生成对抗网络文字生成图像算法综述

邓 博,贺春林,徐黎明,宋兰玉
西华师范大学 计算机学院,四川 南充 637009

摘 要:生成对抗网络是图像合成的重要方法,也是目前实现文字生成图像任务最多的手段。随着跨模态生成研究不断地深入,文字生成图像的真实度与语义相关性得到了巨大提升,无论是生成花卉、鸟类、人脸等自然图像,还是生成场景图和布局,都取得了较好的成果。同时,文字生成图像技术也存在面临着一些挑战,如难以生成复杂场景中的多个物体,以及现有的评估指标不能准确地评估新提出的文字生成图像算法,需要提出新的算法评价指标。回顾了文字生成图像方法自提出以来的发展状况,列举了近年提出的文字生成图像算法、常用数据集和评估指标。最后从数据集、指标、算法和应用方面探讨了目前存在的问题,并展望了今后的研究方向。

关键词:图像合成;生成对抗网络;文字生成图像
文献标志码:A **中图分类号:**TP391.41 doi:10.3778/j.issn.1002-8331.2204-0441

Text-to-Image Synthesis: Survey of State-of-the-Art

DENG Bo, HE Chunlin, XU Liming, SONG Lanyu
School of Computer Science, China West Normal University, Nanchong, Sichuan 637009,China

Abstract: Generative adversarial network is an important method of image synthesis, and the most commonly used method for text to image synthesis. With the deepening of cross-modal generation research, the realism and semantic relevance of text to images have been greatly improved. Good results have been achieved in the synthesis of natural images such as flowers, birds and human faces, as well as in the synthesis of scene graph and layouts. Meanwhile, there are challenges: it is hard to generate multiple objects in a complex scene, and new methods of text to image synthesis cannot be accurately evaluated, new metrics need to be proposed. This paper reviews the development of state-of-the-art text to image methods, and lists methods, datasets and evaluation metrics proposed in recent years. Finally, the existing problems about dataset, metrics, method and application are discussed, and the future research direction is prospected.

Key words: image synthesis; generative adversarial networks; text to image

受益于深度学习的突飞猛进,图像处理技术与计算机视觉应用在近年来取得了较大的进步。图像合成是计算机视觉领域的重要研究主题,生成对抗网络[1](generative adversarial networks,GAN)在图像合成方面取得了重大成果,实现了通过无监督或者有监督的方式训练生成网络,合成高质量的图像。生成对抗网络将生成器合成的图像与来自真实数据集里的图像进行对抗,训练算法生成与真实数据相差无几的图像。条件生成对抗网络[2](conditional generative adversarial networks,cGAN)更进一步地引入条件信息来生成更高质量的指定图像。

文字生成图像(text to image,T2I),即通过技术手段生成能够正确反映文字描述内容的图像。Reed等[3]于2016年展开了第一个使用生成对抗网络进行的T2I任务。以条件生成对抗网络为基础进行扩展,以输入的描述文本为条件信息,在指定的数据集上生成了图像。此后的T2I研究通过改进文本编码,提出专门的损失函数与结构,以及开发新的定量评估指标来优化T2I算法,不断提出新的算法,使得生成图像的质量和语义相关性都取得了巨大提升,能在更复杂的数据集上验证有效性。至今生成对抗网络是T2I任务最常用的方法。

本文回顾了T2I方法自提出以来的发展状况,列举了近年来的T2I算法。首先回顾了T2I最为核心的条件生成对抗网络与文字编码。以是否使用额外监督信号

基金项目:西华师范大学创新团队基金(KCXTD2022-3);博士科研创新项目(21E025)。
作者简介:邓博(1996—),男,硕士研究生,主要研究方向为图像处理,E-mail:1242031892@qq.com;贺春林(1971—),男,硕士,教授,硕士生导师,主要研究方向为图像处理、机器学习;徐黎明(1991—),男,博士,讲师,CCF会员,主要研究方向为图像处理、机器学习;宋兰玉(1995—),女,硕士研究生,主要研究方向为图像处理。
收稿日期:2022-04-25 **修回日期:**2022-08-08 **文章编号:**1002-8331(2022)23-0042-14

为依据[4],将T2I划分为能使用描述文字直接生成图像的方法和使用额外监督作为信号的方法并分别列举,分析其具有的优势与存在的不足之处以及适用的场景。对T2I常用的数据集和评估指标进行说明,并讨论了如今T2I存在的问题,展望今后的研究方向。

1　生成对抗网络

基于深度学习的文本生成图像,可以追溯到使用长短时记忆网络[5](LSTM)以迭代的方式进行手写数字的生成。基于LSTM,DRAW[6]方法将变分自编码器[7](VAE)与空间注意机制相结合。align-DRAW[8]方法进一步将DRAW修改为使用基于自然语言的描述来合成具图像,使用注意模型来计算输入文字与迭代绘制的Patch之间的对齐方式。

生成对抗网络由生成器和判别器组成。生成网络以低维随机噪声作为输入,通过卷积生成图像。判别网络接受来自真实数据集的图像和生成图像进行二分类,生成对抗的过程将产生对抗损失用于生成器和判别器更新。

条件生成对抗网络[2]在生成器与判别器中增加了额外的条件信息,扩展了标准GAN的结构。条件信息可以包括标签等其他模态的数据,以此生成更具有指向性的图像。在T2I任务中,生成对抗网络通过编码器对文字描述进行特征提取并输入生成器,用生成器和判别器进行对抗训练,生成符合描述的高质量图像。

目前GAN是文字生成图像最常用的方法,相较于VAE具有独特的优势。首先GAN没用引入决定性偏置(deterministic bias),生成的图像会更加清晰;其次,GAN可以生成大量数据用于训练,对于训练数据的标签依赖程度很低;同时GAN还具有强大的表达能力,可以在向量空间中执行算数运算,将其转换为对应特征看见内的运算。

损失函数是GAN训练的重要部分,Wang等[9]对GAN的基本损失函数进行了详细的概述和总结。Reed等[3]提出了一种对抗性损失来区分真假图像-文本对,将具有随机描述的真实图像认定为假来鼓励匹配。堆叠结构的T2I算法被提出以后,GAN的损失函数也由整体的对抗损失转换为了由条件损失与非条件损失两部分构成。随着诸如注意力机制、循环一致性、双塔结构和记忆网络等多种方法被引入,新的损失函数也不断被提出,用于不同的T2I任务。

在T2I任务中非常重要的一部分就是对原始文字进行编码,从文本表示创建一个对条件变量有用的特征向量。Reed等[10]使用预训练的Char-CNN-RNN模型获得了文本描述的文字编码;TAC-GAN[11]则使用了跳跃思维向量(skip-thought vectors)来从图像标题中生成文字特征向量;Stack-GAN[12]使用条件增强(CA)来取代了预训练的文字编码器;Souza等提出的句子插值法(SI,2020)在训练中提供连续平滑的向量空间;Attn GAN[14]使用了Bi-LSTM[14]来进行文字编码,通过连接Bi-LSTM的隐藏状态形成每个单词的特征矩阵。每个单词的特征向量进行连接组成句子的全局特征向量;近年提出来的算法多使用BERT[15](如Devlin等[16]):一种预训练的语言表征模型。通过联合调节所有层的左右上下文,生成文本的字词的特征向量,将文本中各字词融合全文语义信息,输出向量表示。

2　文字生成图像算法

本章介绍现有的T2I算法,将生成算法划分为了直接生成算法和使用额外监督的生成算法[4]。这两种方法根据自身的特点与优势,适用于不同的场合。

2.1　直接生成算法

直接生成算法不需要额外的监督信号,对描述文字进行特征提取获得文字向量结合噪声输入生成器直接合成图像。

GAN-INT-CLS[3]是第一个将文本描述作为条件信息加入到图像生成中的应用研究,图1展示了GAN-CLS模型的基本结构,文字描述 t 通过文字编码器进行特征提取产生文字向量 $\varphi(t)$,结合噪声 z 作为生成器的输入,也作为分类条件输入判别器G。判别器G产生的文字-图像匹配损失 L_{match} 取代了原始GAN的对抗损失。GAN-INT-CLS由匹配感知判别器(GAN-CLS)和用多形插值学习(GAN-INT)两种方法构成。GAN-CLS用于判别图像是否按照了文本要求进行生成。多形插值学习(GAN-INT)在训练集标题的文本向量之间进行插值来生成大量额外的文本向量。这些插值的文本向量不需要对应于任何实际的人工书写的文本,因此没有额外的标签成本。

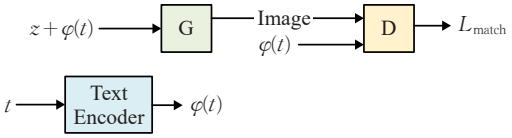

图1　GAN-INT-CLS结构

Fig.1　Structure of GAN-INT-CLS

GAN-INT-CLS生成图像分辨率为64×64,为了进一步提高分辨率,TAC-GAN[11]将AC-GAN[17]用于T2I任务,将类别标签和文字描述向量作为条件信息输入生成器,判别器不仅仅区分真实图像与合成图像,还给它们分配标签。生成图像分辨率提升到了128×128。

相较于多阶段生成判别的算法而言,单阶段生成对抗的算法的结构和训练过程都更加简易,可以更快速地得出结果,修改的难度也相对较低。但是在生成图像的分辨率、精细度和对语义描述反应的准确性方面存在差距。表1列举了部分直接生成的T2I算法。

表1　直接生成T2I算法

Table 1　Direct T2I methods

Input	method
直接生成算法	GAN-CLS[3](2016)
	TAC-GAN[11](2017)
	CVAEGAN[18](2018)
	Text2Scene[19](2018)
堆叠结构	Stack GAN[12](2017)
	Stack ++ GAN[20](2018)
	HD GAN[21](2018)
	Fused GAN[22](2018)
	PPAN[23](2019)
	Perception GAN[24](2020)
注意力机制	AttnGAN[14](2018)
	MANI GAN[25](2020)
	MGD-GAN[26](2021)
	DAE-GAN[27](2021)
	XMCGAN[28](2021)
	Maheshwari[29](2021)
记忆网络	DM-GAN[30](2019)
双塔结构	SD-GAN[31](2019)
	SEGAN[32](2019)
	TIMAM[33](2019)
循环一致性	Mirror GAN[34](2019)
	Chen[35](2019)
	Lao[36](2019)
	CI-GAN[37](2021)
	Das[38](2021)
无条件适应算法	TA-GAN[39](2018)
	Text-SeGAN[40](2018)
	Bridge-GAN[41](2019)
	HF GAN[42](2019)
	ControlGAN[43](2019)
	TextStyleGAN[44](2020)
	TVBI GAN[45](2020)
	TDA NET[46](2020)
	FA GAN[47](2021)
	DF GAN[48](2022)

堆叠结构以Stack GAN[12]为代表，相对于由单一生成器和判别器的网络结构，Stack GAN将生成网络与判别网络由单阶段划分为两阶段，同时使用条件增强(CA)处理文字编码，图2展示了Stack GAN的基本结构，在第一阶段，随机噪声与条件增强后的文字向量共同作为第一阶段生成器G0的输入，粗略描述64×64大小的低分辨率图像。在第二阶段，生成器G1接收上一阶段的生成图像不断对图像进行细化并提升分辨率，生成256×256的图像。Stack++GAN[17]进一步将三组生成器和判别器以树状结构排列，如图3。采用这一结构的GAN的损失函数也产生了相应的变化，由条件损失(对抗损失 L_{adv})和非条件损失(匹配损失 L_{match})两部分组成。

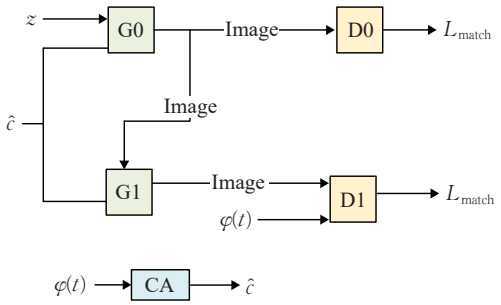

图2　Stack GAN结构

Fig.2　Structure of Stack GAN

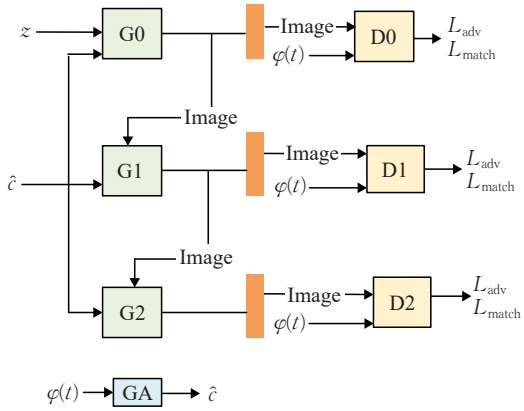

图3　Stack++ GAN结构

Fig.3　Structure of Stack++ GAN

采用堆叠结构，HDGAN[21]以学习从语义空间到图像空间的特征图为目标，用分层嵌套方法生成图像，并使用对目标对抗损失来有效利用图文信息；Perception GAN[24]在堆叠结构的初始图像生成阶段加入了感知理解，在判别器中加入良好的初始图像来改进感知信息。同时引入了第三阶段来改进细化过程；PPAN采用了类似的结构，由一个生成器和三个不同的判别器组成。采用金字塔框架[49-50]，通过横向连接自下而上的路径，将低分辨率但语义强的特征与高分辨率而语义弱的特征结合起来。结合特征感知损失与辅助分类损失完成文字生成图像。

多阶段生成的堆叠结构提升了生成图像的质量，不仅提高了分辨率，还能更加准确地反映文字描述所记载的细节特征，提高了图像与文字间的语义一致性。但这一结构对于初始阶段的生成图像具有较高的依赖性，如果这一阶段的生成结果不好，后续阶段的细化会变得困难，进而影响最终的生成图像质量。

注意力机制结合GAN的方法以AttnGAN为代表。通过关注描述中的相关词，合成图像不同子域的细节。如图4所示，初始阶段的噪声与文字增强向量作为第一阶段生成器的输入，通过卷积产生初始图像 $h0$ 作为下一阶段的初始信息，结合通过注意力模块的词向量 w 共同输入生成器。DAMSM模型计算局部图像与注意力模块和词特征 w 间的相似性损失 (L_{DAMSM}) 以训练生成器。

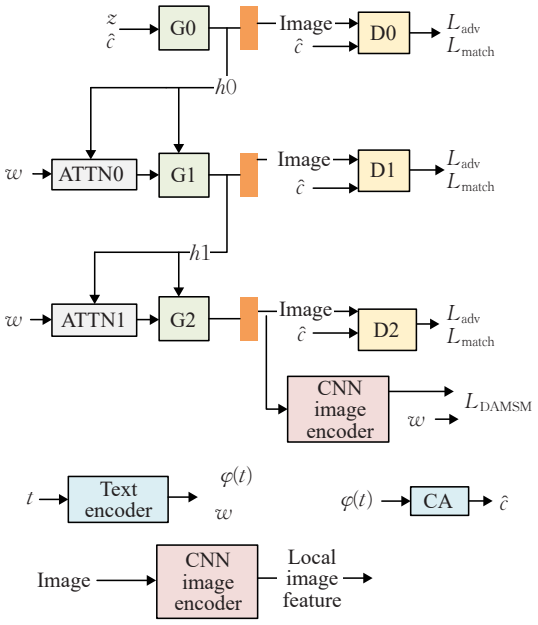

图4　ATTN GAN结构

Fig.4　Structure of ATTN GAN

结合注意力机制,MANIGAN[25]用以在语义层面对图像的一部分进行编辑,使得图像和给定的描述文本相匹配,同时能保留与文本无关的内容;MGD-GAN[26]利用human-part-based discriminator(HPD)和self-cross-attended(SCA)的全局判别器,实现用文字合成行人图像的功能;DAE-GAN[27]通过外表感知的动态重描绘方法来细化图像。对于文字描述中有关物体外表的信息加以利用(如:红眼睛)生成的图像;XMCGAN[28]通过最大化图像和文本之间的信息输出场景;Maheshwari等[29]提出了一种生成对抗网络来为二元分词生成颜色配置文件(color profiles)。将可见的属性和对象组合成不可见的结果(如"红色的"+"鸟"="红色的鸟"),实现以文搜图的跨模态检索。

采用注意力机制到T2I中明显地提升了GAN生成图像与描述文本的语义一致性。但在描述不同的图像内容时,每一个词的重要性并不一样。但是在细化图像的时候所使用的文字描述却是相同的,这并不利于生成图像与语义内容进一步提高关联程度。

采用记忆网络的DM-GAN[30]使用动态记忆模块去提炼图像,包括记忆写入门、关键寻址、权值读取与响应门几个部分。DM-GAN用文字生成图像过程是多阶段的,第一阶段的生成步骤与StackGAN相似。在第二阶段,将上一阶段的生成图像和文字特征输入前,使用记忆写入门来计算一个单词的重要性。关键寻址用于检索相关记忆,值读取则根据相似度概率对价值记忆的加权求和。响应门则用于控制动态的信息流,以及更新图像特征,减轻了多阶段生成对抗对于初始阶段生成图像的依赖。

采用双塔结构的SD-GAN[31]根据两个文本描述是否来自同一个真实图像进行对比,产生的对比损失(图5 $L_{contrast}$)用于计算两个描述间的举例,如果两段文字描述的是同一幅图像,它们的对比损失将会尽可能达到最小。

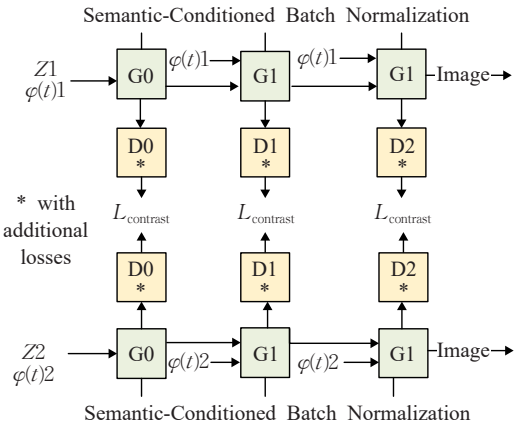

图5　SD-GAN结构

Fig.5　Structure of SD-GAN

同样采用双塔结构的SEGAN[32]是一种基于增强语义的算法。将图像层次的语义一致性引入GAN的训练中,对生成的图像分级,以提高结构一致性,同时建立适应注意权重来增强算法的精确度和稳定性。TIMAM方法[33]使用对抗损失学习模态不变的特征表示,来学习图像和文本级别的区别特征表示。结合BERT模型来提取词向量,用以文字图像匹配域。

采用循环一致性的方法(Mirror GAN[34]、Chen[35]、Lao[36]等)将生成图像还原成原始的描述文本,形成一个循环,如图6所示,MirrorGAN使用语句的文字-图像-文字重描述方法进行T2I。随机噪声结合文字增强向量与注意力模块词向量,经过多阶段生成与对抗,将文字转换为图像,再采用CNN[51]编码器和LSTM对图像进行还原,使得文字生成的图像可以对原始文字进行镜像还原,进一步提高了生成图像与文字描述的语义一致性。

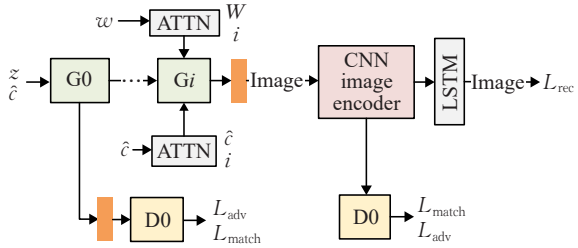

图6　Mirror GAN结构

Fig.6　Structure of Mirror GAN

基于循环一致性方法,CI-GAN[37]在StyleGAN[52]上将输入噪声映射到潜在空间中,生成潜在编码获取更多的语义。同时通过GAN生成图像用反演的方式将生成图像传回潜在空间,获取生成图像的潜在编码,在反演训练中引入循环一致性损失,使两个潜在编码相同并遵循同样的分布;Das等[38]提出的自监督深度学习方法,结

合GAN和基于最大平均差异(MMD)的生成网络进行无监督学习,实现文字和图像的相互转换。

将生成和判别划分为多阶段以提升生成图像的分辨率与精细度。引入额外网络(如注意力、记忆网络等)确保生成图像与描述文字的匹配程度,可以提高生成图像的质量,同时网络结构与训练过程也会变得复杂,计算的时间花销与难度都会增加。

应对上述问题,DF GAN[48]没有引入额外网络,而是用一组生成器和判别器,结合匹配感知的零中心梯度惩罚直接合成图像。其中深度文本图像融合块能够有效地利用文本描述的语义,并在生成过程中对文本和图像特征进行深度融合;HF GAN[42]则对图像不同层次的信息进行收集并进行特征融合,将生成图像自适应融合在一起,使空间粗糙的低分辨率特征图。采用了身份加法、加权加法和快捷连接作为融合方法包含并指示生成图像的整体语义结构,指导细节的生成;TDA NET[46]通过提供的描述文本来填充图像损坏的语义信息,通过双多模注意机制对比得出图与文的剩余部分,提取被损坏区域。同时采用了图像-文本匹配损失法,使得生成图像与文本语义相似性最大化;TA-GAN[39]可生成经过语义处理的图像,根据输入文字创造局部词级判别器以分类属性,使生成器只生成给定文本区域图像,实现以输入文字来修改图像中物体的功能;TVBI GAN[45]针对自然图像可能会模糊,难以从简短的描述中提取出精准语义的问题,提出了利用视觉背后的语义特征来协助文本生成图像的方法;FA GAN[47]结合了自监督判别器和特征感知损失来更好地提取特征表示并合成图像;Text-SeGAN[40]基于TAC-GAN但不进行分类预测。对于给定的图像类别标签生成正向与负向训练实例,根据与类中正示例的语义距离进行负抽样。使生成图像具有更好的多样性,不受模式崩溃的影响;TextStyleGAN[44]在StyleGAN[51]的基础进行了扩展,将其用于T2I,不仅可以生成高分辨率的图像,还可以进行语义操作。使用预训练的图像-文字匹配网络获得文字特征向量,进行拼接后组成句子向量,结合噪声映射到中间的潜在空间。同时对文字和图像特征进行注意引导。使用跨模态匹配损失和跨模态分类损失,结合判别器中的条件与非条件损失进行图文匹配;ControlGAN[43]不仅可以生成图像,还可以通过改变文字描述内容对图像进行操作,例如修改图像的类别、纹理、颜色等,同时不改变其他与之无关的部分。所采用的通道级注意力将语义中有意义的部分与文字描述中的单词联系起来。词级注意空间使生成器能合成最相关单词对应的图像区域;Bridge-GAN[41]采用了一个中间网络,将文字向量和噪声映射到一个过渡空间中,并提出两个新的损失。第一个损失计算中间潜在空间与输入文字向量之间的交互信息,确保文字信息存在于过渡空间里。第二个损失计算生成图像与输入描述间的交互信息,提高语义一致性。

对于算法的选择往往是根据实际的任务需求与应用场景来决定的。需要快速得出训练结果以供后续研究对比或修改的场合,轻量级的T2I算法较为合适。而需要得出精细的生成图像时,则采用结构较为复杂的算法。同时,各种不同的机制与网络之间并不是完全隔绝和对立的,许多算法中都对多种机制和网络的优势进行了结合并投入到T2I任务中。这也是近年T2I算法的发展趋势之一。

2.2 额外监督信号

表2列举了部分使用额外监督信号辅助生成图像的T2I算法,使用额外监督信号可以提升模型生成图像的质量,同时在训练期间,它们也需要更多额外的注释。

表2 采用额外监督信号的T2I算法

Table 2 T2I methods with additional supervision

Input	method
布局	GAWWN[53](2016)
	Layout GAN[54](2019)
	Attn+OP[55](2019)
	Layout2Im[56](2019)
	OP-GAN[57](2020)
	Attrlost GAN[58](2021)
	OC-GAN[59](2021)
掩码	Hong[60](2018)
	OBJ GAN[61](2019)
	LeicaGAN[62](2019)
	Pavllo[63](2020)
	Wang[64](2020)
	AGAN-CL[65](2020)
场景图	Johnson[66](2018)
	Ashual[67](2019)
	PASTE GAN[68](2019)
	Vo[69](2020)
对话	ChatPainter[70](2018)
	Frolov[71](2020)
	VQA-GAN[72](2020)
	Jiang[73](2021)
多个文字描述	C4Synth[74](2018)
	StoryGAN[75](2019)
	RIFE GAN[76](2020)
	Cook GAN[77](2020)
	Wadhawan[78](2020)

2.2.1 布局

GAWWN[53]可以通过给出位置与内容的文字说明,将随机噪声、文字说明和布局作为生成器的输入来生成图像。实现对生成图像中的目标对象姿态与位置进行控制;Layout GAN[54]使用不同种类的二维元素的集合关系进行建模,生成器以一组随机放置的2D图形元素作为输入,使用基于CNN[51]的判别器优化生成布局。

Layout2Im[56]采用了粗糙空间布局(边界框+对象类

别)来生成图像与布局关系;Attrlost GAN[58]实现了布局可重构,使用Attr-ISLA可以对单个对象的外观进行控制。将对抗节点损失用于对象-属性特征上,来鼓励生成器合成反映输入属性的图像。

Hinz等[55]通过向生成器和判别器中添加目标路径,对目标位置和外观进行建模,当对象路径侧重于在有意义的位置生成单个对象时,全局路径生成符合整体图像描述和布局的背景。基于这一原理提出的OP-GAN[57]在生成器和鉴别器的更高层添加额外的对象路径,结合匹配和不匹配的边框,以及图像对使用额外的边框匹配损失,辅助生成图像;OC-GAN[59]提出了场景-图相似度模块(SGSM)用以学习场景中对象之间的空间关系的表示。提高了生成图像的质量,还解决了布局中的"虚假对象"没有边框的问题,以及边框重复导致图像中对象合并的问题。

2.2.2　掩码

由于多数方法并不专门为图像中的对象及其关系建模,在生成复杂场景时会存在困难。OBJ GAN[61]用于以目标为中心的复杂场景文字生成图像。将文本描述和预生成的语义分布,作为图像生成器的输入,为复杂场景进行以对象为中心的文本到图像的合成;Leica GAN[62]有一个专门的学习步骤,文本图像编码器在这一阶段学习语义、纹理和颜色有关的先验知识,文本掩码编码器则学习形状和布局,这些互补的先验被聚合,并用于利用局部和全局特性来逐步创建图像;AGAN-CL[65]用于自动生成真实图像的位置和具体形状,由Attentive GAN和上下文损失算法组成。上下文网络用于生成图像的轮廓,循环转换自动编码器将轮廓转化为真实图像,将轮廓注入生成网络中以引导整个生成网络集中于对象区域,上下文损失与周期一致性损失则连接多领域间的差距;Wang等[64]提出了一种利用语义布局来指导图像合成的具有空间约束的端到端框架,利用语义布局和隐藏的视觉特征逐步生成和细化图像。在每一个阶段,生成器都会生成一个图像和布局,并交由对应的判别器进行判定;Pavllo等[63]提出了一种利用稀疏的实例语义掩码的弱监督方法,从前景分解背景两步生成过程来实现T2I并允许简单地编辑生成图像内容。

2.2.3　场景图

Johnson等[66]用场景图生成图像,还原带有多目标与关系的复杂语句。推理目标之间的关系。使用图卷积处理图,通过预测绑定标注框,分割掩码来计算场景布局,将布局转换图像。Ashual等[67]对这一算法进行了扩展,使用掩码将布局向量与图像外观分离,使用户能够更好地控制布局,并生成与输入场景图更匹配的图像。外观属性既可以从预定义的图像集中选择,也可以从另一个图像中复制。PASTE GAN[68]可以从场景图与裁剪中生成图像,操控物体的视觉外观,生成空间关系。Vo等[69]提出一个端到端网络,从给定的场景图中生成图像。这一网络由两个部分组成:视觉布局模块和Stacking GAN。视觉布局模块使用场景图中所有的关系来预测边框,分别使用个体关系来从初始边框单位中预测输出文本中所有的关系生成布局,每个实体都可以参与多个关系,因此所有关系单元都被统一,并使用卷积LSTM将其转换为视觉关系布局。视觉关系布局反映了场景图中的结构(对象和关系),每个实体对应于一个细化的边界框。在Stacking GAN中使用视觉关系布局来渲染最终图像。

2.2.4　对话

单独的文字描述所包含的信息量并不足以生成具有多个互相作用的场景。ChatPainter[70]将对话信息用于T2I,使用对话数据集与COCO图像进行匹配。

VQA-GAN[72]在AttnGAN-OP[54]的基础上采用问答的方式来设置图像生成器,将图像生成器条件设置在局部相关文本上用问答(QA)编码器输出局部与全局的表示,并以此作为GAN的条件信息分阶段生成图像。同时采用VQA损失鼓励QA对和生成的图像之间的相关性。典型的VQA模型以图像和问题为输入,进行分类训练使正确答案的概率最大化。因此VQA精度可以作为评估输入QA对和生成图像之间的一致性的度量指标;Frolov等[71]提出在不改变体系结构的情况下利用VQA数据。通过简单地连接QA对并将它们作为额外的训练样本和外部VQA损失,在图像质量和图像-文本对齐指标上提高性能。

Jiang等[73]使用GAN在潜在空间中建模一个连续的语义场,实现对生成的图像进行编辑。根据用户输入的请求和语义字段,算法将产生反馈,形成与用户的对话效果,既提高了对于用户需求的理解,也提升了用户体验。

2.2.5　多个文字描述

常见的数据集通常每幅图像包含多个描述,因此输入不止一段描述文字可以提供额外的信息来更好地描述整个场景。RIFE GAN[76]利用训练数据集形成的先验知识来丰富给定的描述,将特征合成图像;Story GAN[75]可以通过故事生成连续图像,以实现故事可视化,给定一个多语句的段落,通过深层的上下文编码器以及故事和图像级别的两个判别器生成系列图像并动态跟踪故事流程;Cook GAN[77]通过烹饪模拟器子网络,对食物图像进行基于配料和烹饪方法之间相互作用的增量变化,不仅能根据食谱描述生成菜肴图像,同时能让菜肴外观根据烹饪动作和配料而变化;Wadhawan等[78]通过增加描述数量,使用结构化的文本描述进行人脸图像生成。

3　评估文字生成图像算法

本章介绍常用于评估T2I算法的评价指标以及数

据集。评估对于衡量和改进算法至关重要。明确一个良好的指标所需要的属性,不仅可以在建立新的指标时作为参考,对以后的相关研究也能有所帮助。

3.1 常用数据集

文字生成图像任务主要在三个数据集上进行验证:CUB_200_2011[79]、Oxford 102 Flowers Dataset[80]和COCO[81]。对于一些用于专门任务的算法,需要使用特定领域的数据集,如CelebA[82]、Recipe1M等[83]。

Oxford 102 Flowers花卉数据集有102个类,每个类别包含40到258张图像。总共8 189张图像,其中训练集7 034张,测试集为1 155张,每张图像拥有10个描述信息;CUB是目前细粒度分类识别研究的基准图像数据集。共有11 788张鸟类图像,200类鸟类。其中训练集有5 994张图像,测试集有5 794张图像,每张图像均提供10个描述以及图像类标记信息,如鸟的绑定标注框,关键部分信息及属性;COCO是一个大型物体检测,分割和字幕数据集,从复杂的日常场景中截取,图像中的目标通过精确的分割进行位置的标定。标注有类别、位置信息,以及图像的语义文本描述。有80个类别,33万张图像,其中20万张有标注,每张图像拥有5个描述。

3.2 图像质量指标

3.2.1 Inception score(IS)

IS[84]用于衡量GAN网络生成图像的质量以及生成图像的多样性。判别器使用预训练的Inception网络[85]对生成图像进行分类。计算其类别标签分布 $P(y|x)$,y 是标签。如果生成的图像有意义,这一分布会有较低的熵。为获取更高质量的生成图像,$P(y|x)$ 越大越好,熵则是越小越好。为综合这两个指标,需要使用KL散度衡量两个分布之间的距离,式(1)中 D_{KL} 为KL散度的计算公式。

$$IS(G)=\mathrm{e}^{E_{x\in p}D_{\mathrm{KL}}(p(y|x)\|p(y))} \tag{1}$$

3.2.2 Fréchet inception distance(FID)

FID[86]在T2I中用于计算真实数据分布和生成图像的数据分布之间的距离,用均值协方差来计算两个分布之间的距离。FID计算公式如下:

$$FID(x,g)=\|\mu_x-\mu_g\|_2^2+\mathrm{tr}(\Sigma_x+\Sigma_g-2(\Sigma_x\Sigma_g)^{\frac{1}{2}}) \tag{2}$$

其中,tr为矩阵的迹,μ 为均值,Σ 是协方差,x 为真实图像,g 为合成图像。FID值越低代表两个分布之间的距离越小。较低的FID值意味着生成图像质量较高。

3.2.3 Sliced Wasserstein distance(SWD)

Wasserstein距离[87]用于计算两个概率分布之间的距离。Rabin等[88]认为将Wasserstein距离用于图像处理十分困难,所以用Sliced Wasserstein距离(SWD)来代替这一度量,在生成图像和真实图像的拉普拉斯金字塔表示中提取的局部块上,计算多尺度统计相似性。通过线性映射得到高维概率分布的一维表示,计算两个概率分布的Wasserstein距离。

3.2.4 GAN-test,GAN-train

在GAN-test与GAN-train中[89]没有设计给出可量化的评价指标,而是计算几个指标并进行对比分析,从而评价GAN的性能。生成图像不仅应该写实,还应该可以识别它来自哪一个类别。设训练集样本 S_{t},验证集为 S_{v},生成的样本集为 S_{g}:(1)在 S_{t} 上训练分类器,在 S_{v} 上计算分类器的准确率,记为GAN-base;(2)在 S_{g} 上训练分类器,在 S_{v} 上计算分类器的准确率,记为GAN-train;(3)在 S_{t} 上训练分类器,在 S_{g} 上计算分类器的准确率,记为GAN-test。

如果一个cGAN已经达到了最优,可以完美捕获目标分布,它生成的图像集 S_{g} 就应该与训练集 S_{t} 有相同的大小。在两个集合上训练的分类器所产生的验证精度也应该是大致相同的。

3.3 图文语义相关性指标

3.3.1 R-precision

R-precision[14]用于度量文本描述和生成的图像之间的视觉与语义间的相似性。使用DAMSM对生成图像与文本进行全局特征向量提取,并计算余弦相似度,再对每一幅图像依照相似度进行降序排序。R-precision通过对提取的图像和文本特征之间的检索结果进行排序。

3.3.2 Visual-semantic similarity(VS)

VS[21]是一个视觉-语义向量模型,用以计算合成图像和输入文本之间的距离。设 $\boldsymbol{x}$ 是生成图像特征向量,$\boldsymbol{y}$ 是文本特征向量,二者距离可计算为 $C(\boldsymbol{x},\boldsymbol{y})$。VS越高,图像与文本语义一致性越好。

$$C(\boldsymbol{x},\boldsymbol{y})=\frac{\boldsymbol{x}\cdot\boldsymbol{y}}{\|\boldsymbol{x}\|_2\cdot\|\boldsymbol{y}\|_2} \tag{3}$$

3.3.3 语义对象精度(SOA)

SOA[57]使用预先训练过的目标检测器来评估生成的图像是否包含图像标题中提到的对象。使用YOLO网络进行识别,YOLO网络的输出结果为SOA-C:一个类别均值,如平均检测到给定目标的多少张图像。SOA-I意为图像均值,平均检测到期望目标的多少张图像。

$$SOA\text{-}C=\frac{1}{|C|}\sum_{c\in C}\frac{1}{|I_c|}\sum_{i\in I_c}YOLOv3(i_c) \tag{4}$$

$$SOA\text{-}I=\frac{1}{\sum_{c\in C}|I_c|}\sum_{c\in C}\sum_{i_c\in I_c}YOLOv3(i_c) \tag{5}$$

其中,目标类 $c\in C$ 和图像 $i_C\in I_C$ 为包含类 c 的目标。监测到 C 类目标 $YOLO(i_C)$ 取值为1,否则为0。

还有一些指标被用于测量生成的图像和文本的相关性。如BLEU[90]、METEOR[91]和CIDEr[92]等。在T2I中,它们被用于验证算法是否产生反映文字描述含义的

图像。

使用自动指标结合人工评估,可以保证T2I算法评估的结果准确。将一组随机的描述文字用于T2I算法中进行图像生成,获得生成图像与对应的文字描述,以人工的方式进行排名,或者选择最好的生成结果。人工评估的意义之一在于获取与人工评判相关的自动指标,使排名更加准确有意义。

目前人工评估没有统一的标准,因此结果难以直接对比。例如Hong等[60]根据文本的相关性对图像排序;Hinz等[57]选择了最符合文字描述的图像。此外并不是所有的指标都适合用于人工排序,如FID、SOA的人工排序与匹配程度相对于IS、R-prec和CIDEr更准确。

4　讨论

本章从几个方面讨论当前T2I算法以及评估指标方面存在的一些问题,归纳目前的研究现状并对今后的研究方向进行概述。

4.1　算法方面

自2016年Reeds等将GAN用于T2I任务以来,如今的T2I算法从单阶段到多阶段生成对抗,综合运用多种损失函数与网络结构(注意力、循环一致性、动态记忆网络、双塔结构等),不论是图像分辨率与精细度还是图像对描述文字的匹配程度都取得了巨大的进步,可以生成花卉、鸟类、人脸等内容的高质量图像。虽然在COCO等更具挑战性的数据集上也有了很大的改进,但所产生的图像(特别是单个对象)缺乏细粒度的细节和清晰度。目前的T2I方法已经适应了非条件图像生成模型,研究更好的条件图像生成的适应性可能比为T2I设计特殊架构更有效。

文字描述中所使用的语法、词汇的位置、数值等信息都会对生成结果造成影响。自AttnGAN提出以来许多算法都使用经预训练的文本编码器来获取文字特征向量。至于向量如何具体影响算法的性能则缺乏研究。虽然有Pavllo等[63]使用基于transformer的编码器获取词向量,Rombach等[93]使用invertible network在BERT与BigGAN[94]间转换来处理T2I。但相关研究进展始终较为缓慢。

目前的T2I算法以GAN为核心,在取得了许多的成果同时,也存在部分暂未解决的问题。如GAN训练过程是漫长且不可控的,有时会产生模式崩溃。虽然关于解决模式崩溃的研究目前已初见成效(Mao[95]、Cha[96]等),但将GAN与其他算法目前已有的成果相结合(VAE、对称蒸馏网络[97-98]、自回归模型[99]、Transformer[100]等),提出新的损失函数[101],可能会有所突破。

虽然现在的数据集常会为一张图像提供多个文字描述,但是这些描述的语义却是相似的。由于少数的句子往往难以描述出复杂的场景,这些描述时常不能提供足够的信息。目前的算法很难直接生成具有多个互相作用的目标场景。因为信息量不足和缺乏对场景和目标的理解,算法不能通过生成完整场景对目标进行建模。虽然已有相关研究(Hong[60]、Li[54]、Pavllo[63]等)提出了不同的方法,分别使用布局、掩码、分层的方法来生成复杂场景,但仍需要在这一方向投入研究以取得突破进展。

4.2　数据集方面

大型高质量的数据集是深度学习方法成功的基础。如今的T2I方法逐渐不用Flowers数据集进行评估了,Flowers与CUB数据集有较高的相似之处,都只包含单一的目标对象。而CUB包含的种类达到200个,远多于Flower的102个类。换言之,在单目标对象数据集上评估T2I算法使用CUB数据集得出的结果就足够具有代表性了。而对于诸如生成人类面部图像的需求,则可以使用CelebA等专门的数据集。对于单目标生成任务,现在许多方法都能生成高分辨率的图像。今后对于T2I算法评估的重点应该放在生成图像与描述文字的一致性上。

可用于描述多目标的复杂场景的数据集通常存在着同样的问题:分辨率较低。目前的T2I算法依然很难训练合成复杂场景下具有多个物体互相作用的图像。对此,将T2I算法的各个模块进行等比例放大(Stap[44]、Zhang[21]等),虽然提高了生成图像的分辨率,但也大大提高了对于硬件的要求,同时需要更多的时间进行训练。在试着提高生成图像的分辨率时,也应该考虑生成图像是否保持了它的真实性。基于图像局部一致性的想法[75],可以尝试为图像的某一部分区域添加文字描述。如Visual Genome[102]数据集包含了对单个图像区域的描述。

文字生成图像依赖于图像描述会存在单向注释问题,即数据集会为一张图像匹配多个描述文本,这可能会让同一个描述文本匹配到了多张不同的图像。Parekh等[103]扩展了COCO数据集的注释,并为存在的"图像-文本"配对。新的配对与描述文本之间提供连续的VS评分。

如今T2I常用的数据集多使用英语描述。为增加其适用性,对于描述文字的多语言研究是必不可少的。尝试其他语言的文字描述并分析其描述方式是否存在差异是有利于文字生成图像方法的泛化的,一个T2I算法应该不必重新训练就可用于多种语言。

4.3　评估方面

对于T2I算法的评估并非轻易,基于目前的技术与知识,用FID来评估生成图像的视觉质量,计算其与真实数据分布的举例是相对可靠的。使用SOA结合人工

方式来评估生成图像与文字描述的图文一致性是相对可靠的。

IS与FID等目前常用的指标都有各自的缺陷存在，使用IS和FID对不同的算法进行评估是不公平的，尤其是对于非GAN算法，在评估过程中往往会被惩罚[104]。今后的评估策略应该尽量做到与算法类型无关。

目前一些算法已经达到了COCO真实图像给出的IS、R-prec和CIDEr的上限（如表3所示）。这种情况下生成图像不够真实，则表明了这些指标并不可靠。IS是可能饱和甚至过拟合的，将batch size放大可以优化这一指标[61]，可以饱和甚至过拟合。而对于R-prec，如果训练和评估算法时使用了相同的文字编码器，这一算法在训练期间就已经达到了过拟合状态[4]。用不同的模型在Conceptual captions[105]上进行训练，FID、SOA、VS都低于真实图像。

表3 部分算法在COCO上的IS、R-prec和CIDEr

Table 3 IS, R-prec and CIDEr on COCO

model	IS	R-prec	CIDEr
REAL IMAGE[57](Hinz,2020)	34.88	68.58	79.5
AttnGAN[14](Xu,2017)	24.76	85.47	69.5
AttnGAN+OP[55](Hinz,2019)	27.88	—	68.9
DM-GAN[30](Zhu,2019)	30.49	88.56	**82.3**
MirrorGAN[34](Qiao,2019)	26.47	74.52	—
SD-GAN[31](Yin,2019)	**35.69**	—	—
Wang[64](2020)	29.03	82.70	—
OP-GAN[57](Hinz,2020)	27.88	**89.01**	81.9
DAE-GAN[27](Ruan,2021)	35.08	—	—

Borji等[106]提出了一个完整的列表，在评估T2I算法生成的图像时，会偏好具有这些特征的算法：(1)具有较好的保真度和多样性；(2)解纠缠表示；(3)带有标注好的绑定边框；(4)与人工评估结果高度一致；(5)具有较低的采样次数和计算复杂度。因此，好的T2I评估，应该从以下几个方面进行：(1)图像的质量与多样性；(2)图像对文字描述内容是否完整反映；(3)对描述中数值与位置等细节信息是否精确反映；(4)生成图像符合是否人类常识；(5)生成图像是否稳定，如果对文字描述内容进行部分替换，生成图像所变化的部分也应该与文字变化相对应；(6)是否具有较好的可解释性，如果生成图像不能较好地反映描述内容，能找到是哪一部分导致了问题出现；(7)评估的过程应该是自动的，不存在必须要手动操作的环节。

IS和FID都使用了在ImageNet[107]上预先训练的Inception网络，在生成有多个对象的复杂场景图像（如COCO）时会出现问题。因此新的指标逐渐被提出和投入使用。例如当文字描述里有与位置相关的信息，则可以使用Scene FID[58]来评估裁剪的目标对象；运用LPIPS[108]评估两组来自相同的文字描述生成的图像之间的多样性。

建立新的T2I评估指标是一件困难的事情，因为对于生成图像与文字描述的语义是否对齐这一概念是没有准确定义的，也很难制定出统一的标准。新指标的提出有利于对T2I算法进行更精确的评估，对进一步研究如何简化评估过程，提高评估效率也有所帮助。

相同的算法常会在不同文献实验中产生不同的结果（表4中*为目标算法，在不同的文献实验中得出了不同分值）。因为实验和评估的标准没有统一，评估的结果会依据图像分辨率和数量等参数发生变化。

表4 相同的算法在不同的实验中结果不同

Table 4 Different result on same model

Model	IS	FID	R-prec
AttnGAN[14](Xu,2017)*	25.89	—	85.47
DM-GAN[30](Zhu,2019)	—	35.49	—
SEGAN[32](Tan,2019)	25.56	34.28	—
Obj-GAN[61](Li,2019)	23.79	28.76	82.98
MirrorGAN[34](Qiao,2019)	—	—	72.13
Huang[109](2019)	—	32.12	—
OP-GAN[57](Hinz,2020)	23.61	33.10	83.80
Wang[64](2020)	23.89	28.76	82.90
CPGAN[110](Liang,2020)	—	—	82.98
Frolov[58](2020)	**26.66**	**27.84**	**83.82**
DM-GAN[30](Zhu,2019)*	30.49	32.64	88.56
Obj-GAN[60](Li,2019)	—	—	82.70
OP-GAN[56](Hinz,2020)	**32.32**	**27.34**	**91.87**
CPGAN[108](Liang,2020)	30.49	—	88.56
Obj-GAN[61](Li,2019)*	30.29	25.64	91.05
OP-GAN[57](Hinz,2020)	24.09	36.52	87.84
Wang[64](2020)	**30.89**	**17.04**	83.00
CPGAN[110](Liang,2020)	30.29	—	**91.05**
OP-GAN[56](Hinz,2020)*	27.88	24.70	89.01
CPGAN[110](Liang,2020)	**28.57**	—	**87.90**

虽然人工评估方式需要耗费大量的时间与精力，具体条件也有可能发生较大变化。但是想要得出可靠的评估结果，自动化评估指标与人工评估方式的结合是必不可少的。但人工评估目前缺乏统一的标准，流程标准化的提出会是很重要的任务（如HYPE[111]）。

不论是采用何种方式进行评估，对于所获得的结果都应该准确描述数据来源。并且要注明是来源于参考文献还是其改进实验。对于人工评估则也该详细描述流程，如采样数量，使用的模型与具体说明等。

4.4 应用方面

在应用层面，T2I往往是由实际需求驱动的，许多行业（如计算机辅助设计、图像编辑）都需要对图像进行精确控制，这也是今后的研究方向之一。对于图像进行局部编辑，在修改要求的内容同时，还需要保留不相关的内容。近年来有不少的成果以多种方式完成了这一功能（Mani GAN[25]、Pavllo等[63]、Jiang等[73]、TA-GAN[39]、ControlGAN[43]）。在图像处理领域也有许多研究涉及通

过文字描述进行图像处理(TDA NET[46]、Dong等[112]、Liu等[113]、Zhu等[114])。

编辑语义特征图和标签对于使用者而言更加灵活。文字描述可以传达较多的内容,今后的T2I算法可以不局限于单一描述,而是从多个不同的描述中提取信息构成整体图像。此外,语音生成图像(S2I)或文字生成视频(Wang等[115]、Balaji等[116]、Deng等[117]、Choi等[118]、Jia等[119]、Suris等[120])和T2I算法的原理具有相似之处。将T2I算法的文字编码器替换为语音编码器可以实现S2I(如Balaji等[116]、Li等[121])。使用文字生成视频也具有较大的研究价值。但是不论使用语音还是文字,生成的视频每一帧都是连续的,对于算法的评估具有一定难度。

5　总结

概述了近年来提出的T2I算法与评估方式,并对存在的问题进行了探讨。在列举T2I算法时依照使用额外监督信号与否进行了划分。直接生成的T2I算法不需要额外的监督信号,可以直接合成图像。包括从单阶段生成对抗的传统算法到以堆叠结构为代表的多阶段生成对抗算法,以及引入了注意力机制、循环一致性结构和双塔结构的算法。使用额外的监督信号的算法在提升性能的同时需要更多注释,包括布局、掩码、对话、场景图等。对于算法的评估则列举了常用数据集与评估指标,随后分别从算法、数据、评估和应用几个方面展开了讨论。最后列举了实际应用中目前的研究成果,采用文字描述对图像进行编辑和修改,以及使用语音描述生成图像已有研究部分成果,尝试使用语音或文字描述生成视频将是今后的研究方向。

参考文献:

[1] GOODFELLOW I,POUGET-ABADIE J,MIRZA M,et al.Generative adversarial nets[C]//Advances in Neural Information Processing Systems,2014:2672-2680.

[2] MIRZA M,OSINDERO S.Conditional generative adversarial nets[J].arXiv:1411.1784,2014.

[3] REED S,AKATA Z,YAN X,et al.Generative adversarial text to image synthesis[C]//International Conference on Machine Learning,2016:1060-1069.

[4] FROLOV S,HINZ T,RAUE F,et al.Adversarial text-to-image synthesis:a review[J].Neural Networks,2021,144:187-209.

[5] HOCHREITER S,SCHMIDHUBER J.Long short-term memory[J].Neural Computation,1997,9(8):1735-1780.

[6] GREGOR K,DANIHELKA I,GRAVES A,et al.DRAW:a recurrent neural network for image generation[C]//International Conference on Machine Learning,2015:1462-1471.

[7] MANSIMOV E,PARISOTTO E,BA J L,et al.Generating images from captions with attention[C]//International Conference on Learning Representations,2016.

[8] KINGMA D P.Max welling auto-encoding variational Bayes[C]//International Conference on Learning Representations,2014.

[9] WANG Z,SHE Q,WARD T E.Generative adversarial networks in computer vision:a survey and taxonomy[J].ACM Computing Surveys,2021,54(2):1-38.

[10] REED S,AKATA Z,LEE H,et al.Learning deep representations of fine-grained visual descriptions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2016:49-58.

[11] DASH A,GAMBOA J C B,AHMED S,et al.TAC-GAN-text conditioned auxiliary classifier generative adversarial network[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision,2017.

[12] ZHANG H,XU T,LI H,el al.StackGAN:text to photo-realistic image synthesis with stacked generative adversarial networks[C]//Proceedings of the IEEE International Conference on Computer Vision,2017:5907-5915.

[13] SOUZA D M,WEHRMANN J,RUIZ D D,et al.Efficient neural architecture for text-to-image synthesis[C]//2020 International Joint Conference on Neural Networks,2020:1-8.

[14] XU T,ZHANG P,HUANG Q,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,2018:1316-1324.

[15] WANG T,ZHANG T,LOVELL B.Faces à la carte:text-to-face generation via attribute disentanglement[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision,2021:3380-3388.

[16] DEVLIN J,CHANG M W,LEE K,et al.BERT:pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies(1)(Long and Short Papers),2019:4171-4186.

[17] ODENA A,OLAH C,SHLENS J.Conditional image synthesis with auxiliary classifier GANs[C]//International Conference on Machine Learning,2016:2642-2651.

[18] ZHANG C,PENG Y.Stacking VAE and GAN for context-aware text-to-image generation[C]//International Conference on Multimedia Big Data,2018:1-5.

[19] TAN F,FENG S,ORDONEZ V.Text2Scene:generating compositional scenes from textual descriptions[C]//Proceedings of the IEEE Computer Vision and Pattern Recognition,2018:6703-6712.

[20] ZHANG H,XU T,LI H,et al.StackGAN++:realistic image synthesis with stacked generative adversarial networks[J].

IEEE Transactions on Pattern Analysis and Machine Intelligence,2018,41(8):1947-1962.
[21] ZHANG Z,XIE Y,YANG L.Photographic text-to-image synthesis with a hierarchically-nested adversarial network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2018:6199-6208.
[22] BODLA N,HUA G,CHELLAPPA R.Semi-supervised fusedGAN for conditional image generation[C]//European Conference on Computer Vision,2018:669-683.
[23] GAO L,CHEN D,SONG J,et al.Perceptual pyramid adversarial networks for text-to-image synthesis[C]//Proceedings of the AAAI Conference on Artificial Intelligence,2019:8312-8319.
[24] GARG K,SINGH A K,HERREMANS D,et al.PerceptionGAN:real-world image construction from provided text through perceptual understanding[C]//2020 Joint 9th International Conference on Informatics Electronics and Vision ICIEV and 2020 4th International Conference on Imaging Vision and Pattern Recognition,2020:1-7.
[25] LI B,QI X,LUKASIEWICZ T,et al.ManiGAN:text-guided image manipulation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2020:7880-7889.
[26] ZHANG S,WANG D,ZHAO Z,et al.MGD-GAN:text-to-pedestrian generation through multi-grained discrimination[C]//Chinese Conference on Pattern Recognition and Computer Vision.Cham:Springer,2021:662-673.
[27] RUAN S,ZHANG Y,ZHANG K,et al.DAE-GAN:dynamic aspect-aware GAN for text-to-image synthesis[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision,2021:13960-13969.
[28] ZHANG H,KOH J Y,BALDRIDGE J,et al.Cross-modal contrastive learning for text-to-image generation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2021:833-842.
[29] MAHESHWARI P,JAIN N,VADDAMANU P,et al.Generating compositional color representations from text[C]//Proceedings of the 30th ACM International Conference on Information and Knowledge Management,2021:1222-1231.
[30] ZHU M,PAN P,CHEN W,et al.DM-GAN:dynamic memory generative adversarial networks for text-to-image synthesis[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2019:5802-5810.
[31] YIN G,LIU B,SHENG L,et al.Semantics disentangling for text-to-image generation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2019:2327-2336.
[32] TAN H,LIU X,LI X,et al.Semantics-enhanced adversarial nets for text-to-image synthesis[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision,2019:10501-10510.
[33] SARAFIANOS N,XU X,KAKADIARIS I A.Adversarial representation learning for text-to-image matching[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision,2019:5814-5824.
[34] QIAO T,ZHANG J,XU D,et al.MirrorGAN:learning text-to-image generation by redescription[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2019:1505-1514.
[35] CHEN Z D,LUO Y.Cycle-consistent diverse image synthesis from natural language[C]//IEEE International Conference on Multimedia & Expo Workshops,2019:459-464.
[36] LAO Q,HAVAEI M,PESARANGHADER A,et al.Dual adversarial inference for text-to-image synthesis[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision,2019:7567-7576.
[37] WANG H,LIN G,HOI S C.Cycle-consistent inverse GAN for text-to-image synthesis[C]//Proceedings of the 29th ACM International Conference on Multimedia,2021.
[38] DAS A S,SAHA S.Self-supervised image-to-text and text-to-image synthesis[C]//International Conference on Neural Information Processing.Cham:Springer,2021:415-426.
[39] NAM S,KIM Y,KIM S J.Text-adaptive generative adversarial networks:manipulating images with natural language[C]//Advances in Neural Information Processing Systems,2018:42-51.
[40] CHA M,GWON Y,KUNG H T.Adversarial learning of semantic relevance in text to image synthesis[C]//Proceedings of the AAAI Conference on Artificial Intelligence,2018:3272-3279.
[41] YUAN M,PENG Y.Bridge-GAN:interpretable representation learning for text-to-image synthesis[J].IEEE Transactions on Circuits and Systems for Video Technology,2019,30(11):4258-4268.
[42] HUANG X Z,WANG M,GONG M.Hierarchically-fused generative adversarial network for text to realistic image synthesis[C]//Conference on Computer and Robot Vision,2019:73-80.
[43] LI B,QI X,LUKASIEWICZ T.Controllable text-to-image generation[C]//Advances in Neural Information Processing Systems,2019.
[44] STAP D,BLEEKER M,IBRAHIMI S.Conditional image generation and manipulation for user-specified content[C]//Proceedings of the IEEE Computer Vision and Pattern Recognition Workshop,2020.
[45] WANG Z,QUAN Z,WANG Z,et al.Text to image syn-

thesis with bidirectional generative adversarial network[C]// Conference on Multimedia and Expo,2020:1-6.

[46] ZHANG L,CHEN Q,HU B,et al.Text-guided neural image inpainting[C]//Proceedings of the 28th ACM International Conference on Multimedia,2020.

[47] JEON E,KIM K,KIM D.FA-GAN:feature-aware GAN for text to image synthesis[C]//International Conference on Image Processing,2021:2443-2447.

[48] TAO M,TANG H,WU S,et al.DF-GAN:deep fusion generative adversarial networks for text-to-image synthesis[C]//Conference on Computer Vision and Pattern Recognition,2022.

[49] LAI W S,HUANG J B,AHUJA N,et al.Deep Laplacian pyramid networks for fast and accurate super-resolution[C]// Proceedings of the IEEE Computer Vision and Pattern Recognition,2017:5835-5843.

[50] LIN T Y,DOLLÁR P,GIRSHICK R B,et al.Feature pyramid networks for object detection[C]//Proceedings of the IEEE Computer Vision and Pattern Recognition, 2017:936-944.

[51] LECUN Y,BOSER B,DENKER J S,et al.Backpropagation applied to handwritten zip code recognition[J]. Neural Computation,1989,1(4):541-552.

[52] KARRAS T,LAINE S,AILA T A.Style-based generator architecture for generative adversarial networks[C]//Proceedings of the IEEE Computer Vision and Pattern Recognition,2018:4401-4410.

[53] REED S E,AKATA Z,MOHAN S,et al.Learning what and where to draw[C]//Advances in Neural Information Processing Systems,2016:217-225.

[54] LI J,YANG J,HERTZMANN A,et al.LayoutGAN:generating graphic layouts with wireframe discriminators[C]// International Conference on Learning Representations, 2019:2-8.

[55] HINZ T,HEINRICH S,WERMTER S.Generating multiple objects at spatially distinct locations[C]//International Conference on Learning Representations,2019.

[56] ZHAO B,MENG L,YIN W,et al.Image generation from layout[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2019: 8584-8593.

[57] HINZ T,HEINRICH S,WERMTER S.Semantic object accuracy for generative text-to-image synthesis[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2020:1552-1565.

[58] FROLOV S,SHARMA A,HEES J,et al.AttrlostGAN: attribute controlled image synthesis from reconfigurable layout and style[C]//DAGM German Conference on Pattern Recognition.Cham:Springer,2021.

[59] SYLVAIN T,ZHANG P,BENGIO Y,et al.Object-centric image generation from layouts[C]//International Conference on Learning Representations,2021:2-7.

[60] HONG S,YANG D,CHOI J,et al.Inferring semantic layout for hierarchical text-to-image synthesis[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2018:7986-7994.

[61] LI W,ZHANG P,ZHANG L,et al.Object-driven text-to-image synthesis via adversarial training[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2019:12166-12174.

[62] QIAO T,ZHANG J,XU D,et al.Learn imagine and create: text-to-image generation from prior knowledge[C]// Advances in Neural Information Processing Systems, 2019:887-897.

[63] PAVLLO D,LUCCHI A,HOFMANN T.Controlling style and semantics in weakly-supervised image generation[C]// European Conference on Computer Vision.Cham:Springer, 2020:482-499.

[64] WANG M,LANG C,LIANG L,et al.End-to-end text-to-image synthesis with spatial constrains[J].ACM Transactions on Intelligent Systems and Technology,2020:1-19.

[65] WANG M,LANG C,LIANG L,et al.Attentive generative adversarial network to bridge multi-domain gap for image synthesis[C]//IEEE International Conference on Multimedia and Expo,2020:1-6.

[66] JOHNSON J,GUPTA A,LI F F.Image generation from scene graphs[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2018: 1219-1228.

[67] ASHUAL O,WOLF L.Specifying object attributes and relations in interactive scene generation[C]//Proceedings of the IEEE International Conference on Computer Vision, 2019:4561-4569.

[68] LI Y,MA T,BAI Y,et al.PasteGAN:a semi-parametric method to generate image from scene graph[C]//Advances in Neural Information Processing Systems,2019:3950-3960.

[69] VO D M,SUGIMOTO A.Visual-relation conscious image generation from structured-text[C]//European Conference on Computer Vision,2020:290-306.

[70] SHARMA S,SUHUBDY D,MICHALSKI V,et al. Chatpainter:improving text to image generation using dialogue[C]//International Conference on Learning Representations,2018.

[71] FROLOV S,JOLLY S,HEES J,et al.Leveraging visual question answering to improve text-to-image synthesis[C]// Proceedings of the Second Workshop on Beyond Vision and Language:Integrating Real-World Knowledge,2020: 17-22.

[72] NIU T,FENG F,LI L,et al.Image synthesis from locally related texts[C]//Proceedings of the International Con-

ference on Multimedia Retrieval,2020:10531-10540.
[73] JIANG Y,HUANG Z,PAN X,et al.Talk-to-edit:fine-grained facial editing via dialog[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision,2021:13799-13808.
[74] JOSEPH K J,PAL A,RAJANALA S,et al.C4Synth:cross caption cycle-consistent text-to-image synthesis[C]//IEEE Winter Conference on Applications of Computer Vision,2018:358-366.
[75] LI Y,GAN Z,SHEN Y,et al.StoryGAN:a sequential conditional GAN for story visualization[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2019:6329-6338.
[76] CHENG J,WU F,TIAN Y,et al.RifeGAN:rich feature generation for text-to-image synthesis from prior knowledge[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2020:10911-10920.
[77] HAN F,GUERRERO R,PAVLOVIC V.CookGAN:causality based text-to-image synthesis[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision,2020:5519-5527.
[78] WADHAWAN R,DRALL T,SINGH S,et al.Multi-attributed and structured text-to-face synthesis[C]//International Conference on Technology Engineering Management for Societal Impact Using Marketing Entrepreneurship and Talent,2020.
[79] WAH C,BRANSON S,WELINDER P,et al.The caltech-UCSD Birds-200-2011 dataset:technical report CNS-TR-2011-001[R].California Institute of Technology,2011.
[80] NILSBACK M E,ZISSERMAN A.Automated flower classification over a large number of classes[C]//2008 Sixth Indian Conference on Computer Vision Graphics and Image Processing I,2008:722-729.
[81] LIN T Y,MAIRE M,BELONGIE S,et al.Microsoft COCO:common objects in context[C]//European Conference on Computer Vision.Cham:Springer,2014:740-755.
[82] LIU Z,LUO P,WANG X,et al.Deep learning face attributes in the wild[C]//International Conference on Computer Vision,2015:3730-3738.
[83] SALVADOR A,HYNES N,AYTAR Y,et al.Learning cross-modal embeddings for cooking recipes and food images[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2017:3020-3028.
[84] SALIMANS T,GOODFELLOW I,ZAREMBA W,et al.Improved techniques for training GANs[C]//Advances in Neural Information Processing Systems,2016.
[85] SZEGEDY C,VANHOUCKE V,IOFFE S,et al.Rethinking the inception architecture for computer vision[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2016:2818-2826.
[86] HEUSEL M,RAMSAUER H,UNTERTHINER T,et al.GANs trained by a two time-scale update rule converge to a local Nash equilibrium[C]//Advances in Neural Information Processing Systems,2017:6626-6637.
[87] ARJOVSKY M,CHINTALA S,BOTTOU L.Wasserstein GAN[C]//International Conference on Machine Learning,2017:214-223.
[88] RANOM J,PEYRE G,DELON J,et al.Wasserstein barycenter and its application to texture mixing[C]//International Conference on Scale Space and Variational Methods in Computer Vision,2011:435-446.
[89] SHMELKOV K,SCHMID C,ALAHARI K.How good is my GAN?[C]//Proceedings of the European Conference on Computer Vision,2018:213-229.
[90] PAPINENI K,ROUKOS S,WARD T,et al.BLEU:a method for automatic evaluation of machine translation[C]//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics,2002:311-318.
[91] LAVIE A,GARWAL A.METEOR:an automatic metric for MT evaluation with high levels of correlation with human judgments[C]//Proceedings of the Second Workshop on Statistical Machine Translation,2007.
[92] VEDANTAM R,LAWRENCE ZITNICK C,PARIKH D.CIDEr:consensus-based image description evaluation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2015.
[93] ROMBACH R,ESSER P,OMMER B.Network-to-network translation with conditional invertible neural networks[C]//Advances in Neural Information Processing Systems,2020:2784-2797.
[94] BROCK A,DONAHUE J,SIMONYAN K.Large scale GAN training for high fidelity natural image synthesis[C]//International Conference on Learning Representations,2018.
[95] MAO Q,LEE H Y,TSENG H Y,et al.Mode seeking generative adversarial networks for diverse image synthesis[C]//Proceedings of the IEEE Computer Vision and Pattern Recognition,2019:1429-1437.
[96] CHA M,GWON Y,KUNG H T.Adversarial learning of semantic relevance in text to image synthesis[C]//Proceedings of the AAAI Conference on Artificial Intelligence,2018:3272-3279.
[97] MENICK J,KALCHBRENNER N.Generating high fidelity images with subscale pixel networks and multidimensional upscaling[C]//International Conference on Learning Representations,2019.
[98] YUAN M,PENG Y.CKD:cross-task knowledge distillation for text-to-image synthesis[J].IEEE Transactions on Multimedia,2019,22(8):1955-1968.
[99] CHEN M,RADFORD A,CHILD R,et al.Generative

pretraining from pixels[C]//International Conference on Machine Learning,2020:1691-1703.
[100] RAMESH A,PAVLOV M,GOH G,et al.Zero-shot text-to-image generation[C]//International Conference on Machine Learning,2021:8821-8831.
[101] LIN T Y,GOYAL P,GIRSHICK R,et al.Focal loss for dense object detection[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2020,42:318-327.
[102] KRISHNA R,ZHU Y,GROTH O,et al.Visual genome:connecting language and vision using crowdsourced dense image annotations[J].International Journal of Computer Vision,2017,123(1):32-73.
[103] PAREKH Z,BALDRIDGE J,CER D,et al.Crisscrossed captions:extended intramodal and intermodal semantic similarity judgments for MS-COCO[C]//Proceedings of Conference of the European Chapter of the Association for Computational Linguistics,2021:2855-2870.
[104] RAVURI S V,VINYALS O.Classification accuracy score for conditional generative models[C]//Advances in Neural Information Processing Systems,2019:12268-12279.
[105] SHARMA P,DING N,GOODMAN S,et al.Conceptual captions:a cleaned hypernymed image alt-text dataset for automatic image captioning[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics(Volume 1:Long Papers),2018:2556-2565.
[106] BORJI A.Pros and cons of GAN evaluation measures[J].Computer Vision and Image Understanding,2019,179:41-65.
[107] DENG J,DONG W,SOCHER R,et al.ImageNet:a large-scale hierarchical image database[C]//IEEE Conference on Computer Vision and Pattern Recognition,2009:248-255.
[108] ZHANG R,ISOLA P,EFROS A A,et al.The unreasonable effectiveness of deep features as a perceptual metric[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2018:586-595.
[109] HUANG W,XU Y,OPPERMANN I.Realistic image generation using region-phrase attention[C]//Asian Conference on Machine Learning,2019:284-299.
[110] LIANG J,PEI W,LU F.CPGAN:content-parsing generative adversarial networks for text-to-image synthesis[C]//European Conference on Computer Vision.Cham:Springer,2020:491-508.
[111] ZHOU S,GORDON M,KRISHNA R,et al.HYPE:a benchmark for human eye perceptual evaluation of generative models[C]//Advances in Neural Information Processing Systems,2019:3449-3461.
[112] DONG H,YU S,WU C,et al.Semantic image synthesis via adversarial learning[C]//IEEE International Conference on Computer Vision,2017:5706-5714.
[113] LIU Y,DE NADAI M,CAI D,et al.Describe what to change:a text-guided unsupervised image-to image translation approach[C]//Proceedings of the ACM International Conference on Multimedia,2020:1357-1365.
[114] ZHU D,MOGADALA A,KLAKOW D.Image manipulation with natural language using two-sided attentive conditional generative adversarial network[J].Neural Networks,2021,136:207-217.
[115] WANG X,QIAO T,ZHU J,et al.S2IGAN:speech-to-image generation via adversarial learning[C]//Proceedings of Interspeech,2020:2292-2296.
[116] BALAJI Y,MIN M R,BAI B,et al.Conditional GAN with discriminative filter generation for text-to-video synthesis[C]//International Joint Conference on Artificial Intelligence,2019.
[117] DENG K,FEI T,HUANG X,et al.IRC-GAN:introspective recurrent convolutional GAN for text-to-video generation[C]//International Joint Conference on Artificial Intelligence,2019:2216-2222.
[118] CHOI H S,PARK C D.From inference to generation:end-to-end fully self-supervised generation of human face from speech[C]//International Conference on Learning Representations,2020.
[119] JIA Y,WEISS R J,BIADSY F,et al.Direct speech-to-speech translation with a sequence-to-sequence model[C]//Interspeech,2019.
[120] SURIS D,RECASENS A,BAU D,et al.A learning words by drawing images[C]//Proceedings of the IEEE Computer Vision and Pattern Recognition,2019:2029-2038.
[121] LI Y,MIN M R,SHEN D,et al.Video generation from text[C]//Conference on Artificial Intelligence,2018:7065-7072.