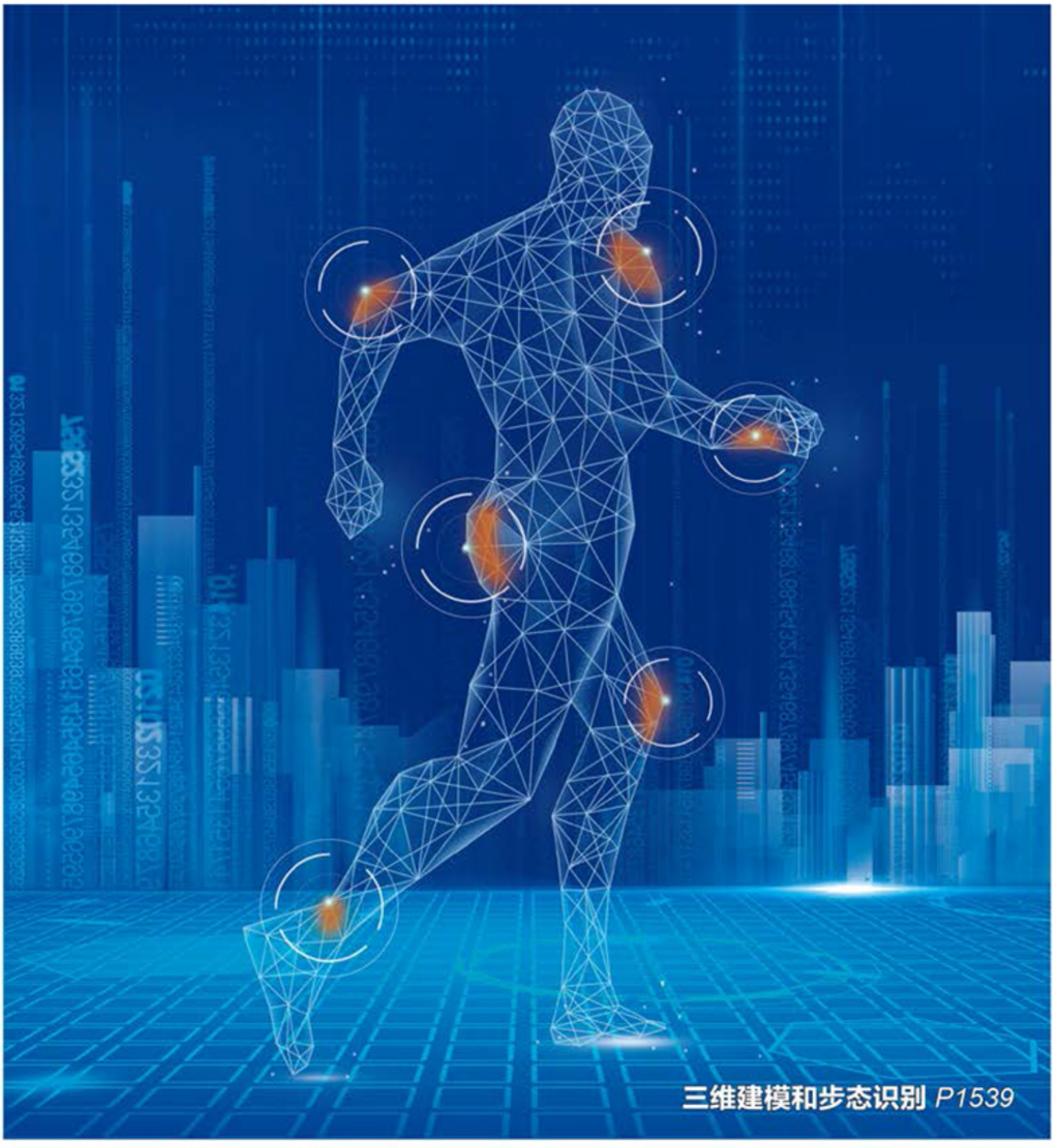
JOURNAL OF IMAGE AND GRAPHICS

主办：中国科学院遥感与数字地球研究所
中国图象图形学学会
北京应用物理与计算数学研究所

中国图象图形学报

2020
08
VOL.25

ISSN1006-8961
CN11-3758/TB



三维建模和步态识别 *P1539*

中国图象图形学报

刊名题字：宋健 | 月刊（1996年创刊）

第25卷第8期（总第292期）
2020年8月16日

中国精品科技期刊
中国国际影响力优秀学术期刊
中国科技核心期刊
中文核心期刊

主管单位 中国科学院
主办单位 中国科学院遥感与数字地球研究所
中国图象图形学学会
北京应用物理与计算数学研究所

主　　编 顾行发
编辑出版 《中国图象图形学报》编辑出版委员会
通信地址 北京市海淀区北四环西路19号
邮　　编 100190
电子信箱 jig@radi.ac.cn
电　　话 010-58887035
网　　址 www.cjig.cn

广告发布登记号 京朝工商广登字20170218号
总 发 行 北京报刊发行局
订　　购 全国各地邮局
海外发行 中国国际图书贸易集团有限公司
（邮政信箱：北京399信箱　邮编：100048）
印刷装订 北京科信印刷有限公司

Journal of Image and Graphics

Title inscription: Song Jian | Monthly, Started in 1996

Superintended by Chinese Academy of Sciences
Sponsored by Institute of Remote Sensing and Digital Earth, CAS
China Society of Image and Graphics
Institute of Applied Physics and Computational Mathematics

Editor-in-Chief Gu Xingfa
Editor, Publisher Editorial and Publishing Board of Journal of Image and Graphics
Address No. 19, North 4th Ring Road West, Haidian District, Beijing, P. R. China
Zip code 100190
E-mail jig@radi.ac.cn
Telephone 010-58887035
Website www.cjig.cn

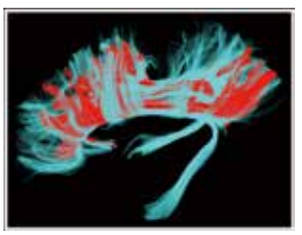
Distributed by Beijing Bureau for Distribution of Newspapers and Journals
Domestic All Local Post Offices in China
Overseas China International Book Trading Corporation
(P.O.Box 399, Beijing 100048,P.R.China))
Printed by Beijing Kexin Printing Co., Ltd.

CN 11-3758/TB
ISSN 1006-8961
CODEN ZTTXFZ

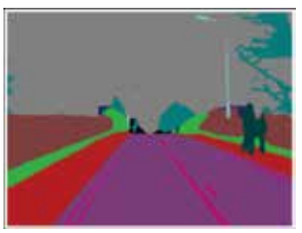
国外发行代号 M1406
国内邮发代号 82-831
国内定价 60.00元

2020年8月 第25卷 第8期（总第292期）
中国图象图形学报
Zhongguo Tuxiang Tuxing Xuebao

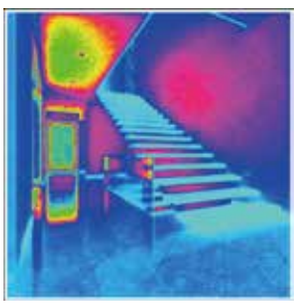
目次



神经纤维跟踪算法研究进展(第1513页)



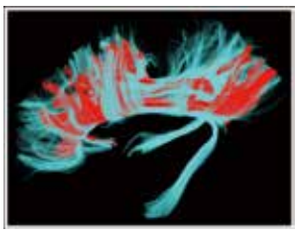
结合注意力机制的双路径语义分割(第1627页)



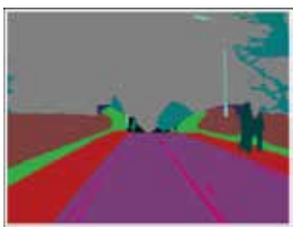
突变策略下多通路Metropolis光照传播(第1658页)

Volume 25, Number 8
Published August 16, 2020

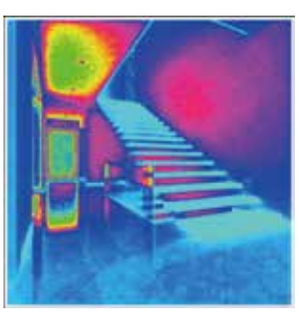
CONTENTS
JOURNAL OF IMAGE AND GRAPHICS



Research progress of neural fiber tracking(P1513)



Two-path semantic segmentation algorithm combining attention mechanism(P1627)



Improved multiplexed Metropolis light transport based on fusion mutation strategy(P1658)

中图法分类号:TP391　　**文献标识码**:A　　**文章编号**:1006-8961(2020)08-1591-13

论文引用格式:Lan H and Liu Q Y. 2020. Image generation from scene graph with graph attention network. Journal of Image and Graphics,25(08):1591-1603(兰红,刘秦邑. 2020. 图注意力网络的场景图到图像生成模型. 中国图象图形学报,25(08):1591-1603)[DOI:10.11834/jig.190515]

图注意力网络的场景图到图像生成模型

兰红,刘秦邑

江西理工大学信息工程学院,赣州　341000

摘　要:**目的**　目前文本到图像的生成模型仅在具有单个对象的图像数据集上表现良好,当一幅图像涉及多个对象和关系时,生成的图像就会变得混乱。已有的解决方案是将文本描述转换为更能表示图像中场景关系的场景图结构,然后利用场景图生成图像,但是现有的场景图到图像的生成模型最终生成的图像不够清晰,对象细节不足。为此,提出一种基于图注意力网络的场景图到图像的生成模型,生成更高质量的图像。**方法**　模型由提取场景图特征的图注意力网络、合成场景布局的对象布局网络、将场景布局转换为生成图像的级联细化网络以及提高生成图像质量的鉴别器网络组成。图注意力网络将得到的具有更强表达能力的输出对象特征向量传递给改进的对象布局网络,合成更接近真实标签的场景布局。同时,提出使用特征匹配的方式计算图像损失,使得最终生成图像与真实图像在语义上更加相似。**结果**　通过在包含多个对象的 COCO-Stuff 图像数据集中训练模型生成 64 × 64 像素的图像,本文模型可以生成包含多个对象和关系的复杂场景图像,且生成图像的 Inception Score 为 7.8 左右,与原有的场景图到图像生成模型相比提高了 0.5。**结论**　本文提出的基于图注意力网络的场景图到图像生成模型不仅可以生成包含多个对象和关系的复杂场景图像,而且生成图像质量更高,细节更清晰。

关键词:场景图生成图像;图注意力网络;场景布局;特征匹配;级联细化网络

Image generation from scene graph with graph attention network

Lan Hong, Liu Qinyi

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

Abstract: **Objective** With the development of deep learning, the problem of image generation has achieved great progress. Text-to-image generation is an important research field based on deep learning image generation. A large number of related papers conducted by researchers have proposed to implement text-to-image. However, a significant limitation exists, that is, the model will behave poorly in terms of relationships when generating images involving multiple objects. The existing solution is to replace the description text with a scene graph structure that closely represents the scene relationship in the image and then use the scene graphs to generate an image. Scene graphs are the preferred structured representation between natural language and images, which is conducive to the transfer of information between objects in the graphs. Although the scene graphs to image generation model solve the problem of image generation, including multiple objects and relationships, the existing scene graphs to image generation model ultimately produce images with lower quality, and the object details are unremarkable compared with real samples. A model with improved performance should be developed to generate high-quality images and to solve large errors. **Method** We propose a model called image generation from scene graphs with a graph attention network (GA-SG2IM), which is an improved model implementing image generation from scene graphs, to gener-

收稿日期:2019-10-10;**修回日期**:2020-01-08;**预印本日期**:2020-01-15

基金项目:国家自然科学基金项目(61762046);江西省自然科学基金项目(20161BAB212048)

Supported by:National Natural Science Foundation of China (61762046)

ate high-quality images containing multiple objects and relationships. The proposed model mainly realizes image generation in three parts: First, a feature extraction network is used to realize the feature extraction of the scene graphs. The attention network of the graphs introduces the attention mechanism in the convolution network of original graphs, enabling the output object vector to have strong expression ability. The object vector is then passed to the improved object layout network for obtaining a respectful and factual scene layout. Finally, the scene layout is passed to the cascaded refinement network for obtaining the final output image. A network of discriminators consisting of an object discriminator and an image discriminator is connected to the end to ensure that the generated image is sufficiently realistic. At the same time, we use feature matching as our image loss function to ensure that the final generated and real images are similar in semantics and to obtain high-quality images. **Result** We use the COCO-Stuff image dataset to train and validate the proposed model. The dataset includes more than 40 000 images of different scenes, where each of them provides annotation information of the borders and segmentation masks of the objects in the image, and the annotation information can be used to synthesize and input the scene graph of the proposed model. We train the proposed model to generate 64 × 64 images and compare them with other image generation models to prove its feasibility. At the same time, the quantitative results of the Inception Score and the bounding box intersection over union (IoU) of the generated image are compared to determine the improvement effects of the proposed model and SG2IM (image generation from scene graph) and StackGAN models. The final experimental results show that the proposed model achieves an Inception Score of 7.8, which increases by 0.5 compared with the SG2IM model. **Conclusion** Qualitative experimental results show that the proposed model can realize the generation of complex scene images containing multiple objects and relationships and improves the quality of the generated images to a certain extent, making the final generated images clear and the object details evident. A machine can autonomously model its input data and takes a step toward "wisdom" when it can generate high-quality images containing multiple objects and relationships. Our next goal is to enable the proposed model for generating real-time high-resolution images, such as photographic images, which requires many theoretical supports and practical operations.

Key words: image generation from scene graphs; graph attention network; scene layout; feature matching; cascaded refinement network

0 引 言

真正的计算机视觉是指计算机不仅能识别图像(兰红和方治屿,2019),而且能创造图像。当计算机能创造图像时,说明计算机真正理解图像。随着深度学习的发展,图像生成技术(强振平 等,2019)得到显著发展。基于深度学习的生成模型主要分为VAE(variational autoencoder)模型(Kingma 和Welling,2013)、Autoregressive 模型(van den Oord 等,2016;Salimans 等;2017)和对抗式生成网络(generative adversarial networks,GANs)模型等(Radford 等,2015;Goodfellow 等,2014;刘哲良等,2019)。VAE 模型通过变分推理方式在潜在向量空间和图像之间联合训练一组编码器和解码器,然后对潜在空间向量进行采样并利用已训练好的解码器输出最终生成图像;Autoregressive 模型将联合建模问题转换为序列问题,每一个像素点的生成都基于所有之前像素点的值;GANs 模型以对抗博弈的方式训练一组生成器和鉴别器,最终生成接近真实的图像。

基于文本描述生成图像是图像生成模型研究的一个主要方向,如果模型能够实现文本到图像的转换,说明模型在语义上理解了图像。Reed 等人(2016a)提出的 GAN-INT-CLS 模型是使用 GANs 模型实现文本序列生成图像的首次尝试,将文本向量作为 GANs 模型的条件输入,较好实现了文本到图像的生成,但主要适用于生成分辨率为 32 × 32 像素的图像。以此为基础,Zhang 等人(2017)提出了 StackGAN 模型,使用两个串联的 GANs 生成了分辨率为 256 × 256 像素的图像。Reed 等人(2016b)提出基于位置约束的文本到图像生成的 GAWWN(learning what and where to draw)模型,通过捕获图像中对象的定位约束,学习图像中对象的边界框,提升生成图像质量。Zhang 等人(2019)提出 StackGAN++模型,使用多对生成器和鉴别器,解决了 StackGAN 模型仅使用两对的局限性,同时为鉴别器增加无条件图像损失,提高了鉴别器的能力。Xu 等

人(2018)提出 AttnGAN 模型,在 StackGAN ++ 模型基础上引入注意力机制(Xu 等,2015)),通过自然语言描述中的相关单词匹配图像不同子区域,同时提出 DAMSM(deep attentional multimodal similarity model)模型计算子图像和相应单词间的损失,使得文本特征具有视觉分辨力。

现有的文本到图像生成模型的研究只适用于包含单个对象的图像数据集,当遇到包含多个对象和关系的复杂场景图像时,生成的图像就会变得混乱。原因在于文本序列是一种线性结构,文本描述中的对象之间较难进行信息传递。Johnson 等人(2018)提出一种基于场景图到图像的生成模型(image generation from scene graph, SG2IM),使用的输入场景图与文本序列相比,更能有力地表示图像中对象之间的结构关系,且更有利于对象之间信息传递。SG2IM 模型利用图卷积网络(Kipf 和 Welling,2016)实现对场景图的特征提取,实现了包含多个对象和关系的图像生成。

虽然 SG2IM 模型在一定程度上解决了文本到图像生成模型的局限性,但是最终生成的图像质量不高,图像中的对象细节不清晰。为此,本文在 SG2IM 模型的基础上,提出基于图注意力网络的场景图到图像生成模型(image generation from scene graph with graph attention network, GA-SG2IM),从 3 个方面提升最终生成图像质量:1)使用图注意力网络(graph attention network, GAT)作为场景图的特征提取网络,以获得具有更强大表示能力的对象特征向量;2)改进对象布局网络中的掩码回归网络和边框回归网络,生成更准确的对象边框和分割掩码;3)采用感知损失作为图像损失函数,使最终生成图像和输入场景图在语义层次上更相似。

1　SG2IM 模型介绍

1.1　模型架构

SG2IM 模型将描述对象及其关系的场景图(scene graph)作为输入,生成与该场景图相对应的逼真图像。模型由生成器网络和鉴别器网络两部分组成,具体架构如图 1 所示。

1.2　子网络

本文提出的 SG2IM 模型的子网络包括图卷积网络、场景布局合成网络、级联细化网络(cascaded refinement network,CRN)和鉴别器网络。

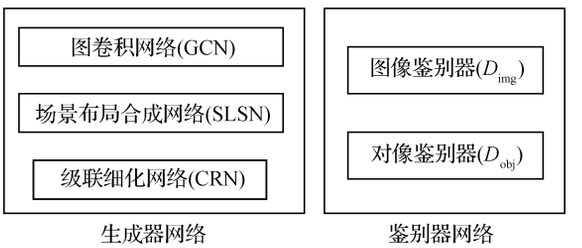


图 1　SG2IM 模型生成器和鉴别器网络组成架构

Fig. 1　Composition architecture of generator and discriminator network of SG2IM model

图卷积网络引入 Kipf 和 Welling(2016)提出的图卷积思想,实现以空间域卷积的方式处理模型输入场景图,目的是得到图中每个对象的抽象向量表示,该嵌入向量聚合了图中其他对象的特征信息。图卷积网络的每一层通过训练 3 个函数得到图中的对象和关系的抽象向量表示,函数的输入为图中的边,输出为边的起始对象(subject)、关系(relationship)和目标对象(object)的向量表示,然后通过一个平均池化函数得到所有对象节点的最终向量表达。通过多个这样的图卷积层之后,每个最终输出对象向量就能沿着图的边聚合其他对象的信息。

场景布局合成网络的目的是为了合成一幅与最终生成图像对应的粗糙场景布局。该场景布局类似生成图像的语义分割图,只包含对象位置信息和边缘轮廓信息,而不包括对象色彩细节。所以要合成一幅足够尊重事实的场景布局,需要预测每个对象的边框和分割掩码。SG2IM 模型通过包含边框回归网络(box regression network)和掩码回归网络(mask regression network)的对象布局网络实现对每个对象的边框和分割掩码的预测,最后将单幅图中所有对象布局合并,得到整个图像的场景布局。

级联细化网络的作用是将合成场景布局转化为生成图像,过程类似图像语义分割(Shelhamer, 2017)的逆过程。该网络通过逐渐为场景布局添加细节信息,按从粗到细的方式生成最终图像。级联细化网络适用于高分辨率、高逼真度的图像生成。Chen 和 Koltun(2017)利用级联细化网络生成了包含几十万个像素点照片般的真实图像。

鉴别器网络与生成器网络对抗训练可以在很大程度上改善生成器网络的输出。SG2IM 模型中的鉴别器由图像鉴别器和对象鉴别器组成。图像鉴别

器可以提升生成图像质量,对象鉴别器网络确保生成图像中的对象足够真实,同时对象鉴别器中引入辅助分类器(Odena 等,2016)确保对象可以被正确分类。

1.3 局限性

SG2IM 模型虽然生成了包含多个对象和关系的复杂场景图像,但是最终生成图像的质量不高。实验表明,只有使用真实标签提供的对象边框和分割掩码生成图像,才能较好区分图中不同对象之间的关系。而使用预测生成的分割掩码和边框生成的图像较为混乱,说明对象布局网络中的边框回归网络不能对场景图中的对象进行较好的定位,而掩码回归网络生成掩码不能较好地表现对象的边缘轮廓信息。所以需要改进对象布局网络以得到更好的对象分割掩码和边框。

2 GA-SG2IM 模型设计

本文提出 GA-SG2IM 模型,旨在生成更高质量的包含多个对象和关系的复杂场景图像,系统结构如图 2 所示。主要分为 4 个部分:1)利用图注意力网络对场景图进行特征提取,得到具有更强表达能力的对象向量;2)使用改进的掩码回归网络和边框回归网络得到更准确的对象掩码和边框,合成更贴近生成图像语义的 2D 场景布局。3)使用级联细化网络实现场景布局到生成图像的对应生成;4)鉴别器网络保证生成图像和图像中的对象足够真实。

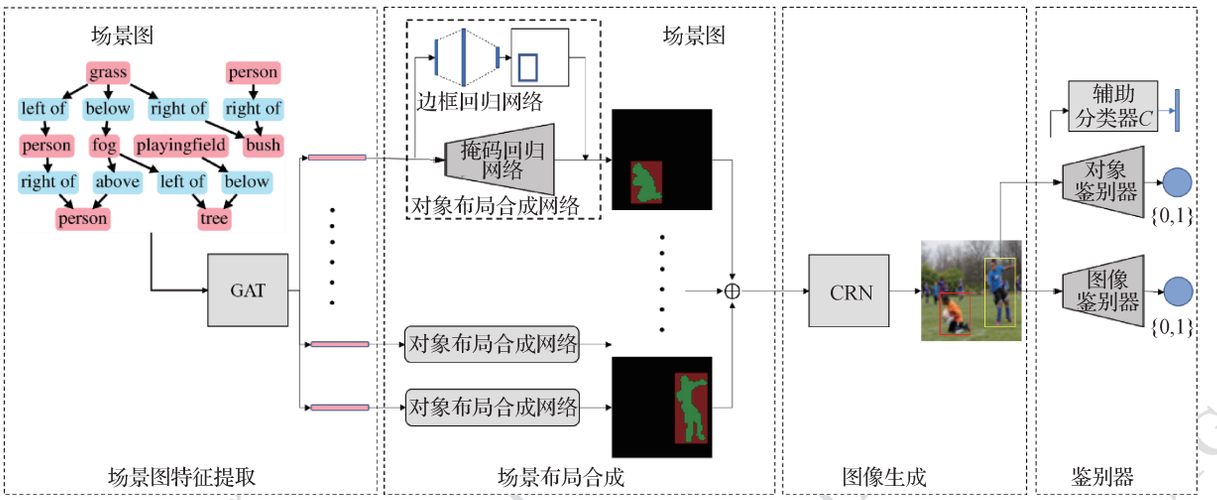


图 2 GA-SG2IM 模型总体结构示意图

Fig. 2 Architecture of GA-SG2IM mode

2.1 基于图注意力网络的场景图特征提取

2.1.1 场景图预处理

模型输入的场景图是由多个实例对象和关系组成的有向图,如图 2 中的输入场景图所示。有向图中每一个节点表示一个实例对象,每一条边表示对象之间的关系。

单个场景图可以用一个元组($\boldsymbol{O},\boldsymbol{E}$)表示,其中 $\boldsymbol{O}=\{o_1,\cdots,o_n\}$ 表示场景图的实例对象集,n 为场景图中的实例对象数;$\boldsymbol{R}=\{r_1,\cdots,r_m\}$ 表示场景图中的对象之间的关系集,m 为场景图中的关系数量;$\boldsymbol{E}=\{(o_i,r_1,o_j),\cdots,(o_p,r_m,o_q)\}$ 表示场景图中所有的有向边,i,j,p,q 表示场景图中的实例对象标号。

使用词嵌入(Chen 等,2015)技术将场景图中的所有对象和关系转换为抽象的嵌入向量,用 $\boldsymbol{h}_o=\{\boldsymbol{h}_{o_1},\cdots,\boldsymbol{h}_{o_n}\}$ 表示对象特征向量集,$\boldsymbol{h}_r=\{\boldsymbol{h}_{r_1},\cdots,\boldsymbol{h}_{r_m}\}$ 表示关系特征向量集。令对象特征向量集和关系特征向量集的特征维度都为 F_1。

2.1.2 GAT 实现场景图特征提取

场景图特征提取网络输出场景图中实例对象的嵌入向量表达,每个向量聚合了所有其他对象和关系的特征信息。

SG2IM 模型中的图卷积网络的每个卷积层通过训练 3 个可学习的函数 g_s、g_p 和 g_o,分别将场景图中每条边(o_i,r_k,o_j)中的 3 个元素对应的低级特征向量表示转换为高级特征向量表示。因此函数输

入都为场景图中边对应的嵌入向量（$\boldsymbol{h}_{o_i}$,$\boldsymbol{h}_{r_k}$,$\boldsymbol{h}_{o_j}$），通过对场景图中的边特征提取，实现沿场景图的边聚合信息，经过多个卷积层后，每个最终输出对象向量聚合了其他所有对象和关系的特征信息。

为了获得更强表达能力的对象嵌入向量，使用图注意力网络作为场景图特征提取网络。图注意力网络在图卷积网络的基础上引入注意力机制，即输出的每个特征向量在聚合邻域节点时，为所有邻域节点分配一个可学习的注意力系数，这样每个对象都能对其所有邻域节点有不同的感知力。具体步骤如下：

1）使用一个共享的参数矩阵 $\boldsymbol{W} \in \mathbf{R}^{F_1 \times F_2}$ 将场景图中所有的对象向量和关系向量转换为更高级的特征向量，保证对象和关系特征向量具有更强的表达能力。然后利用场景图中边的高级特征向量计算对象之间的注意力系数，计算为

$$e_{ij} = \varphi(\boldsymbol{W}[\boldsymbol{h}_{o_i},\boldsymbol{h}_{r_k},\boldsymbol{h}_{o_j}]) \tag{1}$$

式中，e_{ij}为场景图的任意边（o_i,r_k,o_j）中实例对象 o_j 对于实例对象 o_i 的贡献程度，其中注意力计算网络 $\varphi:\mathbf{R}^{3F_2} \rightarrow \mathbf{R}$，[·,·,·]表示张量合并操作。

2）使用 softmax 函数对每个对象的所有邻域节点对象进行标准化操作，使系数在不同节点之间易于比较，具体为

$$\alpha_{ij} = \underset{j}{\text{softmax}}(e_{ij}) = \frac{\exp(e_{ij})}{\sum_{k \in N_i}\exp(e_{ik})} \tag{2}$$

式中，N_i 表示节点 i 的所有一阶邻域节点（包括 i 本身）。获得标准化的注意力系数后，计算与它们对应的节点特征的线性组合，作为每个对象节点的最终输出，即

$$\boldsymbol{h}'_{o_i} = \sum_{j \in N_i}\alpha_{ij}\boldsymbol{W}\boldsymbol{h}_{o_j} \tag{3}$$

3）经过多个图注意力层处理后，每个对象得到其对应的最终输出特征向量，该向量包含场景图中其他所有对象的特征信息。这可以保证相同类别的不同对象有不同的抽象向量表示，用于预测对象布局时可以有不同的输出。图注意力网络结构如图 3 所示。

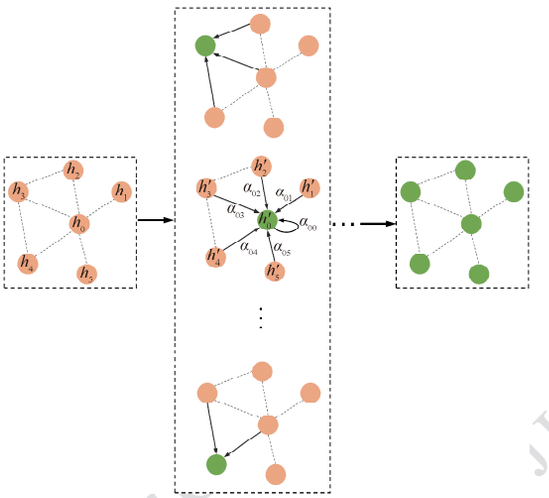


图 3　图注意力网络结构

Fig. 3　Architecture of GAT

2.2　场景布局合成

2.2.1　场景布局合成原理

将场景图转换为图像时需要合成图像对应的场景布局作为过渡，因此将图注意力网络输出的对象向量传递给对象布局网络，预测每个对象的对象布局，然后合并图中所有对象布局，得到图像场景布局。

对象布局网络主要由预测对象边框和分割掩码的边框回归网络，以及掩码回归网络两部分组成。在 SG2IM 模型中，边框回归网络使用两层的 MLP（multi-layer perceptron）预测对象边框的相对图像坐标，即 $\boldsymbol{b}$ =（x_0,y_0,x_1,y_1）。掩码回归网络则使用一系列的上采样层和卷积层预测固定尺寸为 $M \times M$ 的二进制掩码 $\boldsymbol{m}$，然后合成图像的场景布局，具体计算公式为

$$\boldsymbol{L} = \oplus_{i=1}^{n}(\boldsymbol{h}'_{o_i} \otimes \boldsymbol{m}_i) \odot \boldsymbol{b}_i \tag{4}$$

式中，$\boldsymbol{L}$ 表示最终合成场景布局，n 表示一幅图中的实例对象数，$\oplus$ 表示对象布局的合并操作；$\otimes$ 表示对象嵌入向量和掩码张量乘操作，以区分不同对象布局；$\odot$ 表示将分割掩码扩充到对象边框内。

但是 SG2IM 模型实验表明，使用预测生成的对象分割掩码和边框最终生成的图像较为混乱，而使用真实标签提供的分割掩码和边框合成的图像能清晰地分辨图像中的对象和它们之间的关系，说明对象布局网络无法保证生成较高质量的对象布局。

2.2.2　改进的对象布局网络

通过改进对象布局网络中的边框回归网络和掩码回归网络得到更高质量的对象布局，以提升最终生成图像的质量。

1）边框回归网络。SG2IM 模型中的边框回归网络通过直接预测对象边框的 4 个坐标值对图中对象进行定位，其损失函数为坐标差的绝对值。对大小不同的边框，若坐标差相同，则惩罚力度也相同，导致

边框的定位误差过大。例如,1 个像素的坐标差,对长度为 1 的边框影响力是 100%,对长度为 10 的边框影响力只有 10%。

因此在预测边框坐标时,需考虑尺寸比例的问题。本文采用区域卷积网络(region-convolutional neural network, R-CNN)(Girshick 等,2016)中提到的 4 个变换系数 t_x、t_y、t_w、t_h 作为学习目标,实际意义为

$$t_x = \frac{G_x - P_x}{G_w}, \quad t_y = \frac{G_y - P_y}{G_h}$$
$$t_w = \log\left(\frac{G_w}{P_w}\right), \quad t_h = \log\left(\frac{G_h}{P_h}\right) \tag{5}$$

式中,G_x、G_y、G_w、G_h 分别表示真实标签提供的边框中心点的坐标及边框的长宽,P_x、P_y、P_w、P_h 分别表示预测边框的中心点坐标及边框的长宽。

在实际模型训练中,利用线性变换预测 4 个变换系数,即给定对象向量 $\boldsymbol{h}'_{o_i}$,学习一组参数 $\boldsymbol{W} \in \mathbf{R}^{F_2 \times 4}$,得到表示边框变换系数向量 $t_* = \boldsymbol{W}\boldsymbol{h}'_o$。

在训练开始,为所有对象初始化一个边框 $P^0 = (x, y, w, h)$,令 P^0_x,P^0_y 为预生成图像的中心点,P^0_w,P^0_h 取生成图像长宽的一半。然后训练得到的变换系数,使初始边框慢慢向真实标签靠拢,P 的迭代公式可以表示为

$$\begin{cases} P^i_x = P^{i-1}_w t_x + P^{i-1}_x \\ P^i_y = P^{i-1}_h t_y + P^{i-1}_y \\ P^i_w = P^{i-1}_w \exp(t_w) \\ P^i_h = P^{i-1}_h \exp(t_h) \end{cases} \tag{6}$$

整个边框回归网络希望通过学习使得 $P \rightarrow G$,或 $t_* \rightarrow \hat{t}_*$,而 $\hat{t}_*$ 就是为边框回归网络学习提供的真实标签,计算为

$$\hat{t}_x = \frac{G_x - P^0_x}{G_w}, \quad \hat{t}_y = \frac{G_y - P^0_y}{G_h}$$
$$\hat{t}_w = \log\left(\frac{G_w}{P^0_w}\right), \quad \hat{t}_h = \log\left(\frac{G_h}{P^0_h}\right) \tag{7}$$

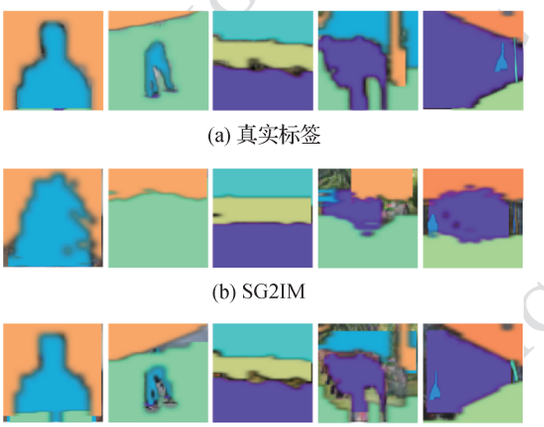
2)掩码回归网络。SG2IM 模型中的掩码回归网络利用一系列转置卷积操作将 1 维的对象向量 $\boldsymbol{h}'_{o_i}$ 转换为统一大小为 $M \times M$ 的对象分割掩码 $\boldsymbol{m}_i$。这是一个低维向量预测高维向量的过程,如果单纯使用生成掩码和真实标签提供掩码的交叉熵作为损失函数训练模型会导致预测的准确率较低。对这种约束的生成问题,引入对抗博弈思想可以有效提高生成质量。因此本文在掩码回归网络后面接一个掩码鉴别器,与掩码回归网络一起构成一个对抗式生成网络(GANs)。

掩码鉴别器与常规的鉴别器一样,是一个二分类网络,主要区分生成掩码和真实掩码。通过对抗博弈训练,掩码鉴别器网络可以更好地学习每个对象的轮廓信息,同时掩码回归网络也能生成与真实标签更相近的对象分割掩码。

2.2.3 对象布局网络训练

对象布局网络引入了对抗博弈的训练思想,因此为了避免前期生成的误差较大的对象布局影响后面网络的训练,需要整个生成网络进行分阶段训练。具体网络结构如图 2 中的场景布局合成网络所示,网络首先生成足够尊重生成图像的场景布局图,然后对其训练以生成高质量的图像。

最终采用改进的边框回归网络和掩码回归网络生成边框和的对象分割掩码,合成的部分场景布局如图 4 所示,为了能更好地区分对象的轮廓,采用非透明的彩色掩码对不同的对象进行标注。



(a) 真实标签

(b) SG2IM

(c) GA-SG2IM(本文)

图 4 生成场景布局样例图

Fig. 4 Examples of synthetic scene layout

((a) real labels; (b) SG2IM; (c) GA-SG2IM (cours))

2.3 基于级联细化网络实现图像生成

级联细化网络由多个功能相同的串联模块组成,每个模块由一个卷积层和上采样层构成。卷积层输入为该模块对应的缩放场景布局图和上一模块的输出,输出为细化后的图像特征集;上采样层接收卷积层输出的特征集,以最近邻插值法实现特征的尺寸倍增。通过这种级联细化,以端对端的方式实现场景布局图到图像的生成,模块越多,最终生成的图像越清晰。级联细化网络的系统结构如图 5 所示,图中 M_i 表示卷积操作,F_i 表示上采样操作,w_i,h_i,d_i,c 表示各模块的输入张量大小。

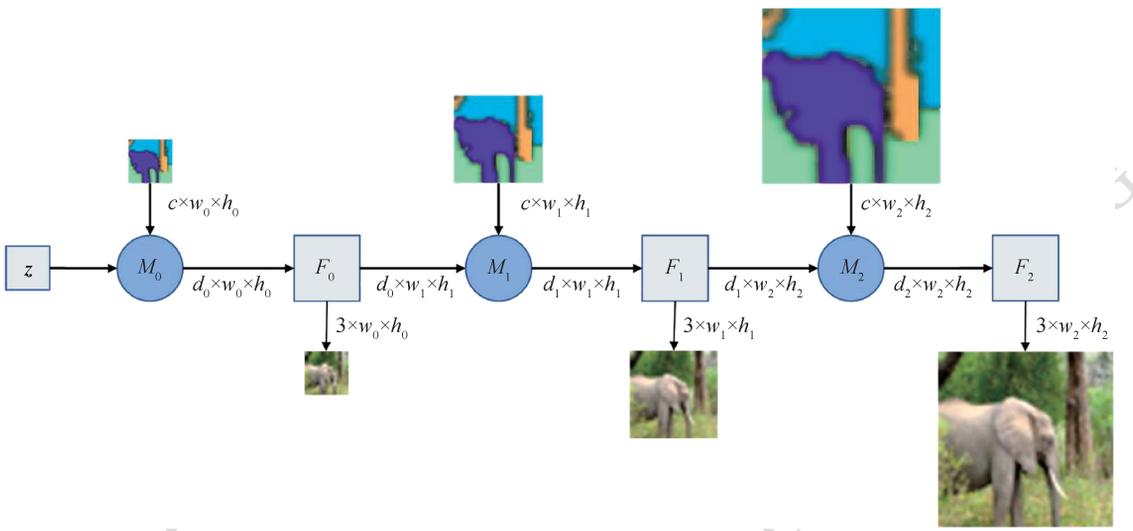


图 5　级联细化网络结构

Fig. 5　Architecture of cascaded refinement networks

级联细化网络的第 1 个模块除了输入缩放的场景布局，还需要初始化一个高斯噪声张量 $z\sim p(z)$ 作为输入；最后 1 个模块的后面接上输出通道为 3 的卷积，即可得到最终输出图像。

2.4　鉴别器网络

鉴别器网络是为了尽可能保证生成器网络输出的真实性，通过联合训练生成器网络和鉴别器网络可以在很大程度上提升生成图像的质量。

通过最大化式(5)训练鉴别器，以判别输入 x 是真实标签还是生成标签。具体为

$$L_{\mathrm{GAN}}=\underset{x\sim p_{\mathrm{real}}}{E}\log D(x)+\underset{x\sim p_{\mathrm{fake}}}{E}\log(1-D(x)) \tag{8}$$

式中，L_{GAN} 表示 GAN 的总损失，$x\sim p_{\mathrm{fake}}$ 表示生成器网络的输出分布，$x\sim p_{\mathrm{real}}$ 表示真实样本分布。同时，在保持鉴别器网络参数不变的情况下，最小化 L_{GAN}（Goodfellow 等，2014）使生成器网络生成可以欺骗鉴别器的输出。通过这种对抗博弈训练，使得鉴别器网络和生成器网络的能力都得以提升，最终生成器生成鉴别器无法区分的图像。

鉴别器网络包括图像鉴别器和对象鉴别器。图像鉴别器的作用是为了保证生成图像整体外观足够真实，输入为生成图像和真实图像，通过一个全卷积网络输出为真或假的概率分布，其网络实现类似 Isola 等人(2017)提出的鉴别器。对象鉴别器的作用是为了保证生成图像中的对象的真实程度，输入为从生成图像或真实图像中裁剪的对象图像，并使用双线性插值法(Jaderberg 等，2015)扩充至同一尺寸。为了保证每个对象都可以正确识别，在对象鉴别器上增加了一个用于判断对象类别的辅助分类器(Odena 等，2016)。

2.5　损失函数

本文模型训练分两个阶段：第 1 阶段通过训练图注意力网络和对象布局网络实现场景图到场景布局的生成；第 2 阶段通过训练级联细化网络和鉴别器网络，实现将场景布局转换为生成图像。不同阶段采用不同的损失函数对网络权重进行更新。

第 1 阶段的网络损失函数由边框损失 L_{box}、掩码损失 L_{mask} 和掩码鉴别器损失 $L_{\mathrm{GAN}}^{\mathrm{mask}}$ 组成，各损失函数计算为

$$\begin{aligned}L_{\mathrm{box}}&=\sum_{i=1}^{n}S_{\mathrm{L1}}(\hat{t}_*-t_*)\\L_{\mathrm{mask}}&=\sum_{i=1}^{n}H(\hat{m}_i-m_i)\\L_{\mathrm{GAN}}^{\mathrm{mask}}&=\sum_{i=1}^{n}C(\hat{m}_i,m_i)\end{aligned} \tag{9}$$

式中，n 表示场景图中实例对象的数量；S_{L1} 表示 Smooth L1 损失函数(Girshick 等，2016)，与 L2 损失函数相比离群点更加鲁棒；H 表示以逐像素交叉熵(pixelwise cross-entropy)方式计算预测掩码和真实掩码之间的差异；C 表示对生成掩码和真实掩码进行二分类的交叉熵损失。

第 2 阶段的网络损失函数由图像损失 L_{img}、图像鉴别器损失 $L_{\mathrm{GAN}}^{\mathrm{img}}$、对象鉴别器损失 $L_{\mathrm{GAN}}^{\mathrm{obj}}$ 和对象分类损失 $L_{\mathrm{GAN}}^{\mathrm{obj}}$ 组成。

图像损失 L_{img} 采用感知损失作为生成图像和真实图像差异的惩罚函数，基本思想是比较视觉感知

网络中的激活函数的输出值,通过缩小特征层次上的差异使生成图像与真实图像在语义上更接近。而实现特征匹配就需要使用预先训练好的视觉感知网络进行特征提取。本文取 VGG(visual geometry group)-19 的部分层作为本文的视觉感知网络。令 $\boldsymbol{\Phi}$ 表示该视觉感知网络,则 $\{\boldsymbol{\Phi}\}_l$ 表示网络层的集合,这些层取自 VGG-19 中的 conv1_2、conv2_2、conv3_2、conv4_2、conv5_2。图像损失计算为

$$L_{\text{img}} = \sum_l \lambda_l \left\| \boldsymbol{\Phi}_l(\boldsymbol{I}) - \boldsymbol{\Phi}_l(\hat{\boldsymbol{I}}) \right\|_1 \tag{10}$$

式中,$\boldsymbol{I}$ 和 $\hat{\boldsymbol{I}}$ 分别表示预测生成图像和真实图像;λ_l 是用来平衡各层对损失的贡献的超参数,它们初始化为每层中元素数量的倒数,通过学习的方式调整,经过一段训练之后,λ_l 逐渐调整将标准化的每项预期贡献提供给总的图像损失。

图像鉴别器损失 $L_{\text{GAN}}^{\text{img}}$ 用于提升生成图像的逼真程度,表示图像鉴别器对生成图像 $\boldsymbol{I}$ 和真实图像 $\hat{\boldsymbol{I}}$ 的二分类概率的交叉熵。对象鉴别损失 $L_{\text{GAN}}^{\text{obj}}$ 用于提升生成图像中对象的逼真程度,表示对象分类器对生成图像中剪辑对象 o_i 和真实图像中剪辑对象 $\hat{o}_i$ 二分类概率的交叉熵。对象分类损失 $L_{\text{AC}}^{\text{obj}}$ 确保图像中的对象可以被正确识别,表示生成图像中剪辑对象的多分类交叉熵。具体分别计算为

$$\begin{aligned} L_{\text{GAN}}^{\text{img}} &= C_1(\hat{\boldsymbol{I}}, \boldsymbol{I}) \\ L_{\text{GAN}}^{\text{obj}} &= \sum_{i=1}^{n} C_1(\hat{o}_i, o_i) \\ L_{\text{AC}}^{\text{obj}} &= \sum_{i=1}^{n} C_2(o_i \mid \boldsymbol{Y}) \end{aligned} \tag{11}$$

式中,n 表示单幅图像中实例对象数,C_1 表示二分类交叉熵损失函数,C_2 表示多分类交叉熵损失函数,$\boldsymbol{Y}$ 表示对象分类类别,共 1 000 个类别。

3 模型实施细节

令图注意力网络中输入对象向量维度 $F_1 = 128$,最终输出向量维度为 $F_2 = 128$,分割掩码输出尺寸 $M = 16$,生成图像的长、宽 $W = H = 64$。所有模型训练采用学习率为 0.000 1 的 Adam(Kingma 和 Ba,2015)作为优化函数,且设置 batch_size 为 32 进行 10 万次迭代。各网络结构设计如下:

1)图注意力网络由 5 个相同的图注意力层串联组成,每层需要训练两个参数化矩阵,分别将对象和关系向量进行特征变换,得到共享参数化矩阵 $\boldsymbol{W} \in \mathbf{R}^{F_1 \times F_2}$ 和注意力系数计算网络 $\varphi: \mathbf{R}^{3F_2} \rightarrow \mathbf{R}$。其中注意力计算网络由两个以 ReLU 函数为激活函数的线性层组成,网络结构如表 1 所示。

表 1 注意力系数计算网络 φ 的网络结构

Table 1 Network architecture of attension coefficient computing network φ

名称	操作
linear(leakyReLU)-1	$3 \times 128 \rightarrow 512$
linear(leakyReLU)-2	$512 \rightarrow 1$

2)对象布局网络由边框回归网络、掩码回归网络和掩码鉴别器网络组成。

边框回归网络结构如表 2 所示,输入对象向量,输出边框的 4 个变换系数。

表 2 边框回归网络结构

Table 2 Network architecture of box regression network

名称	操作
linear(ReLU)	$128 \rightarrow 512$
linear	$512 \rightarrow 4$

掩码回归网络实现对象张量到掩码张量的转换,需要一系列的转置卷积操作,这里采用上采样层(upsample)接上卷积层(conv)实现,其中卷积层都是以 ReLU 做为激活函数的带填充步长为 1 的 3×3 卷积,且引入 Ioffe 等和 Szegedy(2015)提出的 batch normalization 技术,最后一层为了保证输出的掩码值在(0, 1)之间,采用 Sigmoid 函数作为激活函数,具体结构如表 3 所示。

掩码鉴别器网络用于识别真实掩码和生成掩码,因为分类器主要学习掩码的轮廓特征进行分类,因此使用较小的卷积网络即可,具体网络结构如表 4 所示。

3)级联细化网络采用 5 个功能相同的串联模块组成,每个模块由一个上采样层和一个卷积层构成,其中卷积层为以学习率 0.2 的 leakyReLU(Maas 等,2013)为激活函数的带填充步长为 1 的 3 × 3 卷积。令 $\boldsymbol{C} = [1, 1\,024, 512, 256, 128, 64]$,$\boldsymbol{H} = \boldsymbol{D} = [2, 4, 8, 16, 32, 64]$,则第 i 级网络结构如表 5 所示,最后接上一个输出通道为 3 的卷积层得到最终的图像。

4)鉴别器网络包括图像鉴别器和对象鉴别器

表 3　掩码回归网络结构

Table 3　Network architecture of mask regression network

名称	操作
reshape-1	128 → 128 × 1 × 1
upsample-1	128 × 1 × 1 → 128 × 2 × 2
conv(ReLU)-1	128 × 2 × 2 → 64 × 2 × 2
upsample-2	64 × 2 × 2 → 64 × 4 × 4
conv(ReLU)-2	64 × 4 × 4 → 32 × 4 × 4
upsample-3	32 × 4 × 4 → 32 × 8 × 8
conv(ReLU)-3	32 × 8 × 8 →1 6 × 8 × 8
upsample-4	16 × 8 × 8 → 16 × 16 × 16
conv(ReLU)-4	16 × 16 × 16 → 16 × 16 × 16
conv(Sigmoid)	16 × 16 × 16 → 1 × 16 × 16

表 4　掩码鉴别器网络结构

Table 4　Network architecture of mask discriminator

名称	操作
conv(ReLU)-1	1 × 16 × 16 → 64 × 8 × 8
conv(ReLU)-2	64 × 8 × 8 → 128 × 4 × 4
global average pooling	128 × 4 × 4 →128
linear(Sigmoid)	128 → 1

表 5　级联细化网络结构

Table 5　Network architecture of CRN

名称	操作
upsample	$(C_i+128)\times H_i\times W_i\rightarrow (C_i+128)\times H_{i+1}\times W_{i+1}$
conv(leakyReLU)	$(C_i+128)\times H_{i+1}\times W_{i+1}\rightarrow C_{i+1}\times H_{i+1}\times W_{i+1}$

两部分。图像鉴别器类似 Isola 等人(2016)提出的鉴别器,采用带填充步长为 2 的 3 × 3 卷积提取特征,然后输出 8 × 8 网格判断输入为真或者假,其网络结构如表 6 所示。对象鉴别器的输入是从真实图像或生成图像中裁剪下来的对象图像,统一缩放至 32 × 32 像素。输出两个概率分布,其一是判断输入是生成对象还是真实对象的二分类概率分布,其二是对象的具体类别概率分布,其网络结构如表 7 所示,C 为最终分类数。

4　实验结果和评估

使用2017COCO-Stuff数据集训练并验证本文

表 6　图像鉴别器网络结构

Table 6　Network architecture of image discriminator

名称	操作
conv(leakyReLU)-1	3 × 64 × 64 → 64 × 32 × 32
conv(leakyReLU)-2	3 × 32 × 32 → 128 × 16 × 16
conv-1	128 × 16 × 16 → 256 × 8 × 8
conv-2	256 × 8 × 8 → 1 × 8 × 8

表 7　对象鉴别器网络结构

Table 7　Network architecture of object discriminator

名称	操作
conv(leakyReLU)-1	3 × 32 × 32 → 64 × 16 × 16
conv(leakyReLU)-2	64 × 16 × 16 → 128 × 8 × 8
conv	128 × 8 × 8 → 256 × 4 × 4
global average pooling	256 × 4 × 4 → 256
linear-1	256 → 1 024
linear-2	1 024 → 1
linear-3	1 024 → $\lvert C\rvert$

GA-SG2IM 模型。首先利用图像提供的注释信息合成场景图,然后训练模型以生成 64 × 64 像素的图像,并使用 Inception Score(IS)(Salimans 等,2016)评估生成图像质量,以及使用交并比(intersection over union,IoU)评估生成图像中对象的准确度。

4.1　数据集

2017 COCO-Stuff 数据集(Caesar 等,2018)是 MS COCO(Microsoft COCO)数据集(Lin 等,2014)的一个子集。COCO-Stuff 数据集包含 4 万多幅各种场景的图像,且每幅图像都对其中的对象进行了注释,主要包含对象边框信息和分割掩码,通过这些注释信息可以合成模型输入场景图。

具体地,根据对象的图像坐标标注对象之间的相对关系,用 6 个互斥的几何关系 left of、right of、above、below、inside 和 surrounding 构造场景图。同时,为所有的场景图添加一个特殊的 image 对象进行扩充,并且为每个对象和 image 对象之间添加一个特殊的 in image 关系,保证场景图可以将所有对象联系在一起。

实验忽略覆盖面积不到图像 2% 的对象,保留对象数在 3～8 个的图像。最终从 COCO-Stuff 训练集中得到符合要求的 24 972 幅图像作为本文实验

的训练集,从 COCO-Stuff 验证集中选取 1 024 幅符合要求图像作为验证集,2 048 幅图像作为测试集。

4.2 定性结果

本文使用 Pytorch 深度学习框架设计本文的 GA-SG2IM 模型,训练集在单个 1080Ti 显卡上训练,最终得到训练完成的模型。为了验证模型的泛化能力,使用测试集对其进行验证,最终生成的样本图像如图 6 所示。为了更好地进行比较,图6还展示了

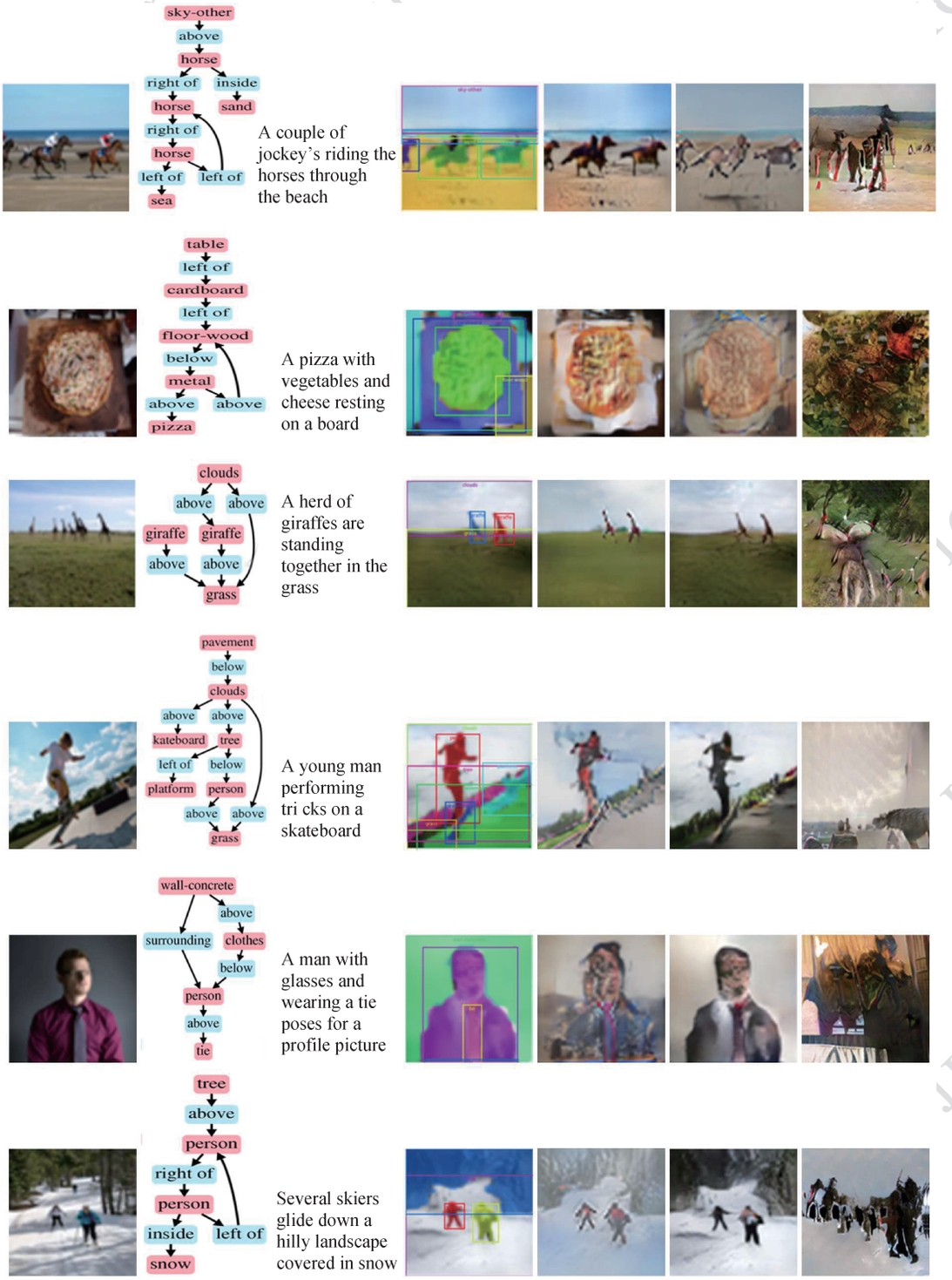


图 6 使用 COCO 测试集生成的图像

Fig. 6 Images generated using the COCO test set

((a) reference images; (b) scene graph; (c) reference text; (d) scene layout; (e) GA-SG2IM(ours); (f) SG2IM; (g) StackGAN)

相应的参考图像、根据参考图像注释信息合成的模型输入场景图、该参考图像对应的其中一个描述图像的参考文本、生成图像对应的 2D 场景布局图(为了更有利于观察,部分图像覆盖率较高的对象掩码不做标注,而标注对象边框,比如天空和草地等)、使用相同场景图输入 SG2IM 模型(Johnson 等,2018)的生成图像和使用参考文本作为输入的 StackGAN 模型(Zhang 等,2017)的生成图像。

从图 6 可以看出,本文 GA-SG2IM 模型生成的图像与参考图像相比,虽然细节上有些不足,但是图像的整体布局和图像中对象的轮廓与参考图像非常相似。与 SG2IM 模型生成的图像相比,本文生成图像更加平滑、清晰度更高,与参考图像的差异更小。而 StackGAN 模型生成图像比较混乱,说明与使用文本描述作为输入的方法相比,使用场景图作为输入的方法更有利于生成包含多个对象和关系的复杂场景图像。

4.3　评价指标

为了更好地对生成图像质量进行评估,本文引入 Inception Score 指标测试生成图像质量,从清晰度和多样性两方面评估生成模型,计算为

$$IS = \exp(\underset{\boldsymbol{x}\sim p_g}{E}\mathrm{KL}(p(\boldsymbol{y}\mid\boldsymbol{x})\|p(\boldsymbol{x}))) \tag{12}$$

式中,$\boldsymbol{x}\sim p_g$ 表示生成器生成图像,KL 表示计算两个概率分布的 KL 散度(Kullback-Leibler divergence),$\boldsymbol{y}$ 是图像经过 Inception V3 网络生成的 1 000 维的输出向量,$p(\boldsymbol{y}\mid\boldsymbol{x})$ 表示图像属于各个类别的概率分布,$p(\boldsymbol{y})$ 表示 N 幅(本文 N = 5 000)图像的平均类别概率分布。

在 COCO 数据集上,将测试集样本随机分为 5 组,并记录各组的均值和标准差。对于 GA-SG2IM 模型和 SG2IM 模型,每个测试集的图像合成 5 幅不同的场景图以生成 5 幅样本图像。对于 StackGAN 模型,每个测试集图像的每条描述文本各生成一幅图像,生成的图像大小为 256 × 256 像素,通过降采样的方式将图像缩小为 64 × 64 像素,最终结果如表 8 所示。

除了查看图像,评估生成图像质量还可以检测模型预测的对象边框,有两种常用的度量方式。一种是计算预测生成对象边框 b 与真实标签提供对象边框 $\hat{b}$ 的交并比,具体为

$$IoU = \frac{b\cap\hat{b}}{b\cup\hat{b}} \tag{13}$$

表 8　不同方法生成 64 × 64 像素图像的 *IS* 比较

Table 8　Comparison of *IS* for 64 × 64 pixels images generated by different methods

方法	*IS*
real images	16.3 ± 0.4
stackGAN	8.4 ± 0.2
SG2IM	7.3 ± 0.1
GA-SG2IM(本文)	7.8 ± 0.1

另一种度量方式是生成对象边框的多样性,即预测的对象边框相对于图中其他对象和关系的变化,用每个类别的对象边界框的位置和面积的标准差进行评估。

表 9 为本文 GA-SG2IM 模型和 SG2IM 模型(Johnson 等,2018)对于预测生成边框的准确性和多样性的评估。其中,R@t 表示对不同 IoU 阈值的对象召回率(recall),用以评估预测对象边框的准确性,σ_x 和 σ_{area} 分别表示所有类别的对象边框位置和面积的标准差的平均值。

表 9　预测边框的准确性和多样性评估

Table 9　Statistics of predicted bounding boxes

方法	R@0.3	R@0.5	σ_x	σ_{area}
SG2IM	52.4	32.2	**0.1**	**0.2**
GA-SG2IM(本文)	**58.7**	**38.4**	**0.1**	**0.2**

注:加粗字体表示各列最优结果。

如果不使用图卷积,那么模型只能为每个对象类别的对象预测一个单独的边界框,而无法实现相同类别不同对象的预测生成,即 $\sigma_x = \sigma_{\mathrm{area}} = 0$。

通过实验数据显示,与 StackGAN 模型相比,本文 GA-SG2IM 模型更有利于包含多个对象和关系的复杂场景图像生成,且生成图像质量更高,此外使用改进的对象布局网络对图像中的对象的位置预测更加准确,对统一类别的不同对象有较好的区分。

5　结　论

针对现有的文本到图像生成模型无法生成包含多个对象和关系的问题,本文提出了基于图注意力网络场景图到图像生成的 GA-SG2IM 模型。首先采用图注意力网络作为输入场景图的特征提取网络;

然后采用改进的对象布局网络预测图像中的对象分割掩码和边框,以得到更符合事实的场景布局;再利用级联细化网络将场景布局转换为真实场景图像,并采用感知损失计算生成图像与参考图像的差异,使得生成图像和参考图像在语义程度上更加相似。

最终定性实验结果表明,本文模型最终生成图像的对象细节更加明显,对象之间的关系更符合事实,说明模型在一定程度上提升了生成图像质量。同时定量实验结果表明,与 Johnson 等人(2018)提出的 SG2IM 模型相比,本文 GA-SG2IM 模型可以得到更高的 *IS*。

本文 GA-SG2IM 模型实现了包含多个对象和关系的复杂场景图像的生成,但是模型输入场景图依赖大量的图像标注信息,对训练数据集中未出现的对象及关系的生成效果较差。因此后续工作将模型直接通过更容易得到的自然语义文本生成复杂场景图像作为研究方向,进一步探索新方法以实现对语义文本中对象的捕捉及它们之间关系的判定。

参考文献(References)

Caesar H, Uijlings J and Ferrari V. 2018. COCO-Stuff: thing and stuff classes in context//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE: 1209-1218 [DOI: 10.1109/CVPR.2018.00132]

Chen Q F and Koltun V. 2017. Photographic image synthesis with cascaded refinement networks//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice: IEEE: 1511-1520 [DOI: 10.1109/ICCV.2017.168]

Chen X X, Xu L, Liu Z Y, Sun M S and Luan H B. 2015. Joint learning of character and word embeddings//Proceedings of the 24th International Conference on Artificial Intelligence. [s. l.]: AAAI Press: 1236-1242

Girshick R, Donahue J, Darrell T and Malik J. 2016. Region-based convolutional networks for accurate object detection and segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(1): 142-158 [DOI: 10.1109/TPAMI.2015.2437384]

Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y. 2014. Generative adversarial nets//Proceedings of the 27th International Conference on Neural Information Processing Systems. Cambridge: MIT Press: 2672-2680

Ioffe S and Szegedy C. 2015. Batch normalization: accelerating deep network training by reducing internal covariate shift//Proceedings of the 32nd International Conference on International Conference on Machine Learning. [s. l.]: ACM: 448-456

Isola P, Zhu J Y, Zhou T H and Efros A A. 2017. Image-to-image translation with conditional adversarial networks//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE: 1125-1134 [DOI: 10.1109/CVPR.2017.632]

Jaderberg M, Simonyan K, Zisserman A and Kavukcuoglu K. 2015. Spatial transformer networks//Advances in Neural Information Processing Systems. [s. l.]: [s. n.]: 2017-2025

Johnson J, Gupta A and Li F F. 2018. Image generation from scene graphs//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE: 1219-1228 [DOI: 10.110/CVPR.2018.00133]

Kingma D P and Ba J. 2015. Adam: a method for stochastic optimization [EB/OL]. [2019-10-01]. https://arxiv.org/pdf/1412.6980.pdf

Kingma D P and Welling M. 2013. Auto-encoding variational bayes [EB/OL]. [2019-10-01]. https://arxiv.org/pdf/1312.6114.pdf

Kipf T N and Welling M. 2016. Semi-supervised classification with graph convolutional networks [EB/OL]. [2019-10-01]. https://arxiv.org/pdf/1609.02907.pdf

Lan H and Fang Z Y. 2019. Image recognition of steel plate defects based on a 3D gray matrix. Journal of Image and Graphics, 24(6): 859-869 (兰红, 方治屿. 2019. 3 维灰度矩阵的钢板缺陷图像识别. 中国图象图形学报, 24(6): 859-869) [DOI: 10.11834/jig.180555]

Lin T Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P and Zitnick C L. 2014. Microsoft COCO: common objects in context//Proceedings of the 13th European Conference on Computer Vision. Zurich: Springer: 740-755 [DOI: 10.1007/978-3-319-10602-1_48]

Liu Z L, Zhu W and Yuan Z Y. 2019. Image instance style transfer combined with fully convolutional network and cycleGAN. Journal of Image and Graphics, 24(8): 1283-1291 (刘哲良, 朱玮, 袁梓洋. 2019. 结合全卷积网络与 CycleGAN 的图像实例风格迁移. 中国图象图形学报, 24(8): 1283-1291) [DOI: 10.11834/jig.180624]

Maas A L, Hannun A Y and Ng A Y. 2013. Rectifier nonlinearities improve neural network acoustic models//ICML Workshop on Deep Learning for Audio, Speech and Language Processing. [s. l.]: [s. n.]: #3

Odena A, Olah C and Shlens J. 2016. Conditional image synthesis with auxiliary classifier GANs//Proceedings of the 34th International Conference on Machine Learning. [s. l.]: JMLR. org: 2642-2651

Qiang Z P, He L B, Chen X and Xu D. 2019. Survey on deep learning image inpainting methods. Journal of Image and Graphics, 24(3): 447-463. (强振平, 何丽波, 陈旭, 徐丹. 2019. 深度学习图像修复方法综述. 中国图象图形学报, 24(3): 447-463) [DOI: 10.11834/jig.180408]

Radford A, Metz L and Chintala S. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. [EB/OL]. [2019-10-01]. https://arxiv.org/pdf/1511.06434.

pdf

Reed S, Akata Z, Yan X C, Logeswaran L, Schiele B and Lee H. 2016a. Generative adversarial text to image synthesis//Proceedings of the 33rd International Conference on International Conference on Machine Learningar. [s. l.]: ACM: 1060-1069

Reed S E, Akata Z, Mohan S, Tenka S, Schiele B and Lee H. 2016b. Learning what and where to draw//Advances in Neural Information Processing Systems. Barcelona, Spain: [s. n.]: 217-225

Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A and Chen X. 2016. Improved techniques for training GANs//Advances in Neural Information Processing Systems. Barcelona, Spain: [s. n.]: 2234-2242

Salimans T, Karpathy A, Chen X and Kingma D P. 2017. PixelCNN++: improving the pixelCNN with discretized logistic mixture likelihood and other modifications [EB/OL]. [2019-10-10]. https://arxiv.org/pdf/1701.05517.pdf

Shelhamer E, Long J and Darrell T. 2017. Fully convolutional networks for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(4): 640-651 [DOI: 10.1109/TPAMI.2016.2572683]

van den Oord A, Kalchbrenner N and Kavukcuoglu K. 2016. Pixel recurrent neural networks[EB/OL]. [2019-10-10]. https://arxiv.org/pdf/1601.06759.pdf

Xu K, Ba J L, Kiros R, Cho K, Courville A, Salakhutdinov R, Zemel R and Bengio Y. 2015. Show, attend and tell: neural image caption generation with visual attention//Proceedings of the 32nd International Conference on Machine Learning. [s. l.]: [s. n.]: 2048-2057

Xu T, Zhang P C, Huang Q Y, Zhang H, Gan Z, Huang X L and He X D. 2018. AttnGAN: fine-grained text to image generation with attentional generative adversarial networks//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE: 1316-1324 [DOI: 10.1109/CVPR.2018.00143]

Zhang H, Xu T, Li H S, Zhang S T, Wang X G, Huang X L and Metaxas D. 2017. StackGAN: text to photo-realistic image synthesis with stacked generative adversarial networks//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice: IEEE: 5907-5915 [DOI: 10.1109/ICCV.2017.629]

Zhang H, Xu T, Li H S, Zhang S T, Wang X G, Huang X L and Metaxas D N. 2019. StackGAN++: realistic image synthesis with stacked generative adversarial networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(8): 1947-1962 [DOI: 10.1109/TPAMI.2018.2856256]

作者简介

兰红，1963 年生，女，教授，硕士生导师，主要研究方向为计算机视觉、图像处理与模式识别。

E-mail: lanhong69@163.com

刘秦邑，通信作者，男，硕士研究生，主要研究方向为计算机视觉。

E-mail: lqy1075@163.com