DOI:10.11918/202412036

基于词级注意力的多样化文本到图像生成

张振兴,杨任农,俞利新,王 瑛,李永林,郭安新

(空军工程大学 空管领航学院,西安 710051)

摘 要:为解决单阶段文本到图像生成技术存在的生成结果单一、精确调控困难等局限性,提出一种基于词级注意力的多样化文本到图像生成对抗网络框架(diverse text-to-image generative adversarial network framework based on word-level attention, WordGAN)。该框架在条件生成对抗网络的基础上,引入双重注意力机制、双残差连接设计和线性层策略。其中,双重注意力机制由条件自适应实例-层归一化、通道注意力模块和像素注意力模块组成,通过量化分析文本描述中各词汇的语义贡献度捕捉网络特征通道及空间像素的权重分布;双残差连接设计用于维持底层视觉特征完整性和支持更深层次的网络扩展;线性层策略利用线性层改变原有图像模式,促进框架探索更多次要模式。在 CUB bird、MSCOCO 和 Oxford-102 基准数据集上进行一系列实验,结果表明:WordGAN 有效地实现了生成图像的多样性,而且不会损害质量和语义一致性,其性能水平优于所对比的文本到图像生成框架。本文提出的基于词级注意力的多样化框架为单阶段文本到图像生成任务研究提供了新的思路,具有实际应用价值。

关键词:词级注意力机制;双残差网络;模式坍塌;单阶段框架;文本到图像生成

中图分类号:V247 **文献标志码**:A **文章编号**:0367-6234(2026)09-0047-13

Diverse text-to-image synthesis based on word-level attention

Zhang Zhenxing, Yang Rennong, Yu Lixin, Wang Ying, Li Yonglin, Guo Anxin

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

Abstract: To address the limitations of single-stage text-to-image generation technology, such as limited diversity of generated results and difficulty in precise control, this paper proposes a diverse text-to-image generative adversarial network framework based on word-level attention (WordGAN). Building on conditional generative adversarial networks, the framework incorporates a dual attention mechanism, a dual residual connection design, and a linear layer strategy. The dual attention mechanism consists of conditionally adaptive instance-layer normalization, a channel-attention module, and a pixel-attention module. It captures the weight distribution of network feature channels and spatial pixels by quantitatively analyzing the semantic contribution of each word in the text description. The dual residual connection design maintains the integrity of low-level visual features and supports deeper network expansion. The linear layer strategy uses linear layers to alter original image patterns, encouraging the framework to explore more secondary patterns. A series of experiments was conducted on the CUB bird, MSCOCO, and Oxford-102 benchmark datasets. The results show that WordGAN effectively achieves diversity in generated images without compromising quality and semantic consistency, and it outperforms other baseline text-to-image generation frameworks. The proposed diverse framework based on word-level attention provides a new approach for research on single-stage text-to-image generation tasks and demonstrates practical application value.

Keywords: word-level attention mechanism; dual-residual network; mode collapse; single-stage architecture; text-to-image generation

文本到图像生成任务能够根据给定的文本描述自动生成高质量且语义一致的图像。这一任务在艺术创作、计算机辅助设计、数据增强、照片编辑等多个应用场景中具有重要价值[1]。近年来,文本到图像生成任务已发展成为连接计算机视觉与自然语言理解两大领域的前沿课题。

为了克服多阶段文本到图像生成框架的局限性,研究者们致力于开发单阶段文本到图像生成框架,其中具有代表性的工作是 DTGAN[1] 和 DF-GAN[2] 框架。这类单阶段框架通过采用一组生成

收稿日期:2024-12-11;**录用日期**:2025-03-06;**网络首发日期**:2025-11-24
网络首发地址:https://link.cnki.net/urlid/23.1235.T.20251121.1910.014
基金项目:西安市青年人才托举计划(0959202513098);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:俞利新,2207621676@qq.com

器/对抗器生成与文本描述语义一致的高质量图像。然而,当前的单阶段文本到图像生成框架仍面临两个关键挑战:1)模式坍塌问题。该问题表现为生成器在给定文本输入时,生成模糊的图像。在严重情况下,条件生成对抗网络(conditional generative adversarial network,cGAN)可能基于同一文本描述生成高度相似的图像集合[3]。如图1所示,尽管引入了随机噪声,DF-GAN仍然难以产生多样性的图像。这种局限性显著降低了生成图像的多样性,进而影响其实际应用价值。以数据增强为例,只有生成具有丰富形态变化的图像,才能有效提升分类模型的鲁棒性。2)现有方法主要依赖全局句子向量进行特征图调制,缺乏对单词级别细粒度信息的有效利用。这种粗粒度的特征表示,限制了模型根据自然语言描述中的具体细节精确控制图像生成过程的能力,难以实现对图像局部特征的精细调控。

基于此,本文提出基于词级注意力的多样化文本到图像生成对抗网络框架(diverse text-to-image generative adversarial network framework based on word-level attention,WordGAN)。该框架包括两种新的词级注意力模块,分别为通道注意力模块(channel-attention module,CAM)和像素注意力模块(pixel-attention module,PAM),两种模块能够引导模型更关注与给定文本描述中重要的单词(如形容词和名词等)有语义联系的关键通道和像素。同时,WordGAN引入一个双残差块和一个线性层,每个残差块包括卷积层、条件自适应实例-层归一化(conditional adaptive instance-layer normalization,CAdaILN)和调制模块,线性层用于增强网络的生成能力。

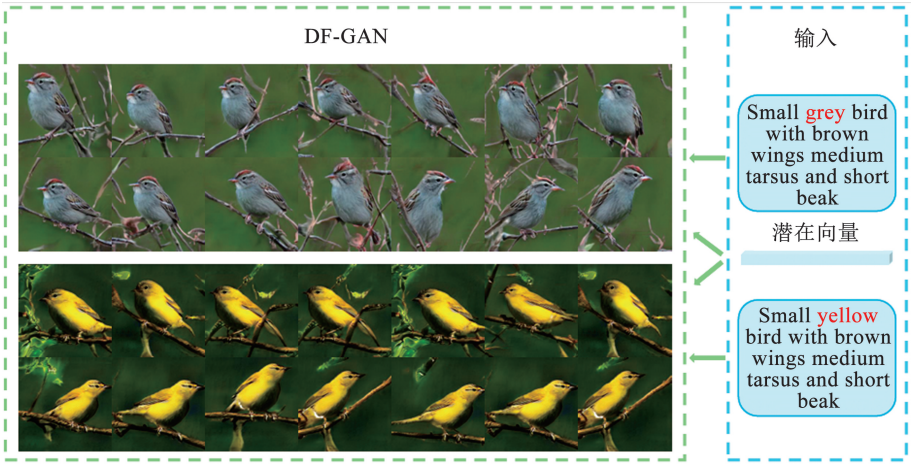


图1　DF-GAN缺乏多样性示例

Fig. 1　Lack-of-diversity examples of DF-GAN

1　基础概念

1.1　cGAN

作为生成对抗网络模型(generative adversarial network,GAN)的衍生模型,cGAN将特定约束信息 $\boldsymbol{c}$(如类别标签、文本描述和低分辨率图像)和随机噪声向量 $\boldsymbol{z}$ 共同作为生成器 G 的输入,进而产生与 $\boldsymbol{c}$ 语义相符的图像[4]。该模型的对抗器 D 负责判断真实图像-条件对$(\boldsymbol{x},\boldsymbol{c})$与生成图像-条件对$(G(\boldsymbol{z},\boldsymbol{c}),\boldsymbol{c})$。cGAN的优化目标 $V(G,D)$ 可表示为

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

式中:$p_{\text{data}}(\boldsymbol{x})$、$p_{\boldsymbol{z}}(\boldsymbol{z})$分别为真实图像 $\boldsymbol{x}$ 的概率分布与潜在向量 $\boldsymbol{z}$ 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,$E_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z}),\boldsymbol{c}\sim p_{\text{c}}(\boldsymbol{c})}[\cdot]$为生成数据分布期望值。

1.2　模式坍塌分析

模式坍塌是指模型未能学习真实数据分布的所有模式,出现生成图像缺乏多样性的现象。对于cGAN来说,模式坍塌指的是根据相同的条件上下文产生的图像是相似的[5]。缺乏多样性问题的主要原因之一是生成器受网络生成能力的限制,仅捕捉真实数据分布的一部分,未能获取全部模态[6]。尽管生成器试图将随机潜在向量映射到原始数据分布中,但这样的映射不是满射的[7]。当设计的模型不够强大时,网络只能捕捉真实数据分布的少数几个模态,并倾向于将不同的输入映射到这些模态。

在单阶段文本到图像生成框架中,缺乏多样性问题变得更加严重[8]。在单阶段框架中,随机潜在向量被用作输入,负责图像变化,而文本特征则作为辅助信息调制视觉特征图,并确定主要的图像内容。为了增强生成图像的语义一致性,生成器通常利用调制模块对每个尺度的特征图进行调制。此作法可

能使网络更倾向于高维结构化的文本上下文特征，进而忽略低维潜在向量。本文旨在增强网络的生成能力，加强随机潜在向量对生成图像的控制，以改善多样性。

2　WordGAN

2.1　WordGAN 的总体框架

WordGAN 的总体框架如图 2 所示。第 1 层利用一个全连接层（fully-connected layer，FC）处理潜在向量 $z \in \mathbf{R}^{100}$，并将结果重塑为初始特征图 $\boldsymbol{F}_0 \in \mathbf{R}^{4\times4\times256}$。之后，$\boldsymbol{F}_0$ 被输入到基本生成器模块，该模块主要包括 7 个双重残差块，接收来自文本编码器的文本嵌入，将其作为额外的条件上下文，以提升生成图像的质量。

双重残差块由两个残差模块组成，每个残差模块由一组卷积层、条件自适应实例-层归一化（conditional adaptive instance-layer normalization，CAdaILN）、ReLU 激活函数和调制模块（通道注意力模块 CAM 和像素注意力模块 PAM）构成。同时，在两个残差模块之间插入 Leaky-ReLU 激活的卷积层[9]和 CAdaILN，以减轻对文本条件信息的控制。双重残差块不仅通过保留比级联卷积更多的原始视觉特征，加速收敛速度，而且能够有效增加网络深度，进而提高网络容量，生成具有更多细节的高质量图像。

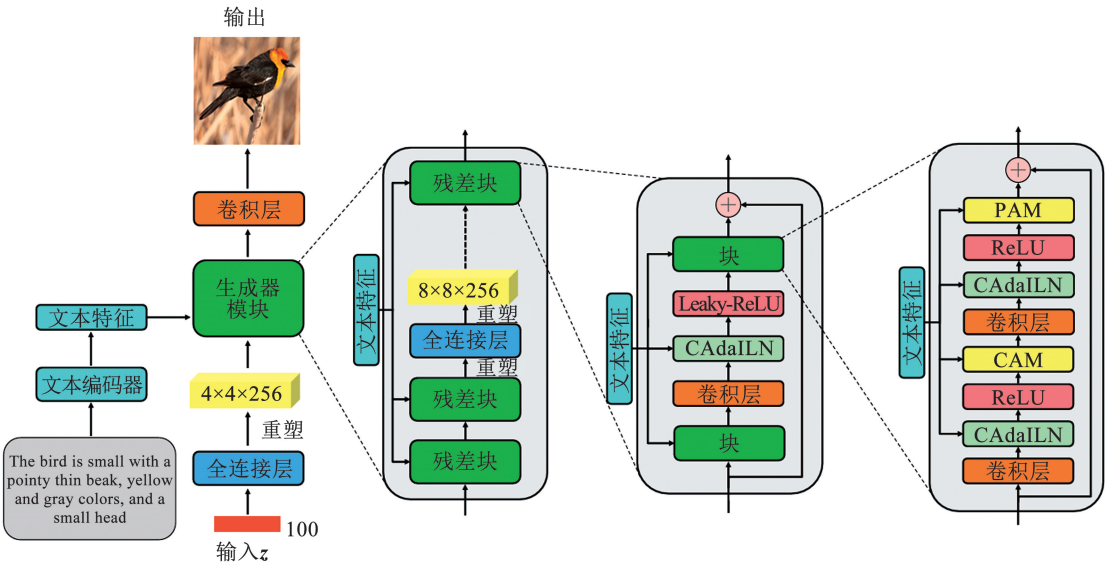


图 2　WordGAN 整体框架

Fig. 2　Overall architecture of the WordGAN

为解决缺乏多样性的问题，本文在网络结构设计上进行了一系列实验，以加强随机潜在向量对视觉特征图的控制，进而增强生成图像变化。实验结果显示，FC 可以显著提升网络的生成能力，促进模型探索真实数据分布的次要模式。因此，本文尝试在文本到图像生成框架中插入一个 FC。实验结果表明，在第 2 个双重残差块之后插入一个 FC，可以在图像质量和多样性上取得最佳的性能。具体来说，将第 2 个双重残差块的输出特征图 $\boldsymbol{F}_2 \in \mathbf{R}^{8\times8\times256}$ 首先调整为 $\mathbf{R}^{16\,384}$ 并输入到一个保持输入特征维度的 FC；然后，$\boldsymbol{F}_2$ 再次被重塑为 $\mathbf{R}^{8\times8\times256}$ 并传递到下一个双重残差块；最后，生成器模块的输出会输入至一个由 tanh 函数激活的卷积层[10]，进而生成最终图像。

通过插入一个 FC，网络无法利用前面特征图的空间二维布局信息，需要将所有必要信息编码为一维向量（嵌入），以此作为后续层中展开至二维的基础。因此，在框架的这个位置，将得到一个可注入随机噪声的表示，进而增加生成模式的多样性。网络难以轻易复制早期特征图中的（部分）二维模式，避免了许多 GAN 框架出现的“查找表”行为。

如果 FC 在网络中位置靠前，其获得的便是早期、粗糙的特征信息，这些信息不足以生成语义一致的输出模式。例如，如果早期的特征图表示的是小鸟组件，网络将无法将其组装成完整的鸟。另一方面，如果 FC 在网络中位置靠后，其接收到的将是几乎完整的模式，进而导致生成图像缺乏多样性。此外，因为 GPU 内存有限，在第 3 个或更深层的残差块之后插入一个 FC 非常困难。

2.2　双重注意力机制

条件上下文和视觉特征图之间的语义关联对图像生成尤为关键。然而，对于文本到图像生成来说，这种关系会变得更加复杂，因为给定的句子包含一系列词汇，这些词汇对生成图像有不同的贡献。例

如,输入文本描述中的形容词会更多地与生成的图像相关联,而不定冠词“the”则影响较小。此外,尽管双重残差结构有利于模型容量和训练稳定性,但可能会引入噪音和冗余信息。基于这些原因,本文设计两种新的词级注意力模块,分别为通道注意力模块 CAM 和像素注意力模块 PAM,用于探索上下文特征与视觉特征图之间的潜在语义关系。CAM 和 PAM 能识别给定文本描述中的重要词汇(如形容词和名词),使网络关注与这些词汇在语义上相关联的关键通道和像素,还可以从通道和空间角度减轻语义无关和冗余特征的影响。

2.2.1 条件自适应实例-层归一化

为提高 GAN 的训练稳定性,大多数现有的文本到图像生成框架采用批归一化(batch normalization, BN)[11]应用于整个批次生成的图像而不是单个图像。然而,BN 的收敛性严重依赖于批次大小。此外,在文本到图像生成领域,BN 的优势并不太明显,因为每个生成图像更多地与给定的文本描述和特征图本身相关联。因此,通过 CAdaILN,在特征映射的层级和通道维度实施归一化操作,并整合从全局语句向量中提取的语义信息调制归一化后的特征表示,如图 3 所示。具体而言,通过两个全连接层矩阵 $\boldsymbol{W}_1$ 和 $\boldsymbol{W}_2$ 将句子向量 $\boldsymbol{s}$ 转换为语言线索 $\boldsymbol{\gamma}\in\mathbf{R}^{1\times1\times C}$ 和 $\boldsymbol{\beta}\in\mathbf{R}^{1\times1\times C}$。同时,对视觉特征图使用实例归一化(instance normalization, IN)和层归一化(layer normalization, LN)进行处理。对 IN 和 LN 的输出 $\hat{\boldsymbol{a}}_{\mathrm{I}}$ 和 $\hat{\boldsymbol{a}}_{\mathrm{L}}$ 进行自适应求和来获得归一化的特征图 $\hat{\boldsymbol{f}}$。然后,利用 $\boldsymbol{\gamma}$ 和 $\boldsymbol{\beta}$ 对 $\hat{\boldsymbol{f}}$ 进行缩放和移位。CAdaILN 过程的数学表达式为:

$$\boldsymbol{\gamma}=\boldsymbol{W}_1\boldsymbol{s} \tag{2}$$

$$\boldsymbol{\beta}=\boldsymbol{W}_2\boldsymbol{s} \tag{3}$$

$$\hat{\boldsymbol{f}}=\boldsymbol{\rho}\odot\hat{\boldsymbol{a}}_{\mathrm{I}}+(1-\boldsymbol{\rho})\odot\hat{\boldsymbol{a}}_{\mathrm{L}} \tag{4}$$

$$\hat{\boldsymbol{a}}=\boldsymbol{\gamma}\odot\hat{\boldsymbol{f}}+\boldsymbol{\beta} \tag{5}$$

式中 $\boldsymbol{\rho}\in\mathbf{R}^{1\times1\times C}$ 为可学习参数,决定 IN 和 LN 的比例,其取值限定在[0,1]。同时,$\boldsymbol{\rho}$ 与生成器参数一起更新。

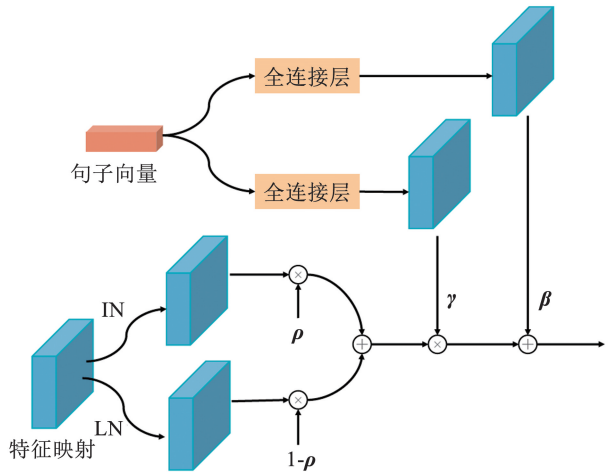


图 3 CAdaILN 流程

Fig. 3 Process of CAdaILN

2.2.2 CAM

CAM 的详细结构如图 4 所示。给定特征图 $\boldsymbol{F}_{\mathrm{c}}\in\mathbf{R}^{H\times W\times C}$($H$、$W$、$C$ 分别表示 $\boldsymbol{F}_{\mathrm{c}}$ 的高度、宽度、通道数),首先采用全局平均池化和最大池化对其进行处理,以提取整体统计信息和显著特征,分别得到通道特征向量 $\boldsymbol{F}_{\mathrm{ca}}\in\mathbf{R}^{1\times1\times C}$ 和 $\boldsymbol{F}_{\mathrm{cm}}\in\mathbf{R}^{1\times1\times C}$。随后,采用两个独立的 1×1 卷积层而不是一个 1×1 卷积处理 $\boldsymbol{F}_{\mathrm{ca}}$ 和 $\boldsymbol{F}_{\mathrm{cm}}$,因为平均池化和最大池化捕获不同的全局空间特征分布。然后,输出维度被转换为 $\mathbf{R}^{1\times C}$ 并在第 1 维上重复 C 次以获得平均池化查询 $\boldsymbol{F}_{\mathrm{caq}}\in\mathbf{R}^{C\times C}$ 和最大池化查询 $\boldsymbol{F}_{\mathrm{cmq}}\in\mathbf{R}^{C\times C}$ 两个特征。以上过程的数学表达式分别为:

$$\boldsymbol{F}_{\mathrm{caq}}=C_{\mathrm{re}}(C_{\mathrm{aq}}(P_{\mathrm{avg}}(\boldsymbol{F}_{\mathrm{c}}))) \tag{6}$$

$$\boldsymbol{F}_{\mathrm{cmq}}=C_{\mathrm{re}}(C_{\mathrm{mq}}(P_{\mathrm{max}}(\boldsymbol{F}_{\mathrm{c}}))) \tag{7}$$

式中:P_{avg} 和 P_{max} 分别表示全局平均池化和最大池化,C_{aq} 和 C_{mq} 表示 1×1 卷积层,C_{re} 表示重塑和重复操作。

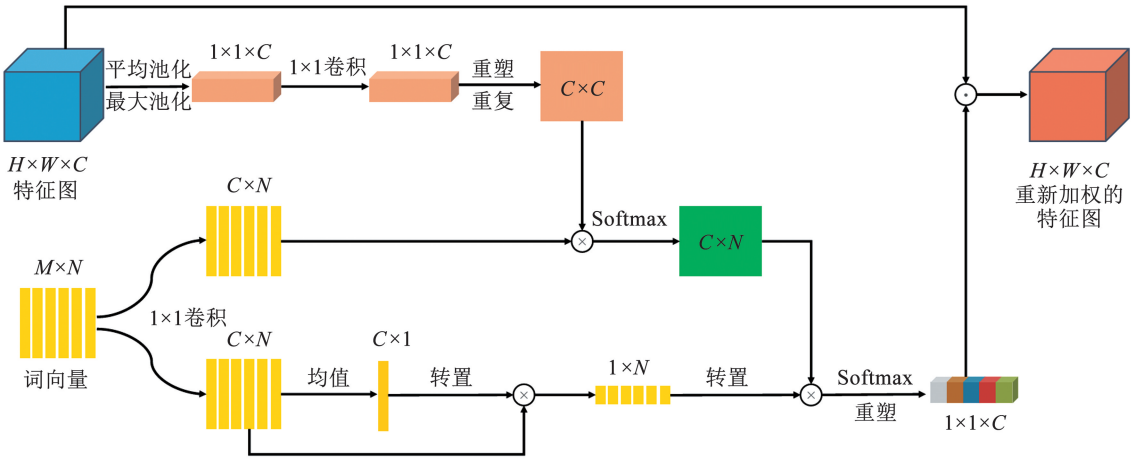


图 4 通道注意力模块示意

Fig. 4 Overview of the channel-attention module

对于单词-上下文向量 $\boldsymbol{E}\in\mathbf{R}^{M\times N}$（$M$ 表示单词嵌入的维度，N 表示给定文本描述中的单词数量），将其传入两个不同的 1×1 卷积层，然后经过 ReLU 激活函数，生成两个上下文向量：键 $\boldsymbol{F}_{\mathrm{ck}}\in\mathbf{R}^{C\times N}$ 和值 $\boldsymbol{F}_{\mathrm{cv}}\in\mathbf{R}^{C\times N}$，其位于视觉特征的共同语义空间中。接着，沿第 2 维计算 $\boldsymbol{F}_{\mathrm{cv}}$ 的均值，并将其维度调整为 $\mathbf{R}^{1\times C}$。同时，将结果与 $\boldsymbol{F}_{\mathrm{cv}}$ 相乘，以获得表示句子中每个单词重要性的上下文注意力图 $\boldsymbol{E}_{\mathrm{ci}}\in\mathbf{R}^{1\times N}$，注意力图中的较大值意味着相应的单词与生成图像更加相关。上下文注意力图的获取过程可表示为：

$$\boldsymbol{F}_{\mathrm{ck}}=C_{\mathrm{ck}}(\boldsymbol{E}) \tag{8}$$

$$\boldsymbol{F}_{\mathrm{cv}}=C_{\mathrm{cv}}(\boldsymbol{E}) \tag{9}$$

$$\boldsymbol{E}_{\mathrm{ci}}=C_{\mathrm{mean}}(\boldsymbol{F}_{\mathrm{cv}})\times\boldsymbol{F}_{\mathrm{cv}} \tag{10}$$

式中：C_{ck} 和 C_{cv} 表示带有 ReLU 激活函数的 1×1 卷积层，C_{mean} 表示平均和重塑操作。

为建模单词-上下文特征和通道之间的语义关联，在查询向量和键 $\boldsymbol{F}_{\mathrm{ck}}\in\mathbf{R}^{C\times N}$ 之间进行点积操作，并应用 Softmax 函数得到上下文通道注意力矩阵，该矩阵表示文本描述中通道与单词之间的语义相似性权重。最后，通过对上下文通道注意力矩阵与 $\boldsymbol{E}_{\mathrm{ci}}$ 的转置进行点积操作，使用 Softmax 函数处理结果，并将其维度调整为 $\mathbf{R}^{1\times1\times C}$，以得到通道注意力权重 $\boldsymbol{W}_{\mathrm{ca}}$ 和 $\boldsymbol{W}_{\mathrm{cm}}$。上述过程的数学表达式为：

$$\boldsymbol{F}_{\mathrm{a}}=\mathrm{Softmax}(D_{\mathrm{is}}(\boldsymbol{F}_{\mathrm{caq}},\boldsymbol{F}_{\mathrm{ck}})) \tag{11}$$

$$\boldsymbol{F}_{\mathrm{m}}=\mathrm{Softmax}(D_{\mathrm{is}}(\boldsymbol{F}_{\mathrm{cmq}},\boldsymbol{F}_{\mathrm{ck}})) \tag{12}$$

$$\boldsymbol{W}_{\mathrm{ca}}=C_{\mathrm{re}}(\mathrm{Softmax}(D_{\mathrm{is}}(\boldsymbol{F}_{\mathrm{a}},\boldsymbol{E}_{\mathrm{ci}}^{\mathrm{T}}))) \tag{13}$$

$$\boldsymbol{W}_{\mathrm{cm}}=C_{\mathrm{re}}(\mathrm{Softmax}(D_{\mathrm{is}}(\boldsymbol{F}_{\mathrm{m}},\boldsymbol{E}_{\mathrm{ci}}^{\mathrm{T}}))) \tag{14}$$

式中：$D_{\mathrm{is}}(\cdot)$ 表示点积操作，$\boldsymbol{F}_{\mathrm{a}}\in\mathbf{R}^{C\times N}$、$\boldsymbol{F}_{\mathrm{m}}\in\mathbf{R}^{C\times N}$ 分别为 $\boldsymbol{F}_{\mathrm{caq}}$、$\boldsymbol{F}_{\mathrm{cmq}}$ 的上下文通道注意力矩阵，$\boldsymbol{W}_{\mathrm{ca}}\in\mathbf{R}^{1\times1\times C}$、$\boldsymbol{W}_{\mathrm{cm}}\in\mathbf{R}^{1\times1\times C}$ 分别为 $\boldsymbol{F}_{\mathrm{a}}$、$\boldsymbol{F}_{\mathrm{m}}$ 的通道注意力权重。

在获取通道注意力分数后，将其与原始特征图相乘，以重新加权视觉特征图，使得网络更多地关注特征图中有用的通道，为其分配更大的权重。同时，设计一种自适应门控方法，动态合并全局平均池化和最大池化的输出特征图。融合特征图 $\boldsymbol{F}_{\mathrm{cw}}\in\mathbf{R}^{H\times W\times C}$ 的获取过程如下：

$$\boldsymbol{F}_{\mathrm{caw}}=\boldsymbol{W}_{\mathrm{ca}}\odot\boldsymbol{F}_{\mathrm{c}} \tag{15}$$

$$\boldsymbol{F}_{\mathrm{cmw}}=\boldsymbol{W}_{\mathrm{cm}}\odot\boldsymbol{F}_{\mathrm{c}} \tag{16}$$

$$\boldsymbol{g}_{\mathrm{ca}}=\sigma(\boldsymbol{C}_{\mathrm{ga}}(C_{\mathrm{re}}(\boldsymbol{F}_{\mathrm{ca}}))+\boldsymbol{C}_{\mathrm{gm}}(C_{\mathrm{re}}(\boldsymbol{F}_{\mathrm{cm}}))) \tag{17}$$

$$\boldsymbol{F}_{\mathrm{cw}}=\boldsymbol{g}_{\mathrm{ca}}\times\boldsymbol{F}_{\mathrm{caw}}+(1-\boldsymbol{g}_{\mathrm{ca}})\times\boldsymbol{F}_{\mathrm{cmw}} \tag{18}$$

式中：⊙表示逐元素相乘，$\boldsymbol{F}_{\mathrm{caw}}$、$\boldsymbol{F}_{\mathrm{cmw}}$ 为重新加权的特征图，$\boldsymbol{g}_{\mathrm{ca}}$ 为用于融合视觉特征图的响应门控，$\boldsymbol{C}_{\mathrm{ga}}$ 和 $\boldsymbol{C}_{\mathrm{gm}}$ 为将通道数量减少至 1 的 FC 矩阵，$\sigma(\cdot)$ 为 Sigmoid 函数。

为了进一步保留基本特征并稳定 cGAN 的学习，采用自适应残差连接生成最终结果 $\boldsymbol{F}_{\mathrm{cu}}\in\mathbf{R}^{H\times W\times C}$。上述过程的数学表达式为

$$\boldsymbol{F}_{\mathrm{cu}}=\gamma_{\mathrm{c}}\times\boldsymbol{F}_{\mathrm{cw}}+\boldsymbol{F}_{\mathrm{c}} \tag{19}$$

式中 γ_{c} 为可学习参数，初始值为 0。

2.2.3　PAM

如上所述，CAM 从通道的角度重新加权视觉特征图。然而，除了视觉特征图中的通道外，像素对质量和语义一致性也非常重要。因此，本文引入 PAM 捕获像素与单词-上下文向量之间的空间相互关系，使得网络能够更多关注重要的像素。值得注意的是，PAM 是在 CAM 的输出结果上执行的，因为特征图中同一通道的像素仍然共享相同的权重。

PAM 的框架如图 5 所示。对于特征图 $\boldsymbol{F}_{\mathrm{s}}\in\mathbf{R}^{H\times W\times C}$，首先在空间维度上对其进行平均池化和最大池化，提取全局特征，得到两个空间特征向量 $\boldsymbol{F}_{\mathrm{sa}}\in\mathbf{R}^{H\times W\times1}$ 和 $\boldsymbol{F}_{\mathrm{sm}}\in\mathbf{R}^{H\times W\times1}$。然后，将 $\boldsymbol{F}_{\mathrm{sa}}$ 和 $\boldsymbol{F}_{\mathrm{sm}}$ 的维度调整为 $\mathbf{R}^{(H\times W)\times1}$，并沿着第 2 维度重复 C 次，得到平均池化查询 $\boldsymbol{F}_{\mathrm{saq}}\in\mathbf{R}^{(H\times W)\times C}$ 和最大池化查询 $\boldsymbol{F}_{\mathrm{smq}}\in\mathbf{R}^{(H\times W)\times C}$。上述过程的表达式为：

$$\boldsymbol{F}_{\mathrm{saq}}=C_{\mathrm{re}}(S_{\mathrm{avg}}(\boldsymbol{F}_{\mathrm{s}})) \tag{20}$$

$$\boldsymbol{F}_{\mathrm{smq}}=C_{\mathrm{re}}(S_{\mathrm{max}}(\boldsymbol{F}_{\mathrm{s}})) \tag{21}$$

式中 S_{avg}、S_{max} 分别表示在空间维度上的平均池化和最大池化。

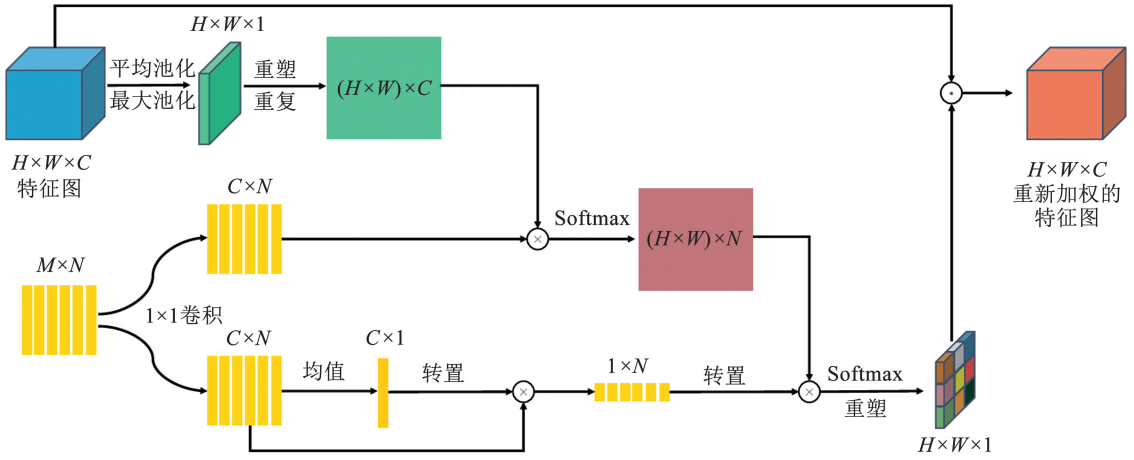


图 5　像素注意力模块示意

Fig. 5　Overview of the pixel-attention module

给定单词–上下文特征 $\boldsymbol{E}\in\mathbf{R}^{M\times N}$,采用与 CAM 相同的方式对其进行处理,产生键 $\boldsymbol{F}_{\mathrm{sk}}\in\mathbf{R}^{C\times N}$、值 $\boldsymbol{F}_{\mathrm{sv}}\in\mathbf{R}^{C\times N}$和上下文注意力图 $\boldsymbol{E}_{\mathrm{si}}\in\mathbf{R}^{1\times N}$。上述过程的数学表达式为:

$$\boldsymbol{F}_{\mathrm{sk}}=C_{\mathrm{sk}}(\boldsymbol{E}) \tag{22}$$

$$\boldsymbol{F}_{\mathrm{sv}}=C_{\mathrm{sv}}(\boldsymbol{E}) \tag{23}$$

$$\boldsymbol{E}_{\mathrm{si}}=C_{\mathrm{mean}}(\boldsymbol{F}_{\mathrm{sv}})\times\boldsymbol{F}_{\mathrm{sv}} \tag{24}$$

式中 C_{sk} 和 C_{sv} 为带有 ReLU 激活函数的 1 × 1 卷积层。

随后,通过对查询与键 $\boldsymbol{F}_{\mathrm{sk}}\in\mathbf{R}^{C\times N}$ 的点积应用 Softmax 函数,得到空间–语义注意力图。空间–语义注意力图表示像素与文本描述中单词之间的语义相似性权重。接着,对空间–语义注意力图和 $\boldsymbol{E}_{\mathrm{si}}$ 的转置进行点积操作,利用 Softmax 函数获得像素级注意力权重,并将其转换为 $\mathbf{R}^{H\times W\times 1}$。像素级注意力权重 $\boldsymbol{W}_{\mathrm{sa}}\in\mathbf{R}^{H\times W\times 1}$、$\boldsymbol{W}_{\mathrm{sm}}\in\mathbf{R}^{H\times W\times 1}$ 的获取过程如下:

$$\boldsymbol{S}_{\mathrm{a}}=\mathrm{Softmax}(D_{\mathrm{is}}(\boldsymbol{F}_{\mathrm{saq}},\boldsymbol{F}_{\mathrm{sk}})) \tag{25}$$

$$\boldsymbol{S}_{\mathrm{m}}=\mathrm{Softmax}(D_{\mathrm{is}}(\boldsymbol{F}_{\mathrm{smq}},\boldsymbol{F}_{\mathrm{sk}})) \tag{26}$$

$$\boldsymbol{W}_{\mathrm{sa}}=C_{\mathrm{re}}(\mathrm{Softmax}(D_{\mathrm{is}}(\boldsymbol{S}_{\mathrm{a}},\boldsymbol{E}_{\mathrm{si}}^{\mathrm{T}}))) \tag{27}$$

$$\boldsymbol{W}_{\mathrm{sm}}=C_{\mathrm{re}}(\mathrm{Softmax}(D_{\mathrm{is}}(\boldsymbol{S}_{\mathrm{m}},\boldsymbol{E}_{\mathrm{si}}^{\mathrm{T}}))) \tag{28}$$

式中:$\boldsymbol{S}_{\mathrm{a}}\in\mathbf{R}^{(H\times W)\times N}$、$\boldsymbol{S}_{\mathrm{m}}\in\mathbf{R}^{(H\times W)\times N}$ 分别为 $\boldsymbol{F}_{\mathrm{saq}}$、$\boldsymbol{F}_{\mathrm{smq}}$ 的空间–语义注意力图,$\boldsymbol{W}_{\mathrm{sa}}\in\mathbf{R}^{H\times W\times 1}$、$\boldsymbol{W}_{\mathrm{sm}}\in\mathbf{R}^{H\times W\times 1}$ 分别为 $\boldsymbol{S}_{\mathrm{a}}$、$\boldsymbol{S}_{\mathrm{m}}$ 的像素级注意力权重。

与 CAM 相同,对像素级注意力得分与原始特征图进行矩阵乘法,以增强视觉特征图的语义一致性。同时,为了保留重要特征,将平均池化和最大池化的重新加权特征图合并,并将其传入一个带有 ReLU 激活函数的 1 × 1 卷积层,以生成合并后的特征图 $\boldsymbol{F}_{\mathrm{sw}}\in\mathbf{R}^{H\times W\times C}$。然后,采用自适应残差连接获得最终输出 $\boldsymbol{F}_{\mathrm{su}}\in\mathbf{R}^{H\times W\times C}$。同时,将经过平均池化与最大池化重新加权后的特征图进行合并,再将其传入一个带有 ReLU 激活函数的卷积层,从而生成合并后的特征图。上述过程的数学表达式为:

$$\boldsymbol{F}_{\mathrm{saw}}=\boldsymbol{W}_{\mathrm{sa}}\odot\boldsymbol{F}_{\mathrm{s}} \tag{29}$$

$$\boldsymbol{F}_{\mathrm{smw}}=\boldsymbol{W}_{\mathrm{sm}}\odot\boldsymbol{F}_{\mathrm{s}} \tag{30}$$

$$\boldsymbol{F}_{\mathrm{sw}}=C_{\mathrm{conv}}([\boldsymbol{F}_{\mathrm{saw}};\boldsymbol{F}_{\mathrm{smw}}]) \tag{31}$$

$$\boldsymbol{F}_{\mathrm{su}}=\gamma_{\mathrm{s}}\times\boldsymbol{F}_{\mathrm{sw}}+\boldsymbol{F}_{\mathrm{s}} \tag{32}$$

式中:$\boldsymbol{F}_{\mathrm{saw}}$ 和 $\boldsymbol{F}_{\mathrm{smw}}$ 为重新加权的特征图;[;] 表示沿通道维度的连接操作;C_{conv} 为带有 ReLU 激活函数的1 ×1卷积操作;γ_{s} 为一个可学习的参数,初始值为 0。

PAM 和 CAM 对于提升生成图像的语义一致性和图像质量十分关键,原因包括以下几点:1) PAM 和 CAM 对每个单词的重要性进行建模,以突出给定文本描述中的重要单词(例如形容词和名词),使生成器更加关注生成图像的语义相关部分;2) 为提升生成效果,结合平均池化和最大池化技术,分别提取多样化的全局空间特征及通道信息;3) 设计一种自适应门控方法,动态合并平均池化和最大池化的输出特征图,以获得更好的性能;4) 将注意力权重应用至所有的通道和像素,以增强多层特征图的语义一致性,并采用自适应残差连接生成最终图像,以保留基本特征并稳定 cGAN 的学习;5) 与先前的方法不同,本文利用注意力模块在通道和空间维度上调制每个尺度特征图;6) 本文提出的 CAdaILN 可灵活控制形状和纹理的变化程度,为注意力模块提供补充。

3 仿真实验

为验证 WordGAN 在图像多样性、质量和语义一致性方面的有效性,在 3 个基准数据集上进行定量和定性实验。

3.1 实验设置

3.1.1 数据集设置

本文在 CUB bird[12]、Oxford-102[13] 和 MSCOCO[14]3 个数据集上评估 WordGAN。这些数据集也被 DF-GAN[2]、StackGAN[15]、AttnGAN[16]、MirrorGAN[17]、ControlGAN[18]、SDGAN[19] 和 DM-GAN[20] 等框架所使用。

1) CUB bird 数据集。该数据集包含 11 788 张图像,其中 8 855 张图像属于训练集,2 933 张图像属于测试集。每张图像包含 10 个对应的文本描述。本文采用 StackGAN 中的方法对 CUB bird 数据集进行预处理。

2) Oxford-102 数据集。该数据集包括 5 878 张训练图像和 2 311 张测试图像,每张图像均附带 10 个文本描述。

3) MSCOCO 数据集。作为一个更具挑战性的数据集,MSCOCO 共包含 123 287 张图像,其中 82 783 张图像用于训练,40 504 张图像用于测试。每张图像均配有 5 条人工标注的文本描述信息。

3.1.2 参数设置

对于文本编码器,本文遵循 DTGAN[1] 的方法,利用预训练的双向 LSTM 获取词向量和全局句子向量。同时,采用 DTGAN 损失函数。潜在向量的维度设置为 100。在训练方面,使用 Adam 优化器训练网络,其中衰减率 $b=(0,0.9)$。同时,采用双时间尺度更新策略[21],分别将生成器和对抗器的学习率设定为 0.000 1 和 0.000 4,批次大小设置为 16。使用 PyTorch 实现 WordGAN,所有实验均在一块 NVIDIA Tesla V100 GPU(32 GB 内存)上进行。

3.1.3　评估指标

采用 3 项通用评估指标进行性能评估，分别为感知分数（inception score，IS）[22]、Fréchet 感知距离（Fréchet inception distance，FID）[23]和可学习感知图像块相似度（learned perceptual image patch similarity，LPIPS）[24]。前两项指标用于评估图像质量，LPIPS 用于评估生成图像的多样性。LPIPS 通过计算生成图像之间的平均特征距离检测多样性，LPIPS 越大，表示生成的图像差异性越大，多样性越强。

对于 MSCOCO 数据集，IS 均值无法评估图像质量，可能会出现饱和甚至过拟合的情况[1-2]。因此，在 MSCOCO 数据集上，本文不采用 IS 均值作为评估指标。

对于 Oxford-102 数据集，由于缺乏比较分数，本文没有列出对应的 FID。在测试时，根据训练集中未出现的文本描述生成了 30 000 张图像用于计算 IS 均值和 FID。

3.2　与先前框架的比较

3.2.1　定量结果

在 CUB bird、Oxford-102 以及 MSCOCO 3 个基准数据集上，将 WordGAN 与现有的单阶段及多阶段条件生成对抗网络文本到图像生成框架进行对比。WordGAN 与其他框架在 CUB bird 和 Oxford-102 数据集上的 IS 均值结果见表 1。由表 1 可以看出，WordGAN 取得了最佳的性能。在 CUB bird 数据集上，相比于 DTGAN 框架，WordGAN 将 IS 均值提高了 0.25；在 Oxford-102 数据集上，相比于 DTGAN 框架，WordGAN 将 IS 均值提高了 0.30。说明 WordGAN 能够生成具有感知合理的图像，且质量优于先前的框架。

表 1　文本到图像生成框架的 IS 均值对比

Tab. 1　Comparison of the IS mean for text-to-image generation frameworks

框架	IS	
	CUB bird	Oxford-102
DTGAN[1]	4.88 ±0.03	3.77 ±0.06
DF-GAN[2]	4.86 ±0.04	3.71 ±0.06
StackGAN[15]	3.70 ±0.04	3.20 ±0.01
StackGAN++[15]	4.04 ±0.05	3.26 ±0.01
AttnGAN[16]	4.36 ±0.03	
MirrorGAN[17]	4.56 ±0.05	
ControlGAN[18]	4.58 ±0.09	
SDGAN[19]	4.67 ±0.09	
DM-GAN[20]	4.75 ±0.07	
WordGAN	**5.13 ±0.05**	**4.07 ±0.04**

WordGAN、StackGAN++、AttnGAN 和 DF-GAN 在 CUB bird 和 MSCOCO 数据集上的 FID 比较结果见表 2。显然，WordGAN 在两个数据集上的 FID 明显低于其他框架，表明 WordGAN 生成的数据分布更接近真实的数据分布。具体来说，在 MSCOCO 数据集上，相比于 DF-GAN，WordGAN 将 FID 降低了 9.24；在 CUB bird 数据集上，相比于 DF-GAN，WordGAN 将 FID 降低了 4.93。

表 2　文本到图像生成框架 FID 对比

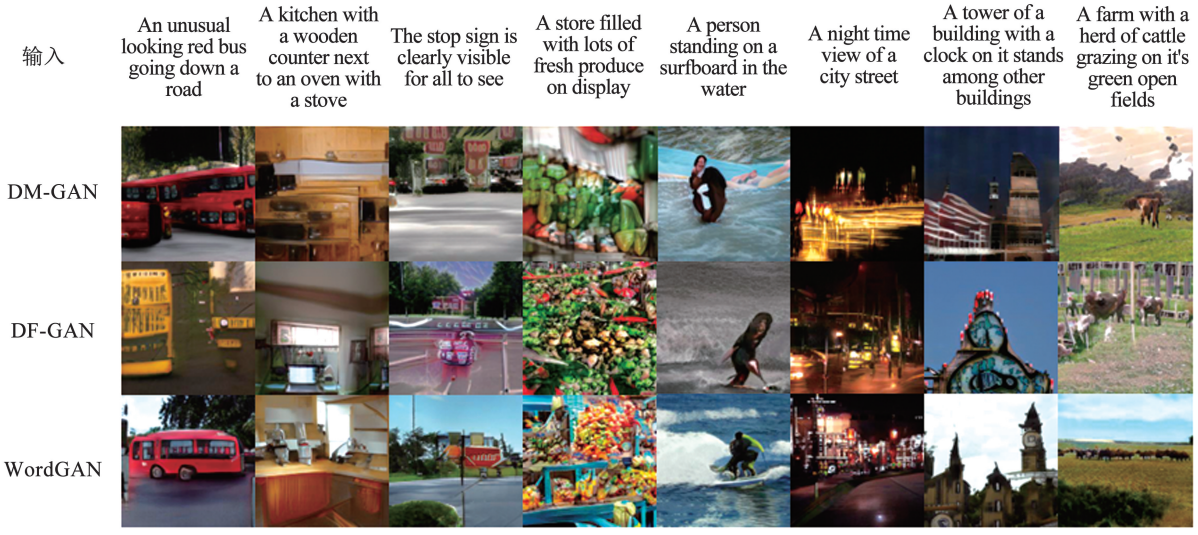
Tab. 2　FID comparison of text-to-image frameworks

框架	FID	
	CUB bird	MSCOCO
DF-GAN[2]	19.24	28.92
StackGAN++[15]	26.07	51.62
AttnGAN[16]	23.98	35.49
WordGAN	**14.31**	**19.68**

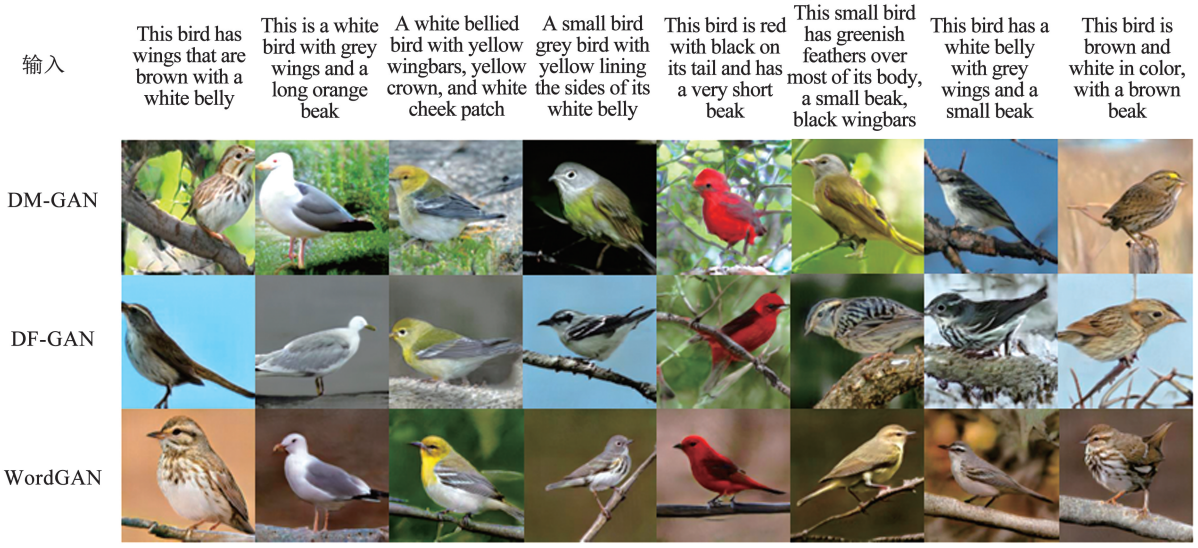
3.2.2　定性结果

DM-GAN、DF-GAN 和 WordGAN 在 MSCOCO 和 CUB bird 数据集上的定性比较如图 6 所示。由图 6 可以看出，WordGAN 能够根据文本描述生成高质量且语义一致的图像。例如，在生成复杂场景方面，WordGAN 在图 6（a）中生成的红色公交车比 DF-GAN 和 DM-GAN 具有更生动的细节（第 1 列），生成了一个具有合理木制柜台的厨房（第 2 列），路上清晰的红色标志（第 3 列），色彩丰富的新鲜蔬菜和水果（第 4 列），一个男人在逼真的海浪上冲浪（第 5 列）以及美丽的夜间街道（第 6 列）。而 DM-GAN 和 DF-GAN 生成的图像对象模糊不清（第 1～3、7、8 列），背景单调（第 4～6 列）。更重要的是，WordGAN 在第 7 列生成了一个令人印象深刻的欧洲钟楼。以上结果表明，具有双重残差结构的 WordGAN 能够捕捉文本描述中的关键词，并突出生成图像的主要对象，生成具有生动细节的高质量多目标场景。

由图 6（b）可以看出，虽然 DM-GAN、DF-GAN 和 WordGAN 都生成了颜色和形状一致的鸟图像，但 WordGAN 更加注重生成图像中与输入文本描述相关的部分，生成了高质量的鸟图像。此外，由 DM-GAN（第 1、2、5、7 列）和 DF-GAN（第 5、7、8 列）生成的一些鸟图像的背景是不合理的，表明通过词级注意力模块，WordGAN 能够缩小视觉特征图与词向量之间的语义差距，生成与文本描述语义一致的高质量图像。



(a) MSCOCO数据集



(b) CUB bird数据集

图 6 DM-GAN、DF-GAN 和 WordGAN 在 MSCOCO 和 CUB bird 数据集上定性结果比较

Fig. 6 Qualitative comparison results of DM-GAN, DF-GAN and WordGAN on the MSCOCO and CUB bird datasets

在 Oxford-102 数据集上，DF-GAN、DTGAN 和 WordGAN 之间的定性比较结果如图 7 所示。由图 7 可以看出，WordGAN 生成的花朵图像细节更生动、颜色更丰富、形状更清晰，相较于 DF-GAN 和 DTGAN 生成的图像更为合理，这证实了 WordGAN 的有效性。近两年，出现了一系列以扩散模型为主的文生图模型，但需要大量的文本-图像对训练，并且需要强大的 GPU 作为算力支撑，不在本文的探讨范围之内。

3.2.3 人类评估

在 CUB bird 和 MSCOCO 数据集上进行人类评估，以测试 DM-GAN[22]、DF-GAN[23] 和 WordGAN 的图像质量和语义一致性。分别从两个数据集中随机选择 100 张图像，在给定相同文本描述的情况下，要求用户根据图像细节和相应的文本描述，从 3 种框架生成的图像中选择最佳图像。此外，为了公平起见，最终得分由 4 名评委计算得出，并由 1 名人员进行监督。在评估过程中，采用双盲评估机制，即评估者不知道图像来源，图像提供者也不知道评估者身份。人类评估结果如图 8 所示。由图 8 可以看出，在两个数据集上，WordGAN 明显优于 DM-GAN 和 DF-GAN，尤其是在 MSCOCO 数据集上，表明了 WordGAN 的有效性。

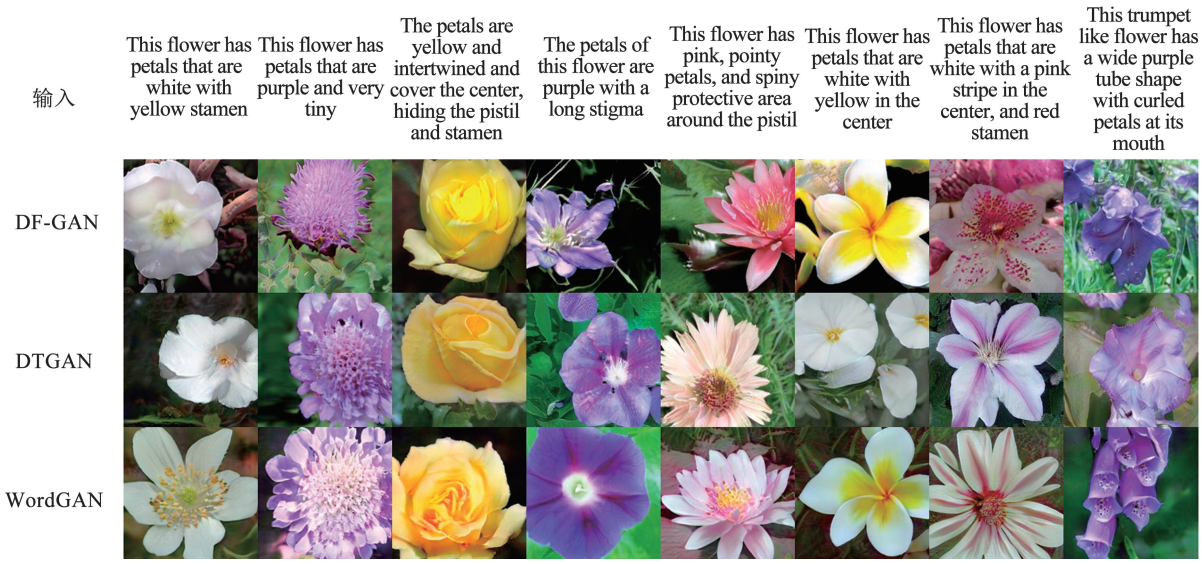


图 7　DF-GAN、DTGAN 和 WordGAN 在 Oxford-102 数据集上的定性结果比较

Fig. 7　Qualitative comparison results of DF-GAN, DTGAN and WordGAN on the Oxford-102 dataset

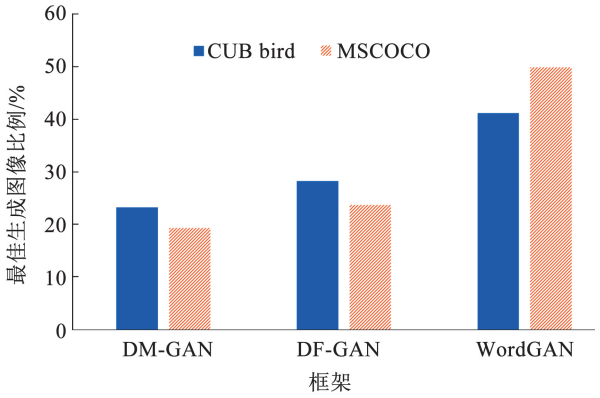


图 8　人类评估结果

Fig. 8　Human evaluation results

3.3　FC 在图像多样性中的有效性

为验证 FC 在提升图像多样性方面的有效性，在 CUB bird、Oxford-102 和 MSCOCO 数据集上进行定量评估，并在 CUB bird 数据集上进行定性比较。

为了更好地表示比较的框架，本文约定 F1 表示 StackGAN++，F2 表示模式寻找生成对抗网络（mode seeking generative adversarial network，MSGAN）[25]，F3 表示 DF-GAN，F4 表示 DTGAN，F5 表示去除 FC 的 WordGAN，F6 表示引入条件增强（conditional augmentation，CA）技术的 F5 框架，F7 表示引入模式寻找（mode seeking，MS）技术的 F5 框架，F8 和 F9 分别表示在 F5 的第 1 个双重残差块之后分别插入 1 个和 2 个 FC 的框架，F10 表示 WordGAN。

根据不同目标设计了 3 组定量比较。第 1 组用于评估单阶段文本到图像框架生成图像的质量和多样性，具体包括 DF-GAN 和 DTGAN，引入 CA 的 StackGAN++[15] 和引入 MS 的 MSGAN。第 2 组用于比较提出的插入 FC 方法与 CA 和 MS。为公平比较，将去除 FC 的 WordGAN 作为基准框架（F5），F6（F5 + CA）和 F7（F5 + MS）分别将 CA 和 MS 添加至 F5 中进行比较。第 3 组用于验证在 F5 中插入 FC 的不同方式的有效性，具体包括 F8（F5 + FC(1)）、F9（F5 + FC(1 + 1)）和 F10（WordGAN）。

定量结果见表 3。在 CUB bird 数据集上，通过比较 F10 与 F1 ~ F4 可以看出，WordGAN 将 LPIPS 从 0.544 提高至 0.682，证实了 WordGAN 在多样性方面的有效性。通过比较 F7 与 F5 可以看出，F7 在 CUB bird 数据集上将 LPIPS 提高了 0.118，但其在 CUB bird 和 Oxford-102 数据集上的 IS 均值分别降低了 0.34 和 0.05，在 CUB bird 数据集上的 FID 增加了 2.06，在 MSCOCO 数据集上的 FID 增加了 1.91。说明虽然引入 MS 可以提高图像的多样性，但可能会降低图像质量。由表 3 还可以看出，F5 的性能优于 F6，说明 CA 对当前单阶段文本到图像方法而言，在提升图像多样性方面可能不太有效。F8 ~ F10 均比 F5 具有更大的 LPIPS，证明了 FC 在提高多样性方面的有效性。此外，F10 的表现优于 F8 和 F9，在 CUB bird 数据集上将 LPIPS 分别提高了 0.103 和 0.027，将 IS 均值分别提高了 0.37 和 0.40，将 FID 分别降低了 4.03 和 4.35；在 Oxford-102 数据集上将 IS 均值分别提高了 0.20 和 0.22，在 MSCOCO 数据集上将 FID 分别降低了 2.48 和 2.87，说明在 F5 的第 2 个双重残差块后插入一个 FC 可以获得最佳的图像质量和多样性表现。综上所述，仅使用一组生成器/对抗器，WordGAN 能根据单个文本描述生成高质量和多样性的图像。

表 3 FC 对图像多样性的有效性结果比较

Tab. 3 Effectiveness comparison of the FC on image diversity

序号	框架	IS		FID		LPIPS
		CUB bird	Oxford-102	CUB bird	MSCOCO	CUB bird
F4	DTGAN[1]	4.88 ±0.03	3.77 ±0.06	16.35	23.61	0.544
F3	DF-GAN[2]	4.86 ±0.04	3.71 ±0.06	19.24	28.92	0.537
F1	StackGAN++[15]	4.04 ±0.05	3.26 ±0.01	26.07	51.62	0.362
F2	MSGAN[25]			25.53		0.373
F5	Baseline	4.91 ±0.06	3.87 ±0.07	16.42	22.53	0.549
F6	F5 + CA	4.72 ±0.06	3.82 ±0.06	17.61	25.23	0.535
F7	F5 + MS	4.57 ±0.05	3.82 ±0.05	18.48	24.44	0.667
F8	F5 + FC(1)	4.76 ±0.05	3.87 ±0.07	18.34	22.16	0.579
F9	F5 + FC(1 + 1)	4.73 ±0.06	3.85 ±0.06	18.68	22.55	0.655
F10	WordGAN	**5.13 ±0.05**	**4.07 ±0.04**	**14.31**	**19.68**	**0.682**

DF-GAN、DTGAN、DF-GAN + FC(2)、DTGAN + FC(2)、F5、F6(F5 + CA)、F7(F5 + MS)、F8(F5 + FC(1))、F9(F5 + FC(1 + 1))和 F10(WordGAN)之间的定性比较结果如图 9 所示。DF-GAN + FC(2)和 DTGAN + FC(2)分别指的是在第 2 个残差块后插入 FC 的 DF-GAN 和 DTGAN。

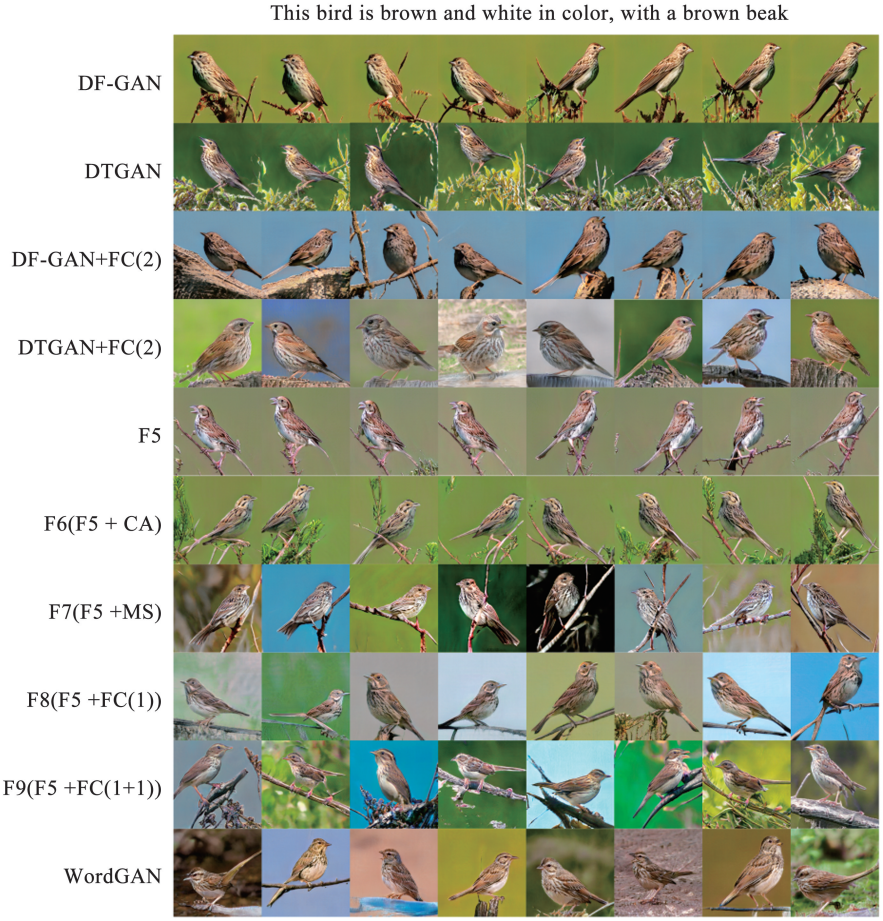


图 9 WordGAN 与其他框架的定性比较结果

Fig. 9 Qualitative compared results of WordGAN and other frameworks

由图 9 可以看出,尽管 DF-GAN(第 1 行)和 DTGAN(第 2 行)在图像质量上表现良好,但生成的鸟的形状相似,背景颜色也相同。然而,在将 FC 插入到框架之后,DF-GAN + FC(2)(第 3 行)和 DTGAN + FC(2)(第 4 行)均生成了多样化的鸟图像(例如,具有不同的形状、方向甚至背景颜色),比

原有框架具有更广泛的应用性，证明所提框架的泛化能力。DF-GAN + FC(2)背景相似的原因可能是调制模块的引入，DF-GAN 中的背景颜色仅取决于文本描述。将 WordGAN(第 10 行)与 F6(F5 + CA)和 F7(F5 + MS)进行比较可以看出，FC 有助于生成具有生动细节的多样化鸟图像，而 CA 并不能提高鸟图像的多样性，MS 可能会影响图像质量。例如在第 10 行，鸟位于树枝或地面上，鸟的方向、背景颜色以及脚的视觉外观均不同。然而，第 6 行的鸟图像仍然具有相似的形状和相同的背景颜色，第7 行的鸟图像有些模糊。与 F5 对比可以发现，F8(F5 + FC(1))、F9(F5 + FC(1 + 1))和 WordGAN 均能生成多样化的图像，进一步证实了 FC 在提升图像多样性方面的有效性。此外，WordGAN 生成的鸟图像比 F6 和 F7 的质量更高，也验证了 WordGAN 的优越性。

为验证 WordGAN 的敏感性，通过修改给定文本描述中的一个词来生成鸟图像。生成图像如图 10所示。由图 10 可以看出，当改变自然语言描述中的颜色属性时，WordGAN 根据修改后的文本进一步生成语义一致的鸟图像，同时保留与未修改部分相关的视觉内容(如形状、位置和纹理)。此外，通过改变潜在向量，WordGAN 生成了具有不同脚、背景颜色、方向和形状的鸟。因此，通过词级注意力模块和插入 FC 方法，WordGAN 能够有效地解耦输入文本描述的属性，准确地控制图像的区域，同时不会损害图像多样性。



图 10　WordGAN 通过改变颜色属性和随机潜在向量生成的图像

Fig. 10　Generated samples of WordGAN by changing the color attribute and the random latent codes

为了更好地理解 WordGAN 如何利用潜在向量实现多样性，在两个随机潜在向量之间进行线性插值，并生成相应的图像。WordGAN 在 CUB bird 数据集上的插值结果如图 11 所示。由图 11 可以看出，脚和图像背景(第 1、4 行)、鸟的形状和位置(第 2 行)、鸟的方向和形状以及脚(第 3 行)均随着潜在向量的变化而逐渐改变。



图 11　WordGAN 两个随机潜在向量线性插值结果

Fig. 11　Linear interpolation results between two random latent vectors in WordGAN

WordGAN 在 MSCOCO 数据集上的插值结果如图 12 所示。由图 12 可以看出，建筑外观(第 1 行)、海滩(第 2 行)、人物(第 2、3 行)、波浪(第 3 行)、草地和动物(第 4 行)均随着潜在向量的变化而连续变化，以上结果显示了前景对象和图像背景的平滑过渡。更重要的是，在进行潜在向量线性插值时，当基于第 1 个和最后 1 个潜在向量得到的图像是高质量的，WordGAN 很可能会产生一系列高质量的图像。



图 12 WordGAN 在 MSCOCO 数据集上两个随机潜在向量线性插值结果

Fig. 12 Linear interpolation results between two random latent vectors of WordGAN on the MSCOCO dataset

4 结 论

1)设计了一个基于词级注意力的多样化文本到图像生成对抗网络框架(WordGAN),该框架通过一组生成器/对抗器能够根据文本描述生成语义一致且高质量的多样化图像。

2)提出了两种新的词级注意力模块:通道注意力模块和像素注意力模块。两种模块通过引导生成器关注文本描述中的关键语义信息,有效提升了生成图像的语义一致性。

3)在网络框架中引入全连接层来增强特征表达能力,该设计有效解决了单阶段框架普遍存在的生成图像多样性不足的问题。通过优化特征空间的分布,使模型能够生成更具多样性的图像。

4)在多个标准数据集上进行实验,结果表明,WordGAN 在图像质量和多样性方面均优于当前框架。尤为突出的是,WordGAN 有效解决了传统文本到图像生成框架中的缺陷,为开发新一代文本到图像生成框架提供了重要理论支撑和实践依据。

参考文献

[1]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

[2]Tao Ming, Tang Hao, Wu Songsong, et al. DF-GAN: a simple and effective baseline for text-to-image synthesis[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Orleans: IEEE, 2022: 16494. DOI: 10.1109/CVPR52688.2022.01602

[3]Che Tong, Li Yanran, Jacob A P, et al. Mode regularized generative adversarial networks[C]//5th International Conference on Learning Representations. Toulon: ICLR, 2017: 3120

[4]胡智超, 余翔湛, 刘立坤, 等. 基于上下文生成对抗网络的时间序列异常检测方法[J]. 哈尔滨工业大学学报, 2024, 56(5): 1
Hu Zhichao, Yu Xiangzhan, Liu Likun, et al. A time series anomaly detection method based on contextual generative adversarial network [J]. Journal of Harbin Institute of Technology, 2024, 56(5): 1. DOI:10.11918/202212029

[5]孙澄, 丛欣宇, 韩昀松. 基于 CGAN 的居住区强排方案生成设计方法[J]. 哈尔滨工业大学学报, 2021, 53(2): 111
Sun Cheng, Cong Xinyu, Han Yunsong. Generative design method of forced layout in residential area based on CGAN[J]. Journal of Harbin Institute of Technology, 2021, 53(2): 111. DOI: 10.11918/201912143

[6]Hua Xia, Rui Ting, Cai Xia, et al. Collaborative generative adversarial network with visual perception and memory reasoning[J]. Neurocomputing, 2020, 414: 101. DOI: 10.1016/j.neucom.2020.06.037

[7]许小伟, 韦道明, 严运兵, 等. 基于改进 ACGAN 的永磁同步电机数据扩张方法[J]. 哈尔滨工业大学学报, 2023, 55(10): 114
Xu Xiaowei, Wei Daoming, Yan Yunbin, et al. Data expansion method of permanent magnet synchronous motor based on improved ACGAN[J]. Journal of Harbin Institute of Technology, 2023, 55(10): 114. DOI:10.11918/202203045

[8]Li Wei, Fan Li, Wang Zhenyu, et al. Tackling mode collapse in multi-generator GANs with orthogonal vectors [J]. Pattern Recognition, 2021, 110: 107. DOI: 10.1016/j.patcog.2020.107646

[9]Xu Jiazheng, Liu Xiao, Wu Yuchen, et al. Imagereward: learning and evaluating human preferences for text-to-image generation[C]//37th Conference on Neural Information Processing Systems. New Orleans: Neural Information Processing Systems Foundation, 2023: 15903

[10]Hao Yaru, Chi Zewen, Dong Li, et al. Optimizing prompts for text-to-image generation[C]//Advances in Neural Information Processing Systems. New Orleans: Neural Information Processing Systems Foundation, 2023: 13215

[11]Hodosh M, Young P, Hockenmaier J. Framing image description as

a ranking task: data, models and evaluation metrics[J]. Journal of Artificial Intelligence Research, 2013, 47(4): 853. DOI: 10.1613/JAIR.3994

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

[13] Xia Weihao, Yang Yujiu, Xue Jinghao, et al. TediGAN: text-guided diverse face image generation and manipulation[C]//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 2256. DOI: 10.1109/CVPR46437.2021.00229

[14] 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

[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] 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

[17] 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

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

[19] 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

[20] 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

[21] Zhou Bolei, Lapedriza A, Xiao Jianxiong, et al. Learning deep features for scene recognition using places database[C]//28th Conference on Neural Information Processing Systems. Montreal: Neural Information Processing Systems Foundation, 2014: 487

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

[23] 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

[24] 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

[25] Qi Mao, Lee H Y, Tseng H Y, et al. Mode seeking generative adversarial networks for diverse image synthesis[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 1429. DOI: 10.1109/CVPR.2019.00152