CN 51-1346/O4 ISSN 1003-501X (印刷版) ISSN 2094-4019 (网络版)

光电工程
Opto-Electronic Engineering

中文核心期刊 中国科技核心期刊
Scopus CSCD

基于YOLACTR的无锚框实例分割算法

梅婷，赵敬伟，林珊玲，谢子昱，林志贤，郭太良

引用本文:

梅婷，赵敬伟，林珊玲，等. 基于YOLACTR的无锚框实例分割算法[J]. 光电工程，2025， **52**(5): 240265.

Mei T, Zhao J W, Lin S L, et al. Anchor-free instance segmentation algorithm based on YOLACTR[J]. *Opto-Electron Eng*, 2025, **52**(5): 240265.

https://doi.org/10.12086/oee.2025.240265

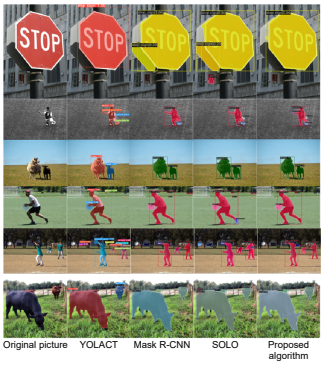
收稿日期: 2024-11-12; 修改日期: 2025-04-07; 录用日期: 2025-04-08

Opto-Electronic Engineering
光 电 工 程

Article
2025 年, 第 52 卷, 第 5 期

DOI: 10.12086/oee.2025.240265 CSTR: 32245.14.oee.2025.240265

基于 YOLACTR 的无锚框实例分割算法



梅 婷[1,2]，赵敬伟[1,2]，林珊玲[2,3*]，谢子昱[1,2]，林志贤[1,2,3]，郭太良[1,2]

[1]福州大学物理与信息工程学院，福建 福州 350116；
[2]中国福建光电信息科学与技术创新实验室，福建 福州 350108；
[3]福州大学先进制造学院，福建 泉州 362200

摘要：针对基于边界框检测的单阶段 YOLACT 算法缺少对感兴趣区域进行定位提取，且两个边界框存在相互重叠而难以区分的问题，基于改进的 YOLACTR 算法，提出一种无锚框实例分割方法，将掩码生成解耦成特征学习和卷积核学习，利用特征聚合网络生成掩码特征，将位置信息添加到特征图，采用多层 Transformer 和双向注意力来获得动态卷积核。实验结果表明，该方法在 MS COCO 公共数据集的掩码精度 (*AP*) 达到 35.2%，相对于 YOLACT 算法，掩码精度提升 25.7%，小目标检测精度提升 37.1%，中等目标检测精度提升 25.8%，大目标检测精度提升 21.9%。相较 YOLACT、Mask R-CNN、SOLO 等方法，所提算法在分割精度和边缘细节保留方面均具有明显优势，特别在重叠物体的分割和小目标检测中表现更为出色，有效解决传统方法在实例边界重叠区域的错误分割问题。

关键词：YOLACT；无锚框实例分割；动态卷积；Transformer

中图分类号：TP391.41 **文献标志码：**A

梅婷，赵敬伟，林珊玲，等. 基于 YOLACTR 的无锚框实例分割算法 [J]. 光电工程，2025，**52**(5): 240265
Mei T, Zhao J W, Lin S L, et al. Anchor-free instance segmentation algorithm based on YOLACTR[J]. *Opto-Electron Eng*, 2025, **52**(5): 240265

Anchor-free instance segmentation algorithm based on YOLACTR

Mei Ting[1,2], Zhao Jingwei[1,2], Lin Shanling[2,3*], Xie Ziyu[1,2], Lin Zhixian[1,2,3], Guo Tailiang[1,2]

[1]College of Physics and Information Engineering, Fuzhou University, Fuzhou, Fujian 350116, China;
[2]Fujian Science & Technology Innovation Laboratory for Optoelectronic Information of China, Fuzhou, Fujian 350108, China;
[3]School of Advanced Manufacturing, Fuzhou University, Quanzhou, Fujian 362200, China

Abstract: Aiming at the problem that the single-stage YOLACT algorithm based on bounding box detection lacks the location and extraction of the region of interest, and the issue that two bounding boxes overlap and are difficult to distinguish, this paper proposes an anchor-free instance segmentation method based on the improved YOLACTR algorithm. The mask generation is decoupled into feature learning and convolution kernel learning, and the feature aggregation network is used to generate mask features. By adding position information to the feature

收稿日期：2024-11-12；修回日期：2025-04-07；录用日期：2025-04-08
基金项目：国家重点研发计划 (2021YFB3600603)；国家自然科学基金青年科学基金项目 (62101132)
*通信作者：林珊玲，526176333@qq.com。

map, multi-layer transformer and two-way attention are used to obtain dynamic convolution kernels. The experimental results show that this method achieves a mask accuracy (*AP*) of 35.2% on the MS COCO public dataset. Compared with the YOLACT algorithm, this method improves the mask accuracy by 25.7%, the small target detection accuracy by 37.1%, the medium target detection accuracy by 25.8%, and the large target detection accuracy by 21.9%. Compared with YOLACT, Mask R-CNN, SOLO, and other methods, our algorithm shows significant advantages in segmentation accuracy and edge detail preservation, especially excelling in overlapping object segmentation and small target detection, effectively solving the problem of incorrect segmentation in instance boundary overlap regions that traditional methods face.

Keywords: YOLACT; anchor-free instance segmentation; dynamic convolution; Transformer

1 引　言

近年来深度学习的发展已经成为新一代信息技术领域的核心板块之一，人工智能已经应用到许多领域之中[1-3]，计算机视觉是人工智能领域中一个主要的、能尽快落地的重要分支，在人工智能领域中有很高的占比。实例分割[4]是计算机视觉领域中具有挑战性的一项新任务，是语义分割和目标检测相结合的综合性任务，需要正确检测图像中所有的实例，并在像素级对每个实例进行标记。即不仅需要对图像中不同类别的实例进行像素级分割，还需要对同一类别的不同实例进行区分[5-7]。

实例分割算法从处理过程可以归纳为两阶段检测算法和单阶段检测算法两大类。两阶段检测实例分割是指实现实例分割主要有两个过程，一是提取物体区域，二是对区域进行实例分割。Mask R-CNN[8]是经典的两阶段检测算法，该算法是在 Faster R-CNN[9]的基础上，在目标分类和边界框回归分支上添加了一个并行的掩码分支来预测分割结果，并将 ROI Pooling 层替换成了 ROI Align，ROI Align 的提出解决了 Faster R-CNN 中 ROI Pooling 的区域不匹配的问题。Cascade Mask R-CNN[10]通过引入多阶段的级联结构，逐步优化分割结果，增强了对复杂场景中实例的分割能力[11]。此外，HTC (hybrid task cascade)[12]进一步将分割任务与其他视觉任务 (如目标检测) 进行联合优化，显著提升了模型的整体性能。两阶段检测算法普遍计算量大，模型复杂度高，为了提高模型的检测精度而大幅降低了检测速度，在实时检测场景中，很难发挥其应用价值。

单阶段检测算法取消了物体区域提取，直接对特征进行检测和分割来获得实例结果。该方法将检测和分类任务集成到单个神经网络中，无需先生成候选区域，简化了检测过程，显著提高了检测速度，减少了计算资源使用。基于锚框的单阶段目标检测方法有 YOLO[13]和 RetinaNet[14]，无锚框的检测方法有 FCOS[15]、CenterNet[16]和 SOLO[17]等。单阶段锚框、无锚框的分割方法大多采用 ResNet[18]与特征金字塔 (feature pyramid network, FPN[19]) 相结合的结构来学习多尺度特征，以实现特征提取的尺度不变性，同时检测器架构简单且无需预先生成锚框的特点，大幅加快算法的运行速度[20]。但是该方法存在尺度对不齐、区域对不齐和任务对不齐的问题。

近年来，无锚框的实例分割方法也逐渐受到关注。PointRend 通过对感兴趣区域的边界进行逐点优化，实现了更精细的分割结果，尤其在边界复杂的实例中表现出色[21]。此外，QueryInst 将 Transformer 架构引入实例分割任务，通过查询机制实现了对实例的高效建模和分割[22]。与这些方法不同，Mask2Former 提出了一种更加统一的框架，旨在处理全景分割、实例分割和语义分割等多种任务[23]。其核心创新在于引入掩码注意力机制 (masked attention)，使得模型能够在预测掩模时，仅关注被遮盖区域，从而更高效地提取局部特征。

尽管当前的实例分割方法取得了较大进展，但依然存在一些问题，特别是在实时性和检测精度方面。其中，YOLACT[24]算法以单阶段检测为基础，结合原型掩码 (prototype masks) 和掩码系数首次实现了实时实例分割[13]。但是该算法仍然在实例检测上存在漏检、检测精度低的问题。同时，YOLACT 算法是基于边界框检测的单阶段检测算法，与两阶段边框检测算法不同，由于缺少对感兴趣区域进行定位提取，直接利用全卷积网络通过聚合的方式进行实例分割，没有考虑到特征所在空间的位置信息。另外，YOLACT 算法采用边界框确定实例范围，与实例本

身的形状是有较大差异的，尤其在两个边界框存在相互重叠而无法区分的时候，边界框与实例形状相比较为粗糙和直接，并不是自然的人眼所观察到的实例的样子。人眼可以通过观察实例的外观、形状或轮廓来判别实例的类别，因此可以采用掩码的方式来进行实例分割，并对特征添加位置信息，以更自然的方式获得更高的检测精度和效果。

因此，在改进 YOLACT 算法的基础上，提出了基于 YOLACTR[25] 算法的无锚框检测的实例分割算法。该算法在保留 YOLACTR 算法网络结构的基础上，不再依赖边界框的检测，而是将掩码生成解耦成特征学习和卷积核学习。其中卷积核学习是通过对来自特征金字塔的输入特征进行随机位置嵌入，并通过 Transformer Layer 将特征信息处理成二维堆叠块，进一步通过检测头生成类别信息和堆叠的卷积核；而特征学习则是利用来自特征金字塔 (FPN) 的部分特征和来自 Transformer Layer 的高层敏感位置特征，通过特征聚合网络和注意力赋予权重，得到多尺度的掩码特征。最后通过动态卷积的方式对掩码特征和对应的卷积进行融合。

2 相关工作

2.1 YOLACT 算法

YOLACT 模型结构如图 1 所示。在 RetinaNet 的基础上进行改进，结合原型掩码和掩码系数实现实例分割，YOLACT 为每张图片生成 k 个原型掩码和 k 个掩码系数，每个掩码系数和原型掩码一一对应，通过将掩码系数和原型掩码进行线性组合，然后通过 Sigmoid 非线性函数来生成最终的分割结果。生成原型掩码和掩码系数的两个分支实际上分别是语义分割和目标检测，通过简洁的语义分割结构和直接的目标检测方式使得 YOLACT 算法实现了实时实例分割。但是其特征提取部分采用 ResNet 与 FPN 相结合的结构，容易造成小目标特征提取不充分，并且类别分支和掩膜分支的预测结果需要按照系数进行叠加，存在两个任务不匹配的问题。

2.2 YOLACTR 算法

针对 YOLACT 算法存在检测精度较低和小目标漏检的问题，提出一种改进的 YOLACTR 算法模型，YOLACTR 的网络模型如图 2 所示。输入图像经过骨干网络 Resnet 和特征金字塔 FPN 获得多尺度特征，接着将实例分割的复杂任务分解成两个并行任务，这些并行分支最终通过极大值抑制、线性组合来获得最终的分割结果。第一个分支是预测头分支，通过 CNN 与 Transformer 相结合，利用 Transformer 关键部件多头注意力，来提高模型的分割精度，对全局特征的不同位置预测实例类别、边界框位置和掩码系数；第二个分支是掩码头分支，通过特征聚合网络聚合不同尺度的实例信息，并通过添加在不同尺度上的 CS 注意力模块，获得更高效的原型掩码。

为进一步改进算法检测精度，提高掩码的准确性，设计无锚框检测算法，基于所提出的 YOLACTR 算法，将掩码生成解耦成特征信息和动态卷积核，利用特征聚合网络生成掩码特征，通过将位置信息添加到特征图，采用多层 Transformer 和双向注意力来获得动态卷积核，以进一步提升掩码检测的准确性。

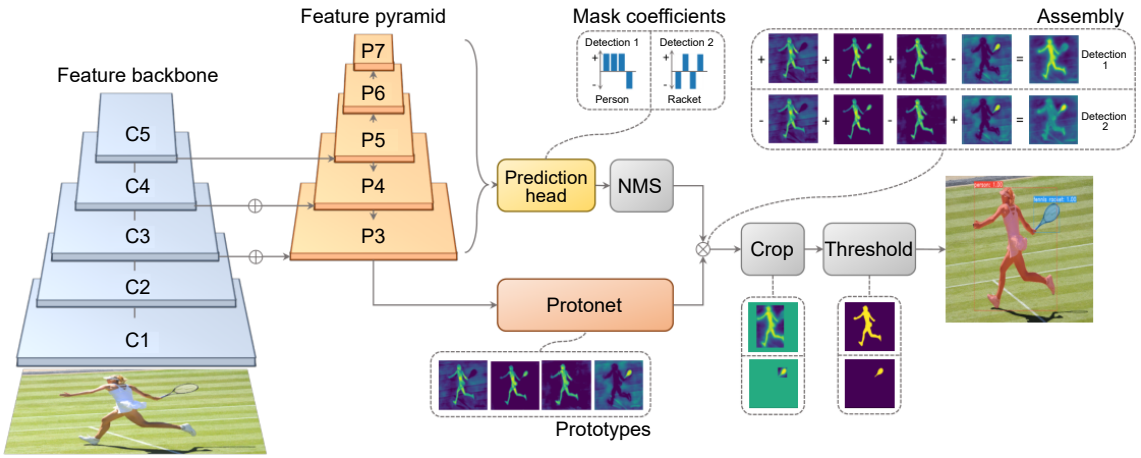


图 1 YOLACT 结构图

Fig. 1 YOLACT structure diagram

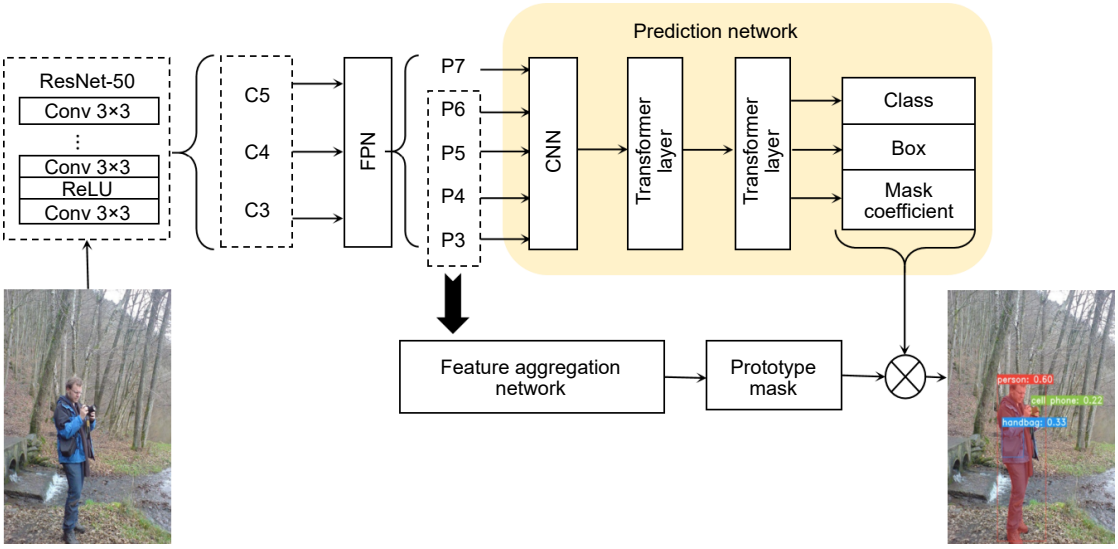


图 2 YOLACTR 网络结构

Fig. 2 YOLACTR network structure

3 无锚框实例分割算法设计

3.1 网络结构

无锚框实例分割算法主要由多尺度特征生成网络、掩模生成网络、头部预测网络和辅助网络结构四部分组成。多尺度特征生成网络由 ResNet 和特征金字塔网络组成，用于生成多尺度图像特征。掩模生成网络将 Transformer 和特征聚合网络相结合，通过位置嵌入和特征融合生成高质量的掩码特征。头部预测网络融合位置信息和 Transformer，生成动态卷积核和实例类别，该部分负责对实例进行分类并生成自适应卷积核，以便后续生成掩码。辅助网络结构用于计算语义分割损失，并有助于提高模型的整体性能和准确性。无锚框实例分割算法的网络模型如图 3 所示。

无锚框实例分割算法主要流程如下所示：

1) 输入图像经过骨干网络 ResNet 和特征金字塔 FPN 获得多尺度特征 P2~P6；

2) 特征 P2~P6 输入到预测网络，对各层网络嵌入随机位置信息，然后送入 6 层 Transformer Layer，然后将特征信息送到分类头和卷积核的函数头，并将附有位置信息的特征层 T5 送到掩码生成网络；

3) 来自 FPN 的特征 P2~P4 和预测网络的特征 T5 输入到掩码生成网络， P2~P4 和 T5 经过特征聚合网络处理获得掩码特征；

4) 最后将获得的掩码特征和对应卷积核进行动态卷积运算，生成最终实例分割掩码。

3.2 随机位置嵌入

YOLACT 算法使用单级边界框检测来定位实例

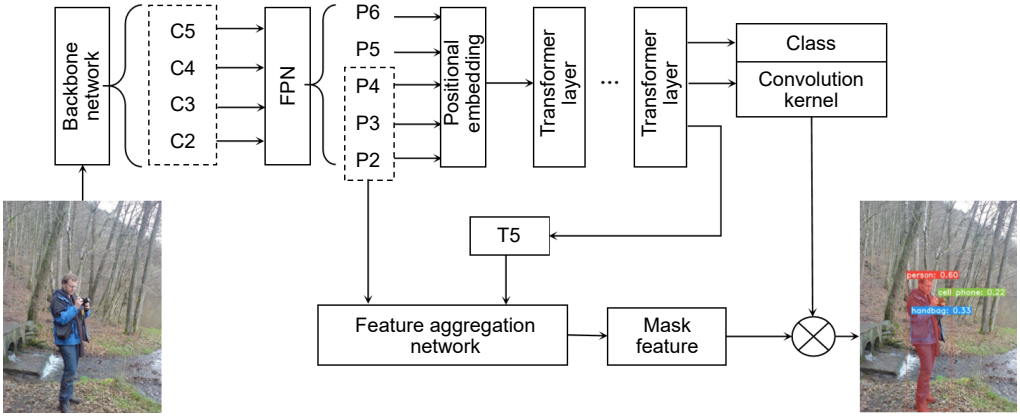


图 3 无锚框实例分割模型结构

Fig. 3 Anchless anchor box example segmentation model structure

的位置信息，但是，此方法并不完全适合实例的真实形状。特别是当图像中存在实例相互遮挡时，实例的真实轮廓可能会因遮挡而失真，从而阻止边界框准确捕获实例的精确位置和形状，边界框检测容易发生冲突，导致检测结果不理想。在上述情况下，YOLACT 算法可能会将遮挡的实例误认为是单独的实例，或者错误地将被某些实例遮挡的区域分配给其他实例，从而导致检测结果不准确。因此，YOLACT 在处理实例遮挡时存在一定的局限性，需要进一步改进以改善检测结果。如图 4 所示，采用无锚框进行检测，即不再使用边界框进行检测，减少了计算边界框位置所需的大量计算资源，但缺少图像的敏感位置信息，因此通过嵌入不同尺寸的位置信息来增加位置灵敏度，通过利用双向注意力来区分不同的实例，从而实现更好的检测效果，其中 μ 为均值。

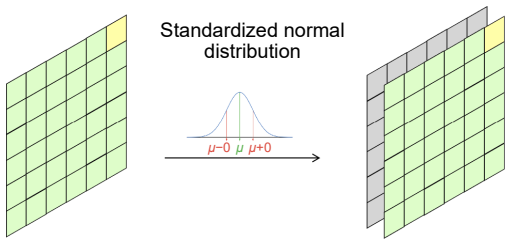


图 4 位置嵌入示意图

Fig. 4 Schematic diagram of the positional embedding

位置编码 (Positional encoding) 在实例分割任务中扮演着重要角色，尤其是在基于 Transformer 的模型中。由于 Transformer 架构本身缺乏对输入数据顺序或空间位置的内在感知能力，因此需要通过位置编码来显式地提供这些信息，从而帮助模型更好地理解图像中像素或特征的空间关系。位置编码包括绝对位置编码、相对位置编码、条件位置编码、上下文位置编码。

采用符合标准正态分布的动态位置信息进行嵌入，通过动态生成位置编码，使其能够适应输入数据的特定分布特性。尽管这种方法并不一定是最优的，但其简单易用，仍能有效达到预期目的。嵌入尺寸与各尺度特征图一致，并将位置信息进行注册，使其像索引一样，可以被其他 Module 检测到。另外，在 Transformer 中，特征图被拆分成 $N\times N$ 的面片堆叠起来作为输入，因此在双向注意力处理中行和列所在的位置嵌入空间为 $1\times N\times C$。

3.3 实例分割网络

无锚框实例分割算法通过将掩码生成解耦成掩码特征学习和对应卷积核学习，两部分是分开进行预测的。由于不再采用边框方式定位掩码位置，为了更进一步得到准确的位置信息和掩码特征，在特征提取网络中，提高对底层特征的利用率，将 P2 特征层融合进来。

首先，对于掩码生成网络，网络结构如图 5 所示。

为了构建实例感知和位置敏感的高分辨率掩码特征表示，新的特征聚合网络将来自 FPN 的 P2 特征层作为最后的融合特征，并将来自预测网络的带有位置信息的特征图 P5 来替代 FPN 生成 P5 特征图，与 FPN 的 P2~P4 相结合进行融合。新的特征网络将以 P3 作为中间融合特征层，这么做的目的是避免特征在长距离的上下采样的过程中损失较多信息。对于 P2~P5 每个尺度的特征图，首先进行卷积核大小为 3×3，步长为 1 的卷积、归一化和 ReLU 非线性激活，得到 P21、P31、P41、P51。将 P41 进行双线性上采样 2×，并输入到 CS 注意力，然后与 P31 进行加和得到 P32。将 P51 进行双线性上采样，继续进行卷积核大小为 3×3，步长为 1 的卷积、归一化和 ReLU 非线性激活操作后，再进行一次双线性上采样，输入 CS 注意力后与 P32 加和得到 P33。将 P21 进行卷积下采

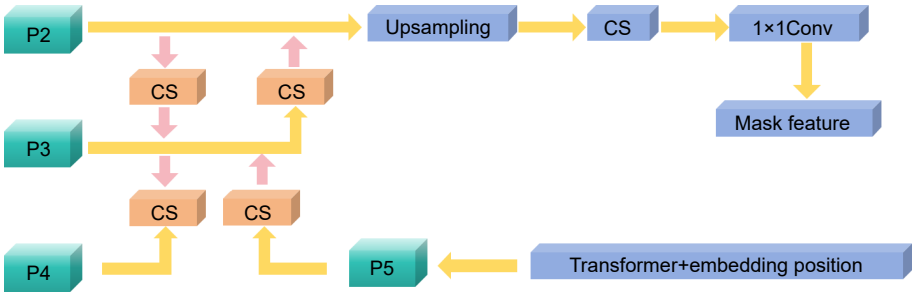


图 5 掩码生成网络结构图

Fig. 5 Structure diagram of mask generation network

样，卷积核大小为 3×3，步长为 2，然后送入 CS 注意力，与 P33 相加后得到 P34。接下来将 P44 进行卷积、双线性上采样、卷积操作，与 P21 进行加和，然后输入到 CS 注意力模块，最后经过 1×1 卷积和上采样后，得到尺寸为 $H\times W$ 的掩码特征图，H 表示最终生成的掩码特征图的空间分辨率的高度；W 表示宽度。

然后，对于预测网络，网络结构如图 6 所示。从特征金字塔得到的 P2~P6 特征图，首先对其进行位置信息嵌入，位置信息通过正态分布进行随机选取，避免位置信息的偶然性。由于添加了位置嵌入信息，所以取消了 Transformer 模块前的卷积操作，同时为了提高检测精度，将 Transformer Layer 增加至 6 层，以获得更好的检测效果。然后对附加位置信息的多尺度特征进行切片处理，使其大小变为 $N\times N$ 的面片，并进行堆叠，将堆叠的特征信息以词嵌入的方式输送到 Transformer 中，经过 Transformer 中的双向注意力处理，得到 $N\times N\times C$（C 为特征通道数）的张量，然后经过不同线性层，用来预测实例类别和卷积核。

最后，对上面得到的统一多尺度的特征图与对应的卷积核通过动态卷积操作来生成掩码，然后经过极大值抑制、裁剪来获得最终结果。其中卷积核 $K\in\mathbb{R}^{N\times N\times M}$（$\mathbb{R}$为实数域），$M$=$\lambda 2C$，$M$ 表示卷积核数量，λ 表示卷积核大小。计算过程定义如下，其中*表示卷积，Y 表示输出特征图，F 表示输入特征图，表达式为

$$Y^{H\times W\times N^2}=F^{H\times W\times C}*K^{N\times N\times M}. \tag{1}$$

3.4 损失函数

使用的损失函数为分类损失 ($L_{\rm cls}$)、掩码损失 ($L_{\rm mask}$) 和语义分割损失 ($L_{\rm se}$) 三种。总的损失函数公式由式 (2) 表示，其中参数 β_1=1、β_2=3、β_3=0.8。

$$L=\beta_1 L_{\rm cls}+\beta_2 L_{\rm mask}+\beta_3 L_{\rm se}. \tag{2}$$

采用 Focal Loss 损失函数，表达式为

$$L_{\rm cls}=-(1-P_i)^{\gamma}\cdot\log(P_i), \tag{3}$$

式中：P_i 为模型对正样本的预测概率；γ 为可调因子。采用 Dice Loss 损失函数，表达式为

$$L_{\rm mask}=1-\frac{2|X\cap Y|}{|X|+|Y|}, \tag{4}$$

式中：$|X\cap Y|$表示 X 和 Y 之间的交集；$|X|$和$|Y|$分别表示 X 和 Y 的元素个数。语义分割损失 $L_{\rm se}$ 采用二元交叉熵损失函数，表达式为

$$L_{\rm se}=-(y\cdot\log(p)+(1-y)\cdot\log(1-p)), \tag{5}$$

式中：y 为标签取值；p 为模型预测为正类的概率。

4 实验结果和分析

4.1 数据集介绍

本研究主要采用 COCO 2017 和 Cityscapes[26-27] 两个数据集进行实验验证。

COCO (Common objects in context) 数据集是计算机视觉领域广泛使用的基准数据集之一，为目标检测、实例分割等任务设计。COCO 2017 包含 118287 张训练图像、5000 张验证图像和约 40000 张测试图像，提供 80 个物体类别的高质量边界框和分割掩码标注。该数据集涵盖日常生活中的常见物体，具有样本数量大、类别丰富、场景复杂等特点，采用平均精度 (mAP) 和平均召回率 (mAR) 等指标进行模型性能评估，是实例分割领域的标准基准。

Cityscapes 数据集专注于城市街景图像分割，包含来自 50 个城市的高分辨率 (2048×1024) 图像。数据集中定义了 30 个类别，分为 8 个大类：平面、人类、车辆、建筑、物体、自然、天空和无效。其中精细标注部分包含 5000 张图像 (2975 张训练、500 张验证、1525 张测试)，粗略标注部分包含 20000 张图像，适合弱监督学习。

在本研究中，主要训练和测试选择了 COCO 2017

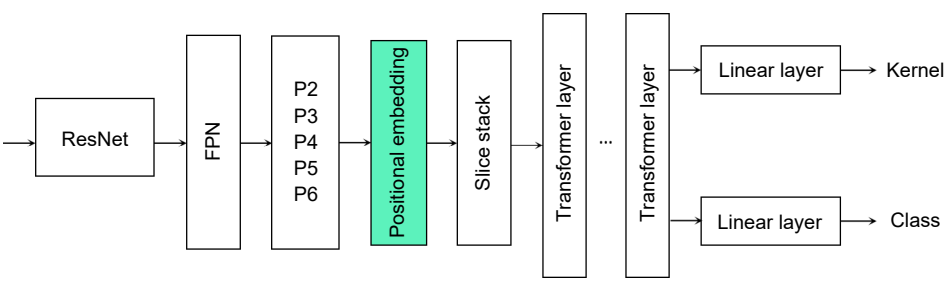


图 6 预测网络结构

Fig. 6 Predictive network structure

数据集，这是一个广泛应用于目标检测和实例分割任务的大规模数据集。为了快速验证模型的有效性并进行消融实验，选择 Cityscapes 数据集。Cityscapes 数据集规模较小，同时提供了高分辨率图像和精细的像素级标注，特别适合实例分割任务的快速实验和模型调试，除去探讨 CS 注意力模块位置的影响外，其他消融实验均采用 Cityscapes 数据集。

4.2 训练策略

所使用的软件和硬件的实验环境配置如表 1 所示。

训练和测试参数配置：使用 ResNet50 作为骨干网络，并使用 ResNet 官方训练模型进行预训练。训练批次大小为 16，采用 SGD 优化策略，初始学习率大小为 0.01，动量为 0.9，权重衰减为 0.0001。共训练 36 个周期，学习率分别在 27 和 33 周期处缩小为 1/10。

数据增强：为提高模型的鲁棒性和泛化能力，对数据进行以下预处理：1) 随机翻转，设置概率为 50%；2) 随机图像输入尺寸，与 YOLACTR 固定输入图像尺寸不同，本文实验设置多个图像尺寸，并进行随机选择。

4.3 消融实验

4.3.1 CS 注意力模块位置的影响

首先对特征网络中 CS 注意力模块的位置进行了探究。在特征聚合网络中，注意力机制可以作用于特征进行采样前或进行采样后，为验证特征网络中 CS 注意力模块不同位置对模型性能的影响，本节探究了在特征采样前后放置注意力机制的效果差异，测试结果如图 7 和图 8 所示。

蓝色曲线表示 CS 注意力关注采样前的精度测试结果，红色曲线表示 CS 注意力关注采样后的精度测试结果。比较结果显示，两种配置对模型最终性能影响相近，但红色曲线 (采样后) 略微平滑于蓝色曲线，收敛更为稳定。综合考虑性能与计算成本，本研究在后续模型训练中采用将 CS 注意力模块置于采样后的配置。

4.3.2 损失函数

损失函数设计对实例分割精度有显著影响。为验证分类损失 (focal loss) 和掩码损失 (dice loss) 对模型性能的独立贡献，设计了消融实验，通过分别替换或移除这些损失函数，观察模型性能的变化。设计了以下三组实验，实验结果如表 2 所示：

1) 替换掩码损失。将 dice loss 替换为 binary crossentropy loss。

2) 替换分类损失。将 focal loss 替换为 crossentropy loss。

3) 完整配置。保留 dice loss 和 focal loss 组合，作为对比基线。

训练过程中的损失变化曲线如图 9 所示，实验结

表 1 实验环境配置

Table 1 Experimental environment configuration

Operating system	Framework	CPU	GPU	Memory	Video memory	Python	CUDNN	CUDA
Ubuntu 20.04.3 LTS	Pytorch	AMD EPYC 7601	NVIDIA GeForce RTX 3090 × 2	32 GB	48 GB	3.8.10	8.0.5	11.0

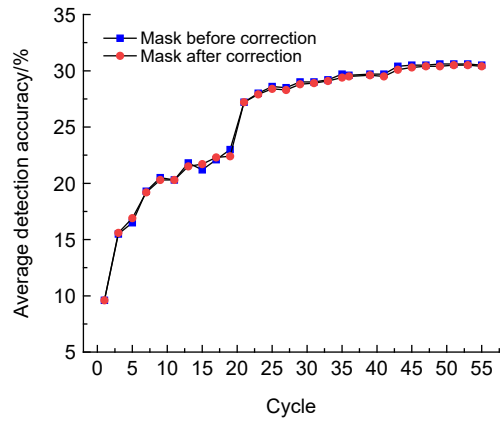


图 7 目标检测结果

Fig. 7 Object detection results

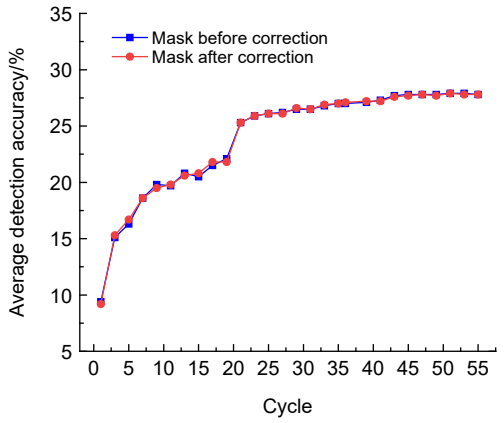


图 8 实例分割检测结果

Fig. 8 Instance segmentation detection results

表 2 不同损失函数配置下的分割结果

Table 2 Segmentation results under different loss function configurations

Loss function configuration	*AP*/%	AP_{50}/%	AP_{75}/%
Replace dice loss	3.0	4.8	3.6
Replace focal loss	11.5	23.8	10.2
Dice loss + focal loss	12.7	26.9	10.9

果清晰表明，完整损失函数配置 (dice loss + focal loss) 下模型性能最佳，总损失下降更快且在训练后期更加稳定。其中掩码损失 dice loss 对实例分割结果的影响尤为显著，替换后 *AP* 值下降了 9.7 个百分点 (从 12.7% 降至 3.0%)。这是因为 dice loss 直接优化分割结果的重叠区域 (IoU)，更适合处理像素级任务中前景与背景区域不平衡的问题。同时，focal loss 通过对难分类样本赋予更高权重，有效提升了模型对小目标和边界区域的分类性能。这一结果验证了所提损失函数设计的合理性和有效性。

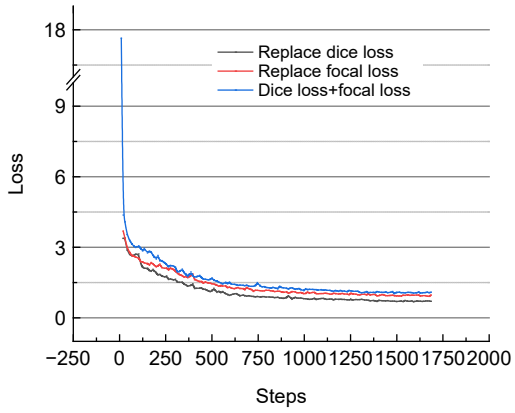


图 9 不同损失函数配置下的损失变化曲线

Fig. 9 Loss variation curves under different loss function configurations

表 3 不同 Transformer 层数的分割结果

Table 3 Segmentation results with different numbers of transformer layers

Transformer layers	*AP*/%	AP_{50}/%	AP_{75}/%	AP_S/%	AP_M/%	AP_L/%
2 layers	12.7	26.9	10.9	1.0	5.1	28.9
6 layers	14.1	29.3	12.4	2.0	6.6	34.7

4.3.3 Transformer 层数对实验结果的影响

Transformer 层数直接影响模型对全局特征的建模能力。为确定最优网络结构，分别训练了 2 层和 6 层 Transformer 的模型，并对比其在 Cityscapes 数据集上的性能表现。本实验在保持其他网络结构和训练参数一致的前提下，分别训练了 2 层和 6 层 Transformer 的模型。在实验中，Transformer 层数的变化仅影响预测网络的特征处理部分，掩码生成网络和其他模块保持不变。所有实验均在 Cityscapes 数据集上进行，训练 36 个 epoch，使用 ResNet50 作为骨干网络。实验结果如表 3 所示。

训练过程中的损失变化曲线和精度变化曲线如图 10 所示。

实验结果显示，6 层 Transformer 模型在所有评估指标上均优于 2 层模型，尤其在小目标检测和大目标检测方面改进显著。这表明增加 Transformer 层数能够增强模型对全局特征的建模能力，提高不同尺度目标的分割精度。从图 10 可以看出，6 层 Transformer 的模型在训练初期的损失下降速度略高于 2 层模型，在后期收敛更稳定，最终精度更高。但两者曲线下降程度接近，是因为增加层数后，模型的参数量增加，优化难度加大，但同时也增强了模型的表达能力。

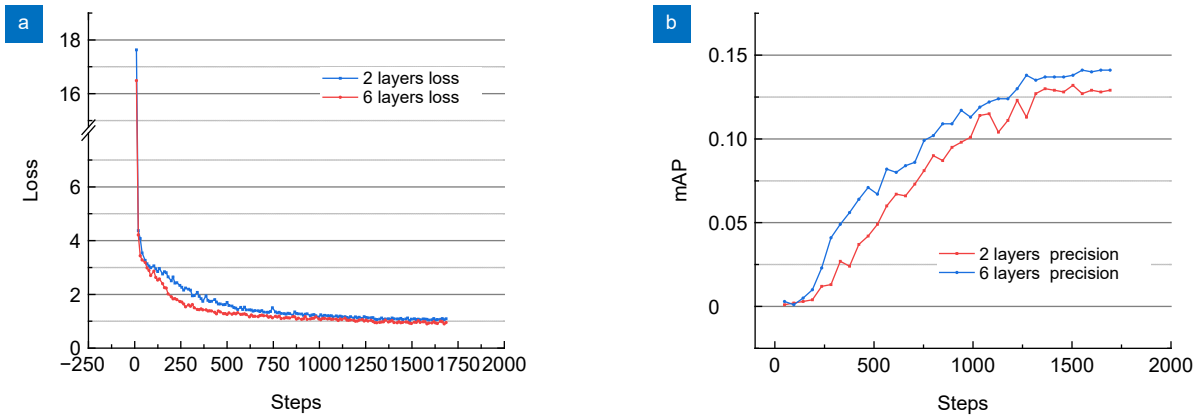


图 10 2 层和 6 层 Transformer 模型的不同曲线。(a) 损失变化曲线；(b) 精度变化曲线

Fig. 10 Different curves for 2- and 6-layer Transformer models. (a) Loss variation curves; (b)Accuracy variation curves

尽管增加 Transformer 层数会导致计算成本的增加，但从实验结果来看，6 层 Transformer 的性能提升显著，小目标检测和高精度分割任务中表现相较 2 层模型提升明显，同时考虑到设备硬件性能约束，再增加额外的层数 (8 层或 12 层) 会产生更大的计算开销，可能导致超过实验设备负载，基于性能和计算成本的权衡，后续选择 6 层 Transformer 用于构建最终模型。

4.4 实验结果

消融实验结果表明，focal loss 和 dice loss 的组合能够显著提升实例分割的精度；同时，增加 Transformer 层数 (从 2 层到 6 层) 能够有效增强模型的全局特征建模能力，提高高 IoU 阈值下的分割精度，在小目标检测中也能够有所提升。这些实验验证了改进算法在损失函数设计和网络结构优化方面的合理性。进一步在 MS COCO 数据集上对改进算法进行了全面的性能评估，在实验设备上，单个模型训练耗时约为 5 d。

构建的无锚框检测算法，不再采用边界框的检测方式，损失函数不再包含边界框损失，新的损失函数主要由分类损失 L_{cls} 和掩码损失 L_{mask} 组成。在训练过程中，损失变化如图 11 所示。

其中，曲线的横坐标是迭代数，纵坐标是损失值。每个 epoch 包含 7330 次迭代，每次迭代数据大小为 16 (总数据集数量=迭代总数×批次大小)。图 11 中损失函数从左到右依次是总损失曲线、类别损失曲线和掩码损失曲线。损失值在 27 个 epoch 处出现下降，这是由于学习率降低为 1/10。通过观察损失下降趋势，在 35~36 个 epoch 附近趋于平稳。

在观察损失下降趋势的同时，算法掩码精度曲线如图 12 所示。随着模型不断训练学习，损失值降低，模型掩码精度呈上升趋势，并在 27 个 epoch 处随学习率变化有较大提升幅度，可以看到在第 33 个 epoch 附近，随着学习率降低掩码精度提升已经不大，并在 35~36 个 epoch 处趋于平稳，因此在第 36 个 epoch 处停止训练。

根据训练结果进行测试，将所提算法与其他算法在 COCO 数据集上进行实例分割，测试结果如表 4 所示，由于测试设备和平台不同，所列部分模型数据

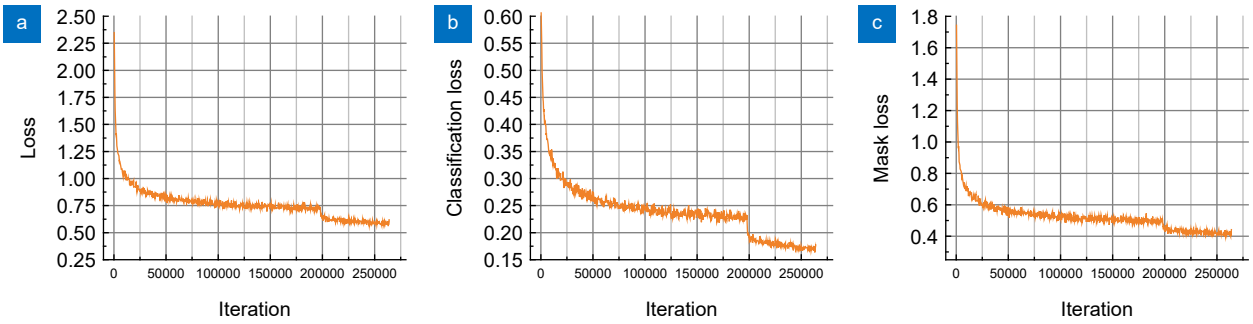


图 11 训练过程中各损失下降示意图。(a) 总损失曲线；(b) 分类损失曲线；(c) 掩码损失曲线

Fig. 11 Schematic diagrams of the decline of each loss during the training process. (a) Total loss curve; (b) Classification loss curve; (c) Mask loss curve

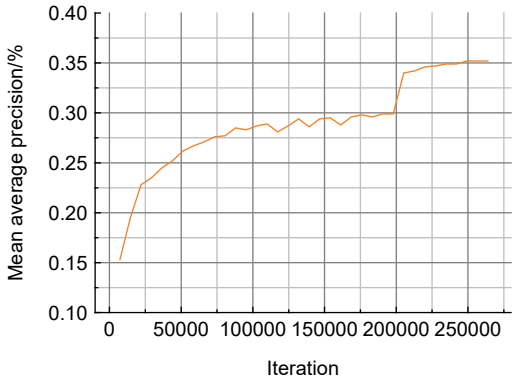


图 12 掩码检测精度上升曲线图

Fig. 12 Mask detection accuracy rise graph

可能和原文报告中的数据有所偏差。

根据表 4 可知，构建无锚框的实例分割检测算法，相对于 YOLACT 算法，掩码精度 (AP) 提升了 25.7%，小目标检测精度 (AP_S) 提升了 37.1%，中等目标检测精度 (AP_M) 提升了 25.8%，大目标检测精度 (AP_L) 提升了 21.9%；与改进算法 YOLACTR 相比，掩码检测精度进一步提升了 21.0%，小目标检测精度提升了 19.6%，其他方面提升也比较大。

这些提升主要得益于所提算法在结构上的创新设计，特别是在多尺度特征融合和关注机制的引入，能够有效地捕捉到不同尺度目标之间的关联，并增强了

表 4　在 COCO 数据集上的实例分割结果

Table 4　Instance segmentation results on the COCO dataset

Network model	*AP*/%	AP_{50}/%	AP_{75}/%	AP_S/%	AP_M/%	AP_L/%
YOLACT	28.0	46.2	29.1	8.9	30.2	47.0
Mask R-CNN	30.5	51.1	32.1	14.2	34.1	43.1
YOLACTR	29.1	48.7	30.0	10.2	31.4	46.8
PolarMask[28]	30.4	51.9	31.0	13.4	32.4	42.8
SOLO	33.1	53.5	35.0	12.2	36.1	50.8
QueryInst	37.5	58.7	40.5	18.4	40.2	57.2
Mask2Former	42.9	65.3	46.0	22.1	46.3	64.8
Proposed algorithm	**35.2**	**55.4**	**37.5**	**12.2**	**38.0**	**57.3**

模型对小目标的感知能力。此外，掩码损失的优化也对精度的提升起到了重要作用，尤其是在处理小目标时，通过改进的掩码损失函数可以更好地引导模型聚焦于目标的细节，减少误检和漏检现象。

相较于采用更复杂结构的 QueryInst 和 Mask2-Former 等先进模型，所提算法采用简洁的无锚框设计，避免了锚框选择和匹配过程中的计算复杂性，在实现显著性能提升的同时保持了较低的计算开销，在性能与复杂度平衡方面具有明显优势。此外，所提出的结构具有很好的算法移植性，易于与其他深度学习框架结合，为后续研究提供了灵活的拓展空间。

改进前后图像检测对比图如图 13 所示，改进后的算法在漏检、重叠处错检以及边缘部位的检测上有明显的进步。对比图中的结果可以看到，所提出的无锚框实例分割算法在细节处理上更加精细，特别是在边缘部分的精确度和实例分割的完整性上，较传统算法有了显著提升。这一改进使得检测结果在处理复杂背景或相似物体时，能够更加准确地分割出目标区域，避免了许多传统算法常见的边缘模糊或目标重叠识别错误的问题。

图 14 和图 15 展示了所提算法与其他算法在相同图像上的检测对比。所提算法能实现精准分割，轮廓

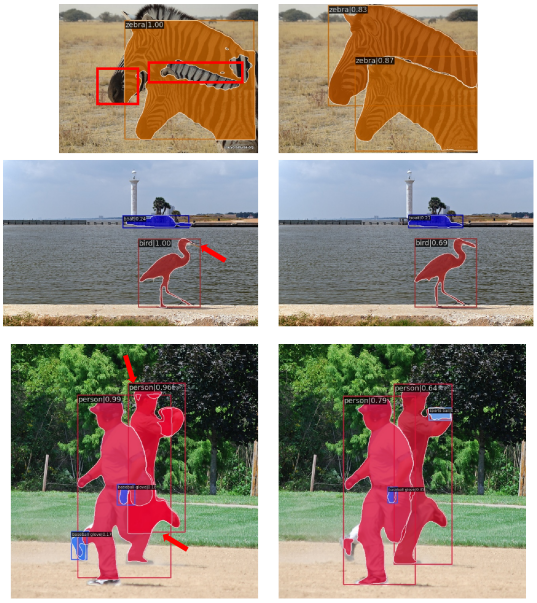


图 13　改进前 (左)、后 (右) 对比图

Fig. 13　Comparison diagrams before improvement (left) and after improvement (right)

图 14　实例分割结果对比图 1

Fig. 14　Comparison diagram of instance segmentation results 1

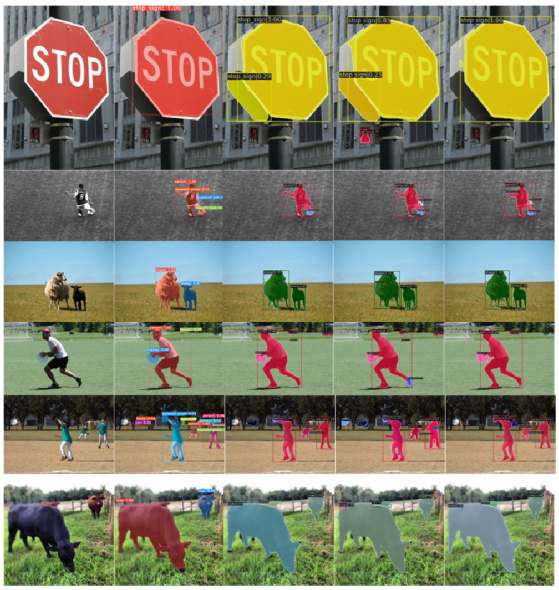


图 15 实例分割结果对比图 2

Fig. 15 Comparison diagram of instance segmentation results 2

清晰且区分度高。以图 14 中的飞机和手机为例，所提算法能更顺滑地勾画目标边缘；图 15 中的球体与人物分割也更为准确，腿部轮廓更贴合实际形态。整体而言，所提算法在分割精度上有显著提升，达到了预期目标。

5 总 结

基于改进的 YOLACT 算法网络结构，提出了无锚框检测的实例分割算法，有效解决了传统边界框检测在实例重叠区域难以区分的问题。该方法将掩码生成解耦成特征学习和卷积核学习，通过随机位置嵌入增强了特征的空间感知能力。在预测网络中，采用多层 Transformer 结构处理位置敏感特征，生成实例类别和动态卷积核；在特征聚合网络中，融合来自 Transformer 的高层语义特征与特征金字塔的多尺度特征，通过 CS 注意力模块优化特征表达，提高了掩码特征质量。最终通过动态卷积方式生成掩码，在边缘细节和小目标检测方面取得了明显进步。

通过消融实验验证了 focal loss 和 dice loss 损失函数组合的有效性，以及增加 Transformer 层数对模型性能的积极影响。所提方法在 MS COCO 公共数据集上的实验结果表明，无锚框检测的实例分割算法成功解决了传统方法中的重叠遮挡问题，显著增强了小目标检测和边缘细节保留能力。与主流实例分割算法相比，所提方法在保持较低计算复杂度的同时，实现了更高的检测精度，特别是在处理具有复杂形状和重叠区域的实例时表现出色。未来研究工作可以通过模型量化和轻量化技术减少计算量，提升推理速度，同时探索无锚框架构与其他实例分割任务的结合，提高算法的通用性和实用价值。

利益冲突: 所有作者声明无利益冲突

参考文献

[1] Zhou T, Zhao Y N, Lu H L, et al. Medical image instance segmentation: from candidate region to no candidate region[J]. *J Biomed Eng*, 2022, **39**(6): 1218−1232.
周涛, 赵雅楠, 陆惠玲, 等. 医学图像实例分割：从有候选区域向无候选区域[J]. 生物医学工程学杂志, 2022, **39**(6): 1218−1232

[2] Pei S W, Ni B, Shen T M, et al. RISAT: real-time instance segmentation with adversarial training[J]. *Multimed Tools Appl*, 2023, **82**(3): 4063−4080.

[3] Hong S L, Jiang Z H, Liu L Z, et al. Improved mask R-CNN combined with Otsu preprocessing for rice panicle detection and segmentation[J]. *Appl Sci*, 2022, **12**(22): 11701.

[4] Wu M J, Zhang Y A, Lin S L, et al. Real-time semantic segmentation algorithm based on BiLevelNet[J]. *Opto-Electron Eng*, 2024, **51**(5): 240030.
吴马靖, 张永爱, 林珊玲, 等. 基于 BiLevelNet 的实时语义分割算法[J]. 光电工程, 2024, **51**(5): 240030.

[5] Su L, Sun Y X, Yuan S Z. A survey of instance segmentation research based on deep learning[J]. *CAAI Trans Intell Syst*, 2021, **17**(1): 16−31.
苏丽, 孙雨鑫, 苑守正. 基于深度学习的实例分割研究综述[J]. 智能系统学报, 2021, **17**(1): 16−31.

[6] Zhang J K, Zhao J, Zhang R, et al. Survey of image instance segmentation methods using deep learning[J]. *J Chin Comput Syst*, 2021, **42**(1): 161−171.
张继凯, 赵君, 张然, 等. 深度学习的图像实例分割方法综述[J]. 小型微型计算机系统, 2021, **42**(1): 161−171.

[7] Minaee S, Boykov Y, Porikli F, et al. Image segmentation using deep learning: a survey[J]. *IEEE Trans Pattern Anal Mach Intell*, 2022, **44**(7): 3523−3542.

[8] He K M, Gkioxari G, Dollár P, et al. Mask R-CNN[C]//*Proceedings of the 2017 IEEE International Conference on Computer Vision*, 2017: 2980–2988. https://doi.org/10.1109/ICCV.2017.322.

[9] Ren S Q, He K M, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]//*Proceedings of the 29th International Conference on Neural Information Processing Systems*, 2015: 91–99.

[10] Cai Z W, Vasconcelos N. Cascade R-CNN: delving into high quality object detection[C]//*Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 6154–6162. https://doi.org/10.1109/CVPR.2018.00644.

[11] Xiao Z J, Tian H, Zhang J H, et al. Fusion of dynamic features enhances remote sensing building segmentation[J]. *Opto-Electron Eng*, 2020, **52**(3): 240231
肖振久, 田昊, 张杰浩, 等. 融合动态特征增强的遥感建筑物分割

[J]. 光电工程, 2020, **52**(3): 240231

[12] Chen K, Pang J M, Wang J Q, et al. Hybrid task cascade for instance segmentation[C]//*Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019: 4969–4978. https://doi.org/10.1109/CVPR.2019.00511.

[13] Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection[C]//*Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 779–788. https://doi.org/10.1109/CVPR.2016.91.

[14] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//*Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014: 580–587. https://doi.org/10.1109/CVPR.2014.81.

[15] Tian Z, Shen C H, Chen H, et al. FCOS: fully convolutional one-stage object detection[C]//*Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision*, 2019: 9626–9635. https://doi.org/10.1109/ICCV.2019.00972.

[16] Zhou X Y, Wang D Q, Krähenbühl P. Objects as points[Z]. arXiv: 1904.07850, 2019. https://arxiv.org/abs/1904.07850.

[17] Wang X L, Kong T, Shen C H, et al. SOLO: segmenting objects by locations[C]//*16th European Conference on Computer Vision*, 2020: 649–665. https://doi.org/10.1007/978-3-030-58523-5_38.

[18] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//*Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 770–778. https://doi.org/10.1109/CVPR.2016.90.

[19] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//*Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 936–944. https://doi.org/10.1109/CVPR.2017.106.

[20] Liu T, Liu H Z, Li X W, et al. Improved instance segmentation method based on anchor-free segmentation network[J]. *Comput Eng*, 2022, **48**(9): 239−247,253.
刘腾, 刘宏哲, 李学伟, 等. 基于无锚框分割网络改进的实例分割方法[J]. 计算机工程, 2022, **48**(9): 239−247,253.

[21] Kirillov A, Wu Y X, He K M, et al. PointRend: image segmentation as rendering[C]//*Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 9796–9805. https://doi.org/10.1109/CVPR42600.2020.00982.

[22] Yang S S, Wang X G, Li Y, et al. Temporally efficient vision transformer for video instance segmentation[C]//*Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022: 2875–2885. https://doi.org/10.1109/CVPR52688.2022.00290.

[23] Cheng B W, Misra I, Schwing A G, et al. Masked-attention mask transformer for universal image segmentation[C]//*Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022: 1280–1289. https://doi.org/10.1109/CVPR52688.2022.00135.

[24] Bolya D, Zhou C, Xiao F Y, et al. YOLACT: real-time instance segmentation[C]//*Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision*, 2019: 9156–9165. https://doi.org/10.1109/ICCV.2019.00925.

[25] Zhao J W, Lin S L, Mei T, et al. Research on instance segmentation algorithm based on YOLACT and Transformer[J]. *Semicond Optoelectron*, 2023, **44**(1): 134−140.
赵敬伟, 林珊玲, 梅婷, 等. 基于 YOLACT 与 Transformer 相结合的实例分割算法研究[J]. 半导体光电, 2023, **44**(1): 134−140.

[26] Cordts M, Omran M, Ramos S, et al. The cityscapes dataset for semantic urban scene understanding[C]//*Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 3213–3223. https://doi.org/10.1109/CVPR.2016.350.

[27] Cordts M, Omran M, Ramos S, et al. The cityscapes dataset[C]//*CVPR Workshop on the Future of Datasets in Vision*, 2015: 1. https://doi.org/10.48550/arXiv.1604.01685

[28] Xie E Z, Sun P Z, Song X G, et al. PolarMask: single shot instance segmentation with polar representation[C]//*Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 12190–12199. https://doi.org/10.1109/CVPR42600.2020.01221.

作者简介

梅婷 (1999-)，女，2020 年毕业于中国福州大学，获学士学位，现攻读电子电路与系统博士学位。目前的研究工作集中在电润湿显示和图像处理的驱动系统。

E-mail: 1004070233@qq.com

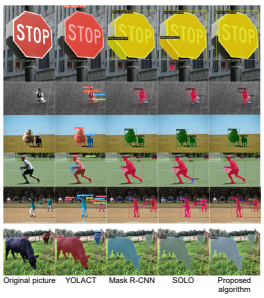
【通信作者】林珊玲 (1991-)，女，2020 年毕业于中国福州大学电子电路与系统专业，博士，现就职于福州大学先进制造学院。主要研究方向为信息显示技术，包括电润湿显示驱动系统、微纳米 LED、图像处理及其他光学信息技术。

E-mail: 526176333@qq.com

扫描二维码，获取PDF全文

Anchor-free instance segmentation algorithm based on YOLACTR

Mei Ting[1,2], Zhao Jingwei[1,2], Lin Shanling[2,3*], Xie Ziyu[1,2], Lin Zhixian[1,2,3], Guo Tailiang[1,2]



Comparison of instance segmentation results 2

Overview: This paper proposes an anchor-free instance segmentation algorithm based on YOLACTR to address the limitations of the single-stage YOLACT algorithm in instance segmentation tasks. Traditional YOLACT algorithms rely on bounding box detection, suffering from precise localization of regions of interest and facing difficulties in distinguishing overlapping instances, which constrains detection accuracy. This research decouples the mask generation process into parallel tasks of feature learning and convolution kernel learning, abandoning traditional bounding box detection methods and adopting a more natural mask representation approach.

In the algorithmic implementation, random positional embedding techniques are employed to enhance the position sensitivity of feature maps, utilizing a six-layer Transformer structure to process spatial information, simultaneously generating dynamic convolution kernels and category information. The feature aggregation network integrates bottom-layer features from the feature pyramid and high-level features from the prediction network, optimizing feature expression capabilities through channel-spatial (CS) attention modules. For the loss function design, the research implements a combination of focal loss for classification tasks and dice loss for mask generation.

The network architecture consists of four primary components: a multi-scale feature generation network utilizing ResNet and feature pyramid networks; A mask generation network combining transformer with feature aggregation; A prediction network incorporating positional information to generate dynamic convolution kernels; Auxiliary network structures to enhance overall performance. This design allows for more effective handling of spatial relationships and instance boundaries compared to traditional anchor-based approaches.

Experimental results on the MS COCO dataset demonstrate that this method achieves a mask accuracy (*AP*) of 35.2%, representing a 25.7% improvement over the YOLACT algorithm. Specifically, the detection accuracy for small targets is improved by 37.1%, for medium targets by 25.8%, and for large target by 21.9%. When compared to algorithms such as Mask R-CNN, YOLACTR, and SOLO, this method shows advantages in segmentation accuracy and edge detail preservation. It performs exceptionally well in handling overlapping objects and small target detection, effectively addressing the segmentation issues in instance boundary overlap regions faced by traditional methods.

This paper effectively overcomes the limitations of traditional bounding box methods by decoupling the mask generation process and introducing anchor-free design, achieving balanced performance in instance segmentation tasks across different scales of objects, particularly improving small target detection capability and boundary differentiation of overlapping objects.

Mei T, Zhao J W, Lin S L, et al. Anchor-free instance segmentation algorithm based on YOLACTR[J]. *Opto-Electron Eng*, 2025, **52**(5): 240265; DOI: 10.12086/oee.2025.240265

Foundation item: National Key Research and Development Program (2021YFB3600603), National Youth Science Foundation (62101132)

[1]College of Physics and Information Engineering, Fuzhou University, Fuzhou, Fujian 350116, China; [2]Fujian Science & Technology Innovation Laboratory for Optoelectronic Information of China, Fuzhou, Fujian 350108, China; [3]School of Advanced Manufacturing, Fuzhou University, Quanzhou, Fujian 362200, China

* E-mail: 526176333@qq.com