also enables us to apply techniques for controlling outputs of DDPMs or improving performance of them in succeeding studies.
- In contrast to prior studies, audio-visual paired data is only required to train the speech encoder, not the whole system. To train the face decoder and super-resolution model, we can use high-resolution face image datasets, which are more effortlessly available and easier to build.

10.21437/Interspeech.2023-1358

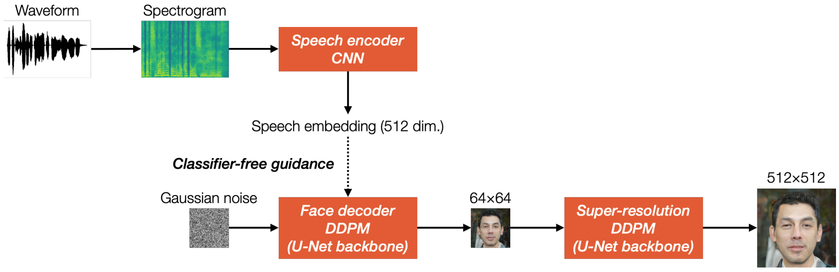


Figure 2: *Inference structure of our speech-to-face conversion system.*

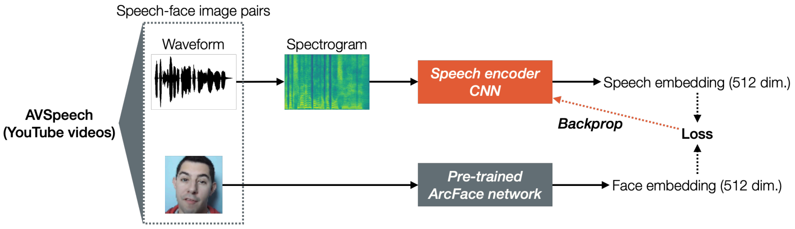


Figure 3: *Training structure of speech encoder.*

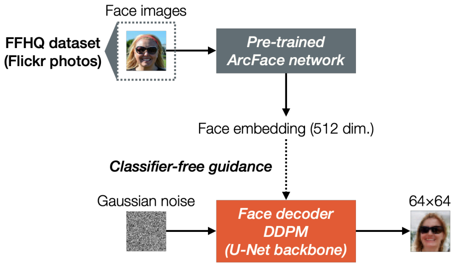


Figure 4: *Training structure of face decoder.*

2. Related work

2.1. Speech-to-face conversion

Speech-to-face conversion is the task of generating face images from speech signals while keeping the speakers' attributes. Oh et al. [5] use a voice encoder that predicts an embedding that is decoded to a normalized face image by an autoencoder-based pre-trained face decoder [21]. The voice encoder is trained to predict embeddings as close to the corresponding face embeddings produced by a pre-trained face recognizer using a paired dataset of speech and face images. Instead of generating normalized face images, as in Oh et al. [5], most subsequent studies generate (non-normalized) face images using GAN-based [6, 7, 9, 11] or autoencoder-based [8, 10] decoders that are trained simultaneously with the other components.

2.2. Denoising diffusion probabilistic models

DDPMs [13, 14] are a class of latent variable models that convert Gaussian noise into samples matching a data distribution via a finite iterative denoising process. Conditional models are possible, for example on class labels, text, low-resolution images, or spectrograms [15, 17, 22, 23, 18, 24, 25, 26]. Let $\mathbf{x}_1, \ldots, \mathbf{x}_T$ be a sequence of latents with the same dimension as the data $\mathbf{x}_0 \sim q(\mathbf{x}_0)$. The forward diffusion process is defined by a variance schedule $\beta_1, \ldots, \beta_T$ and condition $\mathbf{c}$ as

$$q(\mathbf{x}_{1:T} \,|\, \mathbf{x}_0, \mathbf{c}) = \prod_{t=1}^{T} q(\mathbf{x}_t \,|\, \mathbf{x}_{t-1}, \mathbf{c}) \tag{1}$$

where

$$q(\mathbf{x}_t \,|\, \mathbf{x}_{t-1}, \mathbf{c}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}) \tag{2}$$

The reverse backward denoising process is defined as

$$q(\mathbf{x}_{0:T-1} \,|\, \mathbf{x}_0, \mathbf{c}) = \prod_{t=1}^{T} q(\mathbf{x}_{t-1} \,|\, \mathbf{x}_t, \mathbf{c}) \tag{3}$$

and the models are trained to minimize the objective

$$\begin{aligned} &\mathbb{E}_{\mathbf{x}_0, \mathbf{c}, \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), t \sim \mathcal{U}(\{1,\ldots,T\})} [\|\, \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{c}) \,\|_2^2] & (4)\\ &= \mathbb{E}_{\mathbf{x}_0, \mathbf{c}, \boldsymbol{\epsilon}, t} [\|\, \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \mathbf{c}) \,\|_2^2] & (5) \end{aligned}$$

where $\boldsymbol{\epsilon}_\theta$ is a estimator to predict $\boldsymbol{\epsilon}$, $\alpha_t = 1 - \beta_t$, and $\bar{\alpha}_t = \prod_{t'=1}^{t} \alpha_{t'}$.

Classifier-free guidance [20] is a simple and flexible method for training conditional DDPMs. In this method, DDPMs are trained on conditional and unconditional objectives by randomly (e.g., with a 10% probability) dropping $\mathbf{c}$ during training. When sampling, we use the following linear combination of conditional and unconditional $\boldsymbol{\epsilon}$-estimators

$$\tilde{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t, \mathbf{c}) = (1+w)\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{c}) - w\boldsymbol{\epsilon}_\theta(\mathbf{x}_t) \tag{6}$$

where $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t) = \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{c} = \mathbf{0})$ and w is the guidance weight. Setting $w = 0$ means no guidance, and larger $w > 0$ intensify the effect of guidance, resulting in improving sample quality while reducing diversity.

DDPMs have outperformed other state-of-the-art methods recently on image generation tasks [15, 16, 17, 18, 19]. In particular, DDPMs have achieved success on conversion tasks, such as text-to-image [18, 27], text-to-video [28, 29], image-to-image [30], text-to-audio [24], spectrogram-to-waveform [25, 26], and text-to-waveform [22, 23] conversions. In this paper, we apply a text-to-image conversion technique in our speech-to-face conversion system. Specifically, our system generates face images with the guidance of speech embeddings.

3. Model architecture and training

3.1. Model architecture

The proposed model consists of three sequentially connected components: a speech encoder, face decoder, and super-resolution model (Fig. 2). Through this pipeline, an input spectrogram of speech data is converted into a face image.

First, the speech encoder converts an input spectrogram into a speech embedding. We use a convolutional-neural-network-based architecture almost the same[1] to the one used in Oh et al. [5]. Second, the face decoder generates a 64 × 64 face image from Gaussian noise with the guidance of a speech embedding predicted by the speech encoder. The decoder is based on a DDPM, the backbone of which is a time-conditional U-Net [33], conditioned by speech embeddings using the classifier-free guidance. Finally, the super-resolution model upsamples a 64 × 64 face image generated by the face decoder to a 512 × 512 one. We take the same super-resolution architecture used in Saharia et al. [18].

This pipeline architecture enables us to use a face image dataset instead of an audio-visual dataset to train the face decoder and super-resolution model, resulting in generate images in higher resolution than those of prior studies[2].

3.2. Training

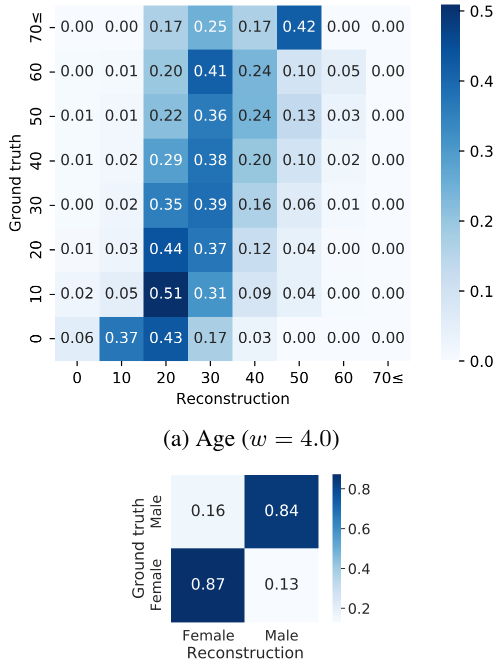
We train the speech encoder, face decoder, and super-resolution model separately.

The speech encoder is trained with an audio-visual dataset such as videos of people talking (Fig. 3). The speaker's face is cropped from a single frame of each video, and the corresponding 512-dimensional face embedding, $\mathbf{v}_f$, is extracted via the ArcFace [31] network. The speech encoder receives an input spectrogram and converts it into a speech embedding $\mathbf{v}_s$, which is expected to approximate $\mathbf{v}_f$ through the training. We choose the cosine distance between $\mathbf{v}_f$ and $\mathbf{v}_s$ as the loss function.

The face decoder and super-resolution model are trained with a high-resolution face image dataset. For the training of the face decoder (Fig. 4), the ArcFace embedding is extracted from each face image. Then the face decoder learns the data distribution of 64 × 64 face images with the guidance of the corresponding face embeddings. For the training of the super-resolution model, each face image is resampled at a 64 × 64 and 512 × 512 image pair, and the model is trained to predict 512 × 512 images from the corresponding 64 × 64 images.

3.3. Implementation details

We select the AVSpeech dataset [34], a large-scale audio-visual dataset from YouTube videos, as the training dataset for the speech encoder. In the manner of Oh et al. [5], up to 6 seconds of audio taken from the beginning of each video in the AVSpeech is transformed into a spectrogram and fed into the speech encoder. If the duration of the video is less than 6 s, the audio is repeated so that it becomes at least 6-s long. All the audio samples are resampled at 16 kHz and converted to single channel samples. Spectrograms are calculated using a short-time Fourier transform with a 25-ms Hann window, 10-ms hop length, and 512 frequency bands. Then, both real and imaginary parts of each spectrogram S are independently compressed as $\operatorname{sgn}(S)\,|S|^{0.3}$, where $\operatorname{sgn}(\cdot)$ denotes the sign function. We use MediaPipe[3] to detect and crop faces from the frames of the AVSpeech videos. Only the frames containing a single face are used for training and testing. As a result, the training, validation, and test sets for the speech encoder have 102 k, 11 k, and 8.2 k spectral-face embedding pairs, respectively. The speech encoder is optimized by LAMB [35], the initial learning rate is 1×10^{-3} with an exponential decay of 0.95 at every 10,000 steps, the batch size is 128, and the speech encoder is trained for 430 k steps.



(a) Age ($w = 4.0$)

(b) Gender ($w = 4.0$)

Figure 5: *Confusion matrices with row-wise normalization comparing classification results on generated images and ground-truth ones for age and gender.*

For the training of the face decoder and super-resolution model, we choose the Flickr-Faces-HQ Dataset [36], a large-scale 1024 × 1024 face image dataset of Flickr photos, as the training dataset. We also use the MediaPipe to detect and crop faces. As a result, the training and validation sets have 60 k and 6.6 k face images, respectively. The hyperparameters for the face decoder and super-resolution model are described in Table 1.

4. Evaluation

4.1. Facial attribute evaluation

To assess how well our models capture facial attributes, we compare the age and gender estimated by inaFaceAnalyzer[4] of the generated images and ground-truth ones. As shown in Table 2, larger values of the classifier-free guidance weight w tend to increase similarity, as is expected considering the function of the classifier-free guidance. Confusion matrices for each of the attributes when the value of w is 4.0 are shown in Fig. 5. As can be seen, the classification results for both age and gender have a certain correlation.

[1] The output dimension of the final layer is changed from 4,096 to 512 because we adopt ArcFace [31] (512 dimensions) as the target face embedding in place of VGGFace [32] (4,096 dimensions).

[2] We attempted to generate face images with the guidance of spectrograms using a DDPM that is trained with an audio-visual dataset, but the quality of final upsampled images were unstable. This may be caused by mismatch between datasets used for training the face decoder and super-resolution model.

[3] `https://google.github.io/mediapipe/`

[4] `https://github.com/ina-foss/inaFaceAnalyzer`

Table 1: *Hyperparameters for the face decoder and super-resolution model.*

	Face decoder	Super-resolution model
Diffusion steps	2,000	2,000
Noise schedule	Cosine	Cosine
Input channels	256	64
Number of ResNet [37] blocks	3, 3, 3, 3	1, 1, 2, 4, 8
Channel multiples	1, 2, 3, 4	1, 2, 4, 8, 16
Cross-attention heads	4	4
Self-attention heads	4	4
Self-attention resolution	2, 4, 8	16
Drop rate for classifier-free guidance	0.1	-
Number of groups for group normalization [38]	32	16
Optimizer	LAMB	LAMB
Batch size	64	128
Learning rate scheduling	Linear warmup steps at first 20 epochs from $\times 0.1$ initial rate	
Peak learning rate	2×10^{-4}	5×10^{-5}
Iteration steps	1.9 M	7.4 M

Table 2: *Correlation coefficients and accuracy for the estimated age and gender of generated images and ground-truth ones for various values of classifier-free guidance weight w.*

w	Corr. coef. (age)	Accuracy (gender)
0.0	0.22	0.83
1.0	0.22	0.85
2.0	0.24	0.85
3.0	0.24	**0.86**
4.0	**0.25**	0.85

4.2. Craniofacial attribute evaluation

We also compare craniofacial attributes. The selected measurements are commonly used ones in the literature to capture ratios and distances in the face[5], and are used in Oh et al. [5]. Because both the generated images and ground-truth ones have various facial angles and sizes, we normalize them using Hsu et al.'s method [39] to obtain the frontal face images, and extract attributes from them using the dlib library[6]. The correlation coefficients of the attributes of the generated images and ground-truth ones are listed in Table 3. As found in Oh et al. [5], the "nasal index" has the largest correlation, albeit our value is smaller. In contrast to the face attribute results, there is no tendency in the results when the value of w is changed.

4.3. Limitations of our system

From the perspective of similarity of generated face images to the ground-truth ones, our system has a limitation compared to several existing studies [5, 6, 7, 9, 11]. This is partially because it is practically impossible to use the information from generated face images or the face decoder to train the speech encoder, as do most of these prior studies, as the inference speeds of DDPMs are much slower than the speed of other models such as GANs and autoencoders. Recent improvements in accelerating inference speeds of DDPMs [40, 41] might help alleviate this problem. Although the similarity is not the main topic of this paper, we could select a more sophisticated audio representations and loss function that does not use any information from other components, such as the ones used in Hong et al. [9] to improve performance. We could of course train the face decoder to use ground-truth or any value of age, gender, or any other attributes such as expression, clothes, as guidance and control generated images if we desired to do so during inference. Such flexibility is an advantage of our approach over the prior studies.

Table 3: *Correlation coefficients of the craniofacial attributes of generated images and ground-truth ones. Following Oh et al. [5], the random baseline is calculated for the "nasal index" by comparing random pairs.*

Face measurement	Correlation	p-value
Upper lip height	0.07	< 0.001
Lateral upper lip height	0.03	> 0.05
Jaw width	0.04	< 0.002
Nose height	0.02	< 0.05
Nose width	0.04	< 0.001
Labio oral region	0.03	< 0.02
Mandibular idx	0.02	< 0.05
Intercanthal idx	0.02	> 0.05
Nasal index	**0.08**	< 0.001
Vermilion height idx	-0.01	> 0.05
Mouth face with idx	0.01	> 0.05
Random baseline	0.02	-

5. Conclusions

We proposed a novel method for speech-to-face conversion using DDPMs. Our system comprises a speech encoder, face decoder DDPM, and super-resolution DDPM, and they are separately trained using an audio-visual dataset (speech encoder) or a high-resolution face image dataset (face decoder and super-resolution model) while prior studies need to be train the whole system using an audio-visual dataset. Although there is a limitation on the similarity of the generated images, we succeeded in building a simple and flexible speech-to-face conversion system based on DDPMs that can generate high-resolution face images. We expect our system to be an stepping stone to apply techniques for controlling outputs or improving performance in succeeding speech-to-face conversion studies.

[5] https://arxiv.org/abs/1901.10436

[6] http://dlib.net

6. References

[1] M. Kamachi, H. Hill, K. Lander, and E. Vatikiotis-Bateson, "'Putting the Face to the Voice': Matching Identity Across Modality," *Current Biol.*, vol. 13, pp. 1709–1714, September 2003.

[2] H. M. J. Smith, A. K. Dunn, T. Baguley, and P. C. Stacey, "Matching Novel Face and Voice Identity Using Static and Dynamic Facial Images," *Atten. Percept. Psychophys.*, vol. 78, no. 3, pp. 868–879, January 2016.

[3] O. Lapteva, *Speaker Perception and Recognition: An Integrative Framework for Computational Speech Processing*. Kassel University Press, 2011.

[4] A. Duarte, F. Roldan, M. Tubau, J. Escur, S. Pascual, A. Salvador, E. Mohedano, K. McGuinness, J. Torres, and X. G. i Nieto, "Wav2Pix: Speech-Conditioned Face Generation Using Generative Adversarial Networks," in *Proc. ICASSP*, May 2019.

[5] T.-H. Oh, T. Dekel, C. Kim, I. Mosseri, W. T. Freeman, M. Rubinstein, and W. Matusik, "Speech2Face: Learning the Face Behind a Voice," in *Proc. CVPR*, June 2019, pp. 7539–7548.

[6] Y. Wen, R. Singh, and B. Raj, "Face Reconstruction from Voice Using Generative Adversarial Networks," in *Proc. NeurIPS*, December 2019.

[7] H.-S. Choi, C. Park, and K. Lee, "From Inference to Generation: End-to-End Fully Self-Supervised Generation of Human Face from Speech," in *Proc. ICLR*, April 2020.

[8] H. Liang, L. Yu, G. Xu, B. Raj, and R. Singh, "Controlled AutoEncoders to Generate Faces from Voices," in *Proc. ISVC*, October 2020.

[9] Z. Hong, J. Wang, W. Wei, J. Liu, X. Qu, B. Chen, Z. Wei, and J. Xiao, "When Hearing the Voice, Who Will Come to Your Mind," in *Proc. IJCNN*, July 2021.

[10] C. Kong, B. Chen, W. Yang, H. Li, P. Chen, and S. Wang, "Appearance Matters, So Does Audio: Revealing the Hidden Face via Cross-Modality Transfer," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 1, pp. 423–436, January 2022.

[11] Y. Bai, T. Ma, L. Wang, and Z. Zhang, "Speech Fusion to Face: Bridging the Gap Between Human's Vocal Characteristics and Facial Imaging," in *Proc. ACMMM*, October 2022, pp. 2042–2050.

[12] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," in *Proc. NIPS*, December 2014.

[13] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep Unsupervised Learning Using Nonequilibrium Thermodynamics," in *Proc. ICML*, vol. 37, July 2015, pp. 2256–2265.

[14] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," in *Proc. NeurIPS*, December 2020.

[15] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image Super-Resolution via Iterative Refinement," *IEEE Trans. Pattern Anal. Mach. Intell.*, September 2022.

[16] J. Ho, C. Saharia, W. Chan, D. J. Fleet, M. Norouzi, and T. Salimans, "Cascaded Diffusion Models for High Fidelity Image Generation," *J. Mach. Learning Res.*, vol. 23, January 2022.

[17] P. Dhariwal and A. Nichol, "Diffusion Models Beat GANs on Image Synthesis," in *Proc. NeurIPS*, December 2021.

[18] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, "Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding," in *Proc. NeurIPS*, November 2022.

[19] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," in *Proc. CVPR*, June 2022, pp. 10 684–10 695.

[20] J. Ho and T. Salimans, "Classifier-Free Diffusion Guidance," in *Proc. Deep Generative Models and Downstream Appl. Workshop*, December 2021.

[21] F. Cole, D. Belanger, D. Krishnan, A. Sarna, I. Mosseri, and W. T. Freeman, "Synthesizing Normalized Faces from Facial Identity Features," in *Proc. CVPR*, July 2017.

[22] M. Jeong, H. Kim, S. J. Cheon, B. J. Choi, and N. S. Kim, "Diff-TTS: A Denoising Diffusion Model for Text-to-Speech," in *Proc. INTERSPEECH*, September 2021, pp. 3605–3609.

[23] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, N. Dehak, and W. Chan, "WaveGrad 2: Iterative Refinement for Text-to-Speech Synthesis," in *Proc. INTERSPEECH*, August 2021.

[24] F. Kreuk, G. Synnaeve, A. Polyak, U. Singer, A. Défossez, J. Copet, D. Parikh, Y. Taigman, and Y. Adi, "AudioGen: Textually Guided Audio Generation," in *Proc. ICLR*, May 2023.

[25] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "DiffWave: A Versatile Diffusion Model for Audio Synthesis," in *Proc. ICLR*, May 2021.

[26] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan, "WaveGrad: Estimating Gradients for Waveform Generation," in *Proc. ICLR*, May 2021.

[27] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models," in *Proc. ICML*, vol. 162, July 2022, pp. 16 784–16 804.

[28] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, "Video Diffusion Models," in *Proc. ICLR*, April 2022.

[29] K. Mei and V. M. Patel, "VIDM: Video Implicit Diffusion Models," in *Proc. AAAI*, February 2023.

[30] C. Saharia, W. Chan, H. Chang, C. A. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi, "Palette: Image-to-image diffusion models," in *Proc. SIGGRAPH*, August 2022.

[31] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," in *Proc. CVPR*, June 2019, pp. 4690–4699.

[32] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep Face Recognition," in *Proc. BMVC*, September 2015, pp. 41.1–41.12.

[33] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Proc. MICCAI*, November 2015, pp. 234–241.

[34] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation," in *ACM Trans. Graph.*, vol. 37, no. 4, August 2018, pp. 1–11.

[35] Y. You, J. Li, S. Reddi, J. Hseu, S. Kumar, S. Bhojanapalli, X. Song, J. Demmel, K. Keutzer, and C.-J. Hsieh, "Large Batch Optimization for Deep Learning: Training BERT in 76 Minutes," in *Proc. ICLR*, April 2020.

[36] T. Karras, S. Laine, and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," in *Proc. CVPR*, June 2019, pp. 4401–4410.

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. CVPR*, June 2016.

[38] Y. Wu and K. He, "Group Normalization," in *Proc. ECCV*, September 2018.

[39] G.-S. Hsu and C.-H. Tang, "Dual-View Normalization for Face Recognition," *IEEE Access*, vol. 8, pp. 147 765–147 775, August 2020.

[40] F. Bao, C. Li, J. Zhu, and B. Zhang, "Analytic-DPM: An Analytic Estimate of the Optimal Reverse Variance in Diffusion Probailistic Models," in *Proc. ICLR*, April 2022.

[41] M. W. Y. Lam, J. Wang, D. Su, and D. Yu, "BDDM: Bilateral Denoising Diffusion Models for Fast and High-Quality Speech Synthesis," in *Proc. ICLR*, April 2022.