S2IGAN: Speech-to-Image Generation via Adversarial Learning

Xinsheng Wang[1,2]*, Tingting Qiao*[2,3]*, Jihua Zhu*[1]**, Alan Hanjalic*[2]*, Odette Scharenborg*[2]

[1]School of Software Engineering, Xi'an Jiaotong University, China.
[2]Multimedia Computing Group, Delft University of Technology, Delft, The Netherlands.
[3]College of Computer Science and Technology, Zhejiang University, China.

`wangxinsheng@stu.xjtu.edu.cn, qiaott@zju.edu.cn, zhujh@xjtu.edu.cn, a.hanjalic@tudelft.nl,`
`o.e.scharenborg@tudelft.nl`

Abstract

An estimated half of the world's languages do not have a written form, making it impossible for these languages to benefit from any existing text-based technologies. In this paper, a speech-to-image generation (S2IG) framework is proposed which translates speech descriptions to photo-realistic images without using any text information, thus allowing unwritten languages to potentially benefit from this technology. The proposed S2IG framework, named S2IGAN, consists of a speech embedding network (SEN) and a relation-supervised densely-stacked generative model (RDG). SEN learns the speech embedding with the supervision of the corresponding visual information. Conditioned on the speech embedding produced by SEN, the proposed RDG synthesizes images that are semantically consistent with the corresponding speech descriptions. Extensive experiments on datasets CUB and Oxford-102 demonstrate the effectiveness of the proposed S2IGAN on synthesizing high-quality and semantically-consistent images from the speech signal, yielding a good performance and a solid baseline for the S2IG task.

Index Terms: Speech-to-image generation, multimodal modelling, speech embedding, adversarial learning.

1. Introduction

The recent development of deep learning and Generative Adversarial Networks (GAN) [1, 2, 3] led to many efforts being carried out on the task of image generation conditioned on natural languages [4, 5, 6, 7, 8, 9]. Although great progress has been made, most of the existing natural language-to-image generation systems use text descriptions as their input, also referred to as Text-to-Image Generation (T2IG). Recently, a speech-based task was proposed in which face images are synthesized conditioned on speech [10, 11]. This task, however, only considers the acoustic properties of the speech signal, but not the language content. Here, we present a natural language-to-image generation system that is based on a spoken description, bypassing the need for text. We refer to this new task as Speech-to-Image Generation (S2IG). This is similar to the recently proposed task of speech-to-image translation task [12].

This work is motivated by the fact that an estimated half of the 7,000 languages in the world do not have written forms [13] (so-called unwritten languages), which makes it impossible for these languages to benefit from any existing text-based technologies, including text-to-image generation. The Linguistic Rights as included in the Universal Declaration of Human Rights state that it is a human right to communicate in ones native language. For these unwritten languages, it is essential to develop a system that bypasses text and maps speech descriptions to images. Moreover, even though existing knowledge and methodology

make 'speech2text2image' transfer possible, directly mapping speech to images might be more efficient and straightforward.

In order to synthesize plausible images based on speech descriptions, speech embeddings that carry the details of semantic information in the image need to be learned. To that end, we decompose the task of S2IG into two stages, i.e., a speech semantic embedding stage and an image generation stage. Specifically, the proposed speech-to-image generation model via adversarial learning (which we refer to as S2IGAN) consists of a Speech Embedding Network (SEN), which is trained to obtain speech embeddings by modeling and co-embedding speech and images together, and a novel Relation-supervised Densely-stacked Generative Model (RDG), which takes random noise and the speech embedding embedded by SEN as input to synthesize photo-realistic images in a multi-step (coarse-to-fine) way.

In this paper, we present our attempt to generate images directly from the speech signal bypassing text. This task requires specific training material consisting of speech and image pairs. Unfortunately, no such database, with the right amount of data, exists for an unwritten language. The results for our proof-of-concept are consequently presented on two databases with English descriptions, i.e., CUB [14] and Oxford-102 [15]. The benefit of using English as our working language is that we can compare our S2IG results to T2IG results in the literature. Our results are also compared to those of [12].

2. Approach

Given a speech description, our goal is to generate an image that is semantically aligned with the input speech. To this end, S2IGAN consists of two modules, i.e., SEN to create the speech embeddings and RDG to synthesize the images using these speech embeddings.

2.1. Datasets

CUB [14] and Oxford-102 [15] are two commonly-used datasets in the field of T2IG [4, 5], and were also adopted in the most recent S2IG work [12]. CUB is a fine-grained bird dataset that contains 11,788 bird images belonging to 200 categories and Oxford-102 is a fine-grained flower dataset contains 8,189 images of flowers from 102 different categories. Each image in both datasets has 10 text descriptions collected by [16]. Since there are no speech descriptions available for both datasets, we generated speech from the text descriptions using tacotron2 [17] which is a text-to-speech system[1].

2.2. Speech Embedding Network (SEN)

Given an image-speech pair, SEN tries to find a common space for both modalities, so that we can minimize the modality gap

---

*Corresponding author.
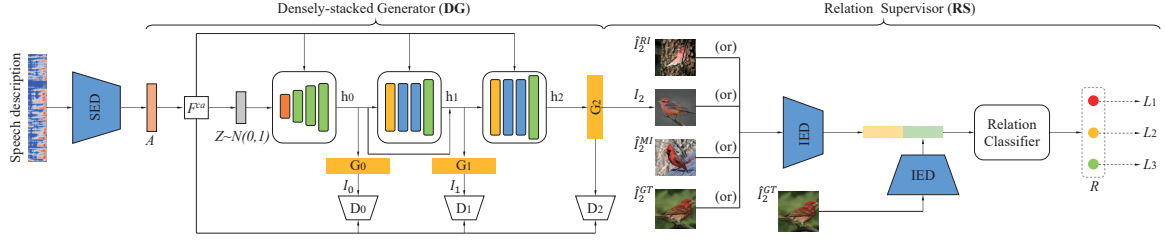
[1]https://github.com/NVIDIA/tacotron2

Figure 1: *Framework of the relation-supervised densely-stacked generative model (RDG). $\hat{I}_2^{RI}$ represents a real image from the same class as the ground-truth image ($\hat{I}_2^{GT}$), I_2 represents a fake image synthesized by the framework. $\hat{I}_2^{MI}$ represents a real image from a different class as $\hat{I}_2^{GT}$. L_i indicates labels for three types of relations. SED and IED are pre-trained in SEN.*

and obtain visually-grounded speech embeddings. SEN is a dual encoder framework, including an image encoder and a speech encoder, which is similar to the model structure in [18].

The image encoder (IED) adopts the Inception-v3 [19] pre-trained on ImageNet [20] to extract visual features. On top of it, a single linear layer is employed to convert the visual feature to a common space of visual and speech embeddings. As a result, we obtain an image embedding V from IED.

The speech encoder (SED) employs a structure similar to that of [18]. Specifically, it consists of a two-layer 1-D convolution block, two-layer bi-directional gated recurrent units (GRU) [21] and a self-attention layer. Finally, speech is represented by a speech embedding A in the common space. The input of the SED are log Mel filter bank spectrograms, which are obtained from the speech signal using 40 Mel-spaced filter banks with 25 ms Hamming window and 10 ms shift.

More details of SEN, including the framework illustration, can be found on the project website[2].

2.2.1. Objective Function

To minimize the distance between a matched pair of an image feature and speech feature while maintaining discrimination of the features compared to features from other bird (CUB) or flower (Oxford-102) classes, matching loss and distinctive loss are proposed.

Matching loss is designed to minimize the distance of a matched image-speech pair. Specifically, in a batch of image-speech embedding pairs $\{(V_i, A_i)\}_i^n$, where n is the batch size, the probability for the speech embedding A_i matching with the image embedding V_i is

$$P(V_i|A_i) = \frac{\exp(\beta S(A_i, V_i))}{\sum_{j=1}^n M_{i,j}\exp(\beta S(A_i, V_j))}, \quad (1)$$

where β is a smoothing factor, set as 10 following [6]. $S(A_i, V_i)$ is a cosine similarity score of A_i and V_i. As in a mini-batch, we only treat (V_i, A_i) as a positive matched pair, therefore we use a mask $M_{i,j} \in \mathbb{R}^{n \times n}$ to deactivate the effect of pairs from the same class. Specifically,

$$M_{ij} = \begin{cases} 0, & \text{if } A_i \text{ matches } V_j \text{ \& } i \neq j, \\ 1, & \text{otherwise}, \end{cases} \quad (2)$$

where A_i matches V_j means they come from the same class. The loss function is then defined as the negative log probability of

$P(V_i|A_i)$:

$$\mathcal{L}_{A-V} = -\sum_{i=1}^n \log P(V_i|A_i). \quad (3)$$

Reversely, we also calculate $\mathcal{L}_{V-A}$ for V_i matching A_i. The matching loss is then calculated as

$$\mathcal{L}_m = \mathcal{L}_{A-V} + \mathcal{L}_{V-A}. \quad (4)$$

Distinctive loss is designed to ensure that the space is optimally discriminative regarding the instance classes. Specifically, both speech and image features in the embedding space are converted to a label space by adding a perception layer, i.e., $\hat{V}_i = f(V_i)$ and $\hat{A}_i = f(A_i)$, where $\hat{V}_i, \hat{A}_i \in \mathbb{R}^N$ and N is the number of classes. The loss function is given by

$$\mathcal{L}_d = -\sum_{i=1}^n \left(\log \hat{P}\left(C_i|\hat{A}_i\right) + \log \hat{P}\left(C_i|\hat{V}_i\right) \right), \quad (5)$$

where $\hat{P}(C_i|\hat{A}_i)$ and $\hat{P}(C_i|\hat{V}_i)$ represent softmax probabilities for $\hat{A}_i$ and $\hat{V}_i$ belonging to their corresponding class C_i.

Total loss for training SEN is finally given by

$$\mathcal{L}_{SEN} = \mathcal{L}_m + \mathcal{L}_d. \quad (6)$$

2.3. Relation-supervised Densely-stacked Generative Model (RDG)

After learning the visually-grounded and class-discriminative speech embeddings, we employ RDG to generate images conditioned on these speech embeddings. RDG consists of two sub-modules, which are a Densely-stacked Generator (DG) and a Relation Supervisor (RS), see Figure 1.

2.3.1. Densely-stacked Generator (DG)

RDG uses the multi-step generation structure [5, 7, 8] because of its previously shown performance. This structure generates images from small scale (low-resolution) to large scale (high-resolution) step by step. Specifically, in our model, 64×64, 128×128, and 256×256 pixel images were generated in multi-steps. To fully exploit the information of the hidden feature (h_i) of each step, we design a densely-stacked generator. With the speech embedding A as input, the generated image in each stacked generator can be expressed as follows:

$$\begin{aligned} h_0 &= F_0\left(z, F^{ca}(A)\right), \\ h_i &= F_i\left(h_0, \ldots, h_{i-1}, F^{ca}(A)\right), i \in \{1, 2\}, \quad (7) \\ I_i &= G_i\left(h_i\right), i \in \{0, 1, 2\}, \end{aligned}$$

where z is a noise vector sampled from a normal distribution. F^{ca} represents Conditioning Augmentation [22, 5] that augments the speech features thus produces more image-speech pairs. It is a popular and useful strategy which is used in most recent text-to-speech generation tasks [9, 6, 7]. h_i is the hidden feature from the non-linear transformation F_i. h_i is fed to the generator G_i to obtain image I_i.

2.3.2. Relation Supervisor (RS)

To ensure that the generator produces high-quality images that are semantically aligned with the spoken description, we propose a relation supervisor to provide a strong relation constraint to the generation process. Specifically, we form an image set for each generated image I_i, i.e., $\{I_i, \hat{I}_i^{GT}, \hat{I}_i^{RI}, \hat{I}_i^{MI}\}$ indicating the generated fake image, the ground-truth image, a real image from the same class as I_i, and a real image from a different randomly-sampled class, respectively. We then define three types of relation classes: 1) a positive relation L_1, between $\hat{I}_i^{GT}$ and $\hat{I}_i^{RI}$; 2) a negative relation L_2, between $\hat{I}_i^{GT}$ and $\hat{I}_i^{MI}$; 3) an undesired relation L_3, between $\hat{I}_i^{GT}$ and $\hat{I}_i^{GT}$. A relation classifier is trained to classify these three relations. We expect the relation between I_i and $\hat{I}_i^{GT}$ to be close to the positive relation L_1, because I_i should semantically align with its corresponding $\hat{I}_i^{GT}$, however, it should not be identical to $\hat{I}_i^{GT}$ to ensure the diversity of the generated results. Therefore, the loss function for training the RS is defined as:

$$L_{RS} = -\sum_{j=1}^{3} \log \hat{P}\left(L_j \,|\, R_j\right) - \log \hat{P}\left(L_1 \,|\, R_{GT-FI}\right), \tag{8}$$

where R_j is a relation vector produced by RS with the input of a pair of images with relation L_j, e.g., $R_1 = RS\left(\hat{I}^{GT}, \hat{I}^{RI}\right)$. R_{GT-FI} is the vector of relation between $\hat{I}_i^{GT}$ and I_i. Note that we apply RS only to the last generated image, i.e., $i = 2$, for computational efficiency.

2.3.3. Objective Function

The final objective function of RDG is defined as:

$$\mathcal{L}_G = \sum_{i=0}^{2} \mathcal{L}_{G_i} + \mathcal{L}_{RS}, \tag{9}$$

where the loss function for the i-th generator G_i is defined as:

$$\begin{aligned}
\mathcal{L}_{G_i} = &- \mathbb{E}_{I_i \sim p_{G_i}}\left[\log D_i\left(I_i\right)\right] + \\
&- \mathbb{E}_{I_i \sim p_{G_i}}\left[\log\left(D_i\left(I_i, F^{ca}(A)\right)\right)\right].
\end{aligned} \tag{10}$$

The loss function for the corresponding discriminator D of RDG is given by:

$$\mathcal{L}_D = \sum_{i=0}^{2} \mathcal{L}_{D_i}, \tag{11}$$

where the loss function for the i-th discriminator D_i is given by:

$$\begin{aligned}
\mathcal{L}_{D_i} = &- \mathbb{E}_{\hat{I}_i \sim p_{data_i}}\left[\log D_i\left(\hat{I}_i\right)\right] + \\
&- \mathbb{E}_{I_i \sim p_{G_i}}\left[\log\left(1 - D_i\left(I_i\right)\right)\right] + \\
&- \mathbb{E}_{\hat{I}_i \sim p_{data_i}}\left[\log D_i\left(\hat{I}_i, F^{ca}(A)\right)\right] + \\
&- \mathbb{E}_{I_i \sim p_{G_i}}\left[\log\left(1 - D_i\left(I_i, F^{ca}(A)\right)\right)\right].
\end{aligned} \tag{12}$$

Here, the first two items are unconditional loss that discriminate the fake and real images, and the last two items are conditional

Table 1: *Performance of S2IGAN compared to other methods.* † *means that the results are taken from the original paper. The best performance is shown in bold.*

Evaluation Metric	Input	CUB (Bird)			Oxford-102 (Flower)		
		mAP	FID	IS	mAP	FID	IS
StackGAN-v2	text	7.01	20.94	4.02±0.03	9.88	50.38	3.35±0.07
MirrorGAN†	text	—	—	4.56±0.05	—	—	–
SEGAN†	text	—	—	**4.67±0.04**	—	—	–
[12]†	speech	—	18.37	4.09±0.04	—	54.76	3.23±0.05
StackGAN-v2	speech	8.09	18.94	4.14±0.04	12.18	54.33	**3.69±0.08**
S2IGAN	speech	**9.04**	**14.50**	4.29±0.04	**13.40**	**48.64**	3.55±0.04

loss discriminating whether the image and the speech description match or not. The I_i is from the model distribution G_i at the ith scale, and $\hat{I}_i$ is from the real image distribution p_{data_i} at the same scale. The generators and discriminators were trained alternately.

2.4. Evaluation Metrics

We use two metrics to evaluate the performance of our SI2GAN model. To evaluate **diversity and quality** of the generated images, we used two popular evaluation metrics for quantitative evaluation of generative models as that in [5]: Inception score (IS) [23] and fréchet inception distance (FID) [24], where, a higher IS means more diversity and a lower FID means a smaller distance between the generated and real image distributions, which indicates better generated images.

The **visual-semantic consistency** between the generated images and their speech descriptions is evaluated through a content-based image retrieval experiment between the real images and the generated images, and evaluated using mAP scores. Specifically, we randomly chose two real images from each class of the test set, resulting in a query pool. Then we used these query images to retrieve generated fake images that belong to their corresponding classes. We used the pre-trained Inception-v3 to extract features of all images. Higher mAP indicates a closer feature distance between fake images and their ground truth images, which indirectly shows a higher semantic consistency between generated images and their corresponding speech descriptions.

3. Results

3.1. Objective Results

We compare our results with several state-of-the-art T2IG methods, including StackGAN-v2 [5], MirrorGAN [7] and SEGAN [9]. StackGAN-v2 is a strong baseline for the T2IG task and provides the effective stacked structure for the following methods. Both MirrorGAN and SEGAN are based on the stacked structure. MirrorGAN utilizes word-level [6] and sentence-level attention mechanisms, and a "text-to-image-to-text" structure for T2IG, and SEGAN also uses word-level attention with extra proposed attention regularization and a siamese structure. In order to allow for a direct comparison on the S2IG task to StackGAN-v2, we reimplemented StackGAN-v2 and replaced the text embedding with our speech embedding. Moreover, we compare our results to the recently released speech-based model by [12].

The results are shown in Table 1. First, our method outperformed [12] on all evaluation metrics and datasets. Compared with the StackGAN-v2 that took our speech embedding as input, our S2IGAN also achieved higher mAP and lower FID on both datasets. These results indicate that our method is effective

in generating high-quality and semantically consistent images on the basis of spoken descriptions. The comparison of our S2IGAN with three state-of-the-art T2IG methods show that the S2IGAN method is competitive, and thus establishes a solid new baseline for the S2IG task.

Speech input is generally considered to be more difficult to deal with than text because of its high variability, its long duration, and the lack of pauses between words. Therefore, S2IG is more challenging than T2IG. However, the comparison of the performances of StackGAN-v2 on the S2IG and T2IG tasks shows that StackGAN-v2 generated better images using speech embeddings learned by our SEN. Moreover, the StackGAN-v2 based on our learned speech embeddings outperforms [12] on almost all evaluation metrics and datasets, except for the slightly higher FID on CUB dataset. Note that [12] takes the native StackGAN-v2 as the generator, which means that the only difference between [12] and the speech-based StackGAN-v2 in Table 1 is the speech embedding method. These results confirm that our learned speech embeddings are competitive compared to text input and the speech embeddings in [12], showing the effectiveness of our SEN module.

3.1.1. Subjective Results

The subjective visual results are shown in Figure 2. As can be seen, the images synthesized by our S2IGAN (d) are photo-realistic and convincing. By comparing the images generated by (d) S2IGAN and (c) StackGAN-v2 conditioned on speech embeddings, we can see that the images generated by S2IGAN are clearer and sharper, showing the effectiveness of the proposed S2IGAN on synthesizing visually high-quality images. The comparison of StackGAN-v2 conditioned on (b) text and (c) speech features embedded by the proposed SEN shows that our learned speech embeddings are competitive compared with the text features embedded by StackGAN-v2, showing the effectiveness of SEN. More results are shown on the project website[2].

To further illustrate S2IGAN's ability to catch subtle semantic differences in the speech descriptions, we generated images conditioned on speech descriptions in which color keywords were changed. As Figure 3 shows, the visual semantics of the generated birds, specifically, the colors of the belly and the wings, are consistent with the corresponding semantic information in the spoken descriptions. These visualization results indicate that SEN successfully learned the semantic information in the speech signal, and that our RDG is capable of capturing these semantics and generating discriminative images that are semantically aligned with the input speech.

3.2. Component analysis

An extensive ablation study investigated the effectiveness of key components of SI2GAN. Specifically, the effects of the densely-stacked structure of DG, RS, and SEN were investigated by removing each of these components respectively. Removing any component resulted in a clear decrease of the generation performance, showing the effectiveness of each component. Details can be found on the project website[2].

4. Discussion and Conclusion

This paper introduces a novel speech-to-image generation (S2IG) task and we developed a novel generative model, called S2IGAN, which tackles S2IG in two steps. First, semantically discriminative speech embeddings are learned by a speech embedding network. Second, high-quality images are generated on the basis
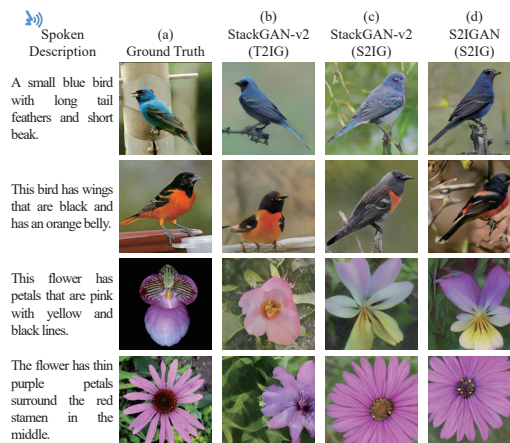

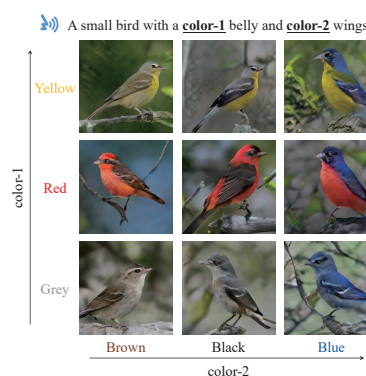Figure 2: *Examples of images generated by different methods.*

Figure 3: *Generated examples by S2IGAN. The generated images are based on speech descriptions with different color keywords.*

of the speech embeddings. The results of extensive experiments show that our S2IGAN has state-of-the-art performance, and that the learned speech embeddings capture the semantic information in the speech signal.

The current work is based on synthesized speech, which makes the current S2IG baseline an upper-bound baseline. The future work will focus on several directions. First, we will investigate this task with natural speech instead of synthesized speech. Second, it will be highly interesting to test the proposed methodology on a true unwritten language rather than the well-resourced English language. Third, we will further improve our methods in terms of efficiency and accuracy, for example, by making end-to-end training more effective and efficient and by applying attention mechanisms to our generator to further improve the quality of the generated images. An interesting avenue for future research would be to automatically discover speech units based on corresponding visual information from the speech signal [25] to segment the speech signal. This would allow us to use segment- and word-level attention mechanisms, which have shown to lead to improved performance on the text-to-image generation task [6], to improve the performance of speech-to-image generation.

5. Acknowledgements

6. References

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[2] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[3] Y. Balaji, M. R. Min, B. Bai, R. Chellappa, and H. P. Graf, "Conditional gan with discriminative filter generation for text-to-video synthesis," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. AAAI Press, 2019, pp. 1995–2001.

[4] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," *arXiv preprint arXiv:1605.05396*, 2016.

[5] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "Stackgan++: Realistic image synthesis with stacked generative adversarial networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 1947–1962, 2018.

[6] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, "Attngan: Fine-grained text to image generation with attentional generative adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1316–1324.

[7] T. Qiao, J. Zhang, D. Xu, and D. Tao, "Mirrorgan: Learning text-to-image generation by redescription," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1505–1514.

[8] G. Yin, B. Liu, L. Sheng, N. Yu, X. Wang, and J. Shao, "Semantics disentangling for text-to-image generation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2327–2336.

[9] H. Tan, X. Liu, X. Li, Y. Zhang, and B. Yin, "Semantics-enhanced adversarial nets for text-to-image synthesis," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 10 501–10 510.

[10] 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 *CVPR*, 2019.

[11] Y. Wen, B. Raj, and R. Singh, "Face reconstruction from voice using generative adversarial networks," in *NeurIPS*, 2019.

[12] J. Li, X. Zhang, C. Jia, J. Xu, L. Zhang, Y. Wang, S. Ma, and W. Gao, "Direct speech-to-image translation," *arXiv preprint arXiv:2004.03413*, 2020.

[13] M. P. Lewis, G. F. Simons, and C. Fennig, "Ethnologue: Languages of the world [eighteenth," *Dallas, Texas: SIL International*, 2015.

[14] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," 2011.

[15] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*. IEEE, 2008, pp. 722–729.

[16] S. Reed, Z. Akata, H. Lee, and B. Schiele, "Learning deep representations of fine-grained visual descriptions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 49–58.

[17] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[18] D. Merkx, S. L. Frank, and M. Ernestus, "Language learning using speech to image retrieval," *arXiv preprint arXiv:1909.03795*, 2019.

[19] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[20] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[21] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[22] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5907–5915.

[23] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Advances in neural information processing systems*, 2016, pp. 2234–2242.

[24] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in neural information processing systems*, 2017, pp. 6626–6637.

[25] D. Harwath and J. Glass, "Towards visually grounded sub-word speech unit discovery," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3017–3021.