软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2024,35(5):2150−2164 [doi: 10.13328/j.cnki.jos.007024]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于条件语义增强的文本到图像生成*

余 凯[1,2], 宾 燚[1], 郑自强[1], 杨 阳[1]

[1](电子科技大学 计算机科学与工程学院, 四川 成都 611731)

[2](电子科技大学 (深圳) 高等研究院, 广东 深圳 518110)

通信作者: 杨阳, E-mail: yang.yang@uestc.edu.cn

摘 要: 文本到图像生成取得了视觉上的优异效果, 但存在细节表达不足的问题. 于是提出基于条件语义增强的生成对抗模型 (conditional semantic augmentation generative adversarial network, CSA-GAN). 所提模型首先将文本进行编码, 使用条件语义增强对其进行处理. 之后, 提取生成器的中间特征进行上采样, 再通过两层 CNN 生成图像的掩码. 最后将文本编码送入两个感知器处理后和掩码进行融合, 充分融合图像空间特征和文本语义, 以提高细节表达. 为了验证所提模型的生成图像的质量, 在不同的数据集上进行定量分析、定性分析. 使用 IS (inception score)、FID (Frechet inception distance) 指标对图像清晰度, 多样性和图像的自然真实程度进行定量评估. 定性分析包括可视化生成的图像, 消融实验分析具体模块等. 结果表明: 所提模型均优于近年来同类最优工作. 这充分验证所提出的方法具有更优性能, 同时能够优化图像生成过程中一些主体特征细节的表达.

关键词: 文本到图像生成; 条件语义增强; 空间-语义融合

中图法分类号: TP391

中文引用格式: 余凯, 宾燚, 郑自强, 杨阳. 基于条件语义增强的文本到图像生成. 软件学报, 2024, 35(5): 2150–2164. http://www.jos.org.cn/1000-9825/7024.htm

英文引用格式: Yu K, Bin Y, Zheng ZQ, Yang Y. Text-to-image Generation with Conditional Semantic Augmentation. Ruan Jian Xue Bao/Journal of Software, 2024, 35(5): 2150–2164 (in Chinese). http://www.jos.org.cn/1000-9825/7024.htm

Text-to-image Generation with Conditional Semantic Augmentation

YU Kai[1,2], BIN Yi[1], ZHENG Zi-Qiang[1], YANG Yang[1]

[1](School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China)

[2](Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China, Shenzhen 518110, China)

Abstract: Text-to-image generation achieves excellent visual results but suffers from the problem of insufficient detail representation. This study proposes the conditional semantic augmentation generative adversarial network (CSA-GAN). The model first encodes the text and processes it using conditional semantic augmentation. It then extracts the intermediate features of the generator for up-sampling and generates the image mask through a two-layer convolutional neural network (CNN). Finally, the text coding is sent to two perceptrons for processing and fusing with the mask, so as to fully integrate the image spatial and text semantics features to improve the detail representation. In order to verify the quality of the generated images of this model, quantitative and qualitative analyses are conducted on different datasets. This study employs inception score (IS) and Frechet inception distance (FID) metrics to quantitatively evaluate the image clarity, diversity, and natural realism of the images. The qualitative analyses include the visualization of the generated images and the analysis of specific modules of the ablation experiment. The results show that the proposed model is superior to the state-of-the-art works in recent years. This fully verifies that the proposed method has better performance and can optimize the expression of main feature details

* 基金项目: 国家自然科学基金 (62102070, U20B2063, 62220106008); 四川省科技计划 (2023NSFSC1392)
余凯和宾燚为共同第一作者.
本文由“多模态协同感知与融合技术”专题特约编辑孙立峰教授、宋新航副研究员、蒋树强教授、王莉莉教授、申恒涛教授推荐.
收稿时间: 2023-04-09; 修改时间: 2023-06-08; 采用时间: 2023-08-23; jos 在线出版时间: 2023-09-11
CNKI 网络首发时间: 2023-11-23

in the image generation process.

Key words: text-to-image generation; conditional semantic augmentation (CSA); spatial-semantic fusion

随着社会进步和科技发展, 人工智能技术的研究和应用日益增加. 国务院于 2017 年发布的《新一代人工智能发展规划》明确指出, 人工智能的快速发展将对人类社会生活和全球格局产生深远而广泛的影响, 因此将其确定为国家发展战略之一, 以充分把握人工智能发展带来的重要战略机遇. 计算机视觉作为人工智能领域的重要的一个分支, 在推动人工智能发展过程中扮演着至关重要的角色. 计算机视觉作为当前的研究热点, 也受到了广泛关注. 通过推动计算机视觉技术的研究和应用, 国家可以提高自身的竞争力, 改善民众的生活质量, 促进社会进步和可持续发展. 如在国家安全与防务中发挥关键作用, 提升边境安全和恐怖主义威胁检测能力. 在经济发展中有广泛应用, 提高生产效率、产品质量, 促进数字经济发展. 在健康医疗领域帮助医生进行准确诊断和治疗. 在支持社会安全和公共服务, 如交通管理、城市规划和公共安全管理, 提高社会整体安全和应急响应能力. 此外, 社会科学研究机构腾讯研究院发布的《AI 生成内容发展报告 2020》中指出, 合成技术具有潜力在影视、娱乐、教育、艺术等多个领域引发变革性影响的重大潜力.

自 Goodfellow 等人[1]提出生成对抗网络后, 图像到图像生成[2,3]、图像风格迁移[4−7]、文本到图像生成[8−10]、图像修复[11]等图像合成技术得到了快速发展. 其中, 由于文本到图像生成拥有强大的合成能力和良好的交互性, 所以备受研究人员的关注, 从而涌现了许多优秀的工作. 但文本到图像生成仍然面临一些重要挑战, 例如, 如何准确地理解文本描述并生成与之匹配且细节清晰的图像仍然是一个难点. 近年来, 为了准确捕捉文本描述与图像内容之间的语义和空间联系, 一些模型[12−14]旨在建立文本语义和图像空间之间的联系. 然而, 当前文本到图像生成技术在生成图像得细节表达方面仍存在一些问题. 例如, 在某些细节场景中生成的图像可能显得和整体图像不和谐, 并且生成的对象缺乏必要的文本语义特征 (如图 1 所示, 红色框标出的是生成鸟爪和生成背景的交界处. 其他方法生成的图像存在目标主体与其他物体不匹配的问题, 如鸟爪细节表达不足、边缘模糊, 而本文方法能有效缓解该问题, 生成的鸟爪细节特征表达充分, 轮廓清晰).

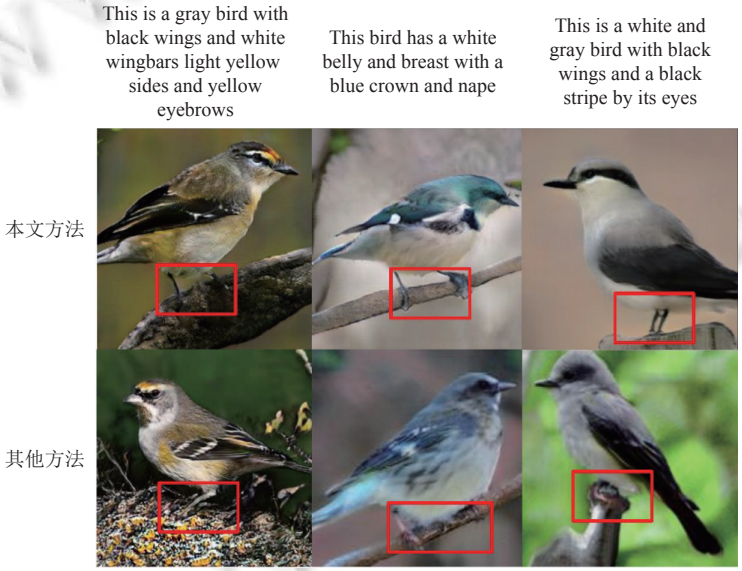


图 1 文本到图像生成示例图

为了解决上述问题, 本文提出了基于条件语义增强的生成方法, 在文本编码时, 进行了条件语义增强, 此外还设计了空间-语义融合模块以充分融合图像的空间特征和文本语义特征, 并设计使用残差生成器 G_0 对生成的细节进行补充. 实验结果表明在图像的质量和细节方面, 本文方法都表现出更优越的性能. 本文的主要贡献如下.

(1) 本文在文本编码后进行条件语义增强, 以在给定少量的文本图像数据对的情况下能提供更多的增强数据,

提高小扰动在语义空间的鲁棒性以达到更准确的图像生成效果并为后续的细节场景提供潜在信息.

(2) 本文通过对中间层特征进行上采样生成图像掩码, 并与增强后的条件语义进行空间-语义融合, 以更好地适配生成符合文本描述的内容, 从而生成具有更丰富、准确细节的图像.

(3) 针对细节表达问题, 本文使用残差生成器 G_0 对其中细节进行补充.

本文第 1 节介绍文本到图像生成相关方法和研究现状. 第 2 节介绍本文所需的基础知识, 包括生成对抗网络和深层注意力多模态相似模型损失. 第 3 节介绍本文构建的基于条件语义增强的文本到图像生成模型. 第 4 节通过对比实验验证了所提模型的有效性. 最后总结全文.

1 相关工作

自从 Reed 等人[15]通过条件生成对抗网络框架首次实现了文本到图像生成以来, 该方向得到了长足的进步. 以 StackGAN[16]为代表的堆叠式生成对抗网络在一定程度上解决了难以生成高分辨率图像的问题, 其第 1 阶段根据文字描述草绘出一张简单形状和基本颜色符合的低分辨率图像, 第 2 阶段把第 1 阶段的输出结果和文字条件作为输入, 生成具有照片真实感的高分辨率的图片. 但该类方法训练需要较大的计算资源, 同时生成图像的过程更加复杂抽象, 存在可解释性较差的问题. 以 DM-GAN[14]为代表的将生成对抗网络与动态记忆组件相结合, 以端到端生成高质量图像为目的, 解决以往方法中很大程度上依赖初始图像的质量的问题. 但该种方法生成过程涉及复杂的动态记忆机制, 导致生成图像的过程中仍存在可解释性问题, 同时训练过程相对复杂且需要消耗大量的计算资源. 以 textStyleGAN[17]为代表的适应无条件模型通过引入额外的风格向量, 可以根据输入风格向量来生成具有不同视觉特征和风格的图像, 这使得用户可以以更精细的方式控制生成图像的外观. 但生成的图像可能受到训练数据的限制, 并且在处理复杂的文本描述时仍然存在难以理解复杂文本和准确生成的困难. 为了使得生成模型具有良好的可解释性和生成图像语义更符合文本描述, 一些细粒度的语义编辑方法[18–20]相继提出. 注意力生成网络 (attentional generative adversarial networks, AttnGAN)[18]使用注意力机制使得本文嵌入的过程中能够实现更加细粒度文本潜入的效果. 该类方法主要使用了多模态相似模型损失函数, 能够利用句子和单词信息计算生成图像与输入文本之间的相似性, 即在神经网络生成全局语义向量的基础上通过加权使得网络关注图像每个区域下最相关的单词. 但模型性能在很大程度上过度依赖于训练数据的质量和多样性. 为了改进模型性能, Huang 等人[19]通过辅助边界框重新定义了目标网格 (object-grid), 并在目标网格和单词词组之间建立额外的关系, 从而将注意力扩展到了网格上. 上述方法虽然生成了较高质量的图像, 但在视觉属性和文本描述的匹配关系存上仍在些许瑕疵. 为了解决这一问题, 可控文本到图像生成 (controllable text-to-image generation, ControlGAN)[21]通过提出一个单词级别和通道注意力驱动生成器以生成文本描述对应的图像. 与之前论文的空间注意力相比, ControlGAN 更加关注颜色信息, 其提出的通道注意力将语义上有意义的区域与相应的单词关联, 同时完成了图像生成任务和属性 (类别, 纹理, 种类等) 编辑. 单词级别的鉴别器为生成器提供了训练信息, 并通过解耦单词与单词之间的相关性以提供生成相应区域的视觉信息. 传统的文本到图像生成方法往往会产生一些语义不连贯或不准确的图像, 如将一只鸟放在水中, 或将一只斑马放在天空中等. 为了解决语义的不连贯或不准确, 基于语义分解的生成对抗网络 (semantics disentangling generative adversarial network, SD-GAN)[22]定义了语义解耦的文本到图像生成方法, 试图将文本和图像中的语义信息分解和控制, 以生成更准确和多样化的图像. 模型充分有效地利用了输入文本的语义, 采用了孪生网络结构从语言描述中提取语义, 进行共享, 但其框架需要消耗大量的计算资源, 而且该模型的生成器是不止一个, 相互之间影响非常大. 虽然 SD-GAN 中提出了单词级别和句子级别条件批量归一化 (conditional batch normalization, CBN), 在图像特征图中注入文本信息, 在图像生成过程中, CBN 只应用了几次, 导致文本特征和图像特征没有充分融合, 生成的图像也会出现模糊, 语义不匹配的问题.

为了解决上述的问题, DF-GAN[12]将框架设计为一阶段的训练模式, 并结合铰链损失 (hinge loss) 和残差结构, 增加了模型训练的稳定性, 可以直接生成高分辨率的图像. 这样做, 一方面节省了大量的计算资源, 另一方面由于设计的网络框架中只有一个生成器, 避免了不同生成器潜在的耦合. 此外, DF-GAN 还设计了一个深度文本图像融

合块模块和目标感知鉴别器 (包括匹配-感知梯度惩罚和单向输出), 前者通过仿射变换将文本信息有效地融入到图像特征中. 后者用于增强语义一致性. 虽然, 这样仿射变换在图像的空间特征上同样有效. 但理想情况下, 文本信息应该只添加到与文本相关的子区域. 此外, 该模型只引入了句子级别的信息, 欠缺细粒度视觉特征合成的能力, 使得图像的细节部分出现不匹配或者细节场景生成质量不佳的情况. 为了解决该问题, Liao 等人[13]提出语义空间感知网络 (semantic-spatial aware GAN, SSA-GAN), 其核心模块为 SSCBN (语义-空间批量归一化)、一个残差块和一个掩码预测器. 其中语义-空间批量归一化主要基于提取的文本特征向量学习语义-空间感知仿射的参数, 之后通过当前的文本图像融合的结果预测掩码映射. 这种仿射变换融合了文本和图像特征, 并在图像特征在语义上与文本保持一致. 然而, 在细节上还是存在一些问题, 如生成的图像在某些细节场景上出现突兀和生成的主体缺少必要的特征. 如图 1 中, 其他方法生成的图像, 存在目标主体与其他物体不匹配、边缘模糊等问题.

与先前的工作[13,15,18]直接将文本进行编码嵌入得到特征表示不同, 本文将编码后的文本特征进行条件增强, 缓解了前者因为维度 (通常是大于 100 的) 过高造成文本语义空间稀疏的问题 (导致生成的图像在某些细节场景上出现突兀和生成的主体缺少必要的特征). 本文通过提供更多的增强训练数据, 提高小扰动在语义空间流形上的鲁棒性以达到更准确的图像生成效果, 并对其中细节进行补充. 另外, 本文提出一个新颖的残差生成器 G_0 对生成图像的细节部分进行补充从而获得更好的细节表征效果.

2 基础知识

本文所提方法主要基于生成对抗网络和深层注意力多模态相似模型损失, 下面就相关概念和基本知识予以介绍.

2.1 生成对抗网络

生成对抗网络 (generative adversarial network, GAN)[1]是由谷歌研究员 Goodfellow 等人于 2014 年基于零和博弈理论[23,24]提出的模型. 模型包括生成器 G 和鉴别器 D, 两者相互对抗训练, 其中生成器 G 生成看起来真实的图像以欺骗鉴别器 D, 而鉴别器 D 则用于判断图像是真实的还是生成的, 其训练架构如图 2 所示, 从高斯分布中随机采样一个噪声信号 z, 经过生成器生成相应的生成数据, 与真实数据同时输入到鉴别器中, 并由判别器对其进行真假判定. 该训练过程可以表示为:

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

其中, x 代表用于训练的真实图像, z 代表随机生成的噪声向量, $G(z)$ 代表模型生成的图像, $D(x)$ 代表对真实图像的鉴别器输出, $D(G(z))$ 代表对生成图像的鉴别器输出. 鉴别器旨在最大化目标函数, 而生成器旨在最小化目标函数. 两者相互对抗, 达到一个纳什均衡点, 从而实现对抗训练.

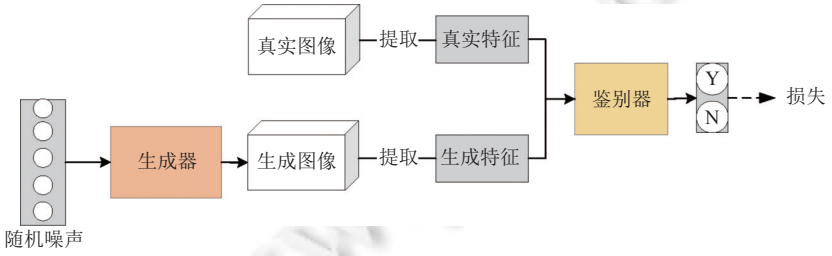


图 2 生成对抗网络框架

然而, 在 GAN 模型训练中, 模式崩溃是一个常见问题[25]. 这种现象可能出现在生成器生成的数据分布变化有限, 导致生成的图像出现重复的视觉特征. 不合理的生成器结构设计、或与生成器不匹配的鉴别器以及不平衡的训练过程都可能导致此现象. 此外, 由于生成器和鉴别器之间的博弈, 训练过程往往难以稳定. 可能会出现生成器过度拟合鉴别器, 或者鉴别器过于强大导致生成的图像缺乏视觉语义. 此外, 鉴别器对生成器的梯度过大或过小也

会导致训练不稳定. 为了解决这些问题, Arjovsky 等人[25]提出了使用 Wasserstein 距离作为生成器和鉴别器之间的度量函数. Wasserstein 距离定义如下:

$$W(P_r,P_g)=\inf_{\gamma\in\Pi(P_r,P_g)}\mathbb{E}_{(x,y)\sim\gamma}[|x-y|] \tag{2}$$

其中, P_r 和 P_g 分别代表真实数据分布和生成数据分布, $\Pi(P_r,P_g)$ 表示两个分布之间的转移概率矩阵, $(x,y)\sim\gamma$ 表示 γ 中的一个样本, $|x-y|$ 表示它们的距离. 相对而言, 使用 Wasserstein 避免了使用传统 JS 散度和 KL 散度存在的问题, 即两个数据分布较远时梯度为 0, 而且, Wasserstein 是一种计算真实数据分布与生成数据之间距离的一种方式, 具有更加平滑的梯度, 使模型具有更佳的收敛性.

在图像生成中, 使用 GAN 框架时同时为避免训练时出现问题, LSGAN[26]是一种较为通用的方法. LSGAN 采用最小二乘损失函数, 更容易收敛到最优解, 与传统 GAN 模型所采用的交叉熵损失函数相比, 避免了饱和效应和梯度消失的问题, 于是提高了模型的稳定性, 进而能够生成更加优异的图像. 同时, 最小二乘损失函数更具有解释性, 因为它是基于真实图像和生成图像之间计算的直接差异, 而不是计算两个分布之间的距离. 其数学表达式为:

$$\min_G\max_D V(D,G)=\frac{1}{2}\mathbb{E}_{x\sim p_{\text{data}}(x)}[(D(x)-1)^2]+\frac{1}{2}\mathbb{E}_{z\sim p_z(z)}[(D(G(z)))^2] \tag{3}$$

其中, $D(x)$ 和 $D(G(z))$ 分别代表对真实数据的鉴别值和生成数据的鉴别值. 这种损失函数的使用改善了生成器的稳定性, 提高了鉴别器的训练速度.

2.2 深层注意力多模态相似模型损失

深层注意力多模态相似模型 (deep attentional multi-modal similarity model, DAMSM)[13]损失是一种用于生成多模态结果的 GAN 损失函数. 它针对图像生成任务中的多模态性质进行了设计, 目的是提高生成的多样性. 该模型将图像子区域和文本单词映射到同一个语义空间, 以度量单词级别的文本图像相似度. 其中, 图像编码器是基于在 ImageNet 数据集上预训练的 Inception V3 模型进行构建的. 利用中间层学习局部特征 $f\in\mathbb{R}^{m\times n}$ (m 是局部特征向量的维数, n 是图像子区域数量) 和全局特征 $\bar{f}\in\mathbb{R}^{2048}$. 然后, 通过一个感知器层, 将图像语义特征转换为文本特征语义空间:

$$v=Wf,\quad \bar{v}=\overline{W}\bar{f} \tag{4}$$

其中, $v\in\mathbb{R}^{D\times 289}$, v_i 是图像对应第 i 个区域的视觉语义特征, $\bar{v}\in\mathbb{R}^D$ 是图像全局语义特征, D 是图像-文本特征空间维度. 文本编码器模型中, 单词特征 $e\in\mathbb{R}^{D\times T}$, 其中 e_i 表示第 i 个单词的特征向量, D 是单词向量的维度, T 是单词向量数量, 全局文本语义特征记为 $\bar{e}\in\mathbb{R}^D$. 首先计算图像子区域和单词之间的相似度:

$$s=e^{\mathrm{T}}v \tag{5}$$

其中, $s\in\mathbb{R}^{T\times 289}$, $s_{i,j}$ 是文本中第 i 个单词和图像中第 j 个子区域之间的内积相似性, 并进行归一化:

$$\bar{s}_{i,j}=\frac{\exp(s_{i,j})}{\sum_{k=0}^{T-1}\exp(s_{k,j})} \tag{6}$$

之后, 基于注意力机制建立一个模型, 查询每个单词区域上下文及其特征向量 c_i 是文本中第 i 个单词对应的图像子区域表示, 为所有视觉区域向量的加权和:

$$c_i=\sum_{j=0}^{n-1}\alpha_j v_j,\ \text{其中,}\ \alpha_j=\frac{\exp(\gamma_1\bar{s}_{i,j})}{\sum_{k=0}^{n-1}\exp(\gamma_1\bar{s}_{i,k})} \tag{7}$$

其中, γ_1 是在计算单词上下文时对相关区域语义特征关注程度. 最终, 用余弦相似性度量 c_i 与 e_i, 即 $R(c_i,e_i)=\frac{c_i^{\mathrm{T}}e_i}{\|c_i\|\|e_i\|}$, 描述整图区域与文本之间的匹配分数为:

$$R(Q,D)=\log\left(\sum_{i=1}^{T-1}\exp(\gamma_2 R(c_i,e_i))\right)^{\frac{1}{\gamma_2}} \tag{8}$$

其中, γ_2 决定相关单词与对应区域的重要程度. 特别地, 当 $\gamma_2 \to \infty$ 时, 原式便趋近于 $\max_{i=1}^{T-1} R(c_i, e_i)$. 基于半监督的学习注意力模型 DAMSM, 其中半监督体现在整图对整文本之间的匹配. 对于图像文本对 $\{(Q_i, D_i)\}_{i=1}^{M}$, 文本中的句子 D_i 和图像 Q_i 之间的后验概率:

$$P(D_i|Q_i) = \frac{\exp(\gamma_3 R(Q_i, D_i))}{\sum_{j=1}^{M} \exp(\gamma_3 R(Q_i, D_i))} \tag{9}$$

其中, 超参数 γ_3 是由实验决定的平滑因子. 在图像-文本对中, 只有 D_i 和 Q_i 配对, 将剩余下的均视为不匹配, 进而, 定义损失函数为:

$$\mathcal{L}_1^w = -\sum_{i=1}^{M} \log P(D_i|Q_i) \tag{10}$$

相对地,

$$\mathcal{L}_2^w = -\sum_{i=1}^{M} \log P(Q_i|D_i) \tag{11}$$

其中, s 为文本中语句中的单词. 重新定义公式 (8) 为 $R(Q,D) = \frac{\bar{v}^{\mathrm{T}}\bar{e}}{\|\bar{v}\|\|\bar{e}\|}$, 替换公式 (9)–公式 (11). 并且使用全局文本特征 $\bar{e}$ 和全局图像特征 $\bar{v}$, 得到句子的损失函数 $\mathcal{L}_1^s$ 和 $\mathcal{L}_2^s$. 那么就将得到 DAMSM 损失:

$$\mathcal{L}_{\mathrm{DAMSM}} = \mathcal{L}_1^w + \mathcal{L}_2^w + \mathcal{L}_1^s + \mathcal{L}_2^s \tag{12}$$

3 基于条件语义增强的文本到图像生成

文本到图像生成是基于生成对抗网络框架, 根据给定的文本描述自动生成相应的图像. 近年来, 一些同期工作 (如 DF-GAN[12]、SD-GAN[22]、SSA-GAN[13]、Text2Shape[27]) 都基于细粒度上做了一些改进, 取得了质量优秀的生成图像. 然而, 这些工作生成的图像在细节上还是存在一些问题, 如目标主体与其他物体不匹配、边缘模糊等. 为了研究这些问题, 本文将生成的掩码可视化后, 实验结果显示这些掩码存在不完整、不准确等问题, 甚至存在部分生成的问题, 经过实验分析, 本文推测可能的原因是由于语义样本稀疏造成的, 语义空间不连续, 从而导致生成的图像在某些细节场景上存在表现与背景信息不协调的问题. 如前文图 1 所示, 根据描述“This is a gray bird with black wings and white wingbars light yellow sides and yellow eyebrows”生成的图像中, 主体与其他物体轮廓的颜色与描述不匹配. 此外, 图 1 中第 3 列生成图像的主体缺少必要的特征, 如即其他方法生成鸟爪与其他物体兼容性差、边缘模糊等问题. 本文针对上述问题, 提出基于条件语义增强的生成对抗网络 (conditional semantic augmentation generative adversarial network, CSA-GAN). 在文本编码时, CSA-GAN 进行了条件语义增强, 以在给定少量的文本图像数据对的情况下能提供更多的增强数据, 从而使得语义空间连续, 提高了生成器在条件数据空间上对小扰动的鲁棒性. 此外, 本文在生成掩码的基础上设计了空间-语义融合模块, 以充分融合图像的空间特征和文本语义特征. 针对细节表达问题, 本文采用残差生成器 G_0 对其进行细节补充. 具体而言, 本文设计了一个残差结构, 将 G_0 和空间-语义融合的输出结果进行叠加, 自适应地将两部分输出融合从而得到细节更丰富、更准确的图像. 整个模型框架图如后文图 3 所示.

3.1 文本编码

本文首先将文本进行嵌入, 使用双向 LSTM 模型对其进行编码得到相应的特征表示. 双向 LSTM 模型由两个单向 LSTM 模型组成, 分别称为前向 LSTM 和后向 LSTM. 通过这两个模型的正向和反向处理, 可以更好地捕捉序列中的上下文信息. 在文本语义表示任务中, 双向 LSTM 输出按时间步进行拼接, 以得到更丰富的单词表示, 从而提高文本信息的表达能力. 双向 LSTM 将输入序列按时间步展开, 每个时间步将输入的单词转换为一个向量表示. 在正向 LSTM 中, 每个时间步的输入是当前单词的向量表示和前一时间步的隐状态, 在反向 LSTM 中, 每个时间步的输入是当前单词的向量表示和后一时间步的隐状态. 在每个时间步, LSTM 单元根据当前输入和前一时间步的状态计算新的隐状态和核状态. 最终, 得到编码后文本信息.

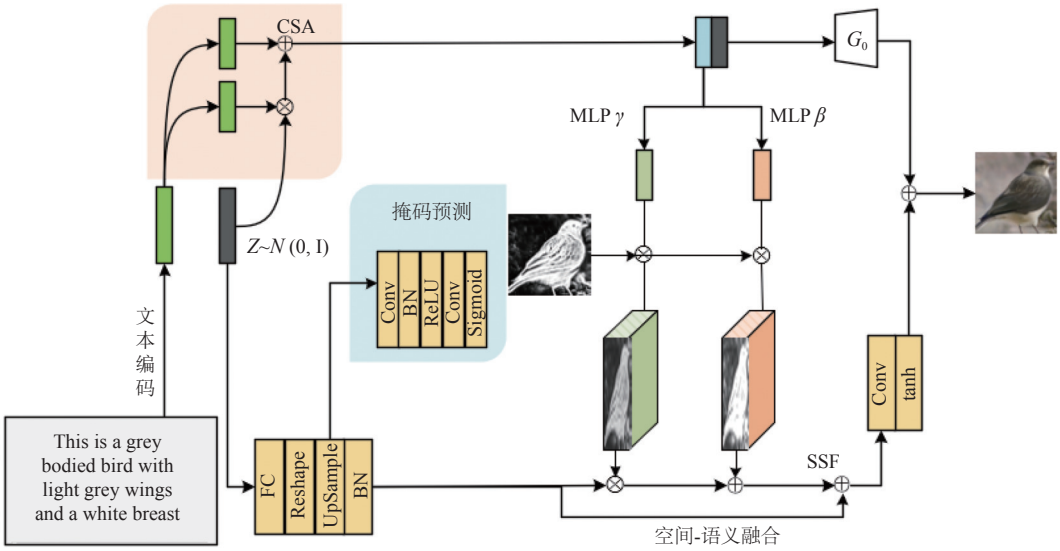


图 3 本文模型框架图

给定一个序列 $t=(t_1,t_2,\ldots,t_L)$, 双向 LSTM 网络最终得到 h_t, 其具体为:

$$h_t=\overrightarrow{h_t}\oplus\overleftarrow{h_t} \tag{13}$$

其中, $\overrightarrow{h_t}$ 和 $\overleftarrow{h_t}$ 分别表示从左往右和从右往左的 LSTM 计算得到的隐藏状态, $\oplus$ 表示连接操作. 由于正向 LSTM 和反向 LSTM 的计算方式一致, 故此处只介绍正向 LSTM. 正向 LSTM 得到隐藏状态序列 $(h_1,h_2,\ldots,h_T)$. 首先, 需要计算输入门的激活值 i_t、遗忘门的激活值 f_t、输出门的激活值 o_t 和细胞状态的更新值 $\widetilde{C}_t$, 计算方式如下:

$$\begin{cases} i_j=\sigma(Wt_j+U_ih_{j-1}+b_i)\\ f_j=\sigma(W_ft_t+U_fh_{j-1}+b_f)\\ o_j=\sigma(W_ot_j+U_oh_{j-1}+b_o)\\ \widetilde{C}_j=\tanh(W_ct_j+U_ch_{j-1}+b_c) \end{cases} \tag{14}$$

其中, t_j 是输入序列中的 j 个元素, h_{t-1} 是上一个时间步的隐状态, W_*,U_*,b_* 是可训练的参数, σ 是 Sigmoid 函数, tanh 是双曲正切函数. 接下来, 需要根据输入门、遗忘门和细胞状态的更新值, 计算新的细胞状态 C_j 和输出 h_j, 计算方式如下:

$$\begin{cases} C_j=f_j\odot C_{j-1}+i_j\odot\widetilde{C}_j\\ h_j=o_j\odot\tanh(C_j) \end{cases} \tag{15}$$

其中, $\odot$ 表示哈达玛积 (Hadamard product).

本文的文本编码器模型参数具体设置为: 隐藏状态特征数量为 256 (句子的文本编码维度数为 256), 单词的特征长度为 18, 参数 drop_prob 设置为 0.5. 最终输出的句子特征与单词特征, 分别记为 $\bar{e}\in\mathbb{R}^{256}$ 和 $e\in\mathbb{R}^{265\times18}$. e 的第 i 列 e_i 是第 i 个词的特征向量.

3.2 条件语义增强

在第 3 节提及到, 本文发现生成的中间掩码不完整, 甚至与文本描述不符合. 通过大量的研究发现, 文本嵌入是以非线性的形式转换为生成条件的潜在变量, 但嵌入特征表示是高维的 (通常其维度都是大于 100). 造成潜在的语义空间不连续, 从而生成器生成的图像效果并不理想. 受 StackGAN++[28]中文本特征多次采样的条件增强思想启发, 为了使得语义空间分布更加稠密, 本文提出了条件语义增强 (conditional semantic augmentation, CSA).

具体说来, 文本描述 t 首先经过文本编码, 得到文本稠密的特征表示 ϕ_t. 在传统的批量归一化的基础上引入了一个新的变量 $\hat{c}$ 作为条件约束. 一些工作, 诸如在文本引导图像修复[11,27]和文本到图像生成任务[16,20], 都采用了将

文本编码后直接固定特征, 并没有充分利用信息. 与这些工作处理文本编码的方式不同, 本文将文本嵌入得到相应的特征表示 ϕ_t 后进行随机采样得到 $\hat{c}$. 具体体现为从一个独立的高斯分布 $\mathcal{N}(\mu(\phi_t), \sum(\phi_t))$ 进行随机采样, 其中平均值 $\mu(\phi_t)$ 和对角协方差矩阵均是在文本嵌入表示 ϕ_t 上计算得到的.

上述的条件语义增强在给定少量的文本图像数据对的情况下能提供更多的增强数据, 从而提高生成器对于语义空间流形上小扰动的鲁棒性, 提高了模型的性能. 为了进一步增强语义空间流形的连续性, 同时为了避免模型过拟合的情况发生, 在目标函数中增加了一项在模型训练时针对生成器的正则化. 具体说来, 本文采用 KL 散度 (Kullback-Leibler divergence) 衡量采样分布和语义空间的相似性:

$$D_{KL}\left(\mathcal{N}(\mu(\phi_t), \sum(\phi_t)) \| \mathcal{N}(0, I)\right) \tag{16}$$

其中, $\mathcal{N}(0, I)$ 为采样分布, $\mathcal{N}(\mu(\phi_t), \sum(\phi_t))$ 为语义空间分布. 条件语义增强引入细微语义扰动, 从而学习更加稳健鲁棒, 语义更加丰富, 能够生成不同外观和姿态的图像, 而不是仅重复生成同一种图像.

3.3 掩码生成

为了更好地适配文本语义和生成图像的空间, 本文将条件语义增强后的文本和生成的掩码进行融合. 本模块将图像空间特征粗略生成轮廓, 即掩码, 用以指导后续语义理解和精细生成, 如语义颜色, 主体边缘分布等. 本文设计了一个简单的两层 CNN 结构 (主要包括了卷积层, 批量归一化层, 激活层等), 生成所需要的语义掩码. 具体说来, 获得中间语义特征后, 使用上采样, 得到与最终生成图像相同尺寸的掩码. 将粗略的掩码映射到图像特征空间得到掩码图 m_i, 其中 m_i 的每个元素 $m_{i,(h,w)}$ 的取值范围在 [0, 1] 之间. 掩码图是基于当前生成的语义特征图进行预测的, 因此直观地引导了当前图像特征图中哪些部分仍需要用文本信息加强. 掩码预测器与整个网络一起训练, 没有特定的损失函数来指导其学习过程, 也没有额外的掩码引导, 因此该模块并不需要额外的计算资源.

3.4 空间-语义融合

为了更加充分融合图像空间特征和文本语义, 提高模型的细节表达能力, 本文提出了空间-语义融合 (spatial-semantic fusion, SSF) 模块. 受 CBN 引入条件的思想启发, 本文将图像空间特征作为条件, 融合图像空间特征和文本语义. 下面将予以具体介绍.

批量归一化是一种常用的神经网络优化技术, 其作用是对神经网络的输入进行标准化, 以减少内部协变量移位 (internal covariate shift) 并加快网络训练. 在深度神经网络中, 不同层之间的输入分布随着网络的训练而不断变化, 导致神经网络学习速度变慢、难以收敛. 批量归一化 (batch normalization, BN) 可以通过在每个 mini-batch 的数据上进行标准化, 使得不同层之间的输入分布更加稳定, 从而加快神经网络的训练速度. 具体来说, 对于每个 mini-batch 的输入 $x = x_1, x_2, \ldots, x_m$, BN 通过以下方式标准化:

$$\hat{x}_i = \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} \tag{17}$$

其中, μ_B 和 σ_B 分别是输入的均值和标准差, ϵ 是非常小的数. 接着, BN 对标准化后的输入进行线性变换:

$$\tilde{x}_i = \gamma \hat{x}_i + \beta \tag{18}$$

其中, γ 和 β 是可学习的缩放因子和平移因子, 分别用来恢复数据的原始分布. 在测试时, 这两个参数 γ 和 β 是固定的. 除了通过训练数据获得固定的 γ 和 β, Dumoulin 等人[20]提出了条件批量归一化: 在 BN 的基础上, 引入了条件信息, 以实现对不同类别或不同样式的数据进行更好的归一化处理. 在传统的 BN 中, 对于每个批次的数据, 都会将其归一化为均值为 0, 方差为 1 的标准正态分布. CBN 的实现方式可以是将条件向量 c 和 BN 的输出拼接在一起, 然后通过一个全连接层来学习条件对归一化的影响从而更加灵活地控制归一化的过程. CBN 的优点在于, 它可以更好地适应不同类别或不同样式的数据, 从而提高模型的泛化能力和鲁棒性. 具体说来, CBN 是通过学习模态参数得到 γ 和 β, 并根据给定条件进行自适应仿射变换. CBN 可以表示为:

$$\tilde{x}_i = \gamma_{\text{mlp}} \hat{x}_i + \beta_{\text{mlp}} \tag{19}$$

其中, γ_{mlp} 和 β_{mlp} 都是在文本特征上通过多层感知器学习而得到的:

$$\gamma_{\text{mlp}} = MLP_{\gamma}(\bar{e}),\ \beta_{\text{mlp}} = MLP_{\beta}(\bar{e}) \tag{20}$$

第 3.3 节介绍了本模型提取的预测掩码. 如果没有信息引导, 网络将无差别地对待所有像素, 从而在某些生成细节上效果不好. 于是本文引入了条件语义增强以生成更多的图像文本对:

$$\tilde{x}_i = m(x,y)(MLP_{\gamma}(\bar{e}) + MLP_{\beta}(\bar{e})),\quad (x,y) \in (h,w) \tag{21}$$

其中, $m(x,y)$ 表示预测掩码 (x,y) 对应的值. 这种方法决定了在哪些区域添加文本信息, 以及需要进行文本于语义信息加强的强度, 模型采样更多, 使得图像的文本语义空间数据流形更连续, 分布更加稠密. 从而生成的特征图和条件文本更加符合要求, 获得更好的生成图像效果. 针对细节表达不足的问题, 在最终生成时用了残差结构 (图 3 中 G_0), 以补充生成图像的细节. 具体说来, 在进行语义-空间融合后, 为了使模型更加注重语义细节的表达, 本文将条件语义增强后进行生成并与空间语义融合后的生成进行加权连接, 以使注重生成细节的表达.

3.5 目标函数

为了使得生成的图像与真实图像更加相似, 本文使用了匹配感知的零中心梯度惩罚 (matching-aware gradient penalty, MA-GP). 该方法旨在提高模型的稳定性和生成质量. 其主要惩罚生成器网络的梯度, 并使用了特殊的距离度量方式以度量真实图像和生成图像的特征分. 具体为:

$$GP_c = \mathbb{E}_{\hat{x}\sim P_{\hat{x}}}[(\|\nabla_{\hat{x}} D_c(\hat{x})\|_2 - 1)^2] \tag{22}$$

其中, D_c 是针对条件 c 的鉴别器, $\hat{x}$ 是条件 c 下的随机插值, $P_{\hat{x}}$ 是插值的概率分布, GP_c 表示对应于条件 c 的梯度惩罚. 通过对所有条件 c 的梯度惩罚求和, 得到 MA-GP:

$$\mathcal{L}_{\text{MA-GP}} = \sum_c w_c GP_c \tag{23}$$

其中, w_c 是条件 c 的权重, 根据条件的重要性进行调整.

此外, 模型将从生成图像和编码文本向量中提取的特征连接起来, 通过两个卷积层计算对抗损失, 鉴别器采用了单向模式[17], 将二者结合提高文本图像语义一致性. 对于鉴别器来说其损失函数如下:

$$\mathcal{L}_{\text{adv}}^{D} = \mathbb{E}_{x\sim p_{\text{data}}}[\max(0,1-D(x,t))] + 0.5\mathbb{E}_{x\sim p_G}[\max(0,1+D(\hat{x},t))] + 0.5\mathbb{E}_{x\sim p_{\text{data}}}[\max(0,1+D(x,\hat{t}))] + \lambda_1\mathcal{L}_{\text{MA-GP}} + \lambda_2 D_{KL} \tag{24}$$

其中, x 为真实图像, t 为其对应的文本描述, $\hat{t}$ 为对应描述之外的描述即不匹配的文本描述, $\hat{x}$ 为模型生成的图像, λ_1 和 λ_2 为需要学习的超参数, $D(\cdot,\cdot)$ 判定文本和图像是否匹配. 具体做法为, 将文本和图像特征映射到同一语义空间, 通过感知器得到最终输出.

为了使得模型生成的文本特征具有视觉辨析力, 本文使用了深层注意力多模态相似模型损失 $\mathcal{L}_{\text{DAMSM}}$. 则生成器的损失函数为:

$$\mathcal{L} = -\mathbb{E}_{x\sim p_G}[D(\hat{x},t)] + \lambda_3\mathcal{L}_{\text{DAMSM}} \tag{25}$$

其中, $\mathcal{L}_{\text{DAMSM}}$ 为公式 (12) 描述的损失函数 (其中 m 设置为 768, n 设置为 289), λ_3 为超参数.

4 实验分析

4.1 实验数据及相关参数介绍

本文的实验选择了 CUB birds 数据集和 COCO 数据集. CUB birds (lltech-UCSD Birds-200-2011) 数据集是专门为鸟类图像识别和检测研究而设计的数据集[29]. 由美国加州理工学院 (Caltech) 和加州大学圣地亚哥分校 (UCSD) 联合创建. 该数据集涵盖了 200 个不同鸟类的 11 788 张图像, 每张图像都有丰富的注释信息, 如鸟类名称、外貌特征和生态习性等. 该数据集适用于各种鸟类图像识别和检测任务, 如图像分类、图像检索、图像生成等. 其优势在于鸟类种类多样, 且每种鸟类的图像都有细致的注释信息, 这为研究者提供了更多的可能性和空间来探索和研究图像识别和检测任务. COCO 数据集中的每张图像都有对应的文本描述, 用于图像描述和文本到图像生成任务. 对于文本到图像生成任务, COCO 数据集提供了每张图像的 5 个不同描述. 因此, 可以根据这些描述生成不同的图像, 实现多样化的图像生成.

本文的实验基于 PyTorch 框架完成. 模型选择优化器为 Adam, 参数设置为 $(\beta_1,\beta_2)\to(0.00,0.09)$, 学习率设置为 $(lr_G,lr_D)\to(0.0002,0.0004)$, 其中, lr_G 和 lr_D 分别表示生成器和鉴别器的学习率. 超参数 $(\lambda_1,\lambda_2,\lambda_3)\to(2,0.01,0.05)$. 批量大小设置为 40, 在 4 块 NVIDIA RTX 3090 上进行试验. 同时, 针对两个不同的数据集, 本文设置了不同的训练批次, CUB birds 设置为 400, COCO 数据集设置为 200.

4.2 定量分析

对于实验分析的定量指标, 本文选择了 IS (inception score) 和 FID, 并选择 DF-GAN[12], AttnGAN[18], StackGAN++[28], SD-GAN[22], DM-GAN[14], ControlGAN[21], SSA-GAN[13]作为基准模型. IS 指标是基于在 ImageNet 数据集上预训练的 Inception 模型的输出概率分布, 用于衡量生成图像真实度, 而 FID 指标则是通过计算生成图像与真实图像在特征空间中的距离, 来评估生成图像与真实图像之间的差异性. FID 指标越小表示生成的图像质量越好. 相关工作[30]指出, FID 指标考虑了生成图像的多样性和生成质量, IS 则注重于生成质量. IS 计算如下:

$$Score_{\mathrm{IS}}=\exp(\mathbb{E}_{x\sim p_G}[D_{\mathrm{KL}}(p(y|x)\|p(y))]) \tag{26}$$

其中, x 代表生成模型生成的图像, y 代表图像 x 的类别标签, $p(y|x)$ 表示给定 x 时, 真实数据集上的类别分布, 通常使用预训练的分类器来估计, $p(y)$ 表示真实数据集上的类别分布, 通常直接从训练数据中统计计算, D_{KL} 为 KL 散度, 表示 x 生成的类别分布与真实数据集上的类别分布之间的差异. FID 计算如下:

$$Score_{\mathrm{FID}}=\|\mu_x-\mu_y\|^2+\mathrm{Tr}(\Sigma_x+\Sigma_y-2(\Sigma_x\Sigma_y)^{1/2})) \tag{27}$$

其中, μ_x 和 μ_y 分别表示真实数据集和生成数据集在 Inception V3 特征空间上的均值向量 Σ_x, Σ_y 分别表示真实数据集和生成数据集在 Inception V3 特征空间上的协方差矩阵. Tr 表示矩阵的迹.

对于 CUB birds 数据集在划分测试集时选取了 2 932 张图片以及对应的文本描述, 本节计算了本文模型与其他模型的 IS 和 FID 指标 (详情见表 1). 从表中可以看出, 本文的模型在 IS 指标上取得了 5.44, 在 FID 指标上取得了 17.04. 即本文提出的模型比基线模型生成质量更佳, 多样性更优秀.

同样地, 得益于 COCO 相对大量的数据, 在划分测试集时选取了 40 469 张图以及对应的文本描述. 本节也对比了本文模型与其他模型的 FID 指标 (具体数值可参见表 2). 根据先前工作[12,13]的实验分析, IS 分数并不能严格对应图像合成质量. 值得注意的是, 较高的 IS 分数通常并不意味着在某些包含许多物体概念的多样数据集 (例如 COCO 数据集) 上具有更好的图像合成质量. 本文将这归因于 IS 分数对物体概念表达能力较差. 因此, 本文不在 COCO 数据集上计算 IS 分数. 从表 2 可以看到, 本文模型在 FID 指标上仍然是最优秀的. 综合表 1 和表 2 来分析: 本文提出的模型在 COCO 数据集上和 CUB birds 数据集上得到的指标均优于其他基线模型/方法.

表 1 本文模型与同类工作在 CUB birds 数据集上定量指标对比

Methods	IS↑	FID↓
StackGAN++[28]	4.04	23.09
AttnGAN[18]	4.36	22.98
ControlGAN[21]	4.58	21.76
SD-GAN[22]	4.67	19.24
DM-GAN[14]	4.75	18.86
DF-GAN[12]	4.86	19.24
SSA-GAN[13]	5.17	17.59
CSA-GAN (本文方法)	**5.44**	**17.04**

表 2 本文模型与同类工作在 COCO 数据集上定量指标对比

Methods	FID↓
StackGAN++[28]	37.59
AttnGAN[18]	35.49
SD-GAN[22]	38.12
DM-GAN[14]	32.64
DF-GAN[12]	28.92
SSA-GAN[13]	19.37
CSA-GAN (本文方法)	**13.59**

4.3 定性分析

生成任务除了第 4.2 节进行的定量指标的计算外, 可视化分析也是生成类分析的一个重要方面. 为此, 本文选取了近年来的几个最优模型 (DF-GAN[12]、SSA-GAN[13]) 可视化结果后作为对比, 并分别在 COCO 数据集上和 CUB birds 数据集上训练这些模型和本文模型, 得到了图 4 和图 5 的可视化结果.

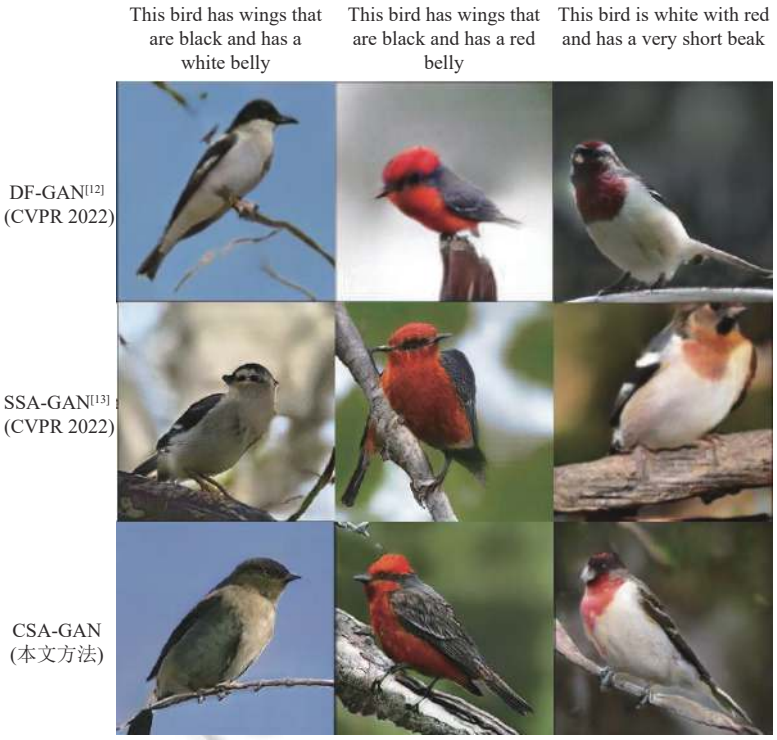


图 4　本文在 CUB birds 数据集生成结果与同类工作对比

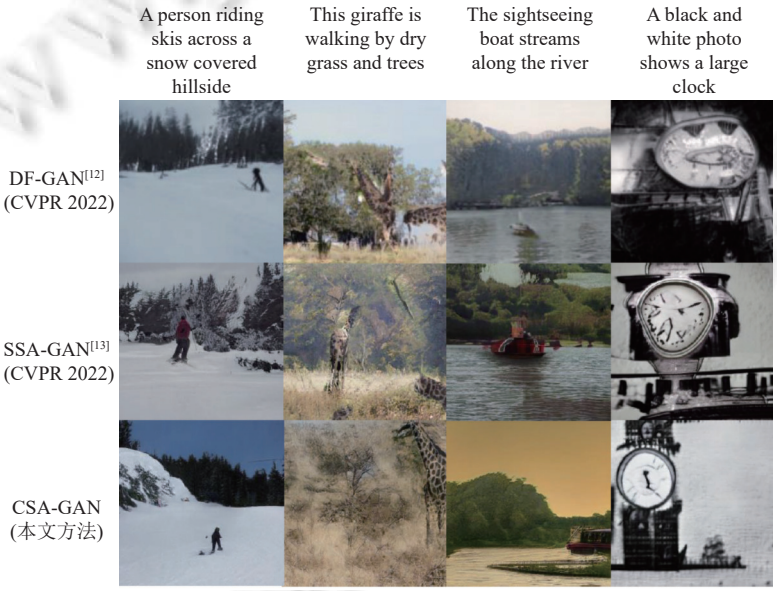


图 5　本文在 COCO 数据集生成结果与同类工作对比

本文模型是为了解决现有模型不能很好地处理前景和背景交界处的视觉上不协调的问题. 该问题是生成类任务一个普遍且常见的问题. 为了解决该问题, 本文模型引入了条件语义增强这一模块更好地适配文本编码特征和生成结果之间的关系. 本文模型所展示的可视化的生成图像也能证实这一点. 图 1 所示的根据文本生成相应图像中, 第 1 行为本文模型测试的结果, 第 2 行为其他模型的测试结果. 图像中用红色的框较为明显地指出了本文模型

在细节处的表现, 即在前景和背景的交界处细节体现更为清晰. 本文模型生成的鸟爪更具体, 而其他模型存在诸多问题. 如第 2 行第 1 列鸟爪附近区域显得很突兀, 还有难以描述的视觉特征. 第 2 行第 2 列图像中鸟爪和所站立的树枝颜色视觉特征几乎没有差别. 第 2 行第 3 列图像中鸟爪几乎没有鸟爪的特征. 这些说明其他模型在这些细节之处是存在问题的. 同样地, 在图 5 中本文模型生成的图像除了在颜色特征上表现优异外, 细节特征优势也是非常明显. 如第 4 列所示 DF-GAN 和 SSA-GAN 生成的钟表均具有较大的形变, 甚至视觉扭曲. 同样地, 相对而言第 2 列中 CSA-GAN 生成长颈鹿具有完整的主体特征.

相对 DF-GAN 和 SSA-GAN, 本文的模型产生的可视化图结果的细节表现得更好. 主要体现在主体和背景交融处显得更加自然和谐, 图 4 是在 CUB birds 上训练的模型, 图 4 中第 1 行是文本描述, 第 2 行、第 3 行分别是 DF-GAN 和 SSA-GAN 根据文本描述而生成的图像结果, 第 4 行为本文模型产生的结果. 文本的形式一般为“A/An/This + object + has/with + 颜色形容词 + 身体的具体某一部分”, 生成的图像是否符合要求的一个重要视觉特征就是符合对应的主体某一部分的颜色特征. 相对而言, 本文模型更加符合文本描述的特征, 例如第 3 列, 生成颜色为白色和红色有非常短的喙的鸟. SSA-GAN 生成的颜色更偏白色和橙色而不是白色和红色. 从颜色这一特征需求而言, 本文所提出的模型更加符合要求, 不难发现本文模型细节特征表现更优.

总体而言, 本节对本文所提出的模型和最近优秀模型在 COCO 和 CUB birds 数据集上的可视化结果进行了对比. 与 DF-GAN 和 SSA-GAN 相比, 本文提出的模型在细节表现方面更优秀, 更加符合文本描述的特征. 本文模型在设计时引入了条件语义增强这一模块更好地适配文本编码特征和生成结果之间的关系, 解决了前景和背景交界处视觉上不和谐问题. 通过生成的图像可以证实, 本文模型在细节处的表现更佳, 生成图像更加真实自然.

4.4 消融实验

为了充分探讨不同模块在本文模型中的贡献, 本文分别针对几个模块做了消融实验, 包括文本使用的 CSA, 上文介绍的 DAMSM 和 SSF. 首先是定量分析, CSA-GAN 是本文模型所提出的完整模型, CSA 是条件语义增强, DAMSM 是用来提高生成图像的质量和多样性的模块. 本文分别针对网络架构中的几个重要模块在 COCO 和 CUB birds 数据集上进行了实验. 具体结果如表 3 所示, 选取了与定量实验分析一样的指标. 从表 3 中数据分析, CSA 在 FID 和 IS 上有着贡献.

表 3 从定量的指标分析了不同的模块在模型中的贡献. 其中, 在 COCO 数据集上 FID 提升较为明显, 为了探索 CSA 为了模型提供的更多的特征信息, 同时为了验证 CSA 在给定少量的文本图像数据对的情况下能提供更多的训练数据对, 本文将生成的掩码进行了可视化, 同时使用了 CAM 算法可视化了更加关注的区域. 在 CUB birds 数据集和 COCO 数据集上可视化结果如图 6 所示, 从左到右依次为文本描述、生成的掩码图、注意力显著图、生成结果图, 从中可以看出本文模型可视化结果关注的更多的部分以区域的形式出现, 其关注的区域相比 SSA-GAN[18]更多, 此外, 生成的掩码的场景更加完善, 由此, 生成的效果更加优秀.

为了进一步验证 CSA 在模型中的作用, 以及 CSA 所对应的生成器 G_0 的作用. 本节设计将本文模型生成的掩码输入到同类工作模型 SSA-GAN, 并进行了定量指标的计算, 其结果如表 4 所示. 在使用本文模型生成的掩码输入到 SSA-GAN 得到的结果与本文差异异常小, 侧面印证了本文生成的中间掩码会正向引导最终的生成图像.

表 3 不同模块对模型的贡献

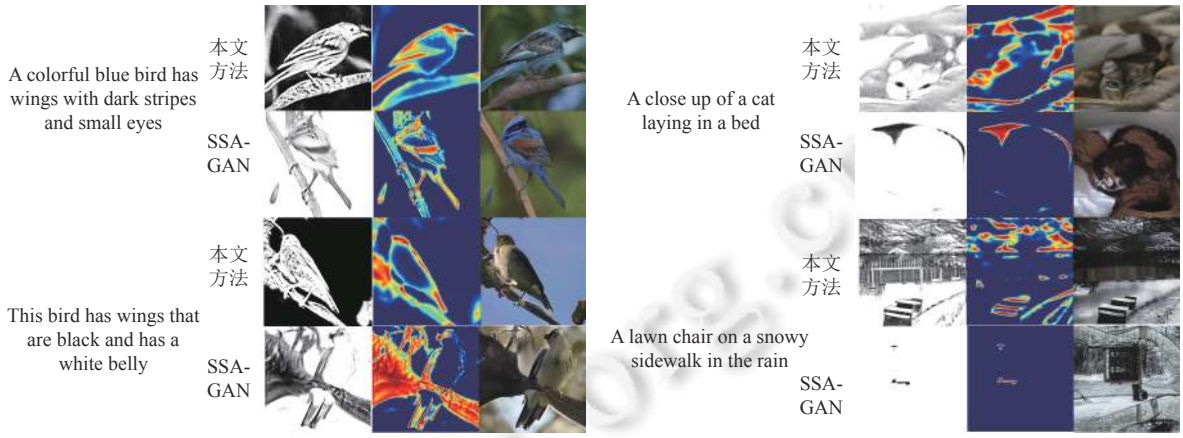
方法	CUB birds		COCO
	IS↑	FID↓	FID↓
CSA-GAN w/o SSF	4.21	17.72	16.26
CSA-GAN w/o DAMSM	4.97	18.61	14.79
CSA-GAN w/o CSA	5.07	18.02	19.65
CSA-GAN (本文方法)	**5.44**	**17.04**	**13.59**

表 4 针对 CSA 的消融实验

方法	CUB birds		COCO
	IS↑	FID↓	FID↓
SSA-GAN w/ fixed mask	5.37	17.48	13.64
SSA-GAN	5.17	17.59	19.37
CSA-GAN w/o G_0	5.19	17.32	13.62
CSA-GAN	**5.44**	**17.04**	**13.59**

从表 4 中可以得出: 本文所使用条件语义增强能够提供更多的信息特征, 生成图像的效果更好. 从表 4 中第 3, 4 行得到的数据表明, 生成 mask 接入到相关工作中, 生成的结果也比原模型的结果指标高很多, 其中 SSA-GAN

w/ fixed mask 代表将本文生成的掩码放入到 SSA-GAN 用于生成新的图像. 图 6 可以直观地观察到生成的掩码更加完整, 模型关注的区域更多, 说明条件语义空间分布更稠密, 数据流形更连续. 图 6 和表 4 证明, 本文模型 CSA-GAN 生成更完整的掩码, 引导后续语义的学习, 进行更精细的生成, 从而最终的生成图像质量更加优秀, 主体细节表达更充分.



(a) CUB birds数据集 (b) COCO数据集

图 6 本文模型生成掩码与 SSA-GAN 对比

5 总 结

为了解决文本到图像生成中细节特征表达不足的问题, 诸如生成的图像在某些细节场景上表现突兀, 本文提出了 CSA-GAN, 通过在文本编码基础上, 对其进行条件语义增强, 并结合残差结构的语义空间融合进行图像生成. 实验结果表明, 本文模型与近年来的几个最优模型相比, 在生成图像的质量和细节处理方面均表现最优. 在定量分析中本文模型的指标 FID 和 IS 均高于同类工作, 定性分析的细节表达也更充分. 此外, 本文还进行了消融实验, 从定量和定性两个方面对模块的有效性进行了论证. 其中定量方面主要计算了评价指标, 定性方面主要包括模型在细节特征 (主要体现在颜色方面) 表达和生成图像在主体特征上的优化 (对比了和同类工作的细节, 如鸟爪的特征). 结果表明: 本文的模型在这两个方面均表现更优, 生成的图像质量更高.

References:

[1] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. In: Proc. of the 27th Int'l Conf. on Neural Information Processing Systems. Montreal: ACM, 2014. 2672–2680. [doi: 10.5555/2969033.2969125]

[2] Chen XY, Xu C, Yang XK, Tao DC. Attention-GAN for object transfiguration in wild images. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 164–180. [doi: 10.1007/978-3-030-01216-8_11]

[3] Chen Z, Luo YD. Cycle-consistent diverse image synthesis from natural language. In: Proc. of the 2019 IEEE Int'l Conf. on Multimedia & Expo Workshops. Shanghai: IEEE, 2019. 459–464. [doi: 10.1109/ICMEW.2019.00085]

[4] Xu XZ, Chang JY, Ding SF. Image style transfering based on StarGAN and class encoder. Ruan Jian Xue Bao/Journal of Software, 2022, 33(4): 1516–1526 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6482.htm [doi: 10.13328/j.cnki.jos.006482]

[5] Xie B, Wang N, Fan YW. Correlation alignment total variation model and algorithm for style transfer. Journal of Image and Graphics, 2020, 25(2): 241–254 (in Chinese with English abstract). [doi: 10.11834/jig.190199]

[6] Cheng MM, Liu XC, Wang J, Lu SP, Lai YK, Rosin PL. Structure-preserving neural style transfer. IEEE Trans. on Image Processing, 2019, 29: 909–920. [doi: 10.1109/TIP.2019.2936746]

[7] Liu MY, Breuel T, Kautz J. Unsupervised image-to-image translation networks. In: Proc. of the 31st Conf. on Neural Information Processing Systems. Long Beach: NIPS, 2017. 700–708.

[8] Li ZL, Zhang SP, Liu Y, Zhang ZX, Zhang WG, Huang QM. Text-driven face image generation and manipulation via multi-level residual mapper. Ruan Jian Xue Bao/Journal of Software, 2023, 34(5): 2101–2115 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6767.htm [doi: 10.13328/j.cnki.jos.006767]

[9] Du PF, Li XY, Gao YL. Survey on multimodal visual language representation learning. Ruan Jian Xue Bao/Journal of Software, 2021, 32(2): 327–348 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6125.htm [doi: 10.13328/j.cnki.jos.006125]

[10] Wu FX, Cheng J. Configurable text-based image editing by autoencoder-based generative adversarial networks. Ruan Jian Xue Bao/Journal of Software, 2022, 33(9): 3139–3151 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6622.htm [doi: 10.13328/j.cnki.jos.006622]

[11] Yu JH, Lin Z, Yang JM, Shen XH, Lu X, Huang TS. Generative image inpainting with contextual attention. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 5505–5514. [doi: 10.1109/CVPR.2018.00577]

[12] Tao M, Tang H, Wu F, Jing XY, Bao BK, Xu CS. DF-GAN: A simple and effective baseline for text-to-image synthesis. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 16494–16504. [doi: 10.1109/CVPR52688.2022.01602]

[13] Liao WT, Hu K, Yang MY, Rosenhahn B. Text to image generation with semantic-spatial aware GAN. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 18166–18175. [doi: 10.1109/CVPR52688.2022.01765]

[14] Zhu MF, Pan PB, Chen W, Yang Y. DM-GAN: Dynamic memory generative adversarial networks for text-to-image synthesis. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 5795–5803. [doi: 10.1109/CVPR.2019.00595]

[15] Reed S, Akata Z, Yan XC, Logeswaran L, Schiele B, Lee H. Generative adversarial text to image synthesis. In: Proc. of the 33rd Int'l Conf. on Machine Learning. New York: JMLR.org, 2016. 1060–1069. [doi: 10.5555/3045390.3045503]

[16] Zhang H, Xu T, Li HS, Zhang ST, Wang XG, Huang XL, Metaxas D. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 5908–5916. [doi: 10.1109/ICCV.2017.629]

[17] Stap D, Bleeker M, Ibrahimi S, ter Hoeve M. Conditional image generation and manipulation for user-specified content. arXiv: 2005.04909, 2020.

[18] Xu T, Zhang PC, Huang QY, Zhang H, Gan Z, Huang XL, He XD. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 1316–1324. [doi: 10.1109/CVPR.2018.00143]

[19] Huang WM, da Xu RY, Oppermann I. Realistic image generation using region-phrase attention. In: Proc. of the 11th Asian Conf. on Machine Learning. Nagoya: PMLR, 2019. 284–299.

[20] Dumoulin V, Shlens J, Kudlur M. A learned representation for artistic style. In: Proc. of the 5th Int'l Conf. on Learning Representations. Toulon: OpenReview.net, 2017. 101–107.

[21] Li BW, Qi XJ, Lukasiewicz T, Torr PHS. Controllable text-to-image generation. In: Proc. of the 33rd Conf. on Neural Information Processing Systems. Vancouver: NeurIPS, 2019. 1–7.

[22] Yin GJ, Liu B, Sheng L, Yu NH, Wang XG, Shao J. Semantics disentangling for text-to-image generation. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 2322–2331. [doi: 10.1109/CVPR.2019.00243]

[23] Von Neumann J, Morgenstern O. Theory of Games and Economic Behavior. Princeton University Press, 1944. 20–60.

[24] Schelling TC. The Strategy of Conflict. Cambridge: Harvard University Press, 1960. 1–50.

[25] Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks. In: Proc. of the 34th Int'l Conf. on Machine Learning. Sydney: JMLR.org, 2017. 214–223. [doi: 10.5555/3305381.3305404]

[26] Mao XD, Li Q, Xie HR, Lau RYK, Wang Z, Smolley SP. Least squares generative adversarial networks. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 2813–2821. [doi: 10.1109/ICCV.2017.304]

[27] Chen K, Choy CB, Savva M, Chang AX, Funkhouser T, Savarese S. Text2Shape: Generating shapes from natural language by learning joint embeddings. In: Proc. of the 14th Asian Conf. on Computer Vision. Perth: Springer, 2019. 100–116. [doi: 10.1007/978-3-030-20893-6_7]

[28] Zhang H, Xu T, Li HS, Zhang ST, Wang XG, Huang XL, Metaxas DN. StackGAN++: Realistic image synthesis with stacked generative adversarial networks. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2019, 41(8): 1947–1962. [doi: 10.1109/TPAMI.2018.2856256]

[29] Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL. Microsoft COCO: Common objects in context. In:

Proc. of the 13th European Conf. on Computer Vision. Zurich: Springer, 2014. 740–755. [doi: 10.1007/978-3-319-10602-1_48]

[30] Frolov S, Hinz T, Raue F, Hees J, Dengel A. Adversarial text-to-image synthesis: A review. Neural Networks, 2021, 144: 187–209. [doi: 10.1016/j.neunet.2021.07.019]

附中文参考文献:

[4] 许新征, 常建英, 丁世飞. 基于 StarGAN 和类别编码器的图像风格转换. 软件学报, 2022, 33(4): 1516–1526. http://www.jos.org.cn/1000-9825/6482.htm [doi: 10.13328/j.cnki.jos.006482]

[5] 谢斌, 汪宁, 范有伟. 相关对齐的总变分风格迁移新模型. 中国图象图形学报, 2020, 25(2): 241–254. [doi: 10.11834/jig.190199]

[8] 李宗霖, 张盛平, 刘杨, 张兆心, 张维刚, 黄庆明. 基于多级残差映射器的文本驱动人脸图像生成和编辑. 软件学报, 2023, 34(5): 2101–2115. http://www.jos.org.cn/1000-9825/6767.htm [doi: 10.13328/j.cnki.jos.006767]

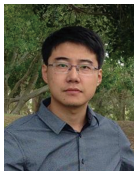
[9] 杜鹏飞, 李小勇, 高雅丽. 多模态视觉语言表征学习研究综述. 软件学报, 2021, 32(2): 327–348. http://www.jos.org.cn/1000-9825/6125.htm [doi: 10.13328/j.cnki.jos.006125]

[10] 吴福祥, 程俊. 基于自编码器生成对抗网络的可配置文本图像编辑. 软件学报, 2022, 33(9): 3139–3151. http://www.jos.org.cn/1000-9825/6622.htm [doi: 10.13328/j.cnki.jos.006622]

余凯(1995－), 男, 硕士生, 主要研究领域为生成对抗网络, 图像生成.

郑自强(1997－), 男, 硕士, 主要研究领域为对抗网络, 图像生成, 条件生成网络.

宾燚(1990－), 男, 博士, CCF 专业会员, 主要研究领域为深度学习, 计算机视觉, 非对称不成对图像的异步生成对抗性网络, 多媒体分析.



杨阳(1983－), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为多媒体分析, 机器学习, 人工智能, 大数据, 计算机视觉, 社交媒体分析.