DOI:10.19651/j.cnki.emt.2212430

无人化起重装卸的目标物实例分割模型研究*

王国桢[1] 卢国杰[2] 王桂棠[3]
(1.广州吉欧电子科技有限公司 广州 510530;2.广州黑格智造信息科技有限公司 广州 510530;
3.广东工业大学机电工程学院 广州 510006)

摘 要:不确定目标物自动识别是研发无人化智能起重装卸系统的关键,目前有效的技术是基于深度学习的实例分割。设计了一个融合 CNN 和 Transformer 的异构特征信息的模块,以解决当前实例分割主干网络存在的提取图像全局上下文特征信息的能力有限、卷积算子难以对感受野的长程相关性进行建模、以及识别纹理特征单一目标时缺乏足够的深度线索等问题。通过利用 Transformer 建模全局依赖关系,并与 CNN 提取局部信息的能力相融合;然后通过引入 Dense RepPoints 检测网络构建了针对不确定目标物的实例分割网络,实现准确分割且能分割其不同表面。应用实验结果表明本方法具有达到很好的实例分割效果,AP 达到 98.82%、mIoU 达到 91.89%,分别比目前同类的研究成果提升了 4.95%和 5.42%。

关键词:不确定目标物;无人化装卸;深度学习;实例分割
中图分类号:TP391.4 **文献标识码**:A **国家标准学科分类代码**:520.6040

Instance segmentation model of uncertain object in unmanned lifting and handling scenarios

Wang Guozhen[1] Lu Guojie[2] Wang Guitang[3]
(1. Guangzhou Geoelectron Technology Co., Ltd., Guangzhou 510530, China;
2. Guangzhou Geoelectron Technology Co., Ltd., Guangzhou 510530, China;
3. School of Electromechanical Engineering, Guangdong University of Technology, Guangzhou 510006, China)

Abstract: Uncertain object auto detection is the key technology of unmanned intelligent lifting handling, and efficient technology recently used is Instance segmentation model based on deep learning. Due to the limited ability of the existing cases to segment the trunk network to extract the global context feature information of lifting scene images, and the difficulty of the convolutional operators in the convolutional neural network-based trunk network to model the long range correlation of the receptive field, and the lack of sufficient depth cues when identifying single targets with texture features, a module is designed to integrate the heterogeneous feature information of CNN and Transformer, and Transformer is used to model the global dependency relationship, and it is integrated with the ability of CNN to extract local information. Then, the Dense RepPoints detection network was introduced to construct the case segmentation network for the complex lifting and loading scenarios, which could accurately segment the loading and unloading objects and different surfaces of the objects. Compared with the most advanced method at present, AP increased by 4.95% to 98.82%, mIoU increased by 5.42% to 91.89%, obtaining a good example segmentation effect, solving the key technical problems of intelligent lifting loading and unloading, thus improving the work efficiency and safety of unmanned lifting loading and unloading logistics transportation, and reducing costs.

Keywords: uncertain object;unmanned lifting handling;deep learning;instance segmentation

0 引 言

传统的货物起重装卸是依靠起重机司机观察识别目标物并操控起重机进行操作。近年起重装卸智能化已实现了预知场景的自动起重装卸操作,提高了物流运输的工作效率和安全性。在不确定场景下实现全自动起重装卸,主要

收稿日期:2022-12-19
* 基金项目:佛山市 2021 年高校教师特色创新研究项目(2021DZXX15)资助

技术瓶颈是对目标货物自动识别检测。目前机器视觉是对目标物进行识别及测量的主要技术[1-4]，但在起重装卸场景出现预先不确定的目标货物、装卸车辆时，该技术应用陷入巨大的困境，其中主要是实例分割的难以精确实现。本文在开发无人化起重装卸智能测控系统项目中，研究了基于深度学习的机器视觉实例分割技术以解决不确定类型和状态的装卸目标物的自动识别检测难题。

实例分割已经成为机器视觉研究中最重要、复杂和具有挑战性的领域之一[5]。实例分割旨在预测对象类别标签和特定像素的对象实例掩码，定位各种图像中存在的不同类别的对象实例。通常实例分割要面对多个物体重叠等复杂背景，这也是其一直是机器视觉的挑战性任务的原因之一。随着深度学习的出现，目前出现了各种实例分割框架，其分割精度迅速提高[6]，并在针对性的场景中得到应用。

常见的深度卷积主干架构[7-10]在图像分类、对象检测、实例分割方面取得了重大进展，在深度学习的各个领域都占据了主导地位[11-12]。大多数主干架构使用多层 3×3 卷积，可以有效地捕获局部信息，但由于货物装卸环境的复杂性，图像全局上下文特征信息的提取能力受限，无法高效编码复杂场景语义信息。

在卷积神经网络（convolutional neural network，CNN）方面，人们已经做出了很大的努力来获取全局上下文信息。如全卷积网络（fully convolutional networks，FCNs）[13]，空洞空间金字塔池（atrous spatial pyramid pooling，ASPP）模块[14]，U-Net[15]、金字塔池化模型（pyramid scene parseing network，PSPNet）[16]等，为了全局聚合局部信息，需要堆叠多个层和分层采样[17]，因此运算效率降低，且利用全局上下文信息的建模能力一般。尽管这些方法确实提高了这些主干网络的性能，但一种用于建模全局（非局部）依赖关系的显式机制可能是一种更强大且可扩展的解决方案。

一些研究人员尝试使用自注意力机制（self-attention）[18]来解决缺乏模型感受野的问题。Fu 等[19]设计了基于自注意力机制的紧凑位置注意模块和紧凑通道注意模块，分别从空间和通道维度对语义相关性进行建模。后来，出现了 Transformer 模型代替深度卷积主干架构。Transformer 模型也采用了编码器-解码器架构，更擅长建模图像全局感受野的长期相关性。因此，使用视觉自注意力模型（vision transformer，ViT）[20]或 Swin Transformer[21]模型替代 CNN 作为主干网络在复杂装卸场景下进行实例分割时发现，Transformer 编码器无法产生令人满意的性能。有研究表明，充分利用远程相关性（即对象之间的距离关系）和局部信息（即同一对象的一致性）是有效获取深度线索的关键能力[22]。

最近，一些研究者提出了 CNN 与 Transformer 的融合模型。CSWin Transformer[23]使用局部增强位置编码（locally-enhanced positional enciding，LePE）来更好地处理局部位置信息，主要解决了全局注意力和局部自注意力的计算成本高导致的 token（令牌）之间的交互域限制问题。DPT[24]采用卷积解码器将不同的图像表示逐渐组合成全分辨率预测。与全卷积神经网络相比，DPT 通过高分辨率处理表示全局感受野的特征，可以提供更细粒度和全局一致的预测。BoTNet（bottleneck transformers for visual recognition[25]）将 CNN 与自我注意相结合，主要采用全局自注意力来替换深度残差网络（deep residual network，ResNet）最后 3 个瓶颈块中的空间 3×3 卷积。Crossformer[26]是提出的跨尺度注意力，用于建立图像中具有较大尺寸差异的对象之间的关系。它可以建立不同尺度特征之间的相互作用。它的核心包括一个跨尺度嵌入层（cross-scale embedding layer，CEL）和长短距离注意力（long short distance attention，LSDA）。

以上方法仅是使用 Transformer 模块简单地替换 CNN 主干网络的卷积模块，而本文拟探索 CNN 分支用于提取局部位置信息，Swin Transformer 分支用于提取全局上下文信息的两分支结构，然后设计融合模块使异构特征信息充分融合，从而组成主干网络。该主干网络通过有效整合远程相关性和局部位置信息，提供全局语义线索和深度线索，在复杂装卸场景下完成强健的语义特征信息和空间特征信息提取。通过引入密集代表点（dense representative points，Dense RepPoints）[27]作为实例分割的检测网络，利用特征信息回归代表点，输入分割掩膜，完成复杂装卸场景目标物的识别。最终通过实验验证方法的有效性。

1　不确定目标物的实例分割主干网络

实例分割网络由主干网络和检测网络组成。主干网络主要负责提取图像的特征信息，检测网络负责学习特征信息并输出目标的掩码。本文主要对主干网络进行了研究。现有主干网络存在 3 个问题，限制了其在起重装卸时分割目标物的有效应用。1）传统的实例分割方法基于手工构建的特征，此类方法的性能依赖于手工设计特征的可靠性及参数选择的合理性。但对于 24 h 室外作业的起重装卸场景，由于目标物外观灰度值巨大差异变化、目标物上不定形的阴影等因素影响，泛化能力差。2）起重装卸场所存在照明不均匀、场景范围大、干扰信息多等问题，基于卷积神经网络的主干网络中的卷积算子难以对感受野的长程相关性进行建模，因此提取起重装卸场景图像全局上下文特征信息的能力有限，无法高效编码不确定场景语义信息，表达能力有待提高。3）本文实例分割的目标不仅要分割出实例，还要在实例上再次分割出不同的表面，如圆桶和方箱的上表面和侧面，但由于其纹理特征单一且存在重叠，现有 Transformer 方法缺乏有效的深度线索编码器，导致分割不同表面准确率较低。尽管深度学习技术总体而言对于目标识别中有明显的优势，但具体应用中由于卷积算子的性

质限制,难以提取复杂场景图像的全局上下文特征信息,缺乏复杂场景图像的空间提取能力。因此,本文着重改进实例分割主干网络,然后引入一种检测网络,构建完整实例分割网络对装卸场景的不确定目标物进行识别。

1.1 CNN 分支构建

CNN 分支使用一个标准的 ResNet 编码器来提取局部信息,这是提取局部特征,挖掘特征深度线索常用的方法。人们在尝试将视觉几何组(visual geometry group, VGG)加深时发现,随着模型的层数越来越多,训练误差往往不降反升、难以收敛,不仅不能提升性能,反而出现了严重的退化问题。ResNet 通过短路机制实现了残差单元进行残差学习,以减轻性能退化。其设计的一个重要原则是当特征图大小降低一半时,特征图的数量增加一倍,这保持了网络层的复杂度。典型的 34 层 ResNet 结构是一个大型(7×7)卷积滤波器,后跟多个残差块,最后是一个全连接层。因此其网络本质上是由残差块的堆叠而成。残差块的结构如图 1 所示,其由 3×3 卷积、批归一化、ReLU 激活函数和残差连接组成。显然,其结构可以分为两部分,其中一部分是将输入 x_l 做残差恒等映射,一部分是将 x_l 与残差恒等映射 $F(x_l,W_l)$ 的结果相加,残差块最终输出 x_{l+1} 至下一层,运算过程可由式(1)所示。

$$x_{l+1}=x_l+F(x_l,W_l) \tag{1}$$

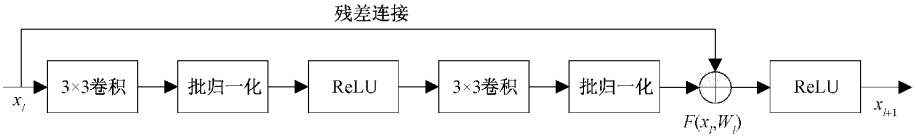


图 1 残差块的结构

1.2 Swin Transformer 分支构建

Swin Transformer(shifted windows transformer)是一个通用的 Transformer 主干。其是为了解决 ViT 架构产生的单一分辨率的特征图和其具有图像大小成二次方的复杂度等不足而提出的,同时实现了 Transformer 使用途径的扩展,从而可以在高分辨率的图像上进行像素级别密集任务的预测。本文利用 Swin Transformer 的先进性能,为网络提高语义特征提取的能力。

如图 2 所示,Swin Transformer 模块的具体实现是通过将 Transformer 模块中的标准多头自注意力(multi-head self attention, MSA)模块替换为基于移动窗口(shifted windows, Swin)的 Transformer 模块而构建的,其他层保持不变。一个 Swin Transformer 模块由一个基于移动窗口的 MSA(SW-MSA)模块组成,然后是一个 2 层 MLP,中间连接具有 GELU 非线性。在 W-MSA、SW-MSA 模块和每个 MLP 之前应用归一化层(layer norm, LN),在每个模块之后均使用残差连接。

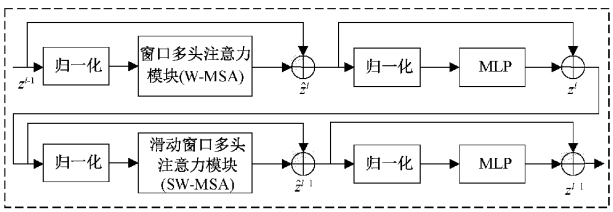


图 2 Swin Transformer 模块结构

1.3 融合模块设计

为了有效地融合来自 CNN 和 Swin Transformer 的编码特征,设计了一个融合模块,其结构如图 3 所示。

首先,将 CNN 分支第 i 层输出的特征矩阵 $x_{CNN,i}$ 和第 $i-1$ 阶段融合的输出特征矩阵 $x_{fusion,i-1}$ 合并(Concat)、插值(Interpolate)、3×3 卷积(Conv)和 ReLU 非线性激活:

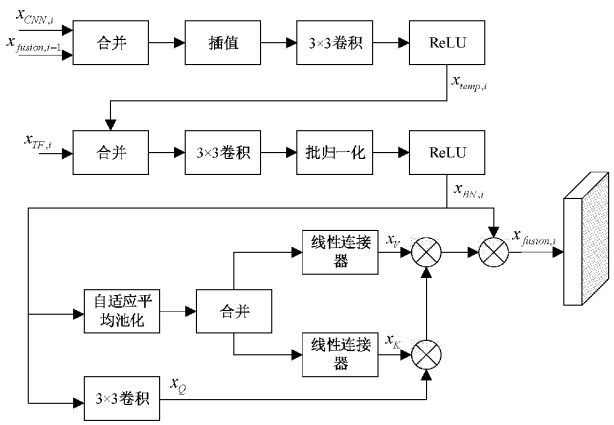


图 3 融合模块结构

$$x_{temp,i}=\mathrm{ReLU}(\mathrm{Conv}(\mathrm{Interpolate}(\mathrm{Concat}(x_{fusion,i-1},X_{CNN,i})))) \tag{2}$$

初步融合后得到中间特征 $x_{temp,i}$,再将其与第 i 层输出的移动窗口 Transformer 分支特征矩阵 $x_{Tf,i}$ 融合,得到归一化后的特征信息 $x_{BN,i}$:

$$x_{BN,i}=\mathrm{ReLU}(\mathrm{BN}(\mathrm{Conv}(Concat(x_{temp,i},X_{Tf,i})))) \tag{3}$$

由于使用多分支结构,特征融合是计算密集型的。为了缓解这个问题,将自适应平均池化(AdativeAvgPool)方法添加到后续融合中,构建每个像素和收敛中心之间的关系:

$$x_K=\mathrm{Linear}(\mathrm{Concat}(\mathrm{AdaptiveAvgPool}(x_{BN,i}))) \tag{4}$$

$$x_V=\mathrm{Linear}(\mathrm{Concat}(\mathrm{AdaptiveAvgPool}(x_{BN,i}))) \tag{5}$$

$$x_Q=\mathrm{Conv}(x_{BN,i}) \tag{6}$$

$$x_{fusion,i}=x(x_Q\otimes x_K)\otimes x_V\otimes x_{BN,i} \tag{7}$$

式中:x_K、x_V、x_Q 分别表示自注意力机制计算的键(Key)、值(Value)、查询(Query)。$x_Q\otimes x_K$ 得到自注意力权重矩

阵，$(x_Q \otimes x_K) \otimes x_V$ 得到自注意力加权特征矩阵，将其和 $x_{BN,i}$ 融合相加得到融合特征矩阵 $x_{fusion,i}$。$x_{fusion,i}$ 中每个位置的特征值是所有位置的特征、局部特征和原始特征的加权和。这个过程实现了分支之间的相似性建模和特征交互。因此，融合模块能够基于全局视图的注意特征图有选择性地聚合上下文信息，并聚合局部信息，弥补 Transformer 分支中空间归纳偏置的不足，从而有效编码远程相关性和深度线索。

1.4　主干网络构建

本文构建的主干网络如图 4 所示。Transformer 分支首先通过 ViT 图块(patch)分割模块将输入的 RGB 图像分割成不重叠的图块。每个图块均被视为一个 token，其特征被设置为原视像素 RGB 值的串联。图块的大小为 4×4，因此每个图块的特征维度为 4×4×3=48。应用这个原始特征值在线性编码层，以将其投影到任意维度，记为 C (C=48)。然后在这些图块上应用几个 Swin Transformer 块。Swin Transformer 块保持图块的数量($H/4 \times W/4$)，并与线性编码层一起被称为“第 1 阶段”。为了产生分层表示，随着网络变得更深，通过图块合并层减少图块的数量。第一个图块合并层连接每组 2×2 相邻图块的特征，并在 4C 维连接特征上应用线性层。这个操作将图块数量减少了 2×2=4 的倍数，即等同实现了 2 倍分辨率下采样，并且输出维度设置为 2C。之后应用 Swin Transformer 块进行特征变换，分辨率保持在 $H/8 \times W/8$。第一个图块合并块和特征转换被称为“第 2 阶段”。该过程重复两次，分别为“第 3 阶段”和“第 4 阶段”，输出分辨率分别为 $H/16 \times W/16$ 和 $H/32 \times W/32$，这些阶段共同产生分层表示。

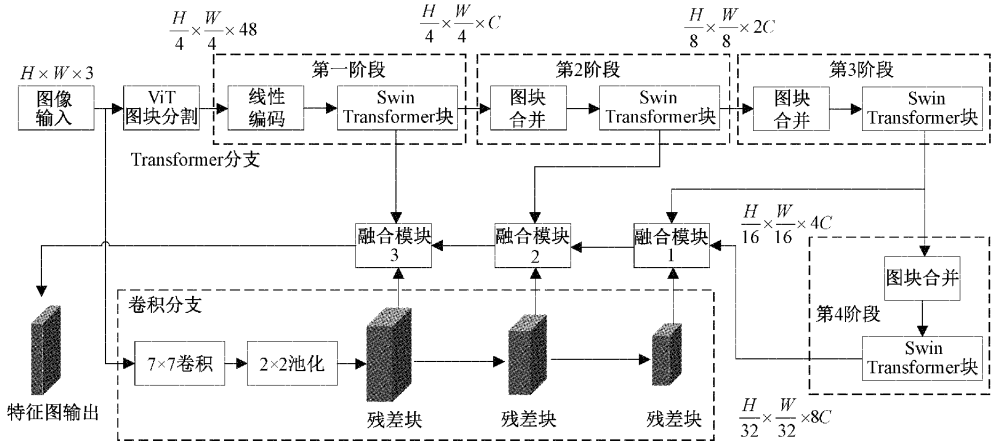


图 4　本文构建的主干网络结构

卷积分支接受图像输入，首先经过一个大型的 7×7 卷积，2×2 池化后输入到 ResNet 残差块，残差块 2 倍分辨率下采样，最终得到各残差块输出。

由于多阶段 Swin Transformer 块产生的分层表示，和 ResNet 输出特征图具有相同的分辨率。因此，容易将多阶段 Swin Transformer 块输出与 ResNet 残差块输出一起输入到各自的融合模块，最终输出特征图。

2　不确定目标物的实例分割检测网络

上一章构建了主干网络，本节引入 Dense RepPoints 模型作为检测网络，最终组成完整的实例分割网络。

2.1　代表点表示

代表点(RepPoints)是一组点，它们学习以一种限制对象空间范围并指示语义上重要的局部区域的方式自适应地将自己定位在对象上，最初用于替代对象矩形框获得更精确的对象表示。RepPoints 的训练是由目标定位和识别目标共同驱动的，这样 RepPoints 就会被真实边界框紧密地绑定，并引导检测器进行正确的目标分类。这种自适应和可微分表示可以在现代目标检测器的不同阶段连贯使用，并且不需要使用锚点在边界框空间上进行采样。

与目标检测不同，实例分割等细粒度几何定位任务通常提供需要精确估计的像素级标注。因此，少量点的表示能力不足，需要更大的点集以及与每个代表点关联的属性向量，最终使用一组自适应代表点 R 用于表示一个对象：

$$R = \{(x_i + \Delta x_i, y_i + \Delta y, a_i)\}_{i=1}^{n} \tag{8}$$

式中：$(x_i + \Delta x_i, y_i + \Delta y_i)$ 是第 i 个代表点；x_i 和 y_i 表示初始化位置；Δx_i 和 Δy_i 是可学习的偏移量；n 是点的数量；a_i 是与第 i 个点关联的属性向量。

2.2　局部特征轮廓细化

预测点的偏移量需要使用单个点的信息来细化点位置，不能直接应用矩形框分类中使用的分组特征，因此采取 Curve-GCN[28] 的方法，即使用局部特征进行轮廓细化。为了在点之间共享特征计算，首先基于图像特征图计算 n 个共享偏移场图。然后对于第 i 个代表点，在第 i 个偏移场的对应位置通过双线性插值直接预测其位置，如图 5 所示。

2.3　代表点采样监督

为了识别点集的类别，通过双线性插值从特征图 F 中

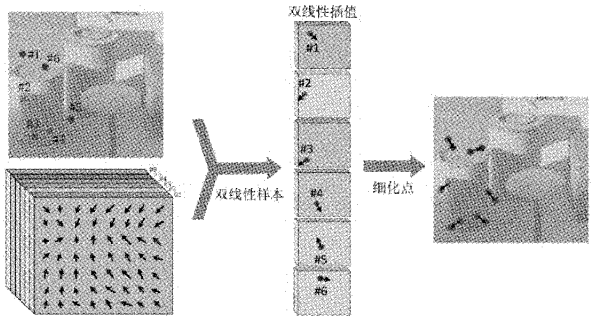


图 5 局部特征轮廓细化方法

提取点 $F(p_i)$的特征,将一个点集 $F(R)$的特征定义为 R 的所有代表点的串联:

$$F(R) = \text{concat}(F(p_1),\cdots,F(p_n)) \tag{9}$$

点集的边界框可以通过转换函数得到。在训练阶段,不需要对代表点进行明确的监督和注释。相反,代表点被框分类损失 L_{cls}^{b} 和框定位损失 L_{loc}^{b} 驱动移动到适当的位置:

$$L = L_{cls}^{b} + L_{loc}^{b} \tag{10}$$

这些代表点适用于同时表示对象类别和准确位置。

在实例分割中,属性可以是一个标量,定义为每个点的前景分数。除了框级分类 L_{cls}^{b} 和定位损失和 L_{loc}^{b} 之外,还引入了点级分类损失 L_{cls}^{p} 和点级定位损失 L_{loc}^{p}:

$$L = L_{cls}^{b} + L_{loc}^{b} + L_{cls}^{p} + L_{loc}^{p} \tag{11}$$

点分类损失 L_{cls}^{p} 和点定位损失 L_{loc}^{p} 用于在训练期间监督不同的片段表示。L_{cls}^{p} 被定义为具有 softmax 激活的标准交叉熵损失函数,其中位于前景中的点被标记为正,否则其标签为负。对于定位监督,可以采用点对点方法,其中每个 ground truth 点都被分配一个精确的几何意义,例如使用 PolarMask[29] 中的极坐标分配方法。每个具有精确几何意义的 ground truth 点也对应密集 RepPoints 中一个固定的索引代表点,使用 $L2$ 距离作为点定位损失 L_{loc}^{p}:

$$L_{point}(R,R') = \frac{1}{n}\sum_{k-1}^{n} \| ((x_i,y_i) - (x_i^i,y_i')) \|_2 \tag{12}$$

式中:$(x_i,\ y_i)\in R$ 和$(x_i',\ y_i')\in R'$分别表示预测点集和真实点集中的点。

然而,为每个点分配精确的几何意义是困难的,并且对于实例分割可能在语义上不准确。因此,需使用点集对点集的监督,而不是监督每个单独的点。点定位损失由监督点集和学习点集之间的倒角距离测量[30-31]:

$$L_{set}(R,R') = \frac{1}{2n}\sum_{i-1}^{n}\min_{j} \| ((x_i,y_i) - (x_i^i,y_i')) \|_2 + \frac{1}{2n}\sum_{j=1}^{n}\min_{i} \| ((x_i,y_i) - (x_i^i,y_i')) \|_2 \tag{13}$$

式中:$(x_i,\ y_i)\in R$,$(x_j',\ y_j')\in R'$。

2.4 不确定目标物的实例分割网络构建与实现

如图 6 所示,构建了不确定起重装卸场景的实例分割网络。其中检测网络可分为 5 个部分:点初始化、密集点

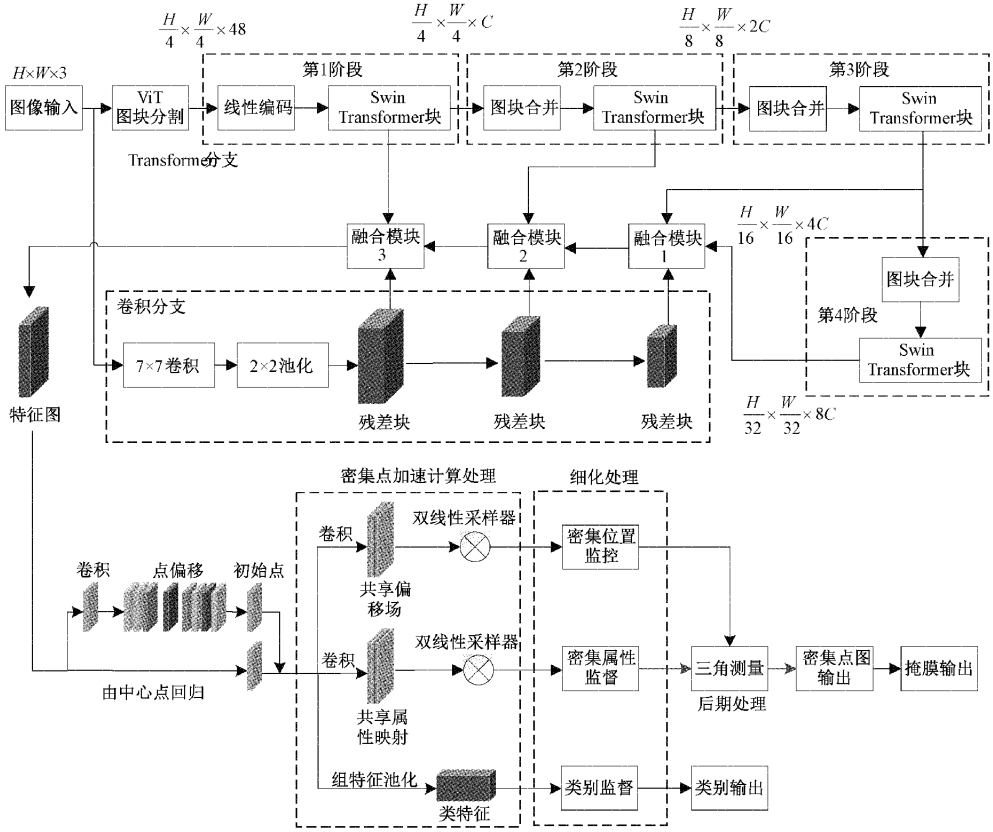


图 6 不确定起重装卸场景的实例分割网络

计算加速处理、细化处理、后期处理和掩膜输出。首先，通过从中心点回归生成初始代表点，并对这些密集初始点进行计算加速处理。具体地，通过引入了组池和共享偏移场来降低计算复杂度，并引入了一个共享属性图来有效地预测一个点是否在前景中。对初始点计算加速后通过密集位置监督、密集属性监督和类别监督进行细化处理。最后使用三角测量的方法进行后期处理，最终输出掩膜结果，实现目标的实例分割。

3　应用试验与分析

本文在一个实际的应用场景中对本研究的实例分割模型进行了实验测试。该场景使用智能门式起重机对装有含辐射性废物的容器进行无人化卸车和搬运深埋。装载容器形状尺寸规格及装车状态各异，运输车辆状态及停车位置不确定，需要动态识别检测，传输数据给起重机及智能吊具进行抓取和搬运。

3.1　数据集与评价指标

针对起重装卸存在的场景各异、目标物训练数据匮乏问题，本项目设计了一种生成对抗网络，用于合成准确的含语义标注和关键点标注的数据集。本文使用了构建的起重装卸目标物（装载含辐射性废物的圆桶和方箱）标注数据集，共 12 000 张进行实验。按 5∶2∶3 的比例划分训练集、验证集与测试集。

3.2　训练设置

模型的超参数设置如表 1 所示。实验中模型训练优化采用 AdamW 方式；使用余弦衰减学习率调度器和 20 个线性预热的 epoch（训练次数）对网络参数进行初始化，正式训练的 epoch 设置为 300；训练批量大小设置为 8；为了让模型能够以较好的速度进行训练，初始学习率设置为 0.001，并使用 0.000 1 的权重衰减。

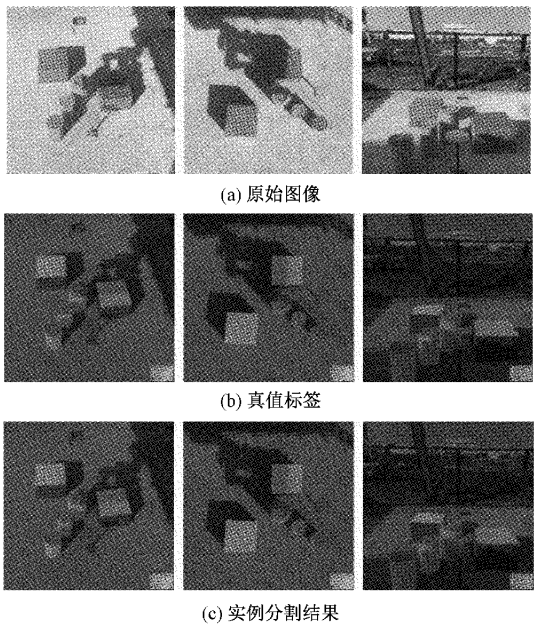
3.3　实验方法及结果与分析

为了探讨主干网络改进对于模型精度和分割完整度的影响，在构建好的起重装卸目标物数据集上应用本文第 2 节研究构建的主干网络和检测网络做了不同场景的多个圆筒和方箱实例分割实验，并与同类的主干网络 ResNet-50、ResNet-101、ResNeXt-101-DCN 和 Swin Transformer 进行对比。检测网络使用原始的代表点检测网络 Dense RepPoints。实验结果应用平均精度（AP）和分割完整度（mIoU，掩码交并比）作为评价指标，并与其他 4 种先进主干网络模型进行对比。模型的类目以及类型/数值由表 1 所示。

表 1　模型的超参数

类目	类型/数值
优化器	AdamW 优化器
学习率	0.001
训练批量	8
轮次	300

图 7 展示了本文主干网络和 Dense RepPoints 检测网络组成的实例分割方法在 3 种卸场景下分割多个圆桶、木箱和圆板的分割效果。由图可见不仅能准确分割目标物，而且还能准确地分割同一目标物的不同表面。因此，该分割结果可直接应用于目标物的定位。



(a) 原始图像

(b) 真值标签

(c) 实例分割结果

图 7　本文实例分割方法实验的分割效果图

主干网络消融实验结果表 2 表明，Swin Transformer 主干网络与 Dense RepPoints 检测网络结合能有效提高分割准确率，达到 93.87%，分割完整度 mIoU 达到 86.47%，证明了 Swin Transformer 骨干网络有效提高了对起重装卸目标的特征提取能力。本文主干网络在 Swin Transformer 上进一步改进，融合卷积分支，AP 值达到 98.82%，分割完整度 mIoU 达到 91.89%。比 ResNeXt-101-DCN、Swin Transformer 分别提高 11.36%、4.95%AP 和 7.98%、5.42%mIoU。由此可见，本文主干网络提升了网络的远程相关性建模能力和深度线索挖掘能力，验证了本文主干网络改进的有效性和先进性，在装卸场景下能达到很好的实例分割效果。

表 2　主干网络消融实验结果及对比

主干网络	检测网络	AP/%	mIoU/%
ResNet-50	Dense RepPoints	79.35	70.72
ResNet-101		81.23	74.69
ResNeXt-101-DCN		87.46	83.91
Swin Transformer		93.87	86.47
本文骨干网络		**98.82**	**91.89**

4　结　　论

本文对无人化起重装卸场景下不确定目标物的实例

分割技术展开研究,通过设计一个融合CNN和Transformer的异构特征信息的模块改进主干网络,提高了主干网络远程相关性编码和深度线索编码的能力,有效提取语义特征信息和空间特征信息。然后通过引入Dense RepPoints检测网络构建了无人化起重装卸场景的实例分割网络,准确地分割目标物且能分割目标物的不同表面,优于目前最先进的方法,解决了起重装卸智能化的目标识别关键技术难题。

参考文献

[1] 郑业. 面向仓储环境的物体抓取与识别技术研究[D]. 哈尔滨:哈尔滨工业大学, 2019.

[2] 崔芳. 面向智能仓储的大类物体识别与检索方法研究[D]. 成都:电子科技大学, 2020.

[3] 张荣旭. 基于视觉的物流配送中心叉车AGV设备场景识别和路径规划[D]. 济南:山东大学, 2021.

[4] 卢国杰,王桂棠,陈泳铮, 等. 基于生成对抗网络的自动装卸目标物标注数据集生成方法[J]. 电子测量技术, 2022, 45(17): 86-93.

[5] 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.

[6] 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, 2022, 17(1): 16-31.

[7] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[J]. Communication of the ACM, 2017, 60(6):84-90.

[8] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 1-9.

[9] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.

[10] TAN M, LE Q. Efficientnet: Rethinking model scaling for convolutional neural networks [C]. International Conference on Machine Learning, 2019: 6105-6114.

[11] ZAIDI S S A, ANSARI M S, ASLAM A, et al. A survey of modern deep learning based object detection models [J]. Digital Signal Processing, 2022, 126, DOI:10.1016/J.DSP.2022.103514

[12] MAHMUD M, KAISER M S, HUSSAIN A, et al. Applications of deep learning and reinforcement learning to biological data[J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(6): 2063-2079.

[13] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 3431-3440.

[14] CHEN L C, PAPANDREOU G, SCHROFF F, et al. Rethinking atrous convolution for semantic image segmentation [J]. ArXiv Preprint, 2017, ArXiv:170605587.

[15] RONNEBERGER O, FISCHER P, BROX T. U-net: Convolutional networks for biomedical image segmentation[C]. International Conference on Medical Image Computing and Computer-assisted Intervention, 2015: 234-241.

[16] ZHAO H, SHI J, QI X, et al. Pyramid scene parsing network[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 2881-2890.

[17] XIE S, GIRSHICK R, DOLLáR P, et al. Aggregated residual transformations for deep neural networks[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 1492-1500.

[18] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30, DOI:10.48550/raXiv:1706.03762.

[19] FU J, LIU J, JIANG J, et al. Scene segmentation with dual relation-aware attention network[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 32(6): 2547-2560.

[20] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[J]. ArXiv Preprint, 2020, ArXiv:201011929.

[21] LIU Z, LIN Y, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 10012-10022.

[22] SAXENA A, CHUNG S, NG A. Learning depth from single monocular images[J]. Advances in Neural Information Processing Systems, 2005: 18, DOI:10.1109/TPAM1.2015.2505283.

[23] DONG X, BAO J, CHEN D, et al. Cswin transformer: A general vision transformer backbone with cross-shaped windows [J]. ArXiv Preprint, 2021, ArXiv:2107.00652.

[24] RANFTL R, BOCHKOVSKIY A, KOLTUN V. Vision transformers for dense prediction [C]. Proceedings of the IEEE/CVF International

Conference on Computer Vision, 2021: 12179-12188.
[25] SRINIVAS A, LIN T Y, PARMAR N, et al. Bottleneck transformers for visual recognition[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 16519-16529.
[26] WANG W, YAO L, CHEN L, et al. Crossformer: A versatile vision transformer based on cross-scale attention [J]. ArXiv Preprint, 2021, ArXiv: 2108.00154.
[27] YANG Z, XU Y, XUE H, et al. Dense reppoints: Representing visual objects with dense point sets[C]. European Conference on Computer Vision, 2020: 227-244.
[28] LING H, GAO J, KAR A, et al. Fast interactive object annotation with curve-gcn[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 5257-5266.
[29] XIE E, SUN P, SONG X, et al. Polarmask: Single shot instance segmentation with polar representation [C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 12193-12202.
[30] RUBNER Y, TOMASI C, GUIBAS L J. The earth mover's distance as a metric for image retrieval[J]. International Journal of Computer Vision, 2000, 40(2): 99-121.
[31] FAN H, SU H, GUIBAS L J. A point set generation network for 3d object reconstruction from a single image[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 605-613.

作者简介

王国桢，硕士研究生，主要研究方向为智能定位传感测量、计算机视觉、深度学习等。
E-mail:635496946@qq.com

卢国杰，硕士研究生，主要研究方向为智能测控、计算机视觉、深度学习等。
E-mail:635496946@qq.com

王桂棠(通信作者)，教授、硕士生导师，主要研究方向为仪器科学与技术、机器视觉等。
E-mail:wanggt@gdut.edu.cn