

## University of Groningen

### Fusion-S2iGan

Zhang, Zhenxing; Schomaker, Lambert

**IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.**

*Document Version*

Early version, also known as pre-print

*Publication date:*

2023

[Link to publication in University of Groningen/UMCG research database](#)

*Citation for published version (APA):*

Zhang, Z., & Schomaker, L. (2023). *Fusion-S2iGan: An Efficient and Effective Single-Stage Framework for Speech-to-Image Generation*. arXiv. <http://arxiv.org/abs/2305.10126v1>

**Copyright**

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

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

**Take-down policy**

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

*Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.*

# Fusion-S2iGan: An Efficient and Effective Single-Stage Framework for Speech-to-Image Generation

Zhenxing Zhang and Lambert Schomaker

**Abstract**—The goal of a speech-to-image transform is to produce a photo-realistic picture directly from a speech signal. Recently, various studies have focused on this task and have achieved promising performance. However, current speech-to-image approaches are based on a stacked modular framework that suffers from three vital issues: 1) Training separate networks is time-consuming as well as inefficient and the convergence of the final generative model strongly depends on the previous generators; 2) The quality of precursor images is ignored by this architecture; 3) Multiple discriminator networks are required to be trained. To this end, 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 given spoken descriptions. Fusion-S2iGan introduces a visual+speech fusion module (VSFM), constructed 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. 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. We conduct a series of experiments on four benchmark data sets, i.e., CUB birds, Oxford-102, Flickr8k and Places-subset. The experimental results demonstrate the superiority of the presented 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.

**Index Terms**—speech-to-image transform, single-stage architecture, generative adversarial network, attention mechanism and fusion module

## I. 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. Consider, for example, the ‘telephone test’, where a human describes a 5-image cartoon or comic strip by speech and the system needs to provide its visual associations to reconstruct the original image input. Incongruencies between original and reconstruction may be a useful source of information for continuous learning, both in terms of visual and textual congruence. As an example, a spoken original cartoon description can be visualized by the speech-to-image GAN, its output can be captioned [2] and this text can be compared to the recognized spoken description. From the textual discrepancies a loss can be computed that is used for training and improving consistency in a cyclic manner.

Moreover, researching speech-to-image synthesis may help cognitive-science researchers to investigate how humans gradually understand the world. Several studies [3]–[5] reveal that human infants know many aspects of their native language by perceiving and analyzing the features of the speech signal such as consonants and vowels. More significantly, infants are able to understand the meanings of numerous ordinary nouns through daily experience with language before they are capable of reading [6]. For example, infants will attend to the banana present in their visual field when hearing ‘Look, a banana!’ from their parents. This illustrates that strong connections between vision and speech signals are being created at this early age and infants acquire the capability of discovering the semantic relationships between the spoken words and the visible objects. 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 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 [7].

The task of speech-to image generation has a variety of advantages over text-to-image synthesis [8], [9], whose goal

Zhenxing Zhang and Lambert Schomaker are with the Bernoulli Institute, University of Groningen, Groningen 9747, The Netherlands (e-mail: z.zhang@rug.nl; l.r.b.schomaker@rug.nl.)

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 [10], [11]. Prosody allows speakers to express the attitude and emotional states, providing key non-verbal cues that help the listener to understand the spoken description [12]. 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 will also be very beneficial for their applications, where users can produce desired photographs by adjusting intonation. Since current text-to-speech synthesis models with deep-learning techniques allow for prosody modulation [11], synthesizing pictures on the basis of emotional features and linguistic information of speech signals should be feasible using such produced training data.

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 3,500 languages lacking orthography or written form [13], 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. Besides, spoken descriptions contain not only the linguistic content but also the non-verbal cues, e.g., emotion, intent and attitude of a speaker, which are significant for generated

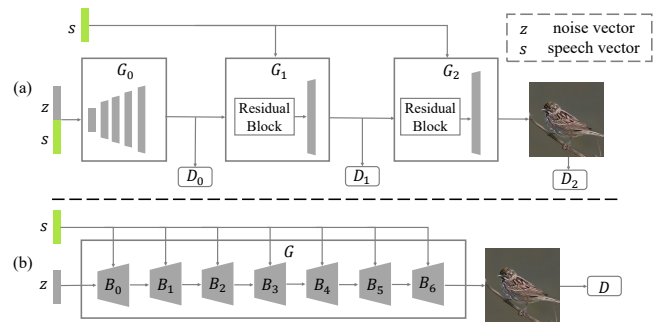


Fig. 1. The comparison between the stacked framework and the proposed architecture. The multi-stage pipeline (a) entails training separate generators to obtain high-quality samples. The presented Fusion-S2iGan (b) is capable of producing visually plausible pictures only employing a single generator/discriminator 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 Section III, and  $G$  and  $D$  are the generator network and discriminator network, respectively.

samples. These features present another crucial challenge for a speech-to-image generation system: *How to effectively inject these features into a neural network which has a large 2D color image as its output?* Note that the common speech-text recognition step is not necessary when using speech-audio descriptions to image generation and emotional features in speech signals would additionally be lost. Unfortunately, there are as yet no suitable data sets with emotional speech in different degrees, describing a corresponding image in each sample. However, as a first step in the direction of exploiting the potential of speech-audio we will focus on non-emotional speech-to-image transforms.

We are not the first to attempt to translate speech signals into high-resolution pictures directly. For example, Li et al. [14] 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. [13] 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. 1(a)), 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 [9]. 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 [8]. Second, the quality of the outputs of the previous generator networks [15] 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 [8]. 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 [8] and DiverGAN [9], 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, we presented dual-attention models, conditional adaptive instance-layer normalization and a new type of visual loss. In DiverGAN, we extended the sentence-level attention models introduced in DTGAN to word-level attention modules, in order 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 in order to overcome the problems of a stacked framework, we introduce a visual+speech fusion module as well as several effective loss functions, which contribute to a new single-stage speech-to-image architecture.

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. 1(b)) for speech-to-image transforms, which is capable of producing high-quality and semantically consistent pictures only using a generator/discriminator pair.
- We design a visual+speech fusion module (VSFM) 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.
- We carry out extensive experiments on four benchmark data sets, i.e., CUB bird [16], Oxford-102 [17], Flickr8k [18] and Places-subset [19]. 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++ [20], Li *et al.* [14] and S2iGAN [13].
- 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 Section VIII-C.

The remainder of this paper is organized as follows. Section II reviews related works. In Section III, we introduce the architecture of the proposed Fusion-S2iGan in detail. In Section IV, we elaborate on the presented visual+speech fusion module. Experimental settings are discussed in Section V and results are reported in Section VI, VII and VIII. Section IX draws the conclusions.

## II. RELATED WORK

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

### A. Text-to-Image Generation

Many text-to-image synthesis approaches are built upon the original conditional generative adversarial network (cGAN) [21] due to its appealing performance. We roughly group them into two categories in terms of the number of the generators and the discriminators they use.

1) *Multi-stage models*: Zhang et al. [20] 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. [22] 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. [23] developed MirrorGAN that introduced an image-to-text model to ensure that synthesized pictures are semantically related to given textual descriptions. Zhu et al. [15] built DMGAN where a dynamic-memory module is applied to enhance the image quality in the initial stage. CPGAN [24] 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) *Single-stage methods*: Reed et al. [25] 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. [26] 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. [8] 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. [9] 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. [27] 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.

### B. Speech-audio-to-Image Synthesis

With the recent rapid advances in generative-adversarial networks (GANs) [28] and conditional GANs (cGANs) [21], 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. [29] made the first attempt to use the cGAN to produce samples on

the basis of music audio. Hao et al. [30] 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. [31] 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. [32] 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. [14] 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. [13] 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 only using a single generator/discriminator pair.

### III. FUSION-S2iGAN FOR SPEECH-TO-IMAGE GENERATION

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.

#### A. Preliminary

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  $S_i = (s_i^1, s_i^2, \dots, s_i^k)$  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  $(\hat{I}_i, s_i)$ .

#### B. Generator

The generator is able to project a speech signal into a photo-realistic and semantically consistent picture, shown in Fig. 2(a). 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 [33].

The designed dual-residual speech-visual fusion block (see Fig. 2(c)) contains two effective and efficient visual+speech fusion modules (see Fig. 2(d)) as well as a suite of ReLU activation functions and convolutional layers. Each visual+speech fusion module comprises Batch Normalization (BN) [34], 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:

$$h_0 = F_0(z) \quad (1)$$

$$h_1 = F_1^{Dual}(h_0, s) \quad (2)$$

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

$$o = G_c(h_7) \quad (4)$$

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 Section IV.

#### C. Discriminator

The architecture of the discriminator network comprises an image encoder, a pre-trained speech encoder and a conditional discriminator network, depicted in Fig. 2(b). 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 [35] 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. 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.

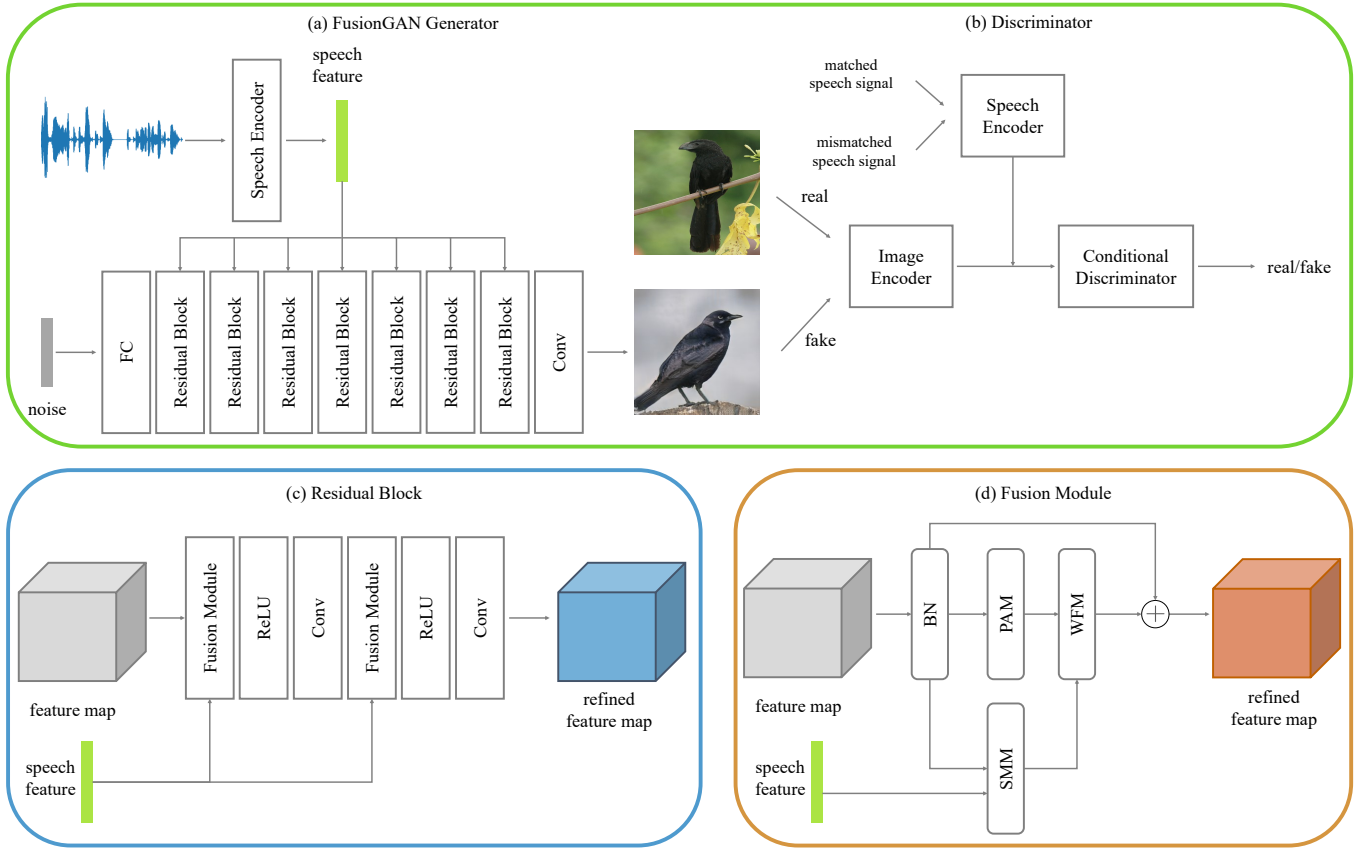


Fig. 2. The overall framework of the proposed Fusion-S2iGAN. *FC* is a dense layer, *Conv* is a convolutional layer, *ReLU* is a ReLU activation function and *BN* is a Batch-Normalization operation. Additionally, *Residual Block* and *Fusion Module* refer to the presented dual-residual speech-visual fusion block (see (c)) discussed in Section III-B and visual+speech fusion module (see (d)) discussed in Section IV, respectively. Furthermore, *PAM*, *SMM* and *WFM* in (d) represent the pixel-attention module, speech-modulation module and weighted-fusion module, respectively, discussed in Section IV. Note that we do not plot the up-sample layers between *Residual Blocks* in (a) due to the limited space.

#### D. Objective Function

An adversarial loss is employed to match generated samples to input speech signals. Inspired by [9], we utilize the hinge objective [36] for stable training instead of the vanilla GAN objective. The adversarial loss for the discriminator is defined as follows:

$$\begin{aligned} \mathcal{L}_{\text{adv}}^D = & \mathbb{E}_{x \sim p_{\text{data}}} [\max(0, 1 - D(x, s))] \\ & + \frac{1}{2} \mathbb{E}_{x \sim p_G} [\max(0, 1 + D(\hat{x}, s))] \\ & + \frac{1}{2} \mathbb{E}_{x \sim p_{\text{data}}} [\max(0, 1 + D(x, \hat{s}))] \end{aligned} \quad (5)$$

where  $s$  is a given spoken caption,  $\hat{s}$  is a mismatched speech-audio description,  $x$  is the real image from the distribution  $p_{\text{data}}$  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 [26] for the discriminator, which applies gradient penalty to real images and input spoken descriptions. The MA-GP Loss is formulated as follows:

$$\mathcal{L}_M = \mathbb{E}_{x \sim p_{\text{data}}} [(\|\nabla_x D(x, s)\|_2 + \|\nabla_s D(x, s)\|_2)^p] \quad (6)$$

For the generator, we apply an adversarial loss and a deep attentional multimodal similarity model (DAMSM) loss [22] to train the network.

#### IV. 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 [8] and then can be made to modulate feature maps in the network. In this section, we develop an effective and efficient visual+speech

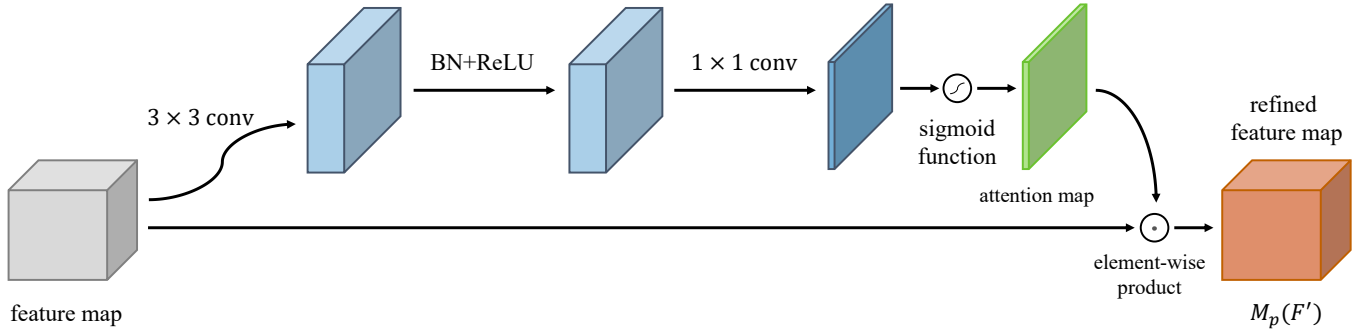


Fig. 3. Overview of the introduced pixel-attention module, which aims to assign larger weights to discriminative and informative pixels. *BN* and *ReLU* refer to Batch Normalization and a ReLU activation function, respectively.  $1 \times 1 \text{ conv}$  and  $3 \times 3 \text{ conv}$  indicate the  $1 \times 1$  and  $3 \times 3$  convolutional operation, respectively.

fusion module (VSFM) to facilitate the visual feature maps with the global speech embedding and yield high-resolution pictures.

#### A. Overall Architecture

The framework of the proposed visual+speech fusion module is shown in Fig. 2(d). 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 VSFM first applies Batch Normalization (BN) [34] 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. After that, the VSFM refines  $F' \in \mathbb{R}^{C \times H \times W}$  using the pixel-attention module (PAM)  $M_p$  and speech-modulation module (SMM)  $M_s$ , respectively. Afterwards, 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:

$$F' = BN(F) \quad (7)$$

$$F'' = F' + M_f(M_p(F'), M_s(F', s)) \quad (8)$$

where *BN* indicates a Batch-Normalization operation. The details of PAM, SMM and WFM will be described in the following subsections.

#### B. 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.

Fig. 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  $3 \times 3$  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  $1 \times 1$  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,

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

where  $f$  is a convolutional layer, *ReLU* is the ReLU function and  $\sigma$  is the sigmoid function. We conduct a matrix multiplication between the original feature map  $F' \in \mathbb{R}^{C \times H \times W}$  and the spatial-attention map  $PA \in \mathbb{R}^{1 \times H \times W}$  to acquire the refined result. Specifically,

$$M_p(F') = F' \odot PA \quad (10)$$

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

#### C. Speech-Modulation Module (SMM)

An adequate modulation method needs to be developed to ensure the semantic consistency and quality of synthesized photographs [9]. 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 [26], 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:

$$WA = MLP(ReLU(MLP(s))) \quad (11)$$

$$BA = MLP(ReLU(MLP(s))) \quad (12)$$

$$M_s(F') = F' \odot WA + BA \quad (13)$$

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

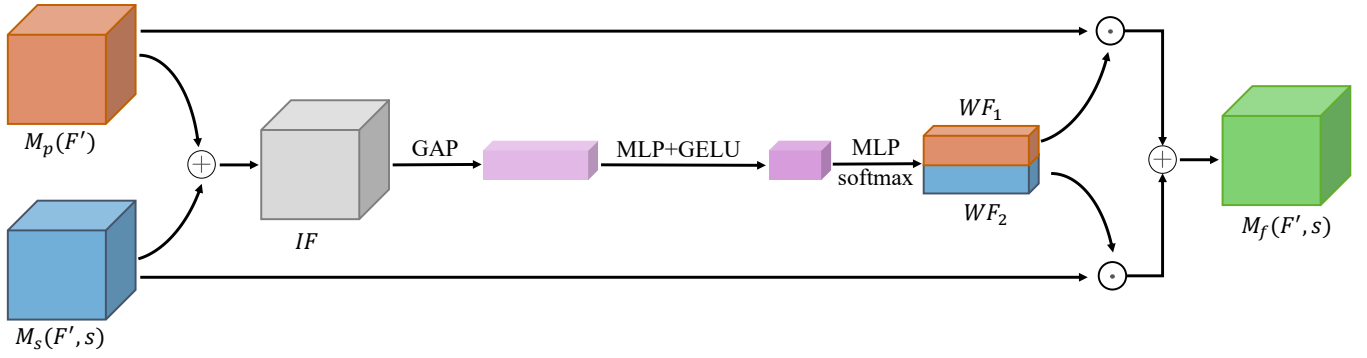


Fig. 4. Overview of the proposed weighted-fusion module, which effectively combines the outputs from the pixel-attention module (PAM) and speech-modulation module (SMM) in an adaptive manner.  $M_p(F')$  and  $M_s(F', s)$  denote the result of PAM and SMM, respectively.  $GAP$ ,  $MLP$  and  $GELU$  represent the global-average pooling, fully-connected perceptron layer and Gaussian Error Linear Unit (GELU) [37], respectively.  $WF_1$  and  $WF_2$  are the channel-aware weight matrix.  $M_f(F', s)$  is the final refined result.

#### D. 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  $WF \in \mathbb{R}^{C \times 2}$ . Afterwards,  $WF$  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,

$$IF = M_p(F') + M_s(F', s) \quad (14)$$

$$IF' = GAP(IF) \quad (15)$$

$$WF = softmax(MLP(GELU(MLP(IF')))) \quad (16)$$

$$WF = [WF_1; WF_2] \quad (17)$$

where  $GAP$  is the global-average pooling and  $GELU$  is the Gaussian Error Linear Unit (GELU) [37].  $WF_1$  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:

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

## V. EXPERIMENTAL SETTINGS

### A. Data Sets

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 set, which are employed by Li *et al.* [14] and S2iGAN [13].

- **CUB bird [16].** 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.* [14] and S2iGAN [13] transform textual descriptions to spoken captions utilizing a text-to-speech method such as Tacotron 2 [38].

- **Oxford-102 [17].** 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 [18].** 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 [39]. We split the Flickr8k data set according to S2iGAN [13].

- **Places-subset [19]** The Places-subset data set is a subset of the Places Audio Caption data set [40], [41], which encompasses real spoken captions of pictures from the Places 205 data set [19]. 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.

### B. Implementation Details

For the speech encoder, following the structure of S2iGAN [13], 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 [42] pre-trained on ImageNet [43] in S2iGAN with the ResNet [44] trained on Places 205 [19]. The dimension of the speech vector is set to 1024. In the experiments, we utilize the Adam optimizer [45] with  $\beta_1 = 0.0$  and  $\beta_2 = 0.9$  to train the networks. Furthermore,



we follow the two timescale update rule (TTUR) [46] 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 [47]. All the experiments are conducted on a single NVIDIA Tesla V100 GPU (32 GB memory).

### C. Evaluation Metrics

We verify the effectiveness of Fusion-S2iGan by computing the following four extensively employed evaluation metrics:

- **Inception score (IS) [48]**. 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 [9].

- **Fréchet inception distance (FID) [42]**. 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(recall)@50 [13]**. The mAP and R(recall)@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%.

## VI. RESULTS ON THE SYNTHESIZED SPEECH DATASETS

### A. Quantitative Results

We compare Fusion-S2iGan with previous single-stage [9], [26], [49] and multi-stage [13]–[15], [20], [22], [23], [50], [51] 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 I. Note that the reported scores in Table I 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 [15] in text-to-image

TABLE I  
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. THE BEST RESULTS OF SPEECH-TO-IMAGE GENERATION ARE IN BOLD.

Datasets	Methods	Input	IS $\uparrow$	FID $\downarrow$	mAP $\uparrow$
CUB	StackGAN++ [20]	text	4.04 $\pm$ 0.05	26.07	7.01
	AttnGAN [22]	text	4.36 $\pm$ 0.03	23.98	–
	MirrorGAN [23]	text	4.56 $\pm$ 0.05	–	–
	ControlGAN [50]	text	4.58 $\pm$ 0.09	–	–
	SDGAN [51]	text	4.67 $\pm$ 0.09	–	–
	DM-GAN [15]	text	4.75 $\pm$ 0.07	16.09	–
	DF-GAN [26]	text	4.86 $\pm$ 0.04	19.24	–
	DTGAN [49]	text	4.88 $\pm$ 0.03	16.35	–
	DiverGAN [9]	text	4.98 $\pm$ 0.06	15.63	–
	Classifier-based [13]	speech	3.68 $\pm$ 0.04	43.76	–
	Li <i>et al.</i> [14]	speech	4.09 $\pm$ 0.04	18.37	–
	StackGAN++ [20]	speech	4.14 $\pm$ 0.04	18.94	8.09
	S2iGAN [13]	speech	4.29 $\pm$ 0.04	14.50	9.04
	<b>Fusion-S2iGan</b>	speech	<b>5.06<math>\pm</math>0.09</b>	<b>13.09</b>	<b>12.12</b>
Oxford-102	StackGAN++ [20]	text	3.26 $\pm$ 0.01	48.68	–
	DF-GAN [26]	text	3.71 $\pm$ 0.06	–	–
	DTGAN [49]	text	3.77 $\pm$ 0.06	–	–
	DiverGAN [9]	text	3.99 $\pm$ 0.05	–	–
	Classifier-based [13]	speech	3.30 $\pm$ 0.06	64.75	–
	Li <i>et al.</i> [14]	speech	3.23 $\pm$ 0.05	54.76	–
	StackGAN++ [20]	speech	3.69 $\pm$ 0.08	54.33	12.18
	S2iGAN [13]	speech	3.55 $\pm$ 0.04	48.64	13.40
	<b>Fusion-S2iGan</b>	speech	<b>3.81<math>\pm</math>0.08</b>	<b>40.08</b>	<b>17.72</b>

generation on the CUB data set, as seen from an improvement of the IS from 4.75 to 4.82 and DTGAN [8] 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.

### B. 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 [13], we utilize the public code <sup>1</sup> 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 <sup>2</sup>, the IS score for S2iGAN was the same for the raw (greyish) output and the enhanced output samples. As can be seen in Fig. 5(a), the shape of birds produced by AttnGAN [22] and S2iGAN [13] (2<sup>nd</sup>, 3<sup>rd</sup>, 6<sup>th</sup> and 8<sup>th</sup> column) is strange, the image details are lost (2<sup>nd</sup>, 3<sup>rd</sup>, 5<sup>th</sup>, 6<sup>th</sup> and 7<sup>th</sup> column) and the backgrounds are blurry (1<sup>st</sup>, 3<sup>rd</sup>, 6<sup>th</sup> and 8<sup>th</sup> 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

<sup>1</sup><https://github.com/xinshengwang/S2iGAN>

<sup>2</sup><https://zenodo.org/record/7014899#.YwN-xHZByUk>



Fig. 5. Qualitative comparison of AttnGAN [22], DTGAN [8] conditioned on the textual descriptions, S2IGAN [13] 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.

distributions compared to AttnGAN and S2IGAN in the 2<sup>nd</sup>, 3<sup>rd</sup>, 6<sup>th</sup>, 7<sup>th</sup> and 8<sup>th</sup> column. Furthermore, as shown in the 1<sup>st</sup>, 2<sup>nd</sup>, 4<sup>th</sup>, 6<sup>th</sup> and 8<sup>th</sup> column, Fusion-S2iGan produces visually plausible birds with more vivid details than AttnGAN and S2IGAN.

The qualitative results of DTGAN [8], S2IGAN [13] and Fusion-S2iGan on the Oxford-102 data set are illustrated in Fig. 5(b), 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 1<sup>st</sup>, 2<sup>nd</sup>, 6<sup>th</sup>, 7<sup>th</sup> and 8<sup>th</sup> column. In addition,

as shown in the 3<sup>th</sup>, 4<sup>th</sup>, 5<sup>th</sup> and 6<sup>th</sup> 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.

## VII. 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.

TABLE II

THE IS, FID AND R@50 OF ATTNGAN [22] BASED ON THE TEXTUAL DESCRIPTIONS, LI *et al.*, STACKGAN++ [20], S2IGAN [13] AND FUSION-S2IGAN CONDITIONED ON THE SPOKEN CAPTIONS ON THE FLICKR8K AND PLACES-SUBSET DATA SETS. THE BEST RESULTS OF SPEECH-TO-IMAGE GENERATION ARE IN BOLD.

Datasets	Methods	Input	IS $\uparrow$	FID $\downarrow$	R@50 $\uparrow$
Flickr8k	AttnGAN [22]	text	12.37 $\pm$ 0.41	84.08	50.40
	StackGAN++ [20]	speech	8.36 $\pm$ 0.39	101.74	16.40
		S2IGAN [13]	speech	8.72 $\pm$ 0.34	93.29
	<b>Fusion-S2iGan</b>	speech	<b>11.70<math>\pm</math>0.45</b>	<b>70.80</b>	<b>34.95</b>
Places-subset	AttnGAN [22]	text	4.59 $\pm$ 0.51	35.59	33.85
	Li <i>et al.</i>	speech	—	83.06	—
	StackGAN++ [20]	speech	3.78 $\pm$ 0.35	47.94	8.87
		S2IGAN [13]	speech	4.04 $\pm$ 0.25	42.09
	<b>Fusion-S2iGan</b>	speech	<b>5.05<math>\pm</math>0.10</b>	<b>25.68</b>	<b>28.37</b>

### A. Quantitative Results

Table II 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.

### B. 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. 6, 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 1<sup>st</sup>, 5<sup>th</sup>, 6<sup>th</sup> and 8<sup>th</sup> column. We can also observe that Fusion-S2iGan synthesizes a man surfing on the realistic sea waves (2<sup>nd</sup> column), two clear persons standing on the beach (3<sup>rd</sup> column), a plausible football player in red and a reasonable background (4<sup>th</sup> column) and a running girl (4<sup>th</sup> column), whereas S2IGAN yields unclear objects (1<sup>st</sup>, 2<sup>nd</sup>, 4<sup>th</sup>, 5<sup>th</sup>, 6<sup>th</sup> and 8<sup>th</sup> column) and blurry backgrounds

(4<sup>th</sup> and 7<sup>th</sup> column). Furthermore, it can be seen that the number of the objects produced by Fusion-S2iGan is correct in the 1<sup>st</sup>, 2<sup>nd</sup>, 3<sup>rd</sup>, 5<sup>th</sup>, 6<sup>th</sup>, 7<sup>th</sup> and 8<sup>th</sup> column.

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

The above analysis indicates that Fusion-S2iGan has the capacity to achieve very good results on the real speech data sets due to the effective speech-visual fusion module and loss functions.

## VIII. ABLATION/SUBSTITUTION TESTS

In order to evaluate the effectiveness of different components in Fusion-S2iGan, we perform a series of ablation/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. 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.

### A. 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) [52], polarized self-attention (PolarizedAttn) [53] and contextual transformer attention (CoTAttn) [54], respectively. The comparison results are depicted in Table III. 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.

### B. 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 IV reports the quantitative results on the CUB and Flickr8k data



Fig. 6. Qualitative comparison of S2IGAN [13] 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.

TABLE III

SUBSTITUTION STUDY ON PIXEL-ATTENTION MODULES. BAM, POLARIZEDATTN AND CoTATTN REFER TO THE BOTTLENECK ATTENTION MODULE [52], POLARIZED SELF-ATTENTION [53] AND CONTEXTUAL TRANSFORMER ATTENTION [54], RESPECTIVELY. PAM INDICATES THE INTRODUCED PIXEL-ATTENTION MODULE. THE BEST RESULTS ARE IN BOLD.

Datasets	CUB			Oxford-102		
	IS $\uparrow$	FID $\downarrow$	mAP $\uparrow$	IS $\uparrow$	FID $\downarrow$	mAP $\uparrow$
BAM [52]	4.35 $\pm$ 0.04	16.23	10.32	3.62 $\pm$ 0.09	42.72	16.35
PolarizedAttn [53]	4.68 $\pm$ 0.04	14.33	10.74	3.75 $\pm$ 0.07	44.20	16.42
CoTAttn [54]	4.70 $\pm$ 0.05	14.07	10.53	3.75 $\pm$ 0.06	41.24	17.28
<b>PAM</b>	<b>5.06<math>\pm</math>0.09</b>	<b>13.09</b>	<b>12.12</b>	<b>3.81<math>\pm</math>0.08</b>	<b>40.08</b>	<b>17.72</b>

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.

C. 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

TABLE IV

SUBSTITUTION TEST ON FUSION METHODS. 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.

Datasets	CUB			Flickr8k		
	IS $\uparrow$	FID $\downarrow$	mAP $\uparrow$	IS $\uparrow$	FID $\downarrow$	R@50 $\uparrow$
Addition	3.70 $\pm$ 0.04	29.11	2.7	9.53 $\pm$ 0.37	100.83	5.98
Product	4.68 $\pm$ 0.04	15.25	11.11	10.23 $\pm$ 0.36	78.03	30.82
<b>WFM</b>	<b>5.06<math>\pm</math>0.09</b>	<b>13.09</b>	<b>12.12</b>	<b>11.70<math>\pm</math>0.45</b>	<b>70.80</b>	<b>34.95</b>

TABLE V

SUBSTITUTION STUDY ON THE VISUAL+SPEECH FUSION MODULE (VSFM). DTGAN AND DF-GAN REFER TO THE SENTENCE-VISUAL FUSION MODULE IN DTGAN [8] AND DF-GAN [26], RESPECTIVELY. VSFM DENOTES THE PRESENTED VISUAL+SPEECH FUSION MODULE. THE BEST RESULTS ARE IN BOLD.

Datasets	CUB			Oxford-102		
	IS $\uparrow$	FID $\downarrow$	mAP $\uparrow$	IS $\uparrow$	FID $\downarrow$	mAP $\uparrow$
DTGAN [49]	4.18 $\pm$ 0.04	24.70	4.92	3.15 $\pm$ 0.04	52.67	13.87
DF-GAN [26]	4.40 $\pm$ 0.03	16.74	8.20	3.58 $\pm$ 0.07	48.37	16.32
<b>VSFM</b>	<b>5.06<math>\pm</math>0.09</b>	<b>13.09</b>	<b>12.12</b>	<b>3.81<math>\pm</math>0.08</b>	<b>40.08</b>	<b>17.72</b>

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 [8] and DFGAN [26] on the CUB and Oxford-102 data sets. The results of the substitution study are displayed in Table V. 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 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.

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 VI. 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.

## IX. 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

TABLE VI

EFFECT OF THE NUMBER OF FUSION MODULES IN THE DUAL-RESIDUAL SPEECH-VISUAL FUSION BLOCK. 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.

Datasets	Flickr8k			Places-subset		
	IS $\uparrow$	FID $\downarrow$	mAP $\uparrow$	IS $\uparrow$	FID $\downarrow$	mAP $\uparrow$
single-module	10.39 $\pm$ 0.34	81.62	32.37	5.02 $\pm$ 0.06	41.67	24.00
<b>dual-module</b>	<b>11.70<math>\pm</math>0.45</b>	<b>70.80</b>	<b>34.95</b>	<b>5.05<math>\pm</math>0.10</b>	<b>25.68</b>	<b>28.37</b>

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.

## REFERENCES

- [1] S. Wang, Y. Yang, Z. Wu, Y. Qian, and K. Yu, "Data augmentation using deep generative models for embedding based speaker recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2598–2609, 2020.
- [2] Y. Zhong, L. Wang, J. Chen, D. Yu, and Y. Li, "Comprehensive image captioning via scene graph decomposition," in *European Conference on Computer Vision*. Springer, 2020, pp. 211–229.
- [3] L. Polka and J. F. Werker, "Developmental changes in perception of nonnative vowel contrasts," *Journal of Experimental Psychology: Human perception and performance*, vol. 20, no. 2, p. 421, 1994.
- [4] J. F. Werker and R. C. Tees, "Cross-language speech perception: Evidence for perceptual reorganization during the first year of life," *Infant behavior and development*, vol. 7, no. 1, pp. 49–63, 1984.
- [5] J. Gervain and J. Mehler, "Speech perception and language acquisition in the first year of life," *Annual review of psychology*, vol. 61, pp. 191–218, 2010.
- [6] 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.
- [7] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.
- [8] Z. Zhang and L. Schomaker, "Dtgan: Dual attention generative adversarial networks for text-to-image generation," in *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021, pp. 1–8.

- [9] —, “Divergan: An efficient and effective single-stage framework for diverse text-to-image generation,” *Neurocomputing*, 2021.
- [10] V. K. R. Sridhar, S. Bangalore, and S. S. Narayanan, “Exploiting acoustic and syntactic features for automatic prosody labeling in a maximum entropy framework,” *IEEE transactions on audio, speech, and language processing*, vol. 16, no. 4, pp. 797–811, 2008.
- [11] C. Du and K. Yu, “Phone-level prosody modelling with gmm-based mdn for diverse and controllable speech synthesis,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 190–201, 2021.
- [12] N. Obin and P. Lanchantin, “Symbolic modeling of prosody: From linguistics to statistics,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 3, pp. 588–599, 2015.
- [13] X. Wang, T. Qiao, J. Zhu, A. Hanjalic, and O. Scharenborg, “Generating images from spoken descriptions,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 850–865, 2021.
- [14] J. Li, X. Zhang, C. Jia, J. Xu, L. Zhang, Y. Wang, S. Ma, and W. Gao, “Direct speech-to-image translation,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 3, pp. 517–529, 2020.
- [15] M. Zhu, P. Pan, W. Chen, and Y. Yang, “Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5802–5810.
- [16] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, “The caltech-ucsd birds-200-2011 dataset,” 2011.
- [17] M.-E. Nilsback and A. Zisserman, “Automated flower classification over a large number of classes,” in *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*. IEEE, 2008, pp. 722–729.
- [18] M. Hodosh, P. Young, and J. Hockenmaier, “Framing image description as a ranking task: Data, models and evaluation metrics,” *Journal of Artificial Intelligence Research*, vol. 47, pp. 853–899, 2013.
- [19] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, “Learning deep features for scene recognition using places database,” *Advances in neural information processing systems*, vol. 27, 2014.
- [20] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, “Stackgan++: Realistic image synthesis with stacked generative adversarial networks,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 1947–1962, 2018.
- [21] M. Mirza and S. Osindero, “Conditional generative adversarial nets,” *arXiv preprint arXiv:1411.1784*, 2014.
- [22] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, “AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1316–1324.
- [23] T. Qiao, J. Zhang, D. Xu, and D. Tao, “Mirrorgan: Learning text-to-image generation by redescription,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1505–1514.
- [24] J. Liang, W. Pei, and F. Lu, “CpGAN: Content-parsing generative adversarial networks for text-to-image synthesis,” in *European Conference on Computer Vision*. Springer, 2020, pp. 491–508.
- [25] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” *arXiv preprint arXiv:1511.06434*, 2015.
- [26] M. Tao, H. Tang, S. Wu, N. Sebe, F. Wu, and X.-Y. Jing, “Df-gan: Deep fusion generative adversarial networks for text-to-image synthesis,” *arXiv preprint arXiv:2008.05865*, 2020.
- [27] Z. Zhang and L. Schomaker, “Optimized latent-code selection for explainable conditional text-to-image GANs,” *arXiv preprint arXiv:2204.12678*, 2022.
- [28] 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.
- [29] 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*, 2017, pp. 349–357.
- [30] W. Hao, Z. Zhang, and H. Guan, “CmCGAN: A uniform framework for cross-modal visual-audio mutual generation,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.
- [31] A. C. 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 *ICASSP*, 2019, pp. 8633–8637.
- [32] T.-H. Oh, T. Dekel, C. Kim, I. Mosseri, W. T. Freeman, M. Rubinstein, and W. Matusik, “Speech2face: Learning the face behind a voice,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 7539–7548.
- [33] K. N. Haque, R. Rana, J. Liu, J. H. Hansen, N. Cummins, C. Busso, and B. W. Schuller, “Guided generative adversarial neural network for representation learning and audio generation using fewer labelled audio data,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2575–2590, 2021.
- [34] S. Santurkar, D. Tsipras, A. Ilyas, and A. Madry, “How does batch normalization help optimization?” in *Advances in Neural Information Processing Systems*, 2018, pp. 2483–2493.
- [35] A. L. Maas, A. Y. Hannun, A. Y. Ng *et al.*, “Rectifier nonlinearities improve neural network acoustic models,” in *Proc. icml*, vol. 30, no. 1. Citeseer, 2013, p. 3.
- [36] J. H. Lim and J. C. Ye, “Geometric gan,” *arXiv preprint arXiv:1705.02894*, 2017.
- [37] D. Hendrycks and K. Gimpel, “Gaussian error linear units (gelus),” *arXiv preprint arXiv:1606.08415*, 2016.
- [38] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, “Natural tts synthesis by conditioning wavenet on mel spectrogram predictions,” in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.
- [39] D. Harwath and J. Glass, “Deep multimodal semantic embeddings for speech and images,” in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 237–244.
- [40] D. Harwath, A. Recasens, D. Surís, G. Chuang, A. Torralba, and J. Glass, “Jointly discovering visual objects and spoken words from raw sensory input,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 649–665.
- [41] D. Harwath, A. Torralba, and J. Glass, “Unsupervised learning of spoken language with visual context,” *Advances in Neural Information Processing Systems*, vol. 29, 2016.
- [42] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.
- [43] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, “Imagenet large scale visual recognition challenge,” *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.
- [44] 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.
- [45] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [46] 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.
- [47] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, “Pytorch: An imperative style, high-performance deep learning library,” in *Advances in neural information processing systems*, 2019, pp. 8026–8037.
- [48] 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.
- [49] Z. Zhang and L. Schomaker, “DtGAN: Dual attention generative adversarial networks for text-to-image generation,” *arXiv preprint arXiv:2011.02709*, 2020.
- [50] B. Li, X. Qi, T. Lukasiewicz, and P. Torr, “Controllable text-to-image generation,” in *Advances in Neural Information Processing Systems*, 2019, pp. 2065–2075.
- [51] G. Yin, B. Liu, L. Sheng, N. Yu, X. Wang, and J. Shao, “Semantics disentangling for text-to-image generation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2327–2336.
- [52] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon, “BAM: Bottleneck attention module,” *arXiv preprint arXiv:1807.06514*, 2018.
- [53] H. Liu, F. Liu, X. Fan, and D. Huang, “Polarized self-attention: towards high-quality pixel-wise regression,” *arXiv preprint arXiv:2107.00782*, 2021.
- [54] Y. Li, T. Yao, Y. Pan, and T. Mei, “Contextual transformer networks for visual recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.