DOI: 10.16452/j.cnki.sdkjzk.2026.01.011　　文章编号:1672-3767(2026)01-0122-09

基于多维空间视觉感知的仓库盘点方法

钟　诚[1],田殿雄[1],赵文炎[1],卢泽钰[1],张良安[2],李　勇[2]
(1.国网冀北电力有限公司唐山供电公司,河北 唐山 063000;2.上海洲固电力科技有限公司,上海 200040)

摘　要:传统人工仓库物资盘点方式效率低下,难以满足现代智慧仓储对物资精准管理的需求。现有的自动识别技术存在易受环境干扰、成本高昂以及在复杂场景下稳定性不足等局限,尤其在动态仓库环境中,物资盘点本质上是一个多维时空信息处理问题,需要有效融合时间、空间与距离等不局限于三维物理结构的高维度抽象特征,同时还需应对遮挡、光线变化和小目标检测等挑战。针对上述问题,本研究提出一种基于多维空间视觉感知的仓库盘点方法(MDSVP-WIM)。该方法通过混合高斯模型动态更新背景来抑制环境干扰,结合二维卷积与图像分割技术提升小目标识别精度,并引入跨帧追踪-联合投票机制,增强系统在动态场景中的稳定性与鲁棒性。在 WareSegNet 与 SpatioTrack-360 数据集上的实验表明,本研究方法的平均精度均值(mAP)分别达到 0.97 和 0.96,F1 值为 0.94 和 0.93,性能显著优于 FCN、PSPNet 等基准模型,为复杂仓储环境下的物资自动化盘点提供了一种高效、稳定的方案。

关键词:智慧仓储;物资盘点;视觉感知;混合高斯模型;跨帧追踪
中图分类号:TB3; U411　　**文献标志码**:A

A warehouse inventory method based on multidimensional spatial visual perception

ZHONG Cheng[1], TIAN Dianxiong[1], ZHAO Wenyan[1], LU Zeyu[1], ZHANG Liang'an[2], LI Yong[2]
(1. Tangshan Power Supply Company of State Grid Jibei Electric Power Co., Ltd., Tangshan 063000, China;
2. Shanghai Zhougu Power Technology Co., Ltd., Shanghai 200040, China)

Abstract: The inefficiency of the traditional manual material inventory method makes it difficult to meet the precise management needs of modern intelligent warehousing. However, the existing automatic recognition technologies have limitations such as susceptibility to environmental interference, high costs, and insufficient stability in complex scenarios. Especially in dynamic warehouse environments, material inventory is essentially a multidimensional spatiotemporal information processing problem, which requires effective integration of high-dimensional abstract features, such as time, space, and distance that are not limited to three-dimensional physical structures, and addresses challenges such as occlusion, light changes, and small object detection. In response to the above issues, this study proposes a multidimensional spatial visual perception based warehouse inventory method (MDSVP-WIM). This method suppresses environmental interference by dynamically updating the background through a Gaussian mixture model, improves the accuracy of small object recognition by combining the techniques of two-dimensional convolution and image segmentation, and enhances the stability and robustness of the system in dynamic scenes by introducing cross-frame tracking joint voting mechanism. The experiments on the WareSegNet dataset and SpatioTrack-360 dataset show that the proposed method, with the mean average precision (mAP) reaching 0.97 and 0.96 respectively and the F1-Score reaching 0.94 and 0.93 respectively, has a significantly better performance than benchmark models such as FCN and PSPNet. This study provides an efficient

收稿日期:2024-12-26
基金项目:山东省自然科学基金项目(ZR2024MF142)
作者简介:钟　诚(1978—),男,吉林长春人,硕士,高级经济师,主要研究方向为物资管理.
赵文炎(1985—),男,湖北仙桃人,硕士,高级工程师,主要研究方向为物资管理、轨迹预测,本文通信作者.
Email:15175579560@139.com

and stable visual perception solution for automated material inventory in complex warehousing environments.

Key words: intelligent warehousing; material inventory; visual perception; Gaussian mixture model; cross-frame tracking

随着物流行业的快速发展,仓储向智能化、自动化方向的深度转型,传统的人工盘点方式已难以满足高效、精准的管理需求[1-2]。尤其是在复杂仓库场景中,高密度货架导致的视觉遮挡、多品类货物混放带来的特征干扰,以及低光照或强反光等环境因素,进一步放大了人工操作的局限性。因此,开发高效的自动仓储物资盘点方法已成为物流仓储行业的迫切需求[3]。

目前主流的自动识别技术中,条码识别和磁卡识别等技术存在纸质标签易损坏、易遮挡以及电子标签成本较高等问题。在此背景下,基于计算机视觉的自动盘点技术正逐渐成为行业研究热点。该技术通过深度学习模型,对视频中货物视觉特征进行自动识别与提取,有效避免了传统标签技术带来的局限性。

基于卷积神经网络(convolutional neural network,CNN)的目标检测算法以物体及边界框为核心表征方式,但存在小目标像素占比低导致特征提取不充分、漏检率高达 30%,以及近距离密集目标边界模糊引发锚框回归重叠、误检率超过 25%等问题[4]。这类传统模型侧重于单帧图像处理,缺乏对跨帧时序信息的有效利用,难以应对货物移动导致的连续帧间特征变化。此外,基于自注意力机制的模型虽然在长距离依赖建模中性能优异,但在仓库场景中对局部细节的捕捉能力通常弱于 CNN 模型[5]。

针对上述问题,本研究提出一种基于多维空间视觉感知的仓库盘点方法(multidimensional spatial visual perception based warehouse inventory method, MDSVP-WIM),其主要创新有:

1) 多维空间去噪模块利用混合高斯(mixture of Gaussian, MoG)模型动态更新背景,能够处理时间、距离等抽象特征,并通过马氏距离计算像素与背景分布的差异,实现前景目标的快速分离,减少动态背景干扰;

2) 多维空间视觉感知模块通过多维卷积并行提取多尺度特征,模块中的多个分支分别从不同角度对图像进行处理,每个分支提取的特征包含了不同的空间信息,从而增强对高维抽象特征的感知能力;

3) 多维空间跨帧跟踪-联合投票模块以边界框几何中心点为核心跟踪锚点,结合视频流连续性实现动态目标的稳定定位,并通过多帧结果融合提升语义判别准确性,有效增强系统在动态场景下对移动货物的跟踪与识别的鲁棒性;

4) 实验结果表明,MDSVP-WIM 在 WareSegNet 和 SpatioTrack-360 数据集上的检测平均精度均值(mean average precision, mAP)分别达到 0.97 和 0.96,F1 值分别为 0.94 和 0.93,性能显著优于全卷积网络(fully convolutional networks,FCN)、金字塔场景解析网络(pyramid scene parsing network, PSP-Net)等基准模型,为复杂仓储环境下实现高效、智能的物资盘点提供了有效的解决方案。

1　相关工作

在仓库自动化管理领域,目标识别与盘点技术的研究主要围绕计算机视觉与深度学习展开。现有方法可归纳为传统特征工程方法、基于 CNN 的目标检测方法和语义分割方法三大类。

早期研究多依赖于人工设计的特征算子提取目标信息。例如,王紫芮等[6]采用尺度不变特征变换方法生成 128 维特征向量,但该方法计算复杂度高,对边缘光滑目标特征响应弱,在动态仓储场景中漏检率高达 40%;Dalal 等[7]提出的定向梯度直方图方法需手动设置滑动窗口参数,在处理密集遮挡目标时易出现特征混淆,误检率达 35%,且难以有效应对多尺度目标共存场景。因此,传统方法因依赖人工特征设计,泛化能力有限,难以适应仓储场景的复杂性与动态性。

近年来,基于 CNN 的目标检测算法逐渐成为主流,其通过卷积层自动学习特征,实现了从人工特征到数据驱动特征的跨越,显著提升了检测精度与泛化能力。Faster R-CNN[8]作为两阶段算法的代表,能够端到端地完成目标定位与分类,在标准尺寸目标上检测精度较高,但对小目标的漏检率也偏高。YOLO 系列[9]采用单阶段架构,检测速度较快,如 YOLOv3 引入多尺度预测机制,提升了对不同尺寸目标的检测能力,但在仓储复杂环境中仍存在因干扰导致的误检问题[10]。

针对上述目标检测的粗粒度缺陷,语义分割技术通过像素级分类提供了更精细化的特征。Shelhamer 等[11]提出的 FCN 以卷积层替代全连接层,支持任意尺寸输入,为语义分割任务奠定了基础,但 FCN 对边界复杂、形状不规则的目标分割效果较差,在仓储异形货物分割中精度仅 60%。Zhao 等[12]提出的 PSP-Net 通过金字塔池化模块融合多尺度上下文信息,但在仓储密集遮挡环境中仍存在特征混淆和误检率高的问题。全景分割网络 UPerNet[13]作为面向全景分割的统一框架,借助编码器-解码器结构与特征金字塔,将仓储场景分割精度提升至 70%,但小目标漏检率仍达 30%。双重注意力网络 DANet[14]通过空间与通道双重注意力机制增强长程依赖建模,但局部细节捕捉能力较弱,计算复杂度较高,相比 UPerNet 精度提高约 3%。专为遥感图像分割设计的多尺度动态图卷积神经网络 DMNet[15]引入动态尺度选择机制,但对小目标特征提取仍不充分,在密集目标场景下误检率超过 25%。

为突破上述瓶颈,研究者进一步提出了通分万物模型 v2(segment anything model v2, SAM v2)[16],该模型是首个将图像分割与视频分割统一到端到端框架中的通用大模型,基于时空记忆注意力机制缓存历史帧键值,能以 44 帧/s 的速度进行零样本掩码输出。动态上下窗口方法(dynamic context-swin v2, DC-Swin v2)[17]通过嵌入动态上下文窗口,可根据目标密度自适应调整大小与步长,在仓储任务中分割精度达到 75%。

2 MDSVP-WIM 方法设计

针对小目标特征提取不充分、复杂边界分割困难等问题,本研究提出 MDSVP-WIM,其整体架构如图 1 所示,由多维空间去噪模块、多维空间视觉感知模块以及跨帧跟踪-联合投票模块构成,输入为多物资监控的视频流,输出识别物资的类别、位置及置信度等信息。

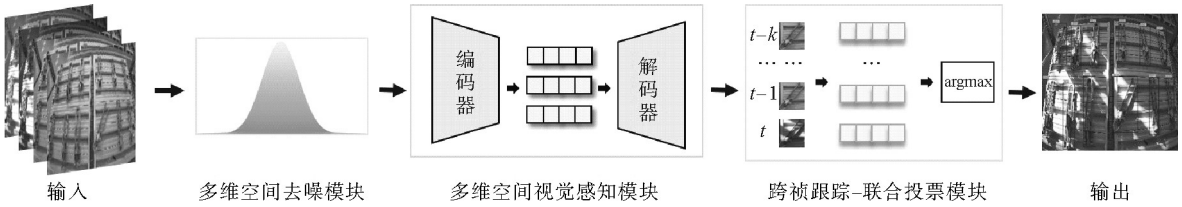


图 1 MDSVP-WIM 架构

Fig. 1 MDSVP-WIM architecture

2.1 多维空间去噪模块

多维空间去噪模块旨在减弱视频流中外部环境因素对检测目标的影响,其核心在于整合时间变化、空间关系和多通道特征等高维度信息,实现对视频流的多维特征建模。已有研究多采用背景减除方法(background subtraction methods, BSM),其核心思想是通过从当前帧中减去背景模型,从而识别出前景目标所对应的像素区域。

与传统方法不同,MDSVP-WIM 的多维空间去噪模块将输入的 RGB 像素数据扩展到长度为 5 的向量数据,该向量的 5 个特征值分别为 RGB 像素数据、像素点的空间距离和像素点在连续两帧(t 和 $t+1$ 时刻)图像中的空间距离变化率。如式(1)所示,$\boldsymbol{L}_{(x,y)}$ 不再是传统的 RGB 像素值,而是当前像素点(x,y)所对应长度为 5 的向量,其中 $\boldsymbol{L}_{(x,y)}$ 中的前 3 个元素分别对应像素点的三通道颜色值,$\boldsymbol{L}_{(x,y)}$ 中的第 4 个元素表示深度传感器获取像素点的空间距离,$\boldsymbol{L}_{(x,y)}$ 中的第 5 个元素表示像素点在连续两帧图像中的空间距离变化率。$\boldsymbol{L}^{\mathrm{bg}}_{(x,y)}$ 为估计的背景模型在 5 个维度上的特征值,$\boldsymbol{F}_{(x,y)}=(F^1_{(x,y)},F^2_{(x,y)},\cdots,F^5_{(x,y)})$用于衡量像素点偏离背景的程度。通过应用阈值函数进行前景分割,得到二值分割结果 $\boldsymbol{B}_{(x,y)}=(B^1_{(x,y)},B^2_{(x,y)},\cdots,B^5_{(x,y)})$,判断依据为 $\boldsymbol{F}_{(x,y)}$ 在 5 个维度上的特征值与预设阈值 T 的大小关系,如式(2)所示。

$$\boldsymbol{F}_{(x,y)}=\boldsymbol{L}_{(x,y)}-\boldsymbol{L}^{\mathrm{bg}}_{(x,y)}。\tag{1}$$

$$B^i_{x,y}=\begin{cases}255, & |F^i_{(x,y)}|>T;\\ 0, & |F^i_{(x,y)}|\leqslant T。\end{cases}\tag{2}$$

式中,$i=1,2,\cdots,5$。

MDSVP-WIM 中的多维空间去噪模块通过融合连续帧序列特征和空间距离维度,减弱背景干扰和对象噪点,增强主体边缘特征。为了便于数据训练,本研究使空间距离变化率服从多元高斯分布来完成背景减除,将每个像素的历史空间距离变化率视为高斯分布的混合值。当新帧在 $t+1$ 时刻出现时,通过计算每个像素与高斯分布的马氏距离来比较背景差异,并据此判断是否更新 $t+1$ 时刻分布的权重。

2.2　多维空间视觉感知模块

多维空间视觉感知模块以去噪后的单帧图像为输入,采用编码器-解码器结构提取多尺度空间特征,并完成像素级语义分割,输出与输入尺寸相同的类别掩码,为后续跨帧跟踪与联合投票提供精确的目标区域依据。

MDSVP-WIM 的空间视觉感知模块采用编码-解码器结构,编码卷积层的内部结构如图 2 所示,其设计遵循"归一化→特征增强→再归一化→特征投影"的流程。首先对输入特征进行批量归一化,然后引入注意力机制模块以增强特征的判别性。该模块采用轻量化深度可分离卷积网络架构,包含 1×1 卷积层、激活函数以及 MDSVP-EM 模块。特征先经过注意力模块处理后,再通过批量归一化层以保持稳定的数值分布,最后输入到前馈神经网络进行多阶特征投影和表达增强。前馈神经网络由 1×1 卷积层、d,3×3 卷积与 MDSVP-EM 构成。其中,1×1 卷积层用于通道调整与特征映射;深度可分离 d,3×3 卷积通过标准 3×3 卷积的中尺度感受野强化局部空间特征提取,d 表示卷积层的输入通道数,深度可分离卷积通过解耦空间与通道进行卷积;MDSVP-EM 则在空间中对特征进行非线性投影与增强调制,结构如图 3 所示。

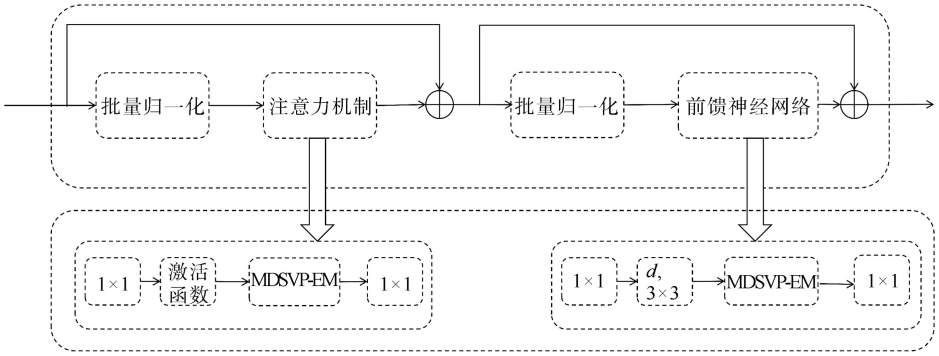


图 2　编码卷积层结构图

Fig. 2　The structure of encoder convolutional layer

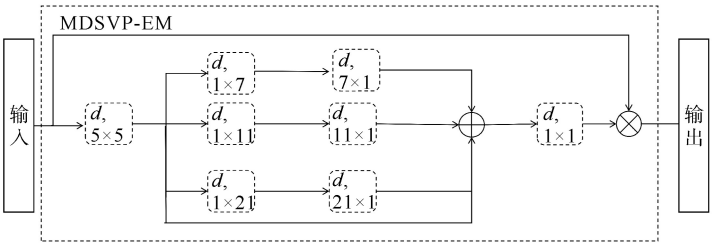


图 3　MDSVP-EM 结构图

Fig. 3　Structure of MDSVP-EM

MDSVP-EM 模块采用卷积并行提取策略,并对关键区域进行显式增强。首先,输入特征通过 5×5 的深度可分离卷积进行初步编码,然后并行输入至不同尺度的条状深度可分离卷积分支中。接着,对多尺度特征进行通道级的加权融合,融合后的特征再通过一个 1×1 的可分离卷积计算每个像素的注意力权重,以评估其在空间表达中的重要性。最后将该注意力图与原始输入特征进行逐元素乘法操作,得到增强后的空间特征表示。其中,条状深度可分离卷积分支的卷积核大小分别为 7、11 和 21。卷积核的参数量和计算量随卷积核的增长呈指数增长,为实现轻量化和高效率,如图 3 所示,采用 7×1 和 1×7 的卷积操作来模拟 7×7 标准二维卷积,同时结合激活函数和深度可分离卷积进行特征融合。通过带有激活函数的

逐通道深度可分离卷积,融合来自不同感受野的通道特征,获得目标各区域像素的注意力分数,最后通过加权计算强化目标主体特征表示,如式(3)和式(4)所示。

$$\boldsymbol{A}_{\mathrm{tt}}=\mathrm{Conv}_{1\times1}\left(\sum_{i=0}^{2}\mathrm{scale}_i\left(\mathrm{DeepConv}(\boldsymbol{X})\right)\right), \tag{3}$$

$$\boldsymbol{O}_{\mathrm{ut}}=\boldsymbol{A}_{\mathrm{tt}}\otimes\boldsymbol{X}。 \tag{4}$$

式中,$\boldsymbol{X}$ 表示 MDSVP-EM 模块的输入特征,即当前层当前时刻每个像素经过逐通道卷积处理后的历史信息;$\boldsymbol{A}_{\mathrm{tt}}$ 和 $\boldsymbol{O}_{\mathrm{ut}}$ 分别表示注意力图和输出特征;$\otimes$表示元素矩阵乘法运算;DeepConv 表示深度卷积;scale_i 表示图 3 中的第 i 个分支。前馈神经网络将这些增强后的特征表示编码为低维特征,同时保留有价值的特征。

为捕捉多层次语义信息,通过跳跃连接融合每一层编码卷积层输出的主体特征表示,并采用上采样策略恢复分辨率。本研究使用结构简单且拟合能力强的多层感知机(multilayer perceptron,MLP)来提取高级语义特征。首先,用一维卷积将编码器最后三层通道的特征全部对齐到相同维度;然后,对每层特征依次采用最近邻上采样进行放大,并与同尺度的跳跃连接特征拼接;拼接后的特征通过深度可分离卷积融合细节与语义信息,再经过一维卷积逐步压缩通道数,在保留空间细节特征的同时减少冗余通道,降低计算量并突出关键语义;最后,使用与两层 1×1 卷积操作等效的 MLP 实现逐像素分类。

2.3 跨帧跟踪-联合投票模块

为提高模型稳定性与鲁棒性,本研究提出目标跨帧跟踪-联合投票模块。该模块利用视频流的连续性,通过跨帧跟踪实时定位目标,并采用联合投票机制融合历史帧信息。具体地说,首先以边界框中心点作为跟踪初始点,确定初始框的中心点坐标,然后在后续帧中根据中心点的位移调整边界框位置。对于一个边界框,其左上角坐标为$(x_{\min},y_{\min})$,右下角坐标为$(x_{\max},y_{\max})$,中心点$(x_{\mathrm{center}},y_{\mathrm{center}})$的计算方法为:

$$(x_{\mathrm{center}},y_{\mathrm{center}})=\left(\frac{x_{\min}+x_{\max}}{2},\frac{y_{\min}+y_{\max}}{2}\right)。 \tag{5}$$

跨帧追踪算法基于目标在连续帧间的位置连续性假设,通过中心点的位移调整边界框位置。在后续帧中,算法根据当前边界框的几何坐标重新计算中心点,并据此调整边界框位置。

为进一步提升类别判断的准确性,采用联合投票机制对多帧中同一跟踪区域的主体特征进行语义判断。该方法采用特征记忆池存储同一跟踪对象在多帧中的区域像素信息,将多次识别结果融合为一个更精确的单一结果。联合投票机制可形式化表示为:

$$\widehat{S}_{i,j}=\arg\max_{C=1}^{k}N_{i,j}(C)。 \tag{6}$$

式中:$\widehat{S}_{i,j}$ 表示像素点(i,j)的联合投票结果,k 为当前分割检测目标的数量;$N_{i,j}(C)$表示在 k 次分割结果中将类别 C 分配给像素(i,j)的次数,argmax 表示选取使 $N_{i,j}(C)$最大的类别 C。

3 实验

3.1 实验数据集

为评估模型性能,本研究采用无人仓库自行采集的 WareSegNet 和 SpatioTrack-360 数据集,主要包含电网物资及零部件图像。数据集总涵盖 120 h 视频、10 368 000 张图像,涉及 10 类物资,场景覆盖多样化的仓储环境。数据集按 75%∶20%∶5%的比例划分为训练集、测试集和验证集。

3.2 评估指标

本研究采用精确度(Precision)、召回率(Recall)、平均精度(mean average precision,mAP)和 F1 值作为定量评估指标,分别记为 P、R、m_{AP} 和 F_1,如式(7)~(11)所示。

$$P=\frac{T_{\mathrm{P}}}{T_{\mathrm{P}}+F_{\mathrm{P}}}, \tag{7}$$

$$R=\frac{T_{\mathrm{P}}}{T_{\mathrm{P}}+F_{\mathrm{N}}}, \tag{8}$$

$$A_P = \int_0^1 P(R)\,\mathrm{d}R, \tag{9}$$

$$m_{AP} = \sum_{i=1}^{C} \frac{A_P(i)}{C}, \tag{10}$$

$$F_1 = 2 \times \frac{P \times R}{P + R} \tag{11}$$

式中:T_P 表示真阳性的数量,F_P 表示假阳性的数量,F_N 表示假阴性的数量,C 为类别总数。

3.3　对比实验

3.3.1　基准方法介绍

在对比实验中,选取以下 7 种主流方法作为基准,与 MDSVP-WIM 进行性能比较。

1) FCN[11]:全卷积神经网络,采用反卷积层实现上采样,能够以端到端方式进行图像语义分割。

2) PSPNet[12]:采用金字塔池化模块聚合多尺度上下文信息,有效融合全局与局部特征,提升场景解析能力。

3) UPerNet[13]:全景分割网络,通过共享的特征提取器和多个任务特定分支实现语义分割与实例分割的融合。

4) DANet[14]:双重注意力网络,基于双重注意力机制,分别在空间和通道维度上建模上下文依赖,增强模型对复杂场景的理解能力。

5) DMNet[15]:多尺度动态图卷积神经网络,为图像分割设计,引入动态尺度选择机制,根据图像特征自适应选择感受野大小,以更好地捕捉不同尺度的目标特征。

6) SAM v2[16]:通分万物模型 v2,在支持零样本迁移的基础上,引入可提示记忆解码器,提升小目标分割与边缘一致性,并通过缓存历史帧键值对,增强跨帧间目标表示的稳定性。

7) DC-Swin v2[17]:基于 Swin Transformer 结构,嵌入动态上下文窗口机制,可根据目标密度自适应调整窗口大小与步长,在保持线性计算复杂度的同时扩大感受野,并结合局部-全局双路径编码提升特征表达能力。

3.3.2　定量分析

本实验围绕检测与识别两项核心任务,评估各方法在仓库盘点场景下的性能表现。7 种对比方法在 WareSegNet 和 SpatioTrack-360 数据集上的评估指标如表 1 所示。

由表 1 可知,FCN 在检测任务中性能较差,其 F1 值仅为 0.31 和 0.33。这主要由于 FCN 在多次降采样过程中丢失了大量细节信息,对大范围上下文的理解能力不足,难以充分提取目标的判别性特征。PSPNet 检测任务相较于 FCN 在误检方面有所改善,且 mAP 在两个数据集上分别为 0.62 和 0.69,但其 F1 值仍偏低,分别为 0.28 和 0.32。在识别任务中,PSPNet 精确度为 0,表明该方法未能正确识别任何目标物品类别。尽管 PSPNet 通过金字塔池化模块扩展了感受野并捕获更多高维特征,但在训练过程中对不确定样本的错误归类导致过拟合问题。因此,尽管其平均精确度高于 FCN,但精确度和 F1 值反而低于 FCN。

UPerNet 作为一种性能优异的语义分割方法,在多任务学习与细节信息保留方面表现突出,其编码器-解码器结构与特征金字塔网络有助于保留更多空间细节,从而提升分割精度。与 UPerNet 相比,DANet 在物品检测精确度上提升约 20%,其检测 mAP 分别达到 0.86 和 0.88,F1 值分别为 0.54 和 0.55,表明该方法在精确度与召回率之间取得了较好的平衡。

DMNet 在识别与检测精确度方面较 DANet 提升有限,但其 F1 值与平均精度均有显著提高。该方法能够动态调整网络结构以适应不同尺度目标,在检测任务中 F1 值分别达到 0.61 和 0.60。相比之下,SAM v2 在检测 mAP 上进一步提升约 8%,识别 mAP 提升约 4%;检测 F1 值提升约 30%,识别 F1 值提升约 15%。SAM v2 借助时空记忆-注意力机制,在实现多尺度特征融合的同时保持跨帧一致性,从而在动态仓库场景中实现精确度、召回率性能平衡。与 SAM v2 相比,DC-Swin v2 在检测 mAP 上略低约 2%,检测 F1 值下降约 2%,识别 F1 值下降约 3%。

表 1 定量实验结果

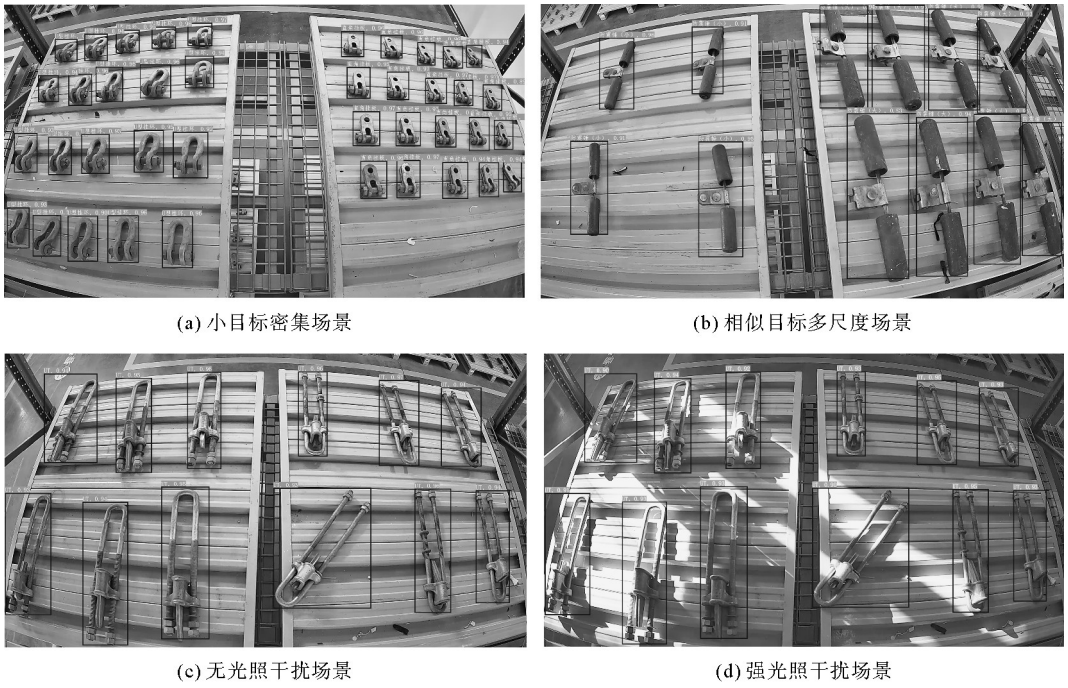
Table 1 Quantitative experiment results

方法	任务	WareSegNet				SpatioTrack-360			
		m_{AP}	R	F_1	P	m_{AP}	R	F_1	P
FCN	检测	0.53	1.00	0.31	0.19	0.55	1.00	0.33	0.20
	识别	0.65	1.00	0.08	0.04	0.59	1.00	0.12	0.07
PSPNet	检测	0.62	1.00	0.28	0.17	0.69	1.00	0.32	0.19
	识别	0.75	0	—	0	0.77	1.00	—	0.02
UPerNet	检测	0.71	1.00	0.32	0.19	0.74	1.00	0.32	0.19
	识别	0.83	1.00	0.03	0.01	0.85	1.00	0.02	0.02
DANet	检测	0.86	1.00	0.55	0.38	0.88	1.00	0.55	0.39
	识别	0.88	0	—	0	0.87	0	—	0
DMNet	检测	0.87	1.00	0.61	0.44	0.88	1.00	0.60	0.44
	识别	0.89	1.00	0.09	0.05	0.91	1.00	0.04	0.03
SAM v2	检测	0.95	0.98	0.92	0.87	0.94	0.97	0.93	0.89
	识别	0.93	1.00	0.77	0.63	0.94	1.00	0.83	0.71
DC-Swin v2	检测	0.93	0.97	0.90	0.84	0.92	0.95	0.89	0.83
	识别	0.91	1.00	0.67	0.50	0.92	1.00	0.80	0.57
MDSVP-WIM	检测	0.97	0.96	0.95	0.93	0.96	0.95	0.93	0.89
	识别	0.97	1.00	0.75	0.60	0.95	1.00	0.73	0.57

MDSVP-WIM 在检测与识别任务中均显著优于上述 7 种对比方法，展现出较好的综合性能。在检测任务中，MDSVP-WIM 在两个数据集上的 mAP 分别达到 0.97 和 0.96，远高于其他方法；在识别任务中，其 mAP 与 F1 值优势明显。需要指出的是，SAM v2 在识别精确度方面略高于本研究模型，这主要源于其零样本迁移能力与基于大规模预训练所带来的泛化性能，使其在相似品类物资的区分上表现更佳。而 MDSVP-WIM 在检测精度上的优势则源于其多维空间去噪模块对仓储环境干扰的有效抑制，以及跨帧跟踪模块对动态目标的稳定定位能力。尽管在识别精确度上略低于 SAM v2，但 MDSVP-WIM 在密集遮挡等复杂仓储场景中表现出更强的检测鲁棒性，更符合实际盘点任务对低误检率与高定位精度的要求。总之，MDSVP-WIM 在保持高召回率的同时，显著提升了检测与识别的精确度与平衡性，其低误报率和高鲁棒性使其成为当前任务中的最优解决方案。

3.3.3 定性分析

本研究进一步通过四类典型复杂场景定性分析验证 MDSVP-WIM 的有效性，包括小目标密集场景、相似目标多尺度识别场景、无光照干扰场景和强光照干扰场景。图 4 中方框标注了模型的目标识别结果，显示在各类复杂场景下其识别精确度均超过 90%。在图 4(a)所示的小目标密集场景中，MDSVP-WIM 基于 BSM 和轻量级编码-解码架构实现了精准识别。图 4(b)展示了在相似目标多尺度识别场景下的实验结果，模型展现出卓越的特征提取与识别能力。为验证模型在外部干扰条件下的鲁棒性，本研究特别设计了强光照变化的实验场景。图 4(c)和图 4(d)分别展示了无光照干扰和强光照干扰条件下的识别结果。实验表明，MDSVP-WIM 在常规条件下具有极高的识别精度。当受到光照等外部干扰时，模型通过跨帧位置跟踪、语义分割及多数投票机制保证识别精度。综上，该模型适用于复杂仓储环境，具有较高的实际应用价值。



(a) 小目标密集场景　(b) 相似目标多尺度场景

(c) 无光照干扰场景　(d) 强光照干扰场景

图 4　定性实验结果

Fig. 4　Results of qualitative experiments

3.3.4　消融实验

为分析各关键模块的贡献,本研究在 SpatioTrack-360 数据集上对 MDSVP-WIM 进行了消融实验,结果如表 2 所示。移除多维空间去噪模块的 MDSVP-WIM_A,在检测与识别任务中各指标均出现下降,表明该模块对提升本研究方法检测与识别性能具有正向作用,能够为后续处理提供更纯净的输入特征。在 MDSVP-WIM_B 中移除多维空间视觉感知模块后,检测与识别指标下降更为显著,说明该模块在提升本研究方法、平均精度方面至关重要,能够为本研究方法提供重要的像素级分类信息,缺少该模块会显著影响本研究方法在复杂场景中的表现。在 MDSVP-WIM_C 中移除跨帧跟踪-联合投票模块后,检测与识别性能出现小幅下降,但整体仍维持在较高水平,表明该模块有助于提升方法在动态场景下的稳定性。

表 2　消融实验结果

Table 2　Results of the ablation experiment

方法	任务	m_{AP}	R	F_1	P
MDSVP-WIM_A	检测	0.90	0.93	0.70	0.56
	识别	0.92	0.60	0.11	0.07
MDSVP-WIM_B	检测	0.89	0.91	0.65	0.51
	识别	0.92	1.00	0.11	0.07
MDSVP-WIM_C	检测	0.91	0.95	0.68	0.53
	识别	0.92	1.00	0.15	0.09
MDSVP-WIM	检测	0.96	0.95	0.92	0.89
	识别	0.95	1.00	0.71	0.57

4　结论

本研究针对仓储盘点任务中存在的小目标易漏检、密集目标易误检以及动态场景适应性差等问题,提出了 MDSVP-WIM。其中,多维空间去噪模块通过引入高斯混合模型实现动态更新背景,融合连续帧间像素变化序列与空间维度信息,有效去除环境干扰,为精准盘点奠定基础;多维视觉感知模块提升多尺度特征提取能力;跨帧跟踪-联合投票模块提升了动态目标的定位稳定性。实验结果表明该方法在检测精度和鲁棒性方面具备显著优势,尤其在复杂场景下表现优异。未来研究可进一步推进方法的轻量化设计,并拓展其在更多工业场景中的应用。

参考文献:

[1] ZHOU B L,ZHAO H,PUIG X,et al. Scene parsing through ADE20K dataset[C]//Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2017:633-641.

[2] ZACCARIA M,MONICA R,ALEOTT J. A comparison of deep learning models for pallet detection in industrial warehouses[C]//Proceedings of the 16th IEEE International Conference on Intelligent Computer Communication and Processing. Piscataway:IEEE Press,2020:417-422.

[3] 刘冰,郭李丽,刘如飞,等.全景影像快速加载与精确量测算法研究[J].山东科技大学学报(自然科学版),2024,43(4):57-66.
LIU Bing,GUO Lili,LIU Rufei,et al. Research on fast loading and accurate measurement algorithm for panoramic images[J]. Journal of Shandong University of Science and Technology(Natural Science),2024,43(4):57-66.

[4] GIRSHICK R,DONAHUE J,DARRELL T,et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2014:580-587.

[5] LINDEBERG T. Scale invariant feature transform[J]. International Journal of Computer Vision,2012,91(3):1-42.

[6] 王紫芮,蒋德钧.基于超低延迟 SSD 的页交换机制关键技术[J].计算机研究与发展,2024,61(3):557-570.
WANG Zirui,JIANG Dejun. Key techniques of swapping mechanism based on ultra-low latency SSD[J]. Journal of Computer Research and Development,2024,61(3):557-570.

[7] DALAL N,TRIGGS B. Histograms of oriented gradients for human detection[C]//Proceedings of the 17th IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2005,1:886-893.

[8] REN S Q,HE K M,GIRSHICK R,et al. Faster R-CNN:Towards real-time object detection with region proposal networks[J]. Advances in Neural Information Processing Systems,2015,28:91-99.

[9] JIANG P Y,ERGU D,LIU F Y,et al. A review of YOLO algorithm developments[J]. Procedia Computer Science,2022,199:1066-1073.

[10] GOODFELLOW I,BENGIO Y,COURVILLE A. Deep Learning[M]. Cambridge:MIT Press,2016.

[11] SHELHAMER E,LONG J,DARRELL T. Fully convolutional networks for semantic segmentation[C]//Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2015:3431-3440.

[12] ZHAO H S,SHI J P,QI X J,et al. Pyramid scene parsing network[C]//Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2017:2881-2890.

[13] XIAO T,LIU Y C,ZHOU B L,et al. Unified perceptual parsing for scene understanding[J/OL]//Lecture Notes in Computer Science,2018. DOI:10.1007/978-01228-1_26.

[14] FU J,LIU J,TIAN H J,et al. Dual attention network for scene segmentation[C]//Proceedings of the 31th IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2019:3146-3154.

[15] LI W S,ZHANG X Y,PENG Y D,et al. DMNet:A network architecture using dilated convolution and multiscale mechanisms for spatiotemporal fusion of remote sensing images[J]. IEEE Sensors Journal,2020,20(20):12190-12202.

[16] RAVI N,GABEUR V,HU Y,et al. SAM 2:Segment anything in images and videos[C]//Proceedings of the Thirteenth International Conference on Learning Representation. Open Directory,2025:571-615.

[17] LIU Z,HU H,LIN Y T,et al. Swin transformer V2:Scaling up capacity and resolution[C]//Proceedings of the 37th IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2025:12009-12019.

(责任编辑:傅 游)