ORIGINAL ARTICLE

Fusion-s2igan: an efficient and effective single-stage framework for speech-to-image generation

Zhenxing Zhang[1] · Lambert Schomaker[1]

Abstract

The goal of a speech-to-image transform is to produce a photo-realistic picture directly from a speech signal. Current approaches are based on a stacked modular framework that suffers from three vital issues: (1) Training separate networks is time-consuming, inefficient and the convergence of the final generative model depends on the previous generators; (2) The quality of precursor images is ignored; (3) Multiple discriminator networks need to be trained. We propose an efficient and effective single-stage framework called Fusion-S2iGan to yield perceptually plausible and semantically consistent image samples on the basis of spoken descriptions. Fusion-S2iGan introduces a visual+speech fusion module (VSFM), with a pixel-attention module (PAM), a speech-modulation module (SMM) and a weighted-fusion module (WFM), to inject the speech embedding from a speech encoder into the generator while improving the quality of synthesized pictures. The PAM module models the semantic affinities between pixel regions and by assigning larger weights to significant locations. The VSFM module adopts SMM to modulate visual feature maps using fine-grained linguistic cues present in the speech vector. Subsequently, the weighted-fusion model (WFM) captures the semantic importance of the image-attention mask and the speech-modulation module at the level of the channels, in an adaptive manner. Fusion-S2iGan spreads the bimodal information over all layers of the generator network to reinforce the visual feature maps at various hierarchical levels in the architecture. A series of experiments is conducted on four benchmark data sets: CUB birds, Oxford-102, Flickr8k and Places-subset. Results demonstrate the superiority of Fusion-S2iGan compared to the state-of-the-art models with a multi-stage architecture and a performance level that is close to traditional text-to-image approaches.

Keywords Speech-to-image transform · Single-stage architecture · Generative adversarial network · Attention mechanism

1 Introduction

The task of speech-to-image generation aims to automatically yield photo-realistic and semantically consistent photographs directly from given speech signals. In recent years, this topic has drawn rapidly growing interest from multidisciplinary communities. It can be potentially used in a wealth of real-world applications, such as creating novel, visually interesting photos providing ideas for visual artists, photograph editing according to the spoken description, generating new data in machine learning [1], when augmentation is needed in training, e.g.,, classifier, and helping disabled persons produce pictures.

The availability of a speech-to-image transform would allow for an AI system to check the visual implications of a spoken narrative.

We think that research on synthetic-image generation directly from speech audio may provide a way of looking into the problems that the human brain is confronted with when forming a visual association on the basis of speech input. Note that in this task, an intermediate textual representation can be avoided (learning to read text happens at a much later stage, in children). In traditional AI, the abstract symbolic nature of text is considered to be an essential property and the pinnacle of intelligent information processing. However, the assumption that decoded speech ('recognized text') is a necessity for image

✉ Zhenxing Zhang
z.zhang@rug.nl

Lambert Schomaker
l.r.b.schomaker@rug.nl

[1] Bernoulli Institute, University of Groningen, Groningen, The Netherlands

Springer

generation has not been proven. From the point of view of deep machine learning, there is no strict necessity to rely on an intermediate symbolic text code. Vectorial representations may also be suitable for abstract representations, as has been shown extensively in recent years [2].

The task of speech-to image generation has a variety of advantages over text-to-image synthesis [3, 4], whose goal is to produce high-resolution image samples semantically aligning with input natural-language descriptions.

Firstly, spoken language may play a more advanced role in controlling and directing an image-generation process than the text modality. Compared to encoded text, speech signals contain, in addition to the pertinent phonemes, rich characteristics that can be roughly divided into timbre, prosody, rhythm, intonation, stressed components, etcetera [5, 6]. Prosody allows speakers to express the attitude and emotional states, providing key non-verbal cues that help the listener to understand the spoken description [7]. For instance, when a person says: 'There is a fire' in a neutral or falling intonation, it is probable that the fire is not big. However, if the speaker says the same sentence in a sharp intonation, then it may be more likely that this is a case of a large, risky fire. The example shows that text alone may not be sufficient to convey subtleties of the speaker's intention. It is possible to introduce non-verbal cues, e.g.,, emotional features, for conditional-image synthesis. Incorporating prosody information into the design of picture-generation systems can also be beneficial in applications, where users can influence desired photographs by adjusting the intonation of the speech probe. Since current text-to-speech synthesis models with deep-learning techniques allow for prosody modulation [8], synthesizing pictures on the basis of emotional features and linguistic information of speech signals should be feasible.

Secondly, the auditory modality is the most commonly used and natural way for humans to communicate information with each other in daily life, while written-form language is more slow and elaborate. For these reasons, compact speech-to-image generation systems may become interesting in comparison to a pipeline of a speech-recognizer followed by a standard text-to-image algorithm.

Thirdly, there are about 3500 languages lacking orthography or written form [9], which makes it impossible to train a text-to-image synthesis model for these languages and thus text-to-image systems cannot benefit these populations. Since acquiring speech signals is relatively easy and not limited by languages, an adequate speech-to-image generation architecture can be designed for these populations. This may be interesting in language research in the humanities. The presence of such tools would allow for an immediate inspection of the correspondence between spoken description and the generated image, by subjects and scholars.

In summary, spoken descriptions are suitable to explicitly and accurately describe an image using linguistic and emotional information, and researching speech-to-image synthesis has practical and scientific implications. Nevertheless, speech-to-image generation remains a considerably challenging cross-modal task. Different from textual descriptions, a speech signal is not discrete but continuous, lacking regular breaks between words, i.e., the spaces in the written-form text. This characteristic will make it difficult for the model to grasp the linguistic information of spoken captions of photographs and pictures while learning the corresponding embeddings.

We are not the first to attempt to translate speech signals into high-resolution pictures directly. For example, Li et al. [10] introduced a multi-stage speech-to-image GAN architecture to produce photo-realistic pictures semantically correlated with the input spoken description. In order to better capture the linguistic information, the researchers adopted a well-trained image encoder as a 'teacher' to train the speech encoder from scratch. Wang et al. [9] designed a new speech-embedding network (SEN) to obtain the speech vector. Furthermore, a relation-supervised densely-stacked generative model is developed to yield high-quality photographs.

These speech-to-image GAN models adopt a multi-stage framework (see Fig. 1a), where several generators and discriminators are employed to produce visually plausible samples. Although this architecture has now acquired promising results in the speech-to-image generation task, there still exist three vital problems. First, this framework entails training separate networks and is therefore inefficient as well as time-consuming [4]. Even worse, it is difficult for the final generator to output perceptually realistic photographs when the earlier generator networks do not converge to a global optimum [3, 11–13]. Second, the quality of the outputs of the previous generator networks [14] is ignored by this architecture. Contextual vectors are not used to enhance and modulate the visual feature maps in the generator for precursor images, which comprises up-sampling operations and convolutional layers [3]. Third, several discriminator networks are required to be trained.

In recent years, we have proposed two novel single-stage text-to-image GAN models, i.e., DTGAN [3] and DiverGAN [4], to address the above-mentioned issues of a multi-stage architecture. Both DTGAN and DiverGAN are capable of adopting a single generator/discriminator pair to produce photo-realistic and semantically correlated image samples on the basis of given natural-language descriptions. In DTGAN, dual-attention models, conditional adaptive instance-layer normalization and a new type of visual loss were presented. In DiverGAN, the sentence-level attention models introduced in DTGAN were
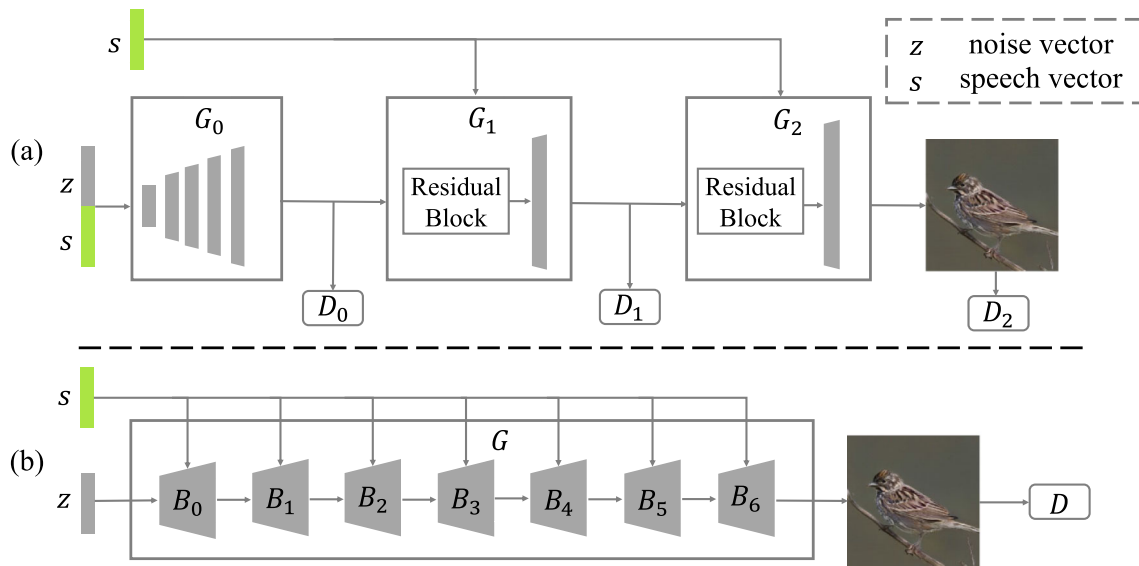
Fig. 1 The comparison between the stacked framework and the proposed architecture. The multi-stage pipeline (**a**) entails training separate generators to obtain high-quality samples. The presented Fusion-S2iGan (**b**) is capable of producing visually plausible pictures only employing a single generator/discriminator pair. In (a), G_0-G_2 are generators and D_0-D_2 are discriminators. In (b), B_0-B_6 are the dual-residual speech-visual fusion blocks discussed in Sect. 4, and G and D are the generator network and discriminator network, respectively

extended to word-level attention modules, to better control an image-generation process using word features. Moreover, we proposed to insert a dense layer into the pipeline to address the lack-of-diversity problem present in current single-stage text-to-image GAN models. Inspired by these previous works and aiming to address the limitations of a stacked framework, our research introduces a visual+−speech fusion module and incorporates several effective loss functions. These advancements collectively contribute to a novel single-stage speech-to-image architecture.

Looking at recent publications on text-to-image generation [15–17], it is clear that this field has become a thriving area of research. Despite the significant progress made by models like the OpenAI 'Dall-E 2' [16] and the stable diffusion model [15] in producing high-quality images, these achievements often come at the cost of requiring billions of images and hundreds of GPU hours for training [15]. Additionally, generating images based on speech descriptions with these models necessitates the extra step of translating speech into text. In contrast, our model offers a direct approach to generating images from speech, while being trained on smaller data sets. This highlights the potential for efficient and effective image generation directly from speech inputs, even with limited training data and using a common single-GPU platform.

The contributions of this paper can be summarized as follows:

- We present a novel effective and efficient single-stage architecture called Fusion-S2iGan (see Fig. 1b) for speech-to-image transforms, which is capable of producing high-quality and semantically consistent pictures only using a generator/discriminator pair.

- A visual+speech fusion module (VSFM) is designed to effectively feed the speech information from a speech encoder to the neural network while improving the quality of generated photographs. More importantly, we spread the bimodal information over almost all layers of the generator. This allows for an influence of the speech over features at various hierarchical levels in the architecture, from crude early features to abstract late features.

- To the best of our knowledge, we are the first to apply (1) the hinge loss, (2) deep attentional multimodal similarity model (DAMSM) loss and (3) matching-aware zero-centered gradient penalty (MA-GP) loss in speech-to-image generation, which are beneficial for the convergence and stability of the generative model.

- Extensive experiments are carried out on four benchmark data sets, i.e., CUB bird [18], Oxford-102 [19], Flickr8k [20] and Places-subset [21]. The experimental results suggest that the proposed Fusion-S2iGan has the capacity to yield better pictures than current multi-stage speech-to-image GAN models such as StackGAN++ [22], Li et al. [10] and S2IGAN [9].

- We explore how far can current single-stage text-to-image methods be used for the speech-to-image transform task, which is depicted in Sect. 5.4.3.

The remainder of this paper is organized as follows. Section 2 reviews related works. In Sect. 4, the architecture of the proposed Fusion-S2iGan is introduced in detail. In Sect. 4.2, we elaborate on the presented visual+speech fusion module. Experimental settings are discussed in Sect. 5 and results are reported in Sects. 5.2, 5.3 and 5.4. Section 6 draws the conclusions.

2 Related works

In this section, research fields related to our work are described, including CGAN in text-to-image generation and speech-audio-to-image synthesis.

2.1 CGAN in text-to-image generation

Text-to-image generation aims to automatically produce visually plausible and semantically consistent image samples, given natural-language descriptions. Many text-to-image synthesis approaches are built upon the original conditional generative adversarial network (cGAN) [23] due to its appealing performance. A cGAN-based text-to-image synthesis architecture is composed of a generator and a discriminator, that are trained to competing goals. The generator model takes a random latent vector and textual vectors as the inputs while outputting the corresponding photograph, attempting to approximate the text-conditioned data distribution. In the meantime, the discriminator network learns to separate the real text-image pair from the fake text-image pair. We roughly group them into two categories in terms of the number of the generators and the discriminators they use.

2.1.1 Multi-stage models

Zhang et al. [22] suggested employing several generators and discriminators to boost image quality and semantic relevance while presenting the first multi-stage text-to-image generation framework named StackGAN. StackGAN serves as a strong basis for the future research. Xu et al. [24] proposed to incorporate a spatial-attention module into the design of a stacked architecture, in order to better bridge the semantic gap between vision and language. Qiao et al. [25] developed MirrorGAN that introduced an image-to-text model to ensure that synthesized pictures are semantically related to given textual descriptions. Zhu et al. [14] built DMGAN where a dynamic-memory module is applied to enhance the image quality in the initial stage. CPGAN [26] designed a memory structure to parse the produced image in an object-wise manner and introduced a conditional discriminator to promote the semantic alignment of text-image pairs.

2.1.2 Single-stage methods

Reed et al. [27] were the first to use a single generator/discriminator pair to yield samples on the basis of natural-language descriptions. However, the resolution of generated pictures is limited owing to the unstable training process as well as the lack of an effective structure. Tao et al. [28] developed a matching-aware zero-centered gradient penalty loss to help stabilize the training and improve the image quality of a single-stage text-to-image GAN model. Zhang et al. [3] presented DTGAN, in which dual-attention models, conditional adaptive instance-layer normalization and a new type of visual loss are designed to generate perceptually realistic images only using a single generator/discriminator pair. Zhang et al. [4] proposed DiverGAN inserting a dense layer into the pipeline to address the lack-of-diversity problem present in current single-stage text-to-image GAN models. Zhang et al. [29] introduced linear-interpolation and triangular-interpolation techniques to explain the single-stage text-to-image GAN model. Moreover, a *Good/Bad* data set was created to select successfully generated images and corresponding good latent codes.

2.2 Speech-audio-to-image synthesis

With the recent rapid advances in generative-adversarial networks (GANs) [30] and conditional GANs (cGANs) [23], speech-to-image generation has made promising advances in image quality and semantic consistency when given speech-audio signals as inputs. Various studies focused on synthesizing images conditioned on the sound of music. Chen et al. [31] made the first attempt to use the cGAN to produce samples on the basis of music audio. In addition, an image-to-sound network is introduced to synthesize music according to instrument pictures. Hao et al. [32] presented a unified architecture (CMCGAN) for audio-visual mutual synthesis. Specifically, CMCGAN incorporated audio-to-visual, audio-to-audio, visual-to-audio and visual-to-visual networks into the pipeline for cyclic consistency and better convenience.

Some publications tried to reconstruct a facial photograph from a short audio segment of speech. Duarte et al. [33] proposed an end-to-end speech-to-face GAN model called Wav2Pix, which has the ability to synthesize diverse and promising face pictures according to a raw speech signal. Oh et al. [34] developed a reconstructive speech-to-face architecture named Speech2Face that contains a voice encoder and a pre-trained face decoder network. The encoder is used to extract face information from the given speech and the decoder aims to reconstruct a realistic face sample.

Different from the above approaches, several papers aimed at translating a spoken description of an image into a high-quality picture directly. Li et al. [10] attempted to apply a multi-stage speech-to-image GAN model to yield perceptually plausible pictures semantically correlated with input speech-audio descriptions. To better acquire the speech embedding, the researchers used a well-trained image encoder as a 'teacher' to train the speech encoder from scratch. Wang et al. [9] proposed S2IGAN where a speech-embedding network (SEN) was designed to obtain the spoken vector and a matching loss and a distinctive loss were presented to train SEN. In addition, a relation-supervised densely-stacked generative model is introduced to produce high-resolution images.

This paper focuses on solving the task of speech-to-image synthesis. In contrast to previous works, our approach aims to generate high-quality images from spoken descriptions, leveraging only a single generator/discriminator pair.

3 Preliminaries

3.1 Conditional generative adversarial network (CGAN)

A CGAN is an advanced version of the GAN, using conditional contexts, such as class labels, textual descriptions, and low-resolution images, in combination with random code as the inputs for G. G yields samples that are correlated with the given conditional contexts, and the aim of D is to distinguish the real pair (x, c) from the fake pair $(G(z, c), c)$. Specifically, the objective $V(G, D)$ of a CGAN can be formulated as follows:

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

where $p_c(c)$ is the distribution of c.

The goal of speech-to-image generation is to yield perceptually plausible samples that are semantically correlated with the linguistic content of input spoken descriptions. Mathematically, let $\{(I_i, S_i)\}_{i=1}^n$ represent a suite of n image-spoken caption pairs for training, where I_i indicates an image and $S_i = (s_i^1, s_i^2, ..., s_i^k)$ refers to a set of k speech-audio descriptions of I_i. The generator of a speech-to-image GAN model aims to synthesize a high-resolution and semantically related picture $\hat{I}_i$ on the basis of a speech signal s_i randomly picked from S_i. In the meantime, the discriminator is trained to separate the real

image-speech pair (I_i, s_i) from the fake image-speech pair $(\hat{I}_i, s_i)$.

3.2 Scaled dot-product attention

In general, an attention function takes a query $q_i \in \mathbb{R}^d$ and a set of key-value pairs as the inputs, and then generates a weighted sum of the values. The weight assigned to each value is determined by a softmax function applied to the dot product between the query and all keys. The method for computing attention weights can be defined as follows:

$$Attention(q_i, K, V) = Softmax\left(\frac{D(q_i, K^T)}{\sqrt{d}}\right) V \qquad (2)$$

where $D(\cdot)$ represents the dot-product operation and *Softmax* denotes the softmax function. $K = (k_1, k_2, ..., k_n), k_i \in \mathbb{R}^d$, and $V = (v_1, v_2, ..., v_n), v_i \in \mathbb{R}^d$ are the keys and the values, respectively. $\frac{1}{\sqrt{d}}$ refers to the scaling factor.

4 The proposed approach

In this section, we discuss the overall architecture of DiverGAN, which can be seen in Fig. 2. After that, a module called visual+speech fusion module (VSFM) is presented as a means of efficiently and effectively combining visual feature maps with the global speech embedding, resulting in high-resolution images. Subsequently, we describe the loss functions used to train our network.

4.1 Overall architecture

The overall framework of Fusion-S2iGan for a speech-to-image transform is presented in Fig. 2. The architecture only consists of a generator and a discriminator. Next, we introduce these two components one by one.

4.1.1 Generator

The generator is able to project a speech signal into a photo-realistic and semantically consistent picture, shown in Fig. 2a. More specifically, the generator network is composed of a dense layer transforming a latent code to the initial feature map and seven dual-residual speech-visual fusion blocks modulating the visual feature map with the spoken vector derived from a speech encoder. The speech encoder is employed to learn the semantic representation and conceptual meaning of a given spoken description that capture the discriminative visual details [35].

The designed dual-residual speech-visual fusion block (see Fig. 2c) contains two effective and efficient
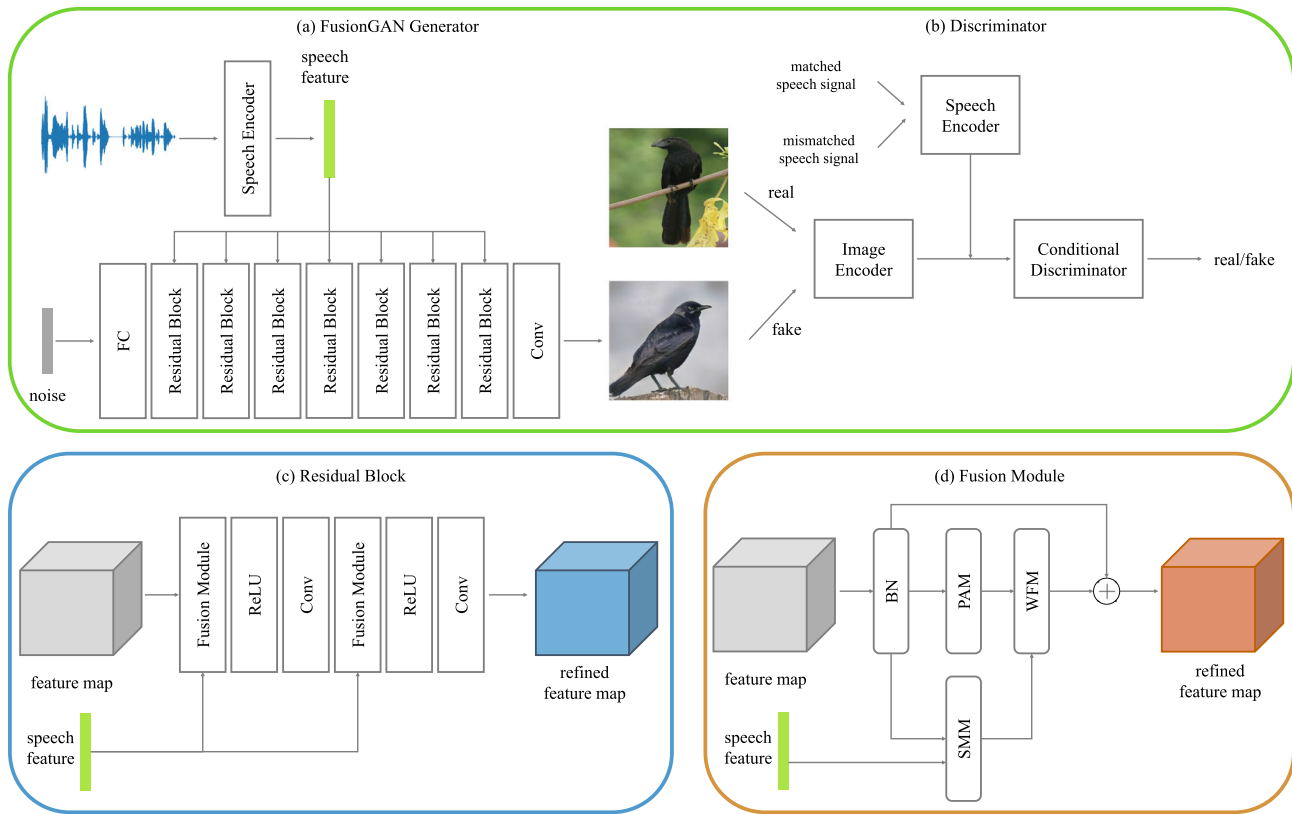
Fig. 2 The overall framework of the proposed Fusion-S2iGan. *FC* is a dense layer, *Conv* is a convolutional layer, *ReLU* is a ReLU activation function and BN is a Batch-Normalization operation. Additionally, *Residual Block* and *Fusion Module* refer to the presented dual-residual speech-visual fusion block (see (**c**)) discussed in Sect. 4.1.1 and visual+speech fusion module (see (**d**)) discussed in Sect. 4.2,

respectively. Furthermore, *PAM*, *SMM* and *WFM* in (**d**) represent the pixel-attention module, speech-modulation module and weighted-fusion module, respectively, discussed in Sect. 4.2. Note that we do not plot the up-sample layers between *Residual Blocks* in (**a**) due to the limited space

visual+speech fusion modules (VSFM) (see Fig. 2d) as well as a suite of ReLU activation functions and convolutional layers. Each VSFM comprises Batch Normalization (BN) [36], a pixel-attention module (PAM), a speech-modulation module (SMM) and a weighted-fusion module (WFM). The dual-residual speech-visual fusion block allows us to easily enhance model capacity by effectively increasing the number of layers, while also stabilizing the training process by maintaining more original features than cascade structures. Benefiting from such dual-residual speech-visual fusion blocks, the generator network has the ability to yield high-quality pictures. The process of synthesizing photographs on the basis of a speech signal is formulated as follows:

$$h_0 = F_0(z) \tag{3}$$

$$h_1 = F_1^{Dual}(h_0, s) \tag{4}$$

$$h_i = F_i^{Dual}(h_{i-1} \uparrow, s) \quad for \quad i = 2, 3, ..., 7 \tag{5}$$

$$o = G_c(h_7) \tag{6}$$

where z is a latent vector randomly sampled from the normal distribution, F_0 is a dense layer, s is the speech embedding from a speech encoder, F_i^{Dual} is the presented dual-residual speech-visual fusion block and G_c is the last convolutional layer. The details of the proposed visual+speech fusion module will be discussed in Sect. 4.2.

4.1.2 Discriminator

The architecture of the discriminator network comprises an image encoder, a pre-trained speech encoder and a conditional discriminator network, depicted in Fig. 2b. The choice of the hyper-parameters in the discriminator is based on our earlier work [3], but textual features are replaced with speech embeddings. To be specific, the image encoder is constructed with one convolutional layer with strides 1 kernel size 3 padding 1 followed by six adaptively successive residual blocks. Each block consists of two convolutional layers, where the first layer with strides 2 kernel size 4 padding 1 is used to reduce the

dimension to half of the input feature map and the second one with strides 1 kernel size 3 padding 1 aims to further distill image features. After each convolutional layer, Leaky-ReLU activation [37] with a slope of 0.2 is utilized to help the training. The number of filters for residual blocks are 64 128 256 512 1024 1024, respectively. In order to stabilize the learning, we incorporate a residual connection into each block. In the process of training, a picture is fed into the image encoder to extract the image features (1024, 4, 4) which are combined with the spoken vector from the speech encoder as the joint embeddings, with a dimension of (2048, 4, 4). After that, the joint features are passed through the conditional discriminator to gain the final conditional score, which is utilized to determine whether the input image-spoken caption pair is real or fake.

4.2 Visual+speech fusion module

It is widely known that the semantic relationships between the visual content of photographs and images and the corresponding conditional contexts, e.g.,, class labels and textual descriptions, play a significant role in an image-generation process. However, this correlation will be more complicated for speech-to-image synthesis, since the speech signal is long and continuous, lacking word boundaries, i.e., the spaces in natural-language descriptions. It is thus very difficult to model the affinities between the linguistic content of spoken descriptions and the visible objects. Moreover, word-level modulation modules fail to fuse spoke information and visual feature maps due to the continuity of speech signals. In this case, a crucial problem is most clearly present for researchers: *how to effectively inject the information from a spoken probe into a neural network which has a large 2D color image as its output?*

The input speech signal needs to be converted to the global speech vector similar to the sentence embedding [3] and then can be made to modulate feature maps in the network. In this section, we develop an effective and efficient visual+speech fusion module (VSFM) to facilitate the visual feature maps with the global speech embedding and yield high-resolution pictures.

4.2.1 Overall architecture

The framework of the proposed visual+speech fusion module is shown in Fig. 2d. Given an intermediate feature map $F \in \mathbb{R}^{C \times H \times W}$ and the speech vector $s \in \mathbb{R}^D$ from a speech encoder as inputs, the process of the VSFM is composed of three steps:

1. The VSFM first applies Batch Normalization (BN) [36] on F, acquiring a new feature map $F' \in \mathbb{R}^{C \times H \times W}$. This may help stabilize the learning of the conditional generative adversarial network (cGAN) and accelerate the training process.

2. $F' \in \mathbb{R}^{C \times H \times W}$ is refined using the pixel-attention module (PAM) M_p and speech-modulation module (SMM) M_s, respectively.

3. The VSFM effectively fuses their outputs $M_p(F') \in \mathbb{R}^{C \times H \times W}$ and $M_s(F', s) \in \mathbb{R}^{C \times H \times W}$ using the weighted-fusion module (WFM) M_f. Meanwhile, a residual connection is employed to get the final enhanced result $F'' \in \mathbb{R}^{C \times H \times W}$.

Note that we spread the bimodal information of the VSFM over all layers of the generator network. This allows for an influence of the speech over features at various hierarchical levels in the architecture, from crude early features to abstract late features. The overall process of the VSFM can be formulated as follows:

$$F' = \mathrm{BN}(F) \tag{7}$$

$$F'' = F' + M_f(M_p(F'), M_s(F', s)) \tag{8}$$

where BN indicates a Batch-Normalization operation. The details of PAM, SMM and WFM will be described in the following subsections.
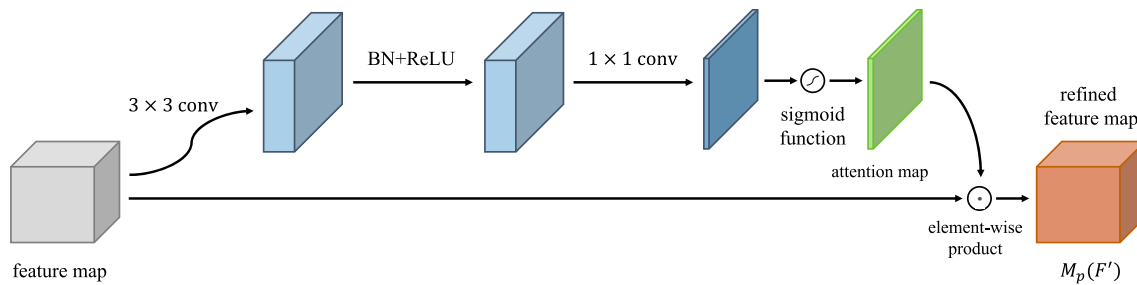
4.2.2 Pixel-attention module (PAM)

Pictures and photographs consist of visual pixels which are considerably significant for image quality. Learning the long-range contextual dependency of each position of a picture is essential for producing perceptually plausible samples. However, convolutional operations can only grasp local relationships between spatial contexts and thus fail to 'see' the entire image field. Hence, the pixel-attention module (PAM) is introduced to effectively model the spatial affinities between visual pixels and enable crucial and informative positions to receive more attention from the generator.

Figure 3 illustrates the process of the pixel-attention module (PAM). For a feature map $F' \in \mathbb{R}^{C \times H \times W}$, we first feed it into a 3×3 convolutional layer followed with BN and a ReLU function to reduce the channel dimension to $\mathbb{R}^{C/r \times H \times W}$. This may integrate and strength the visual feature map across the channel and spatial directions. Subsequently, we use a 1×1 convolutional layer followed with the sigmoid function to process the features to obtain the pixel-attention map $PA \in \mathbb{R}^{1 \times H \times W}$. Mathematically,

$$PA = \sigma(f_1^{1 \times 1}(ReLU(\mathrm{BN}(f_0^{3 \times 3}(F'))))) \tag{9}$$

where f is a convolutional layer, $ReLU$ is the ReLU function and σ is the sigmoid function. We conduct a matrix multiplication between the original feature map $F' \in$

Fig. 3 Overview of the introduced pixel-attention module, which aims to assign larger weights to discriminative and informative pixels. BN and *ReLU* refer to Batch Normalization and a ReLU activation

function, respectively. 1×1 *conv* and 3×3 *conv* indicate the 1×1 and 3×3 convolutional operation, respectively

$\mathbb{R}^{C \times H \times W}$ and the spatial-attention map $PA \in \mathbb{R}^{1 \times H \times W}$ to acquire the refined result. Specifically,

$$M_p(F') = F' \odot PA \tag{10}$$

where $\odot$ is the element-wise multiplication.

4.2.3 Speech-modulation module (SMM)

An adequate modulation method needs to be developed to ensure the semantic consistency and quality of synthesized photographs [4]. To this end, the speech-modulation module (SMM) is introduced to inject the features derived from the speech signal into the network at the proper points in the network architecture.

Inspired by [28], SMM facilitates the visual feature map using detailed linguistic cues captured from the speech embedding s. To be specific, we adopt two dense layers to project s into the linguistic cues $WA \in \mathbb{R}^{C \times 1 \times 1}$ and $BA \in \mathbb{R}^{C \times 1 \times 1}$. After that, WA and BA are employed to scale and shift F'. The process of SMM can be defined as follows:

$$WA = MLP(ReLU(MLP(s))) \tag{11}$$

$$BA = MLP(ReLU(MLP(s))) \tag{12}$$

$$M_s(F') = F' \odot WA + BA \tag{13}$$

where *MLP* is a fully-connected perceptron layer and $M_s(F')$ is the output from SMM.

4.2.4 Weighted-fusion module (WFM)

We do not simply fuse the attended and enhanced visual feature maps from PAM and SMM using an addition or element-wise multiplication, since pixels and speech signals refer to two very different modalities and need to be combined in a more advanced manner. Here, we propose an efficient weighted-fusion module (WFM) to highlight the discriminative and significant regions in an adaptive manner and produce high-quality image samples.

The detailed structure of the weighted-fusion module (WFM) is depicted in Fig. 4. For the outputs $M_p(F')$ and $M_s(F', s)$ from PAM and SMM, respectively, we perform an element-wise addition between them as a first step, acquiring an intermediate feature map $IF \in \mathbb{R}^{C \times H \times W}$. In the next step, IF is used to compute the final weights for both $M_p(F')$ and $M_s(F', s)$. More specifically, we employ global-average pooling (GAP) to process IF to aggregate holistic and discriminative information, thereby obtaining a channel-feature vector $IF' \in \mathbb{R}^{C \times 1 \times 1}$. Subsequently, we feed IF' into two fully-connected perceptron layers, in which the first one is to compress and integrate the channel features and the other aims to recover the channel dimension and capture the semantic importance of the image-attention mask and the speech-modulation module at the level of the channels. After that, we apply a softmax function across the channel dimension to get the contextually channel-aware weight matrix $WF \in \mathbb{R}^{C \times 2}$. Afterward, WF is split into two separate channel-wise weight matrices $WF_1 \in \mathbb{R}^{C \times 1}$ and $WF_2 \in \mathbb{R}^{C \times 1}$. Mathematically,

$$IF = M_p(F') + M_s(F', s) \tag{14}$$

$$IF' = GAP(IF) \tag{15}$$

$$WF = softmax(MLP(GELU(MLP(IF')))) \tag{16}$$

$$WF = [WF_1; WF_2] \tag{17}$$

where *GAP* is the global-average pooling and *GELU* is the Gaussian Error Linear Unit (GELU) [38]. WF_1 and WF_2 are resize into $\mathbb{R}^{C \times 1 \times 1}$ and applied to $M_p(F')$ and $M_s(F', s)$ to get the final fusion result. It is denoted as follows:

$$M_f(F', s) = WF_1 \odot M_p(F') + WF_2 \odot M_s(F', s) \tag{18}$$

4.3 Objective function

An adversarial loss is employed to match generated samples to input speech signals. Inspired by [4], we utilize the hinge objective [39] for stable training instead of the
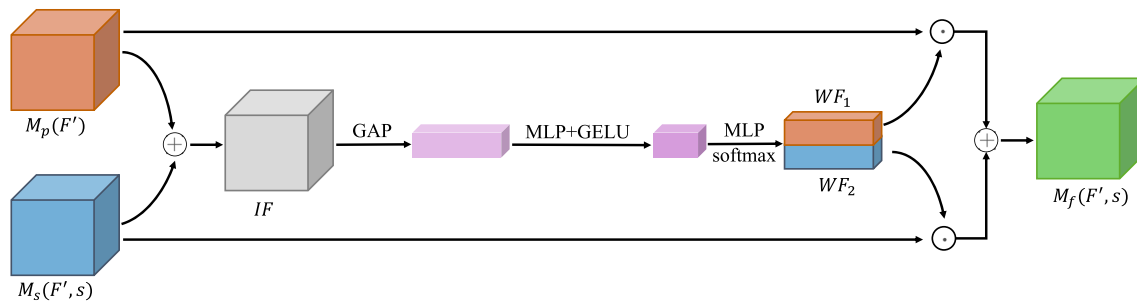
Fig. 4 Overview of the proposed weighted-fusion module, which effectively combines the outputs from the pixel-attention module (PAM) and speech-modulation module (SMM) in an adaptive manner. $M_p(F')$ and $M_s(F', s)$ denote the result of PAM and SMM, respectively. *GAP*, *MLP* and *GELU* represent the global-average pooling, fully-connected perceptron layer and Gaussian Error Linear Unit (GELU) [38], respectively. WF_1 and WF_2 are the channel-aware weight matrix. $M_f(F', s)$ is the final refined result

vanilla GAN objective. The adversarial loss for the discriminator is defined as follows:

$$
\begin{aligned}
\mathcal{L}_{\text{adv}}^D =& \mathbb{E}_{x \sim p_{\text{data}}}[\max(0, 1 - D(x, s))] \\
&+ \frac{1}{2} \mathbb{E}_{x \sim p_G}[\max(0, 1 + D(\hat{x}, s))] \\
&+ \frac{1}{2} \mathbb{E}_{x \sim p_{\text{data}}}[\max(0, 1 + D(x, \hat{s}))]
\end{aligned}
\tag{19}
$$

where s is a given spoken caption, $\hat{s}$ is a mismatched speech-audio description, x is the real image from the distribution p_{data} and $\hat{x}$ is the synthesized sample from the distribution p_G,

To enhance the image quality and semantic consistency of produced pictures, we adopt the matching-aware zero-centered gradient penalty (MA-GP) loss [28] for the discriminator, which applies gradient penalty to real images and input spoken descriptions. The MA-GP Loss is formulated as follows:

$$
\mathcal{L}_M = \mathbb{E}_{x \sim p_{\text{data}}}\left[(\|\nabla_x D(x, s)\|_2 + \|\nabla_s D(x, s)\|_2)^p\right]
\tag{20}
$$

For the generator, we apply an adversarial loss and a deep attentional multimodal similarity model (DAMSM) loss [24] to train the network.

5 Experiments

In this section, a variety of quantitative and qualitative evaluations are conducted to validate the effectiveness of the presented Fusion-S2iGan in yielding visually realistic and semantically consistent images. The evaluations are carried out on two synthesized spoken caption-image data sets, namely Oxford-102 [19] and CUB bird [18], as well as two real spoken caption-image data sets, namely Flickr8k [20] and Places-subset [21]. Specifically, we describe the details of experimental settings in Sect. 5.1. Afterward, our proposed Fusion-S2iGan is compared with previous CGAN-based approaches for text-to-image generation

using the synthesized data sets in Sect. 5.2. In Sect. 5.3, a similar comparison is performed using the real speech data sets. Subsequently, an analysis of the individual contributions from different components of our Fusion-S2iGan is conducted in Sect. 5.4.

5.1 Experimental settings

Datasets To evaluate the proposed Fusion-S2iGan, we perform extensive experiments on two synthesized spoken caption-image data sets and two real spoken caption-image data sets, which are employed by Li et al. [10] and S2IGAN [9].

- *CUB bird* [18]. The CUB data set includes a total of 11,788 images, which is divided into 8,855 training pictures and 2,933 testing pictures. Each picture is accompanied by 10 natural-language descriptions. To evaluate the task of speech-to-image generation, Li et al. [10] and S2IGAN [9] transform textual descriptions to spoken captions utilizing a text-to-speech method such as Tacotron 2 [40].

- *CUB bird* [18]. The CUB data set includes a total of 11,788 images, which is divided into 8,855 training pictures and 2,933 testing pictures. Each picture is accompanied by 10 natural-language descriptions. To evaluate the task of speech-to-image generation, Li et al. [10] and S2IGAN [9] transform textual descriptions to spoken captions utilizing a text-to-speech method such as Tacotron 2 [40]. **Oxford-102** [19]. The Oxford-102 data set is composed of 8,189 pictures, in which 5,878 pictures belong to the training set and the other 2,311 pictures are used for testing. Each picture contains 10 textual descriptions, which are transformed to speech signals in the same way as the CUB data set.

- *Flickr8k* [20]. The Flickr8k data set is a more challenging data set comprising 8,000 scene pictures and each image is paired with 5 real spoken

descriptions collected by [41]. We split the Flickr8k data set according to S2IGAN [9].

- *Places-subset* [21] The Places-subset data set is a subset of the Places Audio Caption data set [42, 43], which encompasses real spoken captions of pictures from the Places 205 data set [21]. It contains a total of 13,803 image-spoken caption pairs belonging to 7 categories, which are divided into 10,933 training paired data and 2,870 testing paired data.

In summary, two synthesized spoken caption-image data sets and two real spoken caption-image data set are adopted to evaluate the proposed Fusion-S2iGan.

Implementation details For the speech encoder, following the structure of S2IGAN [9], we use two 1-D convolution blocks, two bidirectional gated recurrent units (GRUs) and a self-attention module to obtain the speech embedding. To better capture the semantic representation of spoken descriptions on the Places-subset data set, we replace the Inception-V3 [44] pre-trained on ImageNet [45] in S2IGAN with the ResNet [46] trained on Places 205 [21]. Following the settings of S2IGAN [9], the dimension of the speech vector is set to 1024. The training of Fusion-S2iGan is based on our earlier work [3]. Specifically, we utilize the Adam optimizer [47] with $\beta_1 = 0.0$ and $\beta_2 = 0.9$ to train the networks. Furthermore, we follow the two timescale update rule (TTUR) [48] and set the learning rates for the generator and the discriminator to 0.0001 and 0.0004, respectively. We set the batch size to 32. We implement Fusion-S2iGan adopting PyTorch [49]. All the experiments are conducted on a single NVIDIA Tesla V100 GPU (32 GB memory).

Evaluation metrics We verify the effectiveness of Fusion-S2iGan by computing the following four extensively employed evaluation metrics:, i.e., the inception score (IS) [50], Fréchet inception distance (FID) [44] score, mean Average Precision (mAP) and R(ecall)@50 [9], which are used by S2IGAN [9]. The first two are to assess the diversity and perceptual quality, and the last two are to measure the semantic relevancy between synthesized samples and their spoken descriptions.

- *Inception score (IS)* [50]. The IS is obtained by computing the KL divergence between the conditional class distribution and the marginal class distribution. The synthesized pictures are divided into multiple groups and the IS is calculated on each group of photographs, then the average and standard deviation of the score are reported. Higher IS demonstrates better quality and diversity among the generated images [4].
- *Fréchet inception distance (FID)* [44]. The FID calculates the Fréchet distance between the distribution of generated images and the distribution of true data. A

lower FID score means that the generated pictures are closer to the corresponding real pictures.

- *Mean average precision (mAP) and R(ecall)@50* [9]. The mAP and R(ecall)@50 are speech-image retrieval metrics, introduced by S2IGAN to measure the semantic relevancy for the synthesized and real speech data sets, respectively. Higher mAP or R@50 suggest better semantic consistency.

It should be noted that standard deviation on performances cannot be given for all metrics reported (FID and mAP). This is due to the fact that we wanted the measurements to be comparable with other studies. However, given the size of the test sets: $N = 30k$ for the CUB data set and $N = 10k$ for the Oxford-102 data set, we expect the measurements to be fairly reliable. As a comparative illustration: In case of a measured accuracy of 0.80 and a data set of $N = 30k$, at $\alpha = 0.01$, the confidence band would span 0.794 to 0.806, i.e., a deviation of just 0.75%.

5.2 Results on the synthesized speech datasets

5.2.1 Quantitative results

We compare Fusion-S2iGan with previous single-stage [4, 28, 53] and multi-stage [9, 10, 14, 22, 24, 25, 51, 52] cGAN-based methods in text-to-image generation and speech-to-image transforms on the CUB and Oxford-102 data sets. The IS, FID and mAP of Fusion-S2iGan and other compared approaches on the CUB and Oxford-102 data sets are shown in Table 1. Note that the reported scores in Table 1 are based on the results presented in these publications. We can see that Fusion-S2iGan achieves the best scores on speech-to-image synthesis, significantly improving the IS from 4.29 to 4.82 on the CUB data set and from 3.69 to 3.81 on the Oxford-102 data set, reducing the FID from 14.50 to 13.74 on the CUB data set and from 48.64 to 40.08 on the Oxford-102 data set and increasing the mAP from 9.04 to 11.49 on the CUB data set and from 13.40 to 17.72 on the Oxford-102 data set. Notably, Fusion-S2iGan performs better in speech-to-image generation than both DM-GAN [14] in text-to-image generation on the CUB data set, as seen from an improvement of the IS from 4.75 to 4.82 and DTGAN [3] doing text-to-image synthesis on the Oxford-102 data set, here improving the IS from 3.77 to 3.81. The experimental results suggest that Fusion-S2iGan is capable of yielding perceptually plausible pictures with higher quality and better diversity than state-of-the-art speech-to-image transform models and many text-to-image approaches.

Table 1 The IS, FID and mAP of previous text-to-image and speech-to-image approaches and Fusion-S2iGan on the CUB and Oxford-102 data sets

Datasets	Methods	Input	IS ↑	FID ↓	mAP ↑
CUB	StackGAN++ [22]	text	4.04±0.05	26.07	7.01
	AttnGAN [24]	Text	4.36±0.03	23.98	–
	MirrorGAN [25]	Text	4.56±0.05	–	–
	ControlGAN [51]	Text	4.58±0.09	–	–
	SDGAN [52]	Text	4.67±0.09	–	–
	DM-GAN [14]	Text	4.75±0.07	16.09	–
	DF-GAN [28]	Text	4.86±0.04	19.24	–
	DTGAN [53]	Text	4.88±0.03	16.35	–
	DiverGAN [4]	Text	4.98±0.06	15.63	–
	Classifier-based [9]	speech	3.68±0.04	43.76	–
	Li et al. [10]	Speech	4.09±0.04	18.37	–
	StackGAN++ [22]	Speech	4.14±0.04	18.94	8.09
	S2IGAN [9]	Speech	4.29±0.04	14.50	9.04
	Fusion-S2iGan	Speech	**5.06±0.09**	**13.09**	**12.12**
Oxford-102	StackGAN++ [22]	Text	3.26±0.01	48.68	–
	DF-GAN [28]	Text	3.71±0.06	–	–
	DTGAN [53]	Text	3.77±0.06	–	–
	DiverGAN [4]	Text	3.99±0.05	–	–
	Classifier-based [9]	Speech	3.30±0.06	64.75	–
	Li et al. [10]	Speech	3.23±0.05	54.76	–
	StackGAN++ [22]	Speech	3.69±0.08	54.33	12.18
	S2IGAN [9]	Speech	3.55±0.04	48.64	13.40
	Fusion-S2iGan	Speech	**3.81±0.08**	**40.08**	**17.72**

The best results of speech-to-image generation are in bold

5.2.2 Qualitative results

In addition to quantitative experiments, we carry out qualitative comparison on the CUB and Oxford-102 data sets, which is presented in Fig. 5. Note that in order to visualize the synthesized images of S2IGAN [9], we utilize the public code[1] to train the network from scratch due to the lack of available pre-trained models. It was noted that the contrast of the generated images was dull for the S2IGAN code, as given. By enhancing brightness and contrast, this problem could be solved. However, we found this is only needed for human evaluation: On 20 selected images,[2] the IS score for S2IGAN was the same for the raw (grayish) output and the enhanced output samples. As can be seen in Fig. 5a, the shape of birds produced by Attn-GAN [24] and S2IGAN [9] (2nd, 3rd, 6th and 8th column) is strange, the image details are lost (2nd, 3rd, 5th, 6th and 7th column) and the backgrounds are blurry (1st, 3rd, 6th and 8th column). However, Fusion-S2iGan yields more clear and photo-realistic samples than AttnGAN and S2IGAN, which demonstrates the effectiveness of our
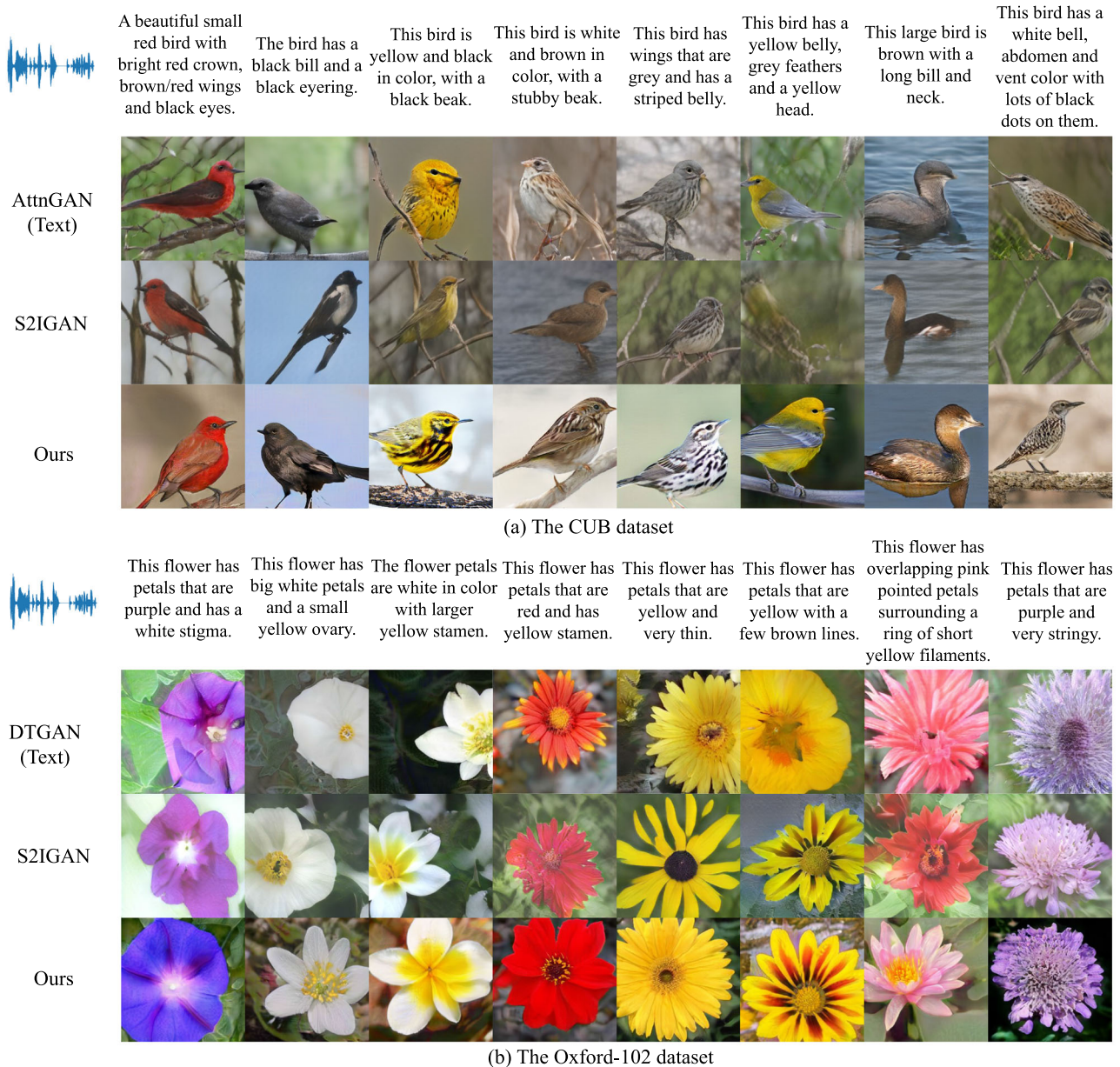
architecture. For example, the birds synthesized by Fusion-S2iGan have a more clear shape and richer color distributions compared to AttnGAN and S2IGAN in the 2nd, 3rd, 6th, 7th and 8th column. Furthermore, as shown in the 1st, 2nd, 4th, 6th and 8th column, Fusion-S2iGan produces visually plausible birds with more vivid details than Attn-GAN and S2IGAN.

The qualitative results of DTGAN [3], S2IGAN [9] and Fusion-S2iGan on the Oxford-102 data set are illustrated in Fig. 5b, suggesting that Fusion-S2iGan is able to adopt a single generator/discriminator pair to produce high-resolution pictures that correspond well to the given speech-audio descriptions. For instance, Fusion-S2iGan yields perceptually realistic flowers with a more vivid shape than DTGAN and S2IGAN in the 1st, 2nd, 6th, 7th and 8th column. In addition, as shown in the 3th, 4th, 5th and 6th column, the flowers generated by Fusion-S2iGan have more clear details and richer color distributions than DTGAN and S2IGAN.

The above experimental results indicate that Fusion-S2iGan equipped with the dual-residual speech-visual fusion blocks has the ability to effectively inject the speech information into the generator network and generate high-quality photographs.

---

[1] https://github.com/xinshengwang/S2IGAN.

[2] https://zenodo.org/record/7014899#.YwN-xHZByUk.

A beautiful small red bird with bright red crown, brown/red wings and black eyes.

The bird has a black bill and a black eyering.

This bird is yellow and black in color, with a black beak.

This bird is white and brown in color, with a stubby beak.

This bird has wings that are grey and has a striped belly.

This bird has a yellow belly, grey feathers and a yellow head.

This large bird is brown with a long bill and neck.

This bird has a white bell, abdomen and vent color with lots of black dots on them.

AttnGAN (Text)

S2IGAN

Ours

(a) The CUB dataset

This flower has petals that are purple and has a white stigma.

This flower has big white petals and a small yellow ovary.

The flower petals are white in color with larger yellow stamen.

This flower has petals that are red and has yellow stamen.

This flower has petals that are yellow and very thin.

This flower has petals that are yellow with a few brown lines.

This flower has overlapping pink pointed petals surrounding a ring of short yellow filaments.

This flower has petals that are purple and very stringy.

DTGAN (Text)

S2IGAN

Ours

(b) The Oxford-102 dataset

Fig. 5 Qualitative comparison of AttnGAN [24], DTGAN [3] conditioned on the textual descriptions, S2IGAN [9] and Fusion-S2iGan on the basis of the speech signals on the CUB and Oxford-102 data sets. The spoken descriptions are shown above the images

5.2.3 Human evaluation

To evaluate the image quality and semantic consistency of StackGAN++ [22], S2IGAN [9] and our Fusion-S2iGan, we conduct a human test on the CUB and Oxford-102 data sets. We selected 100 pictures randomly from each data set. Then, we provided the same spoken captions to users and asked them to choose the best sample synthesized by the three approaches based on image details and the corresponding speech description. In addition, we computed the final scores based on the judgment of two judges to ensure fairness. The results are shown in Fig. 6. We can see that

our presented Fusion-S2iGan performs better than Stack-GAN++ and S2IGAN on both data sets, particularly on the CUB data set, This indicates the effectiveness of our approach.

5.3 Results on the real speech datasets

In addition to the synthesized speech data sets, we also evaluate Fusion-S2iGan on the challenging real speech data sets, i.e., Flickr8k and Places-subset data sets, from both quantitative and qualitative perspectives.
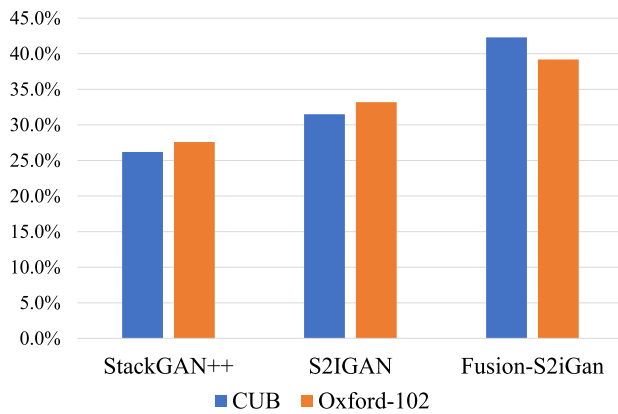
Fig. 6 Human test results (ratio of 1st) of StackGAN++ [22], S2IGAN [9] and our Fusion-S2iGan on the CUB and Oxford-102 data sets

5.3.1 Quantitative results

Table 2 reports the IS, FID and R@50 of Fusion-S2iGan and other compared approaches on the Flickr8k and Places-subset data sets. It can be observed that Fusion-S2iGan performs better than S2IGAN by notably enhancing the IS from 8.72 to 11.70 on the Flickr8k data set and from 4.04 to 5.05 on the Places-subset data set, reducing the FID from 93.29 to 70.80 on the Flickr8k data set and from 42.09 to 25.68 on the Places-subset data set and improving the R@50 from 16.40 to 34.95 on the Flickr8k data set and from 12.95 to 28.37 on the Places-subset data set. Notably, Fusion-S2iGan obtains a remarkably lower FID in speech-to-image synthesis than AttnGAN doing text-to-image generation on both data sets, which demonstrates that the distribution of the samples synthesized by our model is closer to the real data distribution. Specifically, Fusion-S2iGan reduces the FID from 84.08 to 70.80 on the Flickr8k data set and from 35.59 to 25.68 on the Places-subset data set while increasing the IS from 4.59 to 5.05 on the Places-subset data set. We can also see that Fusion-S2iGan gets a lower R@50 score than AttnGAN on both

data set. The reason for this may be that speech signals lack regular breaks between words, i.e., the spaces in the written-form text.

5.3.2 Qualitative results

For qualitative comparison, we visualize the pictures produced by Fusion-S2iGan and S2IGAN and corresponding ground-truth images on the Flickr8k and Places-subset data sets in Fig. 7, which suggests that Fusion-S2iGan is capable of yielding photo-realistic and semantic-consistency photographs when given speech signals. For example, in terms of complex-scene generation, Fusion-S2iGan generates visually plausible dogs with more vivid details and more clear backgrounds than S2IGAN in the 1st, 5th, 6th and 8th column. We can also observe that Fusion-S2iGan synthesizes a man surfing on the realistic sea waves (2nd column), two clear persons standing on the beach (3rd column), a plausible football player in red and a reasonable background (4th column) and a running girl (4th column), whereas S2IGAN yields unclear objects (1st, 2nd, 4th, 5th, 6th and 8th column) and blurry backgrounds (4th and 7th column). Furthermore, it can be seen that the number of the objects produced by Fusion-S2iGan is correct in the 1st, 2nd, 3rd, 5th, 6th, 7th and 8th column.

As can be seen in Fig. 7b, Fusion-S2iGan produces promising and high-quality complex-scene pictures, i.e., a realistic bedroom (1st column), plausible kitchens (2nd and 8th column), high-quality living rooms (3rd and 5th column), visually promising dining rooms (4th and 7th column) and a photo-realistic hotel room (6th column), although the Places-subset data set is very challenging. However, some examples generated by S2IGAN (1st, 4th, 6th, 7th and 8th column) are not plausible. For instance, the shape of the beds is not clear (1st and 6th column) and the color distribution is rough (4th and 8th column).

The above analysis indicates that Fusion-S2iGan has the capacity to achieve very good results on the real speech
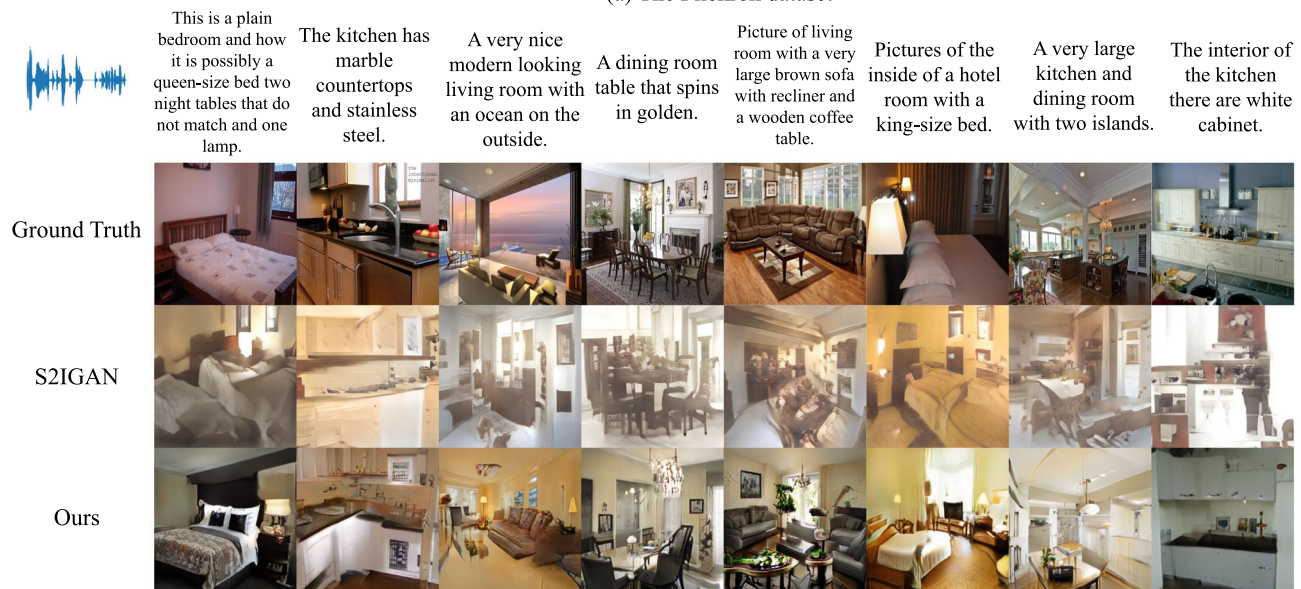
Table 2 The IS, FID and R@50 of AttnGAN [24] based on the textual descriptions, Li et al., StackGAN++ [22], S2IGAN [9] and Fusion-S2iGan conditioned on the spoken captions on the Flickr8k and Places-subset data sets

Datasets	Methods	Input	IS ↑	FID ↓	R@50 ↑
Flickr8k	AttnGAN [24]	Text	12.37±0.41	84.08	50.40
	StackGAN++ [22]	Speech	8.36 ±0.39	101.74	16.40
	S2IGAN [9]	Speech	8.72±0.34	93.29	16.40
	Fusion-S2iGan	Speech	**11.70±0.45**	**70.80**	**34.95**
Places-subset	AttnGAN [24]	Text	4.59±0.51	35.59	33.85
	Li et al.	speech	−	83.06	−
	StackGAN++ [22]	Speech	3.78±0.35	47.94	8.87
	S2IGAN [9]	Speech	4.04±0.25	42.09	12.95
	Fusion-S2iGan	Speech	**5.05±0.10**	**25.68**	**28.37**

The best results of speech-to-image generation are in bold

A small black and white dog running through the grass with a tennis ball in his mouth.

A person rides the waves on a surfboard.

A little girl and boy play in the sand on the beach.

Coach talks with football player, other players and crowd in background.

A dog swimming though water holding a stick.

A dog runs in the snow.

The little girl is running and laughing.

Two dogs playing in a lake.

Ground Truth

S2IGAN

Ours

(a) The Flickr8k dataset

This is a plain bedroom and how it is possibly a queen-size bed two night tables that do not match and one lamp.

The kitchen has marble countertops and stainless steel.

A very nice modern looking living room with an ocean on the outside.

A dining room table that spins in golden.

Picture of living room with a very large brown sofa with recliner and a wooden coffee table.

Pictures of the inside of a hotel room with a king-size bed.

A very large kitchen and dining room with two islands.

The interior of the kitchen there are white cabinet.

Ground Truth

S2IGAN

Ours

(b) The Places-subset dataset

Fig. 7 Qualitative comparison of S2IGAN [9] and Fusion-S2iGan conditioned on the speech signals on the Flickr8k and Places-subset data sets. Ground-truth images are shown above the samples produced by S2IGAN and the spoken descriptions are presented above the ground-truth images

data sets due to the effective speech-visual fusion module and loss functions.

5.3.3 Human evaluation

To further assess the image quality and semantic consistency of StackGAN++, S2IGAN, and our proposed Fusion-S2iGan, we perform a human test on the Flickr8k and Places-subset data sets using the identical process as in Sect. 5.2.3. Figure 8 shows that our approach significantly performs better than StackGAN++ and S2IGAN on both

data sets, demonstrating the superiority of our proposed Fusion-S2iGan.

5.4 Ablation/substitution tests of the proposed approach

In order to evaluate the effectiveness of different components in Fusion-S2iGan, we perform a series of substitution tests on the CUB, Oxford-102 and Flickr8k and Places-subset data sets. Ablation usually refers to the removal of processing steps and evaluating the effects on performance.
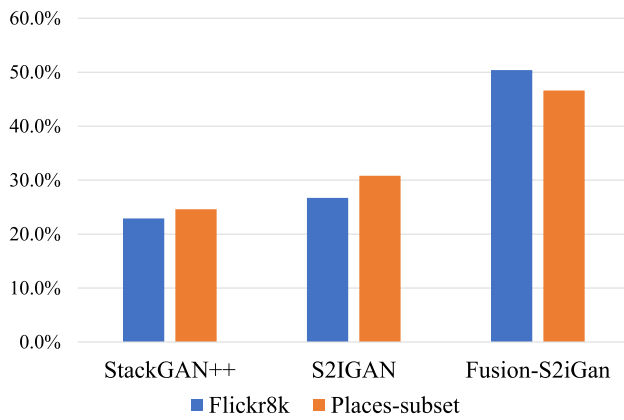
Fig. 8 Human test results (ratio of 1st) of StackGAN++ [22], S2IGAN [9] and our Fusion-S2iGan on the Flickr8k and Places-subset data sets

Here, however, the word 'substitution' is more appropriate, referring to the in-place replacement of a network module by another variant. In the evaluation we used 30k sample images for the CUB data set and 10k samples for the Oxford-102 data set in order to obtain reliable performance estimates.

5.4.1 Effectiveness of the pixel-attention module (PAM)

To validate the effectiveness of the introduced pixel-attention module (PAM), we investigate the performance of Fusion-S2iGan with other attention modules on the CUB and Oxford-102 data sets. Specifically, we replace the PAM in the visual+speech fusion module (VSFM) using the bottleneck attention module (BAM) [54], polarized self-attention (PolarizedAttn) [55] and contextual transformer attention (CoTAttn) [56], respectively. The comparison results are depicted in Table 3. We can see that PAM outperforms BAM, PolarizedAttn and CoTAttn by increasing the IS by 0.12 on the CUB data set and 0.06 on the Oxford-102 data set, reducing the FID by 0.33 on the CUB data set and 1.16 on the Oxford-102 data set and enhancing the mAP by 0.75 on the CUB data set and 0.44 on the Oxford-102 data set. The results demonstrate that

PAM can improve the quality and semantic relevancy of produced pictures, relative to the other methods.

5.4.2 Effectiveness of the weighted-fusion module (WFM)

To further prove the benefits of the proposed weighted-fusion module (WFM), we conduct a substitution test on fusion methods. We replace WFM with the element-wise addition and multiplication operations, respectively. Table 4 reports the quantitative results on the CUB and Flickr8k data sets. It can be observed that WFM performs better than both addition and product operations, significantly improving the IS from 4.68 to 4.82 on the CUB data set and from 10.23 to 11.70 on the Flickr8k data set, reducing the FID from 15.25 to 13.74 on the CUB data set and from 78.03 to 70.80 on the Flickr8k data set and increasing the mAP from 11.11 to 11.49 on the CUB data set and the R@50 from 30.82 to 30.95 on the Flickr8k data set. The above analysis suggests the effectiveness of the presented WFM in comparison to plain additive or product-based weighing.

5.4.3 Effectiveness of the visual+speech fusion module (VSFM)

To verify the effectiveness of the developed visual+speech fusion module (VSFM) and investigate how far can current single-stage text-to-image algorithms be used for speech-to-image generation, we explore the results of Fusion-S2iGan with the sentence-visual fusion modules in DTGAN [3] and DFGAN [28] on the CUB and Oxford-102 data sets. The results of the substitution study are displayed in Table 5. We can see that the VSFM achieves the best score, notably increasing the IS by 0.42 on the CUB data set and 0.23 on the Oxford-102 data set, reducing the FID by 3.00 on the CUB data set and 8.29 on the Oxford-102 data set and increasing the mAP by 3.29 on the CUB data set and 1.40 on the Oxford-102 data set.

It can also be observed that DTGAN does not obtain promising performance on the speech data sets. The reason behind this result may be that the dual-attention module in

Table 3 Substitution study on pixel-attention modules

Datasets	CUB			Oxford-102		
Evaluation Metrics	IS ↑	FID ↓	mAP ↑	IS ↑	FID ↓	mAP ↑
BAM [54]	4.35±0.04	16.23	10.32	3.62±0.09	42.72	16.35
PolarizedAttn [55]	4.68±0.04	14.33	10.74	3.75±0.07	44.20	16.42
CoTAttn [56]	4.70±0.05	14.07	10.53	3.75±0.06	41.24	17.28
PAM	**5.06±0.09**	**13.09**	**12.12**	**3.81±0.08**	**40.08**	**17.72**

BAM, PolarizedAttn and CoTAttn refer to the bottleneck attention module [54], polarized self-attention [55] and contextual transformer attention [56], respectively. PAM indicates the introduced pixel-attention module. The best results are in bold

Table 4 Substitution test on fusion methods

Datasets	CUB			Flickr8k		
Evaluation Metrics	IS ↑	FID ↓	mAP ↑	IS ↑	FID ↓	R@50 ↑
Addition	3.70±0.04	29.11	2.7	9.53±0.37	100.83	5.98
Product	4.68±0.04	15.25	11.11	10.23±0.36	78.03	30.82
WFM	**5.06±0.09**	**13.09**	**12.12**	**11.70±0.45**	**70.80**	**34.95**

Addition and Product represent the element-wise addition and product operation, respectively. WFM denotes the proposed weighted-fusion module. The best results are in bold

Table 5 Substitution study on the visual+speech fusion module (VSFM)

Datasets	CUB			Oxford-102		
Evaluation Metrics	IS ↑	FID ↓	mAP ↑	IS ↑	FID ↓	mAP ↑
DTGAN [53]	4.18 ±0.04	24.70	4.92	3.15±0.04	52.67	13.87
DF-GAN [28]	4.40 ±0.03	16.74	8.20	3.58±0.07	48.37	16.32
VSFM	**5.06±0.09**	**13.09**	**12.12**	**3.81±0.08**	**40.08**	**17.72**

DTGAN and DF-GAN refer to the sentence-visual fusion module in DTGAN [3] and DF-GAN [28], respectively. VSFM denotes the presented visual+speech fusion module. The best results are in bold

Table 6 Effect of the number of fusion modules in the dual-residual speech-visual fusion block

Datasets	Flickr8k			Places-subset		
Evaluation Metrics	IS ↑	FID ↓	mAP ↑	IS ↑	FID ↓	mAP ↑
single-module	10.39±0.34	81.62	32.37	5.02±0.06	41.67	24.00
dual-module	**11.70±0.45**	**70.80**	**34.95**	**5.05±0.10**	**25.68**	**28.37**

Single-module and dual-module indicate the residual structure with a single fusion module and two fusion modules, respectively. The best scores are in bold

DTGAN is not suitable for processing the continuous speech signals and modeling the semantic relationships between vision and speech. The above experimental results indicate that with the VSFM, Fusion-S2iGan is able to yield high-resolution pictures semantically aligning with the input speech-audio descriptions.

5.4.4 Effectiveness of the number of fusion modules in the residual-structure

To evaluate the impact of different number of the fusion modules in the dual-residual speech-visual fusion block, we compare the results of a fusion module (single-module) and two fusion modules (dual-module) on the Flickr8k and Places-subset data sets, as shown in Table 6. We can observe that the proposed dual-residual speech-visual fusion block outperforms the block with a single module on both data set, which validates the effectiveness of the dual-residual structure.

6 Conclusion

In this paper, we propose a unified, novel end-to-end speech-to-image transform architecture named Fusion-S2iGan. Fusion-S2iGan is able to only use a generator/discriminator pair to project a speech signal into a high-quality picture directly. Fusion-S2iGan adopts a new and effective visual+speech fusion module (VSFM) to modulate the visual feature maps with the speech information and boost the resolution of produced samples. More significantly, Fusion-S2iGan spreads the bimodal information over all layers of the generative model. This allows for an influence of the speech over features at various hierarchical levels in the architecture, from crude early features to abstract late features. The hinge objective, deep attentional multimodal similarity model (DAMSM) loss and matching-aware zero-centered gradient penalty (MA-GP) loss are introduced to stabilize the training of the conditional generative adversarial network (cGAN). Fusion-S2iGan is evaluated on four public data sets, i.e., CUB bird, Oxford-102, Flickr8k and Places-subset, outperforming current speech-to-image methods. Moreover, we explore whether

current text-to-image approaches are effective for a speech-to-image transform. Furthermore, our presented VSFM is a general fusion module, and can be easily integrated into current speech-to-image frameworks to improve image quality and semantic consistency. More significantly, our developed architecture overcomes the issues presenting in the existing multi-stage speech-to-image algorithms, and can serve as a strong basis for developing better speech-to-image generation models. In the future, we will investigate how to inject the emotional features derived from the speech signal into the speech-to-image synthesis pipeline and how to combine speech signals with other conditional contexts, e.g.,, sketches, to control the produced pictures. We focus on producing natural-scene pictures from spoken descriptions. However, it is intriguing and attractive to study image generation from other domains such as artistic fonts, drawings, cartoons, sketch and paintings, since they play a significant role in humans' daily life.

Data availibility statement The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable and motivated request.

Declarations

Conflict of interest The authors declared that they have no conflicts of interest in this work.

References

1. Zai El Amri W, Reinhart F, Schenck W (2022) Open set task augmentation facilitates generalization of deep neural networks trained on small data sets. Neural Comput Appl 34(8):6067–6083
2. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781
3. Zhang Z, Schomaker L (2021) Dtgan: Dual attention generative adversarial networks for text-to-image generation. In: 2021 international joint conference on neural networks (IJCNN), pp 1–8. IEEE
4. Zhang Z, Schomaker L (2021) Divergan: An efficient and effective single-stage framework for diverse text-to-image generation. Neurocomputing
5. Yuan Q, Dai Y, Li G (2023) Exploration of english speech translation recognition based on the lstm rnn algorithm. Neural Comput Appl, 1–10
6. Albuquerque RQ, Mello CA (2021) Automatic no-reference speech quality assessment with convolutional neural networks. Neural Comput Appl 33:9993–10003
7. Obin N, Lanchantin P (2015) Symbolic modeling of prosody: from linguistics to statistics. IEEE/ACM Trans Audio Speech Lang Process 23(3):588–599
8. Du C, Yu K (2021) Phone-level prosody modeling with gmm-based mdn for diverse and controllable speech synthesis. IEEE/ACM Trans Audio Speech Lang Process 30:190–201
9. Wang X, Qiao T, Zhu J, Hanjalic A, Scharenborg O (2021) Generating images from spoken descriptions. IEEE/ACM Trans Audio Speech Lang Process 29:850–865
10. Li J, Zhang X, Jia C, Xu J, Zhang L, Wang Y, Ma S, Gao W (2020) Direct speech-to-image translation. IEEE J Sel Top Signal Process 14(3):517–529
11. Kadyan V, Bawa P (2022) Transfer learning through perturbation-based in-domain spectrogram augmentation for adult speech recognition. Neural Comput Appl 34:21015–21033
12. Kadyan V, Dua M, Dhiman P (2021) Enhancing accuracy of long contextual dependencies for Punjabi speech recognition system using deep lstm. Int J Speech Technol 24:517–527
13. Kadyan V, Bawa P, Hasija T (2021) In domain training data augmentation on noise robust punjabi children speech recognition. J Ambient Intell Humaniz Comput 13:2705–2721
14. Zhu M, Pan P, Chen W, Yang Y (2019) Dm-gan: dynamic memory generative adversarial networks for text-to-image synthesis. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5802–5810
15. Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B (2022) High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10684–10695
16. Ramesh A, Dhariwal P, Nichol A, Chu C, Chen M (2022) Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125
17. Gafni O, Polyak A, Ashual O, Sheynin S, Parikh D, Taigman Y (2022) Make-a-scene: Scene-based text-to-image generation with human priors. arXiv preprint arXiv:2203.13131
18. Wah C, Branson S, Welinder P, Perona P, Belongie S (2011) The caltech-ucsd birds-200-2011 dataset
19. Nilsback M-E, Zisserman A (2008) Automated flower classification over a large number of classes. In: 2008 sixth Indian conference on computer vision, graphics & image processing, pp 722–729. IEEE
20. Hodosh M, Young P, Hockenmaier J (2013) Framing image description as a ranking task: Data, models and evaluation metrics. J Artif Intell Res 47:853–899
21. Zhou B, Lapedriza A, Xiao J, Torralba A, Oliva A (2014) Learning deep features for scene recognition using places database. Adv Neural Inf Process Syst 27
22. Zhang H, Xu T, Li H, Zhang S, Wang X, Huang X, Metaxas DN (2018) Stackgan++: Realistic image synthesis with stacked generative adversarial networks. IEEE Trans Pattern Anal Mach Intell 41(8):1947–1962
23. Mirza M, Osindero S (2014) Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784
24. Xu T, Zhang P, Huang Q, Zhang H, Gan Z, Huang X, He X (2018) Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In: Proceedings of

the IEEE conference on computer vision and pattern recognition, pp 1316–1324

25. Qiao T, Zhang J, Xu D, Tao D (2019) Mirrorgan: Learning text-to-image generation by redescription. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1505–1514

26. Liang J, Pei W, Lu F (2020) Cpgan: Content-parsing generative adversarial networks for text-to-image synthesis. In: European conference on computer vision, pp 491–508. Springer

27. Radford A, Metz L, Chintala S (2015) Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434

28. Tao M, Tang H, Wu S, Sebe N, Wu F, Jing X-Y (2020) Df-gan: Deep fusion generative adversarial networks for text-to-image synthesis. arXiv preprint arXiv:2008.05865

29. Zhang Z, Schomaker L (2022) Optimized latent-code selection for explainable conditional text-to-image gans. arXiv preprint arXiv:2204.12678

30. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: Advances in neural information processing systems, pp 2672–2680

31. Chen L, Srivastava S, Duan Z, Xu C (2017) Deep cross-modal audio-visual generation. Proc Themat Workshops ACM Multimedia 2017:349–357

32. Hao W, Zhang Z, Guan H (2018) Cmcgan: A uniform framework for cross-modal visual-audio mutual generation. In: Proceedings of the AAAI conference on artificial intelligence, vol 32

33. Duarte AC, Roldan F, Tubau M, Escur J, Pascual S, Salvador A, Mohedano E, McGuinness K, Torres J, Giro-i-Nieto X (2019) Wav2pix: speech-conditioned face generation using generative adversarial networks. In: ICASSP, pp 8633–8637

34. Oh T-H, Dekel T, Kim C, Mosseri I, Freeman WT, Rubinstein M, Matusik W (2019) Speech2face: Learning the face behind a voice. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 7539–7548

35. Haque KN, Rana R, Liu J, Hansen JH, Cummins N, Busso C, Schuller BW (2021) Guided generative adversarial neural network for representation learning and audio generation using fewer labeled audio data. IEEE/ACM Trans Audio, Speech Lang Process 29:2575–2590

36. Santurkar S, Tsipras D, Ilyas A, Madry A (2018) How does batch normalization help optimization? In: Adv Neural Inf Process Syst pp 2483–2493

37. Maas AL, Hannun AY, Ng AY, et al. (2013) Rectifier nonlinearities improve neural network acoustic models. In: Proceedings of the Icml, 30, 3 . Citeseer

38. Hendrycks D, Gimpel K (2016) Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415

39. Lim JH, Ye JC (2017) Geometric gan. arXiv preprint arXiv:1705.02894

40. Shen J, Pang R, Weiss RJ, Schuster M, Jaitly N, Yang Z, Chen Z, Zhang Y, Wang Y, Skerrv-Ryan R, et al. (2018) Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 4779–4783. IEEE

41. Harwath D, Glass J (2015) Deep multimodal semantic embeddings for speech and images. In: 2015 IEEE workshop on automatic speech recognition and understanding (ASRU), pp 237–244. IEEE

42. Harwath D, Recasens A, Surís D, Chuang G, Torralba A, Glass J (2018) Jointly discovering visual objects and spoken words from raw sensory input. In: Proceedings of the European conference on computer vision (ECCV), pp 649–665

43. Harwath D, Torralba A, Glass J (2016) Unsupervised learning of spoken language with visual context. Adv Neural Inf Process Syst 29

44. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2818–2826

45. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M et al (2015) Imagenet large scale visual recognition challenge. Int J Comput Vision 115(3):211–252

46. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778

47. Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980

48. Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S (2017) Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in neural information processing systems, pp 6626–6637

49. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, et al. (2019) Pytorch: An imperative style, high-performance deep learning library. In: Advances in neural information processing systems, pp 8026–8037

50. Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X (2016) Improved techniques for training gans. In: Advances in neural information processing systems, pp 2234–2242

51. Li B, Qi X, Lukasiewicz T, Torr P (2019) Controllable text-to-image generation. In: Advances in neural information processing systems, pp 2065–2075

52. Yin G, Liu B, Sheng L, Yu N, Wang X, Shao J (2019) Semantics disentangling for text-to-image generation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2327–2336

53. Zhang Z, Schomaker L (2020) Dtgan: Dual attention generative adversarial networks for text-to-image generation. arXiv preprint arXiv:2011.02709

54. Park J, Woo S, Lee J-Y, Kweon IS (2018) Bam: Bottleneck attention module. arXiv preprint arXiv:1807.06514

55. Liu H, Liu F, Fan X, Huang D (2021) Polarized self-attention: toward high-quality pixel-wise regression. arXiv preprint arXiv:2107.00782

56. Li Y, Yao T, Pan Y, Mei T (2022) Contextual transformer networks for visual recognition. IEEE Trans Pattern Anal Mach Intell