DOI: 10.13382/j.jemi.B2407413

基于改进 DETR 的智慧车间人员典型行为识别算法*

何赟泽[1]　谯灵俊[1]　王洪金[1]　马　刚[2]　王耀南[1]
(1. 湖南大学电气与信息工程学院　长沙　410082;2. 湖南红太阳新能源科技有限公司　长沙　410205)

摘　要:生产车间环境复杂,设备众多且人员活动具有高度自主性和不确定性,传统的人工观测方式在面对海量监控数据时,难以实现高效的实时管控。为提高车间人员行为的自动化监测水平,保障生产安全,提出一种基于改进 DETR 的行为识别算法。通过智慧车间的实地调研,采集多种工作行为及异常行为数据,构建车间红外行为数据集,并在此基础上设计改进算法。针对原始算法的不足,引入相对位置编码,并采用空间调制共同注意力机制,旨在提升网络对全局特征中待检测物体的定位精度。此外,通过引入待检测物体的高斯分布权重,使网络解码器更加高效地识别行为特征。实验结果表明,改进后的算法在自建数据集上的识别精度相比原始算法提高了 6.97%,并在公开数据集上同样表现出色。该改进方法不仅为车间人员行为的监控提供了更加高效的解决方案,也为智慧车间的自动化与智能化发展提供了有力的技术支持。

关键词: DETR;行为识别;注意力机制;深度学习;智慧车间;红外数据集
中图分类号: TP391　　**文献标识码**: A　　**国家标准学科分类代码**: 520.20

Typical behavior recognition algorithm for intelligent workshop personnel based on improved DETR

He Yunze[1]　Qiao Lingjun[1]　Wang Hongjin[1]　Ma Gang[2]　Wang Yaonan[1]
(1. School of Electrical and Information Engineering, Hunan University, Changsha 410082, China;
2. Hunan Red Solar New Energy Science and Technology Co., Ltd, Changsha 410205, China)

Abstract: The production workshop environment is complex, with numerous equipment and highly autonomous and uncertain personnel activities. Traditional manual observation methods are difficult to achieve efficient real-time control when facing massive monitoring data. To improve the automation monitoring level of workshop personnel behavior and ensure production safety, a behavior recognition algorithm based on improved DETR is proposed. Through on-site research in the smart workshop, various work behavior and abnormal behavior data were collected to construct an infrared behavior dataset for the workshop, and an improved algorithm was designed based on this. In response to the shortcomings of the original algorithm, relative position encoding is introduced and a spatial modulation joint attention mechanism is adopted to improve the network's localization accuracy of the object to be detected in the global features. In addition, by introducing Gaussian distribution weights of the object to be detected, the network decoder can more efficiently recognize behavioral features. The experimental results show that the improved algorithm has improved recognition accuracy by 6.97% on self built datasets compared to the original algorithm, and also performs well on public datasets. This improvement method not only provides a more efficient solution for monitoring the behavior of workshop personnel, but also provides strong technical support for the automation and intelligent development of smart workshops.

Keywords: DETR; behavior recognition; attention mechanism; deep learning; smart workshop; infrared dataset

收稿日期: 2024-04-07　　Received Date: 2024-04-07
*基金项目:湖南省重点研发计划(2022GK2012)、湖南省科技创新领军人才(2023RC1039)、湖南省自然科学基金杰出青年基金项目(2022JJ10017)资助

0　引　言

在工厂产业升级和智能化的推动下,不少企业引入机器人和人工智能技术以提高生产效率和适应快速变化的市场需求。随着工人活动区域的扩展和行为多样化,传统的安全监控方法既费时又容易疏漏。因此,推进车间数字化与智能化是实现智能制造的重中之重[1]。通过视频监控和人工智能技术的结合,可以实现对工人行为的实时分析和识别。例如,对于涉及明火、抽烟、打电话等危险行为,系统能够迅速做出反应,从而降低事故发生的概率。与传统的人工分析方法相比,这种智能化的安防管控方法不仅更加高效,而且减少了疏漏,为工厂安全提供了更可靠的保障[2]。

行为识别算法可分为基于单帧图像的目标检测方法和基于视频的时空行为检测方法。时空行为检测的缺点在于模型结构较为复杂,必须同时考虑空间和时间的特征,这使得算法的计算量大幅增加,难以达到实时检测的目的。因此,本文将采用目标检测算法,通过学习静态图像特征,输出目标在图像中的位置及类别[3],解决行为识别问题。目前基于深度学习的目标检测算法分为基于锚框或点的单阶段检测器和基于区域的多阶段检测器[4]。2014 年 Girshick 等[5]提出 R-CNN,将神经网络和目标检测联系起来,利用卷积神经网络优秀的特征提取和分类能力,将目标检测转换为候选框问题,并使用线性回归和支持向量机对检测框进行分类修正。在 VOC2012 数据集上,这一方法实现了 53.3%的检测精度。2015 年 Fast R-CNN[6]被提出,实现了同时训练检测器和边界回归器,解决了特征提取中数据冗余问题,进一步提升检测速度。随后 Ren 等提出了双阶段检测器 R-CNN 系列的最快版本,即 Faster R-CNN[7]网络,改变了以往候选区域生成的方式,使用区域生成网络(region proposal network,RPN)生成检测框,在特征提取、分类和回归融合的同时生成候选区域,达到端到端检测的目的,在 VOC2012 数据集上检测精度达到 70.4%,并且检测速度获得大幅提升。2015 年 Redmon 等[8]提出了单阶段检测器 YOLO(you only look once),与 Faster R-CNN 反复训练 RPN 网络和 Fast R-CNN 网络不同,YOLO 仅用一个卷积网络实现端到端目标检测,同时完成物体的定位和分类工作,大幅提升检测速度,达到 45 帧每秒。Liu 等[9]提出了 SSD(single shot multibox detector),采用特征金字塔来提取多个尺度的特征图,大大提升了检测精度。YOLO 系列网络为单阶段目标检测的代表模型,对于不同的检测任务,学者们对其进行改进以满足不同的检测要求。罗国富等[10]将 YOLOv5 网络的特征融合网络以及输出层删除,将得到的模型结构化剪枝后进行知识蒸馏,提升了智能车间工人不安全行为的检测精度;杜闯等[11]用 PP-LCNet 替换 YOLOv5 的原 backbone 部分实现模型轻量化,使模型更容易迁移到嵌入式设备,并在精度与预测速度方面有所提升;Cheng 等[12]通过 CycleGAN 生成合成图像扩展数据集,并在 YOLOv5s 最后一个编码器的前馈网络末端集成了卷积块注意模块,以捕获特征图中更复杂的细节。2020 年,CARION 提出 DETR(detection transformer)[13],与基于区域的卷积 R-CNN 等方法相比,DETR 的不同之处在于它用无序集合预测来解决目标检测问题。具体来说,图像首先通过卷积神经网络进行特征提取得到特征序列。随后在 Transformer [14]中,编、解码器输出具有设置固定长度的无序集合。得到的集合中包含物体类别和坐标。DETR 在大目标检测上效果好,总体性能与精调后的 Faster R-CNN 相当。DETR 将本属于自然语言处理(natural language processing, NLP)领域的 Transformer 跨界到计算机视觉领域[15],开创了目标检测的新范式[16-17]。

为解决可见光图像易受光照、阴影的影响,在低照明度条件下无法准确识别动作,监控画面被遮挡导致的行为识别算法准确率低、召回率差等问题,本文提出一种基于 DETR 的红外行为识别算法。改进原有网络的绝对位置编码和注意力机制,引入相对位置编码学习像素与像素之间的相对距离关系,并且采用空间调制共同注意力机制将目标的查询向量调整到目标中心位置附近。实验证明,改进方法可以实现对车间典型行为的高精度,自动化检测,对于行为识别的研究具有积极意义。

1　原始算法

DETR 是基于 Transformer 编解码器的端到端目标检测算法。与之前的目标检测算法如 YOLO 相比,它不需要先验框等先验知识和约束,也减少了非极大值抑制(non-maximum supression,NMS)消除冗余的边界框等后处理步骤。

DETR 的网络结构简洁,分工明确。由提取图像特征的骨干网络、基于 Transformer 的编、解码器和预测分类头 4 部分组成,如图 1 所示。输入一个 Batch 的数据集图片,由 ResNet50 或者其他 CNN 主干网络进行特征提取,提取后的特征再通过 1x1 的卷积改变通道数后送入 Transformer 中的编解码器中。encoder 阶段在特征图上进行全局分析,便于网络更好的提取不同位置不同物体之间的相互关系。DETR 在 decoder 阶段实现了目标检测任务中目标位置和编码器特征的有效融合过程。在预测模块中,前馈网络(feed forward network,FFN)中的两个全连接层将编码器和解码器之间的注意力层输出进行处理,最终生成目标检测结果。这两个接收来自注意力层

输出的全连接层被称为 class predictor 和 box predictor。class predictor 为每个可能的目标类别输出一个分数,以表示其属于该类别的概率。在训练期间,模型使用交叉熵损失来最小化这些预测与真实类别之间的差异;box predictor 对每个检测框的位置和大小进行预测。位置预测输出左上角和右下角坐标的偏移量。模型使用平滑 L1 损失来最小化这些预测与真实边界框之间的差异。这两个全连接层并行工作,并且在生成最终的结果时合并输出,形成一个元组,其中包含预测的类别置信度和边界框坐标。

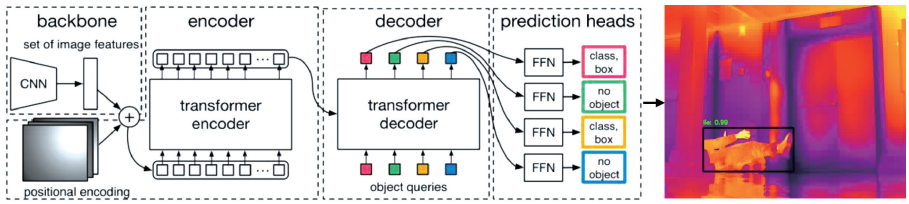


图 1 DETR 网络结构

Fig. 1 DETR framework

对于图像任务,Transformer 的自注意力操作可以获得更大范围的全局信息,相比 CNN 要不断堆叠卷积层来扩大感受野更具优势。DETR 将检测问题视为集合预测,寻找目标与全局上下文的关系,直接输出最终的目标结果。

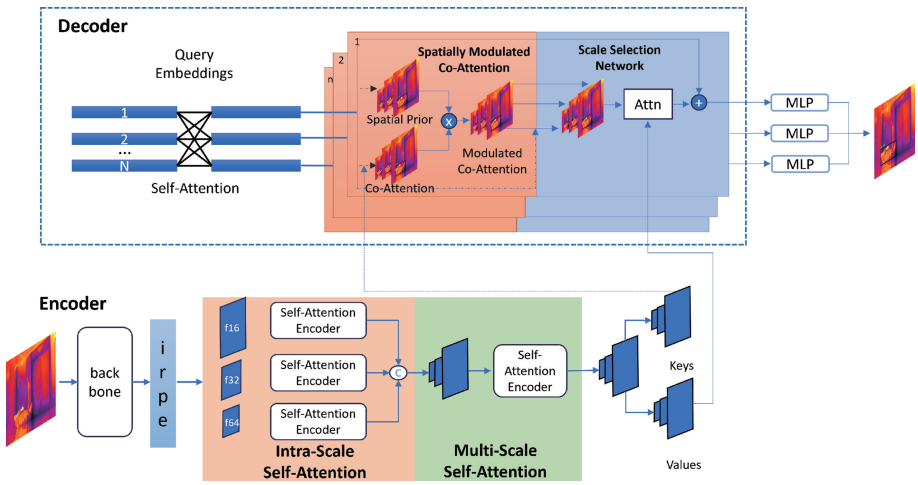


图 2 本文网络框架

Fig. 2 Framework of ours

2 改进算法

DETR 中的绝对位置编码虽然能使输入图像中的每个像素位置都能获得独立的表达,但是它无法利用像素之间的相对距离。这种编码方式忽略了像素点之间的相互联系,可能会影响模型在处理某些任务时的表现,尤其是在需要捕捉像素间复杂关系的场景中。因此本文采用一种相对位置编码(image relative position encoding, IRPE)[18-19]来捕获 token 之间依赖关系。同时采用空间共同注意力机制(spatially modulated co-Attention, SMCA)[20],在 Transformer 的解码器中引入目标的高斯分布权重,使改进后的 SMCA 能够更高效地从全局特征中定位待检测物体的特征,提升预测效果。

2.1 相对位置编码

Self-Attention 将 Query、Key 和 Value 一一映射到输出,对于一个输入序列 $x=(x_1,\cdots,x_n)$,自注意力计算一个输出序列 $z=(z_1,\cdots,z_n)$。公式如式(1)~(3)所示。

$$z_i=\sum_{j=1}^{n}a_{ij}(x_j\boldsymbol{W}^V) \tag{1}$$

$$a_{ij}=\frac{\exp(e_{ij})}{\sum_{k=1}^{n}\exp(e_{ik})} \tag{2}$$

$$e_{ij}=\frac{(x_i\boldsymbol{W}^Q)(x_j\boldsymbol{W}^K)^{\mathrm{T}}}{\sqrt{d_z}} \tag{3}$$

式中:a_i 为经 softmax 计算后的权重系数,$\boldsymbol{W}^{Q,K,V}$ 为每层唯一的参数矩阵。原始 Self-Attention 采用了绝对位置编

码,并添加绝对位置编码输入到 token。

$$x_i = x_i + p_i \tag{4}$$

式中:x_i 为输入 token,p_i 为位置编码,绝对位置编码有多种选择,例如使用不同频率的正弦余弦函数的固定编码,或通过训练参数实现的可学习编码。对于某些任务,元素的相对顺序或距离对结果有重要影响。因此,考虑输入 token 之间的相互关系,引入相对位置编码方法。

$$z_i = \sum_{j=1}^{n} a_{ij}(x_j \boldsymbol{W}^V + \boldsymbol{p}_{ij}^v) \tag{5}$$

$$e_{ij} = \frac{(x_i \boldsymbol{W}^Q + \boldsymbol{p}_{ij}^Q)(x_j \boldsymbol{W}^K + \boldsymbol{p}_{ij}^K)^{\mathrm{T}}}{\sqrt{d_z}} \tag{6}$$

这种方法将不同输入元素 x_i 和 x_j 之间的相对位置编码为向量 $\boldsymbol{p}_{ij}^V$,$\boldsymbol{p}_{ij}^Q$,$\boldsymbol{p}_{ij}^K$,嵌入到自注意模块中。但并未考虑编码信息能否独立于输入 token 来学习。基于此,考虑与 Query、Key 和 Value 的交互,改变式(6)引入新的相对位置编码[19]。

$$e_{ij} = \frac{(x_i \boldsymbol{W}^Q)(x_j \boldsymbol{W}^K)^{\mathrm{T}} + b_{ij}}{\sqrt{d_z}} \tag{7}$$

$$b_{ij} = (x_i \boldsymbol{W}^Q)(\boldsymbol{r}_{ij}^K)^{\mathrm{T}} + (x_i \boldsymbol{W}^K)(\boldsymbol{r}_{ij}^Q)^{\mathrm{T}} \tag{8}$$

式中:$\boldsymbol{r}_{ij}$ 为可训练向量,b_{ij} 有多种变体,例如式(8)为 query 和 key 的相对位置编码。改进位置编码后的自注意力计算流程如图 3 所示。

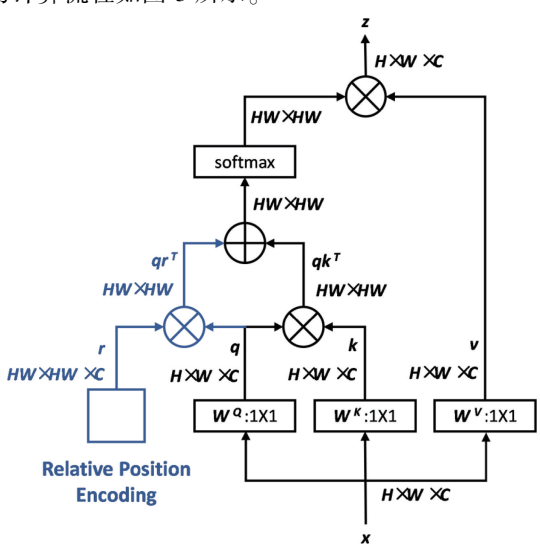


图 3　加入相对位置编码后的自注意力机制

Fig. 3　Self-Attention after adding relative position encoding

输入图像分辨率为 640×512,像素点数量巨大,导致两像素点间的相对位置坐标 i 和 j 的取值范围广。为了减少模型训练时的计算开销,建立分段函数对不同距离的像素点对进行处理,将相对位置转为整数,实现不同的位置像素之间共享编码。该函数如式(9)所示。

$$f(x) = \begin{cases} [x], |x| \leqslant a \\ sign(x) \times \min(\beta, [a + \dfrac{\ln(|x|/a)}{\ln(\gamma/a)}(\beta - a)]), |x| > a \end{cases} \tag{9}$$

式中:[·]是一个四舍五入运算,$sign(\,)$ 决定输出的符号,即输入为正返回 1,输入为负返回-1。a 决定分段点,β 控制$[-\beta,\beta]$范围内的输出,γ 调节对数部分的曲率。

2.2　空间调制共同注意力机制

与传统的基于局部特征的目标检测算法不同,DETR 利用一系列目标查询向量与图像的全局特征进行交互。该方法基于注意力机制,自适应地从图像的不同位置提取物体特征,从而预测物体的边界框坐标及其类别。然而,在这一过程中,与每个目标查询向量交互的特征区域可能包含大量背景或其他无关物体。因此,DETR 的解码器需要更长的训练时间,以使目标查询向量能够更准确地定位物体。

为了解决上述问题,对原有的共同注意力机制进行改进。在编码器中引入对多尺度特征的编码。通过层内自注意力机制(intra-scale self-attention)和层间自注意力机制(multi-scale self-attention)。并在解码器中引入空间调制共同注意力模块,自适应地选择合适尺度的特征,从而高效地编码图像的多尺度信息,提高检测精度。相比于全局特征的查询,改进后的模块将每个物体查询向量的搜索范围聚集到物体中心附近,从而加速查询。即对于每一个给定的物体查询向量,对物体的中心位置以及比例进行预测,生成二维类高斯权重图,预测公式如式(10)、(11)所示。

$$c_h^{\mathrm{norm}}, c_w^{\mathrm{norm}} = sigmoid(\mathrm{MLP}(\boldsymbol{O}_q)) \tag{10}$$

$$s_h, s_w = \mathrm{FC}(\boldsymbol{O}_q) \tag{11}$$

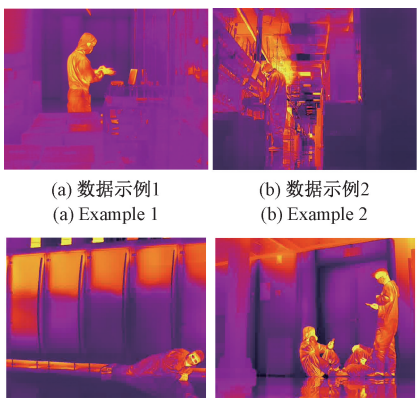
其中,c_h^{norm}、c_w^{norm} 为归一化中心坐标的初始预测,$\boldsymbol{O}_q$ 为查询向量。用预测值动态估计物体的尺度 s_h、s_w 创建二维类高斯分布,然后与共同注意力模型中的注意力矩阵结合,得到空间调制的多头注意力矩阵,强调查询对象中心附近的特征。因此,将每个物体查询向量在共同注意力机制中的搜索范围动态调整到物体中心附近的一定距离内,更好的解决自然世界中复杂目标的长宽比问题,充分学习大目标的信息并且抑制小目标背景信息的影响,从而更全面地提取物体特征。

3　实验

3.1　实验数据集与实验配置

本实验采集设备分辨率为 640×512,帧率 30 fps,在公司光伏智慧车间内进行数据采集。包括工作区域(包含工作设备)、无遮挡宽阔区域等多种拍摄场景,记录了

单人单种行为和多人多种行为。采集到的数据为视频文件,对视频进行抽帧处理。使用 DuplicateCleaner 软件对抽帧后的相似度大于 90%的图片去重,最终经人工筛选得到原始数据集图片。数据集采用人工标注的方式,在审核修正后获得红外数据集 5 913 张。行为类别中危险异常行为包括玩手机、接打电话、摔倒、打架、吸烟、持刀危险以及 3 种工作行为。按照训练集比测试集 4∶1 的比例随机抽取训练集和测试集,得到训练集 4 788 张,验证集 1 125 张,数据集构成如表 1 所示,数据示例如图 4 所示。



(a) 数据示例1 (b) 数据示例2
(a) Example 1 (b) Example 2
(c) 数据示例3 (d) 数据示例4
(c) Example 3 (d) Example 4

图 4 数据示例

Fig. 4 Sample data

表 1 数据集组成

Table 1 Composition of the dataset

类别	标签名	图像数量/张	标签数量/个
玩手机	playphone	881	1 493
接打电话	call	234	261
摔倒	lie	105	105
打架	fight	558	558
吸烟	smoke	906	1 042
持刀危险	danger	733	736
工作行为 1	touch_silicon	149	149
工作行为 2	operate	38	38
工作行为 3	pick	56	56
空白样本	——	1 128	——

本文实验采用的硬件配置为 Intel Xeon(R) CPU E5-2678 v3 @ 2.50GHZ × 48,显卡为 NVIDIA GeForce RTX3090。软件环境为 CUDA11.3,操作系统为 Windows10。网络模型基于 Pytorch1.11.0 框架搭建,Python 版本为 3.8。对不同的网络模型均训练 200 轮。

3.2 评价指标

在目标检测中,通常采用以下多个评价指标评估模型预测效果。

1)混淆矩阵(confusion matrix):以 N 行 N 列的矩阵形式表示数据集的真实值与模型预测值的对应关系,也称为误差矩阵,如图 5 所示。

混淆矩阵		真实值	
		positive	negative
预测值	positive	TP	FP
	negative	FN	TN

图 5 混淆矩阵

Fig. 5 Confusion matrix

其中,TP(真阳性)表示真实值为正,预测值也为正的情况;TN(真阴性)表示真实值与预测值都为负的情况,以上两种属于预测正确的情况。FP(假阳性)表示真实值为负,预测值为正的情况;FN(假阴性)表示真实值为正,预测值为负的情况,这两种代表预测错误。

2)精确度(Precision):真实值为正预测值也为正的部分占所有检测网络认为是正类的比例,公式如式(12)所示。

$$Precision = TP/(TP + FP) \tag{12}$$

3)召回率(Recall):在真实值为正的样本中被预测为正样本的概率,如式(13)所示。

$$Recall = TP/(TP + FN) \tag{13}$$

4)平均精度均值(mean average precision,mAP):将所有类别平均精度值综合加权平均而得到。

5)浮点运算次数(giga floating-point operations per second,GFLOPs):每秒 10 亿次的浮点运算数,用于衡量算法模型的计算量。

3.3 消融实验

为评估不同设计的模块组合对算法性能的优化程度,对同样的预训练权重迁移学习,进行消融实验。DETR 在加入不同的改进策略后实验结果如表 2 所示。第一列为原始 DETR 网络,在未加任何改进措施的情况下 mAP 为 87.01%;在此基础上加入相对位置编码,mAP 提升 2.13%,精确度有所提升,召回率提升 1%,计算量稍有增大;加入空间调制共同注意力机制后,mAP 提升 4.08%,精确度提升 4.13%,召回率提升 2.77%,计算量有所增大;最后使用所有的改进策略总共提升 6.97%,达到 93.98%,精确度达到 95%,召回率为 91.22%,各项指标均为最高。

对同一组验证集进行检测画出混淆矩阵,结果如图 6、7 所示。可以看出改进算法能够有效减少背景误识别,对某些行为的错检也有改善。

为验证网络模型和改进模块在非自建数据集上具有同样的有效性,在公开数据集 Stanford40 上进行相关实验。Stanford40 数据集包括 40 种行为,共计 9 532 张图

表 2　消融实验结果

Table 2　Results of ablation test

方法	第 1 组	第 2 组	第 3 组	第 4 组
DETR	√	√	√	√
irpe		√		√
SMCA			√	√
Precision	88. 35	89. 14	92. 48	95. 00
Recall/%	85. 34	86. 34	88. 11	91. 22
mAP/%	87. 01	89. 14	91. 09	93. 98
GFLOPs	101	104	152	154

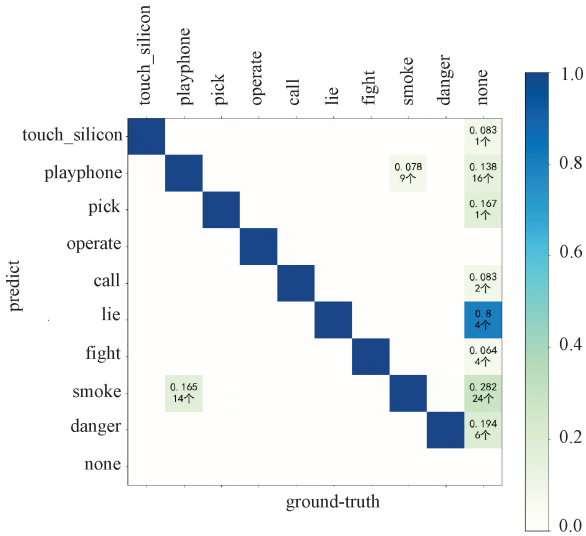


图 6　DETR 混淆矩阵

Fig. 6　Confusion matrix of DETR

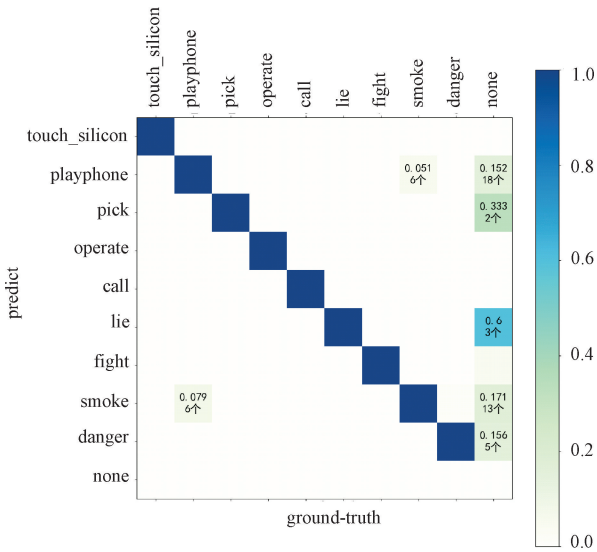


图 7　本文方法的混淆矩阵

Fig. 7　Confusion matrix of ours

片。与红外数据集相比,该数据集分辨率良好。每个行为类都有 100~200 张数量不等的图像,是目前已知的最大规模的静态行为识别数据集[21]。在相同的参数设置下进行实验,结果如表 3 所示。实验结果表明,在公开数据集上,本文的改进策略在上述评估指标上均有提升。

表 3　消融实验结果

Table 3　Results of ablation test

方法	第 1 组	第 2 组	第 3 组	第 4 组
DETR	√	√	√	√
irpe		√		√
SMCA			√	√
Precision	86. 36	86. 64	87. 20	88. 20
Recall/%	57. 56	57. 61	59. 17	59. 21
mAP/%	69. 93	71. 74	72. 61	73. 45

3. 4　对比试验

为验证本文算法的有效性,设置相同的实验配置,利用 Faster R-CNN、SSD、YOLOv5s 主流的单阶段目标检测模型以及百度 Paddledetection 中行为识别模型 PP-human 对自建数据集训练和测试,实验结果如表 4 所示。改进算法的平均精度均值、召回率以及精确度在所测试算法中最高,验证了改进模型的可行性。

表 4　对比实验结果

Table 4　Results of comparative experiments

算法	Precision/%	Recall/%	mAP/%
本文	95. 00	91. 22	93. 98
YOLOv5s	93. 51	89. 34	92. 43
SSD	79. 53	76. 22	78. 87
Faster R-CNN	92. 68	89. 34	91. 26
PP-human	84. 24	86. 00	84. 31

3. 5　实验结果可视化分析

使用融合算法在服务器上对数据集进行测试,各类别检测结果如图 8 所示。各类行为检测精度均在 85%以上,平均精度达到 93. 98%。

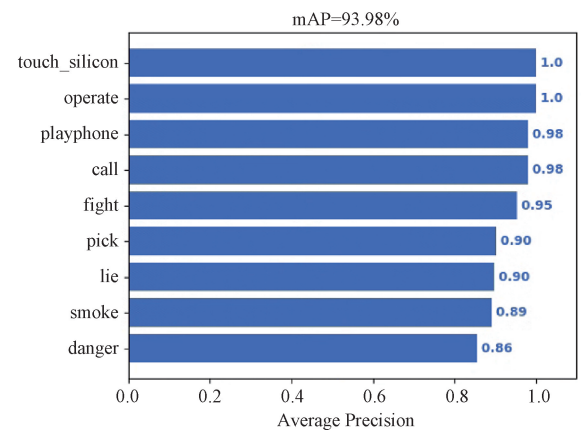


图 8　检测精度

Fig. 8　Detection accuracy

为对比改进后的模型在实际预测中的检测效果,以便更直观观察算法效果的变化,特别选取了相同的图像用改进前后的权重进行预测,结果如图 9、10 所示。可以看到在第 1 列原始算法检测中,第 1 张图片误识别为 fight,而改进算法没有识别出任何动作;原始算法漏检了第 2 张图片左方的目标,而改进算法成功检出;第 3 张图片的目标被遮挡一部分,原始算法并未检出,而改进算法精确检测到 lie 这一动作。

(a) 示例1
(a) Example 1

(b) 示例2
(b) Example 2

(c) 示例3
(c) Example 3

图 9 DETR 检测图像

Fig. 9 Detection Results of DETR

(a) 示例1
(a) Example 1

(b) 示例2
(b) Example 2

(c) 示例3
(c) Example 3

图 10 本文算法检测图像

Fig. 10 Detection results of ours

4 结 论

在智慧车间人员行为识别领域,漏检和误检问题一直是制约算法性能的关键挑战。为解决这一问题,本文引入了基于 DETR 的改进模型,引入相对位置编码以改善原始算法中对像素相对距离关系学习的缺陷,并且采用空间调制共同注意力机制将目标的查询向量调整到目标中心位置附近。通过构建大规模数据集,成功验证了改进算法对检测效果的提升。实验结果表明,所提出的模型平均精度达到了 93.98%,相较于原始算法有了显著提升。为验证改进策略的有效性,本文进行了消融实验,证实了新的模型在解决漏检和误检问题上的优越性。在公开验证集上验证模型的改进策略,结果表明本文的改进方法对公开数据集同样适用。这一研究成果不仅在当前行为识别任务中具有重要实用价值,也为后续相关任务提供了有益的参考。未来研究方向将集中在可见光图像和红外图像融合的探索上,以弥补红外图像分辨率较低的缺陷,从而进一步提高行为识别的效果和精度。这一方面的拓展有望使得智慧车间人员行为识别技术更加全面和可靠。

参考文献

[1] 刘庭煜, 洪庆, 孙毅锋,等. 基于图卷积网络的数字孪生车间生产行为识别方法[J]. 计算机集成制造系统, 2021,27(2):501-509.
LIU T Y, HONG Q, SUN Y F, et al. A method for recognizing digital twin workshop production behavior based on graph convolutional networks[J]. Journal of Computer Integrated Manufacturing Systems, 2021, 27(2): 501-509.

[2] 任丹彤,何赟泽,刘贤金,等. 面向智慧工厂的双光融合车间人员行为识别方法[J]. 测控技术, 2022, 41(8):9-15.
REN D T, HE Y Z, LIU X J, et al. A method for recognizing workshop personnel behavior based on dual-optics fusion for smart factories[J]. Measurement & Control Technology, 2022, 41(8): 9-15.

[3] 苏晨阳,武文红,牛恒茂,等. 深度学习的工人多种不安全行为识别方法综述[J]. 计算机工程与应用, 2024,60(5):30-46.
SU CH Y, WU W H, NIU H M, et al. A review of

worker multiple unsafe behavior recognition methods based on deep learning[J]. Computer Engineering and Applications, 2024, 60(5): 30-46.

[4] 周全,倪英豪,莫玉玮,等. FMA-DETR:一种无编码器的 Transformer 目标检测方法[J]. 信号处理,2024,40(6):1160-1170.
ZHOU Q, NI Y H, MO Y W, et al. FMA-DETR: An encoder-free transformer-based object detection method[J]. Signal Processing, 2024, 40(6): 1160-1170.

[5] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation [J]. IEEE Computer Society, 2014, DOI:10.1109/CVPR.2014.81.

[6] GIRSHICK R. Fast R-CNN [J]. Computer Science, 2015, DOI:10.1109/ICCV.2015.169.

[7] REN S, HE K, 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, 2016, 39(6): 1137-1149.

[8] REDMON J, FARHADI A. YOLO9000: Better, faster, stronge [J]. IEEE, 2017: 6517-6525. DOI: 10.1109/CVPR.2017.690.

[9] LIU W, ANGUELOV D, ERHAN D, et al. Ssd: Single shot multibox detector [C]//Computer Vision-ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I 14. Springer, 2016: 21-37.

[10] 罗国富,王源,李浩,等. 基于改进 YOLOv5s 的智能车间工人不安全行为实时检测方法[J]. 计算机集成制造系统,2024,30(5):1610-1619.
LUO G F, WANG Y, LI H, et al. Real-time detection of unsafe worker behavior in intelligent workshops based on improved YOLOv5s[J]. Journal of Computer Integrated Manufacturing Systems, 2024, 30(5): 1610-1619.

[11] 杜闯,何赟泽,邓海平,等. 基于百度飞桨的面向黑暗环境人员行为检测与身份识别[J]. 电子测量与仪器学报,2023,37(8):21-29.
DU CH, HE Y Z, DENG H P, et al. Personnel behavior detection and identity recognition in dark environments based on baidu paddle [J]. Journal of Electronic Measurement and Instrumentation, 2023, 37(8): 21-29.

[12] CHENG L, HE Y, MAO Y, et al. Personnel detection in dark aquatic environments based on infrared thermal imaging technology and an improved YOLOv5s model[J]. Sensors, 2024, 24(11): 3321.

[13] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers [C]. European Conference on Computer Vision, Springer, 2020: 213-219.

[14] VASWANI A, SHAZEER N, PARMAR N, et al. 2017. Attention is all you need [J]. Advances in neural information processing systems, 30.

[15] 周丽娟,毛嘉宁. 视觉 Transformer 识别任务研究综述[J]. 中国图象图形学报,2023,28(10):2969-3003.
ZHOU L J, MAO J N. A review of vision transformer recognition tasks[J]. Journal of Computer Graphics and Image Processing, 2023, 28(10): 2969-3003.

[16] 李宗刚,宋秋凡,杜亚江,等. 基于改进 DETR 的机器人铆接缺陷检测方法研究[J]. 铁道科学与工程学报,2024,21(4):1690-1700.
LI Z G, SONG Q F, DU Y J, et al. Research on robot riveting defect detection method based on improved DETR[J]. Journal of Railway Science and Engineering, 2024, 21(4): 1690-1700.

[17] 陈洛轩, 林成创, 郑招良,等. Transformer 在计算机视觉场景下的研究综述[J]. 计算机科学,2023,50(12):130-47.
CHEN L X, LIN CH CH, ZHENG ZH L, et al. A review of transformer in computer vision[J]. Computer Science, 2023, 50(12): 130-147.

[18] 张政,何慧. 一种改进的 DETR 输电线通道山火烟雾检测方法[J]. 小型微型计算机系统,2024,45(3):670-675.
ZHANG ZH, HE H. An improved DETR method for detecting transmission line corridor wildfire smoke[J]. Small Microcomputer Systems, 2024, 45(3): 670-675.

[19] WU K, PENG H, CHEN M, et al. Rethinking and improving relative position encoding for vision transformer [C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 10033-10041.

[20] GAO P, ZHENG M, WANG X, et al. Fast convergence of detr with spatially modulated co-attention [C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 3621-3630.

[21] YAO B, JIANG X, KHOSLA A, et al. Human action recognition by learning bases of action attributes and parts[C]. 2011 International Conference on Computer

Vision. IEEE, 2011: 1331-1338.

作者简介

何赟泽,2012 年于国防科学技术大学获得博士学位,现为湖南大学教授,主要研究方向为嵌入式人工智能与边缘计算、红外热成像与机器视觉。

E-mail:hejicker@163.com

He Yunze received his Ph. D. from the University of Defense Science and Technology in 2012. Now a professor at Hunan University, his main research direction is embedded artificial intelligence and edge computing, infrared thermal imaging and machine vision.

谯灵俊,2022 年于湖南大学获得学士学位,现为湖南大学硕士研究生,主要研究方向为深度学习、图像处理。

E-mail:3013277931@qq.com

Qiao Lingjun received his B. Sc. degree from Hunan University in 2022. Now he is a postgraduate of Hunan University. His main research direction is deep-learning and image processing.

王洪金(通信作者),2008 年和 2010 年于湖南大学获得学士学位和硕士学位,2016 年于美国德州农工大学获得博士学位,现为湖南大学副教授,主要研究方向为超分辨热成像与红外多光谱视觉,以及相关图像处理、视觉测量与深度测量及相关机器学习。

E-mail:hjwang_2018@hnu.edu.cn

Wang Hongjin (Corresponding author) received her B. Sc. degree and M. Sc. from Hunan University in 2008 and 2010, and received her Ph. D. degree from Texas A&M University in 2016. Now she is an associate professor in Hunan University. Her main research interests include super-resolution thermal imaging and infrared multispectral vision, and related image process, vision measurement and depth measurement and related machine learning.