# Direct Speech-to-Image Translation

Jiguo Li, Xinfeng Zhang, *Member, IEEE,* Chuanmin, Jia, Jizheng Xu, *Senior Member, IEEE,* Li Zhang, Yue Wang, Siwei Ma, *Senior Member, IEEE,* and Wen Gao *Fellow, IEEE,*

***Abstract*—Direct speech-to-image translation without text is an interesting and useful topic due to the potential applications in human-computer interaction, art creation, computer-aided design. etc. Not to mention that many languages have no writing form. However, as far as we know, it has not been well-studied how to translate the speech signals into images directly and how well they can be translated. In this paper, we attempt to translate the speech signals into the image signals without the transcription stage. Specifically, a speech encoder is designed to represent the input speech signals as an embedding feature, and it is trained with a pretrained image encoder using teacher-student learning to obtain better generalization ability on new classes. Subsequently, a stacked generative adversarial network is used to synthesize high-quality images conditioned on the embedding feature. Experimental results on both synthesized and real data show that our proposed method is effective to translate the raw speech signals into images without the middle text representation. Ablation study gives more insights about our method.**

***Index Terms*—Speech-to-image translation, cross-modal generation, generative adversarial network, teacher-student learning.**

## I. Introduction

IT has been widely accepted by cognitive science community that infants begin learning their native language not by learning words, but by discovering the correlations between the speech signal and visual information [1]. Infants know some aspects of their language by 6–12 months, while they do not understand the common native-language words until 12 months [2]. When communicating with their parents, the infants only receive continuous speech signals from the parents and the visual signals from the surrounding. And the infants can learn the correlation between the high-frequency speech words and the objects or local visual textures. Thus, it is interesting to explore whether a machine can translate the speech signals into images directly, without the help of language words. Translating data between different modalities is a cutting-edge area recently. However, speech-to-image translation has not been well-studied while the similar topic, text-to-image translation, have been investigated in recent literature [3]–[5]. Besides, many languages have no writing form, which calls for the approaches to understand and visualize the speech directly [6]. Not to mention the potential applications in human-computer interaction, art creation and computer-aided design, where speech is the nature input and middle text representation is not necessary. So exploring speech-to-image translation is necessary and meaningful.

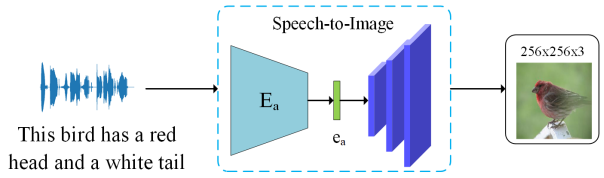


Fig. 1. Illustration for our task: speech-to-image translation without text. **Note: the text is shown only for readability, it is not used in the speech-to-image model.**

As illustrated in Fig. 1, given the raw speech descriptions: "this bird has a red head and a white tail", the corresponding images can be synthesized, which means that the machine has understood the speech signal to some extent and been able to translate the semantic information in the speech signal into the image. Speech and image are in different modalities and the modality gap between these two types of data makes direct speech-to-image translation not trivial. Text-to-image translation [3]–[5], [7] is a closely related topic to ours, which has been investigated for several years. In some text-to-image models, zero-shot learning based methods [7], [8] and generative adversarial networks (GANs) [9] have been used to extract features and synthesize realistic images, respectively. These models generalize better on the new testing classes by leveraging the teacher-student learning to train the text encoder [7], [10]. Compared with the text-to-image translation, speech-to-image translation might be more challenging because the speech signals are continuous, unaligned and noisy. In addition to the text-to-image translation, several models for audio-to-image generation were also presented in recent years. Chen *et al.* [11] and Hao *et al.* [12] synthesized instrument images from different music inputs; Oh *et al.* [13] and Amanda *et al.* [14] reconstructed the human face images from input speech based on the positive correlations between a person's appearance and his voice, and both their frameworks contain a speech encoder and a face decoder. Different from these audio-to-image generation works, which model the acoustic or phonetic information mainly, our speech-to-image translation aims to model the linguistic information in the input speech and translate it into the images. Recent works about audio-

Jiguo Li is with the Key Lab of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China, the University of Chinese Academy of Sciences, Beijing 100049, China, and also with the National Engineering Laboratory for Video Technology, School of Electronic Engineering and Computer Science, Peking University, Beijing 100871, China (e-mail: jiguo.li@vipl.ict.ac.cn)

Xinfeng Zhang is with the School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: xfzhang@ucas.ac.cn)

Siwei Ma, Chuanmin Jia, Wen Gao are with the Institute of Digital Media, School of Electronic Engineering and Computer Science, Peking University, Beijing 100871, China, and are also with the Peng Cheng Lab, Shenzhen, China (e-mail: swma, cmjia, wgao@pku.edu.cn).*(Corresponding author: Prof. Siwei Ma)*

Jizheng Xu, Li Zhang, Yue Wang are with Bytedance Inc. (e-mail: xujizheng, lizhang.idm, wangyue.v@bytedance.com)

visual correlation learning [15], [16] have shown it is feasible to learn the correlation between visual speech descriptions and the objects or local texture in the images, forming the basis of our speech-to-image translation task. In some other topics about speech processing, such as speech-to-speech translation and speech keyword search, recent works [17], [18] have attemped to translate or search the speech without the help of the transcription text, indicating it is feasible to understand the linguistic information in the speech without the help of the middle text representations.

Highly inspired by these previous related works, we design a model to extract features from speech data and train the model in the teacher-student learning manner. In particular, the speech signal is firstly represented as a low-dimensional embedding feature via a speech encoder, then this feature is fed into a conditional generative adversarial network as the condition, and the generator synthesizes the corresponding image with semantic consistency. To the best of our knowledge, our work is the first one to attempt to translate the speech signals into images without the help of text. Compared with the straightforward "two-stage" method, the classifier-based method and the text-to-image models, our method shows better performance than the "two-stage" method and the classifier-based method, even achieves comparable performance to the text-to-image models on the synthesized datasets. Experiments on the real speech data also show the potential for the real application scenarios, like human-computer interaction. etc.

The main contributions of this work can be summarized as follows:

- We propose a framework to translate the speech signals into images directly. Experiments on the synthesized data and real data demonstrate the effectiveness of our proposed framework.
- We train the speech encoder via teacher-student learning that transfers the knowledge in a pretrained image encoder into the speech encoder. Experiments on the synthesized data show that our method can learn the semantic information in the speech descriptions better than the previous classifier-based method, and provide better translation results.
- Ablation study about the loss items, image scales and feature interpolation gives more insights about our method and the speech-to-image translation problem.

The rest of this paper is organized as follows: Section II briefly reviews related works on generative adversarial networks, text-to-image translation, audio-to-image generation, audio-visual correlation learning, teacher-student learning, and direct speech translation, Section III presents our speech-to-image model in detail, and Section IV introduces and analyzes the experimental results on both synthesized and real data, Section V conducts the ablation study. Finally, Section VI concludes this paper.

## II. Related Works

In this section, we review the related works on generative adversarial networks, text-to-image translation, audio-to-image generation, audio-visual correlation learning, teacher-student learning, and direct speech translation.

### A. Generative Adversarial Networks

Generative adversarial networks (GANs) have drawn much attention since it was presented by Goodfellow *et al.* [9] due to its ability to generate high-dimensional data, *e.g.* images. In the model, the generators aim to generate fake data that cannot be separated from the real data, while the discriminators aim to differentiate the generated fake data from the real data. The whole model is optimized via the following loss functions [4], [9]:

$$\min_G \max_D \ \mathbb{E}_{x\sim p_{data}}[\log D(x)] + \mathbb{E}_{z\sim p_z}[\log(1 - D(G(z)))], \tag{1}$$

where $G$ is the generator and $D$ is the discriminator. It is a two-player zero-sum game to arrive a local Nash equilibrium [19], at which neither the discriminator nor the generator can decrease its respective loss. The generator learns a mapping between the noise distribution (*e.g.* the uniform or Gaussian) and the real data (*e.g.* the images or text). When synthesizing images via GANs, attributions [20], text descriptions [10], sketches [21], or images with another style [22] have been used as the conditions to control the appearance of the generated images.

### B. Text-to-Image Translation

Text-to-image translation aims to synthesize images which are semantically consistent with the input text descriptions. It is challenging due to the modality gap between text and images. The computer vision and machine learning community did not pay much attention to this challenging problem until Reed *et al.* [10] used a GAN to synthesize the images conditioned on a low-dimensional representation extracted from the text description. Following Reed's work [10], StackGAN [3] and StackGAN V2 [4] were proposed to generate photo-realistic images up to a resolution of $256 \times 256$ from the text descriptions via a pretrained text encoder [7]. Multi-scale discriminators for increasing resolutions were used in StackGAN and StackGAN V2 to generate images progressively because synthesizing images with a high resolution in one stage had been demonstrated with difficulty [23]. Besides, spatial attention was applied in text-to-image translation [5] by training a multimodal similarity model to calculate the similarity between the word embedding features and the local image features. With text encoder trained by teacher-student learning, these text-to-image translation models generalize well on the new testing classes.

### C. Audio-to-Image Generation

Based on the correlation between audio and images, such as music and instruments, human voices and face appearances, audio-to-image generation aims to generate the images paired with the input audio signals. Chen *et al.* [11] firstly attempted to generate instrument images from the music by leveraging a classifier-based feature extractor and a GAN, but another model is needed if we want to generate music from the image. Hao *et al.* [12] proposed a uniform framework using cycle constraint for the visual-audio mutual generation. Recently,

Oh *et al.* [8] presented a model for generating the face images from a voice using a pretrained face decoder, but the generated results are not sharp due to the concern of privacy. Duarte *et al.* [14] generated sharp face images conditioned on the input speech segmentation using GANs. Different from these previous audio-visual generation works, speech-to-image translation aims to capture the linguistic information in the speech signals and generate images semantically consistent with the input speech descriptions.

### *D. Audio-Visual Correlation Learning*

Audio-visual correlation learning aims to learn a joint embedding feature space over both audio (*e.g.* music, speech, nature sound, *etc.*) and visual (*e.g.* images, videos, *etc.*) data using an embedding alignment model. Based on the prior works on text embedding [24], Harward *et al.* [23] firstly investigated this task to align visual objects and speech signals by a region convolutional neural network (RCNN) [25] and a spectrogram convolutional neural network [26]. Furthermore, Harward *et al.* [27] used vision as an interlingual semantic embeddings of unaligned audios without the use of linguistic transcriptions or conventional speech recognition technologies. Recently, Harward *et al.* [15], [16] operated directly on the image pixels and speech waveforms to associate segments of visual audio captions without relying on any labels, segmentation information or alignment between these modalities. In addition to learning an embedding feature space for speech and images, our work further generates images from the speech embeddings.

### *E. Teacher-Student Learning*

Teacher-student learning is a transfer learning approach, where a pretrained teacher model is used to "teach" a student model [28]. It is widely used in model compression [29], [30] and domain adaption [31]. Reed *et al.* [7] firstly used the GoogLeNet [32] pretrained on ImageNet [33] as the teacher network to learn the deep representation for zero-shot tasks. Following this work, several text-to-image models [3], [4], [10] were proposed for the text-to-image task, based on the same teacher-student learning method to train the text encoder. Recently, teacher-student learning was used to generate the face behind a voice [13]. As a comparison, traditional audio-to-image generation models [11] used a classifier as the feature extractor. In our experiments, we compare the teacher-student learning method with the classifier-based methods and show that the teacher-student learning performs better.

### *F. Direct Speech Translation*

Speech-to-speech translation is one of the most challenging tasks in speech processing and machine learning community, which has tremendous applicable value in our daily life. A speech translation system typically has three components: automatic speech recognition (ASR), machine translation (MT) and text to speech synthesizer (TTS) [34]. Using text as a middle representation to divide this difficult task into three stages have been used for several decades. Recently, with the development of deep learning [35], which shows great potential to model complicated data distribution, researchers have attempted to solve the challenging speech translation without the middle text representation. Bérard *et al.* [36] firstly attempted to build an end-to-end speech-to-text translation system without the text transcription, and showed comparable performance on the synthesized data. Subsequently, Duong *et al.* [37] introduced an attentional model for speech-to-speech translation without speech-to-text transcription, showing the superiority on the low-resources languages. Recently, Jia *et al.* [17] proposed a sequence-to-sequence model to directly translate the speech into another speech, showing comparable performance (only slightly underperform) to a baseline, which cascades of a direct speech-to-text translation model and a text-to-speech synthesis model, on two Spanish-to-English speech translation datasets. These results demonstrated that extracting semantic information from raw speech signals without the middle text representation is practical.

Motivated by the different principles for understanding speech signals, we design a framework to translate the speech signals into images directly, without the help of middle text representation. Specifically, a speech encoder is designed to encode the raw speech signals into a low-dimensional embedding feature. The speech encoder is trained by the teacher-student learning manner via a pretrained image encoder. Subsequently, the speech embedding features are fed into a generator to synthesize images with semantic consistency. Experimental results on both synthesized and real data demonstrate that our proposed model is capable of translating speech into images without the help of middle text representation.

## III. SPEECH-TO-IMAGE MODEL

The modality gap between speech signals and image signals makes it not feasible to directly regress the pixel value from speech signals. Inspired by the common text-to-image architectures [3]–[5], [10], we design a speech encoder to encode the speech signals into a low-dimensional embedding feature. Then this embedding feature is used to synthesize the corresponding images with semantic consistency, The diagram of the proposed algorithm is illustrated in Fig. 2.

### *A. Speech Encoder*

The input raw speech signal is first represented as a time-frequency spectrogram, then it is encoded by our speech encoder into a low-dimensional embedding feature with convolutional neural networks (CNNs) and recurrent neural networks (RNNs). Typically, the speech spectrogram is modeled via an RNN in ASR [38]–[40], however, the long input spectrogram of our model may not convinent enough for RNN model optimization [41]. Inspired by the character-based text embedding architecture [7] and audio-visual cross-modal embedding learning [16], a multi-layer CNN is inserted before the RNN to reduce the signal length, as shown in Fig. 2. Given a speech signal $s_i$ and its spectrogram $s'_i$, the speech embedding $f_{s_i}$ can be obtained:

$$f_{s_i} = E_s(s'_i) = \text{RNN}(\text{CNN}(s'_i)), \tag{2}$$

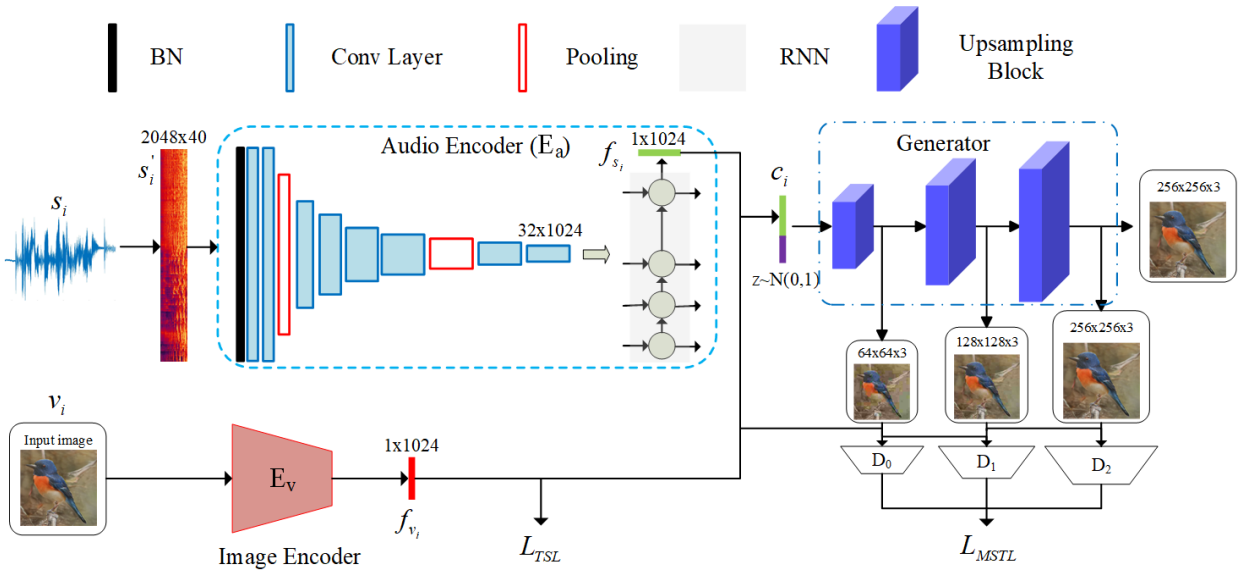


Fig. 2. Illustration for our speech-to-image model: a speech encoder, trained via teacher-student learning, is designed to extract a feature representation from the input speech. After encoding the raw speech signals into a low-dimensional embedding, our model synthesizes images at a resolution of $256 \times 256$ with semantic consistency from the embedding feature. In the figure, $TSL$ denotes teacher-student loss for the speech encoder and $MSTL$ denotes multi-scale triple loss for the generators.

where $E_s$ denotes our speech encoder. The input time-frequency spectrogram is firstly normalized along with frequency in the first layer of the CNN, then the output of the CNN is subsequently encoded as a 1024 dimensional embedding feature by an RNN. The input length of the RNN can be variable, with the only constraint that the input length of the CNNs should be longer than 64 because the model shortens the input sequence by 64 times.

### B. Teacher-Student Learning

The optimization for our speech encoder is not trivial because the ground truth of the speech embedding is not available. Although the class label of the speech description is accessible, there might be generalization problem when the trained model is tested on the new unseen data (new class respect to the training set). Inspired by the cross-modal generation models [3], [7], [13], [16], [23], in this work, we use teacher-student learning [30] to overcome this problem to some extent.

Given an image and its speech description, $(v_i, s_i)$, an image encoder and a speech encoder are used to represent the image/speech as a low-dimensional embedding feature, respectively:

$$f_{v_i} = E_v(v_i) \tag{3}$$

$$f_{s_i} = E_s(s_i'), \tag{4}$$

where $s_i'$ is the time-frequency spectrogram of $s_i$. $E_v$/$E_s$ denotes the image/speech encoder, $f_{v_i}$/$f_{s_i}$ is the low-dimensional embedding feature of the input image/speech. It is worth mentioning that the image encoder is pretrained on a large dataset, such as ImageNet [33], and it is fixed when training the speech encoder. In our model, the pretrained image encoder is the "teacher", while the speech encoder is the "student". The goal of teacher-student learning is to optimize the student model to learn a similar feature space with the teacher model. So the dimension of the student model's feature space should be the same as the teacher's. In our model, GoogLeNet [32] is used as the teacher model to represent the input image with a resolution of $256 \times 256$ as a 1024 dimensional feature. As a result, our speech encoder also encodes the input time-frequency spectrogram into a feature with 1024 dimensions.

### C. Generative Network

The generative network is used to synthesize images conditioned on the speech embedding feature. Following the recent works about text-to-images [3]–[5], we use a stacked conditional GAN, also known as StackGAN v2 [4], to synthesize the images due to its promising performance on generating photo-realistic images. As illustrated in Fig. 2, three branches are used in the generator to synthesize images with a resolution of $256 \times 256$, and three discriminators are used to distinguish the generated images with a resolution of $64 \times 64$, $128 \times 128$, $256 \times 256$ from the real images, respectively. Each upsampling block in the generator contains an upsampling layer and two residual blocks [42] to synthesize details based on the input low-resolution image. Given the condition $c_i$ and the input noise $z_i$ which samples from the Gaussian distribution, the generator synthesizes the fake images:

$$v_{if} = G(c_i, z_i), \tag{5}$$

where $v_{if}$ denotes the synthesized fake images, $G$ denotes the stacked generative network. In our model, the embedding feature of the input speech ($f_{s_i}$ in Fig. 2) is used as the condition $c_i$ for the generator.

### D. Training

The whole model is trained with two steps. Firstly, the speech encoder is optimized with the image encoder. Secondly, the generator is trained conditioned on the speech embedding representations, which are extracted by the pretrained speech encoder.

*1) Training the Speech Encoder:* The speech encoder is trained via teacher-student learning [30], which transfers knowledge from a large model into a small model. In our model, the knowledge in the pretrained image encoder needs to be transferred into our speech encoder. We can optimize the norm to train the speech encoder, however, training student network with the norm optimization alone is slow and unstable [13]. To stabilize and accelerate the training, additional loss items are introduced. Taking inspiration from the text-image embedding learning [7] and audio-to-image generation [13] models, norm loss, jointly embedding loss (JEL), and knowledge distilling loss (KDL) are used in our teacher-student loss (TSL) for training the speech encoder. Given a batch of triplet data $(v_i, s_i, y_i)$, the spectrogram $s'_i$, the image encoder $E_v$ and the speech encoder $E_s$, the objective function between image and speech encoder is defined as:

$$\mathcal{L}_{TSL_i} = \mathcal{L}_{JEL_i} + \lambda_{norm}\mathcal{L}_{norm_i} + \lambda_{KDL}\mathcal{L}_{KDL_i} \tag{6}$$

$$\begin{aligned}\mathcal{L}_{JEL_i} &= \alpha\mathbb{E}_{y_j \neq y_i}[\max(0, f_{s_i}^T f_{v_j} - f_{s_i}^T f_{v_i} + m_{diff})] \\ &+ \beta\mathbb{E}_{y_j = y_i}[\max(0, f_{s_i}^T f_{v_j} - f_{s_i}^T f_{v_i} + m_{same})]\end{aligned} \tag{7}$$

$$\mathcal{L}_{norm_i} = \Sigma_k |f_{s_{ik}} - f_{v_{ik}}| \tag{8}$$

$$\mathcal{L}_{KDL_i} = \mathbb{KL}(softmax(f_{s_i})|softmax(f_{v_i})), \tag{9}$$

where $y_i$ is the class label for $v_i$ and $s_i$, $\lambda_{norm}, \lambda_{KDL}$ are hyper-parameters for fusing the three items. Following [13], $\lambda_{norm}$ and $\lambda_{KDL}$ are tuned to make the gradient magnitudes of the three items with respect to $f_{s_i}$ be with a similar scale at an early training iteration. In $\mathcal{L}_{JEL}$, $m_{diff}/m_{same}$ is the margin for $(v_i, s_i)$ pairs with different/same class label in a batch data, respectively. $\alpha/\beta$ is set to control the inter/intra class distance, respectively. Here, we optimize the 1-norm in $\mathcal{L}_{norm}$. $f_{s_i} = E_s(s'_i)$ is the embedding feature of the input speech description. $f_{v_i} = E_v(v_i)$ is the embedding representation of the input image. In $\mathcal{L}_{KDL}$, $softmax(x_i) = e^{x_i}/\Sigma_j e^{x_j}$.

*2) Training the Generator:* Following [4], the generator is trained by a multi-scale triplet loss (MSTL), which includes three items in each discriminator's loss: conditional item, unconditional item and wrong pair item. Given a triplet $(c_i, v_i, v_i^w)$, where $v_i^w$ denotes the wrong image which belongs to the different classes with $v_i$ :

$$\mathcal{L}_D = \Sigma_{s=1}^3(\mathcal{L}_{D_s}^{cond} + \mathcal{L}_{D_s}^{uncond} + \mathcal{L}_{D_s}^{wrong}) \tag{10}$$

$$\mathcal{L}_{D_s}^{cond} = \mathbb{E}_{(v_i,c_i)\sim p}[D_s(v_i, c_i) + (1 - D_s(G(z, c_i), c_i))] \tag{11}$$

$$\mathcal{L}_{D_s}^{uncond} = \mathbb{E}_{(v_i,c_i)\sim p}[D_s(v_i) + (1 - D_s(G(z, c_i)))] \tag{12}$$

$$\mathcal{L}_{D_s}^{wrong} = \mathbb{E}_{(v_i,c_i)\sim p}[(1 - D_s(v_i^w, c_i))] \tag{13}$$

$$\mathcal{L}_G = \Sigma_{s=1}^3\mathbb{E}_{(v_i,c_i)\sim p}[D_s(G(z, c_i)) + D_s(G(z, c_i), c_i)], \tag{14}$$

where $p$ is the data distribution for the pair $(v_i, c_i)$, $\mathcal{L}_D$ and $\mathcal{L}_G$ both contain three scales. The loss for each discriminator has three items to model both conditional and unconditional distribution. By contrast, the traditional conditional GAN [43] only contains the conditional item $\mathcal{L}_{D_s}^{cond}$, without taking the unconditional distribution into account.

### E. Inference

In the inference phase, the time-frequency spectrograms of the input speech descriptions are encoded as low-dimensional embedding features by the speech encoder. Subsequently, the generator synthesizes images with a resolution of $256 \times 256$ conditioned on the embedding features. The inference can be denoted as:

$$v_{if} = G(z, E_s(s'_i)), \tag{15}$$

where $z$ is the noise vector, $s'_i$ is the spectrogram of the input speech, $v_{if}$ is the synthesized images semantically consistent with the input speech descriptions.

### F. Implementation Details

The raw speech signals are represented as log Mel filter bank spectrograms, following [16]. Specifically, the DC component of each audio is removed via mean subtraction, followed by the pre-emphasis filtering and the short-time Fourier transform (STFT) computation by using a 25 ms Hamming window with 10 ms shift. Then the squared magnitude spectrum of each frame is taken into consideration and the log energies with each of 40 Mel filter bands are computed. As a result, the spectrograms with shape $(\text{band}, \text{frame})$, where the band here is $40$ with variable frame number, can be obtained. When training the speech encoder, we use GoogLeNet [32] as the image encoder. As for the hyperparameters, we set $\lambda_{norm} = 5, \lambda_{KDL} = 1000, m_{diff} = 1, m_{same} = 0.1$ after tuning them to balance the gradients respect to the speech embedding feature.

## IV. Experiments Results and Ayalyses

In this section, we verify the effectiveness of the proposed model for translating speech signals into images without middle text representations and show how well our model can achieve. Firstly, we generate images from the synthesized speech data to demonstrate the effectiveness of our model and compare our model with the straightforward "two-stage" method, traditional classifier-based method, and text-to-image models. Secondly, experiments on real data are conducted to explore our model's robustness to the real noise and the potential for the real application scenarios. Finally, ablation study about the different loss items, image scales, and feature interpolation gives more insights about our model.

### A. Datasets and Metrics

**Datasets.** Several datasets, like Places 205 dataset [44], Caltech-UCSD Birds 200-2011 (CUB-200) [45], Oxford Flower with 102 categories (Oxford-102) [46] and Microsoft COCO [47], are used in the audio-visual correlation learning or text-to-image translation. Harwath *et al.* [16] used the Places 205 dataset with speech descriptions [15], [48] to

learn the association between spoken audio caption segments and nature images portions. Reed *et al.* [7] used CUB-200 to learn deep representations of visual text descriptions. Most investigations about text-to-image translation [3], [4], [10] conducted their experiments on CUB-200, Oxford-102, or COCO dataset. However, these datasets only contain the text descriptions and do not have any speech label, making us difficult to leverage the dataset off-the-shelf to conduct the experiments. Fortunately, with the development of text-to-speech generation [49]–[51] technologies based on deep learning and large-scale labeled speech datasets [52], [53], we can use some mature commercial text-to-speech systems like Baidu TTS eigine[1] or Microsoft Bing TTS[2] to synthesize large-scale high-quality speech from the text descriptions. The synthesized speech data are continuous and unaligned, just like real speech, although they have no background noise and speaker variance.

To compare our model with the text-based models, we use CUB-200 [45] and Oxford-102 [46] dataset to test the performance of our model and synthesize the speech descriptions from the text descriptions via Baidu TTS. As illustrated in Table I,CUB-200 [45] dataset contains 11788 bird images classified into 200 categories, with one bounding box and 10 sentences text descriptions per image. Oxford-102 [46] dataset has 8189 flower images classified into 102 classes. The speech descriptions are synthesized by Baidu TTS engine from the text captions in both datasets. Following [4], we crop all the images to ensure that bounding boxes have greater-than-0.75 object-image-size ratios for CUB-200 dataset. We also conduct experiments on a subset of Places 205 dataset [15], [48] with real speech descriptions to explore the potential for real applications.

**Evaluation Metrics.** Generally, it is difficult to fairly evaluate the generative models. As in [3], [4], [10], we use inception score [54] and Fréchet inception distance [55], [56] to quantitatively evaluate our models. They are formulated as,

$$IS = \exp(\mathbb{E}_x \mathbb{KL}(p(y|\mathbf{x})|p(y))) \tag{16}$$

$$FID = ||m_1 - m_2||_2^2 + \mathrm{Tr}(C_1 + C_2 - 2(C_1C_2)^{\frac{1}{2}}). \tag{17}$$

Inception score (IS) [54] is a metric for both image quality and diversity, which is found correlating well with the human evaluation. The conditional label distribution $p(y|\mathbf{x})$ measures the quality of the generated images, and the images are high-quality when the distribution is with low entropy. The marginal distribution $p(y)$ measures the diversity, and the generated images are more diverse when the distribution is with high entropy. By jointly considering them two together, the KL-divergence of $p(y|\mathbf{x})$ and $p(y)$ provides the evaluation for both image quality and its diversity. When calculating IS in our experiments, we use the Inception-v3 model finetuned on CUB-200 or Oxford-102 dataset following [4]. Fréchet inception distance (FID) [56] measures the distance between the generated and real data. Lower FID means higher similarity for the generated and real data distribution. To calculate the FID between the generated images and the real images, we use all speech labels in the testing set to generate a large number images (*e.g.* 30k for CUB-200, 11k for Oxford-102). When calculating FID in our experiments, we use the Inception-v3 model pretrained on the ImageNet dataset [33] following [56].

[1]https://cloud.baidu.com/product/speech/tts

[2]https://azure.microsoft.com/en-us/services/cognitive-services/text-to-speech/

| Dataset | | Training set | Testing set | Total |
|---|---|---|---|---|
| CUB-200 [45] | class | 150 | 50 | 200 |
| | image | 8855 | 2933 | 11788 |
| Oxford-102 [46] | class | 82 | 20 | 102 |
| | image | 7034 | 1155 | 8189 |

TABLE I
TRAINING/TESTING SET OF CUB-200 [45] AND OXFORD-102 [46] DATASETS.

### B. Experimental Results on Synthesized Data

To demonstrate the effectiveness of our model, experiments are conducted on the synthesized speech data, including CUB-200 [45] and Oxford-102 [46] dataset. The synthesized speech data are continuous and unaligned, although they are well-normalized and less noisy. The CUB-200 dataset contains 200 classes totally, where 150 classes are used to train our model and the rest are used as the testing set. The Oxford-102 dataset contains 102 classes totally, where 82 classes are used for training and 20 classes for testing. The statistical information for training/testing split of these two datasets is shown in Table I. The splitting manner for the training set and testing set follows [7]. The qualitative and quantitative results are illustrated in Fig. 3 and Table II, respectively. Some conclusions can be drawn from the results:

*1) Our model can synthesize images semantically consistent with the input speech:* qualitative results on CUB-200 testing set and Oxford-102 testing set are shown in Fig. 3. For each row, we show the waveform of the input speech description (left) and 8 synthesized images (right) with different input noises conditioned on the same input speech embedding. It is worth mentioning that the transcription results are shown only for readability, and they are not used in our model. The results show that our model can generate realistic images from the input speech description although the input speech is continuous and unaligned. Moreover, the visual information shown in the synthesized images is mostly accordant with the semantic information in the speech descriptions. The generated results from CUB-200 dataest [45] (row 1∼5 in Fig. 3) show that the input speech description controls the color of the generated bird's different parts, such the head, feather, tail, etc. In comparison, the background, gesture, even shape of the generated bird changes when the input noise is different. This means that our model has disentangled the bird's color from the background and gestures to some extent. Similar conclusion also can be drawn from the results of Oxford-102 [46] dataset (row 5∼10 in Fig. 3). The input speech description mainly controls the color of the generated flowers, and the background, size, even the categories of the generated flowers are different when the input noises change. This demonstrates that our model has learned the semantic information in the input speech descriptions to some extent and visualized the semantic information onto the images.


Fig. 3. Qualitive results for CUB-200 [45] testing set (row 1∼5) and Oxford-102 [46] testing set (row 6∼10) (synthesized speech data). Left: the waveforms of the input speech descriptions and its transcription results. Right: 8 images with a resolution of $256 \times 256$ synthesized by our model conditioned on the left input speech description and a random noise vector. **Note that the text is shown only for readability, and it is not used in our model.** Some interesting conclusions can be drawn from the figure: (1) the color of the birds/flowers are consistent with the input speech descriptions and they do not change with the input noise. (2) when fixing the speech descriptions, the backgrounds, gestures (for birds), even the categories (such as row 6) of the generated images change along with the input noise.

*2) Our model is better than the "two-stage" model with text:* To compare our "one-stage" method (translating speech to images without text) with the "two-stage" method (using text as middle representation), we train the "two-stage" model on CUB-200 and Oxford-102 datasets. In our experiments, we use a pretrained ASR model, DeepSpeech[3] [57], [58], to transcribe the speech into text and use the (text, image) paired data to train a text encoder and a generator following [4]. For fair comparison, the generator and the hyper-parameters for its training in the "two-stage" method are the same as our model's. The results of the "two-stage" method on CUB-200/Oxford-102 are shown in the 1st/6th row of Table II, respectively. The "two-stage" model is slightly inferior to our "one-stage" model on both datasets. On the CUB-200 dataset [45], our "one-stage" model performs better by $1.52$ on FID although the IS score of the "two-stage" method is comparable to our model's. Similarly, on the Oxford-102 dataset [46], our "one-stage" method surpasses the "two-stage" method on FID ($54.76$ vs $57.73$), while IS scores of the both models are the same ($3.23$ vs $3.23$). We think the reason why the "two-stage" method performs worse is the word errors in speech recognition.

It is worth mentioning that DeepSpeech might be not a strong baseline in our experiment although DeepSpeech has millions of parameters and is trained on a large dataset, because our speech data are synthesized while the training data of DeepSpeech are real speech data. The word error rates (WERs) of DeepSpeech on our datasets are above $50\%$, showing that DeepSpeech does not perform well on our synthesized speech data. To address this problem, in the following, we compare our proposed model with text-to-image models, which can be viewed as "two-stage" frameworks with an ideal ASR model (WER is $0\%$), to evaluate the performance of our model.

*3) Our method is comparable to the text-to-image methods:* To compare with the upper bound of the "two-stage" method, we compare our method with the text-to-image methods, which can be seen as "two-stage" methods with a perfect ASR model. Two text-to-image methods [3], [4], which use similar structures with our model, are used in the experiments, as shown in Table II. On CUB-200 [45] dataset, our model is slightly inferior to StackGAN-v2 [4] on FID ($18.37$ vs $15.37$), but performs better on IS ($4.09$ vs $4.04$). We think that the reason is that our model generates more diverse results due to the speech signal is higher-dimensional than text signals. Compared with StackGAN-v1 [3], our model performs better with a large gain because we use a stronger generator. On the Oxford-102 dataset [46], our proposed model surpasses the StackGAN v1 on both FID ($54.76$ vs $55.28$) and IS ($3.23$ vs $3.20$), although it only surpasses the StackGAN v2 slightly.

This result has demonstrated that our model performs closely with the upper bound of the "two-stage" models, and our speech encoder has extracted the semantic information in the input speech descriptions into the embedding feature, which is subsequently used as the condition in the generator, just as the text-to-image model in [4].

[3]https://github.com/mozilla/DeepSpeech

| Dataset | Method | Type | IS ↑ | FID ↓ |
|---|---|---|---|---|
| CUB-200 [45] | two-stage | speech | $4.04 \pm .04$ | $20.85$ |
| | classifier-based | speech | $3.68 \pm .04$ | $43.76$ |
| | Ours | speech | $\mathbf{4.09 \pm .04}$ | $\mathbf{18.37}$ |
| | StackGAN-v1 [3] | text | $3.70 \pm .04$ | $51.89$ |
| | StackGAN-v2 [4] | text | $4.04 \pm .05$ | $15.03$ |
| Oxford-102 [46] | two-stage | speech | $3.23 \pm .06$ | $57.73$ |
| | classifier-based | speech | $\mathbf{3.30 \pm .06}$ | $64.75$ |
| | Ours | speech | $3.23 \pm .05$ | $\mathbf{54.76}$ |
| | StackGAN-v1 [3] | text | $3.20 \pm .01$ | $55.28$ |
| | StackGAN-v2 [4] | text | $3.26 \pm .01$ | $48.68$ |

TABLE II
QUANTITATIVE RESULTS OF OUR MODEL ON CUB-200 [45] DATASET AND OXFORD-102 [46] DATASET. WE COMPARE OUR MODEL WITH "TWO-STAGE" METHOD, CLASSIFIER-BASED METHOD [11], AND TEXT-TO-IMAGE MODELS [3], [4]. AS ILLUSTRATED IN THE TABLE, OUR "ONE-STAGE" MODEL PERFORMS BETTER THAN THE "TWO-STAGE" MODEL AND THE CLASSIFIED-BASED MODEL, EVEN COMPARABLY TO THE TEXT-TO-IMAGE MODELS.

*4) Our teacher-student learning method is better than the classifier-based method:* To compare our teacher-student learning method with previous classifier-based methods [11], we train our speech encoder via the classified-based method rather than the teacher-student learning. Specifically, a classifier layer is added after the speech encoder and the speech encoder is trained with the cross entropy loss. To compare fairly, the classifier-based model uses the same feature extractor structure as our model's. In particular, we use the speech encoder in our model as a feature extractor and add a linear layer as the classifier, then train this classifier-based model on the training set. When testing, only the feature extractor is used. By this way, the parameters size and the hyperparameters for classified-based method and our proposed method are the same. The only difference is the training method. The results are illustrated in the 2nd and 7th row of Table II. On the CUB-200 [45] dataset, our model performs rather better than the classifier-based method on both IS ($4.09$ vs $3.68$) and FID ($18.37$ vs $43.76$), indicating that our model's better generalization ability when using the same structure and parameters. On the Oxford-102 [46] dataset, FID score ($54.76$ vs $64.75$) shows our model is better, however, the IS score ($3.23$ vs $3.30$) draws different conclusions. IS takes both the image quality and diversity into account, so these results indicate that our model generates data closer to the real data but less diverse when compared with the classifier-based method on the Oxford-102 dataset. Consequently, our proposed method shows better FID and comparable IS on both CUB-200 [45] and Oxford-102 [46] dataset due to its teacher-student learning method when compared with the classifier-based method.

### C. Experimental Results on Real data

In addition to the synthesized speech data, we also evaluate our model on the real speech data to assess the potential for real applications. Places Audio Captions dataset [16], [48], collected via Amazon Mechanical Turk (AMT), is a real speech dataset for visual descriptions for Places 205 dataset [44].We use a subset of this dataset, which includes


Fig. 4. Experimental results on Placces-Subset (real speech data). Left: the waveforms of the input speech. Right: the synthesized images with a resolution of $128 \times 128$ conditioned on the left input speech. **Note that the text here is shown only for readability, and they are not used in our model.** The errors in the text come from the recognition error.

| Dataset | Method | Type | FID↓ |
|---|---|---|---|
| Places-Subset | two-stage | speech | 64.59 |
| | classifier-based | speech | 232.39 |
| | Ours | speech | 83.06 |

TABLE III
QUANTITATIVE RESULTS OF OUR MODEL ON PLACES-SUBSET DATASET.

13803 paired data with 7 classes[4] (Places-Subset), to evaluate the robustness for the real data of our model. The testing set contains 2870 images, which are randomly selected from the dataset. It is difficult to generate the details for high-resolution images due to the dataset's diversity and the variance between different speakers, so we generate images with a resolution of $128 \times 128$. Some sampled results are shown in Fig. 4. Although the details are not sharp due to the low resolution, we can also see that the color and the scene of the generated images are semantically consistent with the input speech descriptions, which means the model has captured the semantic information in the input speech descriptions to some extent. In addition to visualization examples, we also evaluate the result with objective matrics. Our model achieves 83.06 for FID, as shown in Table III (IS is not used because no finetuned Inception model is available for Places-Subset). Our method is rather better than the classifier-based method (83.06 vs 232.39), demonstrating the effectiveness of the teacher-student learning on the real data. Besides, different from the synthesized datasets, our model performs not as well as the "two-stage" method (83.06 vs 64.59), because the real data are much challenging than the synthesized data to extract the speech semantic feature for our model. These results are encouraging and show the potential for our model to the real scenario, such as human-computer interactions and computer-aided design.

[4]7 classes in Places 205: bedroom, dinette, dining room, home office, hotel room, kitchenette, living room

| Dataset | $\mathcal{L}_{L_{norm}}$ | $\mathcal{L}_{JEL}$ | $\mathcal{L}_{KDL}$ | IS↑ | FID↓ |
|---|---|---|---|---|---|
| CUB-200 [45] | ✓ | | | $3.71 \pm .05$ | 27.45 |
| | ✓ | ✓ | | $4.02 \pm .04$ | **16.75** |
| | ✓ | ✓ | ✓ | $\mathbf{4.11 \pm .05}$ | 18.70 |
| Oxford-102 [46] | ✓ | | | $2.81 \pm .04$ | 66.84 |
| | ✓ | ✓ | | $3.22 \pm .03$ | 58.19 |
| | ✓ | ✓ | ✓ | $\mathbf{3.23 \pm .05}$ | **57.11** |

TABLE IV
ABLATION STUDY RESULTS FOR DIFFERENT LOSS ITEMS OF THE SPEECH ENCODER ON CUB-200 [45] AND OXFORD-102 [46] DATASETS.

| Dataset | Scale | IS ↑ | FID ↓ |
|---|---|---|---|
| CUB-200 [45] | $64 \times 64$ | $3.45 \pm .04$ | 70.18 |
| | $128 \times 128$ | $3.98 \pm .04$ | 28.25 |
| | $256 \times 256$ | $\mathbf{4.11 \pm .05}$ | **18.70** |
| Oxford-102 [46] | $64 \times 64$ | $2.82 \pm .04$ | 70.95 |
| | $128 \times 128$ | $\mathbf{3.27 \pm .06}$ | 63.35 |
| | $256 \times 256$ | $3.23 \pm .05$ | **57.11** |

TABLE V
ABLATION STUDY RESULTS FOR DIFFERENT SCALES FOR THE GENERATOR ON CUB-200 [45] AND OXFORD-102 [46] DATASETS.

## V. ABLATION STUDY

In this section, we conduct ablation study for our model to analyze the different components of the proposed model and the influence of some hyper-parameters as well as the embedding feature space. To compare fairly, we train the models from the scratch for the same iterations with the same hyperparameters except for the specified hyperparameter, such as the loss items for the speech encoders, or the scale of the synthesized images. Specifically, the speech encoders are trained for 100 epochs for both dataset, and the generators are trained for 220k/100k iterations for CUB-200 [45]/Oxford-102 [46] dataset, respectively. The training iterations is set to avoid overfitting or model collapse.

### A. The Loss Function

In the training of our speech encoder, teacher-student loss (TSL) (Eq. 6), including $\mathcal{L}_{L_{norm}}$, $\mathcal{L}_{JEL}$ and $\mathcal{L}_{KDL}$, is used to stabilize and accelerate the training. The experiments are conducted on CUB-200 [45] and Oxford-102 [46] datasets to study how these items influence the final result. To compare fairly, the training hyper-parameters are all set the same except the loss items. The same structure and parameters are used in the generator training, and the generator is with three branches to generate images with a resolution of $256 \times 256$. The weight for $\mathcal{L}_{L_{norm}}/\mathcal{L}_{JEL}/\mathcal{L}_{KDL}$ is set as $5/1/1000$, respectively, to ensure the gradient to the speech embedding is within the similar scale (following [13]). Table IV lists the results on the testing sets of CUB-200 dataset [45] and Oxford-102 dataset [46]. When the only $\mathcal{L}_{L_{norm}}$ is used to train the speech encoder (1st row and 4th row in Table IV), the model performs the worst on both datasets, demonstrating that $L_{norm}$ is not enough to obtain good performance. $\mathcal{L}_{JEL}$ improves

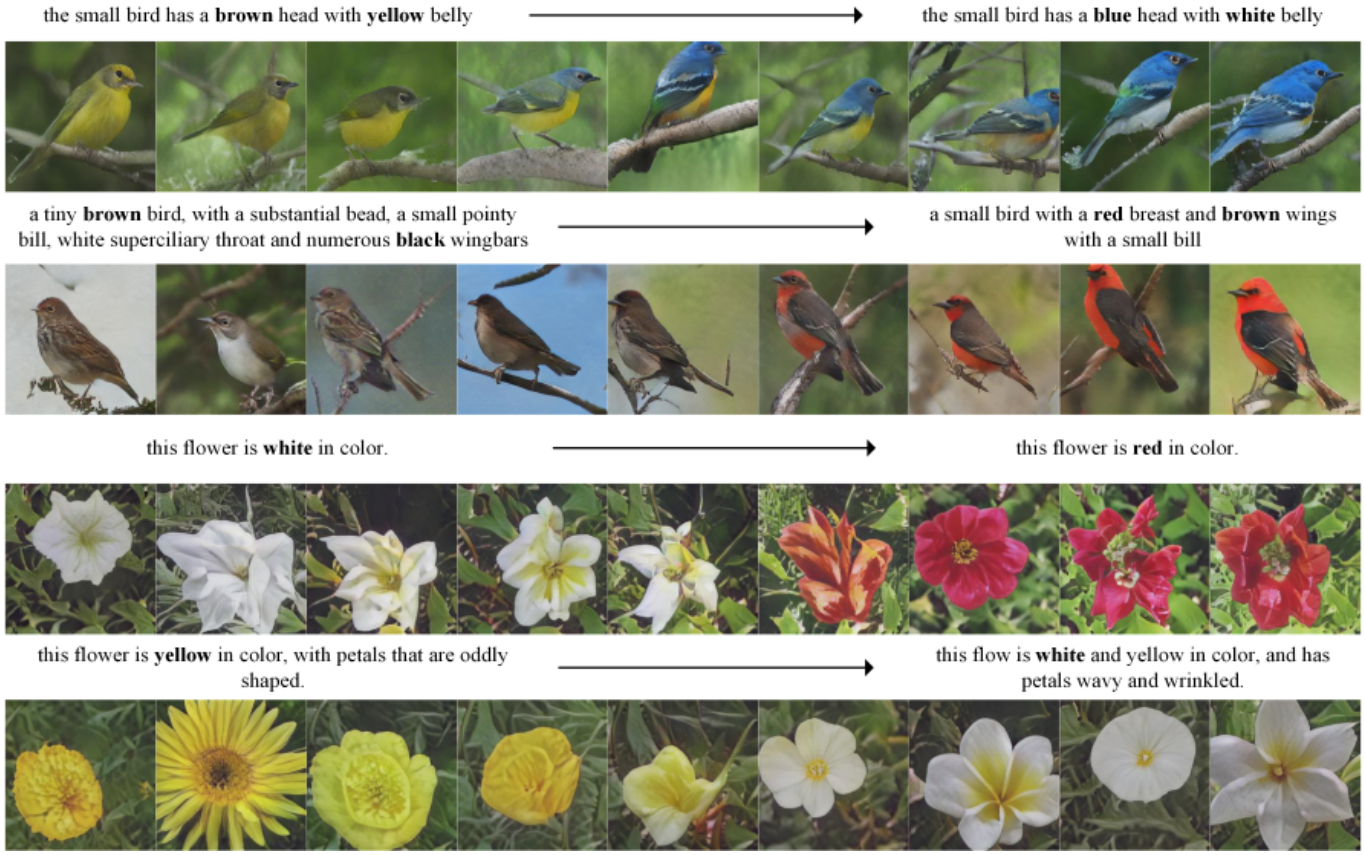


Fig. 5. Feature interpolation results on CUB-200 [45] (row 1-2) and Oxford-102 [46] (row 3-4) dataset. **Note that the inputs are the speech descriptions without the text representations, the text is shown only for readability.** As illustrated in the figure, the generated images are semantically consistent with the input speech descriptions. Moreover, from left to right on each row, the color of the birds/flowers transits gradually due to the feature interpolation, which indicates that our model learns a linear semantic feature space.

the model with a big margin on both datasets (2nd row and 5th row in Table IV), even achieves the best FID on CUB-200 dataset (2nd row in Table IV). As a comparison, $\mathcal{L}_{KDL}$ only boosts the model slightly on Oxford-102 dataset (6th row in Table IV) and even leads an FID drop on CUB-200 dataset (3th row in Table IV). In general, both $\mathcal{L}_{JEL}$ and $\mathcal{L}_{KDL}$ can improve the model more or less.

### *B. Different Scales*

Higher resolution provides us more details, making the generated images more realistic. However, higher resolution increases the complexity of the model, making the model unstable. The scale of the generated images affects the performance of our model in different aspects. So in this subsection, we conduct experiments to study the influences of different scales. Experiments are conducted on CUB-200 dataset [45] and Oxford-102 dataset [46]. All the experiments use the same training hyperparameters and network structure, except for the resolution of the generated images and the discriminator number for the different scales. Specifically, all the experiments use the same pretrained speech encoder to extract embedding features of the speech descriptions. The generator to synthesize images with a resolution of $64^2/128^2/256^2$ uses $1/2/3$ discriminators to model the data distribution, respectively. IS [54] and FID [56] are used to evaluate the performance of the generator. As illustrated in Table V, based on the tree-like structure [4], the higher resolution our model synthesizes, the better performance we can achieve. The only exception is the model for the resolution of $128 \times 128$ obtains the best IS on Oxford-102 dataset (5th row in Table V), however, it only performs better within $1.5\%$ (3.27 vs 3.23) than the model for the resolution of $256 \times 256$. Taking the FID into account, we can still conclude that higher resolution leads to better performance. The similar conclusion is also drawn in the text-to-image translation [3], [4].

### *C. Feature Interpolation*

To study the feature space derived from our speech encoder further, we conduct experiments of feature interpolation on CUB-200 [45] and Oxford-102 [46] datasets to show that the feature space learned by our speech encoder is a linear space to some extent. Specifically, given two embedding features $f_{s_1}, f_{s_2}$, 9 features are sampled by combining $f_{s_1}$ and $f_{s_2}$ linearly: $f_{s_i} = \alpha_i f_{s_1} + (1 - \alpha_i) f_{s_2}, \alpha = 0, 1/8, 2/8, \ldots, 1$. Then these embedding features are fed into the generator to study the semantic transition from $f_{s_1}$ to $f_{s_2}$. As illustrated in Fig. 5, in the first row, the color of the small bird's back shows a smoothing transition from brown to blue, while the color of

belly changes from yellow to white, which exactly shows the semantic described in the input speech descriptions. In the second row, the breast of the small bird changes from brown to red smoothly, just as described in the input speech. The third row and the fourth row are from Oxford-102. Similar to the first two rows, most of the flowers are realistic and the color of the flower transits gradually from the left to the right, although there are some artifacts in some images. The results of feature interpolation verify that our model has learned a linear embedding feature space to some extent.

## VI. Conclusions

In this paper, we have described a new framework to translate the speech signals into the images without the help of middle text representation. We addressed this problem by extracting a low-dimensional embedding feature from the speech descriptions and synthesizing images from this feature via a stacked GAN. We have demonstrated that our proposed model can synthesize images semantically consistent with the input speech description on both synthesized and real data. Moreover, our model performed better than the "two-stage" method and the classifier-based method, even achieved comparable performance to the text-to-image models on the synthesized datasets. We believe that synthesizing images from speech signals without text is a new perspective to understand the semantic information in the speech signals and can open up new research directions.

## VII. Acknowledgment

This work was supported by the National Natural Science Foundation of China and Royal Society (61961130392), National Natural Science Foundation of China (61632001), which are gratefully acknowledged.

## References

[1] E. Bergelson and D. Swingley, "At 6–9 months, human infants know the meanings of many common nouns," *Proceedings of the National Academy of Sciences*, vol. 109, no. 9, pp. 3253–3258, 2012.

[2] D. G. Thomas, J. J. Campos, D. W. Shucard, D. S. Ramsay, and J. Shucard, "Semantic comprehension in infancy: A signal detection analysis," *Child development*, pp. 798–803, 1981.

[3] H. Zhang, T. Xu, and H. Li, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017, pp. 5908–5916.

[4] 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*, pp. 1–1, 2018.

[5] 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[6] D. Tannen, *Spoken and written language: Exploring orality and literacy*. ABLEX Publishing Corporation, 1982, vol. 32.

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

[8] T.-H. Oh, T. Dekel, C. Kim, I. Mosseri, W. T. Freeman, M. Rubinstein, and W. Matusik, "Speech2face: Learning the face behind a voice," *arXiv preprint arXiv:1905.09773*, 2019.

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

[10] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *33rd International Conference on Machine Learning*, 2016, pp. 1060–1069.

[11] L. Chen, S. Srivastava, Z. Duan, and C. Xu, "Deep cross-modal audio-visual generation," in *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*. ACM, 2017, pp. 349–357. [Online]. Available: https://arxiv.org/pdf/1704.08292.pdf

[12] W. Hao, Z. Zhang, and H. Guan, "Cmcgan: A uniform framework for cross-modal visual-audio mutual generation," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[13] T.-H. Oh, T. Dekel, C. Kim, I. Mosseri, W. T. Freeman, M. Rubinstein, and W. Matusik, "Speech2face: Learning the face behind a voice," *arXiv preprint arXiv:1905.09773*, 2019.

[14] 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)*, vol. 3, 2019.

[15] D. Harwath and J. Glass, "Learning word-like units from joint audio-visual analysis," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, 2017, pp. 506–517.

[16] D. Harwath, A. Recasens, D. Suris, G. Chuang, A. Torralba, and J. Glass, "Jointly discovering visual objects and spoken words from raw sensory input," in *The European Conference on Computer Vision (ECCV)*, September 2018.

[17] Y. Jia, R. J. Weiss, F. Biadsy, W. Macherey, M. Johnson, Z. Chen, and Y. Wu, "Direct speech-to-speech translation with a sequence-to-sequence model," *arXiv preprint arXiv:1904.06037*, 2019.

[18] K. Audhkhasi, A. Rosenberg, A. Sethy, B. Ramabhadran, and B. Kingsbury, "End-to-end asr-free keyword search from speech," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1351–1359, Dec 2017.

[19] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems*, 2017, pp. 6626–6637.

[20] R. Huang, S. Zhang, T. Li, and R. He, "Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2439–2448.

[21] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[22] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[23] D. Harwath and J. Glass, "Deep multimodal semantic embeddings for speech and images," in *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*. IEEE, 2015, pp. 237–244.

[24] R. Socher and L. Fei-Fei, "Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 966–973.

[25] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[26] S. Bengio and G. Heigold, "Word embeddings for speech recognition," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[27] D. Harwath, G. Chuang, and J. Glass, "Vision as an interlingua: Learning multilingual semantic embeddings of untranscribed speech," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 4969–4973.

[28] V. Manohar, P. Ghahremani, D. Povey, and S. Khudanpur, "A teacher-student learning approach for unsupervised domain adaptation of sequence-trained asr models," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 250–257.

[29] J. Ba and R. Caruana, "Do deep nets really need to be deep?" in *Advances in neural information processing systems*, 2014, pp. 2654–2662.

[30] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[31] D. Yu, K. Yao, H. Su, G. Li, and F. Seide, "Kl-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7893–7897.

[32] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[33] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. Ieee, 2009, pp. 248–255.

[34] A. Waibel and I. R. Lane, "Enhanced speech-to-speech translation system and methods for adding a new word," Mar. 3 2015, uS Patent 8,972,268.

[35] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.

[36] A. Bérard, O. Pietquin, C. Servan, and L. Besacier, "Listen and translate: A proof of concept for end-to-end speech-to-text translation," *arXiv preprint arXiv:1612.01744*, 2016.

[37] L. Duong, A. Anastasopoulos, D. Chiang, S. Bird, and T. Cohn, "An attentional model for speech translation without transcription," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 949–959.

[38] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*. IEEE, 2013, pp. 6645–6649.

[39] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates *et al.*, "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.

[40] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International Conference on Machine Learning*, 2016, pp. 173–182.

[41] S. Hochreiter, Y. Bengio, P. Frasconi, J. Schmidhuber *et al.*, "Gradient flow in recurrent nets: the difficulty of learning long-term dependencies," 2001.

[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

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

[44] 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, pp. 487–495.

[45] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD Birds-200-2011 Dataset," California Institute of Technology, Tech. Rep. CNS-TR-2011-001, 2011.

[46] M.-E. Nilsback and A. Zisserman, "A visual vocabulary for flower classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, 2006, pp. 1447–1454.

[47] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[48] D. Harwath, A. Torralba, and J. Glass, "Unsupervised learning of spoken language with visual context," in *Advances in Neural Information Processing Systems*, 2016, pp. 1858–1866.

[49] S. Arik, G. Diamos, A. Gibiansky, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou, "Deep voice 2: Multi-speaker neural text-to-speech," *arXiv preprint arXiv:1705.08947*, 2017.

[50] S. Ö. Arık, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, A. Ng, J. Raiman *et al.*, "Deep voice: Real-time neural text-to-speech," in *International Conference on Machine Learning*, 2017, pp. 195–204.

[51] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: 2000-speaker neural text-to-speech," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=HJtEm4p6Z

[52] F. Hernandez, V. Nguyen, S. Ghannay, N. Tomashenko, and Y. Estève, "Ted-lium 3: Twice as much data and corpus repartition for experiments on speaker adaptation," in *Speech and Computer*, A. Karpov, O. Jokisch, and R. Potapova, Eds. Cham: Springer International Publishing, 2018, pp. 198–208.

[53] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 5206–5210.

[54] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Advances in Neural Information Processing Systems*, 2016, pp. 2234–2242.

[55] D. Dowson and B. Landau, "The fréchet distance between multivariate normal distributions," *Journal of multivariate analysis*, vol. 12, no. 3, pp. 450–455, 1982.

[56] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems*, 2017, pp. 6626–6637.

[57] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates *et al.*, "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.

[58] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International Conference on Machine Learning*, 2016, pp. 173–182.