基于Transformer的视觉分割技术进展

李文生[1)] 张 菁[1),2)] 卓 力[1),2)] 吴鑫嘉[1)] 闫 伊[1)]

[1)](北京工业大学信息科学技术学院 北京 100124)

[2)](北京工业大学计算智能与智能系统北京市重点实验室 北京 100124)

摘 要 视觉分割是计算机视觉领域的核心任务，旨在将图像或视频帧中的像素分类以划分成不同区域.得益于视觉分割技术的快速发展，该技术在自动驾驶、航空遥感和视频场景理解等多种应用领域中发挥着关键作用.近年来，基于Transformer的视觉分割技术因具备长程依赖建模能力而备受关注.随着Transformer的模型架构的持续优化与迭代，亟须更全面地理解和认识Transformer在视觉分割领域的已有进展和发展趋势，通过发现现有研究中的不足和挑战，以更深入地探索Transformer的核心理论.为此，本文从图像/视频两个视觉脉络出发，整理、回顾、分析和探讨了近年来基于Transformer的视觉分割相关技术进展，不仅归纳了Transformer的理论框架，还给出了一些应用实例和研究热点，从而做出总结和展望.具体来说，首先梳理了Transformer的背景，包括问题定义、数据集和评估指标、基本结构，其中，问题定义描述了视觉分割在图像/视频任务中的预期目标和结果；数据集和评估指标反映了模型的具体应用场景，以及性能的衡量标准；基本结构则描述了算法的核心模块、实现流程以及各个模块之间的关系.然后，着重阐述了Transformer在图像语义分割、图像实例分割，以及视频语义分割和视频实例分割四个方法体系，并探讨了当前的研究热点.对于图像语义分割任务，分析了Transformer的代表性结构，包括纯Transformer和双分支结构，并以无人机影像非铺装道路分割和遥感图像语义分割为实际应用案例，探讨了Transformer的改进动机与应用效果，并展示了主观结果；图像实例分割总结了常见的非端对端Transformer和端对端Transformer典型结构.视频语义分割主要分为面向精度的和面向效率的Transformer结构，视频实例分割则包括逐帧和逐片段Transformer分割，并以网络直播视频实例分割为应用实例，一方面讨论了可用的数据集、实验参数和评估指标，另一方面，对网络直播视频实例分割主流方法性能进行了评价和分析，展示了一些主观可视化结果.之后，鉴于视觉分割领域的SAM大模型、开放词汇分割、指代分割受到了广泛关注，本文将这些热点问题方法进行了追溯和评述，以期碰撞出视觉分割的新思路和新灵感.最后，尽管基于Transformer在视觉分割技术受到了广泛的关注，但存在的科学问题也逐渐凸显，限制了模型性能与效率的进一步提升，对此本文总结了利用Transformer开展图像/视频语义/实例分割仍需关注的难点问题，并对未来可能的发展方向进行了展望，提供了一些启示供参考.

关键词 视觉分割；Transformer；语义分割；实例分割；自注意力机制

中图法分类号 TP391 **DOI号** 10.11897/SP.J.1016.2024.02760

Overview of Transformer-Based Visual Segmentation Techniques

LI Wen-Sheng[1)] ZHANG Jing[1),2)] ZHUO Li[1),2)] WU Xin-Jia[1)] YAN Yi[1)]

[1)](*School of Information Science and Technology, Beijing University of Technology, Beijing* 100124)

[2)](*Beijing Key Laboratory of Computational Intelligence and Intelligent System, Beijing University of Technology, Beijing* 100124)

Abstract In the field of computer vision, visual segmentation is a fundamental task that categorizes pixels in an image or video frame into distinct regions. Thanks to the significant

收稿日期：2023-12-15；在线发布日期：2024-09-24.本课题得到国家自然科学基金(62471013，61971016)、北京市自然科学基金-市教委联合资助项目(KZ201910005007)资助.**李文生**，博士研究生，主要研究领域为视频语义分割.E-mail：liwensheng@emails.bjut.edu.cn.**张 菁**(通信作者)，博士，教授，中国计算机学会(CCF)会员，主要研究领域为图像\视频处理、深度学习、计算机视觉.E-mail：zhj@bjut.edu.cn.**卓 力**，博士，教授，主要研究领域为图像\视频处理、深度学习、计算机视觉.**吴鑫嘉**，硕士研究生，主要研究领域为遥感图像分割.**闫 伊**，硕士研究生，主要研究领域为遥感图像分割.

development of visual segmentation techniques, it plays a key role in various applications such as autonomous driving, aerial remote sensing, and video scene understanding. In recent years, Transformer-based visual segmentation has attracted much attention because of its long-range dependency modeling capability. With the continuous optimization and updating of Transformer's model architecture, there is an urgent need to more comprehensively understand and recognize the existing progress and development trend of Transformer in field of visual segmentation, and to find out the deficiencies and challenges, so as to explore the core theory of Transformer in a deeper way. To this end, this paper organizes, reviews, analyzes and explores the recent advances in Transformer-based visual segmentation techniques from two visual pipelines of image/video, not only summarizing the theoretical framework of Transformer, but also giving some application examples and research hotspots, so as to make a summary and overlook. Specifically, the background of the Transformer is initially reviewed, including problem definition, datasets, indicators, and the basic structure, in which the problem definition describes the expected goals and results of visual segmentation in image/video tasks; the dataset and indicators respond to the specific application scenarios of the model as well as the performance measures; the basic structure describes the core modules of the algorithm, the implementation process, and the relationship between the individual module. Then, the four methodologies of Transformer are highlighted in detail in terms of image semantic and instance segmentation, as well as the video semantic and instance segmentation, and current research hotspots are discussed. For the task of image semantic segmentation, the representative structures of Transformer are analyzed, including pure Transformer and dual-branch structures, and the motivation and application effect of Transformer's improvement are exhibited and the visual results are shown with the practical application cases of unpaved road segmentation of UAV images and semantic segmentation of remote sensing images, while image instance segmentation summarizes the typical structure of Transformer without/with end-to-end framework. Video semantic segmentation is mainly categorized into accuracy-oriented and efficiency-oriented Transformer structures, while video instance segmentation includes frame-by-frame and segment-by-segment Transformer structure. Notably, video instance segmentation takes livestreaming video instance segmentation as an application example, and not only discusses the available datasets, experimental parameters and indicators, but also evaluates and analyzes the performance of the mainstream methods for livestreaming video instance segmentation, and shows some visual results. Subsequently, for segment anything (SAM), open vocabulary segmentation, and referring segmentation, which are widely concerned in the field of visual segmentation, this paper traces and reviews these hotspots, with a view to colliding new ideas and inspirations in visual segmentation. Finally, although Transformer-based visual segmentation has received widespread attention, the scientific problems have gradually emerged, limiting the further improvement of model performance and efficiency. Finally, this paper summarizes the changeable issues that still need to be addressed in terms of image/video semantic/instance segmentation tasks using Transformer, and looks forward to the potential future development directions to provide some insights for reference.

Keywords visual segmentation; Transformer; semantic segmentation; instance segmentation; self-attention mechanism

1 引 言

作为计算机视觉领域的核心任务之一，视觉分割技术[1]在理解和解释视觉信息方面扮演着至关重要的角色. 该技术将图像或视频帧中的像素进行分类并划分成多个区域，从而提供对场景结构和物体边界的深入理解，不仅有助于识别和理解图像中的各个对象，还能够揭示这些对象之间的空间关系，在机器人导航、视频监控、增强现实等多个领域发挥着不可或缺的作用. 然而，视觉分割技术在实际应用中面临着诸多挑战. 例如，光照条件的多变性和场景的复杂性增加了分割任务的难度. 此外，不同物体之间的视觉特征可能高度相似，这也给精确分割带来了阻碍. 总体而言，复杂的背景、光照变化、遮挡等因素都可能导致分割结果的不准确或不连续. 为了应对这些挑战，研究者们不断探索新的方法和技术，提高视觉分割技术的精确度和鲁棒性成为了研究者关注的焦点.

近年来，基于深度学习的视觉分割方法层出不穷，从卷积神经网络(Convolutional Neural Networks，CNNs)[2]系列一路发展到Transformer系列. CNNs方法利用其独特的卷积层结构有效提取图像局部特征，在图像和视频分割领域得到了广泛应用. 在图像分割过程中，CNNs通过多层卷积和池化操作，能够从简单的边缘和纹理特征逐渐抽象出更复杂的图像内容，如基于神经结构搜索网络的混合CNN-Transformer模型(Hybrid CNN-Transformer Model based on a Neural Architecture Search Network，HCT-net)[3]、融合Transformer和CNN的图像分割结构(Fusing Transformer and CNN Structure for Image Segmentation，DualSeg)[4]等. 在视频分割方面，CNNs通过分析连续帧来理解视频序列中的动态变化，如物体导向的视频分割实例分割方法(Instance Motion for Object-Centric Video Segmentation，InstMove)[5]、城市洪水分割方法(Video Segmentation for Urban Flood Detection and Quantification，V-FloodNet)[6]等. 在我们之前的工作中[7]，基于CNN提出了一种用于视网膜血管分割的双路径网络以提高视网膜血管的完整性，在视网膜血管提取数据集(Digital Retinal Images for Vessel Extraction，DRIVE)和英格兰儿童心脏与健康研究数据集(Child Heart and Health Study in England Dataset，CHASE_DB1)上分别取得了81.9%和86%的准确率. 考虑到图像异常数据的干扰，提出了一种基于图像变换替代任务的图像异常检测方法[8]，并将其应用于颞骨颈静脉球CT图像中骨壁缺失的检测，取得了99.5%的准确率. 尽管CNNs提取的局部特征提高了图像和视频分割的准确性，然而其建模长距离依赖关系的能力有限，这会影响图像/视频分割性能的进一步提升.

近年来，研究人员将Transformer[9]应用于计算机视觉任务中，并逐渐成为一个热门研究方向，已被广泛应用于目标检测[10]、视觉分割[11]、图像/视频生成[12]、视频问答[13]等多个领域. 其核心的自注意力机制(self-attention mechanism)能够有效地捕捉图像中的全局信息，在全局范围内动态地计算图像或视频中不同部分的相互关系，这对于理解复杂的视觉内容至关重要. 在视觉分割领域，Zheng等[14]首次将Transformer应用于图像语义分割，并迅速吸引了广泛的关注. 该方法展示了Transformer在处理图像分割任务时强大的特征表征能力和长距离依赖的建模能力，并获得了广泛的关注. 随后，研究者们开始将Transformer技术应用于更广泛的视觉分割任务，如图像实例分割[15]、视频语义及实例分割[16]等. 可见，Transformer不仅在自然语言处理领域取得了革命性的进展，其在计算机视觉领域的应用也已成为主流趋势.

随着基于Transformer的视觉分割技术的迅速发展，这一领域的研究工作已经取得了前所未有的成果. 一些研究者对这些工作进行了系统的整理和分析，帮助读者快速了解该领域的当前状态、主要发现和理论进展[17]. 这些综述文章通常会根据训练数据中标签的使用方式，将基于Transformer的视觉分割方法划分为基于无监督的、半监督的、有监督的方法[18]. 此外，还有一些工作会专门总结特定领域的Transformer视觉分割方法，为特定应用领域提供了深入的洞察和指导. 尽管这些综述文章为理解Transformer在视觉分割领域的应用提供了宝贵的资源，但目前还缺乏对基于Transformer视觉分割的模型结构优化、自注意力机制改进等热点问题，以及如何实现实时分割、提高模型的泛化能力等难点问题的深入探讨. 在本文中，从图像/视频两个脉络出发，整理了基于Transformer在语义分割和实例分割方法体系，以及相关应用实例，通过系统介绍了各个领域重点关注的热点问题，深入分析了未来的研究方向. 随着研究的不断深入，基于Transformer的视觉分割技术将在未来展现出更加广阔的应用前景和更深远的影响.

2　背　景

本节详细介绍了Transformer在图像和视频处理领域的应用，包括问题定义、数据集和评估指标，以及Transformer的基础架构．问题定义主要阐述语义分割和实例分割的目的和任务；随后展示这两种任务的主要数据集，并探讨用于评估分割性能的关键指标；Transformer基础架构重点介绍自注意力机制和多头注意力等核心组件．

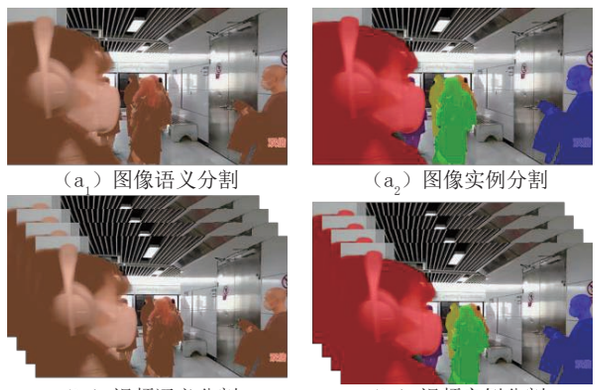
2.1　问题定义

（1）图像分割．给定一个输入图像$\mathbf{I}\in\mathbf{R}^{H\times W\times 3}$，语义分割的目标是输出一组掩码$y_i=(m_i,\ c_i)$，其中$c_i$表示二值图$m_i$的标签，$H\times W$为高度和宽度．每个掩码只包含一个类别标签，其中属于同一类别的所有像素被分配相同的标签，而不考虑它们是否属于同一对象实例[14]．在实例分割任务中，输出的掩码形式为$y_i=(m_i,\ c_i^j)$，其中j表示类别标签为c_i的对象编号[19]．可见，实例分割不仅需要将像素分类到特定类别，还需要区分同一类别对象的不同实例．如图1所示，在语义分割方法中（图1(a_1)），图像中的人物和背景被区分为不同的区域，但图中的两个人物归为同一区域．相比之下，在实例分割中（图1(a_2)）尽管两个人物拥有相同的语义类别，由于它们分别属于不同的实例，因此被标记为独立的区域．

（2）视频分割．给定一个视频片段为$\mathbf{V}\in\mathbf{R}^{T\times H\times W\times 3}$，输出的掩码形式为$\{y_i\}_{t=1}^{T}=\{(m_i,\ c_i^j)\}_{t=1}^{T}$，其中$j$表示类别标签为$c_i$的对象编号[20]．视频实例分割不仅需要区分同一类别对象的不同实例，还要保证同一实例在不同帧上的标签编号保持一致．在视频语义分割（图1(b_1)）和图像语义分割的结果差距不大，都需要将人物和背景区分为不同的区域．然而，视频实例分割（图1(b_2)）在图像实例分割的基础上，需要尽可能使得同一实例在不同视频帧上具有相同的标签类型和编号．

2.2　数据集和评估指标

（1）常用数据集．表1列出了图像和视频分割常用的数据集．常见的图像语义分割数据集，包括波茨坦数据集（ISPRS 2D Semantic Labeling Challenge-Potsdam Data，Potsdam）[21]、法伊欣根数据集（ISPRS 2D Semantic Labeling Challenge-Vaihingen Data，Vaihingen）[22]、视网膜血管提取数据集（Digital Retinal Images for Vessel Extraction，DRIVE）[23]、英格兰儿童心脏与健康研究数据集（Child Heart and Health Study in England Dataset，CHASE_DB1）[24]、遥感影像道路提取数据集（A Challenge to Parse the Earth through Satellite Images，DeepGlobe）[25]、马萨诸塞州道路数据集（Machine Learning for Aerial Image Labeling，Massachusetts）[26]、Pascal可视类别挑战赛数据（The Pascal Visual Object Classes Challenge，Pascal-VOC）[27]、野生物体检测与分割数据集（The Role of Context for Object Detection and Semantic Segmentation in Wild，Pascal Context）[28]、地图远景数据集（The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes，Mapillary）[29]、常见对象数据集（Common Objects in Context，COCO）[30]、场景解析数据集（Scene Parsing through ADE20K Dataset，ADE20K）[31]、城市场景理解数据集（The Cityscapes Dataset for Semantic Urban Scene Understanding，Cityscapes）[32]，其中COCO、ADE20K和Cityscapes



（a_1）图像语义分割　（a_2）图像实例分割　（b_1）视频语义分割　（b_2）视频实例分割

图1　不同视觉任务的比较

表1　图像/视频分割常用的数据集

数据集	训练/测试集	类别数	任务类型
Potsdam	38	6	语义分割
Vaihingen	33	6	语义分割
DRIVE	40	2	语义分割
CHASE_DB1	28	2	语义分割
DeepGlobe	6 k/1 k	2	语义分割
Massachusetts	137/10	2	语义分割
Pascal VOC	1 k/1 k	20	语义分割
Pascal Context	5 k/5 k	59	语义分割
Mapillary	18 k/2 k	65	语义分割
COCO	118 k/5 k	80	语义/实例分割
ADE20k	20 k/2 k	150	语义/实例分割
Cityscapes	3 k/500	19	语义/实例分割
VSPW	198 k/25 k	124	视频语义分割
Youtube-VIS-2019	95 k/14 k	40	视频实例分割

也适用于实例分割.ADE20K是这些数据集中样本量最大的,包含20210张训练图像和2000张验证图像.在视频分割领域,野生视频场景解析数据集(A Large-Scale Dataset for Video Scene Parsing in Wild, VSPW)[33]和YouTube视频实例分割数据集(YouTube Video Instance Segmentation 2019, Youtube-VIS-2019)[20]是最常用的数据集,训练/测试集的数量分别达到了195 k/25 k和95 k/14 k.

(2)常见的指标.图像/视频的语义分割常用的评价指标包括平均交并比(Mean Intersection over Union, mIoU)、平均精确度(Mean Average Precision, mAP)、$F1$分数($F1$ Measure)、像素准确率(Pixel Accuracy, PA)、平均像素准确率(Mean Pixel Accuracy, MPA),以及Kappa系数.其中,mIoU衡量预测分割区域与真实分割区域的重叠程度;mAP评估分割对象位置和类别预测的准确性;F1 Measure是准确率和召回率的加权调和平均,反映模型在正样本检测的性能;PA表示预测类别正确的像素占总像素的比例;MPA计算所有类别的像素精度总和除以类别数;Kappa系数检验预测与实际分类结果的一致性,值在0到1之间.

2.3 Transformer基础架构

Transformer是目前广泛应用于计算机视觉领域的神经网络架构,由编码器和解码器组成[9].这两部分的核心是自注意力机制,赋予其在图像处理任务中的卓越性能.接下来,本文将深入探讨这些模块的功能、结构和它们在图像分割任务中的具体应用.

编码器的主要功能是提取和处理图像特征.它通过自注意力机制捕捉图像中不同区域的关系,帮助模型理解图像的全局上下文信息.编码器由多个相同的层堆叠而成,每层包含多头自注意力机制和多层感知机,每个子层后面都跟着一个残差连接和层归一化.这种结构有助于避免在深层网络中出现梯度消失的问题,使得编码器能够有效地从图像中提取高级特征,为后续的图像分割任务提供必要的信息.

解码器的功能是基于编码器提供的特征信息,生成精确的分割图.它通过关注编码器的输出以及已生成的分割图的部分,逐步构建完整的分割结果.解码器同样由多个层组成,每层包含多头注意力机制和多层感知机制.这种结构使得解码器能够在保持对全局信息关注的同时,逐步细化局部区域的分割结果.

自注意力(Self-Attention, SA)机制使模型能够关注图像中的不同区域,并理解这些区域之间的关系.给定特征$\boldsymbol{X}=[x_1, x_2, \cdots, x_N]\in\boldsymbol{R}^{N\times c}$,其中$N$和$c$分别为特征数量和维度.自注意力机制可表示为:

$$\mathrm{SA}(\boldsymbol{X})=\mathrm{Softmax}\left(\frac{\boldsymbol{X}\boldsymbol{W}_Q(\boldsymbol{X}\boldsymbol{W}_K)^{\mathrm{T}}}{\sqrt{c}}\right)\boldsymbol{X}\boldsymbol{W}_V \quad (1)$$

其中,$\boldsymbol{W}_{\mathrm{Q}}$, $\boldsymbol{W}_{\mathrm{K}}$, $\boldsymbol{W}_{\mathrm{V}}\in\boldsymbol{R}^{c\times c}$为映射矩阵,Softmax为使用的激活函数.

多头注意力机制(Multi-Head Self-Attention, MHSA)由自注意力机制构成,允许模型在处理图像时同时关注多个不同的区域和尺度.这种机制使得模型能够在不同的表示子空间中学习图像的不同特征,从而更全面地理解图像内容.这对于精确地执行图像分割任务至关重要,尤其是在需要区分细微差别的场景中.可表示为

$$\mathrm{MHSA}(\boldsymbol{X})=\mathrm{Concat}\left(\mathrm{SA}(\boldsymbol{X}_1),\cdots,\mathrm{SA}(\boldsymbol{X}_i)\right) \quad (2)$$

多层感知机(Multi-Layer Perceptron, MLP)由两个连续的线性层和非线性激活层组成,其目标是增强注意力层输出的非线性特性.

3 方 法

本节回顾了基于Transformer的视觉分割技术,主要包括语义分割和实例分割,其中,语义分割将视觉数据中的目标划分至不同的类别和区域,实例分割进一步区分同类别的不同个体.按照处理视觉数据的不同,本文以图像/视频为脉络,整理分析了两类视觉数据在语义分割和实例分割的流派方法,如图2所示.对于图像数据,语义分割主要包括纯Transformer和双分支结构两种流派;实例分割在语义分割的基础上进一步区分实例编号,主要包含非端对端和端对端Transformer两种流派;对于视频数据,语义分割主要分为面向精度的和效率的Transformer流派;视频实例分割更加关注实例在时间上的关联问题,现有流派通常从逐帧和逐片段两个方面入手.除此之外,本文展开讨论了图像语义分割在航拍影像的应用,以及视频实例分割在直播视频的应用及面临的问题,最后概述了视觉分割的研究热点问题.

3.1 图像分割

3.1.1 图像语义分割

本节主要介绍基于Transformer图像语义分割

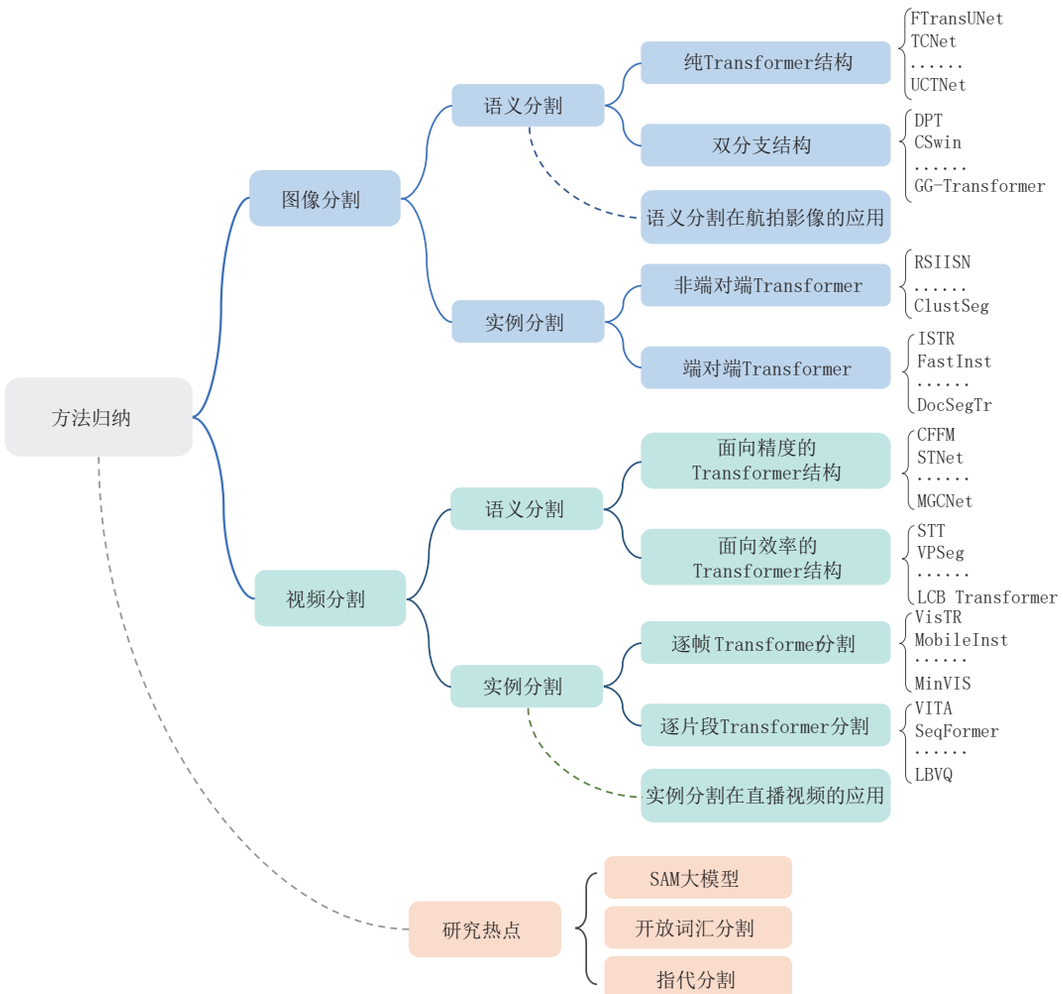


图2　基于Transformer的视觉分割方法归纳图

任务的概念和模型结构，包括纯Transformer结构、双分支结构. 此外，以航拍影像语义分割任务为例，展示了无人机航拍影像非铺装道路分割和遥感语义分割的实际应用性能.

语义分割目标是为图像中的每个像素分配一个语义标签，从而对图像精细化理解. 如图3所示，基于Transformer的语义分割网络通常分为编码器和解码器两部分，其中编码器负责提取特征，解码器则用于生成分割图像. (1)编码器包括图像窗口化操作、编码模块和下采样操作. 首先，执行窗口化操作

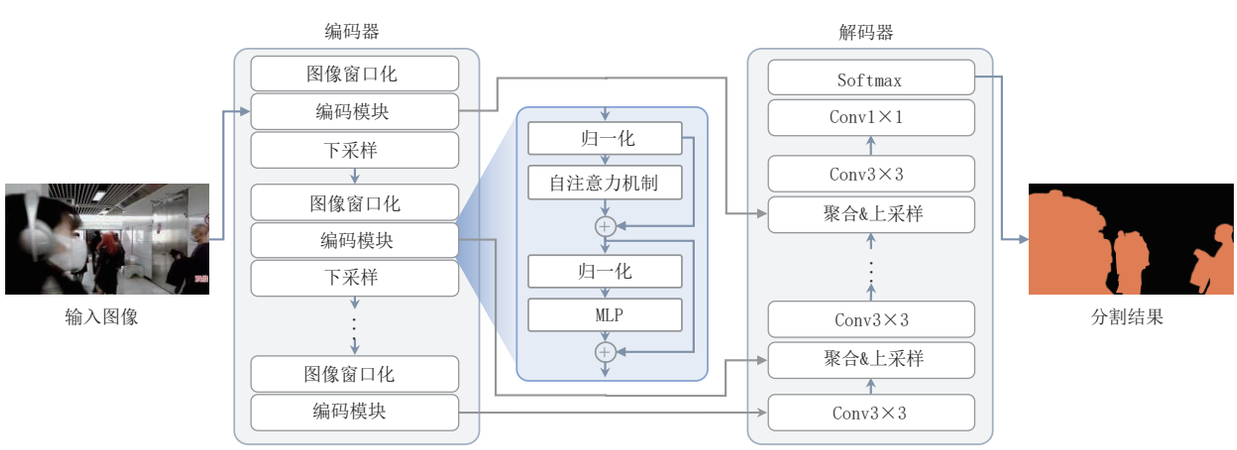


图3　图像语义分割流程图

来将输入图像/特征划分为不同的图像块;然后,利用编码模块提取每个图像块的特征,有效降低了特征提取的计算量.在特征提取过程中,自注意力机制是提取图像特征的核心步骤,而MLP可以对特征进行非线性变换,提升特征的表达能力.(2)解码器利用3×3的卷积操作对特征解码,并利用聚合和上采样操作融合来自不同编码模块的特征,实现图像语义分割.

在最近的研究中,Transformer模型在计算机视觉任务中取得了显著的成功.在这一背景下,基于Transformer重新的语义分割方法(SEgmentation TRansformer, SETR)[14]首次被提出将Transformer模型成功应用于语义分割任务,为该任务带来了新的思路和性能提升.然而,该方法在提高精度的同时也带来了计算复杂度的显著增加.为了解决这一问题,后续工作主要从纯Transformer结构和双分支结构两个方面加以改进.

(1)纯Transformer结构

Transformer的自注意力机制通过Query、Key、Value矩阵间的乘法操作提取图像特征,连续的矩阵乘法导致网络的计算量随着图像尺寸的改变呈现指数级的增大或减小.因此,许多工作尝试将完整的图像划分为局部窗口,并分别提取各个窗口的特征来获取整张图像的语义信息.窗口划分虽然会使自注意力机制执行的数量线性增加,但是每次执行所需的计算量呈现指数级的下降,从整体来看计算量是有所下降的.起初,只是使用等宽高的窗口将图像拆分成多个互不重叠且大小相等的图像块,如密集预测Transformer(Vision Transformers for Dense Prediction, DPT)[34].该方法极大的降低了模型的计算量,但是这种粗糙的窗口化方法存在一定的缺陷,即难以从拆分后的图像中提取全局特征.随后,基于变换窗口机制的视觉Transformer(Hierarchical Vision Transformer using Shifted Windows, Swin Transformer)[35]将两次自注意力操作视为一组特征提取过程.在执行第一次自注意力时使用的传统的窗口划分方法,第二次执行时会将窗口沿水平或竖直方向偏移一段距离并重新拆分窗口.该方法避免了每次只能提取特定局部窗口的语义信息,实现了相邻窗口的信息交流,从而提取全局特征.类似的,全局扫描Transformer(Glance-and-Gaze vision Transformer, GG-Transformer)[36]将空洞卷积的思想融入到传统的窗口划分操作中,设计了间隔采样的窗口划分操作,该操作等同于将图像拆分为多个分辨率更小的图像,在不增加计算量的前提下提取全局信息.有了上述窗口化提取全局信息的基础,为了进一步减少窗口中的冗余信息,基于十字窗口的视觉Transformer(General Vision Transformer Backbone with Cross-Shaped Windows, CSWin Transformer)[37]设计了由水平和竖直条形窗口组合而成的十字窗口,这种方法虽然增加了计算次数,但是明显降低了每次计算的时间复杂度,在降低Transformer模型总体计算量方面取得了显著成效.

还有一些工作尝试在计算自注意力机制的过程中对Query、Key、Value矩阵进行下采样,通过剔除特征中的冗余信息实现在不影响分割精度的前提下降低计算量.在下采样过程中,许多方法基于自适应池化或注意力方法缩减特征图在空间或通道方向的维度来降低下采样带来的信息损失.例如,金字塔视觉Transformer(Pyramid Vision Transformer, PVT)[38]在每次计算自注意力机制的过程中,采用自适应池化方法将Key和Value矩阵下采样为空间尺寸更小的特征图.该方法使得矩阵相乘的计算量与下采样的倍数成正比,同时下采样后的特征图依旧拥有与原特征图相同的感受野范围,因此对分割精度的影响相对较小.还有工作通过减少通道维度来降低计算量,十字窗口全局视觉Transformer(Crisscross-Global Vision Transformer, CGVT)[39]借鉴了压缩提取网络(Squeeze-and-Excitation Networks, SENet)[40]中的通道缩减思想去除特征图中的冗余通道.该方法与PVT较为类似,不同的是CGVT借助SENet保留了通道中更多的语义信息,进一步缓解了降低图像尺寸导致的信息损失.除此之外,稀疏理论也被应用到了Transformer当中,交错稀疏自注意力机制(Interlaced Sparse Self-Attention, ISSA)[41]是其中比较有代表性的工作之一.该工作为了降低自注意力机制的计算量,将原本的密集特征矩阵拆解为两个稀疏矩阵的乘积.稀疏矩阵可以只存储和处理非零元素,从而大大减少存储空间和计算资源的需求.相比于下采样的方法,利用稀疏矩阵代替原特征图可以更轻松地处理大量稀疏数据的场景,对于图像语义分割有较大帮助.不论是采样何种理论框架,在自注意力机制中都可以精炼特征图中的语义信息,从而在保证特征提取能力的同时降低计算成本,为Transformer模型在图像语义分割任务中的应用提供了可能性.

(2)双分支结构

Transformer擅长提取全局信息,一些工作尝试

结合CNN与Transformer来提取不同类型的语义信息，以提升图像分割性能．例如，通过堆叠残差卷积块和Transformer的融合架构（Fusion Architecture via Stacked Residual Convolution Blocks and Transformer，SRCBTFusion-Net）[42]在分割过程中使用残差卷积堆叠模块来提取特征，再用自注意力机制精炼浅层特征，提高了对局部细节和全局语义的理解，有效减少了浅层和深层间的语义混淆对分割结果的影响．Transformer-UNet融合网络（Fusion TransUNet，FTransUNet）[43]也采用组合网络利用精细的浅层特征和上下文深层特征促进模型对图像各区域的感知能力，提高了图像语义分割性能．类似的，多尺度Transformer&CNN融合网络（Multiscale Fusion of Transformer and CNN，TCNet）[44]采用了并行分支结构，分别使用自注意力和ResNet来提取全局和局部信息，构建了一个窗口化的自注意力门控机制，以弥合不同分割区域之间的语义差距．不确定性引导的CNN-Transformer混合网络（Uncertainty-Guided CNN-Transformer Hybrid Networks，UCTNet）[45]引导CNN和Transformer分别感知不同分割区域，以充分发挥各自优势，最大限度地减少特征冗余．上述研究表明，CNN在捕获局部特征和空间信息方面表现优异，而Transformer在处理长程依赖和全局信息时更为有效，结合两者能充分利用CNN在局部处理中的强大能力和Transformer在全局建模中的优势，从而提升模型对图像分割的整体性能．

（3）语义分割在航拍影像的应用

在航拍影像处理领域，将地物目标分割成不同的区域，如建筑物、植被、水体等，能够为城市规划等应用提供有力的数据支持，是一个具有挑战性的任务．

航拍影像中的目标尺寸多样，对分割性能有较大影响，许多工作尝试提升模型在不同尺度下的分割性能．例如，CNN多尺度局部上下文Transformer（CNN and Multi-Scale Local-Context Transformer，CMLFormer）[46]将多尺度水平和竖直条状卷积融入Transformer分割框架中，有效扩展了模型的多尺度感知能力，弥补了分割过程的全局信息损失．CNN多尺度Transformer融合网络（CNN and Multiscale Transformer Fusion Network，CMTFNet）[47]在分割过程中将多尺度信息和通道信息整合到Transformer分割框架中，并构建了多尺度注意力融合模块以自适应地融合深层和浅层特征．强化多尺度表征Transformer（Enhancing Multiscale Representations with Transformer，EMRT）[48]尝试在解码阶段获取多尺度上下文信息，增强分割过程中捕捉不同尺度目标的能力．

此外，Transformer擅长建模丰富的背景信息，虽然可以有效分辨目标，但是缺失的局部信息能力导致分割边缘不完整．为此，许多工作尝试结合局部和全局信息提升模型分割性能．聚类Transformer（Clusterformer）[49]利用从深层解码特征中构建的簇来优化浅层编码特征，提升了遥感图像分割精度．高效的生成对抗Transformer遥感图像分割（Efficient Remote Sensing Segmentation with Generative Adversarial Transformer，Efficient GATrans）[50]在分割过程中引入生成对抗网络来捕捉全局信息，并融合结构相似性损失和对抗性损失的目标函数，实现了高精度语义分割的同时保持较高的效率．结合Swin、卷积，以及全局Transformer的遥感图像语义分割（Swin-Conv-Dspp and Global Local Transformer for Remote Sensing Image Semantic Segmentation，SCG-TransNet）[51]结合了Swin和Deeplabv3＋实现图像分割，增强了局部信息的提取能力，使得在分割过程中可以充分利用遮挡物体的有限像素生成可区分的表示信息，有效缓解因目标遮挡造成的误检甚至漏检情况．混合掩码Transformer（Mixed Mask Transformer，MMT）[52]考虑到分割过程的背景信息通常具有离散分布和局部相似性，因此对于背景区域使用网格采样以最大限度地降低冗余的背景信息干扰．

在我们之前的工作中[53]，提出了一个全局视觉Transformer的无人机航拍影像非铺装道路分割方法（Global Vision Transformer，GVT），首先构建了空洞交叉窗口注意力模块（Dilated Cross Window Attention，DCWin-Attention），用于克服道路的变形和模糊边缘问题；然后引入了一种移位交叉窗口机制，并结合DCWin-Attention来构建DCWin-Attention主干网络，提取具有全局依赖性的多层深度特征．从采集的无人机航拍影像中自建了北京工业大学非铺装道路数据集（Beijing University of Technology Unpaved Road Dataset，BJUT-URD）和遥感影像道路公共数据集DeepGlobe[25]上进行验证，实验细节如下：

（1）数据集．BJUT-URD数据集包括161幅分辨率为6000×4000的无人机航拍影像，并按照8∶2的比例分别用于训练和测试．并对训练集和测试集

进行数据增强，包括旋转、裁剪、平移、翻转和颜色抖动，并将结果缩放到1200×800分辨率．数据增强后，训练集和测试集分别扩展到8167幅和1567幅图像．DeepGlobe道路数据集由6226幅训练图像、1243幅验证图像和1101幅测试图像组成．每幅图像的分辨率为1024×1024．值得注意的是，由于DeepGlobe测试集的标签不对外公开，因此使用DeepGlobe训练集进行训练，验证集进行测试．

(2)实验参数．实验在一台配备16 GB内存、2.1GHz CPU和NVIDIA 2080Ti GPU的电脑上进行．算法框架由Python 3.8和PyTorch 1.7.1以及CUDA 11.1和cuDNN 8.0.5实现．此外，使用AdamW优化器进行训练，其超参数如表2所示．

表2 道路分割方法超参数

学习率	迭代次数	权重衰减	随即裁剪	冻结比例
6×10^{-5}	160 k	0.01	512×512	0.3

(3)评估指标．本实验采用交并比(Intersection of Union，IoU)、$F1$分数、精确率$Precision$和召回率$Recall$来评估每种分割方法在BJUT-URD数据集上的性能．

$$IoU=\frac{|\boldsymbol{Y}\cap\boldsymbol{G}|}{|\boldsymbol{Y}\cup\boldsymbol{G}|} \tag{3}$$

$$Recall=\frac{TP}{TP+FN},Precision=\frac{TP}{TP+FP} \tag{4}$$

$$F1=2\times\frac{Precision\times Recall}{Precision+Recall} \tag{5}$$

其中，$\boldsymbol{Y}$和$\boldsymbol{G}$分别为预测结果和地面实况．$Precision$和$Recall$表示精确率和召回率．TP表示真阳性，TN表示真阴性，FP表示假阳性，FN表示假阴性．DeepGlobe数据集提供了官方评估服务器，通过提交测试结果并返回IoU．

我们比较并分析了不同算法的IoU结果[53]，如图4所示，每16 000次迭代测量子测试集的IoU，并将IoU值可视化，达到了最高，同时训练过程也比其他方法更平滑，这反映了所提工作在非铺装路面分割方面的有效性和优越性．这是因为空洞交叉窗口通过水平和垂直的长距离依赖关系有效地模拟全局依赖关系，此外，将扩张交叉窗与像素区域模块和用于分割道路的移位交叉窗机制相结合，有效弥补了交叉窗造成的信息损失，提升了模型分割准确率．

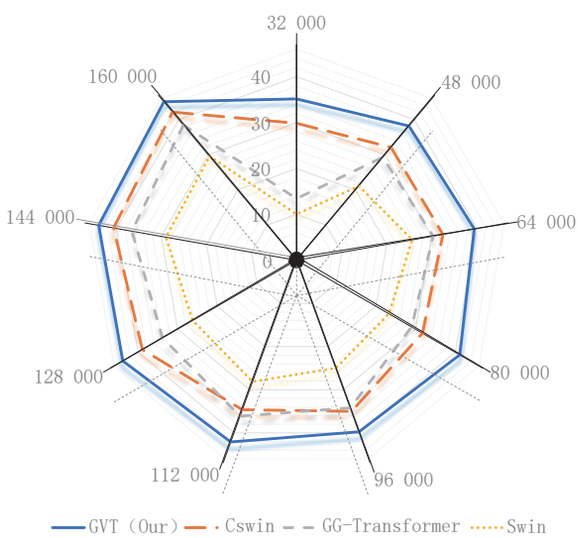


图4 不同方法的IoU结果比较

为了比较道路分割效果，在图5中展示了与数据其他方法的主观结果[53]．其中，第1～2行为自建集主观结果，3～4行是DeepGlobe数据集上的主观结果．可以看出，本方法对道路进行了更完整的分割，还可以有效地处理模糊边缘和遮挡．此外，当道路发生变形时，其他方法可能会将道路与背景混淆，

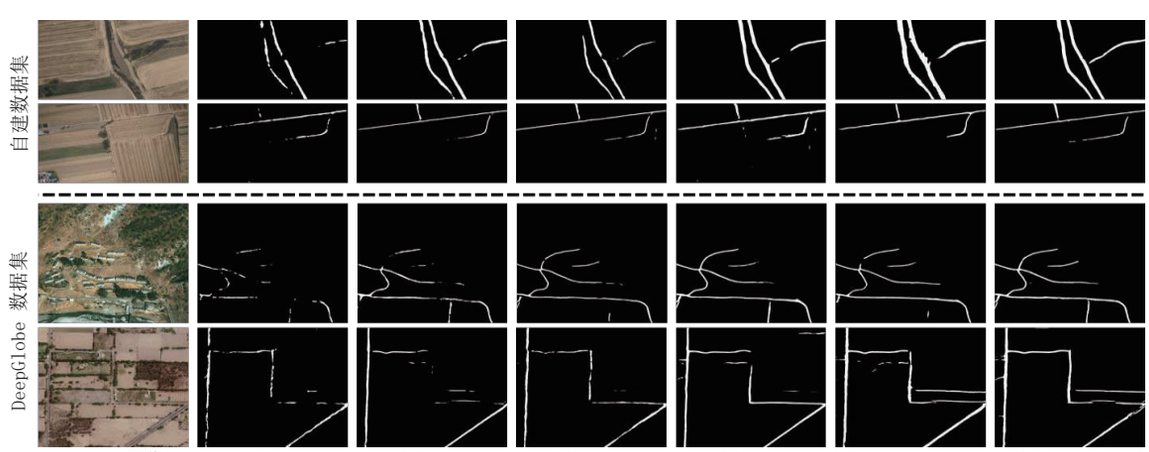


图5 全局视觉Transformer的UAV道路分割主观结果展示[53]

而本方法呈现出更完整的分割结果．因此，利用DCWin-Attention对水平和垂直远程依赖关系进行建模，简化了道路特征的复杂表示．这是因为空洞交叉窗口机制减轻了建模过程中全局依赖的信息丢失，引入了强大的特征学习能力，从而提高了道路分割的性能．

此外，针对高分辨率遥感影像，我们前期提出了一个空间特异性Transformer的语义分割方法[54, 55]，首先将involution和自注意力分支整合构建空间特异性Transformer编码器，提取地物目标多级特征；然后使用大窗口注意力解码器捕获多尺度上下文信息，同时引入地物特征补充分支减少语义信息损失；最后将多尺度特征融合后输入分类器实现高分遥感影像语义分割．所提方法在Potsdam和Vaihingen数据集上进行验证，实验细节如下：

(1)数据集．Potsdam数据集包含来自航空摄像机收集的38个样本，其分辨率为6000×6000．本工作将数据裁剪成600×600分辨率进行训练，并通过旋转、调整大小、水平轴翻转、垂直轴翻转进行数据增强，共15 200张图像，其中80%作为训练集，20%作为测试集．Vaihingen数据集包含来自航空摄像机收集的33个样本，平均分辨率为2494×2064分辨率．对于该数据集，将其裁剪成512×512分辨率进行训练，通过旋转、调整大小、水平轴翻转、垂直轴翻转进行数据增强，共2968张图像，其中80%作为测试集，20%作为训练集．

(2)实验参数．实验使用Ubuntu16.04操作系统，一块NVIDIA 2080Ti GPU，采用PyTorch深度学习框架．训练设置主要遵循数据高效的Transformer，在分辨率为224×224的模型上进行预训练，以减少GPU消耗．然后使用AdamW优化器来指导优化，其超参数与表2相同．

(3)评估指标．除前述F1分数，还使用了平均交并比(Mean Intersection of Union，mIoU)、像素准确率(Pixel Accuracy，PA)、平均像素准确率(Mean Pixel Accuracy，MPA)和*Kappa*系数作为评估指标．

$$mIoU=\frac{1}{K}\sum_{i=1}^{K}IoU_i \tag{6}$$

其中，K表示类别数．

$$PA=\frac{TP+TN}{TP+TN+FP+FN} \tag{7}$$

其中，TP表示真阳性，TN表示真阴性，FP表示假阳性，FN表示假阴性．

*MPA*是类别平均像素精度：

$$MPA=\frac{\mathrm{sum}(P_i)}{N_c} \tag{8}$$

其中，P_i表示每个类别的像素准确率，N_c表示类别数．

*Kappa*系数使用混淆矩阵来衡量分类结果的准确性：

$$Kappa=\frac{p_0-p_e}{1-p_e} \tag{9}$$

其中，p_0和p_e分别指总体分类准确率和偶然一致性误差．此外，参数(Params)和每秒浮点运算次数(Flops)用于评估模型的复杂性．

图6比较了不同编码器和解码器在遥感图像上的分割性能[54, 55]，在引入大窗口注意力解码器后，所提方法达到了最高的87.61% *mIoU*/93.26% *F*1，增加了0.19% *mIoU*/0.06% *F*1．此外，几乎所有类别(不透水表面、建筑、低植被和树木)的分割精度都得到了提高．总之，利用空间特异性Transformer(Spatial-specific Transformer，Spatial-specificT)可以帮助捕获相应地物目标位置的多尺度上下文信息，从而获得更强的多尺度特征学习和表示能力．

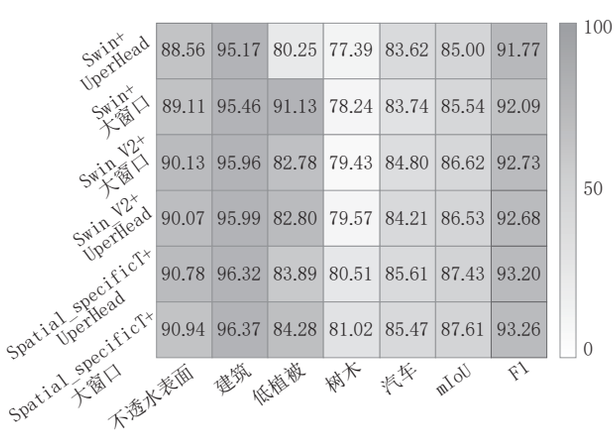


图6　不同编码器和解码器在遥感图像语义分割性能比较

3.1.2　图像实例分割

本节主要解释了基于Transformer图像实例分割的任务目标，归纳了实例分割模型结构的主要组成部分，介绍了基于非端对端的和端对端两个典型的Transformer实例分割网络．

图像实例分割是计算机视觉领域的一项前沿技术，旨在准确区分图像中的每个单独对象，并为这些对象的每个像素分配正确的类别标签．如图7所示，基于Transformer的图像实例分割网络主要包括以下三个部分：(1)编码器沿用了图像语义分割的架

构，包括一个图像窗口化操作、编码模块和下采样操作，其中编码模块用来提取特征；(2)像素解码器与图像语义分割的解码器具有相同的结构，结合3×3的卷积、聚合、上采样操作进行特征融合，解码后的特征会被送入实例解码器进一步精炼，以区分同一类别的不同实例；(3)实例解码器首先初始化固定数量的Query，每个Query表示一个特定实例的语义信息．根据解码器功能与结构的差异，现有方法主要分为非端对端Transformer和端对端Transformer.

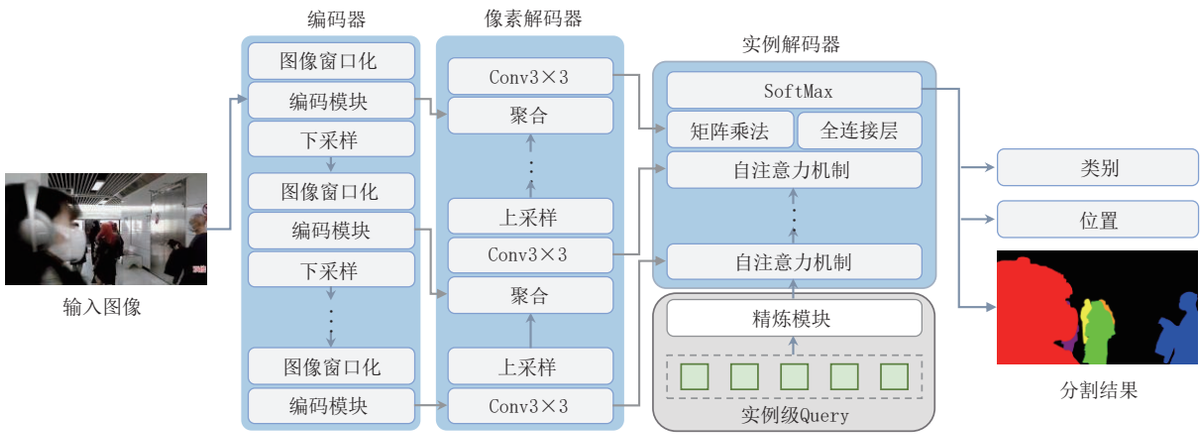


图7 图像实例分割流程图．

(1)非端对端Transformer

最初，基于Transformer的图像实例分割是非端对端的，这些方法通常是基于先检测后分割的框架．例如，遥感图像实例分割网络(Remote Sensing Image Instance Segmentation Network，RSIISN)[56]先利用通用感兴趣区域(Region of Interest，RoI)提取器获取每个目标的局部特征图，然后利用Transformer和CNN对局部特征图进行语义分割，最后合并各个局部特征图的分割结果，实现图像实例分割．虽然这些基于区域的方法可以有效分割并区分不同的实例，但是大量重复检测区域会不可避免的限制模型性能和效率．

一些工作尝试将实例分割拆分为先学习对每个像素进行分类，然后聚类为不同对象簇的分割过程．例如，基于聚类的通用分割方法(Clustering for Universal Segmentation，ClustSeg)[57]利用Transformer初步区分像素类别，迭代使用超像素分割进行聚类以区分并分割不同实例．然而，非端对端的每个检测和分割部分需要单独设计和优化，使得整个流程比较复杂．

(2)端对端Transformer

基于端对端的实例分割工作受到越来越多的关注并逐渐成为主流，这些方法通过初始化固定数量的Query来区分不同的实例，实现端对端的实例分割．例如，掩码注意力Transformer(Masked-Attention Mask Transformer，Mask2Former)[58]和端对端实例分割Transformer(End-to-End Instance Segmentation with Transformer，ISTR)[59]在Transformer编码器中手动初始化了100个Query，并使用自注意力机制将特征图中每个实例的语义信息映射到单独的Query上，实现了端对端的实例分割．还有一些工作先利用Transformer编码器提取场景特征，再在解码器中初始化Query并从场景特征中学习实例信息，如Query Refinement Transformer[60]、基于实例感知的对象分割Transformer(Segmenting Objects with Instance-Aware Transformers，SOIT)[61]、细胞结构检测Transformer(Cell Detection Transformer，Cell-DETR)[62]．端对端的实例分割方法简化了处理流程，通过初始化Query能够直接生成实例分割结果．为了提高图像实例分割的性能，研究者们正致力于增强Query的表征能力．

优化Query的初始化参数是提高Transformer图像实例分割性能的关键步骤．这一过程本质上涉及使用不同策略来获取能够反映实例结构与分布特点的数据，并利用这些数据生成初始化参数．例如，基于Query的实时实例分割方法(Query-based Model for Real-Time Instance Segmentation，FastInst)[63]提出了一种基于实例激活引导的Query初始化方法，首先通过分类头对特征初步分类，然后选择置信度较高的特征作为Queries．这样做的优势在于，它能够更准确地定位到图像中的关键区域，从而为实例分

割提供更精确的起点．文件图像分割Transformer（Document Image Segmentation Transformer，DocSegTr）[64]引入了一个孪生注意力模块，通过横向和纵向注意力来引导网络关注实例的位置和形状，更好地理解实例的空间分布，从而提高分割的精度和可靠性．同时，这些方法使得网络能够从一个更准确的训练起点开始，降低了错误分割的风险．整体而言，优化Query的初始化参数需要综合考虑实例的结构和分布．因此，未来的研究可能会专注于深入分析实例的结构和分布特性，探索如何更有效地利用这些信息来优化Query的初始化．

在实例分割领域，许多精炼Query方法用于提升分割质量．例如，掩码精炼器（Mask Transfiner）[19]利用轻量级全卷积网络生成粗糙的分割结果，通过结合Transformer进行精炼，提升分割结果的连续性和准确性．这种方法有效地融合了FCN的高效性与Transformer的强大处理能力，实现了在低计算成本下提升分割质量的目标．类似的，深度多边形实例分割Transformer（Deep Polygon Transformer for instance segmentation，PolyTransform）[65]专注于实例轮廓的生成和精炼，利用Transformer提高轮廓的精确度，在处理复杂背景或重叠实例时尤为有效．精炼Query的过程显著提高了模型对实例特征的捕捉能力，增强了模型在复杂场景中的鲁棒性，在实例分割领域展现出巨大潜力．未来的研究可能会致力于提高这些方法的效率和准确性，并探索新的技术以应对更多样化的场景和挑战．

3.2　视频分割

3.2.1　视频语义分割

本节首先解释了Transformer在视频语义分割领域的任务动机，然后分析了视频语义分割模型的模型结构，最后总结了该模型中面向精度的以及面向效率的Transformer结构研究进展．

视频语义分割是一种先进的计算机视觉技术，与传统的静态图像分割有显著区别，它处理的是动态连续的图像序列．在视频语义分割中，不仅需要区分每帧视频中的像素类别，还必须考虑帧间的时间连续性和关联性．视频语义分割流程如图8所示，首先使用编码器从视频帧序列中提取特征，通过编码器之间参数共享，减少模型复杂性，同时保障不同帧特征的语义一致性；然后利用特征融合模块获取帧序列时间信息，以捕捉和理解视频中的动态变化；最后将融合后的特征被送入解码器，由解码器生成每个视频帧的语义分割图像．在视频语义分割领域，时间信息的有效融合是决定分割效果的关键因素．因此，如何高效地利用时间信息进行视频语义分割，以及如何处理额外的视频帧特征带来的计算负荷，成为了当前研究的重点问题．为应对这些挑战，研究人员正在探索动态与静态语义信息的有效提取和融合方法，以最大化时间信息的利用．同时，为减轻处理时间信息增加的计算负担，自注意力机制的轻量化研究也越来越受到关注．

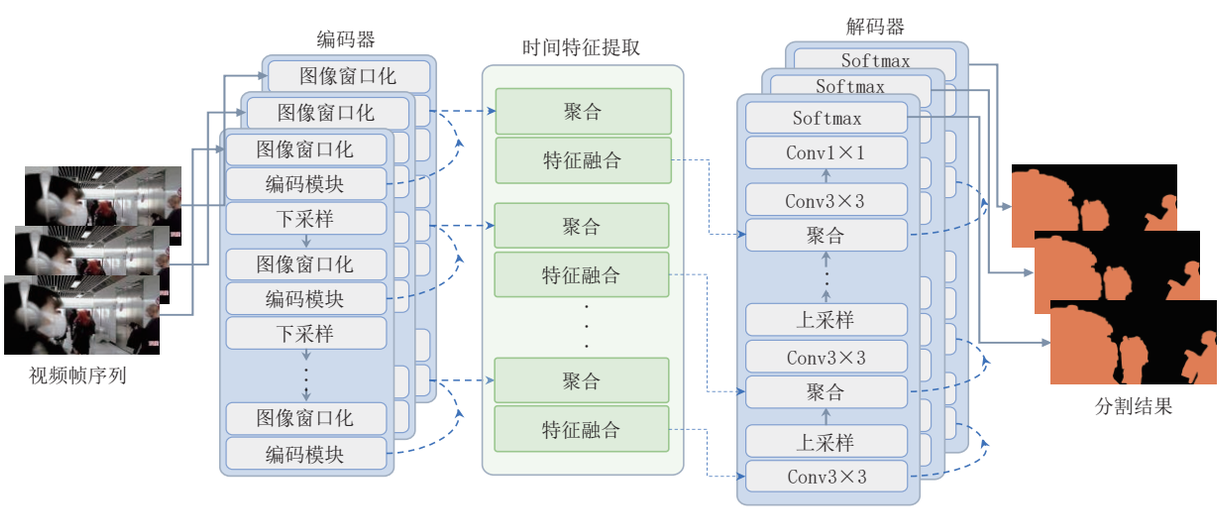


图8　视频语义分割流程图

（1）面向精度的Transformer结构

在视频分割技术的发展中，核心目标是充分利用时间信息来提升性能．近年来的研究工作主要集中在动态和静态信息的提取与融合上，以最大化时间信息的利用．一些工作通过设计特殊的分支或模块来同时编码时间信息和增强单帧图像的语义信

息，如运动状态校准策略(Motion-state Alignment)[66]和记忆增强精炼模块(Memory-Augmented Refinement，MAR)[67]. 这种动静态信息融合方法不仅提高了分割的准确性，还增强了细节捕捉能力，特别适合处理视频中的动态变化. 类似的，由粗到精的特征挖掘策略(Coarse-to-Fine Feature Mining，CFFM)[68]通过粗到细的特征挖掘方法学习并融合静态和动态上下文信息，展现出对复杂场景捕捉的强大能力. 另一方面，基于Transformer的时间信息处理方法，如局部记忆注意力机制(Local Memory Attention Networks，LMANet)[69]和时空语义分割网络[70]，利用Transformer的强大能法专注于提高视频序列的连续性和整体一致性，通过保留上一帧的特征信息并与当前帧的动态信息结合，实现了高效的时间信息融合. 此外，自适应关键帧调度的双分支混合网络(Dual-Branch Hybrid Network of CNN and Transformer with Adaptive Keyframe Scheduling，DHN-AKS)[71]设计了一个基于动态存储矩阵的分割方法，用于存储历史视频帧的动态和静态语义信息，从而捕捉当前帧的时间相关性. 最后，基于Token的帧间信息联合方法，如多粒度上下文网络(Multi-Granularity Context Network，MGCNet)[72]，通过生成每帧特征的Token来实现不同帧信息的联合. 这种方法在处理大量帧的视频序列时表现良好，但可能在实时性方面存在挑战. 总体而言，这些方法通过不同的技术策略实现了静态和动态信息的有效融合，从而更全面地捕捉视频内容的丰富细节，并有效利用时间信息，显著提升了视频分割的准确性.

(2) 面向效率的Transformer结构

在视频语义分割的研究中，轻量化注意力机制的应用是为了降低计算量的一个重要手段. 这些方法可以大致分为两类：一类是通过构建稀疏矩阵来优化注意力机制，另一类是通过多尺度处理和选择性信息筛选来提高效率. 例如，稀疏时间Transformer(Sparse Temporal Transformer，STT)[73]在自注意力机制中构建了Key和Query的稀疏矩阵，有效地过滤了冗余信息，在保持注意力机制核心功能的同时显著减少计算复杂度. 在多尺度处理和选择性信息筛选的方法中，多尺度亲和度聚合策略[74]和选择性令牌屏蔽策略[75]是其中两个比较有代表性的工作. 多尺度亲和度聚合策略通过在不同尺度上聚合特征，允许模型更好地理解场景的整体结构. 选择性令牌屏蔽策略通过去除冗余的语义信息，减少了模型需要处理的数据量，特别适用于那些具有重复或相似场景的视频. 它们都通过筛选和优化注意力机制中的信息来降低计算负担，从而在不牺牲视频分割性能的前提下有效地减少了视频语义分割的计算需求. 此外，还有一些其他降低分割代价的方法，如消失点引导的视频语义分割方法(Vanishing-Point-Guided Video Semantic Segmentation，VPSeg)[76]设计了一个高效的全局上下文-局部细节分割框架，以不同的分辨率自适应地分离全局上下文特征和局部细节特征，降低视频分割的计算成本. 同样的，轻量级卷积扩散Transformer(Lightweight Convolutional Neural Networks with Context Broadcast Transformer，LCB Transformer)[77]在轻量级CNN架构中引入Transformer实现两者之间的优势互补，以平衡视频分割的精度与效率. 总体而言，轻量化注意力机制在视频语义分割领域展现了巨大的潜力. 通过这些创新的方法，研究人员能够在保持高性能的同时，显著降低计算资源的消耗，这对于实时视频处理和资源受限的应用场景尤为重要.

3.2.2 视频实例分割

本节基于Transformer的视频实例分割主要介绍逐帧和逐片段Transformer实例分割，然后以网络直播视频为例，展示视频实例分割的应用效果. 视频实例分割是计算机视觉领域的一个前沿研究方向，用于从视频帧序列中区分出每个独立的实例对象. 在基于Transformer的视频实例分割方法中，通常需要初始化一组固定数量的Query，旨在学习和追踪视频中相互独立的实例. 如图9所示，在视频实例分割过程中，首先利用共享参数的编码器提取输入视频帧的特征. 提取的特征随后被传递到解码器，用于关联不同视频帧中的特征，确保同一实例在各视频帧中的语义保持一致. 视频实例分割的解码器主要分为实例级解码器和视频级解码器两种类型. 在实例级解码器中，不同视频帧的特征通过独立的特征融合模块提取时间信息，同时为每个实例分配一个特定的Query以表征其特征. 最终，自注意力机制用于指导Query学习特征中的实例信息，实现视频实例分割. 在视频级解码器中，初始化一个公共Query，与实例级Query一同送入自注意力机制中，以精炼不同视频帧特征中实例的语义一致性. 这两种方法各有其优势和局限性，目前的研究主要集中在基于实例级和视频级的视频实例分割方法上.

(1) 逐帧Transformer分割

在实例级的视频实例分割中，每个视频帧被分

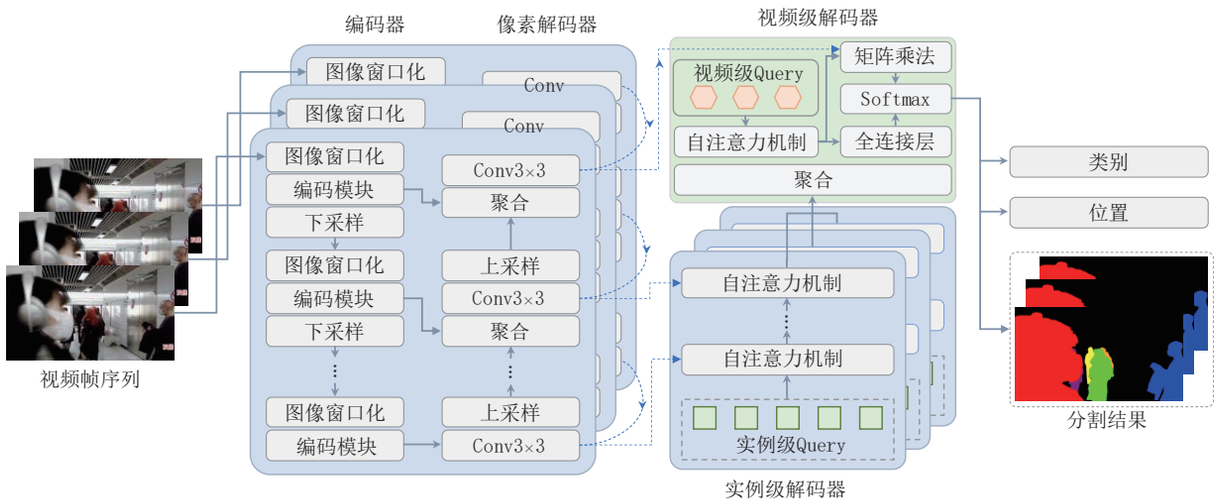


图9　视频实例分割流程图

配一组Query，其中每个Query专注于当前帧内的实例信息．如视频实例分割Transformer（Video Instance Segmentation Transformer，VisTR）[78]和用于视频实例分割的Mask2Former方法（Mask2Former for Video Instance Segmentation，Mask2Former-VIS）[79]等方法采用了Transformer的编码解码结构．在这些方法中，解码器将实例信息编码进Query，并根据Query之间的相似性匹配不同帧间的相同实例．为了增强时间信息融合和动态变化的捕捉能力，许多实例级解码器设计了特征融合模块相邻帧特征的时间信息的策略．例如，时间效率视觉Transformer（Temporally Efficient Vision Transformer，TeViT）[80]通过时间方向的位移操作实现时间信息的交互，并对位移后的特征执行自注意力，以提取实例信息．帧间通信Transformer（Inter-Frame Communication Transformers，IFC）[81]构建了一个针对上一帧特征的记忆库，并利用生成的特征库与当前帧特征融合时间信息，提升了模型捕捉动态变化的能力．移动视频实例分割（Video Instance Segmentation on the Mobile，MobileInst）[82]将双流解码器用于视频实例分割，分别提取全局和局部实例信息，以更少的参数量判断实例的类别和形状．除此之外，还有一些实例级方法利用现有的特征提取策略来提升视频实例分割效果．如可变形视频实例分割（Video Instance Segmentation on the Mobile，DeVIS）[83]采用了可变形注意力机制以强化Query中的实例信息，而精简的视频实例分割框架（Minimal Video Instance Segmentation，MinVIS）[84]则将视频实例分割简化为连续的图像实例分割任务，并使用图像数据集进行训练，仅依据Query中的实例特征进行匹配，提高了实例分割结果的完整性．

（2）逐片段Transformer分割

最近，视频实例分割领域趋向于直接初始化一组视频级的Query，使每个Query能够学习并追踪实例在整个视频序列中的信息，从而增强实例在不同帧间的语义一致性．例如，基于目标token联合的视频实例分割（Video Instance Segmentation via Object Token Association，VITA）[85]编码/解码器．在这种方法中，先通过实例级解码器融合不同帧的实例特征，然后利用视频级定义了一组视频级Query，并构建了专门的视频级解码器将实例在整个时间维度上的信息提取到Query中，避免了实例级Query的匹配过程，简化了关联过程，提升了语义一致性．序列化Transformer（Sequential Transformer，SeqFormer）[86]提出了矩形框Query来预测实例在每帧中的位置和类别，并结合实例级Query和矩形框Query的结果来实现视频实例分割，以提高在每帧中对实例的定位和分类的准确性．此外，学习更好的视频查询索引策略（Learning Better Video Query，LBVQ）[87]通过逐帧和逐片段策略的交互进行实例分割，在时空推理过程中充分利用实例信息．总体而言，这些方法通过不同的技术策略实现了实例信息的精确处理和时间信息的有效融合，从而更全面地捕捉视频内容的动态变化．同时，进一步利用时间维度上的实例信息，显著提升了视频实例分割的准确性．

（3）实例分割在直播视频的应用

网络直播视频分割可以将视频流中的不同目标分割成独立的区域，对于主播和观众的交互、网络监

管等具有重要意义. 如今高分辨率的视频对分割效率具有较大影响，在线视频实例分割框架(Online Video Instance Segmentation Framework，Instance Former)[88]引入了一个新颖的先验传播模块，在分割过程中引入参考点、类得分和实例Query实现连续帧之间的高效信息交互. 离线到在线的知识蒸馏方法(Offline-to-Online Knowledge Distillation，OOKD)[89]将视频分割引入蒸馏框架，实现了将离线模型中更丰富的实例表征到在线模型的过渡. 除此之外，许多工作尝试在提高分割效率的同时增强模型分割的准确率. 例如，学会学习策略(Learning to Learn Better，LLB)[90]设计了鉴别性标签生成模块来分割不同的实例，将背景帧编码为具有鉴别性的目标特征，实现目标信息从历史帧到当前帧的信息融合. 同时，设计了自适应融合模块以将目标信息量跟随到融合后的目标特征中，从而使融合后的目标特征具有更强的鲁棒性. 时间一致的在线视频实例分割(Temporally Consistent Online Video Instance Segmentation，TCOVIS)[91]设计了一种基于全局实例分配策略的分割方法，在考虑整个视频的情况下进行全局最优匹配，并以全局最优目标对模型进行监督，充分利用了时间信息提升视频分割的时间一致性.

在我们之前的工作中[92]，提出了一个直播视频实例分割框架Gp3Former，该框架利用Transformer编码器来增强小实例分割中不同尺度视频特征的表示能力；然后，设计了一个三级联Transformer解码器，在不牺牲场景信息的情况下提取全局、平衡和局部实例特征，以适应直播中不断变化的场景；最后，为了应对直播中的密集实例，在实例关联和分割过程中施加高斯先验，学习一系列跨帧实例的高斯分布. 在自建的北京工业大学直播视频数据集(Beijing Institute of Technology Live Video Dataset，BJUT-LSD)和公共数据集YouTube-VIS 2019[20]上测试了本方法的性能，具体实验细节如下：

(1)数据集. BJUT-LSD数据集包含从斗鱼TV、Bilibili和YouTube收集的71个视频. 每个视频包含20-30个帧，每个视频帧的分辨率为1280×720. 此外，对数据进行了数据增强，共获得526个训练视频和127个测试视频. YouTube-VIS 2019数据集[20]由3471个训练视频、507个验证视频和541个测试视频组成. 每幅图像的分辨率为1280×720.

(2)实验参数. 本框架是使用Python 3.8和PyTorch 1.9.1以及CUDA 11.4和cuDNN 8.4.1实现的，在Ubuntu 16.04操作系统上由配备24 GB内存的NVIDIA 3090 GPU，并使用AdamW优化器对模型进行了优化. 在实验中，保持了与基线一致的超参数设置，以避免不同超参数配置可能产生的潜在偏差，如表3所示.

表3 视频实例分割方法超参数

学习率	迭代次数	权重衰减	随即裁剪	冻结比例
6×10^{-5}	140 k	0.01	640×640	0.3

(3)评估指标. 除前述*mIoU*、*F*1分数、*Kappa*，还使用了平均精确度(Average Precision，AP)、平均召回率(Average Recall，AR)来评估每种方法在数据集上的分割性能.

$$AP=\frac{1}{K}\sum_{i=1}^{K}Precision_i,\ AR=\frac{1}{K}\sum_{i=1}^{K}Recall_i \tag{10}$$

其中，K表示类别数. *Precision*表示当前类精确率，*Recall*表示当前类召回率.

本工作比较了不同方法在多个主干网络上的分割性能. 详情如图10所示[92]，随着主干网结构层数增加，实例分割性能也逐渐提高. 这表明，主干结构的层数越多提取特征信息更丰富，进而提高了分割准确性. 基线值为83.0% *AP*和82.4% *AR*，本方法*AP*和*AR*分别达到了83.4%和83.0%的最大值，高出了0.4%和0.6%. 这归功于本方法能够引入丰富的场景信息，使模型在直播中获得更高的分割性能.

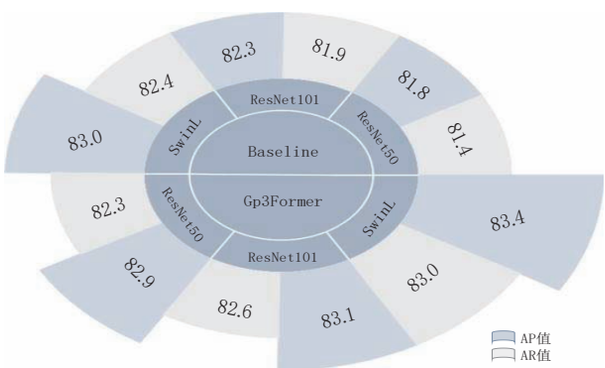
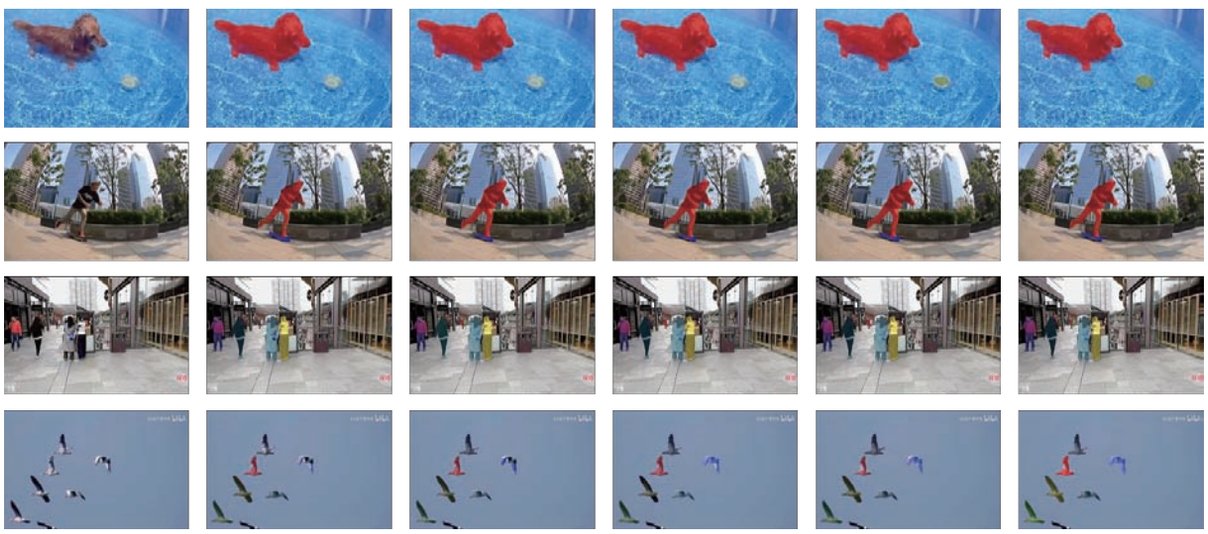


图10 不同方法在多个骨干网络上的分割性能比较

图11展示了本方法与其他方法的主观结果比较[92]. 可以看到，其他方法在分割结果中会将一些实例误识为背景，但本方法提供了更完整的分割. 此外，与SeqFormer和Mask2Former-VIS相比，本方法可以分割密集实例和小实例，而VITA和Mask2Former-VIS会将两个实例视为单个个体. 这是因为视频特征在不同尺度得到增强和集成，三级

联Transformer解码器可以在不牺牲场景信息的情况下提取全局、平衡和局部实例特征. 同时，高斯先验可以学习一系列密集实例帧间的高斯分布，提高实例分割的一致性.



（a）视频帧　（b）IFC　（c）SeqFormer　（d）Mask2Former-VIS　（e）VITA　（f）Gp3Former（Our）

图11　不同方法主观结果展示

3.3　研究热点

3.3.1　SAM大模型

分割一切（Segment Anything，SAM）[93]模型在1100万张图像上训练了超过10亿个掩码，具有强大的分割精度和泛化能力. 受限于SAM的训练成本，现有工作通常直接对SAM进行微调并应用于下游任务. 例如，用于息肉分割的SAM（Transfer SAM for Polyp Segmentation，Polyp-SAM）[94]使用息肉图像微调SAM参数，并用于医学影像息肉分割；类似的，用于土木设施缺陷评估的SAM[95]评估了SAM与U-Net网络对于裂缝分割的效果，并将其应用于图像裂缝检测；此外，异常分割一切（Segment Any Anomaly，SAA+）[96]在SAM中添加了混合快速正则化的零样本快速异常分割，以提高现代基础模型的适应性.

在视频任务中，SAM通常被用于分割视频帧的目标对象，再通过关联目标在不同帧上的分割结果实现视频分割. 例如，分割和追踪一切（Segment and Track Anything，SAM-Track）[97]采用SAM分割视频帧内的目标对象，关联目标在不同视频帧上的位置；类似的，追踪一切模型[98]使用SAM进行掩码预测，实现视频实例分割. SAM具有较强的分割精度和泛化能力，使其可以被应用于各个下游任务中. 然而，该模型具有较高的训练代价限制了对该工作的进一步研究与改进.

3.3.2　开放词汇分割

开放词汇旨在定位和识别标签空间之外的类别，是目前图像视频领域比较热门的研究工作之一. 近年来，随着Transformer的迅速发展，开放词汇受到越来越多的关注.

在图像分割任务中，基于Transformer的融合模块（Transformer-based Fusion Module，Fusioner）[99]将来自语言编码器的类别标签文本特征与输入图像特征对齐，使得Fusioner能够利用语言编码器的泛化能力，分割得到输入文本指定的目标. 零镜头语义分割（Zero-Shot Semantic Segmentation，ZS3）[100]使用文本编码器提取特征对分割区域进行分类，并应用对比语言-图像预训练方法（Contrastive Language-Image Pre-Training，CLIP）视觉编码器为其获取语言对齐的视觉特征. 基于可信标记模块的CLIP（Trusty-Aware Guided CLIP，TagCLIP）[101]在对每个像素进行分类之前明确预测包含对象的像素，从而避免了模型容易将像素误认为新类别的问题.

开放词汇在视频任务中也有一部分相关的工作，例如面向开放词汇的视频实例分割（Towards Open-Vocabulary Video Instance Segmentation，OV2Seg）[102]引入冻结的CLIP主干网络，提出了一种端到端的方法，使用开放式词汇分类器来分割和跟踪未见类别. 另一种开放词汇视频实例分割（Open-Vocabulary Video Instance Segmentation，

OpenVIS)[103]方法尝试分割得到与类别无关的实例区域,然后输入CLIP视觉编码器以判断区域类别.在此基础上,用于视频实例分割的CLIP(Adapting CLIP for Open-Vocabulary Video Instance Segmentation,CLIP-VIS)[104]采用单一的CLIP判断像素类别,并设计了Query匹配规则以自适应地选择最匹配的视频帧,在保障分割性能的前提下简化了分割流程.

3.3.3 指代分割

指代分割是在自然语言描述的指导下,将图像/视频中的目标对象分离出来.这项任务涉及多模态信息融合、自然语言表达的多变性和模型鲁棒性等独特挑战.近年来,深度学习技术的出现带来了解决这些问题的创新思路和方法.

在图像指代分割任务中,先定位后分割策略(Locate-Then-Segment,LTS)[105]将图像分割任务拆分为定位和分割两个阶段.定位阶段使用跨模态交互模块融合语言和视觉特征以获得跨模态表征,分割阶段使用轻量级分割网络,根据检测框和定位阶段获得的跨模态特征生成详精细分割结果.语言感知的视觉Transformer(Language-Aware Vision Transformer,LAVT)[106]在Swin主干网络采用了多层次分割结构,并设计了多模态融合单元来评估单个像素位置与相应语言属性之间的相互关系.掩码标记方法Mask Grounding[107]使用BERT处理指代词汇,并用特殊的掩码标记随机替换一些单词标记.它融合了图像文本特征,并使用分词掩码作为监督信息,以提高语言特征和视觉特征之间的一致性.多标签网络(Multi-Mask Network,MMNet)[108]在分割过程中融合了全局文本和视觉特征,并生成了代表所指表达不同方面的多个Query.这些Query和全局多模态特征被输入视觉语言解码器获得解码特征.这种方法通过自动生成用于监督学习的掩码,降低了注释成本和模型复杂度.

在视频任务中,用于视频分割的语言感知时空协作[109]方法通过跨阶段特征采样和语义传播来分割不同区域,高亮与语言兼容的前景视觉特征,从而促进了视觉和文本在时间和空间维度上的关联与协作.语义辅助对象聚类(Semantic-Assisted Object Cluster,SOC)[110]方法引入了视觉语言对比学习来构建多模态空间,从而将时间建模和跨模态对齐统一到一个简单的分割架构中.频谱引导的多粒度策略(Spectrum-Guided Multi-Granularity,SgMg)[111]通过光谱引导的跨模态融合,引导视觉语言在光谱领域进行全局互动,从而避免指代视频分割过程中出现的特征漂移问题.

指代分割需要语言和视觉之间的有效协调和推理,以准确分割图像中的目标区域.目前,该方法在图像视频分割方面取得了不错的进展,但考虑到模型的通用性和复杂性,依然面临着较大的挑战.

4 总结与展望

4.1 存在的问题

现有的图像语义分割方法通过改进窗口化和注意力机制来平衡长距离依赖和计算效率.这种方法在处理大规模图像数据时表现出一定的优势,但它面临两个关键问题.首先,特定的窗口化设计往往只针对特定的问题场景,这限制了模型在不同类型图像数据上的泛化能力.例如,针对城市街景设计的窗口化策略可能不适用于自然景观的分割,因为这两种场景的视觉特征和结构差异较大;其次,窗口化方法在本质上是通过局部信息间接提取全局特征,这可能影响模型在理解全局上下文时的效果,进而影响收敛速度.

在图像实例分割任务中,Query的作用是区分不同实例,这对于实现高精度分割至关重要.尽管当前方法在使用Query进行实例分割方面取得了显著进展,但仍面临着挑战.例如,图像中一个完整的实例可能被错误地拆分为多个区域,其中每个区域会有一个Query表征.这不仅会影响分割结果的完整性,也降低了整体实例分割的准确性.因此,如何优化Query的分配策略,以确保每个实例被准确且唯一地识别和分割,是图像实例分割领域中一个亟待解决的问题.在视频语义分割任务中,时间信息融合通过分析连续帧的动态变化,有效提升了目标在不同视频帧的语义一致性.然而,当视频中的目标被遮挡,或者多个目标位置过于接近时,这种时间信息融合可能导致目标关联失败.这一现象的根源在于模型的鲁棒性不足,当某一帧中的特征包含冗余或误导性信息时,可能会干扰模型对整个视频序列的分割结果.这不仅影响分割准确性,还可能限制算法在实时场景中的应用.

视频级Query的实例分割方法提升了实例关联的鲁棒性和泛化能力,该方法通过初始化固定长度的Query,有效地关联视频帧中的目标实例.然而,现阶段的视频实例分割主要用于时长较短的视频中,在处理长视频时面临着挑战.长视频需要更长的Query以覆盖整个序列,从而大幅增加计算量,难

以实时捕捉场景变化和目标动态的复杂变化．因此，有必要考虑计算效率和实例关联能力之间的平衡，以满足Transformer处理长视频序列的需求．

4.2 未来的发展方向

Transformer模型通常需要庞大的计算资源和参数量，导致分割效率低下．在图像分割领域中，Transformer中的自注意力机制导致计算代价随输入图像尺寸的增加而呈指数型上升的趋势．此外，在视频分割领域，Transformer巨大的计算量还限制了模型对更长时序信息的关注，同时也降低了处理速度．如何在保持模型性能的同时提高分割效率，是当前的一个重要挑战．

虽然现有的图像分割工作提出了一些方法将计算代价降低至线性复杂度，但是这些方法大多采用局部窗口等减少或限制输入数据的方法，在一定程度上减少了模型可学习的信息量，导致限制了分割性能的上限．因此，下一步工作需要考虑如何在不损失信息量的前提下降低计算复杂度，从而平衡分割精度与效率．

视频领域中目标的时空一致性，即在不同帧之间保持目标的连续性和准确性，是影响模型实际应用效果的关键因素之一．针对此问题，许多工作利用Query学习和关联视频帧中的实例．然而，随着实例的位置变化，不同Query之间关注的区域是否会发生交替或者重叠，这会不可避免地影响模型性能．因此，接下来有必要梳理或者制定更详细的Query训练规则，确保Query之间不会相互影响．

致 谢 本工作受到国家自然科学基金(62471013, 61971016)、北京市自然科学基金-市教委联合资助项目(KZ201910005007)资助．此外，我们向对论文提出宝贵意见的审稿专家们表示衷心的感谢!

参 考 文 献

[1] Yao Z, Wang S, Bao Y. K-Query: Panoptic segmentation with keypoint-based query. Chinese Journal of Computers, 2023, 46(8): 1-16 (in Chinese)
(姚治成，王卅，包云岗．K-Query：基于关键点查询的全景分割方法．计算机学报，2023，46(8)：1－16)

[2] Zhang S, Xu Y, Wu Z, et al. CTC-Net: A novel coupled feature-enhanced Transformer and inverted convolution network for medical image segmentation//Proceedings of the Asian Conference on Pattern Recognition. Kitakyushu, Japan, 2023: 273-283

[3] Yu Z, Lee F, Chen Q. HCT-Net: Hybrid CNN-Transformer model based on a neural architecture search network for medical image segmentation. Applied Intelligence, 2023, 53(17): 19990-20006

[4] Wang J, Zhang Z, Luo L, et al. DualSeg: Fusing Transformer and CNN structure for image segmentation in complex vineyard environment. Computers and Electronics in Agriculture, 2023, 206(1): 107682

[5] Liu Q, Wu J, Jiang Y, et al. InstMove: Instance motion for object-centric video segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 6344-6354

[6] Liang Y, Li X, Tsai B, et al. V-FloodNet: A video segmentation system for urban flood detection and quantification. Environmental Modelling & Software, 2023, 160(1): 105586

[7] Qu Z, Zhuo L, Cao J, et al. TP-Net: Two-path network for retinal vessel segmentation. IEEE Journal of Biomedical and Health Informatics, 2023, 27(4): 1979-1990

[8] Li X, Zhou Y, Yin H, et al. Detecting absence of bone wall in jugular bulb by image Transformation surrogate tasks. IEEE Transactions on Medical Imaging, 2022, 41(6): 1358-1370

[9] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need//Proceedings of the Advances in Neural Information Processing Systems. Long Beach, USA, 2017: 5998-6008

[10] Zhu X, Su W, Lu L, et al. Deformable DETR: Deformable Transformers for end-to-end object detection//Proceedings of the International Conference on Learning Representations. Virtual, 2020: 1-16

[11] Wang H, Zhu Y, Adam H, et al. Max-Deeplab: End-to-end panoptic segmentation with mask Transformers//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 5463-5474

[12] Pan X, Xia Z, Song S, et al. 3D Object detection with pointformer//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 7463-7472

[13] Wang J, Wu Y, Chi H. Embedding VLAD in Transformer for video question answering. Chinese Journal of Computers, 2023, 46(4): 671-689 (in Chinese)
(王继禾，吴颖，迟恒喆．嵌入局部聚类描述符的视频问答Transformer模型．计算机学报，2023，46(4)：671-689)

[14] Zheng S, Lu J, Zhao H, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with Transformers//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 7463-7472

[15] Lazarow J, Xu W, Tu Z. Instance segmentation with mask-supervised polygonal boundary Transformers//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 4382-4391

[16] Ke L, Danelljan M, Ding H, et al. Mask-free video instance segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 22857-22866

[17] Li X, Ding H, Zhang W, et al. Transformer-based visual segmentation: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 1(1): 1-24

[18] Wang W, Zhou T, Porikli F, et al. A survey on deep learning technique for video segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(6): 7099-7122

[19] Ke L, Danelljan M, Li X, et al. Mask transfiner for high-quality instance segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 4412-4421

[20] Yang L, Fan Y, Xu N. Video instance segmentation//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Republic of Korea, 2019: 5188 - 5197

[21] Rottensteiner F, Sohn G, Gerke M, et al. ISPRS 2D semantic labeling challenge-potsdam data, http://www2. isprs. org/commissions/comm3/wg4/2d-sem-label-potsdam.html

[22] Rottensteiner F, Sohn G, Gerke M, et al. ISPRS 2D semantic labeling challenge-vaihingen data, http://www2. isprs. org/commissions/comm3/wg4/2d-sem-label-vaihingen.html

[23] Staal J, Abrämoff M, Niemeijer M, et al. Ridge-based vessel segmentation in color images of the retina. IEEE Transactions on Medical Imaging, 2004, 23(4): 501-509

[24] Owen C, Rudnicka A, Mullen R, et al. Measuring retinal vessel tortuosity in 10-year-old children: Validation of the computer-assisted image analysis of the retina (CAIAR) program. Investigative Ophthalmology Visual Science, 2009, 50(5): 2004-2010

[25] Demir I, Koperski K, Lindenbaum D, et al. DeepGlobe 2018: A challenge to parse the earth through satellite images//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Salt Lake City, USA, 2018: 17200-17209

[26] Mnih V. Machine learning for aerial image labeling, https://www.kaggle.com/datasets/balraj98/massachusetts-roads-dataset

[27] Everingham M, Van Gool L, Williams C, et al. The pascal visual object classes (VOC) challenge. International Journal of Computer Vision, 2010, 88(1): 303-338

[28] Mottaghi R, Chen X, Liu X, et al. The role of context for object detection and semantic segmentation in the wild//Proceedings of The IEEE/CVF Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014: 891-898

[29] Neuhold G, Ollmann T, Rota Bulo S, et al. The mapillary vistas dataset for semantic understanding of street scenes//Proceedings of the IEEE/CVF International Conference on Computer Vision. Venice, Italy, 2017: 4990-4999

[30] Lin T, Maire M, Belongie S, et al. Microsoft COCO: Common objects in context//Proceedings of the European Conference on Computer Vision. Zurich, Switzerland, 2014: 740-755

[31] Zhou B, Zhao H, Puig X, et al. Semantic understanding of scenes through the ADE20K dataset. International Journal of Computer Vision, 2019, 127(1): 302-321

[32] Cordts M, Omran M, Ramos S, et al. The cityscapes dataset for semantic urban scene understanding//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 3213-3223

[33] Miao J, Wei Y, Wu Y, et al. VSPW: A large-scale dataset for video scene parsing in the wild//Proceedings of The IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 4133-4143

[34] Ranftl R, Bochkovskiy A, Koltun V. Vision Transformers for dense prediction//Proceedings of the IEEE/CVF International Conference on Computer Vision. Virtual, 2021: 12179-12188

[35] Liu Z, Lin Y, Cao Y, et al. Swin Transformer: Hierarchical vision Transformer using shifted windows//Proceedings of the IEEE/CVF International Conference on Computer Vision. Virtual, 2021: 10012-10022

[36] Yu Q, Xia Y, Bai Y, et al. Glance-and-gaze vision Transformer//Proceedings of the Advances in Neural Information Processing Systems. Virtual, 2021: 12992-13003

[37] Dong X, Bao J, Chen D, et al. CSWin Transformer: A general vision Transformer backbone with cross-shaped windows//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 12124-12134

[38] Wang W, Xie E, Li X, et al. Pyramid vision Transformer: A versatile backbone for dense prediction without convolutions//Proceedings of the IEEE/CVF International Conference on Computer Vision. Virtual, 2021: 568-578

[39] Deng G, Wu Z, Xu M, et al. Crisscross-global vision Transformers model for very high resolution aerial image semantic segmentation. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61(1): 4404019

[40] Hu J, Shen L, Sun G. Squeeze-and-excitation networks//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 7132-7141

[41] Huang L, Yuan Y, Guo J, et al. Interlaced sparse self-attention for semantic segmentation. Memory, 2019(1): 2500-2510

[42] Chen J, Yi J, Chen A, et al. SRCBTFusion-Net: An efficient fusion architecture via stacked residual convolution blocks and Transformer for remote sensing image semantic segmentation. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61(1): 4411716

[43] Ma X, Zhang X, Pun M, et al. A multilevel multimodal fusion Transformer for remote sensing semantic segmentation. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62(1): 5403215

[44] Xiang X, Gong W, Li S, et al. TCNet: multiscale fusion of Transformer and CNN for semantic segmentation of remote sensing images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2024, 17(1): 3123 - 3136

[45] Guo X, Lin X, Yang X, et al. UCTNet: Uncertainty-guided CNN-Transformer hybrid networks for medical image segmentation. Pattern Recognition, 2024, 152(1): 110491

[46] Wu H, Zhang M, Huang P, et al. CMLFormer: CNN and

multi-scale local-context Transformer network for remote sensing images semantic segmentation. IEEE Journal of Selected Topics in Applied Earth Observations Remote Sensing, 2024, 17(1): 7233-7241

[47] Wu H, Huang P, Zhang M, et al. CMTFNet: CNN and multiscale Transformer fusion network for remote sensing image semantic segmentation. IEEE Transactions on Geoscience Remote Sensing, 2023, 61(1): 2004612

[48] Xiao T, Liu Y, Huang Y, et al. Enhancing multiscale representations with Transformer for remote sensing image semantic segmentation. IEEE Transactions on Geoscience Remote Sensing, 2023, 61(1): 5605116

[49] Liu H, Li W, Jia W, et al. Clusterformer for pine tree disease identification based on UAV remote sensing image segmentation. IEEE Transactions on Geoscience Remote Sensing, 2024, 62(1): 5609215

[50] Qiu L, Yu D, Zhang X, et al. Efficient remote sensing segmentation with generative adversarial Transformer. IEEE Geoscience Remote Sensing Letters, 2024, 21(1): 6000905

[51] Mo Y, Li H, Xiao X, et al. Swin-conv-dspp and global local Transformer for remote sensing image semantic segmentation. IEEE Journal of Selected Topics in Applied Earth Observations Remote Sensing, 2023, 16(1): 5284-5296

[52] Xu Z, Geng J, Jiang W. MMT: Mixed-mask Transformer for remote sensing image semantic segmentation. IEEE Transactions on Geoscience Remote Sensing, 2023, 61(1): 5613415

[53] Li W, Zhang J, Li J, et al. Unpaved road segmentation of UAV imagery via a global vision Transformer with dilated cross window self-attention for dynamic map. The Visual Computer, 2024, 1(1): 1-19

[54] Wu X, Zhang J, Li W, et al. Spatial-specific Transformer with involution for semantic segmentation of high-resolution remote sensing images. International Journal of Remote Sensing, 2023, 44(4): 1280-1307

[55] Wu X, Semantic segmentation technology of high-resolution remote sensing images based on deep learning [Master Thesis], Beijing University of Technology, Beijing, China, 2023 (in Chinese)
(吴鑫嘉，基于深度学习的高分辨率遥感影像语义分割技术[硕士学位论文]，北京工业大学，北京，中国，2023)

[56] Ye W, Zhang W, Lei W, et al. Remote sensing image instance segmentation network with Transformer and multi-scale feature representation. Expert Systems with Applications, 2023, 234(1): 121007

[57] Liang C, Zhou T, Liu D, et al. ClustSeg: Clustering for universal segmentation//Proceedings of the International Conference on Machine Learning. Hawaii, USA, 2023: 20787-20809

[58] Cheng B, Misra I, Schwing A, et al. Masked-attention mask Transformer for universal image segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 1290-1299

[59] Hu J, Cao L, Lu Y, et al. ISTR: End-to-end instance segmentation with Transformers. arXiv preprint arXiv: 2105.00637, 2021

[60] Lu J, Deng J, Wang C, et al. Query refinement Transformer for 3D instance segmentation//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France, 2023: 18516-18526

[61] Yu X, Shi D, Wei X, et al. SOIT: Segmenting objects with instance-aware Transformers//Proceedings of the AAAI Conference on Artificial Intelligence. Virtual, 2022: 3188-3196

[62] Prangemeier T, Reich C, Koeppl H. Attention-based Transformers for instance segmentation of cells in microstructures//Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine. Virtual, 2020: 700-707

[63] He J, Li P, Geng Y, et al. FastInst: A simple query-based model for real-time instance segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 23663-23672

[64] Biswas S, Banerjee A, Lladós J, et al. DocsegTr: An instance-level end-to-end document image segmentation Transformer. arXiv preprint arXiv:2201.11438, 2022

[65] Liang J, Homayounfar N, Ma W, et al. Polytransform: Deep polygon Transformer for instance segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2020: 9131-9140

[66] Su J, Yin R, Zhang S, et al. Motion-state alignment for video semantic segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 3570-3579

[67] Zhuang J, Wang Z, Li J. Video semantic segmentation with inter-frame feature fusion and inner-frame feature refinement. arXiv preprint arXiv:2301.03832, 2023

[68] Sun G, Liu Y, Ding H, et al. Coarse-to-fine feature mining for video semantic segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 3126-3137

[69] Paul M, Danelljan M, Van Gool L, et al. Local memory attention for fast video semantic segmentation//Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems Virtual, 2021: 1102-1109

[70] Grammatikopoulou M, Ricardo S, Bragman F, et al. A spatio-temporal network for video semantic segmentation in surgical videos. International Journal of Computer Assisted Radiology Surgery, 2023, 19(2): 375-382

[71] Liang Z, Dong W, Zhang B. A dual-branch hybrid network of CNN and Transformer with adaptive keyframe scheduling for video semantic segmentation. Multimedia Systems, 2024, 30(2): 67-79

[72] Liang Z, Dai X, Wu Y, et al. Multi-granularity context network for efficient video semantic segmentation. IEEE Transactions on Image Processing, 2023, 32(1): 3163-3175

[73] Li J, Wang W, Chen J, et al. Video semantic segmentation via sparse temporal Transformer//Proceedings of the 29th ACM International Conference on Multimedia. New York, USA,

2021: 59-68

[74] Sun G, Liu Y, Tang H, et al. Mining relations among cross-frame affinities for video semantic segmentation//Proceedings of the European Conference on Computer Vision. Tel Aviv, Israel, 2022: 522-539

[75] Weng Y, Han M, He H, et al. Mask propagation for efficient video semantic segmentation//Proceedings of the Advances in Neural Information Processing Systems. New Orleans, USA, 2023: 131-145

[76] Guo D, Fan D, Lu T, et al. Vanishing-point-guided video semantic segmentation of driving scenes//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2024: 3544-3553

[77] Hu K, Xie Z, Hu Q. Lightweight convolutional neural networks with context broadcast Transformer for real-time semantic segmentation. Image and Vision Computing, 2024, 146(1): 105053

[78] Wang Y, Xu Z, Wang X, et al. End-to-end video instance segmentation with Transformers//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 8741-8750

[79] Cheng B, Choudhuri A, Misra I, et al. Mask2Former for video instance segmentation//Proceedings of the Advances in Neural Information Processing Systems. New Orleans, USA, 2022: 23109-23120

[80] Yang S, Wang X, Li Y, et al. Temporally efficient vision Transformer for video instance segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 2885-2895

[81] Hwang S, Heo M, Oh S W, et al. Video instance segmentation using inter-frame communication Transformers//Proceedings of the Advances in Neural Information Processing Systems. Virtual, 2021: 13352-13363

[82] Zhang R, Cheng T, Yang S, et al. MobileInst: Video instance segmentation on the mobile//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver, Canada, 2024: 7260-7268

[83] Caelles A, Meinhardt T, Brasó G, et al. DeVIS: Making deformable Transformers work for video instance segmentation. arXiv preprint arXiv:2207.11103, 2022

[84] Huang D, Yu Z, Anandkumar A. MinVIS: A minimal video instance segmentation framework without video-based training//Proceedings of the Advances in Neural Information Processing Systems. New Orleans, USA, 2022: 31265-31277

[85] Heo M, Hwang S, Oh S W, et al. VITA: Video instance segmentation via object token association//Proceedings of the Advances in Neural Information Processing Systems. New Orleans, USA, 2022: 23109-23120

[86] Wu J, Jiang Y, Bai S, et al. SeqFormer: Sequential Transformer for video instance segmentation//Proceedings of the European Conference on Computer Vision. Tel Aviv, Israel, 2022: 553-569

[87] Fang H, Zhang T, Zhou X, et al. Learning better video query with SAM for video instance segmentation. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 1(1): 1-13

[88] Koner R, Hannan T, Shit S, et al. Instanceformer: An online video instance segmentation framework//Proceedings of the AAAI Conference on Artificial Intelligence. Washington, USA, 2023: 1188-1195

[89] Kim H, Lee S, Kang H, et al. Offline-to-online knowledge distillation for video instance segmentation//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA, 2024: 158-167

[90] Lan M, Zhang J, Zhang L, et al. Learning to learn better for video object segmentation//Proceedings of the AAAI Conference on Artificial Intelligence. Washington, USA, 2023: 1205-1212

[91] Li J, Yu B, Rao Y, et al. TCOVIS: Temporally consistent online video instance segmentation//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France, 2023: 1097-1107

[92] Li W, Zhang J, Zhuo L. Gp3Former: Gaussian prior tri-cascaded Transformer for video instance segmentation in livestreaming scenarios. IEEE Transactions on Emerging Topics in Computational Intelligence, 2024, 1(1): 1-15

[93] Kirillov A, Mintun E, Ravi N, et al. Segment anything//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France, 2023: 4015-4026

[94] Li Y, Hu M, Yang X. Polyp-SAM: Transfer SAM for polyp segmentation. Medical Imaging: Computer-aided Diagnosis, 2024, 12927(1): 759-765

[95] Ahmadi M, Lonbar A, Sharifi A, et al. Application of segment anything model for civil infrastructure defect assessment. arXiv: 2304.12600, 2023

[96] Cao Y, Xu X, Sun C, et al. Segment any anomaly without training via hybrid prompt regularization. arXiv preprint arXiv: 2305.10724, 2023

[97] Cheng Y, Li L, Xu Y, et al. Segment and track anything. arXiv preprint arXiv:2305.06558, 2023

[98] Yang J, Gao M, Li Z, et al. Track anything: Segment anything meets videos. arXiv preprint arXiv:2304.11968, 2023

[99] Ma C, Yang Y, Wang Y, et al. Open-vocabulary semantic segmentation with frozen vision-language models//Proceedings of the British Machine Vision Conference. London, UK, 2022: 327-332

[100] Ding J, Xue N, Xia G, et al. Decoupling zero-shot semantic segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 11583-11592

[101] Li J, Chen P, Qian S, et al. TagCLIP: Improving discrimination ability of open-vocabulary semantic segmentation. arXiv preprint arXiv:2304.07547, 2024

[102] Wang H, Yan C, Wang S, et al. Towards open-vocabulary video instance segmentation//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France, 2023: 4057-4066

[103] Guo P, Huang T, He P, et al. OpenVIS: Open-vocabulary video instance segmentation. arXiv preprint arXiv: 2305.16835,

2023

[104] Zhu W, Cao J, Xie J, et al. CLIP-VIS: Adapting CLIP for open-vocabulary video instance segmentation. arXiv preprint arXiv:2403.12455, 2024

[105] Jing Y, Kong T, Wang W, et al. Locate then segment: A strong pipeline for referring image segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA, 2021: 9858-9867

[106] Yang Z, Wang J, Tang Y, et al. LAVT: Language-aware vision Transformer for referring image segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 18155-18165

[107] Cheng Y, Zheng H, Han Y, et al. Mask grounding for referring image segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2024: 26573-26583

[108] Yan Y, He X, Wan W, et al. MMNet: Multi-mask network for referring image segmentation. arXiv preprint arXiv: 2305.14969, 2023

[109] Hui T, Liu S, Ding Z, et al. Language-aware spatial-temporal collaboration for referring video segmentation. IEEE Transactions on Pattern Analysis Machine Intelligence, 2023, 45(7): 8646-8659

[110] Luo Z, Xiao Y, Liu Y, et al. SOC: Semantic-assisted object cluster for referring video object segmentation//Proceedings of the advances in Neural Information Processing Systems. New Orleans, USA, 2023: 371-384

[111] Miao B, Bennamoun M, Gao Y, et al. Spectrum-guided multi-granularity referring video object segmentation//Proceedings of the IEEE/CVF International Conference on Computer Vision. Vancouver, Canada, 2023: 920-930

LI Wen-Sheng, Ph. D. candidate. His research interest is video semantic segmentation.

ZHANG Jing, Ph. D., Professor. Her current research interests include image/video processing, computer vision and deep learning.

ZHUO Li, Ph. D., Professor. Her current research interests include image/video processing, computer vision and deep learning.

WU Xin-Jia, master. Her research interest is remote sensing image segmentation.

YAN Yi, master candidate. Her research interest is remote sensing image segmentation.

Background

As one of the tasks in the field of computer vision, visual segmentation plays a crucial role in understanding and interpreting visual information. This technique not only helps to recognize and understand individual objects in a single image, but also reveals the spatial relationships between these objects. However, complex backgrounds, lighting variations, occlusions, and other factors cause visual segmentation tasks to face multiple challenges in practical applications. To cope with these challenges, scholars continue to explore new methods and techniques to improve the accuracy and robustness of segmentation results.

In recent years, a number of visual segmentation methods have emerged, among which the Transformer model has attracted much attention that can dynamically capture the relationships between components in an image or video. With the rapid development of Transformer-based visual segmentation techniques, significant progress has been made in this area of research work. The scholars have systematically organized and analyzed these works to help other researchers quickly track the current state of progress, major findings, and theoretical advances. According to whether labels are used in the training data or not, existing Transformer-based visual segmentation are usually categorized into unsupervised, semi-supervised, and supervised-based methods. In addition, there are also some works summarize and analyze the Transformer-based visual segmentation methods in dedicated fields, such as autonomous driving, aerial remote sensing, and video scene understanding. While these surveys provide valuable and meaningful resources for understanding the application of Transformer in the field of visual segmentation, there is a lack of in-depth exploration of trending research issues such as optimization of model structure, as well as the improvement of self-attention mechanism for Transformer-based visual segmentation.

Therefore, it is an urgent need to more comprehensively understand and recognize the existing progress and development trend of Transformer in field of visual segmentation, and to find out the deficiencies and challenges, so as to explore the core theory of Transformer in a deeper way. This paper organizes, reviews, analyzes and explores the recent advances in Transformer-based visual segmentation techniques from two visual pipelines of image/video, not only summarizing the

theoretical framework of Transformer, but also giving some application examples and research hotspots, so as to make a summary and overlook. Finally, although Transformer-based visual segmentation has received widespread attention, the scientific problems have gradually emerged, limiting the further improvement of model performance and efficiency. Finally, this paper summarizes the changeable issues that still need to be addressed in terms of image/video semantic/instance segmentation tasks using Transformer, and looks forward to the potential future development directions to provide some insights for reference.

This research was supported by the National Natural Science Foundation of China under Grant 62471013, 61971016; in part by the Beijing Municipal Education Commission Cooperation Beijing Natural Science Foundation under Grant KZ201910005007.