DOI: 10.19650/j.cnki.cjsi.J2513963

多模态信息融合下的监控视频人员身份重识别*

吴 军[1],陈 慧[1],徐 刚[2],赵雪梅[1],陈睿星[2]
(1.桂林电子科技大学电子工程与自动化学院 桂林 541004; 2.中国科学院宁波材料技术与工程研究所 宁波 315201)

摘 要:针对目前监控视频人员身份重识别任务难以有效应对低分辨率小目标、人员姿(形)态变化及遮挡检测问题,以 YOLOv9 为基础网络并结合多模态预训练神经网络(CLIP)模型提出一种多模态信息融合下的监控视频人员身份识别新方法,主要涉及两个方面工作:1)通过引入感受野增强模块与可变形卷积计算提高目标人员不同姿态(形)态下的特征检测性能、引入空间增强注意力机制学习特征间的关系以恢复被遮挡人员特征、引入基于归一化高斯距离的损失度量以增强低分辨率目标人员特征检测敏感性等系列模块设计,构建网络 ReID-YOLO 有效增强监控视频在不同姿态、形态及低分辨率、遮挡条件下的目标人员特征检测精度、鲁棒性;2)将 CLIP 跨模态信息融合优势迁移到视频人员身份重识别任务并利用 CLIP 图像-文本信息对齐优势对前一阶段提取的人员目标特征进行身份预测,在利用 ReID-YOLO 人员视觉特征有效区分能力缓解 CLIP 全局场景过度依赖的同时,借助 CLIP 模型场景泛化优势有效克服 YOLO 系列网络在整合场景信息深入解析目标方面的不足,从而整体提高网络模型的监控视频人员身份重识别精度与场景泛化能力。在低分辨率、消融与身份重叠等条件下的实验结果表明,所提方法视频人员身份重识别性能表现出色,优于 YOLO 系列网络及其他 7 个主流的视频人员身份重识别网络模型,具有良好应用前景。

关键词: 视频监控;人员身份识别;YOLO 目标检测;多模态模型 CLIP
中图分类号: TP391.41 TH74 **文献标识码**: A **国家标准学科分类代码**: 420.20

Surveillance video person re-identification under multi-modal information fusion

Wu Jun[1], Chen Hui[1], Xu Gang[2], Zhao Xuemei[1], Chen Ruixing[2]
(*1. School of Electronic Engineering and Automation, Guilin University of Electronic Technology, Guilin 541004, China*;
2. Ningbo Institute of Materials Technology and Engineering, Chinese Academy of Sciences, Ningbo 315201, China)

Abstract: Addressing the challenges of low resolution, severe occlusion and significant changes in personnel pose or shape variations, this paper proposes a new method for personnel re-identification (PR) in surveillance videos based on multimodal information fusion, using YOLOv9 as the backbone network and combining it with The Multi-Modal model CLIP (contrastive language-image pre-training). The method is divided into two stages. In the first stage, a ReID-YOLO network is constructed to enhance person feature detection performance under challenging conditions. A receptive-field enhancement module and deformable convolution are introduced to improve feature extraction for personnel with diverse poses and shapes. A spatially enhanced attention mechanism is employed to model relationships among person features and restore occluded information. In addition, a normalized Gaussian distance-based loss function is designed to increase sensitivity to low-resolution person features. These strategies jointly improve the accuracy and robustness of person feature detection in surveillance videos affected by low resolution, pose variation, shape deformation, and occlusion. In the second stage, the Multi-Modal model CLIP is introduced to improve the overall accuracy and scene generalization ability. By leveraging CLIP′s image-text alignment ability, personnel targets extracted in the first stage are predicted using discriminative features provided by ReID-YOLO. This fusion strategy mitigates CLIP′s excessive reliance on global scene information while compensating for the limited scene-awareness and target semantic parsing capability of YOLO-based networks. Experimental results under challenging conditions such as low

收稿日期:2025-04-23 Received Date: 2025-04-23
*基金项目:国家自然科学基金(42361071,42261061)、广西重点研发计划(桂科 FN2504240020)、宁波市科技计划(2024Z016)项目资助

resolution, ablation studies, and cross-identity scenarios demonstrate that the proposed method achieves outstanding performance in video-based person re-identification. It outperforms YOLO-series networks and seven other state-of-the-art video re-identification models, showing considerable promise for practical applications.

Keywords: video surveillance; person re-identification; YOLO object detection; multimodal model CLIP

0　引　　言

视频人员身份重识别(personnel re-identification, PR)旨在关联不同时刻、不同视角场景视频中同一目标人员的身份[1],在交通[2]、安防[3-4]、农业生产[5]等众多领域具有重要应用价值。本质上,视频 PR 属于特征匹配问题,其基本思想是学习目标人员的视频特征并使同一身份人员在特征空间中接近、不同身份人员则彼此远离[6],该思想与度量学习在行人重识别中的典型目标一致,即通过学习判别性距离度量优化特征空间结构[7]。然而,对于环境复杂的实际监控场景,因视频中人物外观特征受监控视角、光照条件、影像分辨率、背景变化及运动姿态、目标遮挡等干扰因素影响[8],包括卷积神经网络在内的单模态特征匹配方法难以保证目标人员影像特征捕获的一致性、连续性,监控视频 PR 任务实际应用仍面临诸多挑战[9]。已有研究表明,基于梯度算子、注意力机制或多模态证据融合的方法能够在复杂环境下提升特征稳定性与判别性[10]。近年来,无监督行人重识别研究中,钱亚萍等[11]尝试通过多分支结构与局部-全局特征共享提升特征一致性,但仍受到光照与视角变化影响。程思雨等[12]提出的细粒度特征增强与注意力机制虽能在局部区域对齐方面具备优势,但在动态监控环境中仍易受遮挡与外观漂移干扰。胡玉玲等[13]采用的无监督特征对齐方法也受到复杂场景干扰影响。而邓子文等[14]基于深度聚类学习,通过改进特征提取质量、优化密度聚类超参数并引入聚类级记忆字典,有效提升了复杂场景下的识别精度。

从监控场景角度出发,现有的监控视频 PR 网络可概述分为两类:特殊场景视频 PR 和开放场景视频 PR。前者侧重于在环境背景相对单一、光照条件可控且行人姿态相对稳定的场景下进行身份匹配,如室内走廊、体育场馆等某些特定出入口,Mclaughlin 等[15]指出该类场景中摄像机视角固定,常依赖稳定的全局外观特征实现身份匹配。Ristani 等[16]针对多目标多摄像机场景,提出了统一的性能评估指标体系,并构建了用于跨摄像机目标关联的公开数据集,为复杂监控环境下的多目标跟踪与跨视角身份关联研究提供了标准化评测基础;Dou 等[1]提出了一种以身份判别为核心的自监督行人表征学习方法,通过在无人工标注条件下挖掘潜在身份一致性约束,引导模型学习更具判别性的身份特征,从而有效提升行人重识别模型在跨场景和未知域条件下的泛化能力;Wu 等[17]针对采用双流网络学习红外与可见光行人图像的共享特征空间,并利用对比约束缩小跨模态特征差异,实现红外-可见光行人重识别。此类方法因依赖强环境假设而难以泛化。开放场景视频 PR 则面向光照不可控、背景复杂、姿态动态变化的城市街道与交通枢纽等场景,因背景复杂、光照变化大且跨摄像头外观漂移显著,在“野外场景”下的识别精度显著下降[18],Han 等[8]指出跨时空、跨模态干扰会引发显著人员外观特征漂移,导致单模态特征判别性下降;Dou 等[1]通过构建身份感知的自监督学习框架,在特征空间中显式增强同一身份样本之间的一致性约束,使模型能够在缺乏人工标注的情况下学习鲁棒的身份表征,从而提升行人重识别任务的跨域泛化能力;Wu 等[17]融合可见光、红外与深度信息,并借助生成对抗网络补全遮挡区,但仍面临跨模态对齐冗余与长时序关联不稳定等问题[19]。

目前,YOLO 系列网络在视频目标检测方面表现出良好性能,其中:YOLOv9 将通用高效层聚合与加权 Transformer 编码器相结合,使得网络可在非 Transformer 架构效率与 Transformer 全局建模能力之间寻求最优平衡[20];Khanam 等[21]基于 YOLOv8 提出轻量化 YOLOv11,通过简化骨干网络与轻量级注意力大幅降低模型部署门槛;Tian 等[22]提出 YOLOv12 深度融合视觉 Transformer-跨阶段部分连接骨干模块(vision transformer-cross stage partial, ViT-CSP)与动态特征选择 Neck 模块以聚焦多尺度、强遮挡及极端尺度目标的高精度检测等。相比较而言,YOLOv9 优势在于其轻量化全局建模可有效解决局部特征对复杂场景的适配不足问题(如密集人群、重叠目标、相似对象等),是一款均衡型通用目标检测网络;YOLOv11 为适配边缘部署舍弃了 Transformer 模块以获得极致轻量化与高速推理的优势,导致该网络在处理重叠率较高且分辨率较低的数据时,因缺乏全局建模模块而使其误检率会大幅上升;YOLOv12 的设计初衷是适配大规模复杂工业级高端场景,为保证检测精度,仅支持使用高性能图形处理器(graphics processing unit, GPU)或专用人工智能(artificial intelligence, AI)芯片在海量高质量标注数据集上进行模型训练与推理,存在应用的局限性,同时,也有研究通过轻量化与多尺度融合改进 YOLO 系列网络以适应嵌入式场景需求[23]。

整体上,由于 YOLO 系列网络侧重于局部影像特征提取或有限上下文关系、简单语义约束的利用,在整合场

景信息进行深入解析方面仍存在不足,导致其场景泛化能力受限。与此同时,多模态信息融合技术及其预训练模型的出现为解决上述监控视频 PR 问题提供了新的研究方向,以对比语言-图像预训练模型(comparative language image pretraining,CLIP)为例,该模型通过图像-文本间的对比学习展现出强大的多模态理解、场景泛化能力[24]。然而,由于监控场景背景复杂、人员特征不稳定及光照、视角变化因素,传统多模态模型利用方式对于监控视频 PR 存在两方面限制:1)直接处理场景整体信息并进行细粒度特征描述易导致模型无法有效区分场景中的少数关键特征[25](如存在长尾分布的人员目标);2)多模态模型需整合视觉、文本信息以构建统一特征表示,而这种对全局场景的过度依赖导致使其在泛化性能和识别精度上难以满足实际需求。多模态证据融合研究也表明合理设计跨模态特征共享机制有助于提升多模态任务下的稳定性[26]。

针对以上问题,本研究以 YOLOv9 为基础网络并引入 CLIP,提出一种多模态信息融合下的监控视频 PR 新方法,其创新之处主要在于两个方面:1)引入感受野增强模块与可变形卷积计算、空间增强注意力模块及基于归一化高斯距离的损失度量来构建网络 ReID-YOLO(re-identification with YOLO),有效增强视频中目标人员在不同姿态、形态及低分辨率、遮挡条件下的特征检测精度和鲁棒性;2)利用 CLIP 跨模态信息融合优势对前一阶段的人员目标特征成功进行身份(文本)预测,既借助 ReID-YOLO 对人员视觉特征的区分能力缓解 CLIP 对全局场景的过度依赖,又借助 CLIP 模型场景泛化优势有效克服 YOLO 系列网络在整合场景信息深入解析目标方面的不足,从而整体提高网络模型的监控视频 PR 精度与场景泛化能力。

1 人员目标检测网络 ReID-YOLO

本研究以 YOLOv9 为基础网络建立如图 1 所示的监控视频人员目标检测网络 ReID-YOLO,其主要包括 3 个功能模块:1)抗姿态变化检测模块,用于增强视频中目标人员在姿态、形状发生较大变化下的特征检测能力;2)抗遮挡检测模块,用于改善视频中目标人员在遮挡条件下出现的漏检或误检情况;3)低分辨率检测模块,用于增强视频中目标人员在低分辨率条件下的小目标特征检测敏感性。YOLOv9 主要由骨干网络(backbone)、颈部网络(neck)和检测头(detect head)组成,上述 3 个功能模块位于不同部分,其中位于骨干网络与颈部模块间的跨块特征融合机制(cross-block feature fusion,CBFuse)用于实现高层语义信息与低层细节信息充分融合目的。

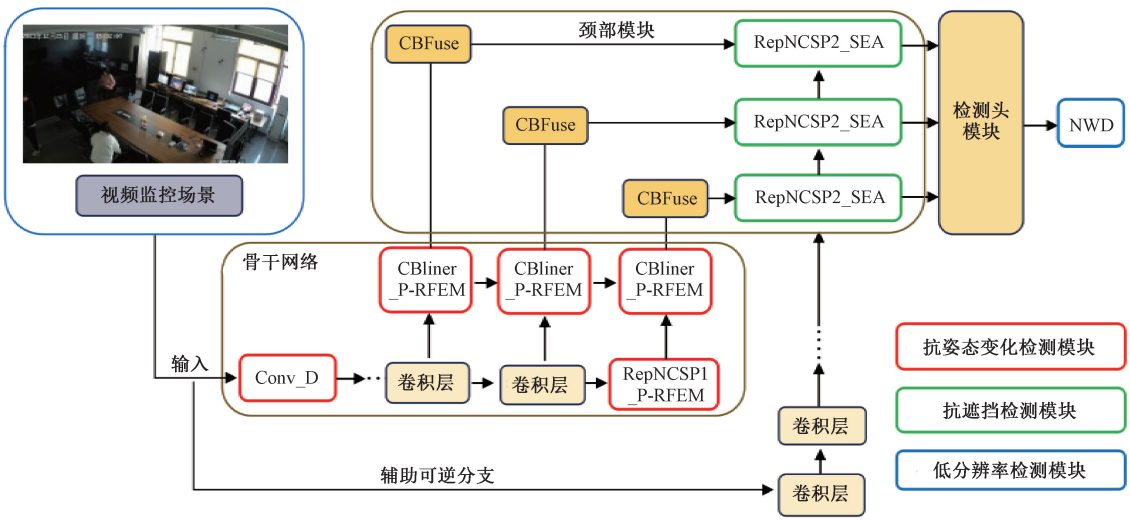


图 1 ReID-YOLO 网络结构

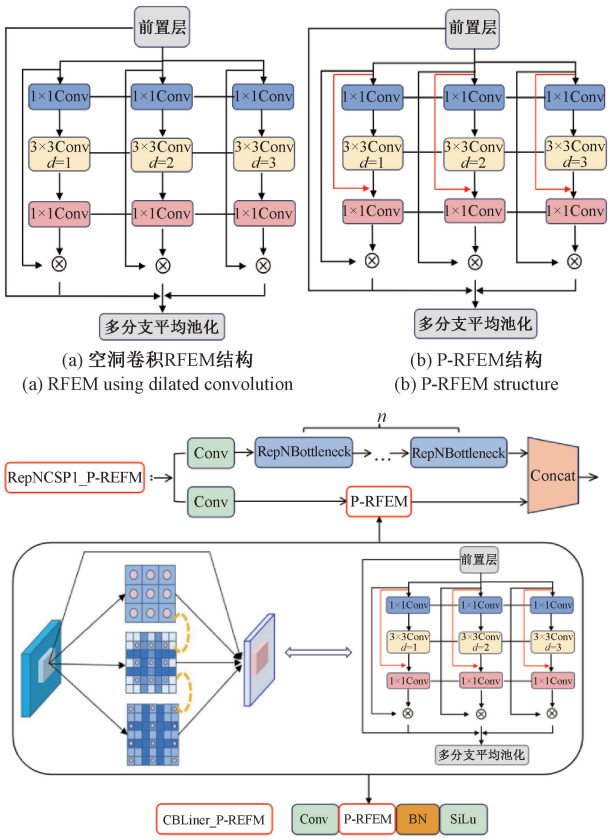
Fig. 1 ReID-YOLO network structure

1.1 抗姿(形)态变化检测

开放场景中目标人物的跑跳、转身、蹲下等各种自然动作会导致其形状、外观特征发生显著差异,这种差异导致原 YOLOv9 网络在面对人物发生姿态、形状变化时存在锚框定位不准确问题,即目标形态、边界与锚框模型所预设的检测框尺度、形状不匹配[27]。针对这一问题,本研究通过引入感受野增强与可变形卷积来实现目标人物的抗姿(形)态变化检测。

受 Inception[27] 与 ResNet[28] 两种架构启发,一些目标检测网络引入感受野增强模块(receptive field enhancement module,RFEM)[27] 以丰富用于预测分类和定位目标特征的感受野信息,并通过插入空洞到卷积核元素间在不增加参数量前提下进一步增加感受野。图 2(a)给出了使用空洞卷积的 RFEM 模块结构,该结构

主要由两部分组成[24]:膨胀卷积多分支和聚集加权层。前者用于捕获多尺度感受野信息及其内部特征依赖关系,使用 3 个不同扩展率的膨胀卷积分支进行卷积操作来捕获多尺度信息,各分支共享权重以减少参数量并使用残差连接以防止梯度爆炸;后者用于收集来自不同分支的信息并为每个分支的特征加权[25],各分支特征经过平均池化层生成具有更大感受野和更丰富上下文信息的特征图。然而,单纯使用空洞卷积可能带来梯度消失或优化困难。计划重参数化卷积结构(planned re-parameterized, PP)[26]设计者认为:当内部最末端的子网络模块引入残差连接时,会与外部主网络模块的残差连接起冲突,从而对整体残差结构产生负面影响[28]。吸收该设计思想,本研究对使用空洞卷积的 RFEM 进行优化并提出计划的感受野增强模块(person-receptive field enhancement module,P-RFEM),见图 2(b),通过使用膨胀卷积充分利用特征图中感受野优势、通过使用合理的残差连接增强梯度反向传播的稳定性并提高特征复用能力,在保持原空洞卷积感受野扩展优势的同时有效缓解梯度消失。



(a) 空洞卷积RFEM结构
(a) RFEM using dilated convolution

(b) P-RFEM结构
(b) P-RFEM structure

(c) 基于P-RFEM的目标人物检测模块
(c) P-RFEM based personnel detection module

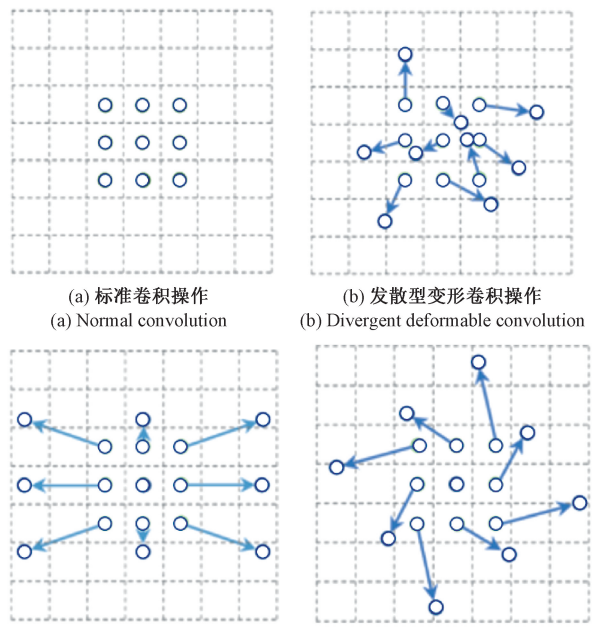
图 2　基于 P-RFEM 的目标人物抗姿态变化检测模块结构

Fig. 2　Personnel feature detection module for resisting posture changes

进一步的,本研究基于 P-RFEM 构建人物抗姿态变化检测模块:RepNCSP1_P-RFEM 和 CBLiner_P-RFEM,如图 2(c)所示,其中:前者旨在充分利用膨胀卷积的感受野扩展优势在深层特征映射阶段加强特征提取与融合,并通过残差结构优化特征传递与梯度传播,进而提高模型对多姿态人员目标的检测能力;后者旨在通过增强多重感受野的特征聚合能力以进一步扩展模型全局感受野,从而在深层特征映射阶段捕捉不同姿态变化的目标时具有更强适应性;模块 RepNBottleneck 用于实现深层语义特征的提取与压缩,为模块 RepNCSP1_P-RFEM 提供更充分的语义支持。两者为极端姿态比例下的人物目标检测可提供多样化的感受野信息,并通过多尺度特征融合提升特征表示能力,从而有助于提升模型多姿态人员目标感知、特征检测能力。

采用固定大小、形状的传统卷积核提取特征难以灵活适应人员姿态动态变化所引起的目标形变,导致目标与锚框的匹配精度下降[29]。针对这一问题,本研究引入可变形卷积(deformable convolution,DCN)以更好地适应目标形变[25]。

与传统卷积操作在网格上仅能提取到矩形框特征不同,DCN 优点在于更准确地提取不同形状的特征,一方面原因在于其拓展了尺度、长宽及旋转变换,使得其卷积位置如图 3 所示是可变形的,其中:图 3(a)为普通卷积操作,图 3(b)~(d)是可变形卷积,图 3(c)、(d)是图 3(b)的



(a) 标准卷积操作
(a) Normal convolution

(b) 发散型变形卷积操作
(b) Divergent deformable convolution

(c) 米字型变形卷积操作
(c) Rice shaped deformable convolution

(d) 旋转型变形卷积操作
(d) Rotating deformable convolution

图 3　普通卷积与可变形卷积对比

Fig. 3　Comparison between standard convolution and deformable convolution

特例。另一方面,传统卷积核权重的固定导致其在处理同一张图不同区域时感受野尺寸相同,而不同区域可能对应不同尺度或不同形变的物体,因此,考虑到目标检测效果很大程度上依赖于用于特征提取的边界框,传统卷积核并非最优,尤其是对于非网格状目标而言。相反的,DCN 能自动调整感受野尺寸以更好地捕捉目标物体的空间结构信息,其关键在于通过引入空间位置偏移向量 $\Delta \boldsymbol{p}_n$ 到传统卷积计算以使其在执行特征提取时可动态调整其感受区域的位置和形状,从而提高模型捕捉动态形变目标的能力。

DCN 计算过程如式(1)所示。

$$y(\boldsymbol{p}_0)=\sum_{\boldsymbol{p}_n\in \mathrm{R}} \cdot w(\boldsymbol{p}_n) \cdot x(\boldsymbol{p}_0+\boldsymbol{p}_n+\Delta \boldsymbol{p}_n) \tag{1}$$

式中:$\{\Delta \boldsymbol{p}_n \mid n=1,2,\cdots,N\}$,$N=|\mathrm{R}|$,$\boldsymbol{p}_n$ 列举了 R 中的所有位置;$\boldsymbol{p}_0+\boldsymbol{p}_n+\Delta \boldsymbol{p}_n$ 表示新位置的坐标,$x(\boldsymbol{p}_0+\boldsymbol{p}_n+\Delta \boldsymbol{p}_n)$ 表示取出该位置的像素值;$w(\boldsymbol{p}_n)$ 表示卷积核相对该位置像素的权重值。迭代相加所有位置的值 $w(\boldsymbol{p}_n)\cdot x(\boldsymbol{p}_0+\boldsymbol{p}_n+\Delta \boldsymbol{p}_n)$ 输出结果像素坐标值。进一步的,本研究基于上述 DCN 构建如图 4 所示的目标人物抗形态变化检测模块 Conv_D,该模块是对基础网络 YOLOv9 中的卷积(Conv)模块进行改进以使卷积层不再仅仅依赖于预设的规则化网格,而是根据目标的实际形态灵活调整其感受区域,从而能有效应对目标动态姿态变化问题。

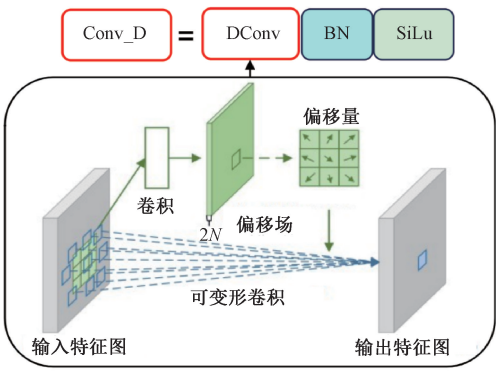


图 4 基于 DCN 的目标人物抗形态变化检测模块结构

Fig. 4 DCN-based personnel feature detection module resisting morphological changes

1.2 抗遮挡检测

监控场景中普遍存在人员部分被遮挡或部分重叠导致的特征丢失问题[29],其中:前者导致基础网络采用锚框机制时错误地将遮挡区域视为背景的一部分,从而发生误检[30];后者导致基础网络检测精度对非极大值抑制[31]阈值非常敏感,从而发生漏检[32]。针对上述问题,本研究引入结构如图 5 所示的空间增强注意力(spatially enhanced attention,SEA)[33]机制以引导网络关注未遮挡部分的关键信息,其中:通道和空间混合模块(channel-spatial mixed module,CSMM)用于实现多尺度目标检测,通过对不同块的多尺度特征处理来学习空间维度和通道间的相关性[34]。

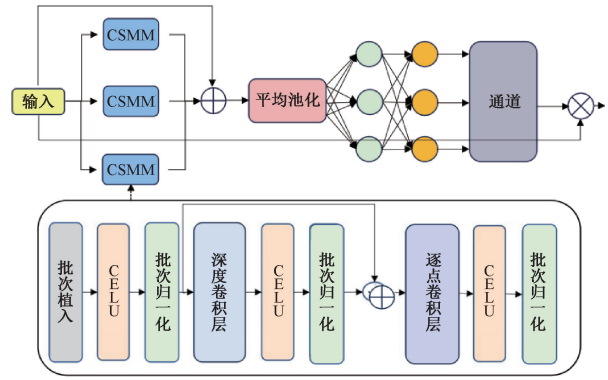


图 5 SEA 结构示意图

Fig. 5 Structure of the spatially enhanced attention (SEA) module

SEA 本质上利用特征图之间的关系来恢复被遮挡的特征,并通过增强未遮挡目标的响应来补偿被遮挡目标的响应损失[33]。本研究构建如图 6 所示的目标人物抗遮挡检测模块:RepNCSP2_SEA,旨在特征融合和多尺度信息提取过程中更好地整合高、低层级特征,从而使模型能更精准地捕捉不同尺度目标的关键信息以提升遮挡目标的检测能力。具体而言,模块整体采用双分支设计,其中:模块 RepNBottleneck 作用同上;辅助分支引入 SEA 可在初步卷积处理后自适应调整特征图中不同空间位置的重要性以引导网络更加关注未被遮挡的区域,从而减轻遮挡区域特征缺失对整体目标识别的影响。进一步的,两分支融合后深层特征与增强空间信息间的结合,不仅能提升目标定位的准确性,并有助于减少无效遮挡特征对目标检测的干扰,从而提升实际场景尤其是人员密集、环境复杂等情况下的人员目标检测精度及抗遮挡的鲁棒性。

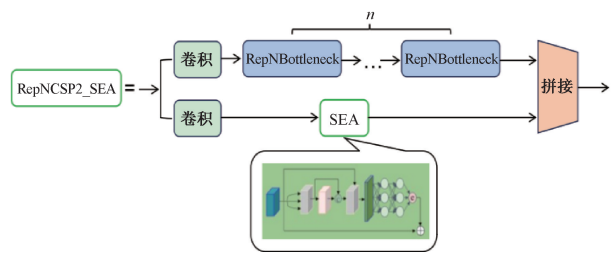


图 6 基于 SEA 的目标人物抗遮挡检测模块结构

Fig. 6 SEA-based personnel feature detection module resisting occlusion

1.3 低分辨率检测

当前目标检测模型性能评估多采用 IoU(intersection over union)指标,但该指标对低分辨率小目标的位置偏差

非常敏感。以方框 A 表示目标的真实边界框、方框 B 表示沿对角线偏差的预测边界框,以横坐标表示方框 A、B 的像素偏差值(个数)、纵坐标表示两者的 IoU 度量值,图 7(a)给出了不同分辨率目标的 IoU 度量变化曲线,其中:方框 A 与 B 的尺度始终保持一致,右上角数值代表方框边长。

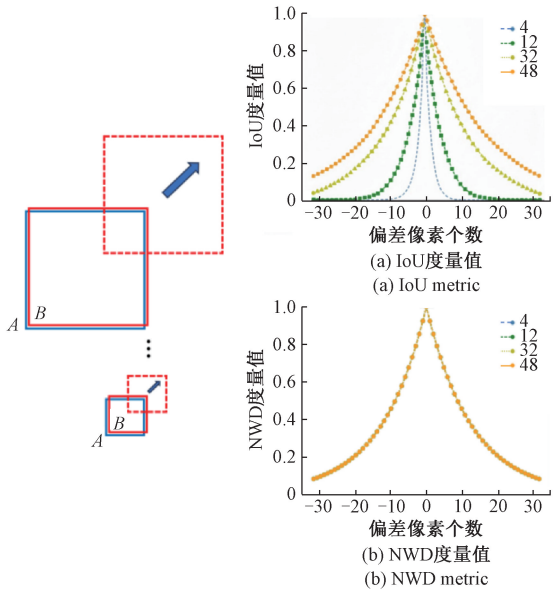


图 7　不同度量下的偏差曲线

Fig. 7　Deviation curve under different measurement

由图 7(a)可看出,相同位置偏差下不同尺度物体的 IoU 差异很大:对于 4 pixels×4 pixels 的低分辨率小目标,微小位置偏差导致 IoU 骤降(偏移一个像素值,从 1 降到 0.55);对于 48 pixels×48 pixels 的正常目标,相同的位置偏差下 IoU 略有下降(从 1 降到 0.95)。上述现象产生的原因有两方面:一方面在于目标边界框位置离散变化的特殊性[24],使得低分辨率小目标的微小位置偏差都会引起锚框标记反转,从而导致其正样本间特征高度相似、用于小目标检测训练的监督信息不足[35];另一方面,虽然动态匹配策略能根据目标的统计特性自适应地获得分配正负样本的 IoU 阈值,但 IoU 对低分辨率目标位置偏差的敏感性使得难以找到一个可提供高质量正负样本的良好阈值[36],故 IoU 用于基于锚框的目标检测器中会大大降低性能。

针对上述问题,本研究以归一化高斯距离(normalized gaussian wasserstein distance,NWD)度量取代标准 IoU 来衡量边界框的相似性,其基本思想是将边界框建模为二维高斯分布并在 NWD 度量下以对应的高斯分布来计算边界框间的相似性[37]。具体而言,边界框的 NWD 计算过程如式(2)所示。

$$W_2^2(N_a,N_b)=\left\|\left(\left[cx_a,cy_a,\frac{w_a}{2},\frac{h_a}{2}\right]^{\mathrm{T}},\left[cx_b,cy_b,\frac{w_b}{2},\frac{h_b}{2}\right]^{\mathrm{T}}\right)\right\|_2^2 \tag{2}$$

式中:N_a 和 N_b 表示边界框 $A=(cx_a,cy_a,w_a,h_a)$ 和边界框 $B=(cx_b,cy_b,w_b,h_b)$ 建模的高斯分布;cx_a、cy_a、w_a、h_a 分别表示边界框的中心坐标、宽度和高度。由于 $W_2^2(N_a,N_b)$ 是一个距离度量,这里使用其指数形式归一化来获得 NWD,其计算过程如式(3)所示。

$$NWD(N_a,N_b)=\exp\left(-\frac{\sqrt{W_2^2(N_a,N_b)}}{C}\right) \tag{3}$$

由上述 NWD 计算可知,其描述边界框中不同像素权重时,边界框中心像素权重最高,其他像素权重从中心到边界逐渐递减,这也就意味着,NWD 可用于度量非重叠小目标间的分布相似性,从而有效避免了对目标边界位置偏差的敏感性。不同分辨率目标的 NWD 度量变化曲线如图 7(b)所示,从图中可看出,与图 7(a)中 IoU 度量小目标出现的曲线骤降现象不同,NWD 的 4 条曲线完全重合,表明 NWD 对于目标尺度不敏感;此外,NWD 度量易嵌入任何基于锚框检测器的正负样本匹配过程,并有助于获得区分小目标正负样本的良好阈值。

2　CLIP 引导的目标人员身份精确识别

旨在将 ReID-YOLO 提取的人员身份局部影像特征输入多模态预训练模型 CLIP 并利用其强大的对比学习能力将人员身份局部影像特征与表征其身份类别的文本特征在同一嵌入空间中对齐,从而自动建立两者映射关系以实现目标人员身份精准判定,其整体工作流程如图 8 所示,主要包含两部分处理:

1)标签文本处理。CLIP 模型通过预训练使图像和相应的文本描述在一个共享的特征空间中对齐。具体而言,待识别人员对应的标签文本构建过程如式(4)所示。

$$T_{\mathrm{body},i}=\{\text{a photo of ID}i\} \tag{4}$$

式中:IDi 为人员身份标识符;$T_{\mathrm{body},i}$ 表示与人员身份 IDi 对应的标签文本,该标签文本提供了人员身份的自然语言描述并在模型预训练过程中作为语义描述与图像信息进行跨模态融合。

2)跨模态融合处理。将第 1 章中 ReID-YOLO 定位出的人员局部图像区域 I_{body} 输入 CLIP 的视觉编码器,从而在高维潜空间中获得一个包含局部区域图像特征的向量表示 F_{body},其计算过程如式(5)所示。

$$F_{\mathrm{body}}=E_v(I_{\mathrm{body}}),I_{\mathrm{body}}=Crop(I,B_{\mathrm{body}}) \tag{5}$$

式中:$B_{\mathrm{body}}\in \mathrm{R}^{H\times W\times X\times Y}$ 表示人员局部图像区域的边界框,H 与 W 分别为局部图像区域的宽度与高度,X 与 Y 分别为人员局部区域的中心坐;E_v 表示 CLIP 的视觉编码器,通常是基于 ResNet-50 或 ViT-B/16 的卷积神经网络;$F_{\mathrm{body}}\in \mathrm{R}^d$ 是 E_v 输出的特征向量,d 为特征空间的维度,该特征向量涵盖目标人员步态特征、上下肢变化等多种准确识别人员身份的重要信息。

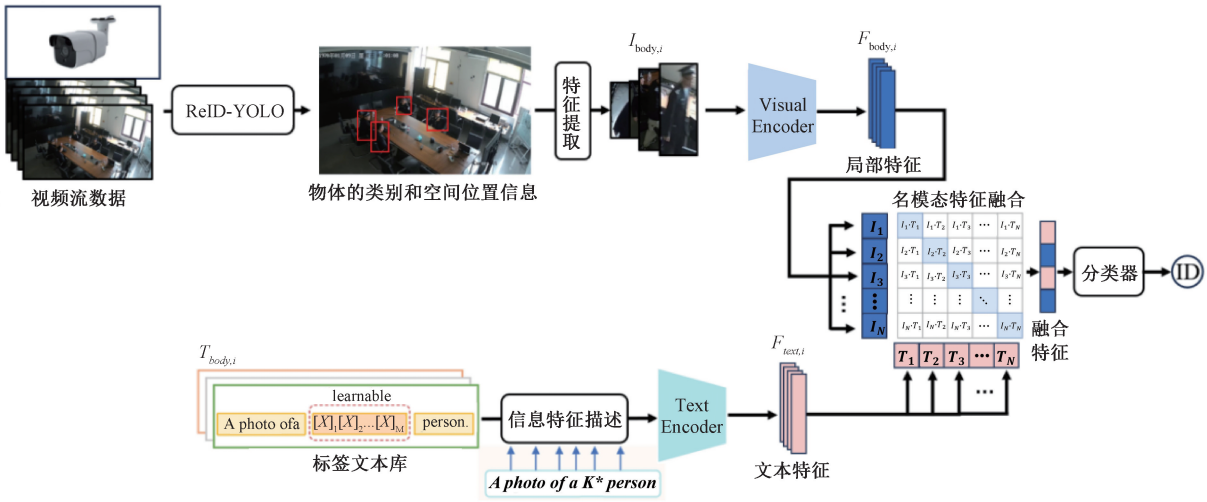


图 8 CLIP 引导下的人员身份识别流程图

Fig. 8 Diagram of personnel identification guided by CLIP

图 9 为某时刻视频人员局部视觉特征与其身份文本标签语义信息的跨模态融合可视化结果,从中可看出,融合后的信息相比于融合前无疑增强了人员视觉特征的表达的一致性、完整性,从而有助于人员身份的精确判定。上述跨模态融合处理的关键在于将从视频图像中提取的人员局部视觉特征与对应的身份文本标签进行有效匹配。为此,CLIP 模型将把每个标签文本输入基于

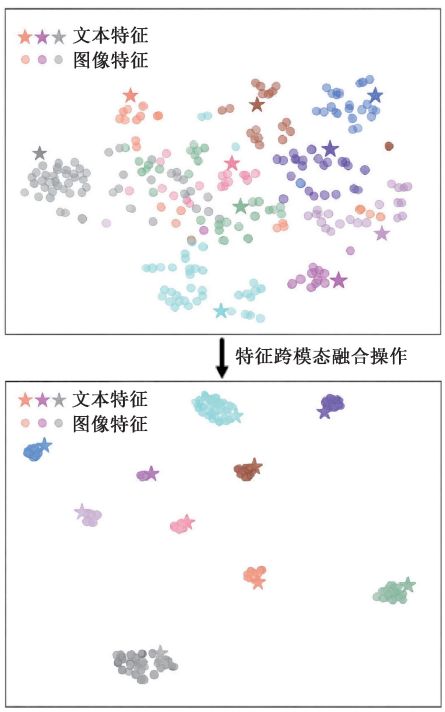


图 9 文本和影像跨模态融合可视化示意图

Fig. 9 Visualization of cross modal information fusion between text and image

Transformer 架构的文本编码器 E_t 进行处理并转换为一个维度固定的文本特征向量,其计算过程如式(6)所示。

$$\boldsymbol{F}_{\text{text}} = E_t(\boldsymbol{T}_{\text{body}}), \boldsymbol{T}_{\text{body}} \in \mathrm{R}^{N\times D} \tag{6}$$

式中:$\boldsymbol{T}_{\text{body}}$ 是每一个身份类别 c_i 的文本描述;N 表示文本描述的长度;D 表示每个词汇的特征维度;$\boldsymbol{F}_{\text{text}} \in \mathrm{R}^d$ 为编码器 E_t 输出的文本特征向量,d 定义同上。因此,分别利用文本编码器与视觉编码器将 $\boldsymbol{T}_{\text{body}}$、$\boldsymbol{I}_{\text{body}}$ 映射到同一潜在的特征空间(即 $\boldsymbol{F}_c \in \mathrm{R}^d$、$\boldsymbol{F}_{\text{body}} \in \mathrm{R}^d$),CLIP 模型成功实现了人员局部视觉特征与对应的身份文本标签的有效匹配。

CLIP 模型通过图像和文本的对比学习来进行大规模的预训练,使具有相似语义的图像和文本在共享特征空间中的距离尽可能接近,而语义不相关的图像与文本则保持较大的距离[38]。依据 CLIP 预训练要求,本研究将输入的人员局部区域图像与其对应的文本描述(如“This is person ID1”)一起进行预训练,并通过最小化匹配图像-文本对的余弦距离实现预训练目的,即图像 $\boldsymbol{I}_{\text{body}}$ 为类别 c_i 的概率计算过程如式(7)所示。

$$P(c_i \mid \boldsymbol{I}_{\text{body}}) = \frac{\exp\left(\dfrac{\text{sim}(E_v(\boldsymbol{I}_{\text{body}}), E_t(\boldsymbol{T}_{\text{body},i}))}{\tau}\right)}{\sum_{j=1}^{K}\exp\left(\dfrac{\text{sim}(E_v(\boldsymbol{I}_{\text{body}}), E_t(\boldsymbol{T}_{\text{body},j}))}{\tau}\right)} \tag{7}$$

$$\text{sim}(\boldsymbol{F}_{\text{body}}, \boldsymbol{F}_{\text{text}}) = \frac{\boldsymbol{F}_{\text{body}} \cdot \boldsymbol{F}_{\text{text}}}{\| \boldsymbol{F}_{\text{body}} \| \ \| \boldsymbol{F}_{\text{text}} \|} \tag{8}$$

式中:$\text{sim}(\cdot)$ 表示方向余弦;$\boldsymbol{F}_{\text{text},i}$ 为每个人员身份类别的文本特征向量;$\boldsymbol{F}_{\text{text},j}$ 是每张图像的文本特征向量;τ 是温度参数;K 是样本数量;$\boldsymbol{F}_{\text{body}}$ 和 $\boldsymbol{F}_{\text{text}}$ 定义同上。

由上述可知,CLIP 引导的视频 PR 过程实质上包含两个阶段:第 1 阶段是利用 ReID-YOLO 实现视频影像中

人员局部视觉特征的鲁棒提取,该阶段聚焦人物面部、轮廓、体型等图像局部特征以尽可能降低全局背景信息干扰、避免无效特征信息对人员身份识别结果的影响,从而能借助 ReID-YOLO 的高召回率克服 CLIP 模型在泛化性要求较高场景中的不稳定性;第 2 阶段则是利用 CLIP 模型在图像与文本间的跨模态信息融合处理能力,整合人员图像局部视觉信息与其身份文本描述语义信息间的构建统一特征表示并将该统一特征表示下的场景泛化优势迁移到人员身份识别任务,从而有效克服 YOLO 系列网络在整合场景局视觉部特征信息进行目标深入解析方面的不足。因此,ReID-YOLO 与 CLIP 的协同有助于整体提高监控视频 PR 精度与场景泛化能力,实现多视角、低分辨监控条件及人员姿态变化、遮挡干扰下的人员身份重识别。

3　实验与分析

由于目前缺乏公开的监控视频 PR 样本数据集,本研究采用 4 台海康威视品牌的监控相机对某会议室场景同时进行多角度摄影以构建网络性能验证所需的样本数据集,其中:4 台监控相机均选用 3.2 mm 焦距镜头并按 30 fps 拍摄频率配置,视频影像分辨率在 480~720 P 之间变化,每 20 帧抽取 1 张作为样本影像数据;每台相机不同时间段采集的视频影像覆盖 2~7 个人员身份,每个人员身份约 400~800 张影像,整个数据共包含约 12 600 张图像(10 000 张为训练集,1 300 张为测试集,1 300 张为验证集),涵盖不同人员身份及其站立、走动、坐下等多种姿态变化,同时视频影像获取过程中人物面部表情、环境灯光及场所设施摆放位置均存在变化以确保场景真实及样本数据的多样性;验证集、测试集尽量包含低分辨率、姿态变化大、存在遮挡的人员目标,如图 10 所示。

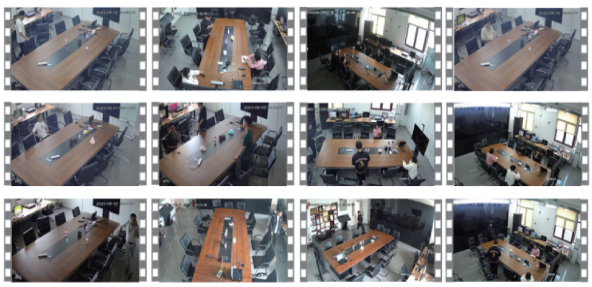


图 10　若干人员难以重新识别的模拟场景

Fig. 10　Diagram of simulated scenarios with several kinds of personnel difficult to be re-identified

本研究监控视频 PR 网络模型训练涉及两个阶段:第 1 阶段 ReID-YOLO 模型训练使用随机梯度下降优化器(stochastic gradient descent, SGD)并设 *batchsize* = 16、*epochs* = 200,学习率为 0.005 且衰减到 0.000 1 时结束训练;第 2 阶段采用 ResNet-50 和 ViT-B/16 预训练模型作为视觉编码器、预训练的 Transformer 作为文本编码器,优化器使用 Adam 且视觉编码器的学习率设为 1×10^{-6}、随机初始化模块的学习率设为 0.001,模型训练 30~80 个 epochs 且每 10 个 epochs 学习率衰减因子设为 1×10^{-8}。上述两阶段框架训练、测试均使用单张显卡 GPU RTX4060,计算环境配置具体见表 1。

表 1　计算环境配置

Table 1　Computing environment configuration

项目	参数
操作系统	Windows 11
GPU	RTX4060 12 G
CPU	Intel i9-13900HX
内存	32 G
Python	3.8
Cuda	11.8

3.1　监控视频 PR 性能评估

依据第 2 章视频 PR 过程对 4 路监控相机拍摄的视频影像数据进行处理,表 2 给出了本研究方法针对不同视角测试集的推理结果统计,并与当前主流的 7 个视频 PR 网络模型对比,包括 AlignGAN (alignment generative adversarial network)[37]、coAtNet[39]、ShuffletNet[40]、ResNet152[41]、DenseNet201[42]、MnasNet[43]、SqueezeNet[44]。其中:同一相机视角表示测试集与训练集为同一相机视角,不同相机视角表示测试集与训练集为不同相机视角;网络性能评价指采用精确率(Precision)、召回率(Recall)和平均准确率(mean average precision, mAP),加粗数值为最优数值。

由表 2 可看出,与传统视频 PR 网络相比,本研究方法整体性能最优,其中:同一相机视角场景中指标 R 与 mAP 数值达到最高水平,分别为 0.93 与 0.91,指标 P 列第 3 位(0.9);不同相机视角场景中指标 R、P 及 mAP 数值均最高。需要指出的是,跨视角(different scene)测试时,包括本研究方法在内的各网络模型性能均出现不同程度的下降,反映出视角差异下行人特征变化对模型泛化能力的巨大挑战。然而,相比于其他网络模型,本研究方法 mAP 下降幅度最低,约为 10%,而 Resnet152 下降约 28%,AlignGAN 下降约 38%,Shufflenet 下降约 40%,Densenet201 下降约 30%,Mnasnet 下降约 44%,Squeezenet 下降约 51%,Coatnet 下降约 39%,表明这些模型在应对不同相机视角时均面临特征不稳定的问题,尤其是 AlignGAN 的 P 和 R 的下降幅度分别达到 65% 和

表 2 推理过程对比结果统计

Table 2 Statistics and comparison of inference results

模型	场景	*P*	*R*	*mAP*
ResNet152	同一相机视角	0.56	0.49	0.53
	不同相机视角	0.15	0.32	0.38
AlignGAN	同一相机视角	0.79	0.85	0.84
	不同相机视角	0.28	0.21	0.52
ShuffleNet	同一相机视角	0.85	0.82	0.64
	不同相机视角	0.51	0.39	0.39
DenseNet201	同一相机视角	0.94	0.91	0.53
	不同相机视角	0.48	0.51	0.37
MnasNet	同一相机视角	0.55	0.37	0.32
	不同相机视角	0.32	0.13	0.18
SqueezeNet	同一相机视角	**0.92**	0.84	0.67
	不同相机视角	0.23	0.29	0.33
CoatNet	同一相机视角	0.82	0.79	0.89
	不同相机视角	0.41	0.38	0.54
本文	同一相机视角	0.90	**0.93**	**0.91**
	不同相机视角	**0.71**	**0.70**	**0.82**

75%,表明该网络性能在跨视角视频 PR 任务中受到严重影响。得益于 CLIP 多模态特征融合优势,本方法在不同视角下仍能保持较高的行人身份重识别性能,视角变换导致的各项指标下降幅度远低于其他模型,表明本研究方法具有更强的泛化能力。

人员身份重识别在相似制服下更具挑战性。针对公开数据集 VS-Clothes[41],表 3 给出了本研究方法在人员相似制服情况下的推理结果统计,并与网络模型 PCB (printed circuit board)[45], HAA (head-shoulder adaptive attention)[46], Pixel[47], CAL (clothes-based adversarial loss)[48],SC-ReID (same-clothes person re-identification)[49]对比,网络性能评价指标采用候选集第 1 次命中率(Rank-1)[48]和 mAP。由表 3 可看出,本研究方法在 VS-Clothes 的 4 个不同相似制服场景下性能最优,相比于次优网络 SC-ReID, mAP 值分别提高了 5.0%、15.2%、10.6% 和 5.8%,原因在于,CLIP 模型通过文本-图像对比学习能充分挖掘两种数据源的互补特性,从而更全面地理解和匹配目标特征,有效提升相似制服人员身份识别精度。此外,在指标 Rank-1 方面,本研究方法 3 个场景均达到了 100%,这表明,本研究方法结合文本数据进行人员图像匹配时几乎在第 1 次检索时就能够准确命中目标,故具有良好的匹配精度且实施效率高。

图 11 进一步给出了本研究方法及其他网络模型针对场景中 4 个人员身份在不同视角、不同分辨率、低光照强度及相似制服条件下的重识别结果,其中:相同标识边界框对应于同一人员身份。从图 11 可看出,除本研究方法外,其他网络模型均存在不同程度的身份确认错误:不同视角条件下,目标特征失真导致传统模型如 PCB、HAA、Pixel 和 CAL,因仅依赖单一模态视觉特征而难以有效建立跨视角的特征对应关系,见图 11(a)~(d);不同分辨率条件下,传统模型如 SC-ReID 则由于缺乏对语义辅助信息的建模能力,在图像细节缺失时因特征表达能力下降而导致目标身份匹配失败,如图 11(e)所示。相反的,本研究借助 CLIP 跨模态融合文本描述与图像特征能力,不仅有效克服视角变化带来的人员身份识别困难,还能利用文本描述信息补充低分辨率图像中丢失细节,从而增强目标识别的鲁棒性,如图 11(f)所示。类似的,本研究方法借助 CLIP 模型通过文本-图像对比学习获得的深层次语义信息可弥补视频图像在低光照强度环境下的光谱信息缺失,或在相似制服条件下更加有效地分辨具有相似外观的目标。总体而言,得益于 CLIP 强大的多模态理解、场景泛化能力及其与 ReID-YOLO 间的互补性,本研究方法能有效提升复杂环境下的视频 PR 精度和鲁棒性。

表 3 VS Clothes 数据集测试结果统计

Table 3 Statistics of test results on VS-clothes dataset

(%)

方法	VS-Clothes (01)		VS-Clothes (02)		VS-Clothes (03)		VS-Clothes (04)	
	Rank-1	*mAP*	*Rank*-1	*mAP*	*Rank*-1	*mAP*	*Rank*-1	*mAP*
PCB	92.0	46.8	90.4	37.5	75.2	26.7	88.3	34.4
HAA	95.2	59.3	94.0	45.7	95.5	58.0	95.3	54.1
Pixel	96.8	65.1	95.2	51.1	94.4	59.6	90.7	54.8
CAL	98.4	52.1	95.8	41.7	91.0	38.7	88.4	43.3
SC-ReID	98.4	72.3	94.0	53.1	96.6	59.9	94.2	55.3
本文	100.0	**77.3**	100.0	**68.3**	90.0	**69.3**	100	**61.1**

3.2 消融实验

1) ReID-YOLO 人员特征检测

为验证低分辨率检测、抗姿(形)态变化检测、抗遮挡检测模块设计在网络 ReID-YOLO 中的作用,本研究将上述模块不同组合引入 YOLOv9 基础模型并利用四路监控相机拍摄的视频影像数据进行消融实验。表 4 分别给出了引入感受野增强模块(RFEM)、计划的感受野增强模块(P-RFEM)、动态卷积(DCN)、空间增强注意力机制(SEA)及归一化高斯距离(NWD)前后的特征检测性能评估指标,其中:评估指标采用特定 IoU 阈值下的 AP (AP_{50}、AP_{75}、$AP_{50:95}$);第 1 行为基础网络人员检测结果统计;第 2~8 行为引入各种模块机制的人员检测结果提升统计。

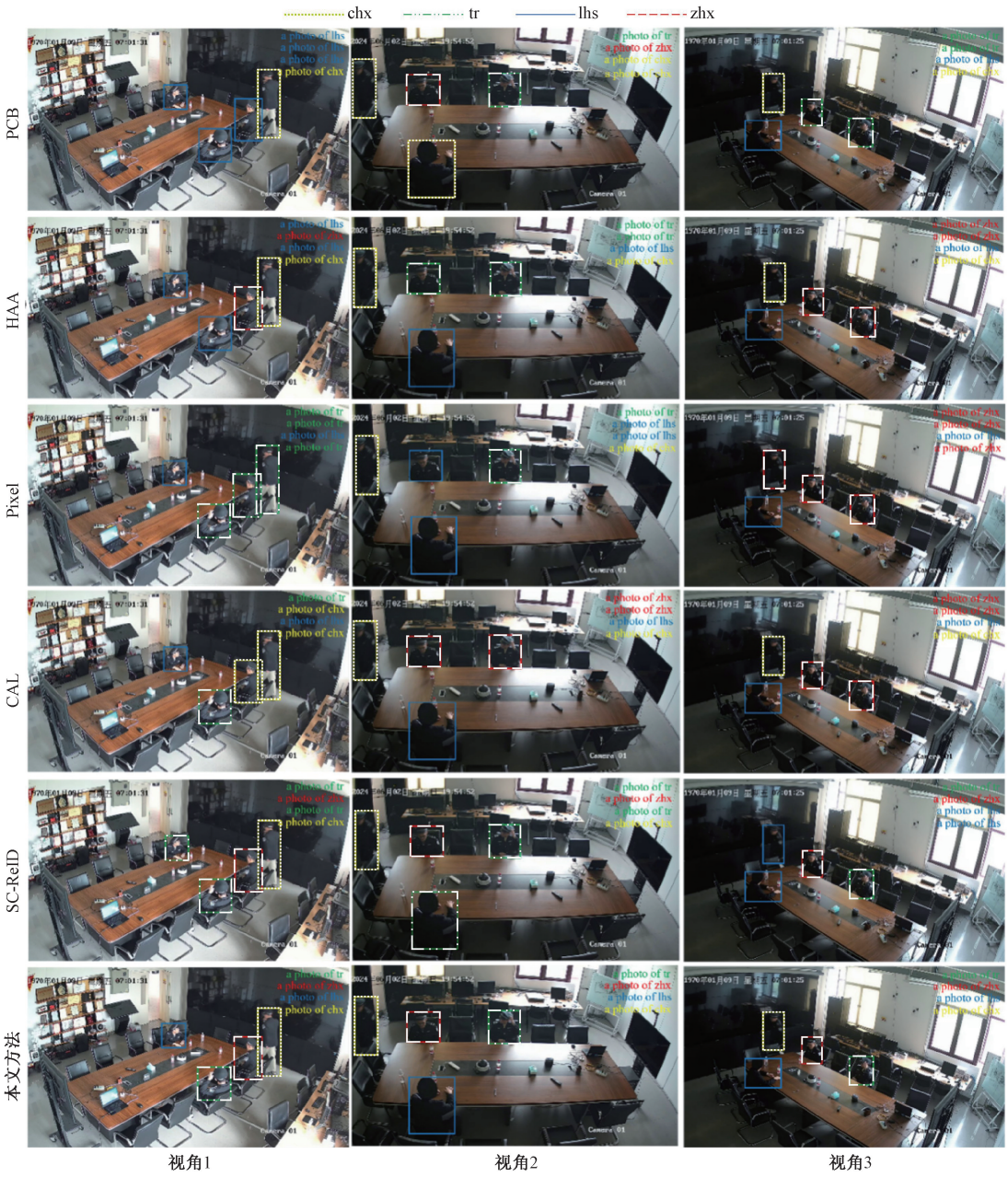


图 11　相似制服场景下 4 类人员身份识别结果对比

Fig. 11　Comparison of identification results for four types of personnel in similar uniforms

表 4　监控视频人员目标检测消融实验结果统计

Table 4　Statistics of surveillance video personnel target detection ablation experiment

模型					评价指标		
+RFEM	+P-RFEM	+DCN	+SEA	+NWD	$AP^{val}_{50:95}$ /%	AP^{val}_{75} /%	AP^{val}_{50} /%
					0. 594	0. 680	0. 751
√					0. 611	0. 678	0. 782
	√				0. 784	0. 808	0. 882
		√			0. 674	0. 743	0. 830
			√		0. 617	0. 723	0. 785
	√	√			0. 791	0. 811	0. 891
	√	√	√		0. 802	0. 843	0. 909
	√	√	√	√	0. 811	0. 852	0. 924

由表 4 可以看出,基础网络单独引入 P-RFEM、DCN、SEA 及 NWD 模块后,其模型精度均有不同程度提升,证明了本研究模块设计的有效性。

其中,引入 P-RFEM 提升精度最高, $AP^{val}_{50:95}$ 从 0.594 提高到 0.784, AP^{val}_{75} 从 0.68 提高到 0.808, AP^{val}_{50} 从 0.751 提高到 0.882,提升幅度分别达 30%、19% 和 17%,原因在于 P-RFEM 通过引入空洞卷积结构与跨尺度残差连接机制,既有效扩大模型的感受野范围,又增强了不同尺度特征间的信息交互与融合能力。具体而言,DCN 相比于标准卷积的优势在于通过可变形采样动态调整卷积核的感受野,从而使模型能更灵活地适应目标的形变特征,故单独引入基础网络后其 $AP^{val}_{50:95}$ 从 0.594 提升至 0.674, AP^{val}_{75} 从 0.680 提升了 0.743, AP^{val}_{50} 从 0.751 提升至 0.830,表明该模块在提升模型对人员外观特征几何变形的适应能力方面发挥了重要作用;SEA 通过注意力机制增强局部关键特征的表达,使模型能够更好地区分被遮挡和未被遮挡的目标区域,故单独引入基础网络后各指标均有小幅提升, AP^{val}_{75} 提升幅度相对较大,原因在于该指标更侧重评估模型在高 IoU 阈值下对目标边界的精确定位能力,而 SEA 在引导网络聚焦于未遮挡区域的判别性特征的同时有效抑制遮挡区域引入的冗余干扰信息,从而有助于在遮挡(高 IoU 阈值条件)场景下实现更为精确的目标轮廓定位。相比于 RFEM 仅通过空洞卷积扩大感受野以获取丰富的全局信息,P-RFEM 还通过使用合理的残差连接增强了梯度反向传播的稳定性并提高特征复用能力,故在处理姿态变化较大的目标时更具优势。相较于单独引入 P-RFEM 或 DCN,P-RFEM 与 DCN 的联合使用使 $AP^{val}_{50:95}$ 平均提升 8.5%, AP^{val}_{75} 提升 4.6%, AP^{val}_{50} 提升 4.1%,其原因在于 P-RFEM 能够增强对复杂姿态的建模能力,而 DCN 则提高了对不规则目标形态的适应性,两者结合可以实现对多样化人员目标的高效检测;P-RFEM、DCN 与 SEA 的联合使用进一步提升模型整体精度,其 $AP^{val}_{50:95}$、AP^{val}_{75}、AP^{val}_{50} 相比于 P-RFEM 与 DCN 的联合使用分别提升 1.4%、3.9% 和 2.0%;P-RFEM、DCN、SEA 与 NWD 的联合使用则使人员特征提取性能达到最佳,证明了本研究模块在复杂场景监控视频 PR 任务下的有效性与鲁棒性。

进一步地,针对“低分辨率小目标”人员,表 5 给出了 NWD 度量下的特征检测性能评估指标,并与其他 4 种度量(IoU, GIoU[50], CIoU[51], DIoU[52])进行对比。评估指标中前 3 列指标采用特定 IoU 阈值下的 AP(AP_{50}、AP_{75}、$AP_{50:95}$);因需对低分辨率小目标检测结果进行对比,后 3 列指标采用 COCO 数据集[52]针对不同目标尺寸(Small、Medium、Large)的 AP(AP_S、AP_M、AP_L)。由该表可看出,传统 IoU 显著低于其他度量方法,各评估指标均处于排名倒数第 1 或倒数第 2。具体而言,GIoU 作为 IoU 的扩展,其检测性能评估指标除 AP^{val}_{75} 外相对于 IoU 均有一定提升,但提升幅度有限,仅指标 AP^{val}_{50} 处于排名第 2 位置,原因在于 GIoU 在 IoU 的基础上额外引入了目标检测框与其最小外边界矩形之间的差异度量,虽然一定程度上缓解了边界框不重叠时的梯度消失问题,但在处理目标尺度变化较大或目标密集分布的情况下,其优化效果依然有限,导致整体性能提升幅度较小;CIoU 整体性能略优于 IoU,指标 AP^{val}_M 和 AP^{val}_L 处于倒数排名第 1 位置,但指标 AP^{val}_S 提高显著(达到 0.601),处于排名第 2 位置,表明 CIoU 在处理目标位置关系和距离差异方面的优化能有效增强模型的精度和鲁棒性;DIoU 各项性能指标均优于 IoU,部分指标性能优于 CIoU 和 GIoU,原因在于 DIoU 在 IoU 的基础上进一步考虑了目标中心点之间的欧氏距离,使得模型在优化目标框时能够同时关注框的重叠程度与中心点相对关系,从而提升目标定位的准确性;NWD 在各个评估指标上均展现出最佳的性能,尤其在 $AP^{val}_{50:95}$、AP^{val}_{75}、AP^{val}_{50}、AP^{val}_S 等关键指标上高于其他度量方法,表明 NWD 能够在复杂的多尺度场景以及不同分辨率下提供更精确的目标识别和定位。

表 5 不同损失度量下的低分辨率人员目标检测结果统计

Table 5 Statistics of personnel target detection under low resolution and different loss measurements (%)

评价指标	$AP^{val}_{50:95}$	AP^{val}_{75}	AP^{val}_{50}	AP^{val}_S	AP^{val}_M	AP^{val}_L
IoU	0.718	0.734	0.776	0.565	0.711	0.825
GIoU	0.721	0.721	0.811	0.578	0.724	0.834
CIoU	0.722	0.745	0.794	0.601	0.707	0.810
DIoU	0.743	0.765	0.791	0.592	0.736	0.842
NWD	0.758	0.785	0.833	0.643	0.748	0.847

针对“姿态变化大”、“形状不规则”、“局部遮挡”、“低分辨率小目标”4 类识别困难的典型人员,图 12 给出了网络 ReID-YOLO 推理结果并与基础网络进行对比。由图 12 可看出,YOLOv9 目标检测易出现漏检或误检(如检测错位或检测框不完整),原因在于其感受野受限,在处理“低分辨率小目标”或“姿态变化大”的行人目标时无法有效捕获完整特征,导致检测失败或框选区域不完整;此外,YOLOv9 对“局部遮挡”目标及“形状不规则”目标检测的鲁棒性较弱,原因在于,信息缺失情况下标准卷积结构难以准确聚焦未遮挡区域,导致检测框漂移或重叠度不足。相比之下,ReID-YOLO 通过 P-RFEM 与 DCN 提高了感受野覆盖范围,通过 SEA 优化特征提取并结合 NWD 优化边界框回归精度,从而有效缓解漏检、

图 12　4 种困难人员网络目标检测结果对比

Fig. 12　Comparison of network target detection results for 4 types of difficult personnel

误检问题,侧面证明了本研究方法视频 PR 任务中应对复杂语义环境的能力。

2) CLIP 人员身份预测

由第 2 章可知,本研究视频 PR 任务实质上包含两个阶段:第 1 阶段是利用 ReID-YOLO 检测视频影像中人员目标视觉特征;第 2 阶段则是利用 CLIP 模型对上一阶段检测到的人员目标进行身份预测(重识别)。为验证 CLIP 模型对视频 PR 任务的有效性,本研究结合 3.1 中不同视角场景划分及 3.2 节第 1 部分 4 类识别困难的典型人员,设计了如表 6 所示 3 种组合下的消融实验,性能评估指标采用精确率(P)。由表 6 可发现,直接使用 CLIP 模型时人员身份预测精度最低,尤其是在不同相机视角场景;与 CLIP 模型相结合进行人员身份预测时,第 1 阶段采用 ReID-YOLO 优于采用 YOLOv9,究其原因在于两个方面:1) CLIP 模型侧重整个场景的全局理解,并未特别关注人员局部图像特征,故依赖于第 1 阶段检测到的视频局部特征对人员身份进行跨模态验证,第 1 阶段的误检、漏检行为将直接导致第 2 阶段身份识别任务失效;2) ReID-YOLO 相比于 YOLOv9,在不同视角、不同分辨率、低光照强度及相似制服条件下检测困难人员特征如“姿态变化大”、“形状不规则”、“局部遮挡”、“低分辨率小目标”方面性能更优,能大幅减少误检、漏检情况,为第 2 阶段 CLIP 多模态身份识别提供了更完整、更精准的人员身份视觉特征信息。总体上,ReID-YOLO 与

表 6　CLIP 人员身份预测消融实验

Table 6　CLIP personnel identity prediction ablation experiments

Clip	+YOLOv9	+ReID-YOLO	低分辨率小目标		姿态变化大		形状不规则		局部遮挡	
			同一相机视角	不同相机视角	同一相机视角	不同相机视角	同一相机视角	不同相机视角	同一相机视角	不同相机视角
√			0.58	0.21	0.47	0.19	0.54	0.22	0.46	0.19
√	√		0.78	0.71	0.81	0.73	0.79	0.69	0.73	0.64
√		√	0.91	0.87	0.92	0.84	0.87	0.83	0.91	0.84

CLIP 两者协同形成优势互补,整体提高复杂场景、语义环境下的监控视频 PR 精度与场景泛化能力,使得在同一视角下 4 类困难人员身份的重识别平均精度达 90%,不同视角下 4 类人员身份重识别平均精度约 85%,均为最高水平,再一次证明了本研究方法的有效性。

图 13 为 CLIP 分别与 ReID-YOLO、YOLOv9 相结合的视频 PR 结果对比示意。

图 13 某时刻视频 4 类困难人员身份重识别结果示意

Fig. 13 Diagram of network video PR results for 4 types of difficult personnel at a certain time

从图 13 可看出,YOLOv9 在特征传递结构与特征上下文理解方面的劣势导致在应对“姿态变化较大”或“局部被遮挡”人员目标时,检测框存在不完整、偏移或过度覆盖背景区域情况,从而为第 2 阶段 CLIP 人员身份识别任务引入大量背景冗余信息的干扰;另一方面,YOLOv9 在多尺度感知能力方面的劣势则导致在应对“形状不规则”或“低分辨率小目标”人员目标时出现漏检现象,甚至无法为第 2 阶段 CLIP 人员身份识别任务提供输入特征。相比之下,ReID-YOLO 在复杂监控场景、条件下的人员目标视觉特征精准提取优势,不仅有效缓解人员目标特征的误检与漏检问题,其与 CLIP 的结合还增强了人员身份重识别多模态系统的整体鲁棒性,从而提升视频 PR 准确性与场景泛化能力。

3.3 鲁棒性实验

为了进一步验证本研究方法在人员身份重识别上的鲁棒性,表 7 给出了本研究方法在人员相似制服情况下的预测结果与预测效率,并与网络模型 YOLOv9,YOLOv11[13]以及 YOLOv12[14]进行对比。需要指出的是,本实验的测试集与训练集为同一相机视角,且 70% 数据作为训练集、10% 数据作为验证集以及 20% 数据作为测试集。网络性能评价指标采用时间(T)、精确率(P)、召回率(R)和 AP(AP_{50}、$AP_{50:95}$),其中加粗数值为最优数值。

表 7 鲁棒性实验对比结果统计

Table 7 Statistics of robustness experiment comparison results

模型	T/s	P	R	AP_{50}	$AP_{50:95}$
YOLOv9	14.48	0.70	0.72	0.76	0.63
YOLOv11	**8.04**	0.56	0.68	0.69	0.57
YOLOv12	60.79	**0.94**	0.84	**0.95**	0.79
本文	16.13	0.92	**0.89**	**0.95**	**0.81**

由表 7 可看出,与最新的 3 个 YOLO 系列网络相比,本研究方法 AP 性能最优,其中:指标 R 与 AP 数值均为最高水平,分别为 0.89、0.95 以及 0.81,指标 T 列第 3 位(16.13 s)。精度指标位列第 2 的 YOLOv12 相比本研究方法在每个批次的训练中需要花费将近 4 倍时间。这印证了 YOLOv12 设计的混合 ViT-CSP 骨干网络不仅可以充分提取局部特征细节信息,还可以有效对全局语义进行建模,进而取得与本研究方法接近的目标识别精度。然而,其复杂的网络结构与视觉 Transformer 模块使得网络的训练成本与计算复杂度成倍上升。另一方面,YOLOv11 除了速度列第 1(每个批次训练耗时仅 8.04 s),在各项精度指标中均为末尾。这也充分说明,当裁剪 CSPNet 的残差块数量并替换大卷积为深度可分离卷积时,网络的参数量将被极限压缩。进一步观察 YOLOv9 的各项参数指标可得,该网络设计的通用高效层聚合模块在不增加计算量的前提下,可以有效增强多尺度特征的关联性。相比 YOLOv8,该网络具有更好的

全局建模能力以及复杂场景局部细节信息的捕捉能力。本研究设计的网络在 YOLOv9 基础上引入感受野增强模块与可变形卷积计算、空间增强注意力模块及基于归一化高斯距离的损失度量模块不仅可以有效提升网络模型的特征检测精度和鲁棒性,还能适应低分辨率、遮挡以及不同姿态形态等复杂环境下人员身份重识别任务。

4　结　　论

视频 PR 对于提升监控系统智能化水平并拓展其应用场景具有重要意义。然而,实际监控场景环境复杂,包括 YOLO 系列网络在内的单模态特征匹配方法在整合场景信息进行人物目标深入解析方面存在不足;与此同时,多模态模型 CLIP 通过图像-文本间的对比学习展现出强大的多模态理解、场景泛化能力,为解决当前视频 PR 问题提供了新的思路。从这一思路出发,本研究结合 YOLOv9 和 CLIP 提出一种多模态信息融合的监控视频 PR 新方法,将 CLIP 跨模态信息融合优势迁移到视频 PR 任务,并引入感受野增强模块与可变形卷积计算、空间增强注意力机制、基于归一化高斯距离的损失度量构建网络 ReID-YOLO 以增强监控视频目标人员特征检测精度、鲁棒性,两者协同既能借助 ReID-YOLO 人员视觉特征区分能力缓解 CLIP 全局场景过度依赖之不足,又能借助 CLIP 模型场景泛化能力克服 YOLO 系列网络在整合场景信息深入解析目标方面的不足,从而整体提高监控视频 PR 精度与场景泛化能力。针对相似制服公开数据集及涵盖“低分辨率小目标”、“姿态变化大”、“形状不规则”、“局部遮挡”等多种识别困难人员身份的室内场景采集视频数据集测试结果证明了所提方法的有效性,优于 YOLO 系列网络及其他 7 个主流的视频 PR 网络,具有良好应用前景。后续将结合更多场景数据对网络进行性能测试并探索在低功耗嵌入式系统上的应用部署与推理实现,为研制高性价比的视频检测设备终端奠定基础。

参考文献

[1] DOU ZH P, WANG ZH D, LI Y L. Identity-seeking self-supervised representation learning for generalizable person re-identification[C]. 2023 IEEE/CVF International Conference on Computer Vision, 2023: 15847-15858.

[2] 侯兴民, 李冉, 张玉洁. 基于脚步诱发结构振动的人员特征身份识别研究[J]. 振动与冲击, 2022, 41(23): 241-248, 292.
HOU X M, LI R, ZHANG Y J. Person identity recognition based on structural vibration induced by footsteps[J]. Journal of Vibration and Shock, 2022, 41(23): 241-248, 292.

[3] 邱杰凡, 周克众, 朱东福, 等. 基于近场特征融合网络的无线身份认证方法[J/OL]. 工程科学与技术, 1-14[2025-09-16].
QIU J F, ZHOU K ZH, ZHU D F, et al. Wireless identity authentication method based on near-field feature fusion network[J/OL]. Engineering Science and Technology, 1-14[2025-09-16].

[4] 袁田, 辛义忠. 结合递归图与 LeNet 网络的足底压力身份识别方法[J]. 仪器仪表学报, 2025, 46(6): 338-347.
YUAN T, XIN Y ZH. Plantar pressure-based identity recognition method combining recurrent plot and LeNet network[J]. Chinese Journal of Scientific Instrument, 2025, 46(6): 338-347.

[5] 赵凯旋, 王锦锦, 高颂, 等. 基于 PointNet++和改进 ConvNeXt 模型的奶牛个体身份识别方法[J]. 农业机械学报, 2025, 56(7): 567-574, 595.
ZHAO K X, WANG J J, GAO S, et al. Cow individual identity recognition based on PointNet++ and improved ConvNeXt model[J]. Transactions of the Chinese Society for Agricultural Machinery, 2025, 56(7): 567-574, 595.

[6] XIANG S CH, QIAN D H, GAO J SH, et al. Rethinking person re-identification via semantic-based domain generalization[J]. ACM Transactions on Multimedia Computing, Communications and Applications, 2023, 20(3): 1-17.

[7] LENG Q. Co-metric learning for person re-identification[J]. Advances in Multimedia, 2018: 1-10.

[8] HAN K, SI CH Y, HUANG Y, et al. Generalizable person re-identification via self-supervised batch norm test-time adaption[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(1): 817-825.

[9] YE M, SHEN J B, LIN G J, et al. Deep learning for person re-identification: A survey and outlook[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(6): 2872-2893.

[10] 李学钊, 王伟, 薛冰. 基于梯度算子和注意力的多模态融合目标检测[J]. 仪器仪表学报, 2024, 45(11): 224-232.
LI X ZH, WANG W, XUE B. Multi-modal fusion object detection based on gradient operator and attention[J].

Chinese Journal of Scientific Instrument, 2024, 45(11): 224-232.

[11] 钱亚萍, 王凤随, 熊磊. 基于局部细化多分支与全局特征共享的无监督行人重识别方法[J]. 电子测量与仪器学报, 2023, 37(1): 106-115.
QIAN Y P, WANG F S, XIONG L. Unsupervised person re-identification based on locally refined multi-branch and global feature sharing[J]. Journal of Electronic Measurement and Instrumentation, 2023, 37(1): 106-115.

[12] 程思雨, 陈莹. 基于 ViT 的细粒度特征增强无监督行人重识别方法[J]. 电子测量与仪器学报, 2024, 38(9): 24-35.
CHENG S Y, CHEN Y. Unsupervised person re-identification method with fine-grained feature enhancement based on ViT[J]. Journal of Electronic Measurement and Instrumentation, 2024, 38(9): 24-35.

[13] 胡玉玲,王鑫依,张一,等. 一种基于注意力的无监督行人重识别方法[J]. 电子测量技术, 2025, 48(10): 161-168.
HU Y L, WANG X Y, ZHANG Y, et al. An attention-based unsupervised person re-identification method[J]. Electronic Measurement Technology, 2025, 48(10): 161-168.

[14] 邓子文, 段勇. 基于深度聚类学习的无监督行人重识别[J]. 电子测量与仪器学报, 2025, 39(3): 208-216.
DENG Z W, DUAN Y. Deep clustering learning-based unsupervised person re-identification[J]. Journal of Electronic Measurement and Instrumentation, 2025, 39(3): 208-216.

[15] MCLAUGHLIN N, DEL RINCON J M, MILLER P. Recurrent convolutional network for video-based person re-identification[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016: 1325-1334.

[16] RISTANI E, SOLERA F, ZOU R, et al. Performance measures and a data set for multi-target, multi-camera tracking[C]. Proceedings of the European Conference on Computer Vision, 2016: 17-35.

[17] WU AN C, ZHENG W SH, YU H X, et al. RGB-infrared cross-modality person re-identification[C]. 2017 IEEE International Conference on Computer Vision, 2017: 5390-5399.

[18] ZHENG L, ZHANG H H, SUN SH Y, et al. Person re-identification in the wild[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017: 1367-1376.

[19] WU Y, LIN Y T, DONG X Y, et al. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018: 5177-5186.

[20] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]. Proceedings of the Neural Information Processing Systems, 2017: 6000-6010.

[21] KHANAM R, HUSSAIN M. Yolov11: An overview of the key architectural enhancements[J]. ArXiv preprint arXiv: 2410.17725, 2024.

[22] TIAN Y J, YE Q X, DOERMANN D. Yolov12: Attention-centric real-time object detectors[J]. ArXiv preprint arXiv: 2502.12524, 2025.

[23] 陈方彬, 赵仲勇, 王建, 等. 基于 YOLO-MCSL 的轻量化智能电能表热缺陷目标检测方法[J]. 仪器仪表学报, 2025, 46(8): 108-119.
CHEN F B, ZHAO ZH Y, WANG J, et al. A lightweight thermal defect detection method for smart electricity meters based on YOLO-MCSL[J]. Chinese Journal of Scientific Instrument, 2025, 46(8): 108-119.

[24] DING SH Y, LIN L, WANG G R, et al. Deep feature learning with relative distance comparison for person re-identification[J]. Pattern Recognition, 2015, 48(10): 2993-3003.

[25] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]. 2017 IEEE International Conference on Computer Vision, 2017: 2999-3007.

[26] 熊鹏文, 胡慕烨, 黄雨轩, 等. 面向小样本纹理分类的多模态证据融合框架[J]. 仪器仪表学报, 2025, 46(6): 154-165.
XIONG P W, HU M Y, HUANG Y X, et al. Small-sample multi-modal evidence fusion framework for texture classification[J]. Chinese Journal of Scientific Instrument, 2025, 46(6): 154-165.

[27] BOSQUET B, CORES D, SEIDENARI L, et al. A full data augmentation pipeline for small object detection based on generative adversarial networks[J]. Pattern Recognition, 2023, 133: 108998.

[28] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection[C]. 29th IEEE Conference on Computer Vision and Pattern Recognition, 2016: 779-788.

[29] CAI ZH W, VASCONCELOS N. Cascade R-CNN: Delving into high quality object detection[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018: 6154-6162.

[30] LAW H, DENG J. CornerNet: Detecting objects as paired keypoints[C]. Proceedings of the European Conference on Computer Vision, 2018: 734-750.

[31] HE K M, ZHANG X Y, REN SH Q, et al. Deep residual learning for image recognition[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.

[32] TAN M X, PANG R M, LE Q V. EfficientDet: Scalable and efficient object detection[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 10781-10790.

[33] RADFORD A, KIM J W, HALLACY J, et al. Learning transferable visual models from natural language supervision[C]. Proceedings of the International Conference on Machine Learning, 2021: 8748-8763.

[34] CARON M, MISRA I, MAIRAL J, et al. Unsupervised learning of visual features by contrasting cluster assignments[C]. Advances in Neural Information Processing Systems, 2020: 9912-9924.

[35] FROME A, CORRADO G S, SHLENS J, et al. Devise: A deep visual-semantic embedding model[C]. Advances in Neural Information Processing Systems, 2013: 2121-2129.

[36] BEDAGKAR-GALA A, SHAH S K. A survey of approaches and trends in person re-identification[J]. Image and Vision Computing, 2014, 32(4): 270-286.

[37] WANG G AN, ZHANG T ZH, CHENG J, et al. RGB-infrared cross-modality person re-identification via joint pixel and feature alignment[C]. 2019 IEEE/CVF International Conference on Computer Vision, 2019: 3623-3632.

[38] ZHU K, GUO H Y, YAN T Y, et al. PASS: Part-aware self-supervised pre-training for person re-identification[C]. European Conference on Computer Vision, 2022: 198-214.

[39] XING J, HAO M ZH, LIU Y T, et al. A CoAtNet-enhanced graphical inversion method with strong generalization for multispectral radiation thermometry[J]. Infrared Physics & Technology, 2025, 150: 105959.

[40] ZHAO H M, GAO Y SH, DENG W. Defect detection using ShuffleNet-CA-SSD lightweight network for turbine blades in IoT[J]. IEEE Internet of Things Journal, 2024, 11(20): 32804-32812.

[41] ALI S, AGRAWAL J. Brain tumor segmentation and detection using deep learning method based on ResNet152[J]. Procedia Computer Science, 2025, 228: 160-169.

[42] MOHANDASS G, KRISHNAN G H, SELVARAJ D, et al. Lung cancer classification using optimized attention-based convolutional neural network with DenseNet-201 transfer learning model on CT image[J]. Biomedical Signal Processing and Control, 2024, 95: 106330.

[43] WANG L, ZHAO Q X, ZAKHAROV M A, et al. Optimizing fault prediction in software based on MnasNet/LSTM optimized by an improved lotus flower algorithm[J]. Egyptian Informatics Journal, 2025, 29: 100623.

[44] ATABEY S, AKAGUNDUZ E. Binary SqueezeNet: Enhancing parameter efficiency in binary neural networks[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025: 4052-4061.

[45] SUN Y F, ZHENG L, YANG Y, et al. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)[C]. Proceedings of the European Conference on Computer Vision, 2018: 480-496.

[46] XU B Q, HE L X, LIAO X Y, et al. Black Re-ID: A head-shoulder descriptor for the challenging problem of person re-identification[C]. Proceedings of the 28th ACM International Conference on Multimedia, 2020: 673-681.

[47] SHU X J, LI G, WANG X, et al. Semantic-guided pixel sampling for cloth-changing person re-identification[J]. IEEE Signal Processing Letters, 2021, 28: 1365-1369.

[48] GU X Q, CHANG H, MA B P, et al. Clothes-changing person re-identification with RGB modality only[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 1060-1069.

[49] WU ZH Y, HU Z R, DING J W. Same-clothes person

re-identification with dual-stream network[J]. Multimedia Systems, 2024, 30(2): 70.

[50] ZHANG X Y, ZHOU X Y, LIN M X, et al. ShuffleNet: An extremely efficient convolutional neural network for mobile devices [C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018: 6848-6856.

[51] SUN Y F, ZHANG L, YANG Y L, et al. Learning part-based convolutional features for person re-identification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(3): 902-917.

[52] DAI Z H, LIU H Y, LE Q V, et al. CoAtNet: Marrying convolution and attention for all data sizes[C]. Advances in Neural Information Processing Systems, 2021: 3965-3977.

作者简介

吴军,1997 年于原武汉测绘科技大学城建学院获得学士学位,2000 年于原武汉测绘科技大学城建学院获得硕士学位,2003 年于武汉大学获得博士学位,现为桂林电子科技大学教授,主要研究方向为数字摄影测量、计算机视觉、图像处理和人工智能。

E-mail:wujun93161@163.com

Wu Jun received his B. Sc. and M. Sc. degrees both from School of Urban Construction, former Wuhan University of Surveying and Mapping Technology in 1997 and 2000, and his Ph. D. degree from Wuhan University in 2003. He is currently a professor at Guilin University of Electronic Technology. His main research interests include digital photogrammetry, computer vision, image process and artificial intelligence.

陈慧,2024 年于东华理工大学获得学士学位,现为桂林电子科技大学硕士研究生,主要研究方向为深度学习、计算机视觉。

E-mail:18870331228@163.com

Chen Hui received her B. Sc. degree from East China University of Technology in 2024. She is currently a master′s student at Guilin University of Electronic Technology. Her main research interests include deep learning and computer vision.

徐刚,2008 年于淮北师范大学获得学士学位,2011 年于桂林电子科技大学获硕士学位,现为中国科学院宁波材料技术与工程研究所副研究员,主要从事机器视觉、人工智能和视频信息处理。

E-mail:xugang@nimte.ac.cn

Xu Gang received his B. Sc. degree from Huaibei Normal University in 2008 and his M. Sc. degree from Guilin University of Electronic Science and Technology in 2011. He is currently an associate professor at Ningbo Institute of Materials Technology and Engineering CAS. His main research interests include machine vision, artificial intelligence and viedo information processing.

赵雪梅,2012 年于辽宁工程技术大学获得学士学位,2017 年于辽宁工程技术大学获工学博士学位,现为桂林电子科技大学副教授,主要研究方向为信息几何、人工智能、深度学习遥感数据处理。

E-mail:zhaoxm@guet.edu.cn

Zhao Xuemei received her B. Sc. and Ph. D. degrees both from Liaoning Technical University in 2012 and 2017. She is currently an associate professor at Guilin University of Electronic Technology. Her main research interests include information geometry, artificial intelligence, and deep learning-based remote sensing data processing.

陈睿星(通信作者),2024 年于桂林电子科技大学获博士学位,现为中国科学院宁波材料技术与工程所助理研究员,主要研究方向为人工智能、深度学习、多模态特征提取。

E-mail:chenruixing@nimte.ac.cn

Chen Ruixing (Corresponding author) received his Ph. D. degree from Guilin University of Electronic Technology in 2024. He is currently an assistant professor at the Ningbo Institute of Materials Technology and Engineering, Chinese Academy of Sciences. His main research interests include artificial intelligence, deep learning, and multimodal feature extraction.