第 58 卷 第 4 期/2021 年 2 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

结合场景描述的文本生成图像方法

黄友文，周斌*，唐欣

江西理工大学信息工程学院，江西 赣州 341000

摘要 研究基于场景描述文本生成对应图像的方法，针对生成图像常常出现的对象重叠和缺失问题，提出了一种结合场景描述的生成对抗网络模型。首先，利用掩模生成网络对数据集进行预处理，为数据集中的对象提供分割掩模向量。然后，将生成的对象分割掩模向量作为约束，通过描述文本训练布局预测网络，得到各个对象在场景布局中的具体位置和大小，并将结果送入到级联细化网络模型，完成图像的生成。最后，将场景布局与图像共同引入到布局鉴别器中，弥合场景布局与图像之间的差距，得到更加真实的场景布局。实验结果表明，所提模型能够生成与文本描述更匹配的图像，图像更加自然，同时有效地提高了生成图像的真实性和多样性。

关键词 图像处理；图像生成；生成对抗网络；场景描述；分割掩模；场景布局

中图分类号 TP391.4 **文献标志码** A **doi**：10.3788/LOP202158.0410012

Text Image Generation Method with Scene Description

Huang Youwen, Zhou Bin*, Tang Xin

School of Information Engineering, Jiangxi University of Science and Technology, Ganzhou, Jiangxi 341000, *China*

Abstract In this paper, a method of generating corresponding images based on scene description text is studied, and a generative adversarial network model combined with scene description is proposed to solve the object overlapping and missing problems in the generated images. Initially, a mask generation network is used to preprocess the dataset to provide objects in the dataset with segmentation mask vectors. These vectors are used as constraints to train a layout prediction network by text description to obtain the specific location and size of each object in the scene layout. Then, the results are sent to the cascaded refinement network model to complete image generation. Finally, the scene layout and images are introduced to a layout discriminator to bridge the gap between them for obtaining a more realistic scene layout. The experimental results demonstrate that the proposed model can generate more natural images that better match the text description, effectively improving the authenticity and diversity of generated images.

Key words image processing; image generation; generative adversarial network; scene description; segmentation mask; scene layout

OCIS codes 100.3010; 100.2980; 150.1135

1 引 言

计算机视觉应用于图像生成、语义分割[1-2]、目标检测[3]等诸多领域。其中通过自然语言描述引导图像生成一直都是图像生成领域的挑战性任务，在实际应用中有广泛的需求，例如广告设计、帮助警察追寻嫌疑犯、漫画生成等。近年来，深度学习的出现促进了自然语言描述引导图像生成的发展，并且已经得到了很大的进展。

现阶段，生成对抗网络(GAN)[4]在图像生成领域已经得到了广泛应用[5]。文本描述引导图像生成是近几年的热门研究领域之一，其主要的任务就是

收稿日期：2020-06-30；**修回日期**：2020-07-28；**录用日期**：2020-08-07

基金项目：江西省教育厅科技项目(GJJ180443)

* **E-mail**：zhoubin_master@163.com

通过一段文本描述生成一张与描述内容相互对应的图片。文本描述引导图像生成方法主要利用 GAN 的原理来完成图像的生成工作。

起初，Reed 等[6]提出 GAN-INT-CLS 网络，GAN-INT-CLS 以条件生成对抗网络(CGAN)[7]为模型主干，将文本描述编码为全局向量，并将其作为生成器和鉴别器的约束。GAN-INT-CLS 有效地生成了分辨率为 64×64 的可信赖图像，但是图像缺少生动的对象细节。随后，Zhang 等[8]为了生成高分辨率的图像，提出了分阶段的堆栈生成对抗网络(StackGAN)模型，StackGAN 的训练策略是先通过文本描述生成包含基本形状、颜色的 64×64 低分辨率图像，再利用生成的低分辨率图像和文本描述修补丢失的细节信息，最后生成 256×256 高分辨率图像。在后续工作中，Zhang 等[9]提出了一种端到端的堆栈生成对抗网络(StackGAN-v2)，StackGAN-v2 将 GAN 扩展成树状结构，利用多个生成器和多个鉴别器进行并行训练，稳定地完成不同分辨率(如 64×64、128×128、256×256)图像的生成。继 StackGAN-v2 之后，Xu 等[10]又在此基础上提出了注意生成对抗网络(AttnGAN)，AttnGAN 在 StackGAN-v2 的基础上增加了注意力机制，着重关注文本描述中的相关单词，并将其编码为单词向量输入到网络模型中，生成器和鉴别器对最相关的单词向量进行精准优化，有效地生成了 256×256 高质量图像。然而，AttnGAN 在处理有多个交互对象的复杂场景时，就会显得十分困难。而后，Johnson 等[11]提出了一种利用场景图生成图像的模型(Sg2im)。Sg2im 通过场景图推断出对象及其关系，对所获得的对象及其关系预测出对象的边界框和分割掩模，得到一个关于文本描述的场景布局，接着将场景布局输入到后续的生成网络中生成相互对应的图像。在复杂场景下，Sg2im 生成的图像更能反映文本描述内容，但是，结果中存在伪影、对象重叠、对象缺失等问题。

为了进一步解决生成图像中伪影、对象重叠、对象缺失的问题，本文在从场景图生成图像的网络模型的基础上提出了一种结合场景描述的 GAN 模型。该模型引入了布局鉴别器[12]，重点关注场景布局与图像之间的差距，弥合此差距，预测出更真实的场景布局，缓解生成图像中出现的伪影、对象缺失现象；同时引入掩模生成网络对数据集进行预处理，生成对象分割掩模向量，将对象分割掩模向量作为约束，通过描述文本训练布局预测网络，从而能更精确地预测出各个对象在场景布局中的具体位置和大小，改善生成图像中出现的多个对象相互重叠现象，提高生成图像的质量。

2 结合场景描述的生成对抗网络

2.1 生成对抗网络

GAN 启发自博弈论中的二人零和博弈[13]。GAN 模型的两位博弈方由生成器 G 和鉴别器 D 充当。生成器 G 的目的是生成看起来真实、与真实样本相似的图像，鉴别器 D 的目的在于尽量分辨输入图像是真实的还是由生成器生成的。在训练过程中，对于优化生成器 G、固定鉴别器 D，生成器 G 捕捉先验分布噪声样本 z，生成类似真实训练样本数据 P_{data}。对于优化鉴别器 D、固定生成器 G，其中鉴别器 D 为一个二分类器，鉴别器 D 尽可能准确区分出真实样本数据 x 和生成样本数据 P_{data}。生成器 G 和鉴别器 D 互相对抗训练的表达式为

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

式中：$V(\cdot)$ 为价值函数。

2.2 结合场景描述的生成对抗网络

为了进一步解决生成图像中伪影、对象重叠、对象缺失等问题，从文本描述中准确地生成相对应的图像，提高生成图像的质量。因此，提出了一种结合场景描述的 GAN 模型，通过图像生成网络 f，将场景图 $\boldsymbol{g}$ 转换为图像，具体如图 1 所示。

首先，场景图 $\boldsymbol{g}$ 由图卷积网络 G_{graph} 处理，图卷积网络 G_{graph} 为每个对象提供嵌入向量，在预处理过程中，由掩模生成网络 G_{mask} 处理真实图像 $\boldsymbol{I}'$，并为每个对象提供分割掩模向量。其次，将每个对象的嵌入向量输入到布局预测网络 G_{layout}，其中 G_{layout} 采用文献[11]中的网络结构，包含 box 回归网络和 mask 回归网络。box 回归网络由多层感知器构成，mask 回归网络由若干个以 Sigmoid 为激活函数的转置卷积构成。接着，将预处理中每个对象的分割掩模向量作为约束训练 G_{layout}，得到对象的预测边界框和分割掩模，以形成图像的预测场景布局。最后，将预测场景布局输入到级联细化网络 G_{refine}[14]，完成图像 $\boldsymbol{I}$ 的生成。

2.3 掩模生成网络

对象掩模信息能够用于场景布局的训练，提高布局准确性。然而，标注对象信息时需要耗费大量的人力物力，因而常用的训练数据集，如 Visual Genome[15]等就没有对象分割掩模信息。为了获取此信息，文献[11]中的生成模型通过一个 mask 回

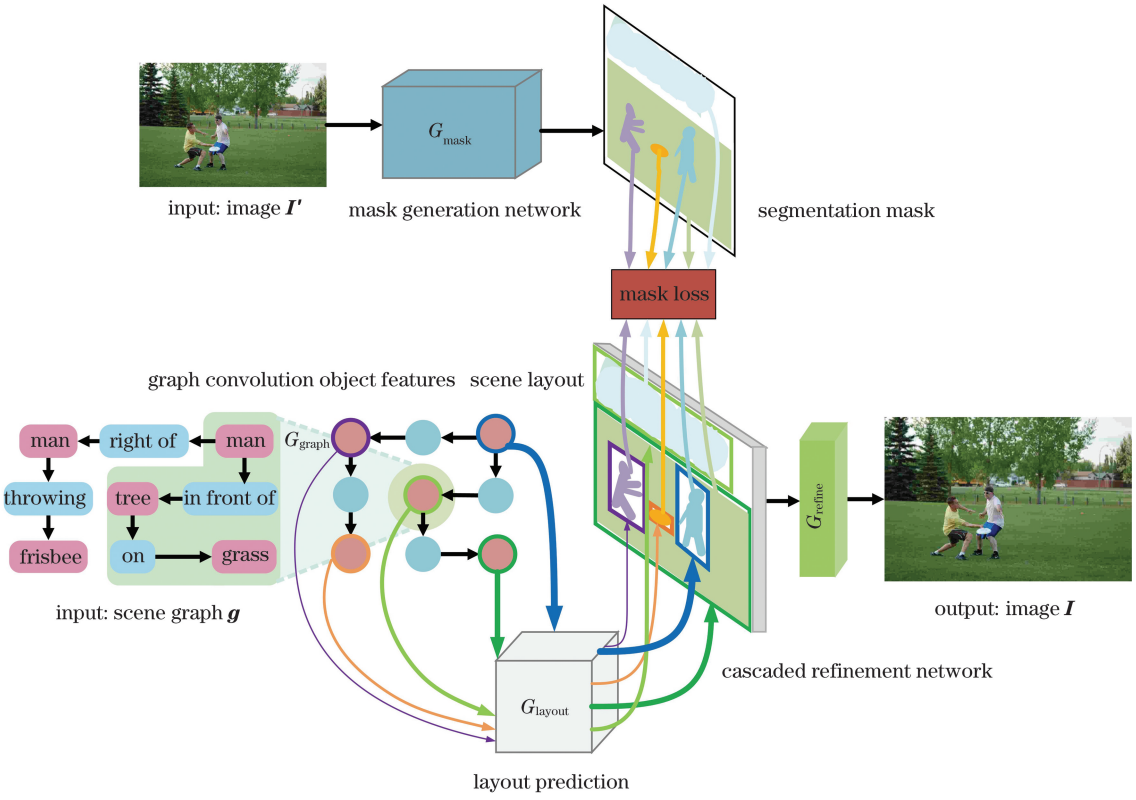

图 1　生成网络模型

Fig. 1　Generation network model

归网络预测图像中每个对象的分割掩模向量，利用布局预测网络预测场景布局。此方法的网络结构简单，没有直接的对象分割信息对网络结构进行约束训练，导致预测的场景布局不够准确，生成图像容易出现对象重叠的现象。基于上述问题，引入图 2 所示的基于 Mask^{x} R-CNN[16] 的掩模生成模型对数据集进行预处理，为数据集中的对象提供分割掩模向量，并将其作为约束训练布局预测网络，从而更精准地预测场景布局，同时免去了费时费力的人工标注过程。

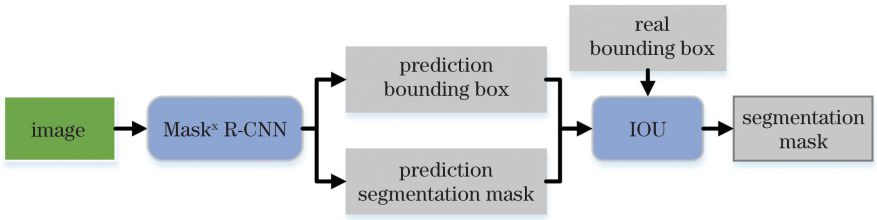

图 2　掩模生成网络

Fig. 2　Mask generation network

该掩模生成网络模型 G_{mask} 包含 2 个部分，如图 2 所示，即 Mask^{x} R-CNN 和 IOU（交并比）筛选单元。

掩模生成网络模型 G_{mask} 的具体过程如图 2 所示。首先，给出真实图像 $\boldsymbol{I}'$，$\boldsymbol{I}'$ 为维度为 $H\times W\times 3$ 的矩阵向量，将 Mask^{x} R-CNN 作用于图像矩阵向量 $\boldsymbol{I}'$。其中，Mask^{x} R-CNN 是由 Hu 等[16] 在 Mask R-CNN[17] 基础上提出的一种新型半监督实例分割方法。该方法使用两个数据集，一个包含掩模标注和框标注，另一个只包含框标注。具体过程是先对图像 $\boldsymbol{I}'$ 进行特征提取，得到图像 $\boldsymbol{I}'$ 的特征图；紧接着通过卷积网络对得到的特征图提取感兴趣区域，并将其传入 box head 网络和 mask head 网络中生成边界框向量和掩模向量，通过权重进行调节，以此得到每个对象的边界框 b_i 和分割掩模向量 $\boldsymbol{m}_i$，该分割掩模向量 $\boldsymbol{m}_i$ 的维度为 $M\times M$，其元素在(0,1)之间；最后，将 Mask^{x} R-CNN 预测的边界框 b_i、分割掩模向量 $\boldsymbol{m}_i$ 及真实边界框 b'_i 传入 IOU 筛选单元。在 IOU 筛选单元中，对 b_i 与 b'_i 进行筛选，表达式为

$$f_{\mathrm{IOU}}=\frac{(x_1-x_0)\cdot(y_1-y_0)\bigcap(x_3-x_2)\cdot(y_3-y_2)}{(x_1-x_0)\cdot(y_1-y_0)\bigcup(x_3-x_2)\cdot(y_3-y_2)},\tag{2}$$

式中：$(x_0,\ y_0)$、$(x_1,\ y_1)$为预测边界框 b_i 的坐标值；$(x_2,\ y_2)$、$(x_3,\ y_3)$为真实边界框 b'_i 的坐标值。若 b_i 与 b'_i 计算得到的 f_{IOU} 大于阈值 t，就输出对象的边界框 b_i 对应的分割掩模向量 $\boldsymbol{m}_i$。

2.4　鉴别器

所提模型中的鉴别网络模型架构如图 3 所示，包含 3 个鉴别器，即布局鉴别器 D_{layout}、图像鉴别器 D_{img}、对象鉴别器 D_{obj}。通过对鉴别器网络 D_{layout}、D_{img} 及 D_{obj} 进行联合训练，图像 $\boldsymbol{I}$ 看起来既逼真、清晰，又包含可识别的对象。

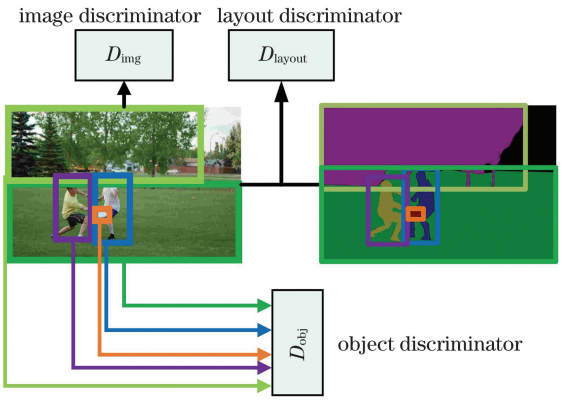

图 3　鉴别网络模型

Fig. 3　Discrimination network model

图像鉴别器 D_{img} 和对象鉴别器 D_{obj} 都是采用文献[11]中的网络结构，其中图像鉴别器 D_{img} 的目的是保证生成的图像具有真实的整体外观，对象鉴别器 D_{obj} 的目的是保证生成图像中的每个对象看起来都是真实的。布局鉴别器 D_{layout} 的任务是判断场景布局是真实还是伪造的。文献[11]中的网络模型所生成的图像会出现伪影、对象缺失现象，因此，所提鉴别网络模型引入了布局鉴别器 D_{layout}，其输入是间隔规则的场景布局块与图像块。

将生成网络与布局鉴别器 D_{layout} 联合进行训练，重点关注并弥合场景布局与图像之间的差距，引导生成网络预测出更加真实的场景布局，有效地缓解生成图像中的伪影、对象缺失现象，从而生成更加清晰且一致的图像。

所提模型中的布局鉴别器由一个带有实例归一化(IN)的多尺度鉴别器[12,18]构成。这个多尺度鉴别器包含全尺度鉴别器和半尺度鉴别器。若仅使用全尺度鉴别器，能够改善伪影现象，但生成的图像缺少更加精细的细节；反之，仅使用半尺度鉴别器，能够缓解对象缺失现象，但生成的图像往往缺失全局一致感。因此，将全尺度鉴别器与半尺度鉴别器组合对生成器进行约束，能够得到更高质量的图像。

全尺度鉴别器与半尺度鉴别器具有图 4 所示的相同网络结构，不同之处在于采用具有不同分辨率的输入，即在两个不同的尺度下工作。首先，将场景布局和图像共同输入到全尺度鉴别器中；其次，同时对场景布局和图像进行下采样，下采样系数为 2；将采样后的场景布局和图像引入到半尺度鉴别器中，从而使布局鉴别器 D_{layout} 在两个不同尺度下能区分场景布局与真实图像。

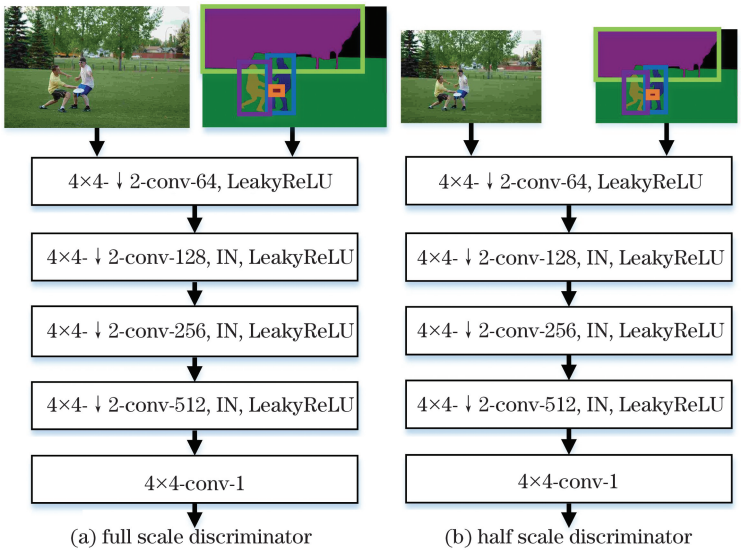

图 4　布局鉴别器

Fig. 4　Layout discriminator

总的说来，生成网络通过鉴别器网络进行对抗性训练，以缓解模型所生成图像出现伪影、对象缺失现象，生成更加清晰并且视觉效果好的图像。因此，布局鉴别器的损失是复合损失，复合损失函数为

$$L_{\text{D-layout}}=L_{\text{real}}-L_{\text{fake-img}}-L_{\text{fake-layout}}, \tag{3}$$

式中：L_{real} 为真布局对抗损失；$L_{\text{fake-img}}$ 为假图像对抗损失；$L_{\text{fake-layout}}$ 为假布局对抗损失。具体的损失函数为

$$L_{\text{real}}=\log D_{\text{layout}}(\boldsymbol{t}',\boldsymbol{I}'), \tag{4}$$

$$L_{\text{fake-img}}=\log\left[1-D_{\text{layout}}(\boldsymbol{t}',\boldsymbol{I})\right], \tag{5}$$

$$L_{\text{fake-layout}}=\log\left[1-D_{\text{layout}}(\boldsymbol{t},\boldsymbol{I}')\right], \tag{6}$$

式中：$\boldsymbol{t}'$ 为根据图像中每个对象的真实边界框与预处理的分割掩模得到的场景布局；$\boldsymbol{t}$ 为根据图像中每个对象的预测边界框与预测分割掩模生成的场景布局。

3 实验及结果分析

实验平台的配置为 Intel Xeon 3104 处理器、16GB 内存、GTX2080Ti 显卡，并使用 64 位操作系统 Ubuntu 18.04 和 Pytorch 深度学习框架。

3.1 数据集

实验采用 Visual Genome 数据集 1.4 版[15]对所提网络模型进行评估。Visual Genome 数据集包含了 108077 张带有场景图注释的图像，其中 80% 的图像数据作为训练集、10% 的图像数据作为验证集、10% 的图像数据作为测试集。

对 Visual Genome 数据集[15]进行预处理，第一，为了使网络充分学习到对象类别和关系类型的特征，只保留训练集中至少出现 500 次的对象类别和至少出现 2000 次的关系类型；第二，将掩模生成网络得到的对象掩模边界框与真实边界框之间的 IOU 作为筛选标准，设定阈值 t，只保留 IOU 大于 t 的对象。显然，阈值 t 越大，掩模信息越准确，对模型的约束效果也越好，但是阈值 t 过大会导致保留的对象太少，对应的对象类别和关系类型也越少。采用不同阈值对数据集进行预处理的结果如表 1 所示。

表 1 不同阈值下的预处理结果

Table 1 Preprocessing results under different threshold values

t	0.3	0.4	0.5	0.6	0.7
Number of objects	156	151	146	127	103
Number of relationship types	38	38	37	30	24

从表 1 中可以看出：相较于 t 为 0.3、0.4 的预处理结果，当阈值 t 为 0.5 时，对象类别数量仅分别减少了 10 个、5 个，关系类型都仅减少了 1 种，这表明阈值为 0.5 时的图像多样性并无太大差别，但掩模信息的准确度提升较大；当阈值 t 为 0.6、0.7 时，虽然掩模信息的准确度获得提升，但是对象类别数量却分别减少了 19 个和 43 个，关系类型分别减少了 7 种和 13 种，这表明 t 为 0.6、0.7 时，图像的多样性较差。因此，选取阈值 t 为 0.5 的网络模型，保留 146 个对象类别和 37 种关系类型。

与此同时，数据集中保留至少有一个关系类型的图像并剔除过小的对象，经过筛选后，训练集剩余 55776 张图像、验证集剩余 4324 张图像、测试集剩余 436 张图像，其中每个图像平均有 10 个物体和 5 种关系类型。

3.2 实验评估指标及结果

为了证明所增加组件的有效性，在对象数量基本相同的情况下，对所提模型、文献[11]中的网络模型 Sg2im、StackGAN[8]及 AttnGAN[10]进行实验对比。采用 IS(inception score)[19]和 FID(Fréchet inception distance)[20]为定量评估指标，结果如表 2 所示。其中 IS 评估指标主要是衡量模型生成图像的多样性，IS 值越大，生成图像的多样性越好；FID 评估指标主要衡量模型生成图像的质量，FID 值越小，生成图像的质量越好。

表 2 IS 和 FID 指标值的对比结果

Table 2 Comparison results of IS and FID values

Model	IS	FID
Real image(64×64)	13.90±0.50	0
Proposed model(no D_{layout})	6.72±0.24	57.48
Proposed model(no G_{mask})	6.69±0.14	61.34
Proposed model(full model)	7.11±0.14	42.20
Sg2im[11]	6.30±0.20	73.39
StackGAN[8]	6.35±0.16	108.68
AttnGAN[10]	6.38±0.22	96.40

表 2 中，no D_{layout} 表示所提模型中去除布局鉴别器 D_{layout}，no G_{mask} 表示所提模型中去除掩模生成网络 G_{mask}，full model 表示所提完整模型。从表 2 可以看出，相较于 Sg2im[11]、StackGAN[8]、AttnGAN[10]，所提模型的 IS 值分别提高了 0.81、0.76、0.73，FID 值分别降低了 31.19、66.48、

54.20。同样地，相较于 Sg2im，所提模型分别在去除布局鉴别器 D_{layout} 和掩模生成网络 G_{mask} 时，IS 值分别提高了 0.42、0.39，FID 值分别降低了 15.91、12.05。结果进一步说明，布局鉴别器 D_{layout} 与掩模生成网络 G_{mask} 对提升图像的生成质量和多样性产生了积极作用。

在前述实验平台下对各模型的图像生成时间进行对比分析，结果如表 3 所示。

表 3　图像生成时间的对比结果

Table3　Comparison results of image generation time

Model	Proposed model(full model)	Sg2im[11]	StackGAN[8]	AttnGAN[10]
Time/s	0.0278	0.0216	0.0634	0.0302

从表 3 可以看出：所提模型与 AttnGAN 模型、Sg2im 模型的图像生成时间相差不大；相比采用两阶段训练的 StackGAN 模型，采用端到端训练的所提模型的图像生成时间更短。这表明，与采用两阶段方式训练的模型相比，采用端到端方式训练的模型运行速度更快。

采用相同文本描述生成了对应的图像，对表 2 中的模型进行定性比较，结果如图 5 所示。

图 5　相同描述的结果对比

Fig.5　Comparison results of same description

所提模型和 Sg2im 生成的图像的最终输出分辨率都为 64×64，StackGAN 和 AttnGAN 生成的都是分辨率为 256×256 图像，利用下采样得到分辨率为 64×64 的图像。从图 5 可以看出：StackGAN 和 AttnGAN 生成的结果与文本描述差别较大，表明这两种模型都难以处理多个对象之间的关系。与 Sg2im 生成的结果相比，所提模型在对象关系的处理方面更强，能够更加精确地预测出各个对象的具体位置和大小，改善了对象重叠现象。例如，在图 5(a)中，文本描述为“斑马在草地上”，Sg2im 中

的斑马位置出现错误；在图 5(b)中，相较于 Sg2im 模型，所提模型得到的图像中，人和木板的位置和大小较为精确；图 5(d)主要表现对象重叠现象，相较于 Sg2im 模型，所提模型中两辆公共汽车未发生重叠，更加接近文本描述。从图 5 还能够看出，所提模型在缓解伪影、对象缺失现象方面的能力要强于 Sg2im。例如，图 5(c)中文本描述所提到的山和图 5(h)中所提到的杯子和玻璃，在所提模型生成的图像中较为完整，没有发生缺失现象；从图 5(e)～(g)可以看出，所提模型缓解了伪影现象，生成的结果较为完整，对象一致并且图像更加清晰。

接着，在逐步增加对象时，对所提模型与 Sg2im 生成的图像进行比较，结果如图 6 所示。

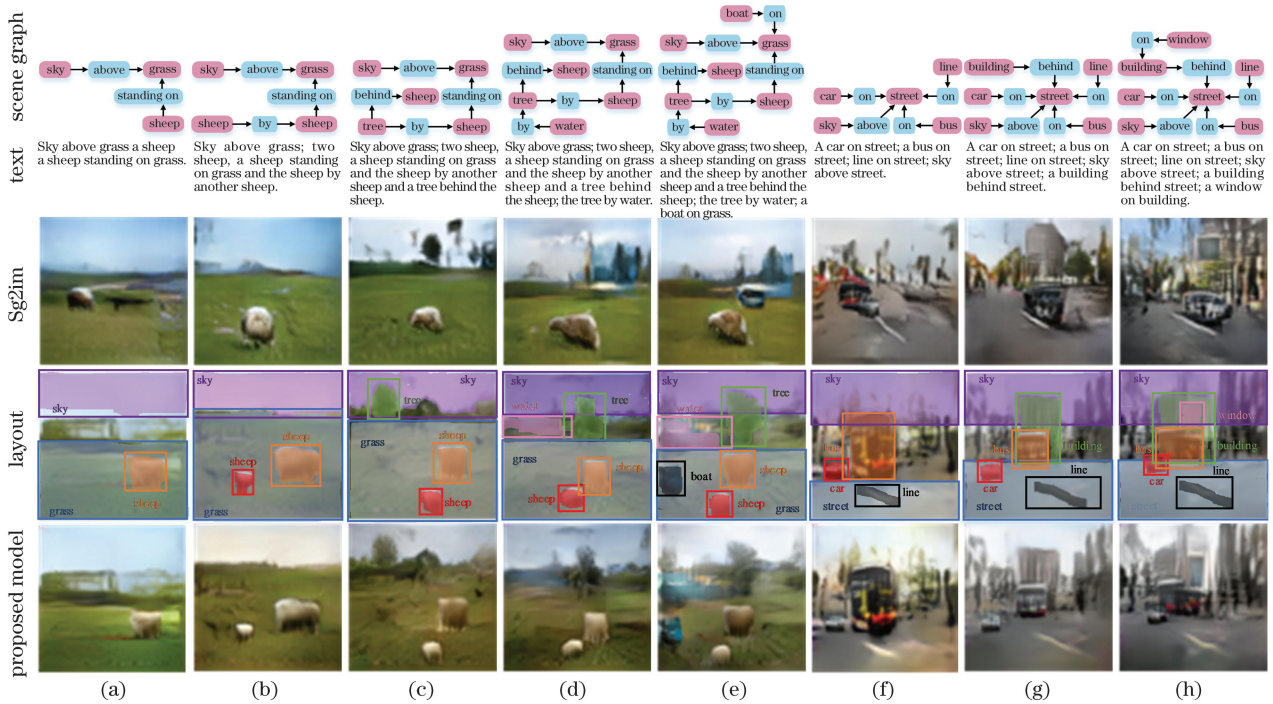

图 6　增加对象后的结果对比

Fig. 6　Comparison results after adding objects

从图 6 可以看出，当文本描述中出现较多对象时，所提模型生成的图像更加真实且一致，没有出现对象缺失或重叠现象。例如，在图 6(b)～(d)中，相较于 Sg2im 模型，所提模型结果中的两只绵羊都未发生重叠，并且图 6(c)、(d)中水的位置和大小更加精确；图 6(g)、(h)中文本描述所提到的汽车，在所提模型生成的图像中较为完整，没有发生缺失现象。

另外，在输入相同的图像时，对所提模型和 Sg2im 模型生成的预测掩模结果进行了比较，结果如图 7 所示。

从图 7 可以看出，本文引入 Mask^{x} R-CNN 对数据进行预处理后，训练布局预测网络，用该网络最终生成的掩模位置和大小更加精确。例如，在图 7(a)中，相较于 Sg2im 模型，所提模型结果中的人和飞机掩模都预测准确，并且飞机的位置和大小更加精确；同时，图 7(b)～(f)中文本描述所提到的对象，所提模型生成的预测掩模更为完整和准确。

4　结　　论

现有的基于场景图的图像生成模型常常会出现伪影、对象重叠、对象缺失等问题，导致生成的图像质量不高，因此，提出了一种结合场景描述的生成对抗网络模型。该模型引入掩模生成网络对数据集进行预处理，并引入布局鉴别器，改善了生成图像的质量，使生成的图像更接近真实图像。实验结果表明，在相同的数据集上，所提模型提高了文本描述中各个对象位置的准确度，预测出了更加真实的场景布局，可以有效地改善生成图像中出现多个对象是相互重叠、对象缺失的问题，缓解了伪影问题，生成更加逼真的图像。

图 7 预测掩模的结果对比

Fig. 7 Comparison results of predicted mask

参 考 文 献

[1] Yue S Y. Image semantic segmentation based on hierarchical context information [J]. Laser & Optoelectronics Progress, 2019, 56(24): 241005.
岳师怡. 基于多层级上下文信息的图像语义分割[J]. 激光与光电子学进展, 2019, 56(24): 241005.

[2] Li L K, Lu C H, Zou B. Research on target detection and feasible region segmentation based on deep learning [J]. Laser & Optoelectronics Progress, 2020, 57(12): 121013.
李立凯, 卢炽华, 邹斌. 基于深度学习的目标检测与可行域分割研究[J]. 激光与光电子学进展, 2020, 57(12): 121013.

[3] Liu W J, Wang F, Qu H C. Object detection model based on multi-scale feature integration[J]. Laser & Optoelectronics Progress, 2019, 56(23): 231007.
刘万军, 王凤, 曲海成. 融合多尺度特征的目标检测模型[J]. 激光与光电子学进展, 2019, 56(23): 231007.

[4] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]//Proceedings of the 27th International Conference on Neural Information Prccessing Systems, December 7-12, 2014, Montreal, Quebec, Canada. New York: Curran Associates, 2014, 2: 2672-2680.

[5] Chen X F, Shen H J, Bian Q, et al. Face image super-resolution with an attention mechanism [J]. Journal of Xidian University (Natural Science), 2019, 46(3): 148-153.
陈晓范, 申海杰, 边倩, 等. 结合注意力机制的人脸超分辨率重建[J]. 西安电子科技大学学报(自然科学版), 2019, 46(3): 148-153.

[6] Reed S, Akata Z, Yan X C, et al. Generative adversarial text to image synthesis[EB/OL]. (2016-06-05) [2020-05-28]. https://arxiv.org/abs/1605.05396.

[7] Mirza M, Osindero S. Conditional generative adversarial nets [EB/OL]. (2014-11-06) [2020-05-28]. https://arxiv.org/abs/1411.1784.

[8] Zhang H, Xu T, Li H S, et al. StackGAN: text to photo-realistic image synthesis with stacked generative adversarial networks [C]//2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 5908-5916.

[9] Zhang H, Xu T, Li H S, et al. StackGAN++: realistic image synthesis with stacked generative adversarial networks [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(8): 1947-1962.

[10] Xu T, Zhang P C, Huang Q Y, et al. AttnGAN: fine-grained text to image generation with attentional generative adversarial networks [C]//2018 IEEE/

CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 1316-1324.

[11] Johnson J, Gupta A, Fei-Fei L. Image generation from scene graphs[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 1219-1228.

[12] Ashual O, Wolf L. Specifying object attributes and relations in interactive scene generation[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27 - November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2019: 4560-4568.

[13] Luo J, Huang J Y. Generative adversarial network: an overview[J]. Chinese Journal of Scientific Instrument, 2019, 40(3): 74-84.
罗佳, 黄晋英. 生成式对抗网络研究综述[J]. 仪器仪表学报, 2019, 40(3): 74-84.

[14] Chen Q F, Koltun V. Photographic image synthesis with cascaded refinement networks[C]//2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 1520-1529.

[15] Krishna R, Zhu Y K, 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.

[16] Hu R H, Dollár P, He K M, et al. Learning to segment every thing[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 4233-4241.

[17] He K M, Gkioxari G, Dollár P, et al. Mask R-CNN[C]//2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 2980-2988.

[18] Wang T C, Liu M Y, Zhu J Y, et al. High-resolution image synthesis and semantic manipulation with conditional GANs[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 8798-8807.

[19] Salimans T, Goodfellow I, Zaremba W, et al. Improved techniques for training GANs[EB/OL]. (2016-06-10)[2020-05-28]. https://arxiv.org/abs/1606.03498.

[20] Heusel M, Ramsauer H, Unterthiner T, et al. GANs trained by a two time-scale update rule converge to a local nash equilibrium[EB/OL]. (2018-01-12)[2020-05-28]. https://arxiv.org/abs/1706.08500.