中图法分类号:TP391.41　　**文献标识码**:A　　**文章编号**:1006-8961(2025)04-1027-14

论文引用格式:Jin T and Hu P Y. 2025. HK-DETR: improved knife-holding dangerous behavior detection algorithm based on RT-DETR. Journal of Image and Graphics, 30(4):1027-1040(金涛, 胡配雨. 2025. 改进实时目标检测Transformer的持刀危险行为检测算法. 中国图象图形学报, 30(4):1027-1040)[DOI:10.11834/jig.240295]

改进实时目标检测Transformer的持刀危险行为检测算法

金涛,胡配雨*

甘肃政法大学人工智能学院,兰州　730070

摘　要:**目的**　在对公安系统网络摄像头获取的视频数据进行分析时,行人危险持刀行为的自动检测面临刀具形状、大小的多样性,以及遮挡和多目标重叠等因素导致的检测精度低、误检率高的挑战。针对上述问题,提出了一种改进实时目标检测Transformer(real-time detection Transformer,RT-DETR)的持刀危险行为检测算法(human-knife detection Transformer,HK-DETR)。**方法**　首先,设计了倒置残差级联模块(inverted residual cascade block,IRCB)作为主干网络中的基本块(BasicBlock),这使得网络更加轻量化,减少了计算冗余,并提高了对全局特征和长距离依赖关系的理解能力;其次,提出了跨阶并行空洞融合网络结构(cross stage partial-parallel multi-atrous convolution,CSP-PMAC),专注于多尺度特征的提取,使模型能有效识别不同大小和角度的刀具;最后,引入了Haar小波下采样(Haar wavelet-based downsampling,HWD)模块来替换原模型中的下采样操作,为多尺度特征融合提供了更丰富的信息。同时,采用了最小点距离交并比(minimum point distance based intersection over union,MPDIoU)损失函数来进一步提升检测性能。**结果**　对比实验表明,与原RT-DETR算法相比,改进后的模型网络参数量下降了25%,精度、召回率、平均精度均值(mean average precision,mAP)分别提高了2.3%、5.5%、5.2%;与YOLOv5m、YOLOv8m和Gold-YOLO-s相比,在模型网络参数量较低的情况下mAP又分别提高了6.3%、5.2%、1.8%。**结论**　本文提出的HK-DETR算法能够有效适应网络摄像头下多种复杂环境的持刀危险行为检测场景,相较于其他参与对比的模型,其性能优势得到了充分的展现。

关键词:持刀行为检测;实时目标检测Transformer(RT-DETR);目标检测;多尺度特征融合;Transformer;危险行为检测

HK-DETR: improved knife-holding dangerous behavior detection algorithm based on RT-DETR

Jin Tao, Hu Peiyu*

College of Artificial Intelligence, Gansu University of Political Science and Law, Lanzhou　730070, China

Abstract: **Objective** In contemporary society, public safety concerns have garnered increasing attention, particularly in crowded venues such as subway stations, railway terminals, and commercial centers, where timely and accurate detection and response to potential threatening behaviors are paramount for maintaining societal stability. The extensive deployment

收稿日期:2024-06-03;**修回日期**:2024-09-05;**预印本日期**:2024-09-12

*通信作者:胡配雨　18225849687@163.com

基金项目:国家自然科学基金项目(62067006);兰州市人才创新创业项目(2020-RC-10);甘肃省高校科研创新平台重大培育项目(2024CXPT-17)

Supported by: National Natural Science Foundation of China (62067006); Lanzhou Talent Innovation and Entrepreneurship Project (2020-RC-10); Major Cultivation Project of Scientific Research and Innovation Platform for Colleges and Universities in Gansu Province (2024CXPT-17)

of network cameras by public security systems serves as a vital surveillance tool, capable of capturing and recording vast amounts of video data in real-time, providing a rich source of information for security analysis. Nevertheless, a pivotal challenge arises when delving into the depths of these video data: the automated detection of pedestrians engaging in dangerous knife-wielding behaviors. The complexity of this task stems primarily from the diversity in knife shapes and sizes, ranging from conventional elongated knives to folding knives and daggers, each exhibiting distinct visual representations in images, posing significant challenges for detection algorithms. Furthermore, occlusions, a common occurrence in real-world surveillance scenarios, including body occlusions between pedestrians, obstructions by trees or buildings, can lead to incomplete target feature information, thereby compromising detection performance. Additionally, multi-object occlusion, prevalent in densely populated areas, where multiple pedestrians or objects overlap in images, exacerbates the difficulty in accurately distinguishing and localizing knife-wielding individuals. To address these issues and enhance the precision and efficiency of detecting dangerous knife-wielding behaviors, this paper proposes an algorithm named human-knife detection Transformer (HK-DETR), which is an improvement upon the real-time detection Transformer (RT-DETR). Building upon the inherent strengths of RT-DETR, HK-DETR incorporates numerous optimizations and innovations tailored specifically to the characteristics of knife-detection tasks. **Method** First, we meticulously designed the inverted residual cascade block (IRCB) as a fundamental building block (BasicBlock) within the backbone network. This innovative design not only achieves a lightweight network architecture, effectively alleviating computational resource scarcity, but also significantly reduces redundant computations. By optimizing the processing flow of feature maps, the IRCB module substantially enhances the backbone network's ability to capture and distinguish diverse features, thereby laying a solid foundation for subsequent complex knife detection tasks. We propose the cross stage partial-parallel multi-atrousconvolution (CSP-PMAC) module, a revolutionary feature fusion strategy. This module directs the network to focus more intently on capturing and integrating multi-scale feature information during the fusion stage, which is pivotal for identifying knives of varying shapes and angles. This design equips the model with exceptional adaptability, enabling it to accurately identify both small knives and large knives, thus significantly improving the model's performance in complex scenarios. In further optimizing the model, we have selected the novel Haar wavelet-based downsampling (HWD) module as a downsampling method to replace the traditional downsampling mechanism within the network. By leveraging its unique hierarchical wavelet decomposition technique, the HWD module effectively diminishes data dimensionality while retaining richer details of object scale variations. This enriches and refines feature representations in subsequent multiscale feature fusion, enhancing the model's robustness in handling scale variations. Finally, to comprehensively enhance detection accuracy, we have adopted the minimum point distance based intersection over union (MPDIoU) loss function. This improved loss function optimizes object localization accuracy by more precisely measuring the overlap between predicted bounding boxes and actual target boxes. It not only considers classification accuracy but also intensifies the pursuit of localization precision, enabling the model to maintain superior detection performance even in the presence of dense or overlapping targets. **Result** Ablation experiments were conducted on the pedestrian knife-carrying dataset, which revealed that each improvement strategy, when applied individually, contributed to a certain degree of performance enhancement for the original RT-DETR model, despite the persistence of challenges such as missed detections and confidence issues in some cases. However, when these improvement strategies were combined, a significant boost in detection performance was achieved. To validate the effectiveness of the proposed model, comparative experiments were performed on the pedestrian knife-carrying dataset. The results demonstrated that compared to the original RT-DETR algorithm, the refined model exhibited a 25% reduction in network parameters while achieving improvements of 2.3%, 5.5%, and 5.2% in accuracy, recall, and mean average precision (mAP), respectively. When benchmarked against YOLOv5m, YOLOv8m, and Gold-YOLO-s, the refined model, with a lower number of network parameters, demonstrated notable mAP enhancements of 6.3%, 5.2%, and 1.8%, respectively. **Conclusion** The proposed HK-DETR algorithm in this paper exhibits remarkable performance advantages in the task of automatically detecting dangerous knife-carrying behaviors of pedestrians in video data captured by public security system network cameras. This algorithm effectively addresses the challenges posed by the diversity of knife shapes and sizes, occlusion, and multi-target overlapping in complex scenarios, while significantly enhancing detection accuracy, recall rate, and mAP through a series of innovative designs. Compared to the original RT-DETR algorithm and

other mainstream detection models such as YOLOv5m, YOLOv8m, and Gold-YOLO-s, HK-DETR achieves notable performance improvements. This result underscores the algorithm's ability to maintain high efficiency and accuracy in diverse and complex environments, offering robust technical support for the field of public security surveillance. Within the realm of public safety, HK-DETR holds immense potential for widespread adoption in surveillance systems of public places like railway stations, airports, subway stations, and shopping malls, enabling real-time detection and early warning of potential knife-related dangers, thereby providing timely and effective information support for law enforcement agencies. Moreover, as technology continues to evolve and mature, the HK-DETR algorithm is poised to expand its reach into other domains, such as intelligent transportation and industrial automation, offering potent solutions to an array of practical problems.

Key words: knife-holding behavior detection; real-time detection Transformer (RT-DETR); object detection; multi-scale feature fusion; Transformer; dangerous behavior detection

0 引　言

随着公安系统的网络摄像头大量安装在公共场所，实现了这些区域的全天候监控，产生了海量的视频数据。为了确保公共安全，需要对数据中危险行为进行实时分析，对危险行为中相关的关键器械实时检测。一些研究团队开始探索持刀的危险行为的识别方法，力求为执法机关工作人员提供预警帮助，确保在危险事件发生时，执法人员能够具有充足的响应时间以迅速应对。行人持刀行为的检测作为人—物交互关系检测的一个重要应用方向（廖越 等，2022），已成为计算机视觉领域的研究热点。行人持刀行为的检测方法主要有以下几种：一种是基于传感器的检测方法，例如使用X射线或扫描仪（Akcay和Breckon，2022），另一种方法是利用卷积神经网络（convolutional neural network，CNN）算法从图像或视频中提取特征，进而实现行人持刀行为的识别和检测。尽管基于CNN的方法在行人持刀行为检测方面取得了显著进展（梁家菲 等，2023），但仍然面临着挑战。行人持刀行为的特征复杂多变，极易受到光照条件、遮挡情况、目标大小及多目标重叠等环境因素的影响，这些因素均可能导致模型性能下降，影响检测结果的准确性和可靠性。因此行人持刀行为检测是一项有意义且有实际应用价值的任务。

基于CNN的目标检测模型通常划分为两阶段检测器和一阶段检测器两大类。两阶段检测器，如基于区域的卷积神经网络Faster R-CNN（Faster region-based convolutional neural network）（Ren 等，2017），首先会生成一系列候选区域，随后对这些区域内的目标进行分类和位置回归；一阶段检测器，如YOLO（you only look once）（Redmon 等，2016）和SSD（single shot multibox detector）（Liu 等，2016），这些算法共同推动了实时目标检测技术的发展。尤其是YOLO系列目标检测算法，长期以来深受多个研究团队的青睐，并得到深入研究。例如，为了解决当前YOLOv5s目标检测网络复杂、参数多、部署所需配置高，难以在嵌入式平台上获得优质识别结果的问题，龙邹荣等人（2023）设计了YOLOv5s_GCB模型，该模型算法在保证检测精确度的同时实现了模型的轻量化，为其在性能较弱的嵌入式平台上的部署与应用提供了一定的理论依据。然而，这些传统的一阶段和两阶段检测器正在逐渐被采用Transformer框架的检测器所超越。Carion等人（2020）提出了一种基于Transformer的端到端目标检测器DETR（detection Transformer），由于该模型其端到端的特性以及利用对象间关系进行全局推理的能力，促进了局部和全局信息之间的融合，在目标检测领域引起了研究者的广泛关注。DETR消除了传统检测流程中的锚点和非极大值抑制（non-maximum suppression，NMS）组件，简化了目标检测过程。然而，DETR存在收敛速度慢和受限于特征图的空间分辨率的缺点。为了缓解这些问题，Zhu 等人（2020）提出Deformable DETR，其注意力模块只关注参考周围的一小组关键采样点。该模块可以自然地扩展到聚合多尺度特征而无需特征金字塔，利用多尺度可变形注意模块代替Transformer注意模块处理特征映射。Wang等人（2022）认为每个可学习的embedding都没有明确的物理意义，使得每个对象查询将不会关注特定的区域，于是引入了基于锚点的对象查询方法，通过锚点的引入，该方法有效地捕获了不同大小和纵横比的物体，提高了边界框和类别信息的预测准确性。Liu等人（2022）将box信息作为DETR中Transformer解

码器查询机制的单元，可以结合上下文进行双重查询的算法，通过这些查询来寻找与box框、上下文有相似性的目标，并逐层动态更新，达到加速网络收敛的效果。Li等人（2024）认为模型收敛缓慢是由于二分匹配的不稳定性导致的，于是在基线DAB-DETR进行改进，将带有噪声的真实框输入Transformer解码器，并训练模型预测其原来的真实框，通过这种方式显著改善了DETR解码器匹配不稳定的问题，让模型更快地收敛。Chen等人（2023）认为DETR中的一对一标签匹配导致监督信号不足，阻碍了模型收敛，引入了一对多标签匹配方法，有效缓解了监督信号不足的问题，并加快了网络收敛速度。综上，不同改进的DETR虽然有效地解决了模型网络收敛慢等一些列问题，但是DETR的高计算成本问题仍然限制了其在实时检测场景中的广泛应用。

百度飞桨团队在2023年提出了一种端到端实时目标检测模型RT-DETR（real-time detetction Transformer）（Zhao等，2023），相较于DETR，RT-DETR以Transformer的并行计算机制为基础，实现了高效处理大量输入数据的能力，从而在保持高准确性的同时，显著提升了目标检测的推理速度。为了进一步提升RT-DETR在持刀行为检测任务上的效率和性能，本文在其基础上进行改进，提出了持刀行为检测算法：HK-DETR（human-knife detection Transformer）。该算法重新设计了主干网络ResNet18的BasicBlock，以促进网络的轻量化以及增强模型对全局特征的理解能力；通过跨阶段局部网络结构，利用并行空洞卷积加强多尺度特征提取；整合小波下采样模块，以解决刀具边界信息丢失的问题；并采用最小点距离交并比（minimum point distance based intersection over union，MPDIoU）损失函数，以提升训练评估的精确度。上述改动综合增强了算法的实时性、准确率和鲁棒性。

1 RT-DETR算法

Transformer模型最初是为自然语言处理（natural language processing，NLP）中的序列到序列任务而设计（Han等，2023），其核心是自注意力机制，该机制允许模型高效地捕捉输入序列中不同位置间的依赖关系，从而更好地处理长距离依赖和序列语义。鉴于Transformer在NLP中的成功应用，研究者们开始探索将其应用于计算机视觉的目标检测任务，并因此提出了RT-DETR模型。

RT-DETR模型主要由4个关键组件构成：骨干网络、高效混合编码器、交并比（intersection over union，IoU）感知查询选择机制以及带有辅助预测头的解码器。其中，高效混合编码器是RT-DETR的核心。该编码器包含两个模块：注意力尺度内特征交互（attention-based intra-scale feature interaction，AIFI）模块和跨尺度特征融合模块（cross-scale feature-fusion module，CCFM）。AIFI模块仅对高层次特征$\boldsymbol{S}_5$进行交互，以减少计算成本并提高对象定位和识别的准确性。CCFM的作用是将两个相邻尺度特征融合为一个新特征。高效混合编码器计算为

$$\boldsymbol{Q}=\boldsymbol{K}=\boldsymbol{V}=f_{\text{Flatten}}\left(\boldsymbol{S}_5\right) \tag{1}$$

$$\boldsymbol{F}_5=f_{\text{Reshape}}\left(f_{\text{Attn}}\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\right)\right) \tag{2}$$

$$\boldsymbol{Output}=f_{\text{CCFM}}\left(\left\{\boldsymbol{S}_3,\boldsymbol{S}_4,\boldsymbol{F}_5\right\}\right) \tag{3}$$

式中，$\boldsymbol{S}_3$，$\boldsymbol{S}_4$，$\boldsymbol{S}_5$为不同尺度的特征图，f_{Flatten}代表将$\boldsymbol{S}_5$展平作为查询$\boldsymbol{Q}$、键$\boldsymbol{K}$、值$\boldsymbol{V}$的初始化，f_{Reshape}代表将经过注意力机制f_{Attn}后的特征恢复为与$\boldsymbol{S}_5$相同的尺寸大小，$\boldsymbol{F}_5$和$\boldsymbol{Output}$分别为f_{Reshape}和f_{CCFM}的输出特征。

高效混合编码器通过尺度内交互和跨尺度融合将多尺度特征转换为一系列图像特征。在训练阶段由于分类分数和位置置信度的分布不一致的问题会导致预测框的选择错误。因此，RT-DETR提出了IoU感知查询选择机制在模型训练期间施加约束，以调整模型偏好，确保高重叠度的预测框获得较高的分类分数。解码器则依据IoU感知机制从编码器输出序列中选取固定数量的高质量图像特征。最终，这些特征通过辅助预测头用于目标类别的识别及边界框的生成。

2 持刀行为检测算法

本文提出的持刀检测算法HK-DETR，是在RT-DETR结构基础上进行改进的模型，其结构如图1所示。为了确保网络尽可能轻量化的同时保证全局特征不丢失，设计了倒置残差级联模块（inverted residual cascade block，IRCB）来更新主干网络ResNet18中的基本块（BasicBlock）。此外，为解决多尺度特征提取问题，以适应不同尺度大小的刀具识

别，提出了跨阶并行空洞融合网络模块（cross stage partial-parallel multi-atrous convolution，CSP-PMAC）。针对刀具边界信息容易丢失的问题，引入了 Haar 小波下采样（Haar wavelet-based downsampling，HWD）模块，该模块能在降低分辨率的同时保留更多细节，特别是刀具的边界信息，这对在复杂光线条件下准确识别刀具至关重要。最后，为了提供更全面、更准确的边界框相似性比较指标，采用了最小点距离交并比（MPDIoU）损失函数，该函数综合考虑了重叠区域、非重叠区域、中心点距离、宽度和高度偏差等多个因素。

2.1 倒置残差级联模块

ResNet18 的 BasicBlock 主要依赖于标准卷积操作，这种设计使得其特征提取能力主要集中于捕捉局部特征，而对于复杂的全局特征捕捉则显得不够高效。在行人持刀检测任务中，行人的身体姿态和手臂位置往往具有明显的预示性动作，手部动作与身体其他部位也存在相应的联系。因此，理解全局特征、长距离依赖关系尤为重要。此外，ResNet18 的计算密集型特性使其在处理大规模图像数据集时面临较高的计算成本。

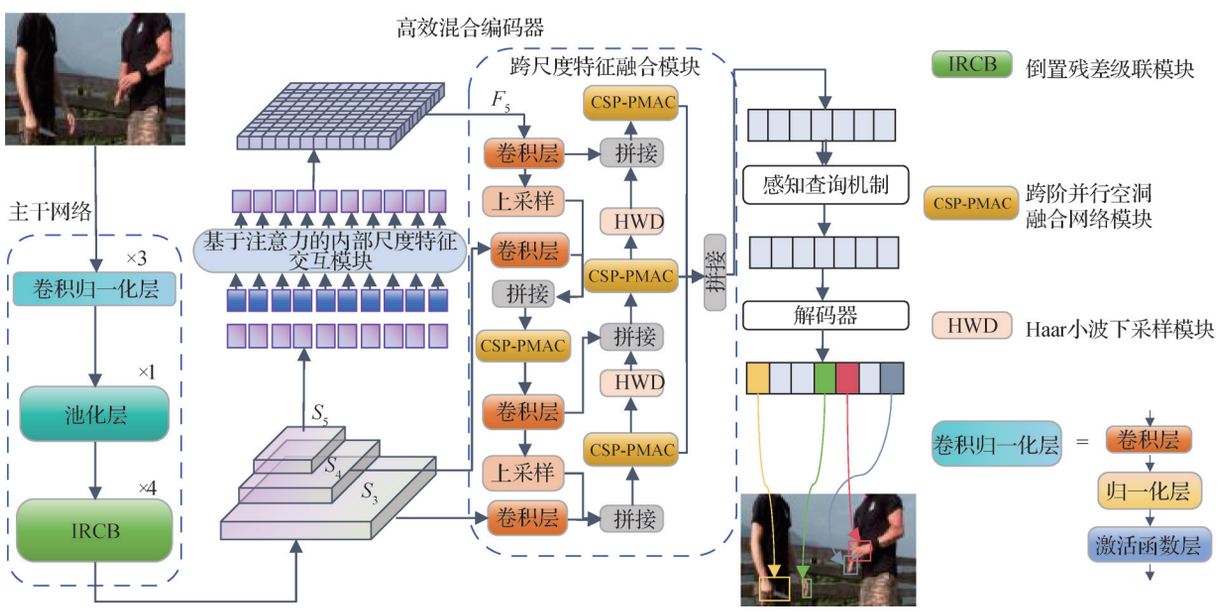


图 1 HK-DETR 模型结构

Fig. 1 Structure of HK-DETR model

为了克服这些局限，受倒置残差移动块（inverted residual mobile block，IRMB）（Zhang 等，2023b）和 EfficientViT 模型中的级联分组注意力机制（cascaded group attention，CGA）启发（Liu 等，2023），提出了一种融合了 IRMB 和改进 CGA（residual cascaded group attention，R-CGA）的倒置残差级联模块 IRCB 来替换原始 BasicBlock，IRCB 结构如图 2(b)(c)所示。特征输入后，会被分别送入 R-CGA 模块和卷积模块（convolution，Conv）。R-CGA 模块以其独特的级联分组以及保留了原始图像特征的方式，有效地实现了注意力头之间的信息解耦，并在不丢失原始特征的前提下逐步精细化了特征，将 R-CGA 模块输出特征与卷积后的输出特征加权相乘，最终再经过一个深度可分离卷积（depth-wise convolution，DW-Conv）模块和残差模块输出最后特征。IRCB 使用 DW-Conv 降低计算量和模型参数的数量，同时利用 R-CGA 注意力机制捕获特征间的全局依赖关系，与 ResNet18 的 BasicBlock 相比，IRCB 可以获取更加多样性的特征。接下来将分别介绍 IRMB 和 R-CGA 模块。

IRMB 是一种结合了 CNN 和 Transformer 架构思想的轻量化网络模块。它巧妙地结合了倒置残差块与注意力模块，实现了对输入特征的精细化和高效化处理。在 IRMB 结构中，特征首先通过一系列卷积层进行初步的特征映射与提取，随后采用高效的多头自注意力（multi-head self-attention，MHSA）和 DW-Conv，并结合跳跃连接，使模型能够聚焦于对任务至关重要的特征信息。最后，通过进一步的卷积和池化操作，得到具有代表性和判别力的输出特征。

CGA 将输入特征分割成多个部分，并在不同的

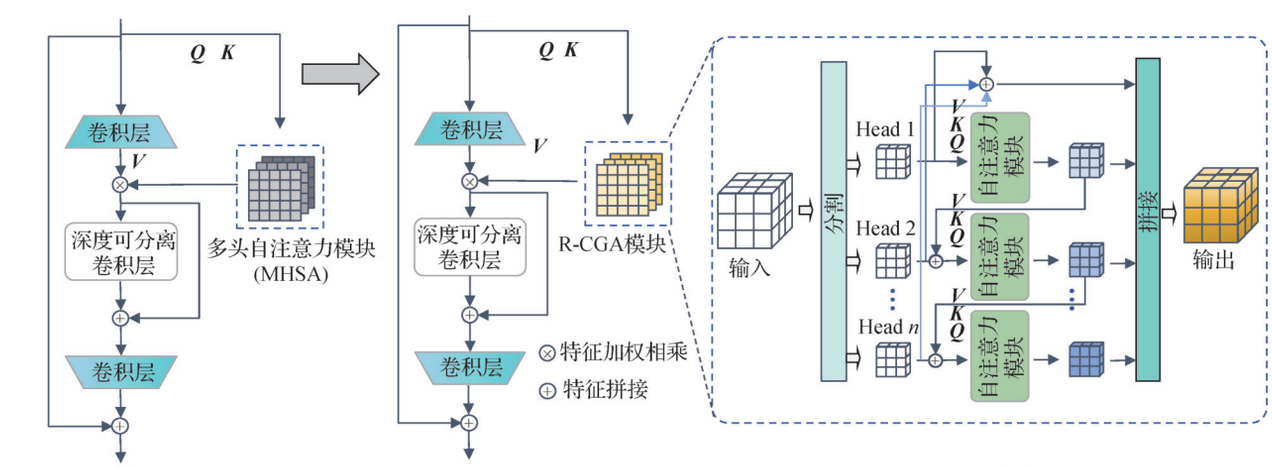
注意力头中独立计算自注意力，同时通过级联的方式逐步细化特征表示。虽然CGA可以有效缓解MHSA中的注意力头冗余问题，但每次分组会导致部分信息模糊化。尽管每个注意力组内部可以加强局部信息的关联性，但在级联过程中，可能会出现全局信息的丢失情况。为了应对这些挑战，对CGA进行了改进，构建了R-CGA。在特征分组后为每个组添加一条支路，使分组特征在输入自注意力模块前通过各自的支路进行特征汇聚，并在最终进行拼接（concatenation，concat）时提供全局信息。其结构如图2（c）所示。R-CGA结构内部计算过程为

$$\tilde{\boldsymbol{X}}_{ij}=f_{\text{Attn}}\left(\boldsymbol{X}_{ij}\boldsymbol{W}_{ij}^{\text{Q}},\boldsymbol{X}_{ij}\boldsymbol{W}_{ij}^{\text{K}},\boldsymbol{X}_{ij}\boldsymbol{W}_{ij}^{\text{V}}\right) \tag{4}$$

$$\boldsymbol{X}'_{ij}=\boldsymbol{X}_{ij}+\tilde{\boldsymbol{X}}_{i(j-1)} \tag{5}$$

$$\tilde{\boldsymbol{X}}_{i+1}=f_{\text{Concat}}\left[\tilde{\boldsymbol{X}}_{ij}+f_{\text{Concat}}\left[\boldsymbol{X}_{ij}\right]_{j=1:h}\right]_{j=1:h}\boldsymbol{W}_{i}^{\text{P}} \tag{6}$$

式中，i表示输入第i个特征，j表示特征输入后的第j个分割，$\tilde{\boldsymbol{X}}_{ij}$表示输入特征$\boldsymbol{X}_i$的第$j$个分割$\boldsymbol{X}_{ij}$和投影矩阵$\left(\boldsymbol{W}_{ij}^{\text{Q}},\boldsymbol{W}_{ij}^{\text{K}},\boldsymbol{W}_{ij}^{\text{V}}\right)$通过$f_{\text{Attn}}$自注意力函数计算的输出。$\boldsymbol{X}'_{ij}$表示通过级联得到的注意力头的输入，具体来说，对于第j个注意力头$\left(j>1\right)$，$\boldsymbol{X}'_{ij}$是原始分割$\boldsymbol{X}_{ij}$与前一个注意力头（第$j-1$个注意力头）的输出$\tilde{\boldsymbol{X}}_{i(j-1)}$之和。$\tilde{\boldsymbol{X}}_{i+1}$将所有注意力头的输出$\tilde{\boldsymbol{X}}_{ij}$进行$f_{\text{Concat}}$拼接，并使用一个线性层$\boldsymbol{W}_i^{\text{P}}$将连接后的输出投影回与输入一致的维度。这里，$h$是注意力头的总数。



(a) 倒置残差移动模块 (IRMB)　(b) 倒置残差级联模块 (IRCB)　(c) 改进的级联分组注意力模块 (R-CGA)

图2　倒置残差级联模块结构图

Fig. 2　Inverted residual cascade block architecture diagram

((a) inverted residual mobile block; (b) inverted residual cascade block; (c) residual cascaded group attention module)

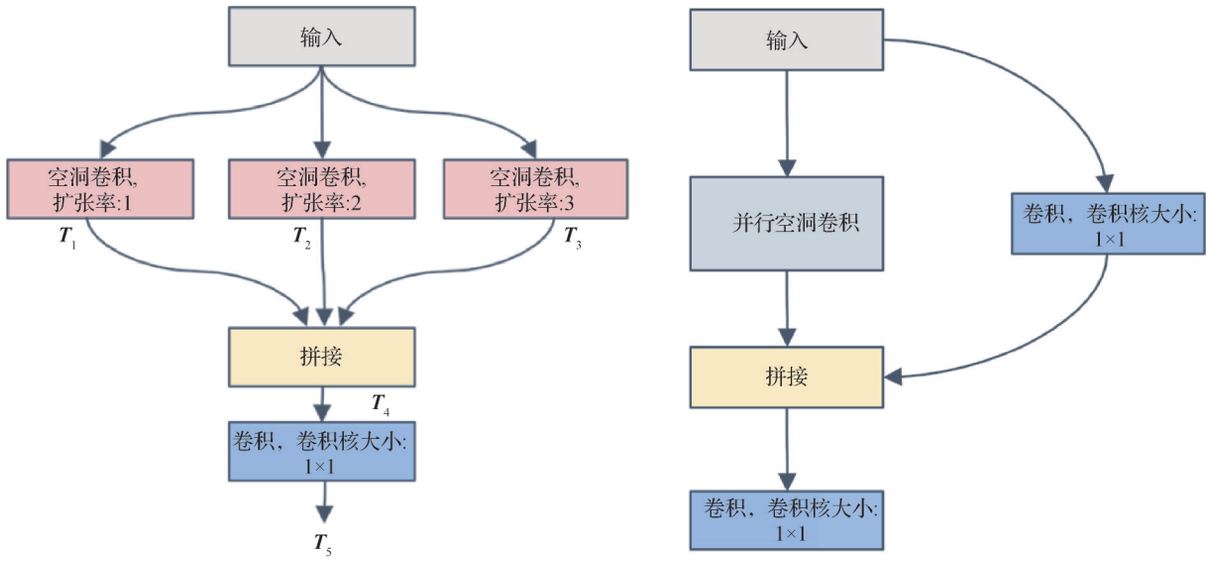
2.2　改进的跨阶段局部网络结构

特征的多样性和有效性对于持刀行为检测模型的性能至关重要。如果网络提取的特征过于单一或冗余，会导致模型性能下降或过拟合的问题。不同尺度的特征对于模型的性能同样重要，为了获得图像的多尺度特征，RT-DETR使用传统的卷积对图像不同尺度特征进行提取。传统卷积在处理较大感受野时需要增加卷积核的大小，从而导致参数数量的急剧增加。因此，设计了并行空洞卷积（parallel multi-atrous convolution，PMAC），其结构如图3（a）所示。进一步地，受跨阶段局部网络结构（cross stage partial，CSP）的启发（Wang等，2020），提出了CSP-PMAC网络结构，增强了模型对图像多样性特征和多尺度特征的获取能力，其结构如图3（b）所示。

PMAC采用了三分支结构。在PMAC结构中，为每个分支分别设置了具有不同扩张率的空洞卷积层，通过调整空洞卷积的扩张率，空洞卷积能够在不增加卷积核大小的前提下有效地扩大网络的感受野（Yu和Koltun，2015），图4展示了空洞卷积不同扩张率之间的区别。

在空洞卷积中，设定$k\times k$大小的卷积核后，通过调整卷积核的空洞数d，可以实现相当于不同大小$k'\times k'$卷积核的卷积效果，当前层空洞卷积的感受野计算为

$$k'=k+(k-1)\times(d-1) \tag{7}$$



(a) 并行空洞卷积(PMAC)　　(b) 改进的跨阶段局部网络(CSP-PMAC)

图 3　PMAC 与 CSP-PMAC 结构图

Fig. 3　Structural diagrams of PMAC and CSP-PMAC ((a) PMAC; (b) CSP-PMAC)

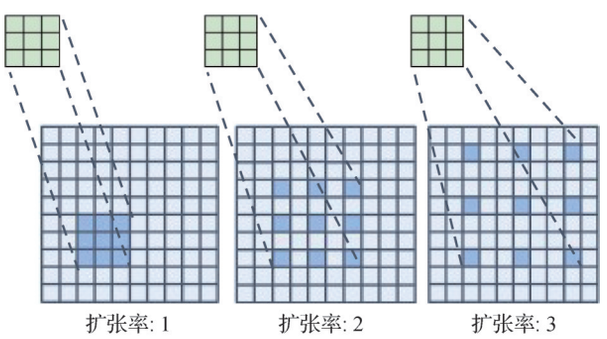


图 4　不同扩张率的空洞卷积图

Fig. 4　Dilated convolution maps with different dilation rates

$$RF_{n+1} = RF_n + (k' - 1) \times \prod_{i=1}^{n} S_i \tag{8}$$

式中，RF_{n+1}表示当前第$n+1$层的感受野，RF_n表示第n层的感受野，S_i表示第i层空洞卷积的步长。

在PMAC层，长为H宽为W的特征图$\boldsymbol{T}$，首先分别通过卷积核都为3×3、步幅为1的3个空洞卷积分支，且3个分支的填充值p和扩充率d分别为1、2和3，输出尺寸为(H, W)的3个特征图$\boldsymbol{T}_1, \boldsymbol{T}_2, \boldsymbol{T}_3$，输出特征图为

$$\boldsymbol{O} = \left\lfloor \frac{i + 2p - k - (k-1) \times (d-1)}{s} \right\rfloor \tag{9}$$

式中，k表示卷积核大小，s表示卷积核滑动时的步长。

然后将3个包含了不同尺度上下文的特征融合输出为特征图$\boldsymbol{T}_4$，最后进行一次普通的卷积操作得到特征图$\boldsymbol{T}_5$。

CSP-PMAC分为两个支路，其中一条支路通过普通的卷积来提取图像特征，确保图像的基本特征信息不会丢失；另一条支路使用PMAC对图像进行多尺度特征的提取。最后，两条支路的输出特征图会通过一个拼接操作进行融合。

2.3　Haar小波下采样模块

下采样操作在CNN中广泛用于聚合局部特征、扩大感受野以及减少计算开销。本算法中下采样主要用于图像不同尺度特征融合前的图像处理，将大尺寸特征图减小到与之特征融合的特征图相同大小，并且要最大程度地提取主要特征信息。RT-DETR使用的传统下采样操作可能导致关键的空间信息丢失，特别是刀具边界信息。Haar小波变换的细节系数能够突出图像中的高频部分，例如边缘和纹理，这对于精准检测刀具的直线和锋利边缘至关重要。引用了一种基于Haar小波变换的下采样模块(Haar wavelet-based downsampling, HWD)(Xu等，2023)，该模块在降低特征图空间分辨率的同时尽可能保留更多的信息。

HWD模块通过应用Haar小波变换，将特征图的空间信息编码到通道维度中，而不丢失任何信息。Haar小波变换将输入的特征图分解为4个部分：低频近似A和水平H、垂直V、对角D这3个方向的高频细节，其中低频近似组件保留了图像的主要信息，而

3个高频细节组件则分别捕获了图像在不同方向上的细节信息，这4个部分对应的函数$\Phi_{\mathrm{A}}(x,y)$、$\psi_{\mathrm{H}}(x,y)$、$\psi_{\mathrm{V}}(x,y)$、$\psi_{\mathrm{D}}(x,y)$分别为

$$\Phi_{\mathrm{A}}(x,y)=\phi_{\mathrm{x}}(x)\times\phi_{\mathrm{y}}(y) \tag{10}$$

$$\psi_{\mathrm{H}}(x,y)=\phi_{\mathrm{x}}(x)\times\psi_{\mathrm{y}}(y) \tag{11}$$

$$\psi_{\mathrm{V}}(x,y)=\psi_{\mathrm{x}}(x)\times\phi_{\mathrm{y}}(y) \tag{12}$$

$$\psi_{\mathrm{D}}(x,y)=\psi_{\mathrm{x}}(x)\times\psi_{\mathrm{y}}(y) \tag{13}$$

式中，x，y表示二维空间中的坐标，$\phi_{\mathrm{x}}(x)$和$\psi_{\mathrm{x}}(x)$分别表示沿x方向的一维尺度函数和小波函数，而$\phi_{\mathrm{y}}(y)$和$\psi_{\mathrm{y}}(y)$分别是沿y方向的对应一维尺度函数和小波函数。

经过小波变化后的特征图通过一个包含标准卷积层、批量归一化层和ReLU（rectified linear unit）激活函数的特征表示学习块来进一步处理这些特征，调整通道数并过滤冗余信息，以便后续层更有效地学习特征。上述Haar小波下采样模块结构如图5所示。

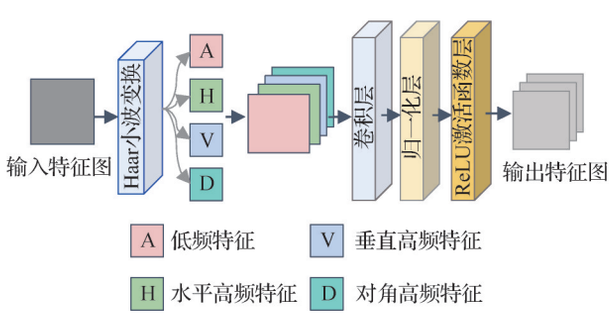


图5 Haar小波下采样结构

Fig. 5 Structure of Haar wavelet downsampling

2.4 MPDIoU损失函数

目标检测算法中，边界框用于表示目标的位置和大小，以使算法能够将目标与背景或其他对象进行区分。边界框通常由4个坐标值表示，分别代表了矩形框的左上角和右下角的位置，这种以坐标表示边界框的方法是一种通用的方式。交并比（IoU）损失函数常用于评估算法的性能和衡量不同检测框与标注框之间的匹配程度。传统的IoU损失函数通过计算两个框的交集与并集的比率来评估两个框之间的重叠度，然而，IoU损失函数有一些局限性。例如，IoU损失函数在处理小目标和大目标时存在不适应的问题。对于小目标，由于它们的边界框区域较小，即使有一些微小的误差，也可能会导致IoU值显著下降，从而影响模型的训练；对于大目标，由于边界框区域较大，即使有较大的误差，IoU值可能仍然很高，导致模型对区域的定位和检测能力的缺失。目前主流的目标检测算法中，通常使用完全交并比（complete IoU，CIoU）（Zheng等，2022）损失作为损失函数。CIoU损失函数对小目标的检测相对于IoU更加稳健，它通过考虑边界框的尺寸和位置来减少微小的边界框偏差对IoU值的影响，提高了对小目标的检测能力。CIoU的损失函数为

$$L_{\mathrm{CIoU}}=1-IoU+\frac{\rho^2(b,b^{\mathrm{gt}})}{c^2}+\alpha V \tag{14}$$

式中，b和b^{gt}分别代表预测框的中心坐标和真实框的中心坐标，ρ表示了这两个中心点之间的欧几里得距离，即预测框的中心点与真实框的中心点之间的直线距离。通过c来表示预测框和真实框的最小外接框的对角线距离。α是用于做trade-off的参数，V是用来衡量长宽比一致性的参数，α与V相乘就可以将预测框纵横比拟合真实框的纵横比考虑进去。尽管CIoU损失在实际应用中能够提升目标检测算法的性能和鲁棒性，但在某些特定情况下，如当预测框与边界框具有相同的纵横比，但宽度和高度值完全不同时，CIoU无法有效地衡量目标之间的准确性。这导致CIoU在这些情况下的性能下降。本文采用最小点距离交并比（MPDIoU）（Ma和Xu，2023）来代替CIoU，MPDIoU通过研究水平矩阵的几何特征，提出了一种新的基于最小点距离的边界框相似性比较指标，该指标包含了重叠和不重叠区域、中心点距离、宽度和高度偏差。MPDIoU损失函数为

$$d_1^2=\left(x_1^{\mathrm{pred}}-x_1^{\mathrm{gt}}\right)^2+\left(y_1^{\mathrm{pred}}-y_1^{\mathrm{gt}}\right)^2 \tag{15}$$

$$d_2^2=\left(x_2^{\mathrm{pred}}-x_2^{\mathrm{gt}}\right)^2+\left(y_2^{\mathrm{pred}}-y_2^{\mathrm{gt}}\right)^2 \tag{16}$$

$$MPDIoU=IoU-\frac{d_1^2}{h^2+w^2}-\frac{d_2^2}{h^2+w^2} \tag{17}$$

$$L_{\mathrm{MPDIoU}}=1-MPDIoU \tag{18}$$

式中，x_1^{pred}，y_1^{pred}和x_2^{pred}，y_2^{pred}分别表示预测边界框左上角坐标和右下角坐标；x_1^{gt}，y_1^{gt}和x_2^{gt}，y_2^{gt}分别表示真实边界框左上角坐标和右下角坐标。d_1^2和d_2^2分别表示预测边界框与真实边界框左上角点与右下角点坐标差值的平方和，即左上角点与右下角点的距离的平方，h和w表示输入图像的高度和宽度。MPDIoU示意图如图6所示。

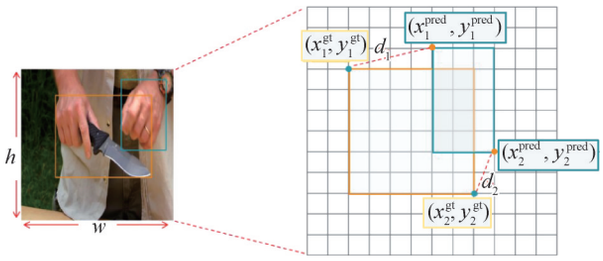


图 6　MPDIoU 示意图

Fig. 6　Diagram of MPDIoU

图 7　行人持刀检测数据集样例

Fig. 7　Sample dataset for pedestrian knife-carrying detection

3　实验结果与分析

3.1　行人持刀数据集

持刀行为检测的实际应用通常在网络摄像头拍摄的图像上，但从这些摄像头收集数据可能涉及隐私问题，同时，公开的持刀数据也相当稀缺。因此，本文自建了一个行人持刀数据集，该数据集来源于人体检测数据集，包括 VOC（visual object classes）、COCO（common objects in context）、MPII（max planck institute for informatics）以及互联网获取的图像数据。这些数据源提供了广泛的真实场景、多样的环境以及多种对象类别。首先挑选出所需要的持刀图像，然后对数据集进行了细致的清理和过滤，以消除噪声和异常值。最终数据集共包含 11 188 幅图像，并按照 8∶2 的比例将其划分为训练集和验证集。其中，训练集包含 8 951 幅图像，验证集包含 2 237 幅图像。所有图像均进行了 Labelme 标注，人体手持刀具标记为 Knife，不持刀具标记为 NULL。数据集样例如图 7 所示，该数据集满足了实验训练需求。此外，为了进一步评估自建数据集上训练的模型在实际网络摄像头监控环境中的泛化表现，采用高清网络摄像头，采集并最终选取 100 幅行人持刀图像并进行标注，为对比实验中不同模型泛化测试提供了数据支持。

3.2　实验配置

持刀检测实验中，采用了基于 Ubuntu 18.04.1 LTS 的操作系统，使用了 NVIDIA A30 型号的 GPU，该 GPU 拥有 24 GB 的显存。深度学习模型的开发与训练则依赖于 PyTorch 框架，版本为 2.2.1，Python 语言的版本是 3.9，CUDA 为 11.3 版本。在模型训练阶段，输入图像尺寸设定为 640 × 640 像素，每个训练批次（batch）包含 4 幅图像。网络的初始学习率设置为 0.000 1，具体参数设置如表 1 所示。模型训练的交并比损失（IoU loss）和分类损失（classification loss）如图 8 所示。从图中可以观察到，训练周期在第 140 个 epoch 之后，IoU loss 和 classification loss 的曲线趋于平稳，模型开始收敛。最终的实验结果在第 150 个 epoch 获得。

表 1　参数设置

Table 1　Parameter setting

参数名	参数值
优化函数	optimizer = “AdamW”
初始学习率	lr = 0.000 1
动量	momentum = 0.9
迭代次数	epoch = 150
小批量数据	bacth_size = 4

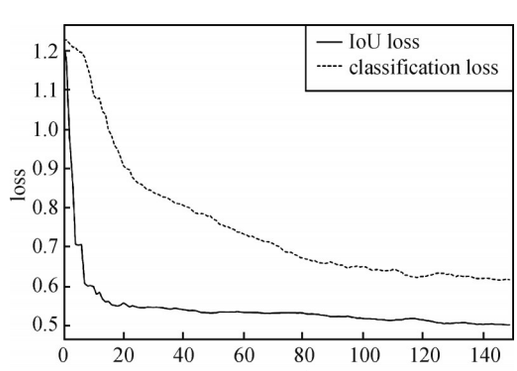


图 8　损失函数变化曲线

Fig. 8　Change curves of loss function

3.3　评价指标

为了确保模型性能的精确评估，本研究选取了多项指标以综合衡量预测精确度。这些指标包括：

精确度(precision, P)、召回率(recall, R)、参数量(parameter)、10亿浮点运算数(giga floating point operations, GFLOPs)以及平均精确度均值(mean average precision, mAP)。precision描述了在模型预测为正样本的实例中,真正为正样本的比例。精确度高意味着预测为正样本的实例中,真正的正样本占比高,即误报率低。recall用于衡量模型在预测正例方面的完备性,召回率高意味着模型能识别出更多的真实正样本,即模型是否能找到数据集中的所有目标。mAP是数据集中所有类别的AP(average precision)的平均值,它包括了不同交并比IoU阈值下的结果,本文设置IoU阈值为0.5,如果预测框与某个真实框的IoU大于或等于预设的阈值,则认为这个预测框是一个有效检测,所以用mAP@0.5对模型进行评估。GFLOPs精确量化了硬件在1 s内所能执行的浮点运算次数,进而间接地反映了模型在特定硬件平台上的运行效率。较高的GFLOPs值通常表明硬件具备更快的处理速度,能够高效地应对大量的计算任务。parameter则直接体现了模型的存储空间大小。因此,在目标检测任务中,综合考虑GFLOPs和parameter两个指标,有助于评估和优化模型在特定硬件上的性能表现。

3.4 损失函数的选择

为了验证不同边界框回归损失函数的效果,以通过CSP-PMAC模块和HWD模块对RT-DETR颈部网络改进后的模型为基础,在该模型上展开不同边界框回归损失函数之间的对比。选择了CIoU、EIoU(Yang等,2021)、Wise-IoU(Tong等,2023)、Inner-SIoU(Zhang等,2023a)和MPDIoU进行实验对比。实验结果如表2所示,CIoU、EIoU、Wise-IoU、Inner-SIoU和MPDIoU的mAP@0.5值分别为83.9%、84.4%、85.2%、84.3%和85.5%。可以观察到MPDIoU损失函数的性能明显优于其他损失函数,获得了最高的mAP@0.5值85.5%。与原始IoU相比,MPDIoU的mAP@0.5值提高了1.6个百分点,因此,为了达到更好的检测效果,选择MPDIoU作为所提模型的边界框回归损失函数。

表2 不同损失函数的比较

Table 2 Comparison of different loss functions

损失函数	mAP@0.5/%
CIoU	83.9
EIoU	84.4
Wise-IoU	85.2
Inner-SIoU	84.3
MPDIoU	**85.5**

注:加粗字体表示最优结果。

3.5 消融实验

为了验证融合了CSP-PMAC、HWD和IRCB模块对改进模型准确率和速率的有效性,本文在自建行人持刀数据集上进行消融实验,实验结果如表3所示。从表3可以观察到,当单独使用CSP-PMAC模块后,模型的准确度提高1.2%,但是相较于基线模型参数量几乎不变,这得益于提出的并行空洞卷积在获取不同尺度特征时并不会显著增加网络的参数量。当同时使用CSP-PMAC模块和HWD模块后,

表3 消融实验结果

Table 3 Ablation study results

CSP-PMAC	HWD	IRCB	P/%	R/%	mAP@0.5/%	参数量	GFLOPs
–	–	–	87.3	78.1	82.9	19 874 328	56.9
√	–	–	88.5	80.4	84.2	19 876 888	58.2
–	√	–	87.7	78.8	83.4	19 218 968	55.6
√	√	–	88.8	78.9	85.5	19 482 648	57.7
–	–	√	87.5	80.4	84.3	15 283 768	46.8
√	–	√	86.5	76.7	82.7	15 286 328	48.0
–	√	√	88.4	79.3	83.5	**14 628 408**	**45.5**
√	√	√	**90.6**	**83.6**	**89.1**	14 892 088	47.5

注:加粗字体表示各列最优结果,“√”表示采用,“–”表示未采用。

mAP@0.5 提升了 2.6%，同时，精确度提升了 1.5%，充分表明了与原网络结构中的下采样方法相比，HWD 下采样方法更加优越，下采样后保留的特征信息更加丰富。当 3 个模块同时引入后，模型参数量得到了显著下降，参数量较基线模型下降了 25%。最终，本文所提模型的精确度涨至 90.6%，召回率和 mAP@0.5 分别增至 83.6% 和 89.1%。实验发现，每一项改进策略在单独应用时都对原始 RT-DETR 模型有一定的性能提升，尽管在某些情况下仍存在一些挑战，如漏检和置信度问题。然而，当这些改进策略共同作用时，检测性能得到了显著的提升，这充分证明了所提出的改进策略在行人持刀行为检测任务中的有效性。

3.6　对比实验

为了验证本文方法的先进性，在行人持刀数据集上与 SSD、YOLOv3-tiny（Redmon 和 Farhadi，2018）、YOLOv5m、YOLOv8m、Gold-YOLO-s（Wang 等，2023）、YOLOv5s_GCB 和 RT-DETR 这几种主流目标检测方法开展了对比实验。此外，为了评估这些模型在网络摄像头的泛化能力，在由高清网络摄像头采集的行人持刀图像数据集上进行测试，以更全面地展示各模型的性能。各训练好的模型处理测试图像后生成的 Grad-CAM 图像如图 9 所示，不同模型在行人持刀数据集上的对比结果如表 4 所示。不同模型在网络摄像头数据集上的对比结果如表 5 所示。

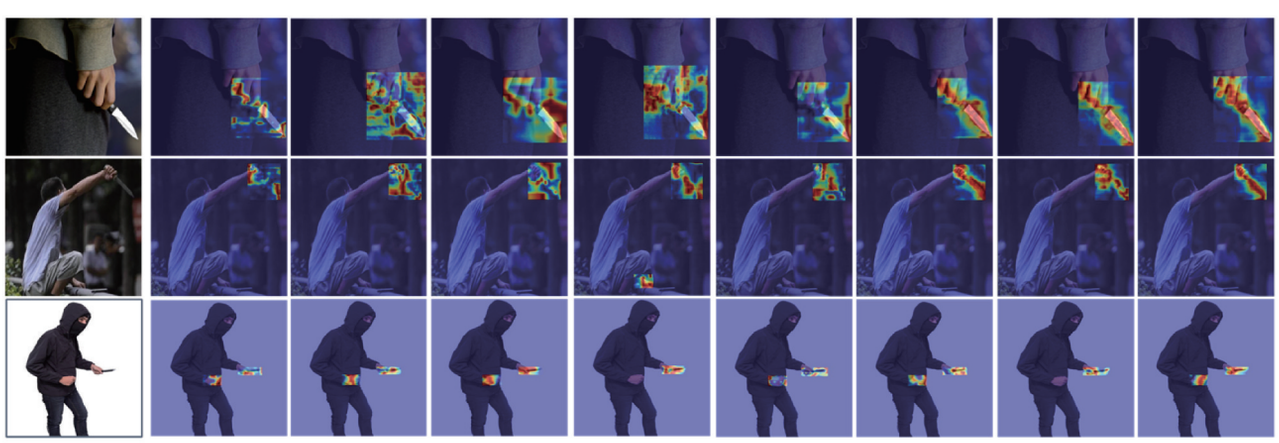


图 9　不同模型输出的 Grad-CAM 图像

Fig. 9　Grad-CAM images from different models

表 4　不同模型在行人持刀数据集上的实验结果

Table 4　Experimental results of different models on the pedestrian knife-carrying dataset

方法	P/%	R/%	mAP@0.5/%	参数量/M	GFLOPs
SSD	67.6	45.5	69.2	–	–
YOLOv5s_GCB	71.2	62.3	68.5	**4.9**	**12.3**
YOLOv3-tiny	77.7	64.6	74.5	12.1	18.9
YOLOv5m	85.8	75	82.8	25	64
YOLOv8m	85.5	76.7	83.9	25.8	78.7
Gold-YOLO-s	89.2	80.5	87.3	21.7	45.6
RT-DETR	88.3	78.1	83.9	19.9	56.9
HK-DETR（本文）	**90.6**	**83.6**	**89.1**	14.9	47.5

注：加粗字体表示各列最优结果，“–”表示未在数据集上测试。

表5　不同模型在网络摄像头数据集上的实验结果

Table 5　Experimental results of different models on the webcam dataset

模型	mAP@0.5/%
YOLOv5m	80.3
YOLOv8m	81.9
Gold-YOLO-s	84.2
RT-DETR	82.7
HK-DETR(本文)	**85.5**

注：加粗字体表示最优结果。

图9中的3个场景经过数个模型测试所呈现的Grad-CAM图像中高亮区域凸显了网络关注的重点。显而易见，HK-DETR模型相较于其他模型，其热力图不仅覆盖了刀具的更大范围，而且亮度更高、焦点更为集中。这得益于CSP-PMAC、IRCB以及HWD下采样模块的同时应用，使得模型能够更精准地锁定目标，体现了模型的高效与精确。然而，刀具检测并非易事，实际场景中的光线不足、行人姿态多样、尺度变化和遮挡等复杂因素都可能影响模型的性能。可通过对多尺度图像训练来应对尺度变化，整合上下文或位置信息来减少遮挡影响，同时优化模型以满足实时需求，从而有效应对这些挑战。

表4数据显示，HK-DETR模型在自建行人持刀数据集上的目标检测精度达到了90.6%，比YOLOv3-tiny高出12.9%，比YOLOv5m高出4.8%，比Gold-YOLO-s高出1.4%，比YOLOv8m高出5.1%，比RT-DETR高出2.3%。在精度方面，HK-DETR模型优势显著。本文方法不仅提高了识别精度，还显著降低了模型参数量，仅为YOLOv8m的57.7%，并且在召回率和mAP等评价指标上也有不同程度的提升。综上所述，本文算法在确保检测精度的同时，轻量化了模型，减少了运算量，提高了检测速度，具有较高的实用价值。

表5数据显示，HK-DETR模型在网络摄像头数据集上的mAP达到了85.5%，表现最佳，这反映了其在网络摄像头下具备一定的泛化能力。

为了进一步展示所提方法在网络摄像头下的检测性能，将基线模型RT-DETR和改进模型HK-DETR在网络摄像头数据集的检测结果进行展示并与实际标注进行对比，如图10所示。图中可以看到，样本2和样本4处RT-DETR模型出现了漏检和误检的错误，而本文模型完成了所需的检测任务。

图10　RT-DETR与HK-DETR（本文模型）网络摄像头数据集测试结果对比

Fig. 10　Comparison of test results on the webcam dataset between RT-DETR and HK-DETR ((a) ground truth; (b) RT-DETR; (c) HK-DETR (ours))

4　结　论

本文提出了一种基于RT-DETR的改进算法，旨在提升行人持刀行为检测的准确性和效率。通过引入IRCB模块，优化了网络的主干结构，使其在处理全局特征和长距离依赖关系时更为高效。同时，CSP-PMAC结构的加入强化了模型对不同尺度刀具的识别能力，而HWD下采样模块则有效保留了刀具边界等关键信息。此外，MPDIoU损失函数的应用进一步提高了模型在边界框定位准确性上的表现。虽然该模型在自建持刀行为数据集上取得了很好的

效果，但在现实场景中，由于数据集的多样性有限，其性能会受到限制。未来的工作应侧重于收集更多复杂人群环境的持刀数据集，反映更广泛的场景因素，以进一步完善模型在复杂和动态环境中的泛化能力。同时，也可以考虑采用无监督或弱监督学习方法，从而达到缓解数据集缺乏的问题。

参考文献（References）

Akcay S and Breckon T. 2022. Towards automatic threat detection: a survey of advances of deep learning within X-ray security imaging. Pattern Recognition, 122: #108245 [DOI: 10.1016/j.patcog.2021.108245]

Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A and Zagoruyko S. 2020. End-to-end object detection with transformers//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 213-229 [DOI: 10.1007/978-3-030-58452-8_13]

Chen Q, Chen X K, Wang J, Zhang S, Yao K, Feng H C, Han J Y, Ding E R, Zeng G and Wang J D. 2023. Group DETR: fast DETR training with group-wise one-to-many assignment//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 6610-6619 [DOI: 10.1109/ICCV51070.2023.00610]

Han K, Wang Y H, Chen H T, Chen X H, Guo J Y, Liu Z H, Tang Y H, Xiao A, Xu C J, Xu Y X, Yang Z H, Zhang Y M and Tao D C. 2023. A survey on visual Transformer. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(1): 87-110 [DOI: 10.1109/TPAMI.2022.3152247]

Li F, Zhang H, Liu S L, Guo J, Ni L M and Zhang L. 2024. DN-DETR: accelerate DETR training by introducing query denoising. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(4): 2239-2251 [DOI: 10.1109/TPAMI.2023.3335410]

Liang J F, Li T, Yang J Q, Li Y N, Fang Z W and Yang F. 2023. Video anomaly detection by fusing self-attention and autoencoder. Journal of Image and Graphics, 28(4): 1029-1040 (梁家菲, 李婷, 杨佳琪, 李亚楠, 方智文, 杨丰. 2023. 融合自注意力和自编码器的视频异常检测. 中国图象图形学报, 28(4): 1029-1040) [DOI: 10.11834/jig.211147]

Liao Y, Li Z M and Liu S. 2022. A review of deep learning based human-object interaction detection. Journal of Image and Graphics, 27(9): 2611-2628 (廖越, 李智敏, 刘偲. 2022. 基于深度学习的人—物交互关系检测综述. 中国图象图形学报, 27(9): 2611-2628) [DOI: 10.11834/jig.211268]

Liu S L, Li F, Zhang H, Yang X, Qi X B, Su H, Zhu J and Zhang L. 2022. DAB-DETR: dynamic anchor boxes are better queries for DETR [EB/OL]. [2024-05-18]. https://arxiv.org/pdf/2201.12329.pdf

Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C Y and Berg A C. 2016. SSD: single shot MultiBox detector//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 21-37 [DOI: 10.1007/978-3-319-46448-0_2]

Liu X Y, Peng H W, Zheng N X, Yang Y Q, Hu H and Yuan Y X. 2023. EfficientViT: memory efficient vision transformer with cascaded group attention//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 14420-14430 [DOI: 10.1109/CVPR52729.2023.01386]

Long Z R, Cai L F, Ye B Q, Tang B, Zhao M F, Tang Y L, Wang J X, and Zhou M. 2023. Research on improved YOLOv5s lightweight target detection algorithm. Journal of Chongqing University of Technology (Natural Science Edition), 37(23): 244-251 (龙邹荣, 蔡林峰, 叶彬强, 汤斌, 赵明富, 唐跃林, 王建旭, 周密. 2023. 改进YOLOv5s的轻量化目标检测算法研究. 重庆理工大学学报(自然科学), 37(23): 244-251 [DOI: 10.3969/j.issn.1674-8425(z).2023.12.028]

Ma S L and Xu Y. 2023. MPDIoU: a loss for efficient and accurate bounding box regression [EB/OL]. [2024-05-18]. https://arxiv.org/pdf/2307.07662.pdf

Redmon J, Divvala S, Girshick R and Farhadi A. 2016. You only look once: unified, real-time object detection//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 779-788 [DOI: 10.1109/CVPR.2016.91]

Redmon J and Farhadi A. 2018. YOLOv3: an incremental improvement [EB/OL]. [2024-05-18]. https://arxiv.org/pdf/1804.02767.pdf

Ren S Q, He K M, Girshick R and Sun J. 2017. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6): 1137-1149 [DOI: 10.1109/TPAMI.2016.2577031]

Tong Z J, Chen Y H, Xu Z W and Yu R. 2023. Wise-IoU: bounding box regression loss with dynamic focusing mechanism [EB/OL]. [2024-05-18]. https://arxiv.org/pdf/2301.10051.pdf

Wang C C, He W, Nie Y, Guo J Y, Liu C J, Han K and Wang Y H. 2023. Gold-YOLO: efficient object detector via gather-and-distribute mechanism//Proceedings of the 37th International Conference on Neural Information Processing Systems. Red Hook, USA: Curran Associates Inc.: 51094-51112

Wang C Y, Liao H Y M, Wu Y H, Chen P Y, Hsieh J W and Yeh I H. 2020. CSPNet: a new backbone that can enhance learning capability of CNN//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Seattle, USA: IEEE: 1571-1580 [DOI: 10.1109/CVPRW50498.2020.00203]

Wang Y M, Zhang X Y, Yang T and Sun J. 2022. Anchor DETR: query design for Transformer-based detector//Proceedings of the 36th AAAI Conference on Artificial Intelligence. Palo Alto, California USA: AAAI, 2022: 2567-2575 [DOI: 10.1609/aaai.v36i3.20158]

Xu G P, Liao W T, Zhang X, Li C, He X W and Wu X L. 2023. Haar wavelet downsampling: a simple but effective downsampling module for semantic segmentation. Pattern Recognition, 143: #109819 [DOI: 10.1016/j.patcog.2023.109819]

Yang Z M, Wang X L and Li J G. 2021. EIoU: an improved vehicle detection algorithm based on VehicleNet neural network. Journal of Physics: Conference Series, 1924(1): #012001 [DOI: 10.1088/1742-6596/1924/1/012001]

Yu F and Koltun V. 2015. Multi-scale context aggregation by dilated convolutions [EB/OL]. [2024-05-18]. https://arxiv.org/pdf/1511.07122.pdf

Zhang H, Xu C and Zhang S J. 2023a. Inner-IoU: more effective intersection over union loss with auxiliary bounding box [EB/OL]. [2024-05-18]. https://arxiv.org/pdf/2311.02877.pdf

Zhang J N, Li X T, Li J, Liu L, Xue Z C, Zhang B S, Jiang Z K, Huang T X, Wang Y B and Wang C J. 2023b. Rethinking mobile block for efficient attention-based models//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 1389-1400 [DOI: 10.1109/ICCV51070.2023.00134]

Zhao Y, Lv W Y, Xu S L, Wang G Z, Wei J M, Dang Q Q, Liu Y and Chen J. 2023. DETRs beat YOLOs on real-time object detection [EB/OL]. [2024-05-18]. https://arxiv.org/pdf/2304.08069.pdf

Zheng Z H, Wang P, Ren D W, Liu W, Ye R G, Hu Q H and Zuo W M. 2022. Enhancing geometric factors in model learning and inference for object detection and instance segmentation. IEEE Transactions on Cybernetics, 52(8): 8574-8586 [DOI: 10.1109/tcyb.2021.3095305]

Zhu X Z, Su W J, Lu L W, Li B, Wang X G and Dai J F. 2020. Deformable DETR: deformable transformers for end-to-end object detection [EB/OL]. [2024-05-18]. https://arxiv.org/pdf/2010.04159.pdf

作者简介

金涛，男，副教授，主要研究方向为计算机视觉、边缘计算和人工智能机器人技术。E-mail: jt6642@gsupl.edu.cn

胡配雨，通信作者，男，硕士研究生，主要研究方向为计算机视觉、目标检测与识别和深度学习。

E-mail：18225849687@163.com