TU Delft

Delft University of Technology

Generating Images from Spoken Descriptions

Wang, Xinsheng; Qiao, Tingting; Zhu, Jihua; Hanjalic, Alan; Scharenborg, Odette

DOI
10.1109/TASLP.2021.3053391

Publication date
2021

Document Version
Accepted author manuscript

Published in
IEEE/ACM Transactions on Audio Speech and Language Processing

Citation (APA)
Wang, X., Qiao, T., Zhu, J., Hanjalic, A., & Scharenborg, O. (2021). Generating Images from Spoken Descriptions. *IEEE/ACM Transactions on Audio Speech and Language Processing*, *29*, 850-865. Article 9333641. https://doi.org/10.1109/TASLP.2021.3053391

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

Generating Images from Spoken Descriptions

Xinsheng Wang, Tingting Qiao, Jihua Zhu*, *Member, IEEE* Alan Hanjalic, *Fellow, IEEE,* Odette Scharenborg, *Senior Member, IEEE*

***Abstract*—Text-based technologies, such as text translation from one language to another, and image captioning, are gaining popularity. However, approximately half of the world's languages are estimated to be lacking a commonly used written form. Consequently, these languages cannot benefit from text-based technologies. This paper presents 1) a new speech technology task, i.e., a speech-to-image generation (S2IG) framework which translates speech descriptions to photo-realistic images 2) without using any text information, thus allowing unwritten languages to potentially benefit from this technology. The proposed speech-to-image framework, referred to as S2IGAN, consists of a speech embedding network and a relation-supervised densely-stacked generative model. The speech embedding network learns speech embeddings with the supervision of corresponding visual information from images. The relation-supervised densely-stacked generative model synthesizes images, conditioned on the speech embeddings produced by the speech embedding network, that are semantically consistent with the corresponding spoken descriptions. Extensive experiments are conducted on four public benchmark databases: two databases that are commonly used in text-to-image generation tasks, i.e., CUB-200 and Oxford-102 for which we created synthesized speech descriptions, and two databases with natural speech descriptions which are often used in the field of cross-modal learning of speech and images, i.e., Flickr8k and Places. Results on these databases 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.**

I. Introduction

THE task of conditional image generation is to synthesize images on the basis of given information, e.g., labels or descriptions. The development of deep learning and Generative Adversarial Networks (GANs) [1], [2] brought a revolution to various computer vision tasks based on conditional image generation methods, such as image-to-image translation [3], [4] and facial editing [5], [6]. Recently, many efforts have been carried out on the task of image generation conditioned on natural language [7], [8], [9], which allows users to synthesize images by describing what they want to generate. This task has potential applications in image retrieval, automated image editing, and forensics. Typically, natural language-to-image generation systems use text descriptions as their input. This is referred to as Text-to-Image Generation (T2IG). We present a novel natural language-to-image generation task that is not only based on a spoken description of what needs to be generated but also bypasses the need for text: Speech-to-Image Generation (S2IG). Fig. 1 illustrates this new task. This task is similar to the independently and simultaneously proposed task of speech-to-image translation task [10].

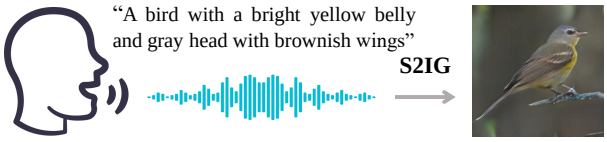


Fig. 1: Illustration of image generation conditioned on the speech description. The image is generated by the proposed S2IGAN. Text is shown only for the convenience of showing the speech content.

Our motivation for bypassing text is the fact that an estimated half of the 7,000 world's languages lack a commonly used written form [11] (these are 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 one's native language. This right is obviously not achieved for unwritten languages. In the case of natural language-to-image generation, in order to bypass text, this means that a system needs to be developed that maps spoken descriptions to images without the need for an intermediate textual representation. Moreover, even though existing knowledge and methodology make "speech2text2image" transfer possible, directly mapping speech to images might be more efficient and straightforward.

Synthesizing plausible images based on spoken descriptions is a challenging task as the speech signal is a long, continuous audio signal without spaces between words, i.e., unlike written text in alphabetical languages such as English. The lack of word boundaries makes it harder to learn speech embeddings that capture the details of the semantic information in a stretch of speech describing an image, e.g., the word-level attention mechanism in the DAMSM module of AttnGAN [9], a useful strategy that implicitly connects words to the object regions in the images for the text-to-image task, cannot be used in a speech-to-image system. Moreover, the long sequence of the speech signal requires a speech encoder that is different from a text encoder. To tackle this challenge, we decompose the task

Xinsheng Wang is with the School of Software Engineering, Xi'an Jiaotong University, Xi'an 710049, China, and also with the Multimedia Computing Group, Delft University of Technology, 2628 CD Delft, The Netherlands. He is supported by the China Scholarship Council (CSC). Email: wangxinsheng@stu.xjtu.edu.cn

Tingting Qiao is with the College of Computer Science and Technology, Zhejiang University, Hangzhou 310007, China, and also with the Multimedia Computing Group, Delft University of Technology, 2628 CD Delft, The Netherlands. Email: qiaott@zju.edu.cn

Jihua Zhu is with the School of Software Engineering, Xi'an Jiaotong University, Xi'an 710049, China. Email: zhujh@xjtu.edu.cn

Alan Hanjalic and Odette Scharenborg are with the Multimedia Computing Group, Delft University of Technology, 2628 CD Delft, The Netherlands. Email: A.Hanjalic@tudelft.nl, O.E.Scharenborg@tudelft.nl.

* Corresponding author.

of S2IG into two stages, i.e., a speech semantic embedding stage and an image generation stage. Specifically, we propose a speech-to-image generation system based on adversarial learning. This system consists of a Speech Embedding Network (SEN) trained to obtain speech embeddings by co-embedding speech and images into a common embedding space, and a novel Relation-supervised Densely-stacked Generative Model (RDG), which takes random noise and the speech embedding as input to synthesize photo-realistic images in multiple steps. We refer to this model as S2IGAN (Speech-to-Image Generative Adversarial Network).

In order to generate images directly from the speech signal bypassing the need for text, specific training material which consists of speech and image pairs is required. Unfortunately, no such database, with the right amount of data, exists for any unwritten language, as unwritten languages are not only low-resourced but also under-researched. We therefore had to revert to testing our proof-of-concept S2IGAN model on a well-resourced, well-researched language for which appropriate databases do exist, i.e., we used four English benchmark databases: two databases that are commonly adopted in the T2IG task, i.e., CUB-200 [12] and Oxford-102 [13], for which we use synthesized spoken descriptions, and two databases with real speech descriptions that are often used in cross-modal speech-to-image retrieval tasks, i.e., Places [14] with speech descriptions collected by [15] and Flickr8k [16] with speech descriptions collected by [17]. The benefit of using English as our working language is that we can compare our S2IG results to T2IG results, which are primarily based on the English language, in the literature.

Preliminary work was presented in [18] in which we tested our model on the synthesized speech databases. Here, we present a complete description of the model and extensive testing of our model including on real, natural speech. Moreover, we present a detailed ablation study. The contributions of this work are as follows:

- We propose S2IGAN which consists of two effective modules to tackle Speech-to-image generation, i.e., a Speech Embedding Network which learns discriminative speech embeddings and a Relation-supervised Densely-stacked Generative Model generating high-quality images that are semantically consistent with the input spoken descriptions.
- Thorough experiments on four benchmark databases demonstrate that our SI2GAN is a solid baseline for the S2IG task.
- An extensive ablation study shows that the proposed speech embedding model, the novel densely-stacked generator structure, and the proposed relation supervisor in the generative model are effective, indicating the S2IGAN is well-designed.

The rest of this paper is organized as follows: Section II reviews related work. Section III introduces the databases adopted in this work and describes the proposed approach in detail. Section IV presents extensive experimental results and evaluation. The ablation study is also presented here. Section V discusses the performance and limitations of the proposed method. Finally, the paper concludes in Section VI. Examples of the synthesized images and their corresponding spoken descriptions as well as the synthesized speech caption databases and the source code of S2IGAN can be found on the project page[1].

II. Related Works

The development of GANs [1], [19] led to the promotion of various conditional image generation tasks, of which the text-to-image generation [7], [8], [9] is most related to our S2IG task. This section reviews related work on natural language-to-image generation. Additionally, we review related research on visually-grounded speech embedding learning, which is an important part of our S2IGAN. Please note that although our work is related to existing sound-to-image/video generation tasks, e.g., voice-to-face inference [20], [21], [22], [23], audio-driven talking face generation [22], [24], [25], and body movement generation based on music [26], we will not review these here because these sound-to-image tasks do not take into account the semantics in the audio as does our task. We refer the interested reader to a review of these related studies which can be found in [27].

A. Natural language-to-image generation

The GAN-based text-to-image model was first introduced by [28], in which the text descriptions were encoded as vector representations which were used as input of a conditional GAN [2]. Since then, many strategies have been proposed to improve the performance of GAN-based models on the T2IG task [7], [8], [9], [29].

The stacked structure proposed in [7], [30] showed good performance in synthesizing high-resolution images and has been one of the most popular strategies in most recent T2IG frameworks [8], [29], [31]. In the stacked structure, images are generated using multiple steps from low-resolution to high-resolution. T2IG performance using the stacked structure [7] as the basic structure is further improved [8], [31] by adding a word-level attention mechanism [9]. Recently, more complex and effective structures for T2IG were proposed [8], [29], [32], [33]. For instance, Qiao et al. [8] proposed a model structure similar to the encoder-decoder, called MirrorGAN that tried to recover the text descriptions based on the synthesized images to guarantee semantic consistency between the text description and generated image. In [29], a Siamese structure was introduced into the T2IG, and it showed that the Siamese structure trained with contrastive loss can bring clear improvements for T2IG. To generate more photo-realistic complex images, the Obj-GANs [33] decomposed the T2IG into two stages: the first stage generates a semantic layout with bounding boxes, object shapes, and class labels, and the second stage synthesizes photo-realistic images.

However, all the natural language-to-image generation research mentioned above is based on written language, i.e., text descriptions. The task most related to our work is presented in [10], in which the authors adopted the teacher-student structure

[1]https://xinshengwang.github.io/projects/S2IGAN/

to learn speech embeddings and used StackGAN-v2 [7] as the generator to generate images from the content of the speech signal. In our work, we will compare our method with this recently proposed method.

Encouraged by the good performance of the stacked structure [7], in our S2IGAN, we use a similar stacked generator structure to synthesize images from coarse to fine. Compared with the original structure [7], our model is densely-stacked. Moreover, a relation supervisor to ensure that the generated images are semantically aligned with the spoken descriptions is proposed (see Section III for details).

B. Visually-grounded speech embedding learning

Learning an effective speech embedding that carries detailed semantic information is crucial for speech-conditioned image generation. However, in the speech community, learning speech embeddings typically uses text transcriptions as supervisory information [34], [35]. However, as explained, the S2IG framework does not use textual information. Instead, the supervisory information consists of the corresponding images of the speech descriptions. Inspired by human infants' ability to learn spoken language by listening and paying attention to the concurrent speech and visual scenes, several recent methods [15], [36], [37], [38], [39], [40] have been proposed to learn speech models grounded by visual information.

Harwath et al. [36] demonstrated that semantic information can be learned from speech-image pairs without the need for text. They embedded the global features of speech and images in a common embedding space. The semantic correspondences between the images and the spoken descriptions learned by the model were then evaluated in a cross-modal speech-based image retrieval task. Adding a self-attention mechanism in the speech encoder to get speech embeddings was found to improve performance in a similar cross-modal retrieval task [37]. Recently, Ilharco et al. [38] trained both an image encoder and a speech encoder from scratch based on a large-scale database and showed the best performance on the speech-image cross modal retrieval task to date.

In this work, we also adopt the visually-grounded method [36], [37], [38] to learn speech embeddings. Specifically, a dual encoder framework, which consists of a speech encoder and an image encoder, is proposed which allows for the speech encoder and image encoder to be trained in a self-supervised way with speech caption-image pairs as training data. With speech as input, the trained speech encoder is used to extract the speech embeddings. Different from most of the previously described visually-grounded speech embedding learning methods in which the traditional triplet loss function was adopted to train the model [15], [36], [37], [39], [40], in this work we propose a more effective matching loss which is similar to the masked margin softmax loss [38] that can make full use of negative samples within a mini-batch. Furthermore, previous studies only focused on databases with scene images, primarily Flickr8k [16], while no fine-grained image databases, e.g., CUB-200 [12], in which images from different categories share similar semantics, have been considered. In this study, in order to learn speech embeddings of fine-grained image captions, an additional distinctive loss is designed.

TABLE I: Statistics of CUB-200 [12] and Oxford-102 [13]

Database	Statistics	Training set	Test set	Total
CUB-200 (Bird) [12]	class	150	50	200
	image	8855	2933	11788
Oxford-102 (Flower) [13]	class	82	20	102
	image	7034	1155	8189

III. Speech-to-image Generation with S2IGAN

Given a spoken description of an image, our goal is to generate an image that is semantically aligned with the content of the input spoken utterance. To this end, S2IGAN consists of two modules, i.e., an SEN (see Section III-B) to create the speech embeddings and an RDG (see Section III-C) to synthesize the images using these speech embeddings. To train this model, databases with image-speech pairs are required (see Section III-A).

A. Databases

1) Synthesised speech caption-image pairs: CUB-200 [12] and Oxford-102 [13] are two commonly-used databases in T2IG tasks. Both databases are fine-grained image databases with corresponding text captions for each image. As shown in Table I, CUB-200 is a bird database that contains 11,788 bird images from 200 classes (bird species), while Oxford-102 is a flower database that contains 8,189 flower images from 102 classes (flower species). Each image in both the CUB-200 and the Oxford-102 databases has 10 text descriptions. Following [28], [41], we split the databases into class-disjoint training and test sets. Specifically, the CUB-200 training set consists of 150 classes, and the test set of 50 classes non-overlapping with the training classes. Similarly, the Oxford-102 database has 82 training classes and 20 test classes. This class-disjoint split method requires the model to have the ability to infer unseen classes based on the knowledge learned from seen classes. This can be seen as a kind of zero-shot learning, and both databases with their disjoint classes for training and testing are commonly used in zero-shot learning of image recognition [42], [43], [44].

CUB-200 and Oxford-102 do not contain spoken captions. Spoken captions for the images are obtained using text-to-speech (TTS) technology. Specifically, we synthesize the spoken captions from the text captions using Tacotron 2 [45] with WaveGlow [46] which transforms mel-spectrograms to high quality speech. This TTS system is pre-trained on LJSpeech [47] which consists of 13,100 short audio clips of a single speaker reading passages from 7 non-fiction books. The pre-trained model can be found on Github[2], and the synthesized speech caption databases can be directly downloaded from the project page[1].

[2]https://github.com/NVIDIA/tacotron2



Fig. 2: Image examples from Flickr8k.

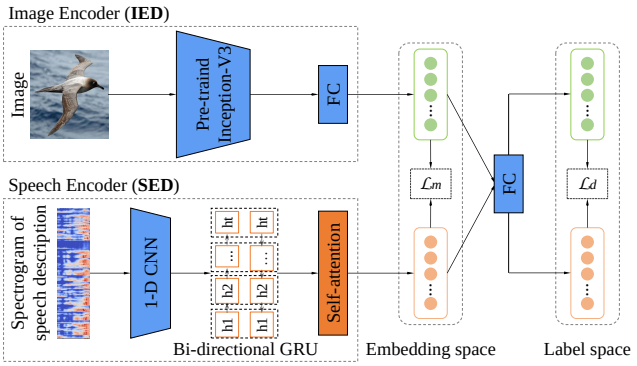


Fig. 3: Framework of the speech embedding network (SEN).

2) Real speech caption-image pairs: Two databases containing real speech are used: the Places Audio Caption database [15], [36] and the Flickr8k database [16]. The Places Audio Caption database is a spoken caption database collected via Amazon Mechanical Turk. It is a real speech database that contains spoken captions of images from the Places [14] image database. The Places Audio Caption database consists of 403,385 image-caption pairs belonging to 205 scene categories. In this database, each image has one corresponding spoken caption. The work by [10], which is most closely related to our work, uses the Places database. In order to make a direct comparison with [10] on the task of S2IG, we use the same subset of the Places database as used in [10]. Specifically, this subset (referred to as the Places-subset hereafter) consists of 7 scene categories: bedroom, dinette, dining room, home office, hotel room, kitchenette, and living room, with a total of 13,803 image-spoken caption pairs for training and 2,870 image-spoken caption pairs for testing.

Flickr8k [16] is another scene image database. It consists of 8,000 images that depict a variety of scenes and situations. Each image is paired with five different text captions describing the image. The spoken captions were collected by [17] from Amazon Mechanical Turk workers who were asked to pronounce the original written captions. The data is split according to [48], with a training set of 6,000 images and 1,000 images both for the development and test set.

Compared to the Places-subset, Flickr8k is a more challenging database for the S2IG task. In the Places-subset, each scene has around 2,000 image-captions pairs, while Flickr8k does not have defined scene categories, and some images may contain unique scenes in the database. Fig. 2 shows several image samples from Flickr8k showing the wide range of scenes with limited relations between the images. This great diversity in scenes makes it hard to generate an image given a new scene description. However, we would like to test our method on this challenging database, in order to investigate what can be learned from such a diverse scene image database.

B. The Speech Embedding Network (SEN)

Given an image-speech pair, the SEN tries to find a common space for the speech and image modalities, so that we can minimize the modality gap and obtain visually-grounded speech embeddings. As shown in Fig. 3, the SEN is a dual encoder framework, consisting of an image encoder and a speech encoder. These two encoders embed the input images and speech into an embedding space, such that image representations in this space can be used as supervision information to train the speech encoder, and vice versa.

The image encoder (IED) adopts the Inception-v3 [49] pre-trained on ImageNet [50] to extract visual features. On top of it, a linear layer is employed to convert the visual feature to a common space of visual and speech embeddings. The input size of this linear layer is 2048 and the output size in the common embedding space is 1024. As a result, we obtain an image embedding $\boldsymbol{x}_i$ which is a 1024-d vector from the image encoder.

The speech encoder (SED) employs a structure similar to that of [37]. Specifically, it consists of a two-layer 1-D convolution block, two bi-directional gated recurrent units (GRU) [51] and a self-attention layer. The 1-D convolutional block consists of two 1-D convolutional layers with 40 input, 64 hidden and 128 output channels. The size of the hidden layer of the bi-directional GRU was 512 and the size of the output was 1024 by concatenating the bidirectional representations. Finally, a stretch of speech is represented by a speech embedding $\boldsymbol{y}_i$ in the common space.

The input of the speech encoder of the SEN are log Mel filter bank spectrograms, which are obtained from the speech using 40 Mel-spaced filter banks with 25 ms Hamming window and 10 ms shift. The final speech embedding $\boldsymbol{y}_i$ is computed by the self-attention module via a weighted sum over all the hidden states of the GRU. First, the attention vector $\boldsymbol{w}_t$ of each hidden state $\boldsymbol{h}_t$ is calculated by

$$\boldsymbol{w}_t = \text{softmax}\left(\boldsymbol{W}_2 \tanh\left(\boldsymbol{W}_1 \boldsymbol{h}_t + \boldsymbol{b}_1\right) + \boldsymbol{b}_2\right), \tag{1}$$

where $\boldsymbol{W}_i$ and $\boldsymbol{b}_i$ are weights and biases in linear transformers. The unit numbers of $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are 128 and 1024 respectively. The speech embedding $\boldsymbol{y_i}$ is then obtained via the sum over the Hadamard product between all hidden nodes and their corresponding attention vectors:

$$\boldsymbol{y_i} = \sum_t \boldsymbol{w}_t \circ \boldsymbol{h}_t. \tag{2}$$

1) Objective Function: To minimize the distance between a matched image and speech feature pair, a matching loss is defined. Moreover, for the fine-grained databases, maintaining the discrimination between embeddings of different classes

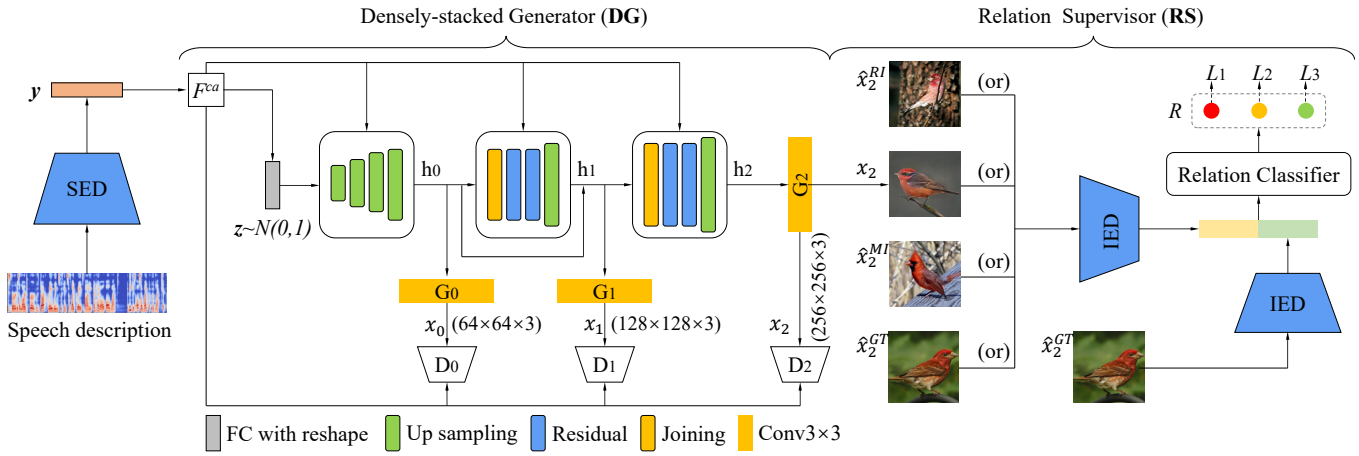


Fig. 4: Framework of the relation-supervised densely-stacked generative model (RDG). $\hat{x}_2^{RI}$ represents a real image from the same class as the ground-truth image ($\hat{x}_2^{GT}$), x_2 represents a fake image synthesized by the framework. $\hat{x}_2^{MI}$ represents a real image from a different class as $\hat{x}_2^{GT}$. L_i indicates labels for three types of relations. SED and IED are pre-trained in SEN.

(e.g., bird species) is important. To this end, a distinctive loss is proposed for the fine-grained databases. Thus, the total loss for the fine-grained databases consists of the matching loss and the distinctive loss, while the total loss for the other two databases consists only of the matching loss.

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

$$P\left(\boldsymbol{x}_i|\boldsymbol{y}_i\right) = \frac{\exp\left(\beta S\left(\boldsymbol{y}_i, \boldsymbol{x}_i\right)\right)}{\sum_{j=1}^{n} \boldsymbol{M}_{i,j} \exp\left(\beta S\left(\boldsymbol{y}_i, \boldsymbol{x}_j\right)\right)}, \tag{3}$$

where β is a smoothing factor, and set to 10 following [9]. $S\left(\boldsymbol{y}_i, \boldsymbol{x}_i\right)$ is a cosine similarity score of $\boldsymbol{y}_i$ and $\boldsymbol{x}_i$. As in a mini-batch, we only treat $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ as a positive matched pair, therefore we use a mask $\boldsymbol{M}_{i,j} \in \mathbb{R}^{n\times n}$ to deactivate the effect of pairs from the same class. Specifically,

$$\boldsymbol{M}_{ij} = \begin{cases} 0, & \text{if } \boldsymbol{y}_i \text{ matches } \boldsymbol{x}_j \ \&\ i \neq j, \\ 1, & \text{otherwise}, \end{cases} \tag{4}$$

where $\boldsymbol{y}_i$ matches $\boldsymbol{x}_j$ means they come from the same class (for the fine-grained databases), or the caption is one of the image's captions (for Place-subset and Flickr8k). The loss function is then defined as the negative log probability of $P\left(\boldsymbol{x}_i|\boldsymbol{y}_i\right)$:

$$\mathcal{L}_{\boldsymbol{y}-\boldsymbol{x}} = -\sum_{i=1}^{n} \log P\left(\boldsymbol{x}_i|\boldsymbol{y}_i\right). \tag{5}$$

Reversely, we can also calculate $\mathcal{L}_{\boldsymbol{x}-\boldsymbol{y}}$ for $\boldsymbol{x}_i$ matching $\boldsymbol{y}_i$. Then the matching loss is calculated as

$$\mathcal{L}_m = \mathcal{L}_{\boldsymbol{y}-\boldsymbol{x}} + \mathcal{L}_{\boldsymbol{x}-\boldsymbol{y}}. \tag{6}$$

Distinctive loss is designed to ensure that the space is optimally discriminative regarding the instance classes, i.e., ensuring that the learned embeddings can distinguish between different classes (e.g., bird species). Here, we take the classification objective function as the distinctive loss. Specifically, with $f(.)$ representing a linear transformer that transfers both speech and image embeddings in the common embedding space to a label space, in which the representation of an image or a stretch of speech represents the probability distribution for each class label, *i.e.*, $\hat{\boldsymbol{x}}_i = f\left(\boldsymbol{x}_i\right)$ and $\hat{\boldsymbol{y}}_i = f\left(\boldsymbol{y}_i\right)$, where $\hat{\boldsymbol{x}}_i, \hat{\boldsymbol{y}}_i \in \mathbb{R}^N$ and N is the number of classes. The loss function is given by

$$\mathcal{L}_d = -\sum_{i=1}^{n} \left(\log \hat{P}\left(c_i|\hat{\boldsymbol{y}}_i\right) + \log \hat{P}\left(c_i|\hat{\boldsymbol{x}}_i\right)\right), \tag{7}$$

where $\hat{P}(c_i|\hat{\boldsymbol{y}}_i)$ and $\hat{P}(c_i|\hat{\boldsymbol{x}}_i)$ represent softmax probabilities for $\hat{\boldsymbol{y}}_i$ and $\hat{\boldsymbol{x}}_i$ belonging to their corresponding class c_i.

Total loss for training SEN for fine-grained databases is finally given by

$$\mathcal{L}_{SEN} = \mathcal{L}_m + \mathcal{L}_d. \tag{8}$$

Since Flickr8k does not contain class information, the distinctive loss cannot be applied. Places-subset does contain different scene classes, however, each image is unique and cannot be described by captions of other images even if they are from the same scene class (e.g., for CUB-200, two bird images that belong to the same bird species can be described by the same caption, because birds from the same class have the same appearance). Thus for Flickr8K and Places-subset, the total loss only contains the matching loss:

$$\mathcal{L}_{SEN} = \mathcal{L}_m. \tag{9}$$

C. The Relation-supervised Densely-stacked Generative Model (RDG)

After training the SEN, speech embeddings can be extracted on the basis of the input speech. Then, we employ the RDG to generate images conditioned on these speech embeddings. The RDG consists of two sub-modules: a Densely-stacked Generator (DG) and a Relation Supervisor (RS). As shown in Fig. 4, the DG consists of three pairs of generators (see G_0, G_1, and G_2) and discriminators (see D_0, D_1, and D_2) for different image scales, and the RS consists of a relation classifier. The SED and IED are the speech encoder and image encoder pre-trained in the SEN, respectively. Compared to

the generative model of StackGAN-v2, the proposed densely-stacked structure and relation supervisor are the two main differences. The motivation and details for these proposed modules are explained in the next two subsections.

1) Densely-stacked Generator (DG): The RDG uses the multi-step generation structure [7], [8], [29] (see also Section II-A). In order to create 256×256 pixel images, this structure generates images from small scale (low-resolution) to large scale (high-resolution) in three steps: i.e., following [7], [8], [29], 64×64, 128×128, and 256×256 pixel images were generated. In contrast to existing T2IG systems [7], [8], [29], in which the stacked-generators are stacked in parallel, we propose a densely-stacked generator structure which allows the information of the hidden feature (h_i) in each step to be conveyed forward more effectively. With the speech embedding $\boldsymbol{y}$ as input (to avoid confusion with the stage subscript i, a speech embedding $\boldsymbol{y}$ no longer has the subscript in the following part), the generated image in each stacked generator can be expressed as follows:

$$\begin{aligned} h_0 &= F_0\left(\boldsymbol{z}, F^{ca}(\boldsymbol{y})\right), \\ h_i &= F_i\left(h_0, \ldots, h_{i-1}, F^{ca}(\boldsymbol{y})\right), i \in \{1,2\}, \\ x_i &= G_i\left(h_i\right), i \in \{0,1,2\}, \end{aligned} \tag{10}$$

where $\boldsymbol{z}$ is a noise vector sampled from a normal distribution. h_i is the hidden feature from the non-linear transformation F_i. h_i is fed to generator G_i to obtain image x_i. Note that h_{i-1} is not only conveyed to F_i, but also conveyed to F_{i+1} via the residual connection that fuses h_{i-1} with h_i (see also the arrow going from h_0 to h_2 in Fig. 4). F^{ca} represents Conditioning Augmentation [30], [7] that augments the speech embeddings. This embedding augmentation method is a popular and useful strategy that is used in the most recent text-to-image generation tasks [9], [31], [8]. Specifically, with the mean μ and the standard deviation σ which are functions of speech embedding $\boldsymbol{y}$, the new sampled speech embedding is formulated as

$$F^{ca}(\boldsymbol{y})=\mu+\sigma\cdot\mathcal{N}(0,I). \tag{11}$$

2) Relation Supervisor (RS): In the MirrorGAN [8], the synthesized images are recovered back to text descriptions to guarantee semantic consistency between the text description and the synthesized image. However, recovering the spoken caption from an image is not an easy feat. A second possible method to guarantee semantic consistency would be to use a matching loss between the descriptions and the corresponding generated images such as DAMSM in AttnGAN [9]. However, DAMSM uses word-level attention which requires word-level segments while no easily identified word-level segments in the speech signal exist. Consequently, the word-level attention mechanism in DAMSM cannot be implemented in our speech-to-image task. When only the matching loss of DAMSM without the attention mechanism is used, only the matching of global features of images and speech is considered for the speech-to-image generation task. Although evaluation metrics, e.g., Inception score (IS) (see the details of evaluation metrics in Section IV-A1), based on the global features of the images could benefit from this loss, it is expected that this global feature constraint would make the generated image unnatural. This hypothesis was tested in our experiments. The details of the DAMSM-based method can be found in Section IV-D2.

Therefore, for our S2IG task, we propose a module, named relation supervisor, to provide a strong relation constraint to the generation process in order to generate high-quality images that are semantically aligned with the spoken descriptions.

For the fine-grained databases CUB-200 and Oxford-102, we form an image set for each generated image x_i, i.e., $\{x_i, \hat{x}_i^{GT}, \hat{x}_i^{RI}, \hat{x}_i^{MI}\}$ indicating the generated fake image, the ground-truth image, a real image from the same class as x_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 $\hat{x}_i^{GT}$ and $\hat{x}_i^{RI}$; 2) a negative relation L_2, between $\hat{x}_i^{GT}$ and $\hat{x}_i^{MI}$; 3) an undesired relation L_3, between $\hat{x}_i^{GT}$ and itself, i.e., $\hat{x}_i^{GT}$. The relation classifier is trained to classify these three relations. We expect the relation between x_i and $\hat{x}_i^{GT}$ to be close to the positive relation L_1, because x_i should semantically align with its corresponding $\hat{x}_i^{GT}$, however, it should not be identical to $\hat{x}_i^{GT}$ to ensure the diversity of the generation. Therefore, the loss function for training the RS is defined as:

$$L_{RS} = -\sum_{j=1}^{3} \log \hat{P}\left(L_j \left| R_j \right.\right) - \log \hat{P}\left(L_1 \left| R_{GT-FI} \right.\right), \tag{12}$$

where R_j is a relation vector produced by RS with the input of a pair of images with relation L_j, *e.g.*, $R_1 = RS(\hat{I}^{GT}, \hat{I}^{RI})$. R_{GT-FI} is the vector of relation between $\hat{x}_i^{GT}$ and x_i.

For the other two databases, i.e., Places-subset and Flickr8k, each speech description only matches with one image, and no $\hat{x}_i^{RI}$ exists in these databases. The positive relation L_1 is defined as the relation between $\hat{x}_i^{GT}$ and itself, i.e., $\hat{x}_i^{GT}$, while no undesired relation is defined. The loss function for these two databases is defined as:

$$L_{RS} = -\sum_{j=1}^{2} \log \hat{P}\left(L_j \left| R_j \right.\right) - \log \hat{P}\left(L_1 \left| R_{GT-FI} \right.\right). \tag{13}$$

Note that we apply the RS only to the last generated pixel image, i.e., $i = 2$, for computational efficiency.

3) Objective Function: The final objective function of the RDG is defined as:

$$\mathcal{L}_G = \sum_{i=0}^{2} \mathcal{L}_{G_i} + \mathcal{L}_{RS}, \tag{14}$$

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

$$\begin{aligned} \mathcal{L}_{G_i} = &-\mathbb{E}_{x_i\sim p_{G_i}}\left[\log D_i\left(x_i\right)\right] + \\ &-\mathbb{E}_{x_i\sim p_{G_i}}\left[\log\left(D_i\left(x_i, F^{ca}(\boldsymbol{y})\right)\right)\right], \end{aligned} \tag{15}$$

where $\mathbb{E}_{x_i\sim p_{G_i}}$ is the expected value over all generated data instances. The first item is the unconditional loss and the second one is the conditional loss. The loss function for the corresponding discriminator D of RDG is given by:

$$\mathcal{L}_D = \sum_{i=0}^{2} \mathcal{L}_{D_i}, \tag{16}$$

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

$$\begin{aligned} \mathcal{L}_{D_i} = &-\mathbb{E}_{\hat{x}_i\sim p_{data_i}}\left[\log D_i(\hat{x}_i)\right] + \\ &-\mathbb{E}_{x_i\sim p_{G_i}}\left[\log\left(1 - D_i(x_i)\right)\right] + \\ &-\mathbb{E}_{\hat{x}_i\sim p_{data_i}}\left[\log D_i(\hat{x}_i, F^{ca}(\boldsymbol{y}))\right] + \\ &-\mathbb{E}_{x_i\sim p_{G_i}}\left[\log\left(1 - D_i(x_i, F^{ca}(\boldsymbol{y}))\right)\right]. \end{aligned} \tag{17}$$

Here, the first two items are unconditional loss that discriminate the fake and real images, and the last two items are conditional loss discriminating whether the image and the speech description match or not. The x_i is from the model distribution G_i at the i^{th} scale, and $\hat{x}_i$ is from the real image distribution p_{data} at the same scale. The generators and discriminators were trained alternately.

IV. Experiments

A. Experimental Setup

1) Evaluation Metrics: We adopted several evaluation metrics to verify the performance of the proposed method for different aspects.

Diversity and quality of the generated images: Inception score (IS) [19] and fréchet inception distance (FID) [52] are two popular evaluation metrics for quantitative evaluation of generative models. Following [7], we used these two evaluation metrics to evaluate the diversity and quality of the generated images.

The IS is a metric for both image quality and diversity, and is defined as

$$IS = \exp\left(\mathbb{E}_x D_{KL}(p(l|x)\|p(l))\right), \tag{18}$$

where x is a generated image, and l is the label predicted by the inception model [19]. Higher IS means the model can generate more diverse and meaningful images. In the experiments, IS for databases of Place-subset and Flickr8k was calculated with the Inception-v3 pre-trained on ImageNet [50]. For CUB-200 and Oxford-102, to make a fair comparison with other methods, we used the fine-tuned Inception-v3 provided by [30]. To measure FID, we used the pre-trained inception model to extract image features. Then, the generated data distribution was modeled by a multidimensional Gaussian distribution with mean μ_x and covariance Σ_x. Correspondingly, the real data distribution was modeled by a multidimensional Gaussian distribution with mean $\mu_{\hat{x}}$ and covariance $\Sigma_{\hat{x}}$. The FID between the generated data and real data is computed as

$$\mathrm{FID}(x, \hat{x}) = \|\mu_x - \mu_{\hat{x}}\|_2^2 + \mathrm{Tr}\left(\Sigma_x + \Sigma_{\hat{x}} - 2\left(\Sigma_x \Sigma_{\hat{x}}\right)^{\frac{1}{2}}\right). \tag{19}$$

The FID measures the difference between the generated and real Gaussians, and a lower FID means a smaller distance between the generated and real image distributions, which indicates better generated images.

Semantic consistency between the generated images and their spoken descriptions: To evaluate the visual-semantic similarity between the generated images and their spoken descriptions, we conducted a content-based image retrieval experiment between the real images and the generated images. Specifically, for the Places-subset database and the Flick8k database, all ground-truth images in the test set were used as queries to retrieve the corresponding synthesized images. The retrieval performance was evaluated with R(e-call)@50 which indicates the percentage of the queries for which at least one matching synthesized image is found among the top-50 retrieval results. For the fine-grained databases, i.e., CUB-200 and Oxford-102, two queries from the real images of each class were randomly chosen because each class has many images. Since in the fine-grained databases each class has numerous images, a ground-truth real image should retrieve all matching images. The retrieval performance was then evaluated with the mean Average Precision (mAP), which is defined as

$$\mathrm{mAP} = \frac{\sum_{q=1}^{Q} \mathrm{AP(q)}}{Q}, \tag{20}$$

where Q is the number of queries, AP is Average Precision which is defined as

$$\mathrm{AP} = \frac{\sum_{k=1}^{n} (P(k) \times \mathrm{rel}(k))}{\text{number of relevant images}}, \tag{21}$$

where k is the rank in the sequence of retrieved images, n is the number of retrieved images, $P(k)$ is the percentage of matched images in the top-k retrieved images, and $rel(k)$ is an indicator function equaling 1 if the retrieved image at rank k is a correct image matching the ground-truth image.

In the image retrieval task, the images were represented by the features extracted from the Inception-v3 pre-trained on ImageNet [50]. Higher R@50 and mAP indicate a closer feature distance between the fake and their ground truth images, which indirectly shows a higher semantic consistency between the generated images and their corresponding spoken descriptions.

Quality of the speech embedding: In the ablation study, to evaluate the effectiveness of the distinctive loss on training the speech embedding network, and also to investigate the relation between the image generation performance and speech embedding performance, the speech embedding network was evaluated using a cross-modal image retrieval task. Specifically, we used the spoken captions as queries to retrieve images from the same class as described in the spoken caption. Since the distinctive loss is only used for fine-grained databases (see Section III-B1), in which each class has numerous images, a spoken caption should retrieve all matching images. The retrieval performance is then evaluated with mAP. Higher mAP means a better retrieval result, indicating better performance on embedding the speech information.

B. Results on the synthesized speech databases

The fine-grained databases CUB-200 and Oxford-102 give us the chance to perform a direct comparison with existing text-to-image generation methods because these databases are commonly used in T2IG tasks. In this section, we present the performance of our method on the fine-grained databases but using synthesized spoken descriptions rather than textual captions as input to our system. The performance is evaluated both objectively (see Section IV-B1) by comparing our system with several state-of-the-art T2IG methods and subjectively by visually inspecting the generated images and comparing these to the ground truth (see Section IV-B2).

1) Objective Results: We compare our results on the S2IG task with several state-of-the-art T2IG methods, i.e., StackGAN-v2 [7], AttnGAN [9], MirrorGAN [8] and SEGAN [31] on the T2IG task. StackGAN-v2 is a strong baseline for the T2IG task and provides the effective stacked structure for AttnGAN, MirrorGAN, and SEGAN. AttnGAN uses a word-level attention mechanism for the T2IG task. MirrorGAN uses

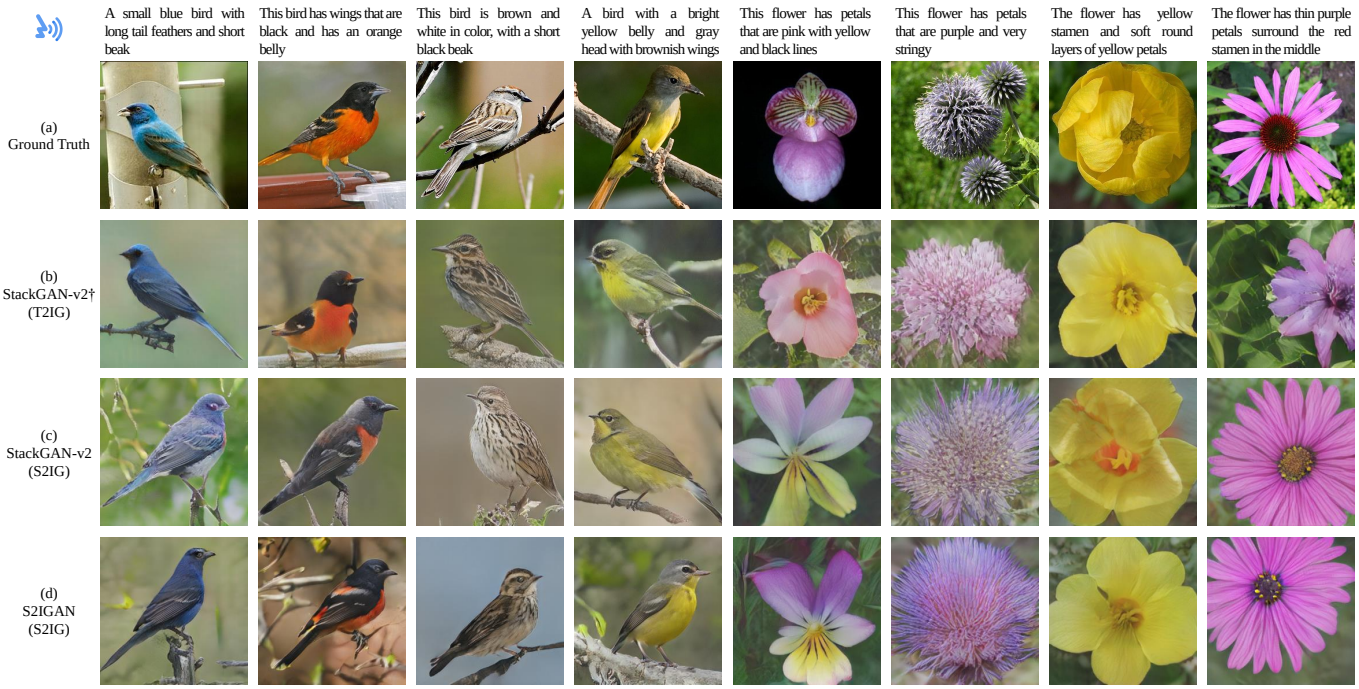


Fig. 5: Examples of images generated by different methods. † indicates the reproduced results of StackGAN-v2. The input is a spoken description. The textual descriptions are not available during training nor testing and shown here just for convenience for the reader.

TABLE II: Performance of S2IGAN compared to StackGAN-v2 [7] and [10] on the S2IG and T2IG tasks, and compared to AttnGAN [9], MirrorGAN [8], and SEGAN [31] on the T2IG task. The Classifier-based method is the baseline taken from [10]. † means our reproduced results of StackGAN-v2. The best performance is shown in bold.

Database	Method	Input	mAP	FID	IS
CUB-200 (Birds)	StackGAN-v2 [7]	text	—	15.30	4.04±0.05
	StackGAN-v2 [7]†	text	7.01	20.94	4.02±0.03
	AttnGAN [9]	text	—	—	4.36±0.03
	MirrorGAN [8]	text	—	—	4.56±0.05
	SEGAN [31]	text	—	—	**4.67±0.04**
	Classifier-based	speech	—	43.76	3.68±0.04
	Li *et al.* [10]	speech	—	18.37	4.09±0.04
	StackGAN-v2 [7]	speech	8.09	18.94	4.14±0.04
	S2IGAN	speech	**9.04**	**14.50**	4.29±0.04
Oxford-102 (Flowers)	StackGAN-v2 [7]	text	—	48.68	3.26±0.01
	StackGAN-v2 [7]†	text	9.88	50.38	3.35±0.07
	Classifier-based	speech	—	64.75	3.30±0.06
	Li *et al.* [10]	speech	—	54.76	3.23±0.05
	StackGAN-v2 [7]	speech	12.18	54.33	**3.69±0.08**
	S2IGAN	speech	**13.40**	**48.64**	3.55±0.04

this word-level attention mechanism and additionally uses a "text-to-image-to-text" structure for T2IG. SEGAN also uses the word-level attention mechanism but with an extra proposed attention regularization and a Siamese structure. Additionally, we directly compare StackGAN-v2 to our S2IGAN on the S2IG task. In order to do so, we reimplemented StackGAN-v2 and replaced the text embedding with our speech embedding. Finally, we compare our results to the recently released speech-based model by [10] and the baseline used in that study, referred to as Classifier-based, in which the speech encoder was trained with the cross-entropy loss without using visual information. Specifically, in the Classifier-based method, a classifier layer is added after the speech encoder. The speech encoder is then trained to classify the categories of the speech captions. The generator in the Classifier-based method is StackGAN-v2 [7]. This last comparison allows us to show the importance of using the relationship between the images and speech to learn the speech embeddings.

Table II shows the results of our S2IGAN compared to those of other methods for both the S2IG (with input of speech) and the T2IG (with input of text) tasks on the two fine-grained databases CUB-200 and Oxford-102. The performance is evaluated in terms of mAP, FID, and IS. Focusing specifically on the speech-based image generation task, we see that our S2IGAN method outperformed the other speech-based method of [10] and the Classifier-based method on all evaluation metrics and databases. Specifically, our method obtains 21.1% and 11.2% FID relative improvements on CUB-200 and Oxford-102, respectively, compared to [10] and 4.8% and 9.9% higher IS scores than [10] on CUB-200 and Oxford-102, respectively.

Compared to the StackGAN-v2 [7] that took our speech embedding as input, our S2IGAN achieved better mAP and FID on both databases, and also a higher IS for CUB-200. These results indicate that the proposed generative model, with the proposed densely stacked structure and the proposed relation supervisor, is effective in generating high-quality and semantically consistent images on the basis of spoken descriptions.

Speech input is generally considered to be more difficult to deal with than text because of its inherent high variability, its long duration, and the lack of pauses between words, which also means that certain standard techniques used for T2IG cannot be used for S2IG. Therefore, S2IG is

more challenging than T2IG. However, the comparison of the performances of StackGAN-v2 [7] on the S2IG and T2IG tasks shows that StackGAN-v2 [7] generated better images using speech embeddings learned by our SEN compared to their text embeddings. Moreover, the StackGAN-v2 [7] based on our learned speech embeddings outperforms [10], which also uses StackGAN-v2 as the generator but uses a different speech embedding method, on almost all evaluation metrics and databases, except for the slightly higher FID on the CUB database. These results confirm that our learned speech embeddings are competitive compared to the text embeddings in [7] and the speech embeddings in [10], showing the effectiveness of our SEN module. The better performance of the speech embeddings compared to the text embeddings of StackGAN-V2 [7] does not necessarily mean that speech as input is superior to text input for the speech-to-image task. Instead these results show that when the word-level segments in the text are underutilized and only global embeddings are learned from text as is the case in StackGAN-v2, it is possible to learn speech embeddings that are comparable to text embeddings by designing a good speech embedding network, e.g., our SEN.

When we compare the performance of the AttnGAN-based methods (i.e., AttnGAN [9], MirrorGAN [8], and SEGAN [31]) on the T2IG task with the performance of the proposed S2IGAN on the S2IG task, we see that the S2IGAN on the S2IG task still has a large gap to bridge to perform as well as the AttnGAN-based methods on the T2IG task. This difference in performance can and should be attributed to the lack of pauses between words in speech which forestalls the use of word-level attention in S2IGAN, which has proven useful for T2IG.

In short, the comparison of our S2IGAN with two speech-based models show that the S2IGAN method is competitive, and thus establishes a solid new baseline for the S2IG task. Nevertheless, compared to word-level attention mechanism-based T2IG methods on the T2IG task, the S2IGAN still has a performance gap to bridge.

2) Subjective Results: Fig. 5 shows examples of images generated by (b) StackGAN-v2 with the original text embeddings as input, (c) StackGAN-v2 with our speech embeddings as input, and (d) our S2IGAN with speech as input for the spoken descriptions for different birds and flowers shown in the first row. Row (a) shows the corresponding real images. 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 the speech embeddings, we can see that the images generated by S2IGAN are clearer and sharper.

Note that (c) StackGAN-v2 and (d) S2IGAN are both based on the speech embeddings learned by our speech embedding network. The difference is the proposed generative model with the newly proposed densely stacked structure and relation supervisor compared to the original stacked-structure in StackGAN-v2. These results show the effectiveness of the proposed RDG 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 in [7], again showing the effectiveness of the SEN module.

In order to quantitatively investigate the subjective quality of the images generated by the StackGAN-v2 and our S2IGAN on the speech-to-image task, we performed a human perceptual rating experiment on the Oxford test dataset, similar to that by [8]. Moreover, to compare the performance of the proposed Relation Supervisor with a speech-image matching loss on the S2IG task (see Section III-C2 and in Section IV-D2 for the objective results), a variation of S2IGAN, named DAMSM-based, was also included in the human perceptual study.

Sixty-four subjects (age range: 18-40; ratio male-female: 11:8; ratio graduate student-employees: 8:1) from three universities[3] with different professional backgrounds participated in an online rating study. They participated voluntarily and were not paid for participation.

The task for the participants in the rating experiment consisted of two parts: 1) a quality test which compares the quality of the images generated by the different methods; 2) a semantic consistency test which compares the semantic consistency of the images generated by the different methods with their ground-truth images. For the tests, we randomly selected 3 spoken captions from each of the 20 classes in the Oxford test dataset and generated their corresponding images with StackGAN-v2, S2IGAN, and DAMSM-based. This resulted in 60 images generated by each model. These (3 models $\times$ 60 images =) 180 images were used for the rating of both the quality and the semantic consistency.

In the quality test part, the participants were shown the three images generated by the three models on the basis of the same caption next to each other on a screen and were asked to choose the best quality image from these three generated images. In total, the participants were asked to rate 60 sets of three generated images (i.e., 3 captions from 20 classes). After the quality test part, participants were shown the same 60 sets of generated images again in the semantic consistency test part. However, in the semantic consistency test, in addition to the three generated images, a fourth image, the ground-truth image, was added. Here, the participants were asked to choose the generated image that was most consistent with the ground-truth image.

Fig. 6 shows the percentage of the total votes the generated images of each model obtained for the quality test (on the left) and the semantic consistency test (on the right). The images generated by our S2IGAN were most often chosen as the best image compared to the images of both StackGAN-v2 and the DAMSM-based method, both in terms of quality of the generated images and the semantic consistency between the ground-truth image and the generated image. A one-way ANOVA confirmed these results: there was a significant difference in the number of votes the images of a particular model obtained in the quality test ($F(2,177) = 43.58$, $p < .001$) and for the semantic consistency test ($F(2,177) = 21.22$, $p < .001$), meaning that the quality and the semantic consistency of the three models differed significantly. Post-

[3]Delft University of Technology, the Netherlands, Eindhoven University of Technology, the Netherlands, and Xi'an Jiaotong University, China.

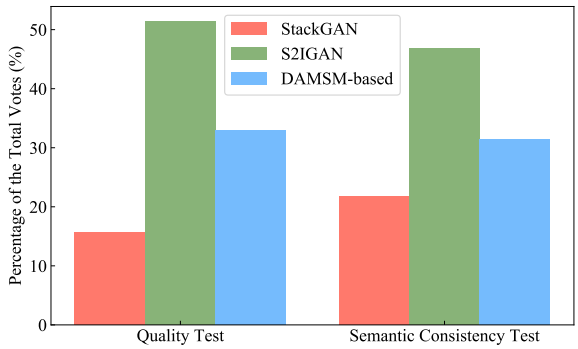


Fig. 6: The results (in terms of percentage of the total votes) for the generated images of the three models for the quality test on the left and the semantic consistency test on the right from the human rating test.

hoc comparisons using the Tukey HSD test showed that all models differed significantly from one another in the quality test and the semantic consistency test (all $p < .001$; except for the difference between StackGAN-v2 and DAMSM in the semantic consistency test: $p = .038$).

To illustrate S2IGAN's ability to catch subtle semantic differences in the spoken descriptions, we generated images conditioned on spoken descriptions in which we changed color keywords. Fig. 7 shows synthesized bird images of S2IGAN on the basis of the spoken input "A small bird with a (color on y axis) belly and (color on x axis) wings". As can be seen in Fig. 7, 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 fine-grained 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 spoken description.

C. Results on the real speech databases

Real speech is more variable in terms of the pronunciation of sounds due to differences in speakers, speaking rate, pitch, etc., compared to synthesized speech. Moreover, real speech databases contain more background noise than the synthesized speech databases. The Places-subset and Flikcrk8k with real speech captions make it possible to test S2IGAN's performance on the S2IG task with real spoken descriptions as input. Moreover, compared to the fine-grained databases, the images in these two databases are more complex, which give us the chance to investigate S2IGAN's ability to synthesize complex scene images. The performance of our method on the complex scene image databases using natural spoken descriptions is evaluated both objectively (see Section IV-C1, in which we also compare our method with other methods) and subjectively (see Section IV-C2).

1) Objective Results: We compare our S2IGAN to StackGAN-v2 for the S2IG task on the two scene image databases with natural spoken descriptions, i.e., Places-subset and Flickr8k, and to [10] on the Places-subset. Moreover, considering the good performance of the AttnGAN-based methods

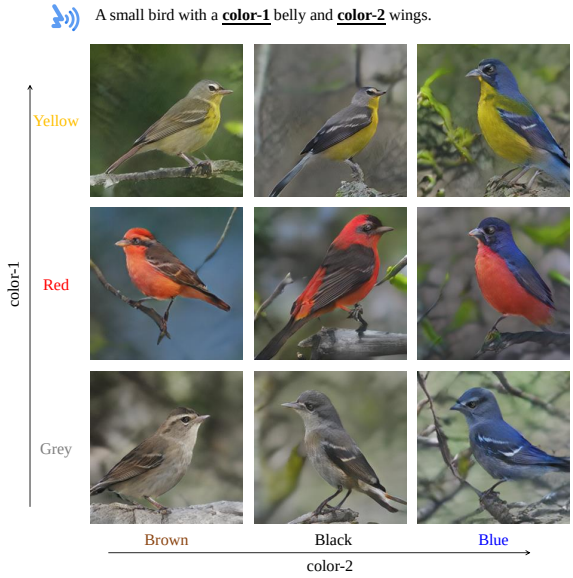


Fig. 7: Generated examples by S2IGAN. The generated images are based on speech descriptions with different color keywords. The input is a spoken description. The textual descriptions are not available during training nor testing and shown here just for convenience.

TABLE III: Performance of S2IGAN compared to AttnGAN [9] and StackGAN-v2 [7] on the Places-subset and Flickr8k databases, and compared to [10] on the Places-subset database. The best performance is shown in bold. † means the text is transcribed from speech using an automatic speech recognition system.

Database	Method	Input	R@50	FID	IS
Places-subset	AttnGAN [9]	text†	**33.85**	**35.59**	**4.59±0.51**
	Li *et al.* [10]	speech	—	83.06	—
	StackGAN-v2 [7]	speech	8.87	47.94	3.78±0.35
	S2IGAN	speech	12.95	42.09	4.04±0.25
Flickr8k	AttnGAN [9]	text†	**50.40**	**84.08**	**12.37±0.41**
	StackGAN-v2 [7]	speech	16.40	101.74	8.36±0.39
	S2IGAN	speech	16.40	93.29	8.72±0.34

[9], [8], [31] on the T2IG task, we take the AttnGAN [9] to create an upper bound performance on the two real speech databases. As the AttnGAN [9] is designed for the T2IG task, we first transcribed the spoken captions to text using an automatic speech recognition (ASR) system[4] built with Kaldi [53]. This ASR system consists of a hybrid factorized time-delay neural network (TDNN-F) [54] acoustic model (AM) and a four-gram language model (LM), both trained using the 960-hour Librispeech English database [55]. The automatically transcribed text is then used as input to the AttnGAN.

Table III shows the results on the Places-subset and Flickr8k databases. The performance is measured in terms of R@50, FID, and IS. The results of StackGAN-v2 are obtained with our speech embeddings. Please note, [10] only reports FID performance on the Places-subset database.

On the Places-subset, huge improvements in the image quality (shown by a decrease on FID) are observed for the speech embedding-based StackGAN-v2 [7] compared to [10]. This improvement over [10] is solely driven by our speech

[4]https://kaldi-asr.org/models/m13

encoder, which shows the good performance of our method on obtaining speech embeddings. Additionally, the proposed S2IGAN outperforms StackGAN-v2 [7] with a substantial margin on all evaluation metrics. Specifically, our method achieves 46.0%, 12.3%, and 6.9% relative improvements on R@50, FID, and IS respectively over StackGAN-v2. These results show the good performance of our proposed generative model (RGD) with the densely-stacked structure and relation supervisor, compared to the original stacked structure of StackGAN-v2.

On Flickr8k, S2IGAN again obtains similar or better results than StackGAN-v2 on all metrics. The results on both real speech databases confirm that the proposed S2IGAN achieves state-of-the-art S2IG performance not only on synthesized speech but also on real speech.

Comparing the transcription-based AttnGAN with the speech-based methods shows a large superiority for the transcription-based model, especially on the measure of R@50. However, in terms of efficiency, our S2IGAN outperforms the transcription-based AttnGAN: the average time required by our S2IGAN and the transcription-based AttnGAN to generate an image of the Flickr8k testing set was 23.3 ms and 54.1 ms, respectively. The superiority of the transcription-based AttnGAN on the used measures can be explained as follows: 1) the well-trained ASR system alleviates the effect of noisy and content-irrelevant information in the speech signal during the transcription process while in the speech-based methods this information is at least partially embedded in the speech embeddings, and 2) unlike the speech-based methods, the transcription-based method allows for the use of the word-level attention mechanism, which can connect a word within a sentence to a specific object in an image. Note that a well-trained ASR system depends on a huge number of transcriptions, which means that the transcription-based method cannot be used for unwritten languages, as was the objective of our research. The performance of AttnGAN given in Table III is only to show an upper bound rather than give a fair comparison. The analysis of the efficiency showed the good efficiency of the proposed S2IGAN.

2) Subjective Results: To show what images can be generated by S2IGAN after training on complex scene image databases with natural speech as input, Fig. 8 presents several examples of synthesized images on the Places-subset (top half of figure) and Flickr8k (bottom half of figure) and their respective ground truth images. The spoken captions are shown above the images. The colored words in the written descriptions indicate those corresponding objects that are well synthesized (blue), objects that are reflected semantically but not photo-realistically (green), and objects that are not synthesized (red), respectively.

As seen in the images in the top half of Fig. 8, scenes in the images of the Places-subset are much more complicated than those in the fine-grained databases. Many different objects can appear within one image. For instance, an image of a dining room may consist of paintings, tables, chairs, and lamps. Despite the complicated scenes, S2IGAN is still able to generate images that are semantically consistent with the corresponding spoken descriptions. As Fig. 8 shows, the synthesized images on the Places-subset can reflect scenes described by the speech signal relatively well. The generated images can easily be labelled as, e.g., a hotel room, a bedroom, a living room, and a kitchen, showing that S2IGAN is able to synthesize photo-realistic images in the S2IG task with real speech as input.

Nevertheless, details in the scenes are generated less well. Although each scene category has around 2,000 images in the Places-subset, each image is unique and images belonging to the same scene category consist of different specific objects, resulting in a diversity that makes it hard to capture every object described in the spoken captions. For instance, in the second spoken caption of the Places-subset in Fig. 8, there are three objects: bedroom, ceiling fan, and lamp, while only two appear in the generated image. This is especially the case for small objects, e.g., a ceiling fan, a chandelier, and the monitor. In these synthesized images, some of those small objects (in red color in the captions) failed to be synthesized. Additionally, some small objects (in green color in captions) can be reflected to some extent but are not presented well.

The lower part of Fig. 8 are synthesized examples for Flickr8k. As stated in Section III-A2, Flickr8k is a highly diverse database with various scenes but limited examples per scene, making S2IG quite a challenging task on this database. However, from the synthesized images we can see that the basic scenes described in the spoken description, e.g., waves, water, snow, forest, and grass (indicated with the blue color), can be successfully synthesized by S2IGAN. Furthermore, despite that foregrounds in these generated images are not so photo-realistic, they still show (some) semantic consistency with the corresponding spoken descriptions (see the green words in the spoken captions). These results indicate that SEN is capable of learning semantic information from such a diverse database with real speech descriptions.

D. Ablation Study

The effectiveness of the key components of S2IGAN, i.e., the densely-stacked structure of the image generator, the relation supervisor, the speech embedding network, and the distinctive loss in SEN are investigated separately. Because the synthesized images on the two fine-grained databases are more photo-realistic than those on the scene image databases, and consequently their results are more stable, these two databases are primarily used for the ablation study.

The results are shown in Table IV, which shows the S2IG performance on CUB-200 and Oxford-102 of variants of S2IGAN in which different parts are removed (indicated with "w/o" in Table IV). As an end-to-end training method is generally considered to be more friendly to users, the performance of S2IGAN trained in an end-to-end way is also evaluated. The relation supervisor is slightly different for the fine-grained databases compared to the complex scene databases because there is no second image that can be described by a given speech caption in the Places-subset and Flickr8k (see Eq. 12 and Eq. 13). So, the performance of the RS is also investigated on the Places-subset and Flickr8k (see Table V). Moreover, to compare the performance of the proposed RS and a speech-image matching loss, the performance of DAMSM-based method is also shown in Table IV (see Section III-C2).

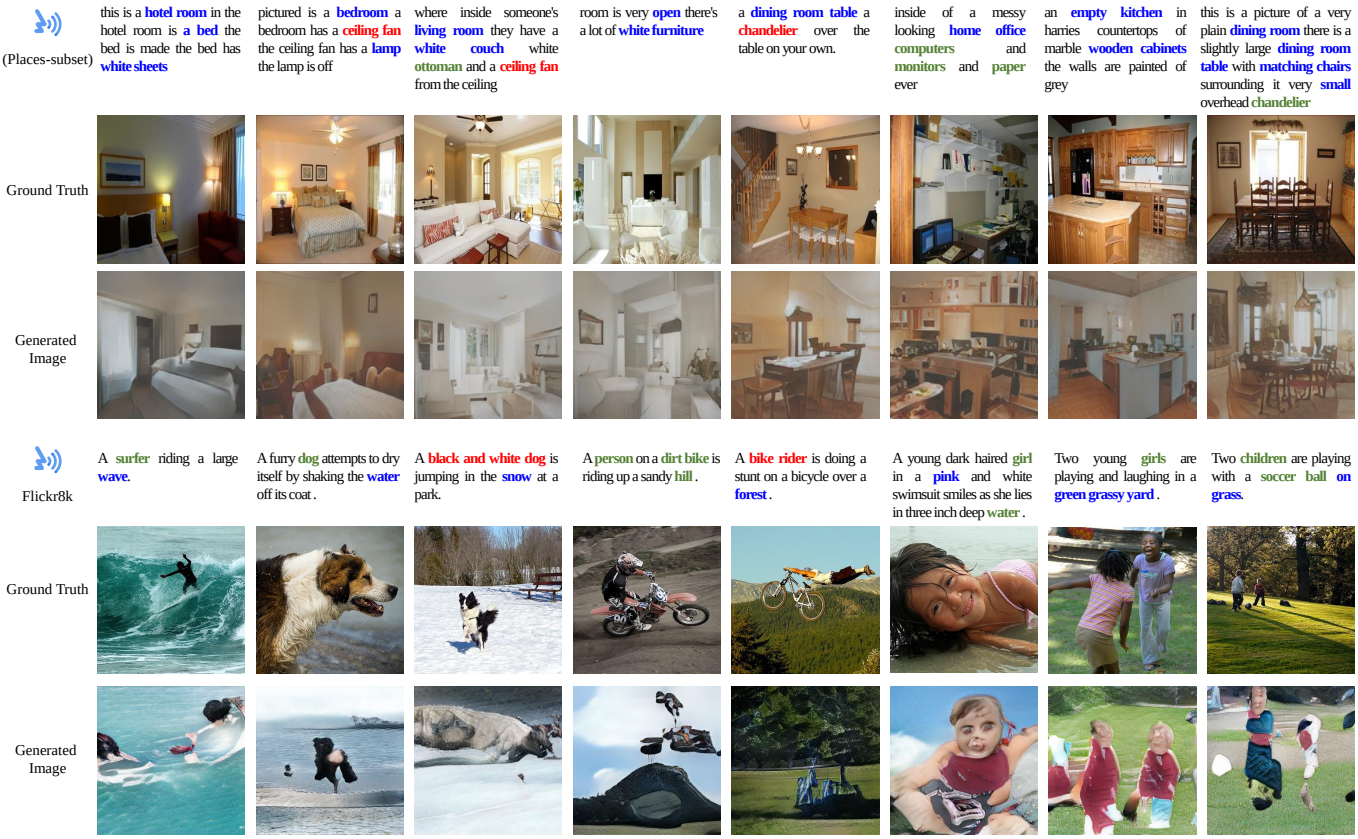


Fig. 8: Examples of images generated by S2IGAN and their ground truth images on the basis of the Places-subset (top half of figure) and Flickr8k (bottom half of figure). The input is a spoken description (shown above the images). The textual descriptions are not available during training nor testing and are shown here for convenience only. The colored words in the textual descriptions indicate that the corresponding objects are well synthesized (blue), the objects are reflected semantically but not photo-realistically (green), and the objects are not synthesized (red), respectively.

TABLE IV: Ablation study results of S2IGAN; w/o means without. The best results are shown in bold.

Database	CUB-200 (Bird)			Oxford-102 (Flower)		
Evaluation Metric	mAP	FID	IS	mAP	FID	IS
S2IGAN w/o Dense	8.66	17.58	4.19±0.04	13.13	64.37	3.68±0.05
S2IGAN w/o RS	8.54	15.59	4.14±0.05	12.86	53.24	3.70±0.08
S2IGAN w/o SEN	2.91	19.56	3.49±0.05	7.38	67.60	2.77±0.04
S2IGAN end-to-end	7.38	21.54	4.29±0.06	12.47	51.88	3.55±0.08
S2IGAN	9.04	**14.50**	**4.29±0.04**	**13.40**	**48.64**	3.55±0.04
DAMSM-based	**9.60**	15.06	4.26±0.05	12.46	55.70	**3.70±0.04**

1) Effect of the densely-stacked structure of the image generator: We evaluated the effect of the densely-stacked structure of the DG by changing it with the traditional stacked structures as in StackGAN-v2 (see S2IGAN w/o Dense in Table IV) while keeping the proposed relation supervisor and speech embedding method. Comparing the results to S2IGAN, S2IGAN w/o Dense shows a performance drop for most evaluation metrics on both databases. Specifically, there is a decrease in mAP and a clear increase in FID when the densely stacked structure is removed, confirming the effectiveness of the proposed densely-stacked structure for image generation. Note that the dense connection method or the residual connection method is a common strategy in the neural networks, and here, for the first time, the proposed method shows the usefulness of this strategy for the stacked structure in the image generation task.

2) Effect of the Relation Supervisor: The effect of the relation supervisor (see RS in Fig. 4 and loss function presented in Eq. 12-13) on training the generators were investigated for the synthetic speech and real speech databases separately. The effect of the RS can be observed by comparing the performances of S2IGAN and S2IGAN w/o RS in Tables IV and V for the synthesized and real speech databases, respectively. The results show that the RS module leads to improvements in FID and IS (except for the Oxford-102 database) for all databases. Moreover, the addition of the RS leads to a higher mAP for the synthesized speech databases and an improved R@50 for the Places-subset database. No difference in R@50 for Flickr8k was observed. These results indicate the effectiveness of the RS in ensuring the semantic consistency between the synthesized images and the corresponding spoken descriptions.

To further investigate the effectiveness of our RS, we compared it with the commonly used, intuitive way, to perform the semantic consistency constraint, i.e., the image-speech matching loss in AttnGAN [9], introduced in Section III-C2, to replace our RS module, which we refer to as the DAMSM-based method. Due to the lack of word-level segments in the speech signal, the word-level attention mechanism in the

TABLE V: Effect of the Relation Supervisor on the Places-subset and Flickr8k; w/o means without. The best results are shown in bold.

Database	Places-subset			Flickr8k		
Evaluation Metric	R@50	FID	IS	R@50	FID	IS
S2IGAN w/o RS	11.98	45.34	4.01±0.28	**16.40**	93.84	8.72±0.34
S2IGAN	**12.95**	**42.09**	**4.04±0.25**	**16.40**	**93.29**	**9.32±0.56**

TABLE VI: Comparison of Relation Supervisor and DAMSM. The best results are shown in bold.

Database	CUB-200 (Bird)			Oxford-102 (Flower)		
Evaluation Metric	mAP	FID	IS	mAP	FID	IS
DAMSM-based	**9.60**	15.06	4.26±0.05	12.46	55.70	**3.70±0.04**
S2IGAN	9.04	**14.50**	**4.29±0.04**	**13.40**	**48.64**	3.55±0.04

original DAMSM [9] was dropped. Comparing the results of S2IGAN with those of the DAMSM-based method in Table IV shows that, in agreement with our hypothesis (see Section III-C2), the DAMSM-based method achieves comparable results to our S2IGAN (RS-based method) on all performance measures, indicating that the DAMSM image-speech matching loss is helpful from the perspective of those objective evaluation metrics. However, again as hypothesized, the human perceptual rating study (see Section IV-B2) showed that the quality of the images generated by the DAMSM-based method were significantly lower than those generated by our S2IGAN. The images generated by the DAMSM-based method were found to be less clear and less sharp, indicating that our proposed RS is better than the image-speech matching loss method.

The lower quality of the generated images by the DAMSM-based method is likely due to the following reason. In the original DAMSM module in [9], the word-attention mechanism is an important module to align the regional information from an image with the corresponding words in the written caption. However, when the DAMSM is implemented for the speech-to-image task without the word-level attention mechanism, DAMSM only considers the matching relation between images and speech from the perspective of the global features. Because all the standard evaluation metrics are calculated on the basis of the global image features, the constraint of DAMSM on the global image features does not lead to a decrease in those performance measures compared to our proposed RS. However, this matching constraint between speech and images in the embedding space ignores the local details of images but only works on a semantic level, which could affect the quality of the generated images. In contrast, our proposed RS does not compare two different modalities, i.e., speech and image, but rather aims to distinguish different images in which the local details of images would also play an important role, consequently leading to a better quality of the generated images compared to the DAMSM-based method.

3) Effect of the speech encoder network: In order to investigate the effect of the SEN on our S2IGAN, we forewent the pretraining of the speech encoder (SED in Fig. 4) (see Fig. 3). In practice, two end-to-end training methods were performed. One is the S2IGAN w/o SEN in Table IV. In this method, the speech encoder is not pre-trained. Instead, the speech encoder is trained along with the generative model in an end-to-end way with a loss function for the generative model (Eq. 14 and Eq. 16). In the other end-to-end training method (S2IGAN end-to-end in Table IV), the loss function for training SEN is adopted while training the speech encoder and generative model simultaneously. As can be seen, S2IGAN w/o SEN shows a substantial reduction in performance on all three measures. The S2IGAN end-to-end method also shows a slight decrease on all metrics on both databases. These results indicate a well-trained speech encoder is quite important for S2IG, and also shows the necessity to pre-train the speech encoder with SEN in our S2IGAN system. Moreover, these results as well as other results of the proposed S2IGAN, for the first time, show that the visually-grounded method can be used to learn speech embeddings for S2IG task.

4) Effect of the distinctive loss in the speech encoder network: Lastly, we investigated the effect of the distinctive loss $\mathcal{L}_d$ in Eq. 8 on training the SEN on the fine-grained databases (the distinctive loss $\mathcal{L}_d$ was not used for training the real speech databases) by removing $\mathcal{L}_d$ from Eq. 8 (i.e., the speech embeddings are no longer mapped into the label space in Fig. 3). The effect of the distinctive loss on the SEN is evaluated using a speech-based image retrieval task with the learned image embeddings and speech embeddings, which will inform us about the discriminating ability of the learned embeddings. Moreover, we investigated the effect of the distinctive loss on the relationship between the performance of the speech embedding network and the S2IG tasks, which will be evaluated on image generation.

Table VII shows the speech-based image retrieval performance, which is evaluated with mAP, and the image generation performance, which is evaluated with mAP, FID, and IS. Training the SEN without using $\mathcal{L}_d$ results in a performance drop in terms of mAP on both databases for both the SEN and the full S2IGAN (although this drop is marginal for the Oxford-102 database for the SEN). Importantly, a better performance of SEN always led to an increase in the performance of S2IGAN, showing the importance of learning a good speech embedding for the task of image generation.

To better understand the role of the loss function, we visualized the speech feature distributions produced by the SEN trained with and without $\mathcal{L}_d$ using t-SNE [56]. The t-SNE visualizes the distribution of the embeddings in a two-dimensional space via dimensionality reduction. Speech embeddings extracted from the speech encoder in the SEN trained with and without distinctive loss are visualized in Fig. 9. For ease of inspection, the presented data are from 10 randomly selected classes from the CUB-200 test database, and only the first spoken caption for each image is selected. In this figure, each point indicates a speech embedding, and points with the same color are from the same class. The class IDs are those from the CUB-200 database [12]. The distance between different points (i.e., embeddings) indicates the relative similarity of the embeddings. Smaller distances indicate more similar embeddings. An ideal speech encoder should cluster embeddings from the same class close together while embeddings from different classes should be further

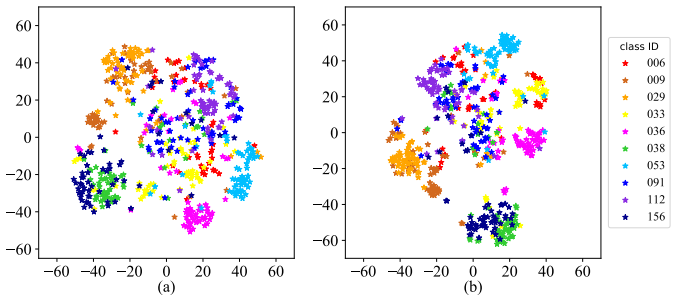


Fig. 9: Visualization of the distribution of the speech embeddings created by SEN without (a) and with (b) distinctive loss. For ease of inspection, the presented data are from 10 randomly selected classes (with IDs) from the CUB-200 test database.

apart. From this figure we can see that the use of $\mathcal{L}_d$ increases the distance between different classes, which is important to create semantic-discriminative speech embeddings for fine-grained databases. For instance, after training SEN using $\mathcal{L}_d$, class 112 was no longer mixed with other classes. So, both the objective results and the visualizations show that $\mathcal{L}_d$ is critical for learning more discriminative speech embeddings, which further helps S2IGAN to generate better semantically-consistent images on fine-grained databases.

TABLE VII: Effect of the distinctive loss $\mathcal{L}_d$ in SEN. mAP of SEN is calculated on the real test images.

Database	$\mathcal{L}_d$	SEN	S2IGAN		
		mAP	mAP	FID	IS
CUB-200 (Bird)	w/o	23.68	6.80	16.66	4.19±0.05
	w/	**24.24**	**9.04**	**14.50**	**4.29±0.04**
Oxford-102 (Flower)	w/o	41.85	10.03	69.47	3.35±0.08
	w/	**41.86**	**13.40**	**48.64**	**3.55±0.08**

V. DISCUSSION

The proposed S2IGAN was evaluated on four benchmark databases for the task of speech-to-image generation: 1) two fine-grained databases commonly used databases for the text-to-image generation task, i.e., CUB-200 and Oxford-102. These databases allow us to perform a direct comparison of our S2IGAN with state-of-the-art T2IG methods. Because the fine-grained databases do not have spoken captions, we synthesized spoken captions based on the available text captions using a TTS system. 2) Two complex scene databases, which contain images and corresponding natural spoken captions. These databases allow us to investigate the ability of S2IGAN to synthesize images using natural speech as input, and provide a different and more challenging type of images (compared to the fine-grained databases).

Results on the fine-grained databases show that our proposed S2IGAN not only performs well on the task of fine-grained image generation but also achieves state-of-the-art performance. Its success is due to the good performance of both the speech embedding network and the generative model, specifically the newly proposed densely stacked structure and the newly proposed relation supervisor. The distinctive loss used in training the speech embedding network for the S2IG task (see Section IV-D4) was found to be highly important for training a good speech encoder for the fine-grained image databases, which subsequently was shown to be important in generating better images.

The results obtained on the two complex scene image databases with natural spoken captions, i.e., the Places-subset and Flickr8k, show that our S2IGAN can capture the semantics in natural speech and can use this to generate images that are semantically consistent with the spoken descriptions at the global level, although these images are of a lower quality than those based on the fine-grained databases due to details in the scenes being less well synthesized or missing entirely. This lesser performance is, on the one hand, due to the much higher variability of the speech signal in real speech compared to the synthesized speech used for the fine-grained image databases. On the other hand, it is due to the use of complex scene images for training for which only a few (or only one in the case of Flickr8k) images exist per class. Specifically, for the Places-subset, although the generated images were photo-realistic, S2IGAN was found to miss small objects when synthesizing images. While for Flickr8k, which is an even more diverse database, S2IGAN generated the basic scenes but failed to synthesized photo-realistic objects in the images. This weakness of S2IGAN on such complex scene images might be (partly) alleviated by adding additional information, such as bounding boxes as used in [33] for text-to-image generation on complex scene image databases.

The superiority of the text-to-image AttnGAN [9] on the two real-speech databases suggests that word-level attention might be important for the image generation task. In the future, we will therefore try to incorporate segment- or word-level attention into our model to improve the image quality and accuracy. An interesting approach for this would be to automatically discover speech units based on corresponding visual information from the speech signal [57] to segment the speech signal, which would allow us to use segment- and word-level attention mechanisms to improve the performance of speech-to-image generation. An alternative, simpler way to implement the word-level attention mechanism could be using the downsampling strategy to get temporal speech representations instead of global feature vectors.

Finally, we should note that due to database limitations, all databases adopted in this study are English. For future work, it will be highly interesting to test the proposed methodology on a truly unwritten language rather than the well-resourced English language.

VI. CONCLUSION

This paper introduced 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 the speech embedding network. Second, high-quality images are generated on the basis of the speech embeddings extracted by the speech encoder in the speech embedding network. The results of extensive experiments on four commonly-used databases show that our S2IGAN has state-of-the-art performance on

the task of S2IG. When trained on databases with a finite number of classes and multiple images per class (i.e., fine-grained image databases), the S2IGAN can generate high-quality photo-realistic images. When trained on databases that contain a more diverse set of scenes, the S2IGAN is able to generate global scenes but misses details.

ACKNOWLEDGMENT

This work has been partially supported by the National Key R&D Program of China (No. 2018AAA0102504). The authors also thank all participants for evaluating the generated images in the human perceptual rating experiment.

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.

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

[3] Z. Yi, H. Zhang, P. Tan, and M. Gong, "Dualgan: Unsupervised dual learning for image-to-image translation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[4] C. Wang, C. Xu, C. Wang, and D. Tao, "Perceptual adversarial networks for image-to-image transformation," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 4066–4079, 2018.

[5] R. Wu, G. Zhang, S. Lu, and T. Chen, "Cascade ef-gan: Progressive facial expression editing with local focuses," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2020.

[6] G. Zhang, M. Kan, S. Shan, and X. Chen, "Generative adversarial network with spatial attention for face attribute editing," in *European conference on computer vision*, 2018.

[7] 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.

[8] 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.

[9] 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.

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

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

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

[13] 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.

[14] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Advances in neural information processing systems*, 2014.

[15] D. Harwath, A. Recasens, D. Surís, G. Chuang, A. Torralba, and J. Glass, "Jointly discovering visual objects and spoken words from raw sensory input," in *European conference on computer vision*, 2018.

[16] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *Journal of Artificial Intelligence Research*, vol. 47, pp. 853–899, 2013.

[17] D. Harwath and J. Glass, "Deep multimodal semantic embeddings for speech and images," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015.

[18] X. Wang, T. Qiao, J. Zhu, A. Hanjalic, and O. Scharenborg, "S2IGAN: Speech-to-Image Generation via Adversarial Learning," in *Proc. Interspeech 2020*, 2020, pp. 2292–2296.

[19] 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.

[20] Y. Wen, B. Raj, and R. Singh, "Face reconstruction from voice using generative adversarial networks," in *Advances in neural information processing systems*, 2019.

[21] A. Duarte, F. Roldan, M. Tubau, J. Escur, S. Pascual, A. Salvador, E. Mohedano, K. McGuinness, J. Torres, and X. Giro-i Nieto, "Wav2pix: speech-conditioned face generation using generative adversarial networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[22] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019.

[23] H.-S. Choi, C. Park, and K. Lee, "From inference to generation: End-to-end fully self-supervised generation of human face from speech," *arXiv:2004.05830*, 2020.

[24] H. Huang, Z. Wu, S. Kang, D. Dai, J. Jia, T. Fu, D. Tuo, G. Lei, P. Liu, D. Su *et al.*, "Speaker independent and multilingual/mixlingual speech-driven talking head generation using phonetic posteriorgrams," *arXiv preprint arXiv:2006.11610*, 2020.

[25] L. Song, W. Wu, C. Qian, R. He, and C. C. Loy, "Everybody's talkin': Let me talk as you want," *arXiv preprint arXiv:2001.05201*, 2020.

[26] H.-Y. Lee, X. Yang, M.-Y. Liu, T.-C. Wang, Y.-D. Lu, M.-H. Yang, and J. Kautz, "Dancing to music," in *Advances in Neural Information Processing Systems*, 2019, pp. 3586–3596.

[27] H. Zhu, M. Luo, R. Wang, A. Zheng, and R. He, "Deep audio-visual learning: A survey," *arXiv preprint arXiv:2001.04758*, 2020.

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

[29] 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.

[30] 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.

[31] 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.

[32] Q. Lao, M. Havaei, A. Pesaranghader, F. Dutil, L. D. Jorio, and T. Fevens, "Dual adversarial inference for text-to-image synthesis," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[33] W. Li, P. Zhang, L. Zhang, Q. Huang, X. He, S. Lyu, and J. Gao, "Object-driven text-to-image synthesis via adversarial training," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019.

[34] S. Palaskar, V. Raunak, and F. Metze, "Learned in speech recognition: Contextual acoustic word embeddings," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019.

[35] A. Haque, M. Guo, P. Verma, and L. Fei-Fei, "Audio-linguistic embeddings for spoken sentences," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019.

[36] D. Harwath, A. Torralba, and J. Glass, "Unsupervised learning of spoken language with visual context," in *Advances in neural information processing systems*, 2016.

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

[38] G. Ilharco, Y. Zhang, and J. Baldridge, "Large-scale representation learning from visually grounded untranscribed speech," *arXiv:1909.08782*, 2019.

[39] H. Kamper, G. Shakhnarovich, and K. Livescu, "Semantic speech retrieval with a visually grounded model of untranscribed speech." *IEEE/ACM Trans. Audio, Speech & Language Processing*, vol. 27, no. 1, 2019.

[40] O. Scharenborg, L. Besacier, A. Black, M. Hasegawa-Johnson, F. Metze, G. Neubig, S. Stüker, P. Godard, M. Müller, L. Ondel *et al.*, "Speech technology for unwritten languages," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 964–975, 2020.

[41] 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.

[42] X. Wang, S. Pang, and J. Zhu, "Domain segmentation and adjustment for generalized zero-shot learning," *arXiv:2002.00226*, 2020.

[43] Y. Xian, S. Sharma, B. Schiele, and Z. Akata, "f-vaegan-d2: A feature generating framework for any-shot learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019.

[44] J. Li, M. Jing, K. Lu, Z. Ding, L. Zhu, and Z. Huang, "Leveraging the invariant side of generative zero-shot learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019.
[45] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018.
[46] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019.
[47] K. Ito, "The lj speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.
[48] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015.
[49] 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.
[50] 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.
[51] 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:1406.1078*, 2014.
[52] 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.
[53] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *Proc. ASRU*, 2011.
[54] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks," in *Proc. INTERSPEECH 2018*, 2018, pp. 3743–3747.
[55] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.
[56] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, 2008.
[57] D. Harwath and J. Glass, "Towards visually grounded sub-word speech unit discovery," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019.