DOI: 10.13382/j.jemi.B2206089

基于空间特征融合的车间作业工具检测算法*

王 呈[1] 黄义超[1] 杨桂锋[2]
(1. 江南大学物联网工程学院 无锡 214122;2. 无锡威孚高科技集团股份有限公司 无锡 214031)

摘 要:手和工具的交互是区分车间人员作业行为的关键信息。为防止泵件装配工序错漏,达到实时监测的目的,提出基于空间特征融合的车间作业工具检测算法。首先,为了提高对目标的定位能力和检测精度,基于帧差法分割前景中的手部运动区域,获得具有运动空间特征的纹理图像,结合装配过程的 RGB 图像构成目标检测网络的双通道输入。设计空间感知模块实现双通道输入的空间特征融合,获得全局空间信息。利用特征增强模块融合全局空间信息和深层语义信息,加强显著位置的特征响应。然后,采用 ESNet(enhance shuffleNet)重构主干网络,基于深度可分离卷积实现多尺度特征提取,提高检测速度。最后,针对图像背景中局部元素变化问题,采用 CutOut 数据增强方法,提高模型抗干扰能力。实验结果表明,本文所提算法有效降低了误检率,较传统 YOLOv5s 的 mAP 提高 6.4%,能够快速准确检测车间人员作业时使用的工具。

关键词: 工具检测;帧差法;双通道;注意力
中图分类号: TP391;TN9 **文献标识码**: A **国家标准学科分类代码**: 520.6040

Workshop tool detection algorithm based on spatial feature fusion

Wang Cheng[1] Huang Yichao[1] Yang Guifeng[2]
(1. School of Internet of Things Engineering, Jiangnan University, Wuxi 214122, China;
2. Wuxi Weifu High-Technology Group Cooperation, Wuxi 214031, China)

Abstract: The interaction of hands and tools is the key information to distinguish the behavior of workers. To prevent the errors and omissions of the process in the assembly of pumps and achieve the purpose of real-time monitoring, a workshop tool detection algorithm based on spatial feature fusion is proposed. First, in order to improve the localization ability and detection accuracy of the object, the hand motion region in the foreground is segmented based on the frame difference method to obtain a texture map with hand spatial information, which is combined with RGB images of the assembly process to form a dual channel inputs to the object detection network. The spatial perception module is designed to realize the spatial feature fusion of the dual channel inputs and obtain the global spatial information. The feature enhancement module is proposed to mix the global spatial information and deep semantic information to enhance the feature response at salient locations. Then, the ESNet (enhance shuffleNet) is used to reconstruct the backbone network and form a multi-scale feature extraction module by deep separable convolution to improve the detection speed. Finally, in view of the local elements change both in the foreground and the background, the CutOut data enhancement method is used to improve the anti-interference capability. The experimental results show that the proposed algorithm can effectively reduce the false detection rate, and improve the mAP by 6.4% compared with the traditional YOLOv5s. The method can quickly and accurately detect the tools used by shop workers.

Keywords: tool detection; frame difference method; dual channel; attention

收稿日期: 2022-12-07 Received Date: 2022-12-07
*基金项目:国家自然科学基金面上项目(61973137)、近地面探测技术重点实验室预研基金(6142414220203)项目资助

0 引 言

手和工具交互是区分车间人员装配行为的重要信息。泵件装配过程中,包含使用喷油枪喷油、拧螺丝、安装 Mprop 配件等基本步骤,各步骤使用工具不同,准确检测使用工具对于判别错乱工序和违规操作具有重要作用[1]。泵件装配使用的工具具有数量多、尺度变化大等特点,且场景信息复杂,难以基于传统手工特征实现工具检测[2]。近年来,深度学习方法因卷积神经网络具有提取图像的空间和语义特征的能力,能够有效降低背景干扰,被广泛应用于不同尺度的目标检测中[3]。

基于深度学习的目标检测算法以 Faster-RCNN、YOLO、SSD 为代表分为双阶段和单阶段算法,算法中包含主干部分、连接层和预测层 3 个部分。其中主干部分用以提取输入图片特征,连接层融合不同尺度的特征,增强特征表现。在预测层中,以 Faster RCNN 为代表的双阶段检测算法利用区域推荐网络生成推荐框。以 YOLO、SSD 为代表的单阶段检测算法直接在不同尺度特征图上生成候选框,摒弃了区域推荐网络,在推理速度上优于双阶段检测算法。

针对手部接触目标检测问题,文献[4]利用 Faster-RCNN[5]目标检测网络,将手部接触物体统一视为活跃目标进行训练,模型能够快速回归活跃目标位置,但仍需对目标进行二次分类,得到目标具体类别。文献[6]以 Faster-RCNN 目标检测网络为基线模型,并构建一类车辆模型装配数据集,检测作业过程中手持工件的类别和位置。该研究中定义的手部作业行为识别场景与工业环境有较大差别,未考虑实际环境干扰和检测实时性。

车间作业的复杂环境会对视觉检测造成干扰,包括装配作业人员的更换,光照的改变和无关物件的出现造成图像前后景的信息变化,导致检测网络对目标的定位能力下降,影响检测效果。为提高模型对前景中手持工具的定位能力,考虑到连续帧图像中包含手与工具交互的运动信息,能够用于表征作业时手部活动的空间位置。构建双通道目标检测网络融合运动信息[7]并引导检测网络定位手部接触目标,可以提高检测精度。为获得连续帧图像的运动信息,文献[8]基于图像序列中像素在时间域上的变化以及相关性,生成光流图表征运动信息。文献[9]基于帧差法过滤背景像素,获得具有目标运动信息的前景图像。为构建双通道检测网络,文献[10]在 SSD 模型上构建并行的 VGG 特征提取网络,分别提取深度图像信息和 RGB 图像信息,将深度信息特征与 RGB 图像特征进行融合,提高目标检测精度。文献[11]在 YOLOv5 检测模型中加入语义分割子网络,用以区分图像中的目标和背景,将网络输出作为空间注意力图反馈给 YOLOv5 网络的检测层,增强图像兴趣区域的特征响应,提升小目标检测精度。针对作业环境中出现无关物件造成的背景局部元素变化,导致模型误检的问题,为提高模型抗干扰能力,学者通过 Cutout[12]、Random Erasing、GridMask 等基于区域随机删除的数据增强方法模拟噪声,引导模型关注全局空间,避免过拟合,或采用 CutMix、Mosaic 方法将多张图片剪切成一张,生成复杂的输入图片,增强模型鲁棒性[13]。但是,随机删除方法存在过度删除目标区域问题,造成关键像素或完整目标被删除,丢失上下文信息,使图像完全变为噪声[14],影响检测效果。

实时检测手持工具对算法运行速度有较高要求,为提高目标检测算法的实时性,文献[15]在 YOLOv4[16]基础上引入 MobileNetV3[17]主干网络,通过深度可分离卷积[18]轻量化主干网络提高检测速度。文献[19]提出特征高复用思想,设计 Ghost 卷积,对既有特征进行线性增广,用更少的参数生成更多的特征。文献[20]提出一种基于注意力机制的特征融合模块,结合通道和空间注意力赋予特征不同权重,在不增加参数规模的情况下,增强对图像细节的关注,提升目标检测精度。文献[21]结合深度可分离卷积、高复用特征提取思想和注意力机制,设计 ESNet(enhance shuffleNet)主干网络用于轻量级目标检测,取得优异的实时性能。

综上,针对面向装配工序步骤识别的目标检测问题需要解决算法的准确性、实时性、鲁棒性问题。本文将泵件装配 RGB 图像和手部运动前景图像组成 YOLOv5 检测网络的双通道输入,通过融合手部运动前景图像中的空间特征,强化模型学习和推理,提升检测精度,并结合轻量化卷积思想,提高检测速度。基于 CutOut 数据增强方法,平衡删除信息和保留信息,提升模型鲁棒性。

1 YOLOv5 算法简介

YOLOv5 的主干网络部分由卷积层和 CSP 模块组成,卷积层通过卷积和归一化后利用 SiLU 函数实现非线性激活。CSP 模块通过加强浅层和深层特征的直接传递减少参数量,一定程度上缓解梯度消失问题。特征融合部分使用特征金字塔(feature pyramid network,FPN)+路径聚合网络(path aggregation network,PAN)结构,充分融合浅层空间特征和高层语义特征,加强对不同尺度目标的检测能力。使用 Mosaic 方法实现数据增强,通过选取 4 张图片拼接成一张用于丰富训练数据,提升小目标的检测精度。

2　基于空间特征融合的车间作业工具检测算法

2.1　改进的双通道空间特征融合网络

目标检测网络多使用特征金字塔结构实现特征融合。本文针对手持工具检测问题,充分利用手部运动信息,基于传统特征金字塔,在 YOLOv5 的特征融合网络中设计手部区域分割模块(hand segmentation module, HSM)、空间感知模块(diff-convolutional block attention module, Diff-CBAM)、特征增强模块(feature enhance-CSP, FE-CSP),如图 1 所示。首先,通过手部运动区域分割模块过滤输入图像(input image)的背景信息,获得具有手部运动空间特征的纹理图像(DiffVein image),与主干网络(BackBone)的特征映射构成空间感知模块的双通道输入。然后,利用空间感知模块实现主干网络中的浅层空间特征和纹理图像中的运动空间特征融合,获得全局空间信息。最后,通过特征增强模块平滑深层特征的上采样(Upsample),融合全局空间信息和深层特征的语义信息,提高对手持工具的定位能力。

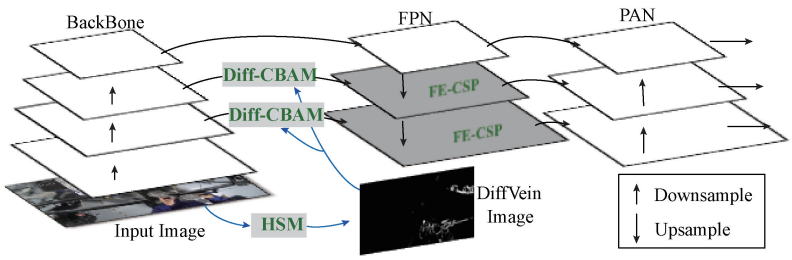


图 1　改进的双通道特征融合网络

Fig. 1　Improved dual channel feature fusion network

1)基于帧差法的手部运动区域分割模块

车间作业人员泵件装配过程图如图 2 所示。在图像序列的前后帧中,手部区域像素由于手和工具的交互动作,存在运动变化特点。因此,基于像素的时间差分并采用帧差法,能够得到记录像素运动过程的纹理图像,分割手部运动区域。

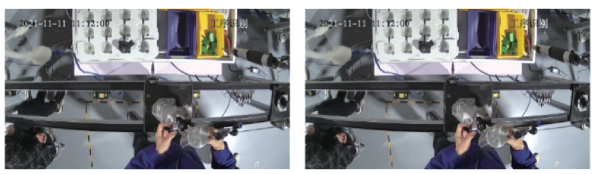


图 2　泵件装配过程前后帧图像

Fig. 2　Front and back frames of pump assembly

帧差法计算过程如下,首先将前后帧 RGB 图像对应的像素值相减得到差分图像,对于取自第 i 帧和第 $i+1$ 帧的两张 RGB 图像 $\boldsymbol{I}_i$ 和 $\boldsymbol{I}_{i+1}$,其中前后帧时间间隔为 200 ms,通过式(1)获得 RGB 差分图像 $\boldsymbol{I}_{diff}$:

$$\boldsymbol{I}_{diff} = |\boldsymbol{I}_{i+1} - \boldsymbol{I}_i| \tag{1}$$

利用灰度化得到灰度图 $\boldsymbol{I}'$,去除图像中的色彩信息并保留纹理,如图 3(a)所示。最后基于阈值化分割灰度图的前后景,进一步过滤背景中的像素。阈值化将像素值变化大于阈值的区域标记为前景,小于阈值的区域标记为背景,增强前后景对比度。本文通过最大类间方差法(Otsu)基于自适应阈值分割,对灰度图 $\boldsymbol{I}'$ 分割前后的效果如图 3 所示。

(a) 分割前
(a) Before the Otsu threshold

(b) 分割后
(b) After the Otsu threshold

图 3　阈值分割前后效果

Fig. 3　The effect before and after the Otsu threshold

Otsu 从 $g \in [0, 255]$ 的灰度值分布中,寻找使类间方差最大化的 g_s 作为阈值,阈值函数 δ 和类间方差 σ^2 的表达式分别如下:

$$\delta(\boldsymbol{I}'_{x,y}) = \begin{cases} 255 & \boldsymbol{I}'_{x,y} > g_s \\ 0 & \boldsymbol{I}'_{x,y} \leqslant g_s \end{cases} \tag{2}$$

$$\sigma^2(g) = \psi_1(m_1 - m_G)^2 + \psi_2(m_2 - m_G)^2 \tag{3}$$

式中:m_1 和 m_2 为前后景像素灰度均值,m_G 为像素全局均值,ψ_1 和 ψ_2 分别为以灰度值 g 为阈值分割的前后景像素分布概率。图中灰度为 i 的像素分布概率 $p_i = n_i/N$,n_i 表示图像中灰度为 i 的像素数量,N 表示图像中像素总数。当阈值为 g 时,图像被分为前后景的概率 ψ_1 和 ψ_2 分别为 $\sum_{i=0}^{g} p_i$ 和 $\sum_{i=g+1}^{255} p_i$,m_1 和 m_2 为前后景像素灰度均值,分别表示如下:

$$m_1 = \frac{\sum_{i=0}^{k} i \times p_i}{\psi_1}, m_2 = \frac{\sum_{i=k+1}^{255} i \times p_i}{\psi_2} \tag{4}$$

基于 δ 过滤背景信息,获得分割后纹理图像 $\boldsymbol{I}_d$,构成目标检测网络的双通道输入,如图 3(b)所示。图中存在两处明显的像素密集区域,包括作业人员手部运动区域,车间作业台侧上方受弹性装置连接的闲置工具悬挂区域。基于帧差法区域分割后能充分抑制背景信息,为检测网络提供空间兴趣区域,引导对手持工具的定位。

2)基于全局注意力的空间感知模块

为提高模型对检测目标空间位置的感知,本文基于传统的 CBAM[22] 注意力机制,设计基于全局注意力的空间感知模块(Diff-CBAM),融合纹理图像 $\boldsymbol{I}_d$ 表征的空间兴趣区域信息。Diff-CBAM 模块如图 4 所示。

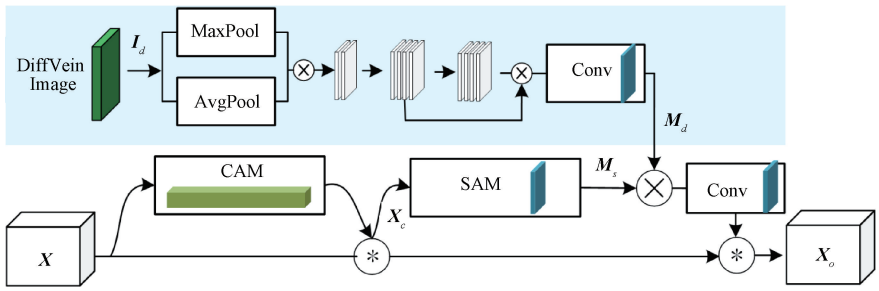


图 4 Diff-CBAM 模块

Fig. 4 Diff-CBAM module

对于输入特征图 $\boldsymbol{X} \in \boldsymbol{R}^{H\times W\times C}$,其中 H、W 和 C 分别是高、宽和通道数量,Diff-CBAM 首先通过通道注意力(channel attention module,CAM)模块捕捉跨通道信息,得到维度为 $C \times 1 \times 1$ 的通道特征权重 $\boldsymbol{M}_c$,与特征图相乘获得具有全局通道关键度的输出 $\boldsymbol{X}_c$:

$$\boldsymbol{X}_c = \boldsymbol{M}_c(\boldsymbol{X}) \cdot \boldsymbol{X} \tag{5}$$

通过空间注意力模块(spatial attention module,SAM)压缩 $\boldsymbol{X}_c$ 的通道信息,利用自适应平均池化 F^s_{avg} 和最大池化 $F^s_{\max}$,经卷积 $f^{7\times7}$ 变为通道维度为 1 的特征图,采用 Sigmoid 函数 δ 实现非线性激励,获得空间自注意力权重 $\boldsymbol{M}_s$:

$$\boldsymbol{M}_s = \delta(f^{7\times7}[F^s_{avg}(\boldsymbol{X}_c);F^s_{\max}(\boldsymbol{X}_c)]) \tag{6}$$

然后,提取 DiffVein 图像 $\boldsymbol{I}_d$ 中的手部运动空间信息,将 $\boldsymbol{I}_d$ 通过自适应最大池化和平均池化,池化效果如图 5 所示。

(a) DiffVein示意图 (a) DiffVein image (b) 平均池化效果 (b) Adaptive mean pool (c) 最大池化效果 (c) Adaptive max pool

图 5 DiffVein 图像及池化效果图

Fig. 5 Pool of DiffVein image

对池化特征图通过残差卷积模块 g 编码,再经过降维获得维度为 $H \times W \times 1$ 的包含运动信息的空间特征权重 $\boldsymbol{M}_d$:

$$\boldsymbol{M}_d = \delta(f^{7\times7}(g([F^s_{avg}(\boldsymbol{I}_d);F^s_{\max}(\boldsymbol{I}_d)]))) \tag{7}$$

拼接两个空间特征权重 $\boldsymbol{M}_s$ 和 $\boldsymbol{M}_d$,经过卷积和激励后融合为 $H \times W \times 1$ 的全局空间注意力 $\boldsymbol{M}_{sd}$,将其与经过通道注意力加权的特征层 $\boldsymbol{X}_c$ 相乘得到空间感知模块的输出特征图 $\boldsymbol{X}_o$:

$$\boldsymbol{X}_o = \boldsymbol{M}_{sd} \cdot \boldsymbol{X}_c = \delta(f^{3\times3}[\boldsymbol{M}_d;\boldsymbol{M}_s]) \cdot \boldsymbol{X}_c \tag{8}$$

3)基于坐标位置注意力的特征增强模块

为充分融合高层语义特征和包含手部运动信息的空间特征,提高对手部接触工具的定位能力,在特征金字塔不同层级间设计一类基于坐标位置注意力(coordinate attention module,CA)[23] 的特征增强模块(FE-CSP),模块如图 6 所示。

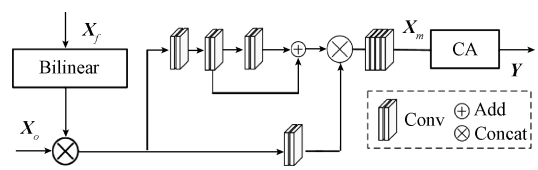


图 6 特征增强模块

Fig. 6 Feature enhance module

FE-CSP 由双线性插值上采样(bilinear upsample)、CSP 卷积模块、CA 注意力模块组成。双线性插值上采样结合被插值点周围多个特征值的特征信息,能够平滑高层语义特征的分辨率重构,减少边缘细节特征的丢失。利用 YOLOv5 中的 CSP 模块融合高层语义特征 $\boldsymbol{X}_f$ 和包含手部运动信息的全局空间特征 $\boldsymbol{X}_o$,获得融合特征 $\boldsymbol{X}_m$ 。通过 CA 注意力模块增强 $\boldsymbol{X}_m$ 中显著位置的特征响应,提高手持工具检测精度,CA 模块如图 7 所示。

CA 模块在通道注意力中嵌入位置信息,沿输入特征图的高 h 、宽 w 两个空间方向池化聚合特征,将特征图编码为 2 个维度为 $C \times H \times 1$、$C \times 1 \times W$ 的特征向量 $\boldsymbol{g}^h$ 和

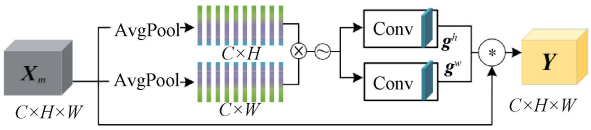


图 7　坐标位置注意力模块

Fig. 7　Coordinate attention module

$\boldsymbol{g}^w$。对于特征图 $\boldsymbol{X}_m$ 第 c 通道的上坐标位置为 (i,j) 的值 $\boldsymbol{x}_c(i,j)$，加权后的输出 $\boldsymbol{Y}$ 对应的值 $\boldsymbol{y}_c(i,j)$ 为：

$$\boldsymbol{y}_c(i,j)=\boldsymbol{x}_c(i,j)\times\boldsymbol{g}_c^h(i)\times\boldsymbol{g}_c^w(j) \tag{9}$$

2.2　基于 ESNet 的特征提取网络

通过构建双通道特征金字塔实现特征融合虽然提高了模型对目标位置空间的特征响应，但会产生额外的耗时。为了提高手持工具检测实时性，引入 ESNet 替换 YOLOv5 主干网络。ESNet 使用深度可分离卷积组成多尺度特征提取模块，采用通道划分（channel split）、通道重组（channel shuffle）和 Ghost 卷积思想，结合通道注意力机制，具有多尺度特征提取和高效率特征利用的优势。ESNet 如图 8 所示。

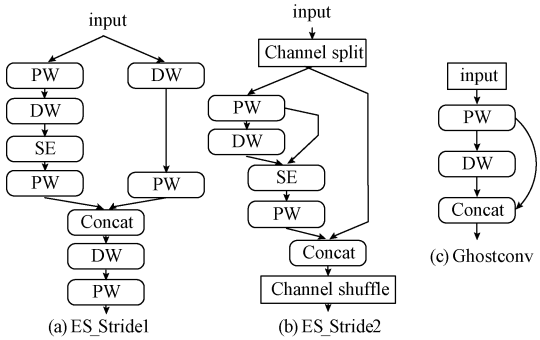


图 8　ESNet 主要模块

Fig. 8　Main module of ESNet

深度可分离卷积由逐通道卷积（depth wise convolution，DW）和逐点卷积（point wise convolution，PW）组成。DW 基于分组卷积实现特征通道上的空间变换，本文使用卷积核为 5×5 的 DW 卷积丰富感受野，PW 基于卷积核大小为 1×1 的点卷积，实现跨通道间的信息交流。在步长为 1 的模块（ES_Stride1）的残差连接中，通过改变 DW 卷积和 PW 卷积的先后顺序，实现多尺度提取特征。在步长为 2 的模块（ES_Stride2）中，利用通道稀疏连接方式降低卷积层共享参数使用 DW 和 PW 卷积组成 Ghost 卷积[19]，通过特征复用减少特征通道间存在的冗余信息。模块分别如图 8（a）、（b）、（c）所示。基于 ESNet 改进的主干网络如图 9 所示。改进主干网络和特征金字塔后的 YOLOv5 整体网络结构如图 10 所示。

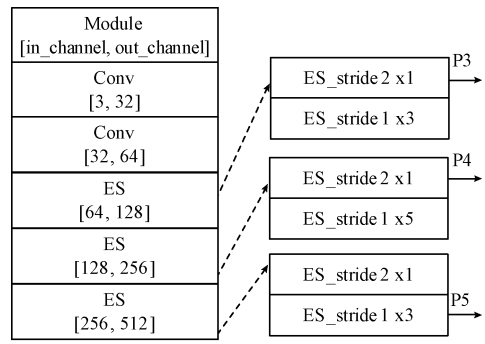


图 9　基于 ESNet 的主干网络结构

Fig. 9　Backbone network based on ESNet

2.3　基于 CutOut 的数据增强方法

针对作业环境中出现无关物件造成的局部元素变化，产生干扰信息，导致模型误检的问题，本文基于 CutOut 数据增强方法，通过添加随机噪声块增强训练样本，CutOut 方法如图 11（a）所示。CutOut 数据增强方法

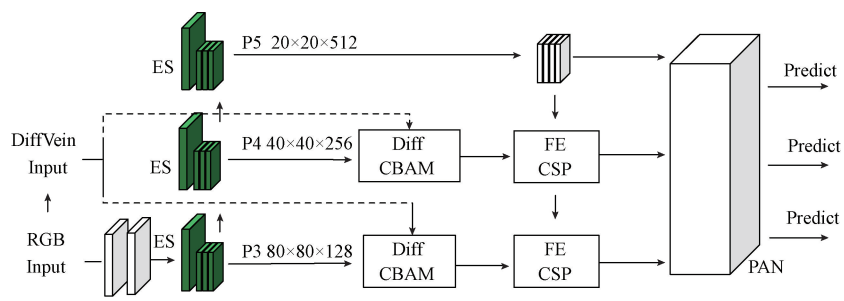


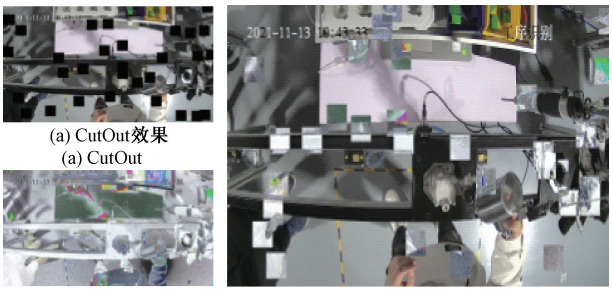
图 10　改进主干网络和特征融合的 YOLOv5 网络结构

Fig. 10　YOLOv5 network architecture diagram with improved backbone and feature fusion

虽然能够减少对小部分特定视觉特征的依赖，但可能过度删除目标区域，造成关键像素或完整目标被删除，丢失工具和手部接触部分语义信息，本文针对 Cutout 方法丢失上下文信息的问题，利用高斯噪声填充 CutOut 的随机块增强手持工具数据集。

高斯噪声对图像像素影响效果如图 11（b）所示。本文在图像局部添加随机噪声块，提高训练样本的像素复杂度，提升模型对场景局部元素变化的鲁棒性。同时保

留工具与手部接触部分的纹理信息,避免出现由于关键信息丢失,造成检测精度下降的问题。使用高斯噪声填充块的 CutOut 方法如图 11(c)所示。



(a) CutOut效果
(a) CutOut

(b) 高斯噪声效果
(b) Gauss noise

(c) 高斯噪声填充CutOut填充块效果
(c) CutOut with Gauss noise

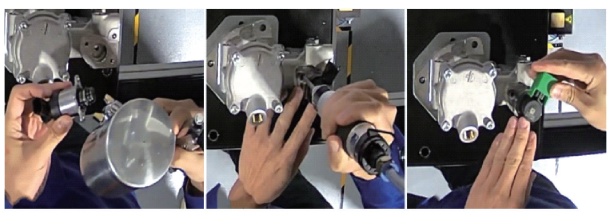
图 11 CutOut 和高斯噪声效果示意

Fig. 11 Demonstration of CutOut augment and Gauss noise effect

3 实验与分析

3.1 数据集和训练样本构建

本文利用对手持工具的检测来完成对喷油枪喷油,使用螺丝刀,安装绿色 Mprop 配件 3 个动作的工序识别,喷油枪 o1、螺丝刀 o2 和绿色 Mprop 配件 o3 为 3 类检测目标。数据集取自合作企业无锡威孚高科技集团有限公司的泵件装配过程。车间作业时使用工具和泵件装配操作台分别如图 12 和 13 所示。手持工具检测及其标注方式如图 14 所示,锚框中仅包含与手部接触的工具。数据集中共包含 3 608 张图片。将其按 6∶4 的比例划分为训练集和验证集,如表 1 所示。



(a) 喷油枪
(a) Fuel injection gun

(b) 螺丝刀
(b) Screwdriver

(c) 绿色Mprop配件
(c) Green Mprop parts

图 12 检测工具示意图

Fig. 12 Demonstration of detecting tools

表 1 各类别样本数量

Table 1 Numbers of samples in each category

Data set	o1	o2	o3
Train data	305	989	857
Val data	201	620	636

本文利用 YOLOv5 算法中的 Mosaic 数据增强方法来提升小目标检测精度,Mosaic 方法将 4 张图片拼接成 1 张,效果如图 15(a)所示。为了保证 Mosaic 增强后的装

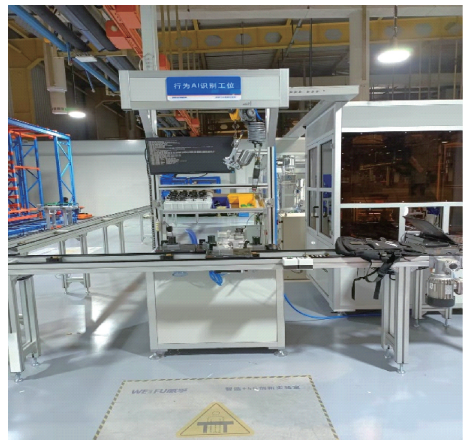
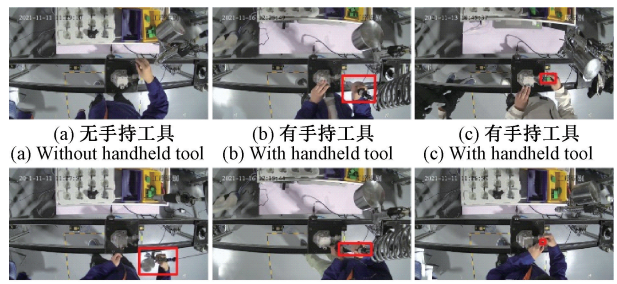


图 13 5G-AI 行为识别-工序识别作业台

Fig. 13 5G-AI Process recognition workstation



(a) 无手持工具
(a) Without handheld tool

(b) 有手持工具
(b) With handheld tool

(c) 有手持工具
(c) With handheld tool

(d) 喷油枪
(d) Fuel injection gun

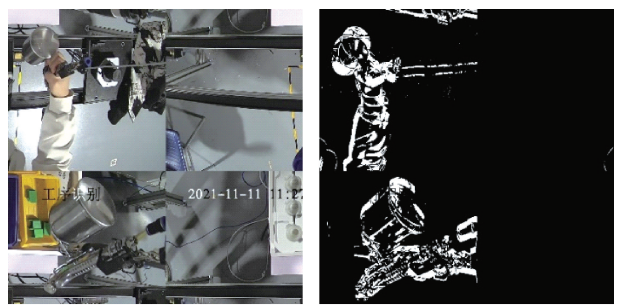
(e) 螺丝刀
(e) Screwdriver

(f) 绿色Mprop配件
(f) Green Mprop parts

图 14 数据集示意图

Fig. 14 Demonstration of the dataset

配图像和 DiffVein 图像区域对齐,在构建训练样本时,对输入的 DiffVein 图片进行 Mosaic 处理,实现效果如图 15(b)所示。



(a) RGB Image + Mosaic

(b) DiffVein Image + Mosaic

图 15 融合 Mosaic 增强的训练样本

Fig. 15 Train data with mosaic augment

3.2 实验环境与评价指标

服务器端实验环境配置如表 2 所示。对改进前后的网络进行 150epoch 训练,其中 Batch Size=32。采用随机梯度下降法优化,初始学习率设置为 0.01,使用余弦退火动态调整学习率,冲量大小为 0.9,权重衰减设置为

0.000 5。采用平均精确率(average precision,AP)、mAP(mean average precision)、精确率(precision,P)、召回率(recall,R)评价模型准确性[24],通过每秒钟检测帧数(Frame/s)评价模型实时性。

表 2　实验环境

Table 2　Experimental environment

类别	环境条件	类别	环境条件
CPU	E5-4650 V3@ 2.10 GHz	CUDA 版本	CUDA 11.1
显卡	NVIDIA Corporation GM200 [GeForce GTX TITAN X]	深度学习框架	Pytorch
内存	16 GB	运行环境	Linux
系统	Centos7	脚本语言	Python3.8

3.3　数据增强对比实验

为验证数据增强方法的效果,选择去除 Mosaic 数据增强的 YOLOv5s 模型(Model 1)作为基线模型(Baseline)进行对比实验,模型 2 和 3 分别通过 Mosaic 和 CutOut 完成数据增强。模型 4 融合 CutOut 和 Mosaic 方法。AP 值、精确率和召回率结果如表 3 所示。表中 o1、o2 和 o3 分别代表检测目标喷油枪、螺丝刀和绿色 Mprop 配件。

由表 3 可知,使用 Mosaic 方法增强数据,mAP 值较 Baseline 由 0.838 上升到 0.864,上升了 2.6%。检测喷油枪(o1)和绿色 Mprop 配件(o3)的 AP 值分别从 0.712 上升到 0.730,从 0.819 上升到 0.888,尤其是检测绿色 Mprop 配件这类较小目标上升了约 7%。表明针对手持工具检测,利用 Mosaic 数据增强方法能有效提升小目标的检测能力。

表 3　数据增强实验对比

Table 3　Comparison of data augment

Model	Augment	AP0.5				Precision	Recall
		o1	o2	o3	mAP		
1	—	0.712	0.982	0.819	0.838	0.855	0.843
2	Mosaic	0.730	0.973	**0.888**	0.864	0.912	0.840
3	CutOut	**0.757**	0.985	0.808	0.850	0.909	0.851
4	CutOut+Mosaic	0.749	**0.987**	0.887	**0.874**	**0.915**	**0.856**

使用 CutOut 方法增强数据,检测喷油枪和螺丝刀的 AP 值分别较 Baseline 从 0.712 提升到 0.757,从 0.982 提升到 0.985。mAP 由 0.838 提高到 0.850,提高了 1.2%,精确率由 0.855 提高到 0.909,提高了 5.4%。对图片添加随机噪声验证模型抗干扰能力,使用 CutOut 增强前后的模型检测效果如图 16 和 17 所示。改进前的模型将背景的绿色像素噪点块误检为绿色 Mprop 配件。改进后的模型提高了检测精度,并消除了绿色噪点块干扰下的误检现象。说明 CutOut 方法增强训练数据集能够有效提高检测模型的抗干扰能力,减少由于局部元素变化导致的误检问题。

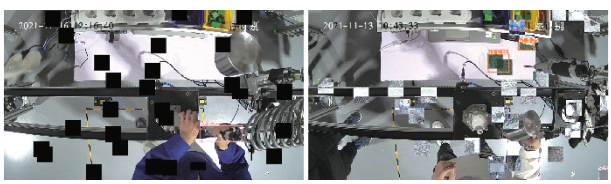


图 16　数据增强前检测图

Fig. 16　Detect results before CutOut augment

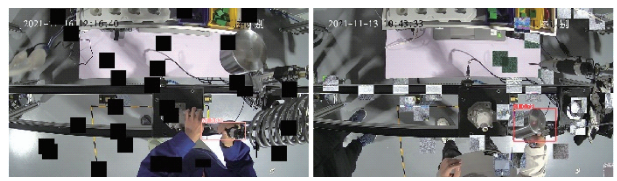


图 17　数据增强后检测图

Fig. 17　Detect results after CutOut augment

3.4　模块消融实验

为了测试融合运动空间信息的双通道输入在手持工具检测中的效果,在 YOLOv5 网络中嵌入 Diff-CBAM 模块,融合 DiffVein 纹理图像的空间特征,采用不同数据增强方法后进行对比实验,结果如表 4 所示。由表 4 可知,改进的双通道 YOLOv5s 使 mAP 值较改进前由 0.874 提升到 0.919,上升 4.5%。召回率由 0.856 提升到 0.897,上升 4.1%。同时对比表 3 和 4 可以发现,引入双通道前的各模型检测螺丝刀的平均精确率都维持在 0.97 以上,而检测喷油枪平均精确率都较低,原因是由于喷油枪结构较其他工具都更为复杂,作业时尺度和形体变换也较大,且喷油枪样本数量较少,相对螺丝刀和绿色 Mprop 配件更难检测。引入双通道后的模型有效提升了对喷油枪的检测效果,使检测喷油枪的 AP 值由 0.74 左右的上升到 0.86 左右,上升 12%,继而提高所有类别的 mAP。

改进双通道前后模型的 Precision-Recall 曲线变化如图 18 所示,改进后模型的 Precision-Recall 曲线积分面积有了明显增加,在提高喷油枪召回率的阈值下,能够保持更高的精确率。改进前后的检测效果如图 19 所示,改进

表 4 双通道输入消融实验对比

Table 4 Comparison of dual channel input ablation experiments

Model	Augment	AP0.5				Precision	Recall
		o1	o2	o3	mAP		
YOLOv5s	CutOut+Mosaic	0.749	**0.987**	0.887	0.874	0.915	0.856
YOLOv5s+Diff-CBAM	—	0.851	0.978	0.832	0.887	0.899	0.879
YOLOv5s+Diff-CBAM	Mosaic	0.863	0.978	0.882	0.908	0.917	0.888
YOLOv5s+Diff-CBAM	CutOut	0.865	0.982	0.836	0.894	0.920	0.885
YOLOv5s+Diff-CBAM	CutOut+Mosaic	**0.879**	0.983	**0.894**	**0.919**	**0.922**	**0.897**

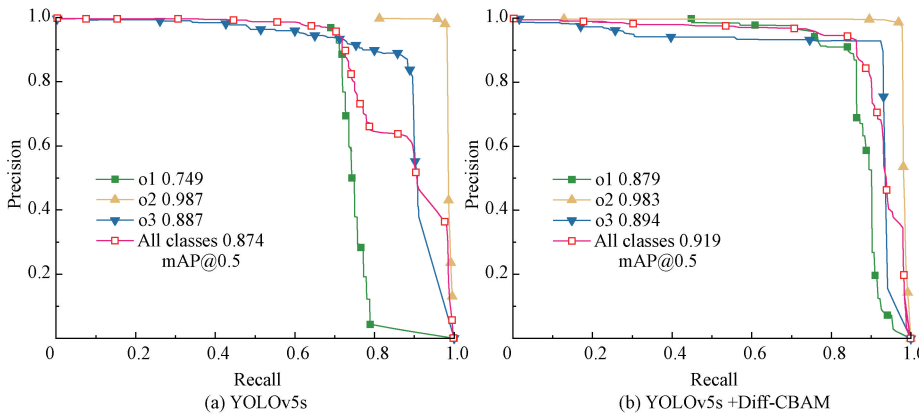
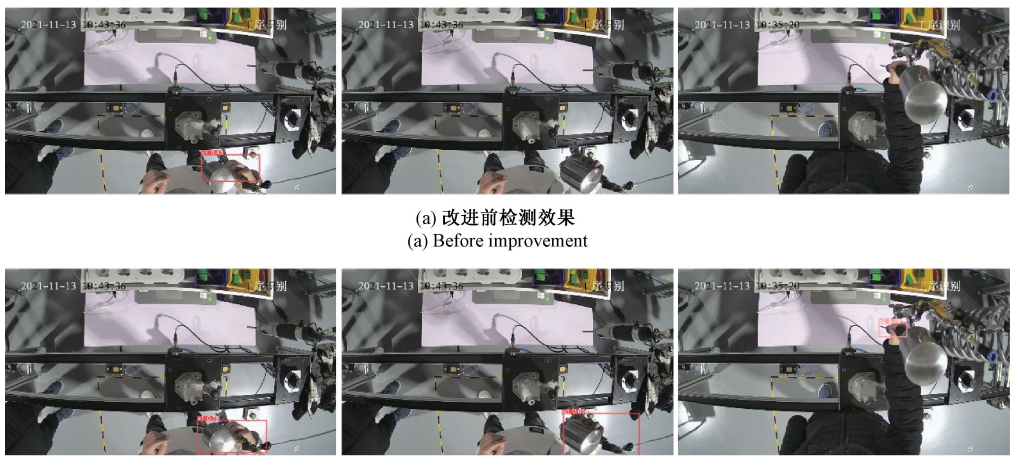


图 18 Precision-Recall 对比曲线

Fig. 18 Comparison curves of Precision-Recall



(a) 改进前检测效果

(a) Before improvement

(b) 改进后检测效果

(b) After improvement

图 19 双通道输入改进前后检测效果

Fig. 19 Detecting results before and after dual channel improvement

后的模型引入手部作业的运动信息,通过捕获前后帧中像素的变化,提高了模型对工具全局空间的定位能力,减少了环境光照、手部接触部分遮挡和替换操作人员和对工具检测的影响。针对喷油枪这类尺度变换较大的目标,有效改善网络丢失手部接触目标局部关键信息的问题,提高了检测精度。

为验证改进各模块等对模型的影响,对双通道模型中的主干网络和特征金字塔中的主要模块进行消融对比。通过替换 ESNet 主干网络和原 YOLOv5 主干网络;在特征增强模块 FE-CSP 中,基于 CSP 基础结构,分别加入双线性插值上采样和 CA 注意力机制完成消融实验。实验结果如表 5 所示。

表 5　网络主要模块消融实验对比

Table 5　Comparison of network main module ablation experiments

Diff-CBAM	ES	FE-CSP		AP0. 5				Precision	Recall	Frame/s
		Bilinear	CA	o1	o2	o3	mAP			
×	×	×	×	0. 749	**0. 987**	0. 887	0. 874	0. 915	0. 856	74. 9
×	√	×	×	0. 748	0. 986	0. 885	0. 873	0. 899	0. 868	**138. 8**
√	×	×	×	0. 879	0. 983	0. 894	0. 919	0. 922	0. 897	56. 5
√	√	×	×	0. 881	0. 985	0. 884	0. 917	0. 919	0. 884	86. 3
√	√	√	×	0. 885	0. 984	0. 888	0. 919	0. 922	0. 895	84. 8
√	√	×	√	0. 886	0. 981	0. 909	0. 925	0. 923	0. 907	78. 6
√	√	√	√	**0. 889**	0. 982	**0. 913**	**0. 928**	**0. 930**	**0. 908**	77. 4

由表 5 可知，通过 ESNet 替换主干网络，对比原 YOLOv5s 网络，在保持检测平均精确率的基础上，每秒检测帧数由 74. 9 上升到 138. 8。对比双通道 YOLOv5s 网络，每秒检测帧数由 56. 5 上升到 86. 3，并且 mAP 仅下降了 0. 2%，说明 ESNet 综合深度可分离卷积，Ghost 卷积思想和通道注意力针对轻量级检测任务，能够有效提高实时性。引入 FE-CSP 特征增强模块后，虽然每秒检测帧数从 86. 3 减少到 77. 4，略有下降，但在 mAP 和精确率、召回率指标上分别提高到了 0. 928 和 0. 930、0. 909，说明 FE-CSP 通过结合双线性插值上采样和 CA 注意力模块能够更充分传递特征，提升检测精度。实验证明各改进方法均有一定效果，与未使用 ESNet 和 FE-CSP 的双通道模型相比，在检测速度略有下降的情况下，检测喷油枪的 AP 值由 0. 879 上升到 0. 889，检测绿色 Mprop 配件的 AP 值由 0. 894 上升到 0. 913，mAP 由 0. 919 上升到 0. 928，权衡速度精度指标，通过 ESNet 和 FE-CSP 模块能够提升模型检测性能。

基于双通道网络，引入 ESNet 和 FE-CSP 后模型的 Precision-Recall 曲线变化如图 20 所示，可见，模型进一步提高了绿色 Mprop 配件和喷油枪在高召回率阈值下的检测精确率。改进前后的检测效果如图 21 和 22 所示，改进后模型能够更细致的拟合目标工具区域位置，减少细节特征丢失，有效提升检测精度。

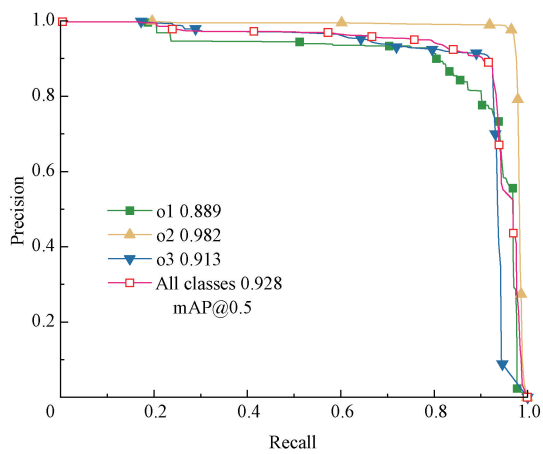


图 20　改进后模型 Precision-Recall 曲线

Fig. 20　Precision-Recall curve of improved model

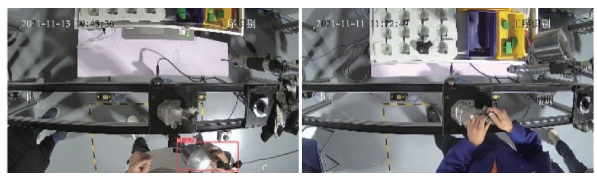


图 21　引入 ESNet 和 FE-CSP 模块前检测效果

Fig. 21　Detecting results before introducing ESNet and FE-CSP module

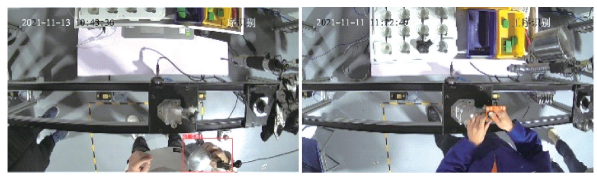


图 22　引入 ESNet 和 FE-CSP 模块后检测效果

Fig. 22　Detecting results after introducing ESNet and FE-CSP module

3. 5　目标检测算法性能对比实验

为验证本文改进算法的性能，分别选用传统的 Faster-RCNN、YOLOv4、YOLOv4-tiny、YOLOv5s 和改进算法进行推理速度和 mAP 等指标的测试对比，结果如表 6 所示。可以看出，本文提出的算法 mAP 值达到 92. 8%，与传统的 YOLOv5s 相比提高了 6. 4%，并且使得模型体积缩减至 8. 6 MB。由于提取运动信息存在额外耗时，运行速度不及 YOLOv4-tiny 算法，但综合检测平均精确率和速度指标，改进算法较 Faster-RCNN、YOLOv4 等传统目标检测算法有明显提升。

3. 6　装配工序步骤识别

结合本文提出的车间作业工具检测算法检测结果完成装配工序步骤识别过程如下，获得当前时刻 t 的手持工具类别和位置信息 $O(t)$ 后，利用工具位置信息，进一步过滤纹理图中远离工具的噪点，对时刻 t 的纹理图像

分类得到姿态信息 $M(t)$。连续帧检测图像和纹理图像分别如图 23 和 24 所示。基于连续帧的手持工具信息和姿态信息判别工序步骤。算法被集成在边缘计算设备上,获得较好效果。

表 6 目标检测算法性能对比

Table 6 Comparison of different object detection algorithm performance

Net	Augment	Backbone	Size/MB	Frame/s	mAP0.5/%
Faster-RCNN	—	ResNet50	108.7	8.4	83.5
YOLOv4	Mosaic	CSPDarkNet53	238.5	21.2	87.1
YOLOv4-tiny	Mosaic	CSPDarkNet53	22.3	**79.6**	84.9
YOLOv5m	Mosaic	CSPDarkNet53	42.2	68.3	88.8
YOLOv5s	Mosaic	CSPDarkNet53	14.1	74.9	86.4
Ours	CutOut+Mosaic	ESNet	**8.6**	77.4	**92.8**

图 23 连续帧工具检测图像

Fig. 23 Tool detect images of continuous frames

图 24 连续帧纹理图像

Fig. 24 DiffVein image of continuous frames

4 结 论

本文为区分装配工序步骤,针对泵件装配时使用的工具,提出了基于空间特征融合的车间作业工具检测算法方法。首先,为利用目标的运动信息引导检测网络定位目标,基于帧差法采用基于像素的时间差分,提取手部运动的空间信息;设计 Diff-CBAM 全局注意力模块,采用融合双通道空间特征的策略提高模型对检测目标空间位置的感知;基于 FE-CSP 模块增强深层语义信息和浅层空间信息的融合能力,提高手持工具检测精度。并利用 ESNet 替换主干网络,基于深度可分离卷积和特征高复用提升检测速度。最后,采用 CutOut 数据增强方法,借助其随机裁剪思想防止过拟合,提高抗干扰能力。实验结果表明,本文所提算法 mAP 达到 92.8%,较 YOLOv5s 原网络提高 6.4%,每秒检测帧数达到 77.4 帧。综合准确性和实时性,较 YOLOv5s、YOLOv5m、YOLOv4-tiny、Faster-RCNN 等传统目标检测算法有明显提升,满足车间作业快速、准确检测操作工具实现工序识别的需求。为进一步提高模型性能,后期可以进一步优化运动区域分割算法,提高算法运行速度和对工具的定位能力。扩充样本数量,增加作业工具的种类进行实验和改进,提高模型精度。

参考文献

[1] 王佳铖, 鲍劲松, 刘天元,等. 基于工件注意力的车间作业行为在线识别方法[J]. 计算机集成制造系统, 2021, 27(4): 1099-1107.
WANG J CH, BAO J S, LIU T Y, et al. Online method for worker operation recognition based on attention of workpiece [J]. Computer Integrated Manufacturing Systems, 2021, 27(4): 1099-1107.

[2] 吴培良, 隰晓珺, 杨霄,等. 一种基于联合学习的家庭日常工具功用性部件检测算法[J]. 自动化学报, 2019, 45(5): 985-992.
WU P L, XI X J, YANG X, et al. An algorithm for affordance parts detection of household tools based on joint learning [J]. Acta Automatica Sinica, 2019, 45(5): 985-992.

[3] 张静, 刘凤连, 汪日伟. 智能装配中基于 YOLOv3 的工业零件识别算法研究[J]. 光电子·激光, 2020, 31(10): 1054-1061.
ZHANG J, LIU F L, WANG R W. Research on industrial parts recognition algorithm based on YOLOv3 in intelligent assembly[J]. Journal of Optoelectronics · Laser, 2020, 31(10): 1054-1061.

[4] SHAN D, GENG J, SHU M, et al. Understanding human hands in contact at internet scale[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020: 9866-9875.

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

[6] RAGUSA F, FURNARI A, LIVATINO S, et al. The meccano dataset: Understanding human object interactions from egocentric videos in an industrial-like domain [C]. Proceedings of the IEEE/CVF Winter Conference on Applications of Coputer Vision, 2021: 1569-1578.

[7] 胡海苗, 沈柳青, 高立崑,等. 运动信息引导的目标检测算法[J]. 北京航空航天大学学报, 2022, 48(9): 1710-1720.

HU H M, SHEN L Q, GAO L K, et al. Object detection algorithm guided by motion information[J]. Journal of BeiJing University of Aeronautics and Astronautics, 2022, 48(9): 1710-1720.

[8] FEICHTENHOFER C, PINZ A, ZISSERMAN A. Convolutional two-stream network fusion for video action recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 27-30.

[9] 任克强, 高晓林. 基于五帧差和二维 Renyi 熵的运动目标检测[J]. 电子测量与仪器学报, 2015, 29(8): 1179-1186.
REN K Q, GAO X L. Moving object detection based on five-frame difference and two-dimensional Renyi entropy[J]. Journal of Electronic Measurement and Instrumentation, 2015, 29(8): 1179-1186.

[10] 周永福, 李文龙, 胡冉冉. 多尺度特征融合的双通道 SSD 行人头部检测算法[J]. 激光与光电子学进展, 2021, 58(24): 383-394.
ZHOU Y F, LI W L, HU R R. Two-channel SSD pedestrain head detection algorithm based on multi-scale feature fusion[J]. Laser & Optoelectronics Progress, 2021, 58(24): 383-394.

[11] 吴萌萌, 张泽斌, 宋尧哲,等. 基于自适应特征增强的小目标检测网络[J/OL]. 激光与光电子学进展: 1-14[2022-12-02].
WU M M, ZHANG Z B, SONG Y ZH, et al. Small object detection network based on adaptive feature enhancement[J/OL]. Laser & Optoelectronics Progress: 1-14[2022-12-02].

[12] DEVRIES T, TAYLOR G W. Improved regularization of convolutional neural networks with cutout[J]. arXiv preprint, 2017, arXiv: 1708. 04552.

[13] 史朋飞, 韩松, 倪建军,等. 结合数据增强和改进 YOL Ov4 的水下目标检测算法[J]. 电子测量与仪器学报, 2022, 36(3): 113-121.
SHI P F, HAN S, NI J J, et al. Underwater object detection algorithm combining data enhancement and improved YOLOv4[J]. Journal of Electronic Measurement and Instrumentation, 2022, 36(3): 113-121.

[14] ZHONG Z, ZHENG L, KANG G, et al. Random erasing data augmentation[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 13001-13008.

[15] 伍济钢, 成远, 邵俊,等. 基于改进 YOLOv4 算法的 PCB 缺陷检测研究[J]. 仪器仪表学报, 2021, 42(10): 171-178.
WU J G, CHENG Y, SHAO J, et al. A detect detection method for PCB based on the improved YOLOv4[J]. Cninese Journal of Scientific Instrument, 2021, 42(10): 171-178.

[16] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. YOLOv4: Optimal speed and accuracy of object detection[J]. arXiv preprint, 2020, arXiv: 2004. 10934.

[17] HOWARD A G, ZHU M, CHEN B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint, 2017, arXiv: 1704. 04861.

[18] 侯维岩, 靳东安, 王高杰,等. 基于嵌入式系统的智能售货柜目标检测算法[J]. 电子测量与仪器学报, 2021, 35(10): 217-224.
HOU W Y, JIN D AN, WANG G J, et al. Object detection of intelligent vending cabinet via embedded system[J]. Journal of Electronic Measurement and Instrumentation, 2021, 35(10): 217-224.

[19] HAN K, WANG Y, TIAN Q, et al. Ghostnet: More features from cheap operations[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020: 1580-1589.

[20] 朱书勤. 基于注意力融合网络的 RGB-D 目标检测算法[J]. 电子测量技术, 2021, 44(9): 110-115.
ZHU SH Q. RGB-D target detection algorithm based on attention fusion network[J]. Electronic Measurement Technology, 2021,44(9): 110-115.

[21] YU G, CHANG Q, LV W, et al. PP-PicoDet: A better real-time object detector on mobile devices[J]. arXiv preprint, 2021, arXiv: 2111. 00902.

[22] WOO S, PARK J, LEE J Y, et al. Cbam: Convolutional block attention module[C]. Proceedings of the European Conference on Computer Vision, 2018: 3-19.

[23] HOU Q, ZHOU D, FENG J. Coordinate attention for efficient mobile network design[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2021: 13713-13722.

[24] 侯学良, 单腾飞, 薛靖国. 深度学习的目标检测典型算法及其应用现状分析[J]. 国外电子测量技术, 2022, 41(6): 165-174.
HOU X L, SHAN T F, XUE J G. Analysis of typical target detection algorithm based on deep learning and its application status[J]. Foreign Electronic Measurement Technology, 2022, 41(6): 165-174.

作者简介

王呈,2014 年于北京交通大学获得博士学位,现为江南大学物联网工程学院副教授,主要研究方向为非线性系统建模与控制、机器学习与数据挖掘。
E-mail:wangc@ jiangnan. edu. cn

Wang Cheng received his Ph. D. degree from Beijing Jiaotong University in 2014. He is currently an associate professor at Jiangnan University. His main research interests include modeling and control of nonlinear system, machine learning and data mining.