DOI:10.11918/202403057

基于单阶段框架的语音到图像生成方法

张振兴,杨任农,张 滢,李永林,赵 雨,左家亮

(空军工程大学 空管领航学院,西安 710051)

摘 要:为解决多阶段语音到图像生成过程中生成器缺乏对图像进行集成和细化的过程,框架训练多个对抗器网络而增加复杂性、训练成本等问题,提出一种单阶段语音到图像生成对抗网络(single-stage speech-to-image generative adversarial network, Single-S2IGAN)框架。该框架在条件生成对抗网络基础上,引入视觉-语音融合模块(visual speech fusion module, VSFM),VSFM由批归一化(batch normalization, B_N)、像素注意力模块(pixel-attention module, M_{PA})、语音调制模块(speech modulation module, M_{SM})和加权融合模块(weighted fusion module, M_{WF})组成。其中,B_N用于稳定网络训练;M_{PA}对空间像素区域间的语义关联进行建模,并为关键位置分配更高的权重;M_{SM}利用语音向量中的细粒度语言线索调制视觉特征映射;M_{WF}以自适应的方式捕捉M_{PA}和M_{SM}在通道级别的语义重要性。在CUB bird、Oxford-102、Flickr8k和Places-subset 4个基准数据集上进行一系列实验,结果表明:Single-S2IGAN能够生成高分辨率且与输入语音描述语义一致的图像;与现有其他单阶段框架和多阶段框架方法相比,Single-S2IGAN表现出更好的性能,且其性能水平接近文本到图像生成的方法。本文所提Single-S2IGAN框架为语音到图像生成任务研究提供了新的思路。

关键词:语音到图像生成;单阶段框架;生成对抗网络;注意力机制;融合模型

中图分类号:V247 **文献标志码**:A **文章编号**:0367-6234(2025)09-0115-10

Speech-to-image generation method based on a single-stage framework

Zhang Zhenxing, Yang Rennong, Zhang Ying, Li Yonglin, Zhao Yu, Zuo Jialiang

(School of Air Traffic Control and Navigation, Air Force Engineering University, Xi'an 710051, China)

Abstract: To address the challenges in multi-stage speech-to-image generation process, such as the generator's lack of integration and refinement of images, as well as the increased complexity and training costs associated with training multiple adversarial networks, we propose a single-stage speech-to-image generative adversarial network framework (Single-S2IGAN). This framework builds upon conditional generative adversarial networks by introducing a visual speech fusion module (VSFM), which consists of batch normalization (B_N), a pixel-attention module (M_{PA}), a speech modulation module (M_{SM}), and a weighted fusion module (M_{WF}). B_N is used to stabilize the training of the network; M_{PA} models the semantic relationships between spatial pixel regions and assigns higher weights to key positions. M_{SM} modulates visual feature maps using fine-grained linguistic cues from the speech vector; and M_{WF} adaptively captures the semantic importance of M_{PA} and M_{SM} at the channel level. A series of experiments were conducted on four benchmark datasets: CUB bird, Oxford-102, Flickr8k, and Places-subset. The results indicate that Single-S2IGAN can generate high-resolution images that are semantically consistent with the input speech descriptions. Compared to both single-stage and multi-stage frameworks, Single-S2IGAN demonstrates superior performance, with levels of effectiveness approaching those of text-to-image generation methods. The proposed Single-S2IGAN framework offers a new perspective for research into speech-to-image generation tasks.

Keywords: speech-to-image generation; single-stage framework; generative adversarial network; attention mechanism; fusion model

语音描述能够借助语言和情感信息准确地描述图像的内容,语音到图像生成任务旨在根据给定的语音信号自动生成既逼真又语义相符的图像。近年来,因其巨大的应用前景,在跨学科领域中引起了广泛关注。针对语音到图像生成任务,Li等[1]提出了一种多阶段语音到图像生成对抗网络(generative

收稿日期:2024-03-25;**录用日期**:2024-04-29;**网络首发日期**:2024-11-29
网络首发地址:https://link.cnki.net/urlid/23.1235.t.20241128.1820.002
基金项目:国家自然科学基金(71501184,62106284);陕西省自然科学基金(2021JQ370)
作者简介:张振兴(1993—),男,博士,讲师;杨任农(1969—),男,教授,博士生导师
通信作者:张 滢,2207621676@qq.com

adversarial network, GAN) 方法, 以生成与输入语音描述语义相关的逼真图像。为更好地捕捉语言信息, 引入一个预训练的图像编码器作为“教师”, 用以指导语音编码器的训练。Wang 等[2] 设计了一个新的语音嵌入网络, 用于获取语音特征向量。此外, 他们还开发了一个基于关系监督的多阶段生成模型, 以产生高质量的图像。然而, 语音到图像生成是一项极具挑战性的跨模态任务。与文本描述相比, 语音信号具有连续性而非离散性, 缺乏文本中单词间规则的间隔, 即空格。这一特性使得模型在尝试学习对应的语义表示时, 难以捕获描述图像内容的语音信息。

当前的语音到图像 GAN 模型采用多阶段框架 (见图 1(a)), 通过多个生成器 ($G_0 \sim G_2$) 和对抗器 ($D_0 \sim D_2$) 生成视觉合理的样本。尽管这种框架在语音到图像生成任务中取得了令人满意的结果, 但仍存在 3 个关键问题: 1) 与单阶段框架相比, 训练多个网络会增加计算时间, 并可能影响生成模型的收敛性和稳定性[3]。更糟糕的是, 如果前面的生成器未能收敛到全局最优解, 那么后续的生成器网络可能难以得到进一步提升[4]。2) 该框架忽视了早期生成器生成图像的质量。初期的生成器网络主要由上采样层和卷积层构成, 没有利用自然语言描述对图像进行集成和细化[4-5]。3) 该框架需要训练多个对抗器网络, 增加了模型的复杂性和训练成本。

基于此, 本文提出了单阶段语音到图像生成对抗网络 (single-stage speech-to-image generative adversarial network, Single-S2IGAN) 框架, 如图 1(b) 所示。该框架仅有单组生成器 (G) 和对抗器 (D), 利用预训练语音编码器提取语音嵌入作为输入特征。在生成器中融入了视觉 - 语音融合模块 (visual speech fusion module, VSFM) 和双残差层 ($B_0 \sim B_6$), 以有效地将语音特征融入生成器网络, 确保生成图像与给定语音信号语义相符, 进而提升生成样本的质量。最后, 在多种数据集上进行仿真测试, 以验证该框架的性能。

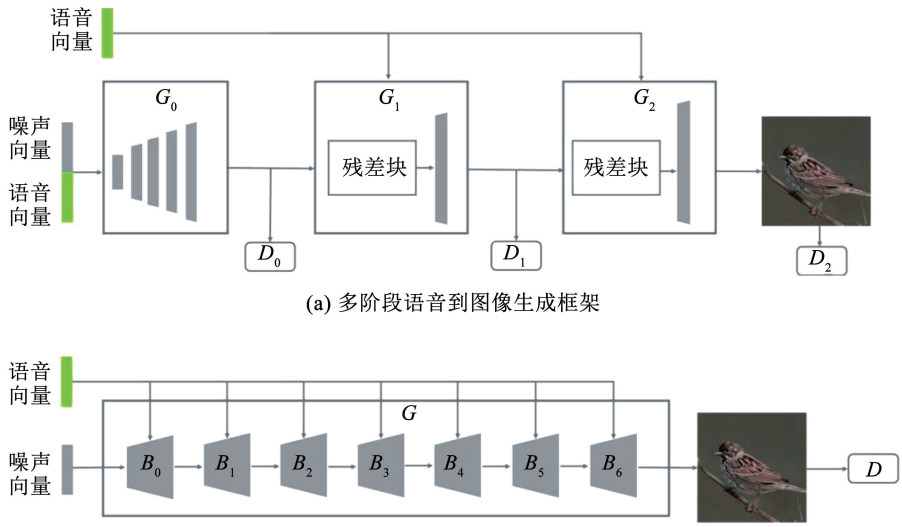


图 1 多阶段框架与所提出的单阶段框架比较

Fig. 1 Comparison between multi-stage framework and proposed single-stage framework

1 条件生成对抗网络

作为 GAN 的一种扩展, 条件生成对抗网络 (conditional generative adversarial network, cGAN) 以条件上下文 $\boldsymbol{c}$ (如类别标签、文本描述和低分辨率图像等) 和随机向量 $\boldsymbol{z}$ 作为生成器 G 的输入, 并输出与条件 $\boldsymbol{c}$ 语义相关的图像。同时, 对抗器 D 负责区分真实图像 - 上下文对 $(\boldsymbol{x},\boldsymbol{c})$ 和生成图像 - 上下文对 $(G(\boldsymbol{z},\boldsymbol{c}),\boldsymbol{c})$。cGAN 的目标函数 $V(G,D)$ 为

$$\min_G \max_D V(G,D) = E_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x},\boldsymbol{c})] + E_{\boldsymbol{z}\sim p_{\text{z}}(\boldsymbol{z}),\boldsymbol{c}\sim p_{\text{c}}(\boldsymbol{c})}[\log(1-D(G(\boldsymbol{z},\boldsymbol{c}),\boldsymbol{c}))] \tag{1}$$

式中: $p_{\text{data}}(\boldsymbol{x})$ 和 $p_{\text{z}}(\boldsymbol{z})$ 分别为真实数据 $\boldsymbol{x}$ 的分布和随机潜在向量 $\boldsymbol{z}$ 的分布, $p_{\text{c}}(\boldsymbol{c})$ 为 $\boldsymbol{c}$ 的分布, $E_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\cdot]$ 为真实数据分布期望值, $E_{\boldsymbol{z}\sim p_{\text{z}}(\boldsymbol{z}),\boldsymbol{c}\sim p_{\text{c}}(\boldsymbol{c})}[\cdot]$ 为合成数据分布期望值。

在语音到图像生成任务中, $\boldsymbol{c}$ 是给定的语音描述。假设 $\{(\boldsymbol{I}_i,\boldsymbol{C}_i)\}_{i=1}^n$ 表示用于训练的由 n 个图像 - 语音对组成的集合, 其中 $\boldsymbol{I}_i$ 代表一张图像, $\boldsymbol{C}_i=(\boldsymbol{c}_i^1,\boldsymbol{c}_i^2,\cdots,\boldsymbol{c}_i^k)$ 包含 k 个语音描述。生成器 G 的目标是根据从 $\boldsymbol{C}_i$ 中随机选择的语音描述 $\boldsymbol{c}_i$, 生成一个视觉逼真且语义一致的图像 $\hat{\boldsymbol{I}}_i$。同时, 训练对抗器 D 以区分真实的图像 - 语音对 $(\boldsymbol{I}_i,\boldsymbol{c}_i)$ 和生成的图像 - 语音对 $(\hat{\boldsymbol{I}}_i,\boldsymbol{c}_i)$。

2　Single-S2IGAN

Single-S2IGAN 的整体框架如图 2 所示，该框架仅包括一个生成器和一个对抗器。本节分为 4 部分，分别详述 Single-S2IGAN 框架中的生成器（图 2(a)）、对抗器（图 2(b)）、视觉融合模块（图 2(c)和图 2(d)）以及损失函数。

2.1　生成器

生成器能够将语音信号映射为逼真且语义一致的图像，框架如图 2(a)所示。其中，全连接层将潜在向量转换为初始特征图，双残差视觉－语音融合模块利用语音向量调制视觉特征图。语音编码器用于学习给定语音描述的语义表示和概念意义，以捕获重要的视觉细节描述。双残差视觉－语音融合模块（图 2(c)）包含两个有效的 VSFM（图 2(d)）以及一系列 ReLU 激活函数和卷积层。

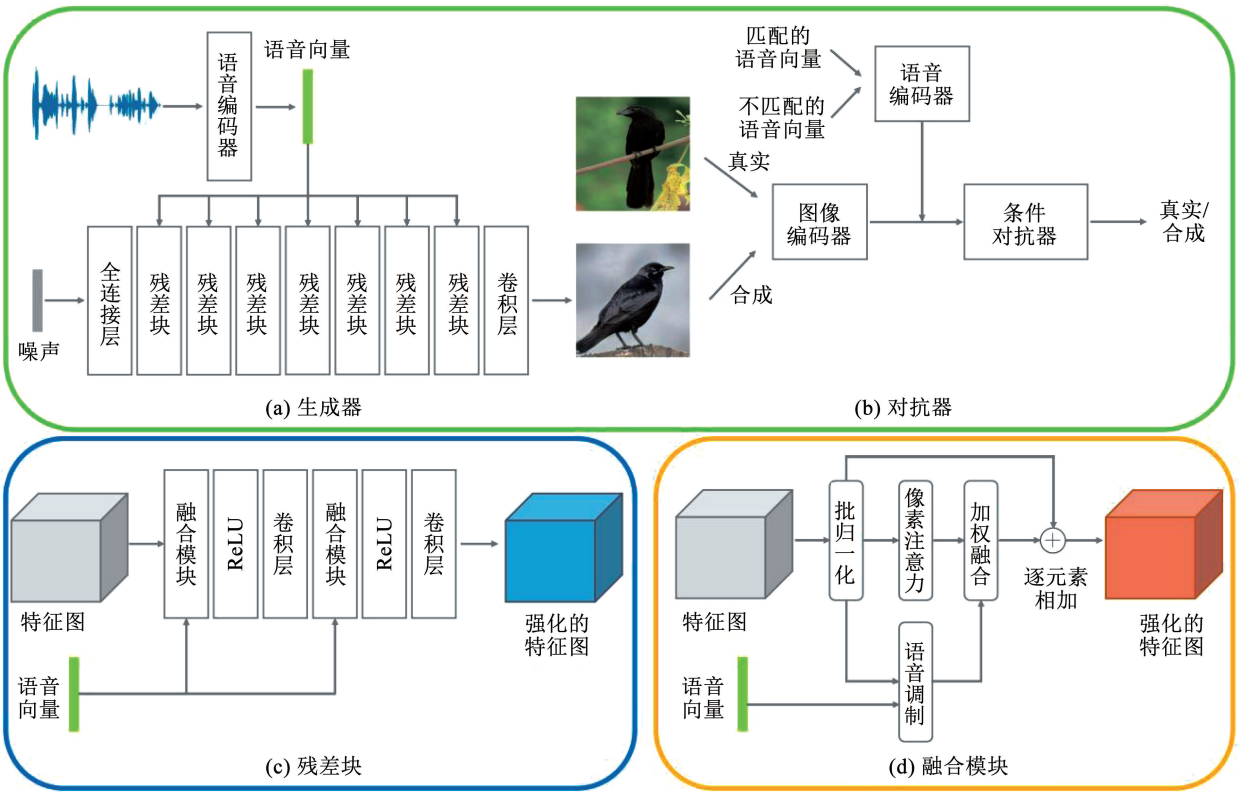


图 2　Single-S2IGAN 的整体框架

Fig. 2　Overall framework of Single-S2IGAN

VSFM 包括批归一化（batch normalization，B_N），像素注意力模块（pixel-attention module，M_{PA}），语音调制模块（speech modulation module，M_{SM}）和加权融合模块（weighted fusion module，M_{WF}）。双残差视觉－语音融合模块允许通过有效增加网络层数来增强模型容量，同时通过保留比串联结构更多的原始特征稳定训练过程。借助双残差视觉－语音融合模块，生成器网络能够产生高质量的图像。语音信号生成图像的过程为：

$$\boldsymbol{h}_0 = F_0(\boldsymbol{z}) \tag{2}$$

$$\boldsymbol{h}_1 = F_1^{\text{Dual}}(\boldsymbol{h}_0, \boldsymbol{s}) \tag{3}$$

$$\boldsymbol{h}_i = F_i^{\text{Dual}}(\boldsymbol{h}_{i-1}\uparrow, \boldsymbol{s}),\ i = 2,3,\cdots,7 \tag{4}$$

$$\boldsymbol{I}_o = L_f(\boldsymbol{h}_7) \tag{5}$$

式中：F_0 为全连接层，$\boldsymbol{s}$ 为由语音编码器获得的语音向量，F_i^{Dual} 为提出的双残差视觉＋语音融合模块，$\boldsymbol{h}_0 \sim \boldsymbol{h}_7$ 为特征映射图，L_f 为最后一个卷积层，$\boldsymbol{I}_o$ 为最终生成的图像。

2.2　对抗器

对抗器网络包括图像编码器、预训练语音编码器和一个条件对抗器网络，见图 2(b)。对抗器中超参数的选择基于文献[4]，但文本特征为语音向量。具体来说，图像编码器由一个卷积层（步长为 1，内核大小为 3，填充为 1）及 6 个自适应连续残差块组成。每个块包含两个卷积层，其中第 1 层步长为 2，内核大小为 4，填充为 1，用于将输入特征图的维度减少到一半；第 2 层步长为 1，内核大小为 3，填充为 1，目的是进一步提炼图像特征。在每个卷积层之后，采用斜率为 0.2 的 ReLU 激活函数帮助训练网络。残差块的通道数量分别为 64、128、256、512、1 024、1 024。为了稳定学习，在每个块中引入一个

残差连接。在训练过程中,将图像输入到图像编码器以提取图像特征(1 024,4,4),然后将这些特征与语音向量结合作为联合嵌入,维度为(2 048,4,4)。最后,将联合特征通过条件对抗器,获得最终的条件分数,用于确定输入的图像 – 语音描述真伪。

2.3 视觉 – 语音融合模块(VSFM)

在图像生成过程中,图像的视觉内容与相应的条件上下文(如类别标签和文本描述等)之间的语义关系具有重要的作用。然而,对于语音到图像生成来说,语音信号时间长且连续,缺少词汇边界,即自然语言描述中的空格。因此,建模语音描述的语言内容与图像内容之间的关联性非常困难。此外,由于语音信号的连续性,单词级调制模块无法将语音信息和视觉特征图融合在一起。在这种情况下,研究如何有效地将语音描述中的信息注入到以二维彩色图像为输出的神经网络中十分重要。

因此,设计了 VSFM,将输入的语音信号转换为类似于句子嵌入的全局语音向量,使用语音特征来调制网络中的视觉特征图,进而生成高分辨率图像。

2.3.1 整体框架

VSFM 的框架如图 2(d)所示。以给定中间特征图 $\boldsymbol{F}\in\mathbf{R}^{C\times H\times W}$ 和语音编码器获得的语音向量 $\boldsymbol{s}$ 作为输入,VSFM 的过程包括 3 个步骤:1)在 $\boldsymbol{F}$ 上应用批归一化,得到新的特征图 $\boldsymbol{F}'\in\mathbf{R}^{C\times H\times W}$。这有助于稳定条件生成对抗网络的学习过程,并加快训练速度。2)分别使用 M_{PA} 和 M_{SM} 强化 $\boldsymbol{F}'\in\mathbf{R}^{C\times H\times W}$。3)VSFM通过 M_{WF},有效地融合注意力模块的输出 $M_{\mathrm{PA}}(\boldsymbol{F}')\in\mathbf{R}^{C\times H\times W}$ 和语音调制模块的输出 $M_{\mathrm{SM}}(\boldsymbol{F}',\boldsymbol{s})\in\mathbf{R}^{C\times H\times W}$。同时,采用残差连接获得最终的结果 $\boldsymbol{F}''\in\mathbf{R}^{C\times H\times W}$。

值得注意的是,Single-S2IGAN 将 VSFM 的双模态信息应用于生成器网络的所有层中,允许语音影响框架各个分层级别上的特征(从早期的粗糙特征到后期的抽象特征),以提升生成图像的质量。VSFM 的整体过程数学表达式为

$$\boldsymbol{F}'=B_{\mathrm{N}}(\boldsymbol{F}) \tag{6}$$

$$\boldsymbol{F}''=\boldsymbol{F}'+M_{\mathrm{WF}}(M_{\mathrm{PA}}(\boldsymbol{F}'),M_{\mathrm{SM}}(\boldsymbol{F}',\boldsymbol{s})) \tag{7}$$

2.3.2 像素注意力模块(M_{PA})

图像是由视觉像素组成的,像素对图像质量有着重要的影响。学习图像中不同位置之间的上下文关系对于生成高质量的样本至关重要。然而,卷积操作只能捕捉空间上下文之间的局部关系,无法“看到”整个图像场景。因此,引入 M_{PA},用于建模视觉像素之间的空间关联,并获取空间注意力图以微调特征图,引导生成器关注图像的关键区域。

M_{PA} 处理流程见图 3。首先,将特征图 $\boldsymbol{F}'\in\mathbf{R}^{C\times H\times W}$ 输入到一个 3 × 3 的卷积层中,并经过批归一化和 ReLU 函数处理,将通道维度减少至 $\mathbf{R}^{C/r\times H\times W}$。这可以整合并加强通道和空间方向上的视觉特征图。随后,使用一个 1 × 1 的卷积层和 Sigmoid 函数处理特征图,得到像素注意力图 $\boldsymbol{P}_{\mathrm{K}}\in\mathbf{R}^{1\times H\times W}$。数学表达式为

$$\boldsymbol{P}_{\mathrm{K}}=\sigma(f_1^{1\times1}(R(B_{\mathrm{N}}(f_0^{3\times3}(\boldsymbol{F}'))))) \tag{8}$$

式中:$\sigma(\cdot)$ 为 Sigmoid 函数,$f_1^{1\times1}$、$f_0^{3\times3}$ 分别为 1 × 1、3 × 3 卷积层,$R(\cdot)$ 为 ReLU 函数。

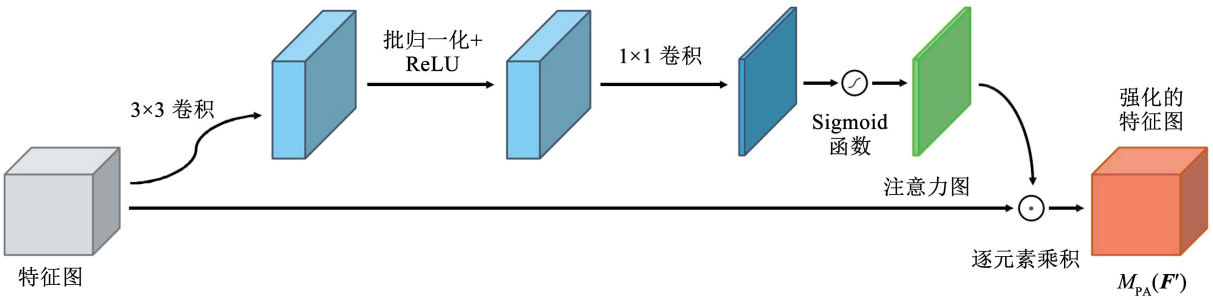


图 3 像素注意力模块示意

Fig. 3 Schematic of pixel-attention module

对特征图 $\boldsymbol{F}'\in\mathbf{R}^{C\times H\times W}$ 和像素注意力图 $\boldsymbol{P}_{\mathrm{K}}\in\mathbf{R}^{1\times H\times W}$ 进行矩阵相乘,以获得最终的结果。数学表达式为

$$M_{\mathrm{PA}}(\boldsymbol{F}')=\boldsymbol{F}'\odot\boldsymbol{P}_{\mathrm{K}} \tag{9}$$

2.3.3 语音调制模块(M_{SM})

针对语音到图像生成任务,需要开发一种新的并且有效的调制方法,以确保生成图像的语义一致性和质量[3]。为此,引入 M_{SM}。M_{SM} 能控制视觉特征图的形状和纹理变化,缩小视觉与语音间的语义鸿沟,利用从语音嵌入中提取的详细语言线索来调整和变换视觉特征图。具体来说,采用两个全连接层将 $\boldsymbol{s}$ 转换为语言线索 $\boldsymbol{W}_{\mathrm{K}}\in\mathbf{R}^{C\times1\times1}$ 和 $\boldsymbol{B}_{\mathrm{K}}\in\mathbf{R}^{C\times1\times1}$,再用 $\boldsymbol{W}_{\mathrm{K}}$ 和 $\boldsymbol{B}_{\mathrm{K}}$ 缩放和平移特征图 $\boldsymbol{F}'$。M_{SM} 的过程定义为

$$\boldsymbol{W}_{\mathrm{K}}=F_0(R(F_0(\boldsymbol{s}))) \tag{10}$$

$$\boldsymbol{B}_{\mathrm{K}}=F_0(R(F_0(\boldsymbol{s}))) \tag{11}$$

$$M_{\mathrm{SM}}(\boldsymbol{F}') = \boldsymbol{F}' \odot \boldsymbol{W}_{\mathrm{K}} + \boldsymbol{B}_{\mathrm{K}} \tag{12}$$

式中：$\boldsymbol{F}_0$ 为全连接层，$M_{\mathrm{SM}}(\boldsymbol{F}')$ 为 M_{SM} 的输出特征图。

2.3.4　加权融合模块（M_{WF}）

由于像素和语音信号属于两种不同的模态，简单地使用加法或逐元素乘法融合 M_{PA} 和 M_{SM} 的输出很难起作用，所以需要开发一种更好的方式进行组合。因此提出 M_{WF}，从通道角度学习 M_{PA} 和 M_{SM} 的语义重要性，以自适应的方式强调重要的图像区域，从而进一步提升生成图像的分辨率。

M_{WF} 的详细结构如图 4 所示。首先，使用逐元素加法处理 M_{PA} 和 M_{SM} 的输出 $M_{\mathrm{PA}}(\boldsymbol{F}')$ 和 $\mathrm{M}_{\mathrm{SM}}(\boldsymbol{F}', \boldsymbol{s})$，得到中间特征图 $\boldsymbol{I}_{\mathrm{F}} \in \mathbf{R}^{C\times H\times W}$。采用全局平均池化（global average pooling，P_{GA}）处理 $\boldsymbol{I}_{\mathrm{F}}$，以获得整体和重要的信息，从而获得通道特征向量 $\boldsymbol{I}'_{\mathrm{F}} \in \mathbf{R}^{C\times 1\times 1}$。随后，将 $\boldsymbol{I}'_{\mathrm{F}}$ 输入到两个全连接层中，第 1 个用于压缩和整合通道特征，第 2 个用于恢复通道维度，并在通道级别捕获图像注意力和语音调制模块的语义重要性。最后，在通道维度应用 Softmax 函数，得到上下文感知的通道权重矩阵 $\boldsymbol{W}_{\mathrm{F}} \in \mathbf{R}^{C\times 2}$。将 $\boldsymbol{W}_{\mathrm{F}}$ 分成两个独立的权重矩阵 $\boldsymbol{W}_{\mathrm{F1}} \in \mathbf{R}^{C\times 1}$ 和 $\boldsymbol{W}_{\mathrm{F2}} \in \mathbf{R}^{C\times 1}$。数学表达式为

$$\boldsymbol{I}_{\mathrm{F}} = M_{\mathrm{PA}}(\boldsymbol{F}') + M_{\mathrm{SM}}(\boldsymbol{F}', \boldsymbol{s}) \tag{13}$$

$$\boldsymbol{I}'_{\mathrm{F}} = P_{\mathrm{GA}}(\boldsymbol{I}_{\mathrm{F}}) \tag{14}$$

$$\boldsymbol{W}_{\mathrm{F}} = S_{\mathrm{om}}(F_0(G_{\mathrm{L}}(F_0(\boldsymbol{I}'_{\mathrm{F}})))) \tag{15}$$

$$\boldsymbol{W}_{\mathrm{F}} = [\boldsymbol{W}_{\mathrm{F1}}; \boldsymbol{W}_{\mathrm{F2}}] \tag{16}$$

式中：$S_{\mathrm{om}}(\cdot)$ 为 Softmax 函数，G_{L} 为高斯误差线性单元。

将 $\boldsymbol{W}_{\mathrm{F1}}$ 和 $\boldsymbol{W}_{\mathrm{F2}}$ 维度调整为 $\mathbf{R}^{C\times 1\times 1}$，并应用于 $M_{\mathrm{PA}}(\boldsymbol{F}')$ 和 $M_{\mathrm{SM}}(\boldsymbol{F}', \boldsymbol{s})$ 以获得最终的融合结果。数学表达式为

$$M_{\mathrm{WF}}(\boldsymbol{F}', \boldsymbol{s}) = \boldsymbol{W}_{\mathrm{F1}} \odot M_{\mathrm{PA}}(\boldsymbol{F}') + \boldsymbol{W}_{\mathrm{F2}} \odot M_{\mathrm{SM}}(\boldsymbol{F}', \boldsymbol{s}) \tag{17}$$

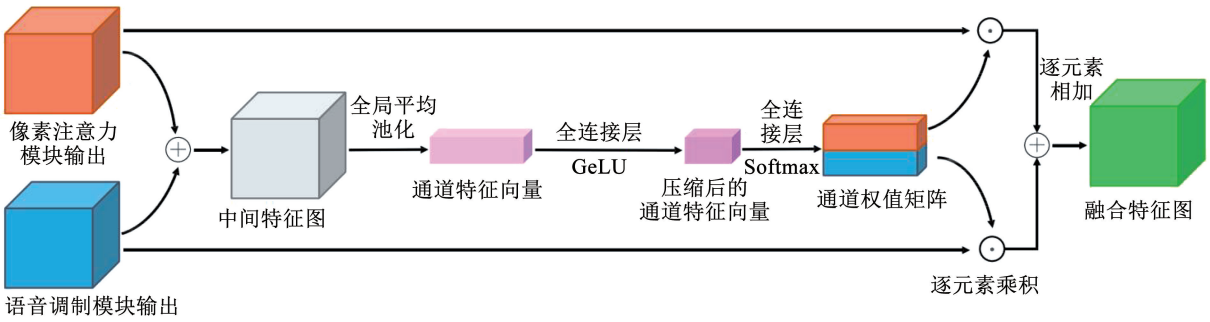


图 4　加权融合模块结构

Fig. 4　Structure of weighted fusion module

2.4　损失函数

对抗损失函数用于确保生成的样本与输入的语音信号语义一致。受文献[6]启发，Single-S2IGAN 利用铰链损失函数稳定网络的训练过程。对抗器的对抗损失函数定义为

$$\begin{aligned} L_{\mathrm{adv}} = {} & E_{\boldsymbol{x}\sim p_{\mathrm{data}}}[\max(0, 1 - D(\boldsymbol{x}, \boldsymbol{s}))] + \\ & \frac{1}{2}E_{\hat{\boldsymbol{x}}\sim p_{\mathrm{G}}}[\max(0, 1 + D(\hat{\boldsymbol{x}}, \boldsymbol{s}))] + \\ & \frac{1}{2}E_{\boldsymbol{x}\sim p_{\mathrm{data}}}[\max(0, 1 + D(\boldsymbol{x}, \hat{\boldsymbol{s}}))] \end{aligned} \tag{18}$$

式中：$\hat{\boldsymbol{s}}$ 为不匹配的语音描述，$\hat{\boldsymbol{x}}$ 为从分布 p_{G} 采样的生成样本。

采用匹配感知零中心梯度惩罚损失函数[6]训练对抗器，匹配感知零中心梯度惩罚损失函数对真实图像和输入的语音描述应用梯度惩罚，以提高生成图像的质量和语义一致性。匹配感知零中心梯度惩罚损失函数的数学表达式为

$$L_{\mathrm{M}} = E_{\boldsymbol{x}\sim p_{\mathrm{data}}}[(\| \nabla_x D(\boldsymbol{x}, \boldsymbol{s}) \|_2 + \| \nabla_s D(\boldsymbol{x}, \boldsymbol{s}) \|_2)^p] \tag{19}$$

式中：$\nabla_x D(\cdot)$、$\nabla_s D(\cdot)$ 分别为对抗器 D 对 $\boldsymbol{x}$、$\boldsymbol{s}$ 的梯度值，$\|\cdot\|_2$ 为欧几里得范数，p 为指数系数。

对于生成器，应用对抗损失和深度注意力多模态相似性模型损失训练网络。

3　仿真实验与结果分析

为验证 Single-S2IGAN 的性能，全面评估其有效性，设计了仿真实验，并对结果进行了定量和定性分析。首先，介绍实验设置，包括实验所需的数据集、实验参数细节和评价指标。然后，将本文方法与其他几种基于 cGAN 的文本到图像生成和语音到图像生成方法在 4 个数据集上进行指标对比。

3.1　实验设置

3.1.1　数据集

选用 4 个广泛应用于测试语音到图像生成方法的数据集[1-2]，即两个生成语音 - 图像数据集 CUB bird[7]、Oxford-102[8] 以及两个真实语音 - 图像数据集 Flickr8k[9]、Places-subset[10]。

CUB bird 数据集包括 11 788 张图像，分为

8 855 张训练图像和 2 933 张测试图像。每张图像都具有 10 个自然语言描述。为了评估语音到图像生成任务,S2IGAN[2] 利用文本转语音方法将文本转换为语音描述,以构建生成语音 – 图像数据集。

Oxford-102 数据集由 8 189 张图像组成,其中 5 878 张属于训练集,另外 2 311 张属于测试集。每张图像包含 10 个文本描述,文本转语音方法将这些文本转换为语音信号。

Flickr8k 数据集是一个更具挑战性的数据集,包括 8 000 张场景图像,每张图像配有 5 个真实语音描述。

Places-subset 数据集具有 13 803 个真实语音 – 图像对,涵盖了 7 个类别,分为 10 933 个训练数据对和 2 870 个测试数据对。

3.1.2 实验参数细节

实验平台选用 NVIDIA Tesla V100 GPU 显卡,内存为 32 GB,操作系统为 Ubuntu;所有方法均基于 PyTorch 深度学习框架实现;语音编码器使用两个 1-D卷积块、两个双向门控循环单元和 1 个自注意力模块以获得语音嵌入,语音向量的维度为1 024;选用 Adam 优化器训练网络,控制一阶矩和二阶矩估计的指数衰减率 β_1 和 β_2 分别设为 0 和 0.9;遵循两时间尺度更新规则[11-12],将生成器和对抗器的学习率分别设为 0.000 1 和 0.000 4;批次大小设为 32。

3.1.3 评价指标

采用以下 4 种评价指标进行性能评估:感知分数(IS)[13]、Fréchet 感知距离(FID)[14]、平均精度均值(mAP)和召回率@50[2](R@50)。前两个指标用于评估多样性和图像质量,后两个用于衡量生成样本与其语音描述之间的语义相关性。其中,IS、mAP 和 R@50 指标越高,表示方法的性能越好,而 FID 指标则相反。

3.2 生成语音 – 图像数据性能分析

3.2.1 定量分析

为了更加客观地展示 Single-S2IGAN 的有效性,在 CUB bird 和 Oxford-102 数据集上,将 Single-S2IGAN 与单阶段[3-4,6] 和多阶段[1-2,5,12,15-18] 基于 cGAN 的文本到图像生成和语音到图像生成方法进行比较,结果详见表 1。由表 1 可以看出:与最优的语音到图像生成框架相比,在 CUB bird 数据集上,Single-S2IGAN 将 IS 均值提高了 0.77,FID 降低了 1.41,mAP 提高了 3.08;在 Oxford-102 数据集上,Single-S2IGAN 将 IS 均值提高了 0.12,FID 降低了 8.56,mAP 提高了 4.32。在 CUB bird 数据集上,与 DM-GAN[5] 实现文本到图像生成方法相比,Single-S2IGAN 将 IS 均值提高了 0.31;在 Oxford-102 数据集上,与 DTGAN[4] 文本到图像生成方法相比,Single-S2IGAN 将 IS 均值提高了 0.04。说明 Single-S2IGAN 能够产生感知上合理且质量更高、多样性更好的图像,优于最好的语音到图像生成和许多文本到图像生成方法。同时,Single-S2IGAN 达到了与文本到图像生成方法相近的性能。

表 1 文本与语音到图像生成方法的 IS、FID、mAP 对比

Tab. 1 Comparison of IS, FID, and mAP for text and speech-to-image generation methods

数据集	方法名称	输入	IS	FID	mAP
CUB bird	StackGAN++[15]	文本	4.04 ±0.05	26.07	7.01
	AttnGAN[12]	文本	4.36 ±0.03	23.98	
	MirrorGAN[16]	文本	4.56 ±0.05		
	ControlGAN[17]	文本	4.58 ±0.09		
	SDGAN[18]	文本	4.67 ±0.09		
	DM-GAN[5]	文本	4.75 ±0.07	16.09	
	DF-GAN[6]	文本	4.86 ±0.04	19.24	
	DTGAN[4]	文本	4.88 ±0.03	16.35	
	DiverGAN[3]	文本	4.98 ±0.06	15.63	
	Classifier-based[2]	语音	3.68 ±0.04	43.76	
	Li 等提出的方法[1]	语音	4.09 ±0.04	18.37	
	StackGAN++[15]	语音	4.14 ±0.04	18.94	8.09
	S2IGAN[2]	语音	4.29 ±0.04	14.50	9.04
	Single-S2IGAN	语音	**5.06 ±0.09**	**13.09**	**12.12**
Oxford-102	StackGAN++[15]	文本	3.26 ±0.01	48.68	
	DF-GAN[6]	文本	3.71 ±0.06		
	DTGAN[4]	文本	3.77 ±0.06		
	DiverGAN[3]	文本	3.99 ±0.05		
	Classifier-based[2]	语音	3.30 ±0.06	64.75	
	Li 等提出的方法[1]	语音	3.23 ±0.05	54.76	
	StackGAN++[15]	语音	3.69 ±0.08	54.33	12.18
	S2IGAN[2]	语音	3.55 ±0.04	48.64	13.40
	Single-S2IGAN	语音	**3.81 ±0.08**	**40.08**	**17.72**

3.2.2　定性分析

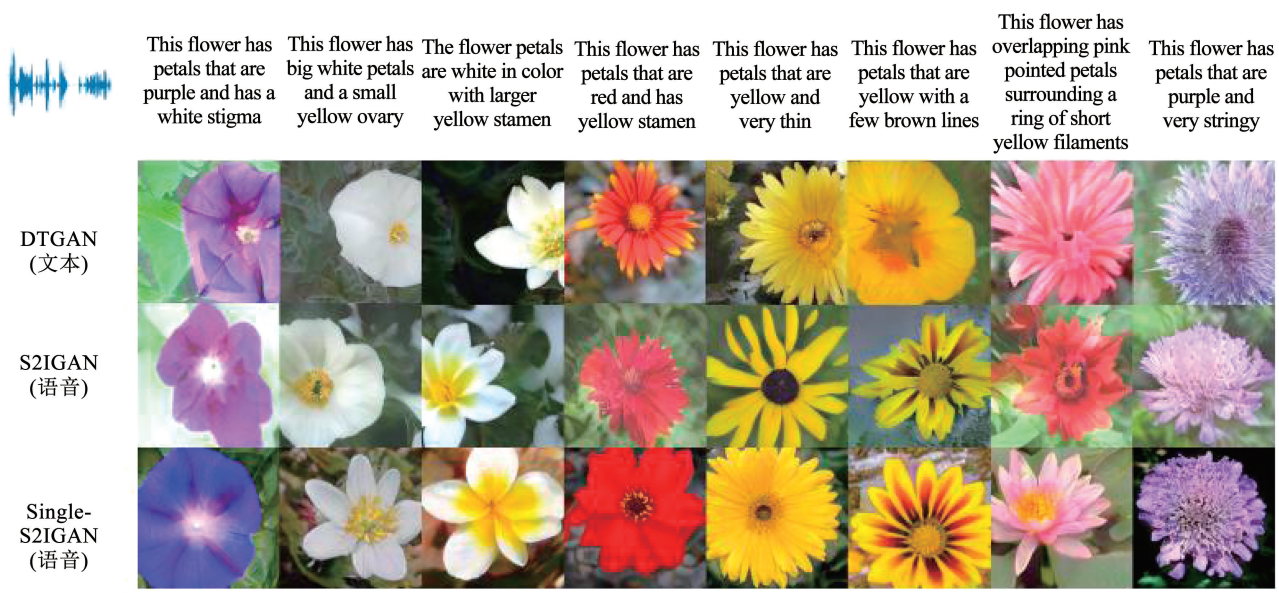
为了更加直观地展示 Single-S2IGAN 的有效性，在 CUB bird 和 Oxford-102 数据集上进行定性比较，结果如图 5 所示。图 5(a)、图 5(b)中第 1 行均为语音对应的文本描述，第 2 行分别为 AttnGAN[12] 和 DTGAN[4] 基于文本描述生成的图像，第 3 ~ 4 行分别为 S2IGAN[2]、Single-S2IGAN 基于语音描述生成的图像。

由图 5(a)可以看出：AttnGAN 和 S2IGAN 生成的鸟形状奇怪(第 2、3、6 和 8 列)，图像细节丢失(第 2、3、5 ~ 7 列)，背景模糊(第 1、3、6 和 8 列)。与 AttnGAN 和 S2IGAN 相比，Single-S2IGAN 生成的鸟具有更清晰的形状和更丰富的色彩分布(第 2、3、6 ~ 8 列)，在视觉上(第 1、2、4、6 和 8 列)更真实，并具有更生动的细节。因此，相比于 AttnGAN 和 S2IGAN，Single-S2IGAN 能够生成更清晰、更逼真的图像。

由图 5(b)可以看出：在第 1、2、6 ~ 8 列中，Single-S2IGAN 生成了比 DTGAN 和 S2IGAN 在视觉上更逼真、形状更生动的花朵；在第 3 ~ 6 列中，Single-S2IGAN 生成的花朵具有更清晰的细节和更丰富的色彩分布。说明 Single-S2IGAN 能够仅用单组生成器/对抗器生成与给定语音描述相符的高分辨率图像。



(a) CUB bird数据集



(b) Oxford-102数据集

图 5　Single-S2IGAN 与其他方法在 CUB bird 和 Oxford-102 数据集上的可视化结果

Fig. 5　Visual results of Single-S2IGAN compared to other methods on CUB bird and Oxford-102 datasets

上述实验结果表明,借助双重残差视觉 + 语音融合模块,Single-S2IGAN 能够有效地将语音信息注入生成器网络并生成高质量图像。

为了验证 Single-S2IGAN 的效率,使用该方法生成了 1 000 张图像,平均用时 128 s,说明 Single-S2IGAN 能够实时基于语音产生图像。为了进一步降低方法参数量,后期可以使用迁移学习的方法,使用预训练的 Single-S2IGAN 框架监督训练另一个由简单神经网络构成的生成器。

3.2.3 人类评估

为了评估 StackGAN + +、S2IGAN 和 Single-S2IGAN 的图像质量和语义一致性,在 CUB bird 和 Oxford-102 数据集上进行人类测试。具体而言,从每个数据集中随机选取 100 张图像,然后向测试人员提供相同的语音描述,并要求其根据图像内容和相应的语音描述选择最佳的生成样本。为了确保公平性,最终得分以两位评委的计算为准。测试结果如图 6 所示。由图 6 可以看出,Single-S2IGAN 在 CUB bird 数据集上最佳生成样优于 S2IGAN 11% 左右,在 Oxford-102 数据集上最佳生成样优于 S2IGAN 6% 左右,表明 Single-S2IGAN 生成图像的质量更高,且与语义的一致性更好。

3.3 真实语音 – 图像数据性能分析

除了对生成语音 – 图像数据性能分析以外,本文还在更具挑战性的真实语音数据集上,即 Flickr8k 和 Places-subset 数据集上,对 Single-S2IGAN 进行了定量和定性评估。

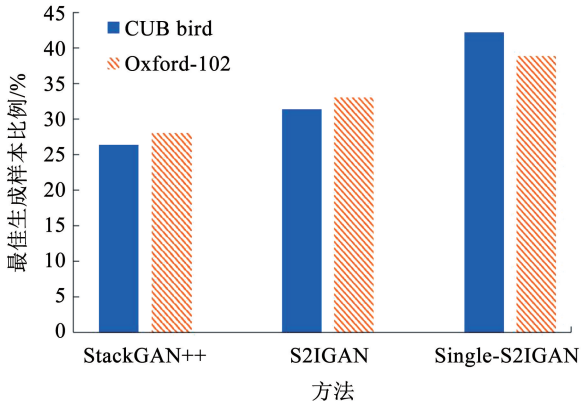


图 6 StackGAN + +, S2IGAN 和 Single-S2IGAN 的最佳生成样本比例

Fig. 6 Optimal generation sample ratios for StackGAN + +, S2IGAN, and Single-S2IGAN

3.3.1 定量分析

Single-S2IGAN 与其他方法在 Flickr8k 和 Places-subset 数据集上的 IS、FID 和 R@50 对比结果见表 2。由表 2 可以看出:在 Flickr8k 数据集上,与 S2IGAN[2] 相比,Single-S2IGAN 将 IS 均值提高了 2.98,FID 降低了 22.49,R@50 提高了 18.55;与 AttnGAN[12] 相比,Single-S2IGAN 将 FID 降低了 13.28。在 Places-subset 数据集上,与 S2IGAN[2] 相比,Single-S2IGAN 将 IS 均值提高了 1.01,FID 降低了 16.41,R@50 提高了 15.42;与 AttnGAN[12] 相比,Single-S2IGAN 将 FID 降低了 9.91,同时将 IS 均值提高了 0.46。说明 Single-S2IGAN 的性能优于其他方法,且生成的样本分布更接近真实数据分布。

表 2 文本与语音到图像生成方法的 IS、FID 和 R@50 对比

Tab. 2 Comparison of IS, FID, and R@50 for text and speech-to-image generation methods

数据集	方法名称	输入	IS	FID	R@50
Flickr8k	AttnGAN[12]	文本	12.37 ±0.41	84.08	50.40
	StackGAN++[15]	语音	8.36 ±0.39	101.74	16.40
	S2IGAN[2]	语音	8.72 ±0.34	93.29	16.40
	Single-S2IGAN	语音	**11.70 ±0.45**	**70.80**	**34.95**
Places-subset	AttnGAN[12]	文本	4.59 ±0.51	35.59	33.85
	Li 等提出的方法[1]	语音		83.06	
	StackGAN++[15]	语音	3.78 ±0.35	47.94	8.87
	S2IGAN[2]	语音	4.04 ±0.25	42.09	12.95
	Single-S2IGAN	语音	**5.05 ±0.10**	**25.68**	**28.37**

3.3.2 定性分析

Single-S2IGAN 和 S2IGAN 在 Flickr8k 和 Places-subset 数据集上生成的图像以及对应的真实图像如图 7 所示。图 7 中第 1 行是语音对应的文本描述,第 2 行是数据集中的真实图像,第 3 ~ 4 行分别是 S2IGAN[2]、Single-S2IGAN 基于语音描述生成的图像。由图 7(a)可以看出:当提供语音信号时,在生成复杂场景方面,S2IGAN 产生的对象不清晰(第 1、2、4 ~ 6 和 8 列)且背景模糊(第 4 和 7 列);与 S2IGAN 相比,Single-S2IGAN 生成的对象在背景上

更清晰，视觉上更合理，细节上更生动逼真，而且生成的目标数量正确（第 1 ~ 3、5 ~ 8 列）。

由图 7（b）可以看出：尽管 Places-subset 数据集非常具有挑战性，但 Single-S2IGAN 生成了令人期待且高质量的复杂场景图像，即逼真的卧室（第 1 列）、合理的厨房（第 2 和 8 列）、高质量的客厅（第 3 和 5 列）、视觉上令人期待的餐厅（第 4 和 7 列）以及逼真的酒店客房（第 6 列）。然而，S2IGAN 生成的一些例子（第 1、4、6 ~ 8 列）是不合理的，如床的形状不清晰（第 1 和 6 列）、色彩分布粗糙（第 4 和 8 列）。



(a) Flickr8k数据集

(b) Places-subset数据集

图 7　Single-S2IGAN 与 S2IGAN 在 Flickr8k 和 Places-subset 数据集上的可视化结果

Fig. 7　Visual results of Single-S2IGAN and S2IGAN on Flickr8k and Places-subset datasets

以上分析表明，借助有效的视觉 - 语音融合模块和损失函数，Single-S2IGAN 能够在真实语音数据集上取得非常好的结果。

3.3.3　人类评估

为了进一步评估 StackGAN++、S2IGAN 和 Single-S2IGAN 的图像质量和语义一致性，在 Flickr8k 和 Places-subset 数据集上进行了人类测试，测试结果如图 8 所示。由图 8 可以看出，在 Flickr8k 数据集上，Single-S2IGAN 优于 S2IGAN 25% 左右；在 Places-subset 数据集上，优于 S2IGAN 18% 左右，体现出了 Single-S2IGAN 在真实语音 - 图像生成方面的优越性。

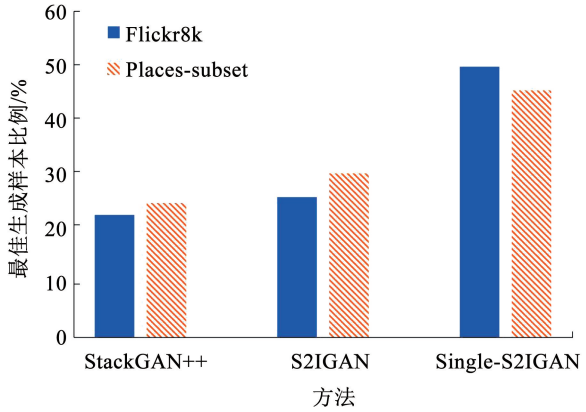


图 8 StackGAN++, S2IGAN 和 Single-S2IGAN 的最佳生成样本比例

Fig. 8 Optimal generation sample ratios for StackGAN++, S2IGAN, and Single-S2IGAN

4 结 论

1)提出了一种新的单阶段语音到图像生成对抗网络(Single-S2IGAN)框架,该框架通过一组生成器/对抗器,能够根据语音描述生成高分辨率且语义一致的图像,实现了语音到图像生成的高效性。

2)设计一种视觉 - 语音融合模块(VSFM),该模块由批归一化、像素注意力模块、语音调制模块和加权融合模块组成。VSFM 利用语音信息调制视觉特征图,显著提升了生成样本的分辨率,进而增强了图像的视觉表现。

3)提出将双模态信息应用于生成器网络的所有层,使得语音特征能够影响框架中各个层次的特征表达,极大地提高了生成图像的质量和语义一致性。

4)相较于当前的多阶段语音到图像生成方法,Single-S2IGAN 展现出了显著的优势,且其性能接近文本到图像生成方法。更重要的是,Single-S2IGAN 克服了现有多阶段语音到图像方法中存在的问题,为构建更先进的语音到图像生成模型奠定了基础。

参考文献

[1] LI Jiguo, Zhang Xinfeng, Jia Chuanmin, et al. Direct speech-to-image translation[J]. IEEE Journal of Selected Topics in Signal Processing, 2020, 14(3): 517. DOI: 10.1109/JSTSP.2020.2987417

[2] Wang Xinsheng, Qiao Tingting, Zhu Jihua, et al. Generating images from spoken descriptions[J]. IEEE/ACM Transactions on Audio Speech and Language Processing, 2021, 29: 850. DOI:10.1109/TASLP.2021.3053391

[3] Zhang Zhenxing, Schomaker L. DiverGAN: an efficient and effective single-stage framework for diverse text-to-image generation[J]. Neurocomputing, 2022, 473: 182. DOI:10.1016/j.neucom.2021.12.005

[4] Zhang Zhenxing, Schomaker L. DTGAN: dual attention generative adversarial networks for text-to-image generation[C]//Proceedings of the International Joint Conference on Neural Networks. Shenzhen: IEEE, 2021: 1. DOI: 10.1109/IJCNN52387.2021.9533527

[5] Zhu Minfeng, Pan Pingbo, Chen Wei, et al. DM-GAN: aynamic memory generative adversarial networks for text-to-image synthesis[C]//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 5802. DOI: 10.1109/CVPR.2019.00595

[6] Tao Ming, Tang Hao, Wu Songsong, et al. DF-GAN: deep fusion generative adversarial networks for text-to-image synthesis[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Orleans: IEEE, 2023: 3419. DOI: 10.1109/CVPR52688.2022.01602

[7] Wah C, Branson S, Welinder P. The caltech-UCSD birds-200-2011 dataset[J]. California Institute of Technology, 2010, 11(2): 135

[8] Nilsback M, Zisserman A. Automated flower classification over a large number of classes[C]//Proceedings of the 6th Indian Conference on Computer Vision, Graphics and Image Processing. Bhubaneswar: IEEE, 2008: 722. DOI: 10.1109/ICVGIP.2008.47

[9] Hodosh M, Young P, Hockenmaier J. Framing image description as a ranking task: data, models and evaluation metrics[J]. Journal of Artificial Intelligence Research, 2020, 47(4): 853. DOI: 10.1613/JAIR.3994

[10] Zhou Bolei, Lapedriza A, Xiao Jianxiong, et al. Learning deep features for scene recognition using places database[J]. Advances in Neural Information Processing Systems, 2018, 27(12): 1023. DOI: 10.1613/AIPS.2014.10046

[11] Gan Yanfen, Chung J, Young J, et al. A novel fusion algorithm for copy-move forgery detection[C]//Proceedings of the 7th International Conference on Virtual Reality and Visualization. Zhengzhou: IEEE, 2017: 203. DOI:10.1109/ICVRV.2017.00048

[12] Xu Tao, Zhang Pengchuan, Huang Qiuyuan, et al. AttnGAN: fine-grained text to image generation with attentional generative adversarial networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 1316. DOI:10.1109/CVPR.2018.00143

[13] Salimans T, Goodfellow I, Zaremba W, et al. Improved techniques for training GANs[C]//30th Annual Conference on Neural Information Processing Systems, NIPS 2016. Barcelona: Neural Information Processing Systems Foundation, 2016: 2234

[14] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 2818. DOI: 10.1109/CVPR.2016.308

[15] Zhang Han, Xu Tao, Li Hongsheng, et al. StackGAN++: realistic image synthesis with stacked generative adversarial networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(8): 1947. DOI:10.1109/TPAMI.2018.2856256

[16] Qiao Tingting, Zhang Jing, Xu Duanqing, et al. MirrorGAN: learning text-to-image generation by redescription[C]//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 1505. DOI:10.1109/CVPR.2019.00160

[17] Li Bowen, Qi Xiaojuan, Lukasiewicz T, et al. Controllable text-to-image generation[C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems, NeurIPS 2019. Vancouver: Neural Information Processing Systems Foundation, 2019: 2065

[18] Yin Guojun, Liu Bin, Sheng Lu, et al. Semantics disentangling for text-to-image generation[C]//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 2327. DOI: 10.1109/CVPR.2019.00243