一种基于对比学习大模型的视觉定位方法

陆庆阳[1],袁广林[2*],朱 虹[2],秦晓燕[2],薛模根[2,3]
(1. 中国人民解放军陆军炮兵防空兵学院研究生大队,安徽合肥 230031;2. 中国人民解放军陆军炮兵防空兵学院信息工程系,安徽合肥 230031;3. 偏振光成像探测技术安徽省重点实验室,安徽合肥 230031)

摘 要: 一阶段视觉定位方法由于其快速性而受到广泛关注,该方法利用图像与文本的融合特征预测目标框,但是现有方法在特征融合前没有进行图像与文本特征的对齐,限制了视觉定位的精度. 为了解决这一问题,本文提出一种基于对比学习大模型的视觉定位方法. 该方法采用基于对比学习的大规模预训练模型CLIP(Contrastive Language-Image Pre-training)提取图像和文本特征,利用Transformer编码器融合图像文本特征,使用多层感知机和融合特征预测目标框. 该方法能够解决视觉定位方法上述不足的原因在于:借助CLIP模型的编码器可以提取高度语义对齐的图像和文本特征,同时使用全局注意力交互融合图像与文本的上下文特征. 在5个数据集上,对本文提出的方法进行实验验证,实验结果表明:相比于现有视觉定位方法,本文方法取得了综合精度的提升.

关键词: 视觉定位;对比学习;变换器;注意力;大模型;对齐
中图分类号: TP391.4 **文献标识码:** A **文章编号:** 0372-2112(2024)10-3448-11
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20230364

A Visual Grounding Method with Contrastive Learning Large Model

LU Qing-yang[1], YUAN Guang-lin[2*], ZHU Hong[2], QIN Xiao-yan[2], XUE Mo-gen[2,3]
(1. *Graduate brigade, PLA Army Academy of Artillery and Air Defense, Hefei, Anhui* 230031, *China*;
2. *Department of Information Engineering, PLA Army Academy of Artillery and Air Defense, Hefei, Anhui* 230031, *China*;
3. *Anhui Province Key Laboratory of Polarization Imaging Detection Technology, Hefei Anhui* 230031, *China*)

Abstract: The one-stage visual grounding method has received widespread attention due to its speed, which uses fused features of images and text to predict target boxes. However, existing methods do not align image and text features before feature fusion, which limits the accuracy of visual grounding. To solve this problem, this paper proposes a visual grounding method based on contrastive learning large model. This method extracts features of image and text with CLIP (Contrastive Language-Image Pre-training) which is a large-scale pre-trained model based on contrastive learning. It uses Transformer encoders to fuse the image-text features and predicts target boxes using multi-layer perceptron and fused features. The method can overcome the above shortcomings for the following reasons: It can extract highly aligned image-text features in semantics via the CLIP encoders. Meanwhile, it uses global attention to interactively fuse contextual features of images and text. The proposed method was experimentally validated on five datasets, and the experimental results show that compared to existing visual grounding methods, the proposed method has achieved an improvement in overall accuracy.

Key words: visual grounding; contrastive learning; Transformer; attention; large model; align

1 引言

视觉定位(visual grounding)是计算机视觉领域一项重要的任务,其目的是在图像中定位语言指定的目标. 视觉定位在智能检测、智能监控、无人驾驶等方面具有广泛的应用前景. 近年来,得益于神经网络技术的不断提高和发展,视觉定位有了较大进步,已经提出一些方法,但是视觉定位是一项具有挑战性的任务,目前仍然面临诸多难题.

概括来说,现有视觉定位方法主要包括两阶段方法[1~10]和一阶段方法[11~18]两类. 两阶段方法首先生成一些区域建议,然后将区域建议与语言表达进行匹配,最后根据最优匹配得到目标定位框. 两阶段方法存在两个问题. 一是定位性能取决于区域建议的质量,如果在区域建议阶段中没有检测到目标,则不可能对其进

收稿日期:2023-04-21;修回日期:2023-10-18;责任编辑:王天慧
*通讯作者:袁广林

行定位. 二是由于要进行区域建议生成和跨模态相似性计算,导致其计算成本较高. 为了克服两阶段方法存在的问题,研究者提出了一阶段视觉定位方法. 一阶段方法首先融合视觉与语言特征,然后利用融合特征预测目标框,其关键问题是图像与文本信息的融合. 近年来提出的基于Transformer的一阶段视觉定位方法[13~18]提升了视觉定位性能,是目前研究的热点. 但是这些方法在用Transformer编码器或解码器融合图像与文本特征之前,没有进行图像与文本特征的对齐,这限制了视觉定位精度. 另一方面,由于使用Transformer模型使得这些方法的训练需要更多的数据,并且训练收敛速度较慢.

受到CLIP[19]大规模预训练模型在动作识别等视觉任务中应用的启发,本文针对上述问题提出了一种基于对比学习大模型的视觉定位方法(Visual Grounding Method with CLIP, CLIPVG). 该方法采用CLIP模型的图像编码器和文本编码器提取特征,使用Transformer编码器融合图像文本特征,利用融合特征和多层感知机定位目标. 在5个视觉定位数据集(如RefCOCO等,见5.2节)上,对提出方法进行了实验验证,实验结果表明:相比于现有视觉定位方法,本文方法取得了综合精度的提升.

本文的主要贡献和创新点如下:

(1)提出一种基于Transformer架构的视觉定位方法,该方法的特征提取和融合均使用了Transformer架构,保证了特征提取模块和特征融合模块的结构一致性,从而能够充分融合图像和文本特征.

(2)提出利用基于对比学习的大规模预训练CLIP模型提取图像和文本特征,保证了提取的图像和文本特征处于同一语义空间,有利于两种模态信息的融合.

(3)提出利用预训练的CLIP模型初始化特征提取模块并且在训练时固定其网络参数,这大大减少了整个模型的训练参数,提高了训练速度.

2　相关工作

根据视觉定位的核心流程,现有视觉定位方法可以分为两类:两阶段方法和一阶段方法,下面对这两类方法进行概括介绍.

2.1　两阶段方法

视觉定位的早期工作主要是两阶段方法,在第一阶段,通过建议生成[20]或者目标检测[21]产生大量区域建议,在第二阶段,根据图像和文本特征之间的相似性选择最佳的候选建议作为定位结果. Liu等人[22]和Zhang等人[2]提出利用卷积神经网络和长短时网络分别提取图像和文本特征实现视觉定位. Hu等人[3]和Ye等人[4]使用模块化网络将文本表达式分解为主题、位置和对象间关系等不同的模块化组件,然后将每个组件与图像进行匹配实现视觉定位. Liu等人[5]和Zhuang等人[6]利用注意力机制关注关键词和图像区域实现视觉定位. Wang等人[7]和Yang等人[8]提出基于图的两阶段视觉定位方法,该方法利用对象关系图学习发现文本表达的相关对象. Cirik等人[9]和Liu[10]等人利用外部语言解析器增强文本的表示能力,从而提升两阶段方法的视觉定位性能. 虽然上述两阶段方法取得了令人鼓舞的结果,但仍存在两方面的缺点:一是提取密集的区域建议计算成本高,难以实现快速的视觉定位;二是区域建议的质量是影响视觉定位准确性的重要因素,如果提取的区域建议中没有要定位的目标,则视觉定位将失败.

2.2　一阶段方法

为了克服两阶段方法的不足,受一阶段目标检测的启发,研究者提出了一阶段视觉定位方法. 与两阶段方法不同,一阶段方法直接将文本特征与图像特征融合,利用融合特征直接预测目标区域的边界框. Yang等人[11]使用DarkNet[22]和Bert[23]分别提取图像和文本特征,利用卷积操作融合图像和文本特征,采用YOLOv3[22]检测器的输出头定位目标. Liao等人[24]将一阶段视觉定位建模为一个相关过滤过程. Yang等人[25]的ReSC方法提出一个递归子查询构造框架,在图像和查询中进行多次推理,从而减少推理混淆的情况. Huang等人[12]提出的LBYL-Net采用DarkNet和Bi-LSTM分别提取图像和文本特征,设计一个语言描述指导下特征融合模块交互融合图像和文本特征,使用YOLOv3[22]检测头定位目标.

近年来,受到Transformer[26]在目标检测[27]中应用的启发,研究者提出了基于Transformer的一阶段视觉定位方法. Deng等人[13]提出的TransVG利用Transformer编码器提取图像和文本特征,用一个有6层Transformer编码层的特征融合模块交互图像与文本特征并学习一个特征向量用于视觉定位. Du等人[14]提出的VGTR和Zhu等人[15]提出SeqTR利用基于Transformer的编解码网络实现视觉定位,其中编码器用于图像文本信息融合,解码器提取特征用于视觉定位. Yang等人[16]提出的VLTVG利用Transformer编码器和Bert[23]分别提取图像和文本特征,利用基于多头注意力的视觉语言验证模块、语言引导的上下文编码器和多阶段跨模态解码器交互融合图像和文本特征用于视觉定位. 为了应对基于Transformer的视觉定位的过拟合问题,Qu等人[17]提出一种SiRi训练机制提升视觉定位的精度. Li等人[18]利用Transformer实现特征编码器、跨模态交互器和模态无关解码器三个模块组成的视觉定位方法,取得较好视觉定位效果. 从上述分析可以看

出：虽然基于Transformer的一阶段视觉定位方法提升了视觉定位性能，但是这些方法[13~18]在用Transformer编码器或解码器融合图像与文本特征之前，没有进行图像与文本特征的对齐，这限制了视觉定位精度的提升；另一方面，使用Transformer模型使得这些方法的训练需要更多的数据，并且训练收敛速度较慢.

3 对比学习预训练大模型

2021年，Radford等人[19]提出了语言图像对比学习预训练CLIP模型，该模型使用对比学习和大规模图像文本对（4亿个图像文本对）学习图像文本特征，在动作识别等下游任务上取得了优异效果. 如图1所示，CLIP模型由文本编码器和图像编码器组成，其中文本编码器由N层Transformer编码层[26]组成，图像编码器采用ResNet-50[28]或者ViT[29]. 文本信息经过BPE[30]编码并在首尾加入sot和eot两个词向量作为文本编码器的输入词向量. 文本词向量通过文本编码器后，eot所对应的输出eot′表示文本全局特征. 图像编码器采用ViT，输入图像分割成块加上位置编码形成图像词向量，在图像词向量头部加入cls词向量形成图像编码器的输入，图像词向量经过ViT后，cls对应的输出cls′表示图像全局特征. 计算eot′和cls′的余弦相似度，使用交叉熵损失函数训练CLIP网络模型.

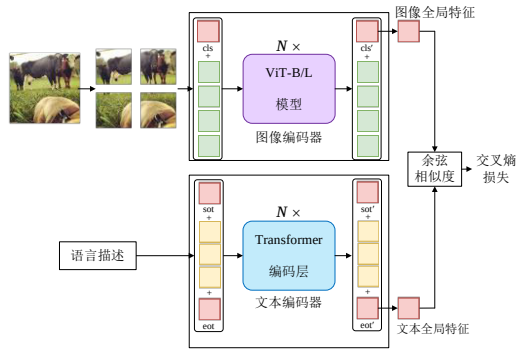


图1 CLIP模型的组成与结构

为了说明CLIP模型学习的视觉特征具有视觉定位能力，本文利用Grad-CAM++[31]方法生成图像的激活图，并根据激活图得到目标框，如图2所示. 具体方法是：首先使用图像编码器中对比损失的梯度与学习到的图像特征做哈达玛（Hadamard）积运算得到激活值，然后将激活值恢复成二维激活图，其次根据设定的阈值将激活图二值化，最后根据二值化图像拟合目标框.

图2中第1列为原始图像，第2列为原始图像的激活图，第3列为原始图像与激活图的复合图，第4列为视觉定位结果和目标真值（绿色为定位框，红色为真值框）. 从中可以看出：利用CLIP模型学习的图像特征激活图的语义信息具有较强的显著性，可以利用激活图进行视觉定位.

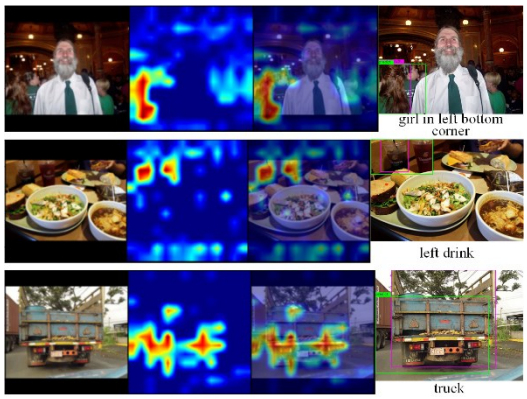


图2 激活图与视觉定位结果

4 视觉定位方法

由第3节的分析可知：语言图像对比学习预训练CLIP大模型学习到的视觉特征具有视觉定位能力，但是直接用于视觉定位精度不高，原因在于CLIP模型的训练目标不是视觉定位，文本和图像之间的信息交互不足. 因此在CLIP模型的基础上，根据视觉定位目标设计文本图像特征融合模块和目标框预测头，从而形成一体化网络用于视觉定位，下面对此进行详细阐述.

4.1 网络结构

由图1中CLIP模型的组成与结构可知：CLIP模型的最后一层嵌入向量不但包含图像和文本特征，而且保留了位置区域信息. 因此可以在视觉定位损失的监督下，用Transformer对CLIP模型学习的图像和文本特征进一步交互融合，学出用于视觉定位的特征向量. 基于此思路，本文设计的视觉定位网络CLIPVG模型的组成与结构如图3所示，包括文本图像特征提取器、文本图像特征融合器和目标框预测头.

（1）文本图像特征提取器. 文本图像特征提取器负责文本与图像信息的特征提取，如图3所示，它包括文本编码器和图像编码器两部分，其结构与CLIP模型的文本编码器和图像编码器结构相同，由12层Transformer编码层[27]构成. 文本编码器将文本$\boldsymbol{l}$编码成文本特征$\boldsymbol{z}_l \in \mathbb{R}^{C_l \times N_l}$，其中$C_l$和$N_l$分别是文本特征维度和数量. 图像编码器将图像$\boldsymbol{v} \in \mathbb{R}^{3 \times H_0 \times W_0}$编码为图像特征$\boldsymbol{z}_v \in \mathbb{R}^{C_v \times N_v}$，其中$C_v$和$N_v$是图像特征的维度和数量.

（2）文本图像特征融合器. 文本图像特征融合器的功能是融合文本和图像特征，由两个线性映射层和一个文本图像特征融合模块组成，其中文本图像特征融合模块由N层Transformer编码层[27]组成. 文本特征$\boldsymbol{z}_l$和图像特征$\boldsymbol{z}_v$经过线性映射层得到文本嵌入$\boldsymbol{p}_l \in \mathbb{R}^{C_p \times N_l}$和图像嵌入$\boldsymbol{p}_v \in \mathbb{R}^{C_p \times N_v}$. 将$\boldsymbol{p}_l$、$\boldsymbol{p}_v$和可学习的编码向量$\boldsymbol{p}_{\text{reg}}$拼接在一起，形成文本图像特征融合器的输入$\boldsymbol{x}_0 =$

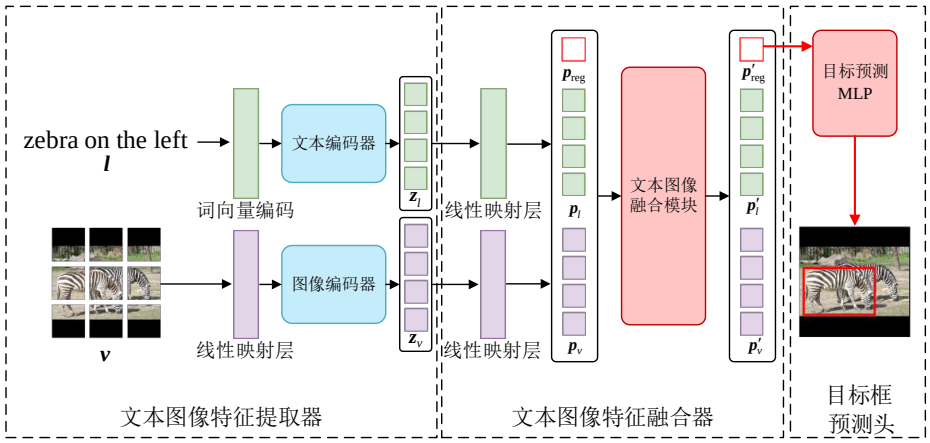


图3　CLIPVG的组成与结构

$(\boldsymbol{p}_{\text{reg}},\boldsymbol{p}_l,\boldsymbol{p}_v)$,其中,$\boldsymbol{p}_{\text{reg}}$在训练开始时随机初始化并随着模型一起优化.$\boldsymbol{x}_0$经过文本图像特征融合模块,得到融合特征$\boldsymbol{x}_1=(\boldsymbol{p}'_{\text{reg}},\boldsymbol{p}'_l,\boldsymbol{p}'_v)$,其中,$\boldsymbol{p}_{\text{reg}}$对应的输出$\boldsymbol{p}'_{\text{reg}}$汇集了文本和图像上下文信息的融合表示,用于目标框的预测.

(3)目标框预测头.目标框预测头是一个具有2层隐藏层和1个输出层的多层感知机,其功能是利用融合特征$\boldsymbol{p}'_{\text{reg}}$回归目标框信息.具体方法是将$\boldsymbol{p}'_{\text{reg}}$送入多层感知机中,由多层感知机回归输出目标框的中心坐标,以及宽和高.

4.2　损失函数

网络训练的目的是得到目标框的中心坐标以及宽和高,此问题利用回归技术实现,因此利用回归损失作为损失函数.由于Smooth $L1$损失计算简单,有利于训练快速收敛,GIoU损失具有尺度不变性,这两个损失可以互补.因此,本文采用Smooth $L1$损失与GIoU损失的组合作为网络训练的损失函数.假设$\boldsymbol{b}=(x,y,w,h)$为标注框,其中,(x,y),w,h分别是标注框的中心坐标、宽和高,$\hat{\boldsymbol{b}}=(\hat{x},\hat{y},\hat{w},\hat{h})$为预测框,其中,$(\hat{x},\hat{y})$,$\hat{w},\hat{h}$分别是预测框的中心坐标、宽和高,则损失函数表示如下:

$$L=L_{\text{smooth-}L1}(\boldsymbol{b},\hat{\boldsymbol{b}})+\lambda L_{\text{giou}}(\boldsymbol{b},\hat{\boldsymbol{b}}) \tag{1}$$

其中λ为平衡两种损失的权重参数,本文在实验中设置为1.

4.3　实现细节

网络训练开始,用预训练的CLIP模型[19]的文本编码器和图像编码器初始化CLIPVG模型的文本编码器和图像编码器,在训练过程中固定其权重参数.CLIPVG模型的规模相较于CLIP模型较小,因此在实验中使用精度更高的数据类型FP32,而非原CLIP模型中的FP16. CLIP模型中图像编码器ViT有多个不同的版本,本文采用基础模型ViT-B/16和大模型ViT-L/14-336进行实验.在使用ViT-B/16模型时,输入图像归一化为224×224并划分成16×16大小的图像块,文本图像特征融合器的线性映射层的输出大小设置为512. 在使用ViT-L/14-336模型时,输入图像归一化为336×336并划分成14×14大小的图像块,文本图像特征融合器的线性映射层的输出大小设置为1 024. 网络训练的学习率为10^{-4},对于RefCOCO+数据集总共训练180轮,并在第120轮后学习率以0.1系数进行衰减.对于其它数据集,每个数据集共训练100轮,并在第80轮后学习率以0.1系数衰减.为了避免过拟合,在文本图像特征融合模块的每个Transformer编码层的多头自注意力层和FFN层均使用0.1的dropout. 模型训练的数据增强方法与文献[11,24,25]相同.

5　实验与分析

5.1　实验条件

本文实验仅使用一张NVIDIA GeForce RTX3090 GPU,相较于大模型的训练与优化,硬件要求较低.在系统软件方面,操作系统要求ubuntu20.04版本以上,Python版本为3.7.12,pytorch版本为1.7.1,CUDA版本为11.0.

5.2　数据集

本文利用ReferItGame、Flickr30K Entities、RefCOCO、RefCOCO+和RefCOCOg等5个数据集进行实验.数据集ReferItGame[32]中的图像来自数据集SAIAPR-12[33],共有20 000张图像,其中训练集包含54 127个图像文本对,验证集包含5 842个图像文本对,测试集包含60 103个图像文本对.数据集Flickr30K Entities[34]是由数据集Flickr30K[35]通过短语标注改造得到,共包含31 783张图像和42.7万个文本表述,其中训练集包含427 193个图像文本对,验证集包含14 433个图像文本对,测试集包含14 481个图像文本对.数据集RefCOCO、RefCOCO+和RefCOCOg均由2014年版

COCO数据集[36]加上相应文本表述构建而成. RefCOCO数据集[37]包含19 994图像和50 000个被描述的物体,文本表述共有142 210条,其训练集包含120 624个图像文本对、验证集包含10 834个图像文本对、测试集A包含5 657个图像文本对、测试集B包含5 095个图像文本对. RefCOCO+数据集[37]包含19 992图像和49 856个被描述的物体,全部表述共有141 564条,其训练集包含120 191个图像文本对、验证集包含10 758个图像文本对、测试集A包含5 726个图像文本对和测试集B包含4 889个图像文本对. RefCOCOg数据集[38]包含25 799个图像和49 822个被描述的物体,每个物体多条表述,共有95 010个图像文本对. 该数据集有RefCOCOg-google[38]和RefCOCOg-umd[39]两种划分方式. RefCOCOg-google的训练集和验证集分别包含85 474和9 536个图像文本对. RefCOCOg-umd的训练集、验证集和测试集分别包含80 512、4 896和9 602个图像文本对.

5.3 消融实验

为说明图像编码器、文本图像特征融合模块中的通道数和Transformer编码层数对视觉定位精度的影响,本文利用RefCOCO数据集进行了消融实验,视觉定位精度采用准确率进行定量评价.

5.3.1 图像编码器实验分析

对采用不同图像编码器的CLIPVG进行对比实验,实验采用的文本编码器是12层Transformer编码器. 表1给出了不同图像编码器的实验结果,可以看出:当图像编码器与文本编码器结构一致时,视觉定位准确率优于异构的编码器结构,也就是图像编码器采用Transformer比采用ResNet的准确率高,对于同一类型的图像编码器随着网络深度的增加,视觉定位准确率也在增加.

表1 不同图像编码器实验结果

图像编码器	定位准确率		
	验证集	测试集A	测试集B
ResNet-50	53.19	61.09	58.71
ResNet-101	55.49	62.24	47.14
RN50×64	58.54	66.64	50.97
ViT-B/32	70.03	76.67	61.14
ViT-B/16	78.54	84.28	70.21
ViT-L/14-336	81.08	86.55	73.15

5.3.2 文本图像特征融合模块通道数实验分析

在文本图像特征融合模块中,采用不同通道数的CLIPVG进行对比实验,实验采用的图像编码器是ViT-B/16. 表2给出了文本图像特征融合模块中不同通道数的实验结果,可以看出:采用256通道的准确率最低,采用512通道的准确率最高,采用1 024通道的准确率处于中间水平,但是不同通道数对定位的准确率影响不大.

表2 文本图像特征融合模块中不同通道数实验结果

融合模块通道数	定位准确率		
	验证集	测试集A	测试集B
256	81.05	86.51	72.93
512	81.97	87.20	73.62
1 024	81.63	86.79	73.56

5.3.3 文本图像特征融合模块中的Transformer编码层实验分析

在文本图像特征融合模块中,采用不同Transformer编码层的CLIPVG进行对比实验,实验采用的图像编码器是ViT-B/16,融合模块通道数均为512. 表3给出了文本图像特征融合模块中编码层数实验结果,可看出:随着Transformer编码层数的增加,定位准确率逐渐提高.

表3 文本图像特征融合模块中Transformer编码层数实验结果

Transformer编码层数	定位准确率		
	验证集	测试集A	测试集B
1	52.71	59.50	44.77
3	79.57	84.25	70.50
6	81.63	86.79	73.56
9	83.01	87.32	75.23

5.4 对比实验

5.4.1 定量实验

定量实验中分别从效率和精度两个方面衡量本文方法的性能. 本文方法选用CLIPVG-B6、CLIPVG-L6、CLIPVG-B9和CLIPVG-L9进行对比实验. CLIPVG-B6和CLIPVG-B9的骨干网络是ViT-B/16,其特征融合模块的Transformer编码层数分别是6和9. CLIPVG-L6和CLIPVG-L9的骨干网络是ViT-L/14-336,其特征融合模块的Transformer编码层数分别是6和9. 表4给出了本文方法与代表性方法TransVG在训练参数量、FlOPs、推理时间三个方面进行的比较结果.

本文所有方法的训练参数量均小于TransVG的训练参数量,即本文方法的训练时间较短. 在FLOPs方面,本文方法CLIPVG-B6和CLIPVG-B9小于TransVG,本文方法CLIPVG-L6和CLIPVG-L9大于TransVG. 在推理时间方面,与TransVG相比,本文方法CLIPVG-B6和CLIPVG-B9的推理时间较短,本文大型模型CLIPVG-L6和CLIPVG-L9的推理时间较长.

精度对比实验将本文视觉定位方法CLIPVG与目前先进视觉定位方法进行对比分析. 实验数据集是ReferItGame、Flickr30K Entities、RefCOCO、RefCOCO+和RefCOCOg. 视觉定位精度采用准确率进行定量评价.

表4　模型计算效率对比实验结果

模型	训练参数量/M	FLOPs/G	推理时间/ms
TransVG-rn50	149.52	40.48	30.84
TransVG-rn101	168.46	70.78	37.40
CLIPVG-B6(本文)	20.11	16.75	17.87
CLIPVG-B9(本文)	29.57	18.48	19.99
CLIPVG-L6(本文)	54.75	137.90	42.46
CLIPVG-L9(本文)	79.95	146.16	46.66

表5给出了在数据集ReferItGame和Flickr30K Entities上的实验结果,可以看出:在ReferItGame和Flickr30K Entities数据集上,本文方法CLIPVG-L9的准确率最优.

表5　ReferItGame和Flickr30K数据集实验结果(红色、蓝色和绿色依次表示前3名)

方法类型	模型	骨干网络	ReferIt-Game	Flickr-30K
两阶段方法	CMN(2017)[3]	VGG16	28.33	—
	VC(2018)[2]	VGG16	31.13	—
	MAttNet(2018)[40]	ResNet101	29.04	—
	Similarity Net(2019)[41]	ResNet101	34.54	60.89
	CITE(2018)[42]	ResNet101	35.07	61.33
	PIRC(2018)[43]	ResNet101	59.13	72.83
	DDPN(2018)[44]	ResNet101	63.00	73.30
一阶段方法	SSG(2018)[45]	DarkNet53	54.24	—
	ZSGNet(2019)[46]	ResNet50	58.63	63.39
	FAOA(2019)[11]	DarkNet53	60.67	68.71
	RCCF(2020)[24]	DLA34	63.79	—
	ReSC-Large(2020)[25]	DarkNet53	64.60	69.28
	LBYL-Net(2021)[12]	DarkNet53	67.47	—
Transformer方法	VGTR(2022)[14]	ResNet101	—	75.32
	SeqTR(2022)[15]	DarkNet53	69.66	81.23
	TransVG(2022)[13]	DETR	70.73	79.10
	CMI(2023)[18]	ResNet101	71.07	79.15
	VLTVG(2022)[16]	ResNet101	71.98	79.84
	CLIPVG-B6(本文方法)	ViT-B/16	66.66	78.92
	CLIPVG-L6(本文方法)	ViT-L/14-336	69.71	81.67
	CLIPVG-B9(本文方法)	ViT-B/16	68.48	79.74
	CLIPVG-L9(本文方法)	ViT-L/14-336	72.36	82.95

表6给出了在数据集RefCOCO、RefCOCO+和RefCOCOg上的结果,可以看出:在这些数据集的4个验证集和5测试集上,本文方法CLIPVG-L9的准确率有7个排名第1,有2个排名第3,本文方法CLIPVG-L6的准确率有7个排名第2.

通过上述对比实验可以看出:基于Transformer的方法普遍优于其他的方法,而相比于其它基于Transformer的方法,本文方法取得了综合精度的提升. 其原因在于:本文方法使用了基于对比学习的CLIP网络参数初始化图像与文本编码器网络参数,并在整体网络训练中冻结. 对比学习大模型CLIP使得提取的图像和文本特征处于同一语义空间,这有利于后期特征融合实现视觉定位,而其他方法在图像与文本特征融合前未进行特征的对齐.

5.4.2　定性实验

图4给出了本文方法CLIPVG-L6对16张图像的定位结果(绿色为定位框,红色为真值框). 它们分别从目标属性识别、同类物体区分、空间关系理解、复杂语言描述理解和复杂图像背景干扰抑制等角度展现本文方法性能. 图4(a)和图4(d)的语言描述中包含"eating"和"yellow"分别指示出目标的动作属性和颜色属性,这证明本文方法对目标属性的识别能力较强. 图4(e)~

表6 RefCOCO、RefCOCO+和RefCOCOg数据集实验结果(红色、蓝色和绿色依次表示前3名)

方法类型	模型	骨干网络	RefCOCO			RefCOCO+			RefCOCOg		
			验证集	测试集A	测试集B	验证集	测试集A	测试集B	验证集-g	验证集-u	测试集-u
两阶段方法	CMN(2017)[3]	VGG16	—	71.03	65.77	—	54.32	47.76	57.47	—	—
	VC(2018)[2]	VGG16	—	73.33	67.44	—	58.40	53.18	62.30	—	—
	ParalAttn(2018)[6]	VGG16	—	75.31	65.52	—	61.34	50.86	58.03	—	—
	MAttNet(2018)[40]	ResNet-101	76.65	81.14	69.99	65.33	71.62	56.02	—	66.58	67.27
	LGRANs(2019)[7]	VGG16	—	76.60	66.40	—	64.00	53.40	61.78	—	—
	DGA.(2019)[8]	VGG16	—	78.42	65.53	—	69.07	51.99	—	—	63.28
	RvG-Tree(2019)[47]	ResNet-101	75.06	78.61	69.85	63.51	67.45	56.66	—	66.95	66.51
	NMTree(2019)[10]	ResNet-101	76.41	81.21	70.09	66.46	72.02	57.52	64.62	65.87	66.44
	CM-Att-Erase(2019)[5]	ResNet-101	78.35	83.14	71.32	68.09	73.65	58.03	—	67.99	68.67
一阶段方法	SSG(2018)[45]	DarkNet-53	—	76.51	67.50	—	62.14	49.27	47.47	58.80	—
	FAOA(2019)[11]	DarkNet-53	72.54	74.35	68.50	56.81	60.23	49.60	56.12	61.33	60.36
	RCCF(2020)[24]	DLA-34	—	81.06	71.85	—	70.35	56.32	—	—	65.73
	ReSC-Large(2020)[25]	DarkNet-53	77.63	80.45	72.30	63.59	68.36	56.81	63.12	67.30	67.20
	LBYL-Net(2021)[12]	DarkNet-53	79.67	82.91	74.15	68.64	73.38	59.49	62.70	—	—
Transformer方法	VGTR(2022)[14]	ResNet-101	79.30	82.16	74.38	64.40	70.85	55.84	64.05	66.83	67.28
	TransVG(2022)[13]	DETR	81.02	82.72	78.35	64.82	70.70	56.94	67.02	68.67	67.73
	SeqTR(2022)[15]	DarkNet-53	83.72	86.51	81.24	71.45	76.26	64.88	71.50	74.86	74.21
	VLTVG(2022)[16]	ResNet-101	84.77	87.24	80.49	74.19	78.93	65.17	72.98	76.04	74.18
	CMI(2023)[18]	ResNet-101	81.92	83.40	77.37	68.49	72.18	60.30	68.39	69.08	69.04
	CLIPVG-B6(本文方法)	ViT-B/16	81.63	86.79	73.56	69.67	79.22	55.61	68.91	71.26	70.94
	CLIPVG-L6(本文方法)	ViT-L/14-336	85.54	89.29	77.78	76.19	84.09	63.86	76.12	77.02	76.71
	CLIPVG-B9(本文方法)	ViT-B/16	83.01	87.32	75.23	70.02	80.89	57.01	70.91	72.86	71.95
	CLIPVG-L9(本文方法)	ViT-L/14-336	86.12	89.41	78.51	77.51	85.02	64.02	77.01	78.08	77.31

(*j*)的图像存在多个同类物体“giraffe”,“dog”和“bike”,自然语言的不同描述让对象间产生区别,这证明本文方法能在同类物体间加以区分. 图4(*b*)和图4(*c*)展示出本文方法对“bottom right corner”的方位关系和“second from left”的顺序关系有较好的理解能力. 图4(*g*)、图4(*k*)和图4(*l*)等结果均有较为复杂的自然语言描述,说明本文方法能够从复杂描述中解析出关键语义信息实现定位. 图4(*m*)~(*p*)中图像背景较复杂给定位带来干扰,本文方法能够有效抑制背景干扰. 从这些结果中可看出:尽管图像中存在相似物体和复杂背景干扰等困难,但是本文方法能够准确地定位目标.

图5给出了本文方法CLIPVG-L6部分代表性的失败定位结果(绿色为定位框,红色为真值框). 图5(*a*)的语言描述“left”仅指出方位并没有指明目标种类属性,图5(*b*)由于建筑物“building”所对应的地基边界不清导致了定位范围无法准确预测,图5(*c*)中密集物体遮挡与目标语言描述不明确对定位造成困难,图5(*d*)中语言描述“18”提供的信息有限,缺乏对目标的种类属性的表达. 从中可以看出:查询文本描述存在着语义歧义和定位范围边界模糊等问题,这是造成本文方法失败的主要原因.

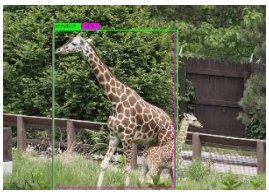
图6给出了本文方法CLIPVG-L6的zero-shot视觉定位结果. 在实验中,使用了一些训练数据集中没有的目标进行视觉定位测试. 从图6中视觉定位结果可以看出本文方法对于开放世界中的一些场景和概念,具有zero-shot视觉定位能力,其原因在于本文方法使用的大规模预训练CLIP模型带来的能力.

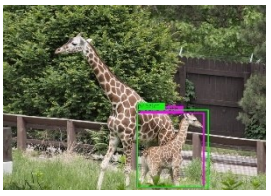

(*a*) Dude eating in white

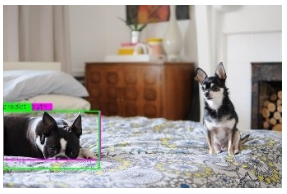

(*b*) Bottom right corner man

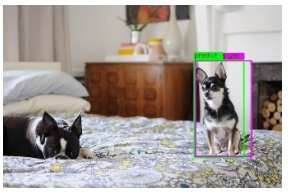

(*c*) Person second from left

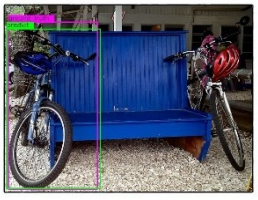

(*d*) Yellow guy

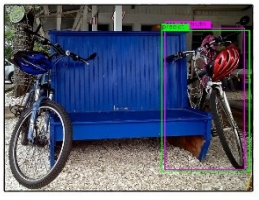

(*e*) Mommy giraffee

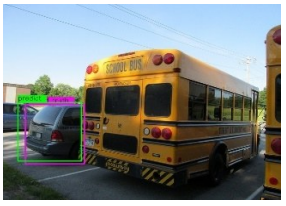

(*f*) Baby giraffe

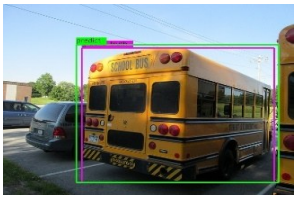

(*g*) A dog which is lying on the bed with standing an-other one


(*h*) Small black and tan dog sitting up


(*i*) Bike on left


(*j*) Right bike


(*k*) Gray van with tinted windows in parking lot


(*l*) A school bus that is of a grey car and another bus


(*m*) Big donut slicet


(*n*) Right bike

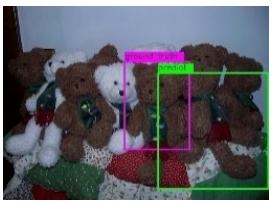

(*o*) A portion of food with carrots and cheese slices


(*p*) A sweet foo in an orange bowl

图4　本文方法CLIPVG-L6对16张图像的视觉定位结果

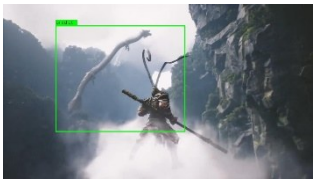

(*a*) Left


(*b*) Building


(*c*) Front right


(*d*) 18

图5　本文方法的失败定位结果


(*a*) A white dragon in the sky


(*b*) A astronaut on the bottom


(*c*) Sun WuKong


(*d*) Flying space station

图6　本文方法的zero-shot视觉定位结果

6 结论

针对现有视觉定位方法的不足,本文提出一种基于对比学习大模型的一阶段视觉定位方法. 该文方法的主要创新之处是提出利用基于对比学习的大规模语言视觉模型提取文本和图像特征并进行特征的对齐,然后再利用Transformer交互融合对齐的图像文本特征进行视觉定位,可以利用较少的GPU硬件实现快速的模型训练. 在5个基准数据集上对本文方法进行了实验验证和分析,实验结果表明本文方法的综合精度优于现有方法.

参考文献

[1] LIU J Y, WANG L, YANG M H. Referring expression generation and comprehension via attributes[C]//2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2017: 4866-4874.

[2] ZHANG H W, NIU Y L, CHANG S F. Grounding referring expressions in images by variational context[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 4158-4166.

[3] HU R H, ROHRBACH M, ANDREAS J, et al. Modeling relationships in referential expressions with compositional modular networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2017: 1115-1124.

[4] YE J B, LIN X, HE L, et al. One-stage visual grounding via semantic-aware feature filter[C]//Proceedings of the 29th ACM International Conference on Multimedia. New York: ACM, 2021: 1702-1711.

[5] LIU X H, WANG Z H, SHAO J, et al. Improving referring expression grounding with cross-modal attention-guided erasing[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2019: 1950-1959.

[6] ZHUANG B H, WU Q, SHEN C H, et al. Parallel attention: A unified framework for visual object discovery through dialogs and queries[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 4252-4261.

[7] WANG P, WU Q, CAO J W, et al. Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2019: 1960-1968.

[8] YANG S B, LI G B, YU Y Z. Dynamic graph attention for referring expression comprehension[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2019: 4644-4653.

[9] CIRIK V, BERG-KIRKPATRICK T, MORENCY L P. Using syntax to ground referring expressions in natural images[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI, 2018: 6756-6764.

[10] LIU D Q, ZHANG H W, ZHA Z J, et al. Learning to assemble neural module tree networks for visual grounding [C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2019: 4673-4682.

[11] YANG Z Y, GONG B Q, WANG L W, et al. A fast and accurate one-stage approach to visual grounding[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2019: 4682-4692.

[12] HUANG B B, LIAN D Z, LUO W X, et al. Look before you leap: Learning landmark features for one-stage visual grounding[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 16888-16897.

[13] DENG J J, YANG Z Y, CHEN T L, et al. TransVG: End-to-end visual grounding with Transformers[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2021: 1769-1779.

[14] DU Y, FU Z H, LIU Q J, et al. Visual grounding with Transformers[C]//2022 IEEE International Conference on Multimedia and Expo (ICME). Piscataway: IEEE, 2022: 1-6.

[15] ZHU C Y, ZHOU Y Y, SHEN Y H, et al. SeqTR: A simple yet universal network for visual grounding[M]//Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2022: 598-615.

[16] YANG L, XU Y, YUAN C F, et al. Improving visual grounding with visual-linguistic verification and iterative reasoning[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 9499-9508.

[17] QU M X, WU Y, LIU W, et al. SiRi: A simple selective retraining mechanism for Transformer-based visual grounding[M]//Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2022: 546-562.

[18] LI K, LI J X, GUO D, et al. Transformer-based visual grounding with cross-modality interaction[J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2023, 19(6): 1-19.;

[19] RADFORD A, KIM J W, HALLACY C, et al. Learning

transferable visual models from natural language supervision[C]//Proceedings of the 38th International Conference on Machine Learning, ICML 2021. New York: ICML, 2021: 8748-8763.

[20] UIJLINGS J R R, VAN DE SANDE K E A, GEVERS T, et al. Selective search for object recognition[J]. International Journal of Computer Vision, 2013, 104(2): 154-171.

[21] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[22] REDMON J, FARHADI A. YOLOv3: An incremental improvement[EB/OL]. (2018-04-08)[2022-09-27]. https://arxiv.org/abs/1804.02767.

[23] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional Transformers for language understanding[C]//Proceedings of NAACL-HLT. Stroudsburg: ACL, 2019: 4171-4186.

[24] LIAO Y, LIU S, LI G B, et al. A real-time cross-modality correlation filtering method for referring expression comprehension[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 10880-10889.

[25] YANG Z Y, CHEN T L, WANG L W, et al. Improving one-stage visual grounding by recursive sub-query construction[M]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020: 387-404.

[26] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 6000-6010.

[27] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with Transformers[M]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020: 213-229.

[28] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 770-778.

[29] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale[EB/OL]. (2020-10-21)[2022-03-09].https://arxiv.org/abs/2010.11929, arXiv: 2010.11929.

[30] SENNRICH R, HADDOW B, BIRCH A. Neural machine translation of rare words with subword units[EB/OL]. (2016-06-10)[2022-10-24]. https://arxiv.org/abs/1508.07909.

[31] CHATTOPADHAY A, SARKAR A, HOWLADER P, et al. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks[C]//2018 IEEE Winter Conference on Applications of Computer Vision (WACV). Piscataway: IEEE, 2018: 839-847.

[32] KAZEMZADEH S, ORDONEZ V, MATTEN M, et al. ReferItGame: Referring to objects in photographs of natural scenes[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg, PA, USA: Association for Computational Linguistics, 2014: 787-798.

[33] ESCALANTE H J, HERNÁNDEZ C A, GONZALEZ J A, et al. The segmented and annotated IAPR TC-12 benchmark[J]. Computer Vision and Image Understanding, 2010, 114(4): 419-428.

[34] PLUMMER B A, WANG L W, CERVANTES C M, et al. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models[C]//2015 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2015: 2641-2649.

[35] YOUNG P, LAI A, HODOSH M, et al. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions[J]. Transactions of the Association for Computational Linguistics, 2014, 2: 67-78.

[36] JOSEPH C R. Common objects in context dataset mirror [EB/OL]. [2014-09-01][2022-10-25].https://pjreddie.com/projects/coco-mirror/.

[37] YU L C, POIRSON P, YANG S, et al. Modeling context in referring expressions[M]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2016: 69-85.

[38] MAO J H, HUANG J, TOSHEV A, et al. Generation and comprehension of unambiguous object descriptions[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 11-20.

[39] NAGARAJA V K, MORARIU V I, DAVIS L S. Modeling context between objects for referring expression understanding[M]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2016: 792-807.

[40] YU L C, LIN Z, SHEN X H, et al. MAttNet: Modular attention network for referring expression comprehension [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 1307-1315.

[41] WANG L W, LI Y, HUANG J, et al. Learning two-branch neural networks for image-text matching tasks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(2): 394-407.
[42] PLUMMER B A, KORDAS P, KIAPOUR M H, et al. Conditional image-text embedding networks[M]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018: 258-274.
[43] KOVVURI R, NEVATIA R. PIRC Net: Using proposal indexing, relationships and context for phrase grounding [M]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019: 451-467.
[44] YU Z, YU J, XIANG C, et al. Rethinking diversified and discriminative proposal generation for visual grounding [EB/OL]. (2018-05-09) [2022-06-28]. https://arxiv. org/abs/1805.03508.
[45] CHEN X, MA L, CHEN J, et al. Real-time referring expression comprehension by single-stage grounding network[EB/OL]. (2018-12-09) [2022-09-28]. https://arxiv.org/abs/1812.03426.
[46] SADHU A, CHEN K, NEVATIA R. Zero-shot grounding of objects from natural language queries[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2019: 4694-4703.
[47] HONG R, LIU D, MO X, et al. Learning to compose and reason with language tree structures for visual grounding [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(2): 684-696.

作者简介

陆庆阳 男,1994年生,安徽合肥人. 陆军炮兵防空兵学院硕士研究生,主要研究方向为计算机视觉领域的多模态目标跟踪及视觉计数.
E-mail: lqy465813@163.com

袁广林 男,1973年生,河南周口人,博士,教授. 主要从事计算机视觉、机器学习及其应用方面的研究.
E-mail: yuangl_plus@126.com

朱 虹 女,1987年生,河北博野人,硕士,目前正在国防科技大学攻读博士学位,主要研究方向为视觉定位和视觉跟踪.
E-mail: candy_zhuhong@126.com

秦晓燕 女,1980年生,安徽淮北人,副教授. 主要研究方向为目标检测、机器学习及应用.
E-mail: xiaoyanqin_hf@163.com

薛模根 男,1964年生,安徽合肥人,博士,教授. 现任中国人民解放军陆军炮兵防空学院正教授、安徽省偏振成像探测技术重点实验室主任. 主要从事图像处理、光电检测和物体跟踪方面研究.