

联合跟踪的特征聚合:遮挡场景下的视频目标检测任务

王家辉, 王洪玉, 郝应光

引用本文

王家辉, 王洪玉, 郝应光. [联合跟踪的特征聚合:遮挡场景下的视频目标检测任务](#)[J]. 计算机科学, 2026, 53(8): 165-173.

WANG Jiahui, WANG Hongyu, HAO Yingguang. [Feature Aggregation with Joint Tracking:Video Object Detection in Occlusion Scenarios](#) [J]. Computer Science, 2026, 53(8): 165-173.

相似文章推荐 (请使用火狐或 IE 浏览器查看文章)

Similar articles recommended (Please use Firefox or IE to view the article)

[改进YOLOv11n的无人机航拍图像小目标检测模型](#)

Improved YOLOv11n Model for Small Target Detection in UAV Aerial Images

计算机科学, 2026, 53(8): 201-208. <https://doi.org/10.11896/jsjcx.250800100>

[基于DETR改进的CAD图元识别方法](#)

Improved DETR-based Method for CAD Graphic Element Recognition

计算机科学, 2026, 53(8): 139-147. <https://doi.org/10.11896/jsjcx.251100082>

[基于改进YOLOv5的库室装备检测](#)

Armory Equipment Detection Based on Improved YOLOv5

计算机科学, 2026, 53(6A): 250800049-6. <https://doi.org/10.11896/jsjcx.250800049>

[基于RGB-IR多模态融合的番茄小目标检测](#)

RGB-IR Multi-modal Fusion-based Tomato Small Object Detection

计算机科学, 2026, 53(6A): 250700173-8. <https://doi.org/10.11896/jsjcx.250700173>

[基于分阶段训练策略与多尺度特征融合的目标检测方法](#)

Object Detection Method Based on Phased Training Strategy and Multi-scale Feature Fusion

计算机科学, 2026, 53(6A): 250700088-7. <https://doi.org/10.11896/jsjcx.250700088>

联合跟踪的特征聚合:遮挡场景下的视频目标检测任务

王家辉 王洪玉 郝应光

大连理工大学信息与通信工程学院 辽宁 大连 116024

(wjhdut1008@163.com)

摘要 视频目标检测(VID)的主要任务是利用视频帧之间的时序信息来辅助检测。与图像目标检测不同,视频目标检测能够利用丰富的时序上下文信息,通过参考其他帧的特征来增强当前帧的检测效果。因此,现有方法普遍致力于聚合多帧间的特征,以构建时序增强的特征表示,从而提升检测的准确性。然而,在视频序列中,目标常因遮挡、运动模糊等因素出现外观退化,导致单帧检测器难以生成高质量的候选区域,使得后续的跨帧特征聚合过程失效。为此,提出一种联合跟踪与检测的视频目标检测方法,旨在改善候选区域缺失导致的聚合失败问题。具体而言,该方法首先引入轻量级跟踪器,并设计特征选择模块,以协同利用检测信息与跟踪器提供的运动先验;其次,采用加权框融合(WBF)算法对检测框与跟踪预测框进行合并,以生成在遮挡等挑战场景下更为可靠的候选区域;最后,提取这些高质量候选区域及其对应轨迹上的特征,送入后续模块进行时序聚合。在ImageNet VID数据集上的实验结果表明,所提方法有效增强了模型在遮挡场景下的鲁棒性,展现出更高的检测精度。

关键词: 视频目标检测;目标检测;目标跟踪;特征聚合;加权框融合

中图分类号 TP391

Feature Aggregation with Joint Tracking: Video Object Detection in Occlusion Scenarios

WANG Jiahui, WANG Hongyu and HAO Yingguang

School of Information and Communication, Dalian University of Technology, Dalian, Liaoning 116024, China

Abstract Video object detection(VID) aims to exploit temporal information between video frames to assist object detection. Unlike still-image detection, VID can leverage rich temporal context by referencing features from other frames to enhance the detection performance. Therefore, existing methods generally focus on aggregating features across multiple frames to construct temporally enhanced feature representations and improve detection accuracy. However, in video sequences, objects often undergo appearance degradation due to factors such as occlusion and motion blur, making it difficult for single-frame detectors to generate high-quality candidate regions. This, in turn, can cause the subsequent cross-frame feature aggregation to fail. To address this issue, this paper proposes a tracking-assisted video object detection method designed to mitigate aggregation failures caused by missing candidate regions. Specifically, it introduces a lightweight tracker and designs a feature selection module to jointly utilize detection outputs and motion priors provided by the tracker. It then employs the WBF(Weighted Box Fusion) algorithm to merge detection boxes with tracking predictions, producing more reliable candidate regions under challenging conditions such as occlusion. Finally, it extracts these high-quality candidate regions and their corresponding trajectory features for subsequent temporal aggregation. Experiments on the ImageNet VID dataset demonstrate that the proposed method effectively enhances robustness in occluded scenarios and achieves higher detection accuracy compared with previous VID approaches.

Keywords Video object detection, Object detection, Object tracking, Feature aggregation, Weighted box fusion

1 引言

近年来,目标检测任务作为计算机视觉领域的基础任务^[1],发展迅速。其目标是同时定位和分类图像中的目标。针对视频数据进行目标检测,最直接的方法是将连续的视频帧逐帧输入目标检测器进行独立检测,获得检测结果。但这种做法忽略了视频帧之间连续的时序信息和运动信息^[2],在

面对视频中的图像质量下降等问题时,检测性能往往难以保证。视频帧中常见的图像退化现象包括运动模糊、形变和遮挡等,这些问题显著增加了检测难度。在此情况下,仅依赖当前帧的图像信息难以准确检测目标。考虑到目标在连续视频帧中具有时序连续性,利用历史帧信息辅助当前帧预测,成为提升检测精度的重要途径。因此,视频目标检测任务应运而生。该任务的关键即为,如何有效利用视频帧之间的时序信

到稿日期:2025-07-22 返修日期:2025-11-26

基金项目:大连市科技创新基金(2022JJ11CG002)

This work was supported by the Dalian Science and Technology Innovation Fund(2022JJ11CG002).

通信作者:王洪玉(whyu@dlut.edu.cn)

息来提升性能^[3]。

视频帧虽可能受到模糊、遮挡、形变等干扰,但其蕴含丰富的时序关联信息。如何从其他帧中为目标帧获取支持信息,对提高视频目标检测的准确性具有关键意义。目前,主流的视频目标检测方法多集中于,利用帧间时序信息进行跨帧特征聚合,从而提升检测精度。典型的视频目标检测方法通过连接边界框来连接静止目标检测器的输出形成 tubelet^[4],并在同一 tubelet 中细化结果。后续出现的特征聚合方法,首先通过初步检测取得目标帧的候选区域,对于目标帧中的一个候选区域,计算候选区域与其他帧的特征的注意力权重,根据计算出的权重对参考特征进行加权聚合,利用得到的新特征得到检测结果。

然而,此类方法高度依赖初步检测得到的候选区域,其聚合准确性建立在候选区域完整的前提上。在严重遮挡等情况下,仅依赖初步检测往往难以获取全部潜在的目标候选区域。此外,这类方法在特征聚合时需遍历参考帧的全部特征,可能引入无关信息权重,导致得到次优检测结果。

针对上述问题,本文设计了一种联合目标跟踪的视频目标检测框架,主要工作如下:

1)针对在严重遮挡时无法获取可靠候选区域导致特征聚合失败的问题,在传统视频目标检测算法的基础上引入跟踪器,提出了一个能联合使用跟踪预测信息和目标检测信息的特征选择模块,并使用加权框融合算法融合跟踪预测信息和目标检测信息;

2)提出了一种利用目标轨迹特征的聚合方法,进一步抑制在传统遍历方法中非目标特征的干扰,提升聚合特征的质量;

3)通过在 Imagenet VID 数据集上进行的实验验证了所提方法的有效性,相比以往的视频目标检测方法,本文方法表现出更高的检测精度,增强了目标检测模型在遮挡场景下的鲁棒性。

2 相关工作

视频目标检测网络通常构建于静态图像目标检测器的基础之上,并结合时序建模、目标跟踪机制及特征增强等方法,以提升其在复杂动态场景中的鲁棒性与检测精度。本章将对与本文研究相关的视频目标检测、目标跟踪等部分展开介绍。

2.1 基于注意力的视频目标检测

视频目标检测 (Video Object Detection, VOD) 任务不仅要求在每一帧中准确检测目标,还需充分利用视频中连续帧之间的时序信息,以提升检测精度。在早期研究中,由于两阶段目标检测器在第一阶段能够有效区分前景与背景,从而在前期便过滤掉背景区域,这为视频目标检测利用时序信息进行特征聚合提供了有利条件。因此,大量早期 VOD 方法均采用两阶段检测器作为基础框架。注意力机制^[5] (Attention Mechanism) 一经提出,便迅速被引入计算机视觉任务。在图像目标检测中,典型的注意力机制包括 SE^[6] 和 CBAM^[7]。近年,Transformer 架构(如 DETR^[8]) 在目标检测任务中引入自注意力机制 (Self-Attention), 通过构建特征之间的全局关联,提升了检测器的上下文建模能力。由于自注意力机制具

有强大的时序信息提取能力,因此被广泛应用于视频目标检测中。Faster R-CNN^[9] 是视频目标检测任务的基础检测器。典型的视频目标检测方法 FGFA^[10] (Flow-Guided Feature Aggregation) 以 Faster R-CNN 为基础检测器,引入光流估计^[11],通过引导历史帧特征向关键帧对齐,实现跨帧特征融合,提升了检测器对模糊帧的鲁棒性。近年来,大量方法尝试通过注意力机制或内存模块实现更有效的跨帧特征聚合。例如,MEGA^[12] 提出的全局与局部的记忆增强模块,可同时建模视频中长期和短期的上下文信息;TROIA^[13] 则基于 ROI-Align^[14] 技术,实现目标区域的精细特征对齐与融合;QueryProp^[15] 提出轻量级查询传播模块,仅针对高置信度目标传播特征信息,降低了计算成本,提高了聚合效率。此外,SELSA^[16] 采用图神经网络^[17] (GNN) 建模目标间关系,实现了更结构化的时序特征融合。TransVOD^[18] 是第一个基于时空 Transformer 架构的端到端视频目标检测网络,实现了端到端的检测。Eventful TransformersVOD^[19] 则在此基础上提出了一种基于差值的策略,用于稀疏更新注意力值结果,加快推理速度。另外,受扩散模型的启发,DiffusionVID^[20] 网络通过细化随机噪声框的方式,以获得视频序列中的原始对象框,有效地从视频中的退化图像中细化目标检测框。STFVOD^[21] 提出了一种空融合框架用于提升视频目标检测的性能,通过关注时间和空间特性来改进检测结果。

2.2 基于跟踪的视频目标检测任务

目标跟踪^[22] (Object Tracking) 任务是在视频中对同一目标在不同帧之间建立身份关联,从而得到目标的轨迹信息。典型方法如 ByteTrack^[23],采用基于高分和低分检测框的联合匹配策略,提升了在遮挡环境下的跟踪鲁棒性。其他方法如 CenterTrack^[24],FairMOT^[25] 等,通过回归目标的位置信息并学习其运动偏移,实现了检测与跟踪一体化。

由于检测任务和跟踪任务具有相似性,因此通过跟踪提取时序信息来辅助检测,也是视频目标检测任务中的一种常见做法。D&T^[26] (Detect and Track) 将检测与跟踪结合,借助跟踪信息优化目标检测的连续性。TFVOD^[27] 提出了一个即插即用的卷积回归跟踪器,通过运动偏移预测目标位置来优化对目标的定位,从而提升检测精度。GMVOD^[28] 设计了全局存储和局部连续性网络,提高了检测精度,并且设计了目标跟踪器,利用视频的局部连续性来跟踪从关键帧到非关键帧的对象,加快推理速度。2023 年提出的 YOLOV^[29] 方法则基于一阶段目标检测器构建,通过特征选择降低计算量,将跨帧聚合机制引入一阶段目标检测器,提高了检测精度和推理速度。最近提出的 FAIM^[30] 在特征聚合时引入目标分割思想进行实例级特征聚合,进一步排除背景对聚合的干扰。

然而,上述方法大多依赖检测器生成的候选框进行特征聚合。在严重遮挡等复杂情况下,仅依赖检测器的预检测结果往往无法覆盖全部的潜在目标区域。针对这一问题,本文进一步探索检测与跟踪任务的结合,将检测器的预检测结果与跟踪器的预测结果进行融合,从而优化候选区域的生成与筛选策略。

3 本文方法

本文提出了联合跟踪的特征聚合视频目标检测方法,通

过联合跟踪任务和视频目标检测任务,进一步改善视频中存在遮挡、形变等图像退化因素导致的检测失败问题。

3.1 模型框架

鉴于视频数据既存在多种图像退化问题,又蕴含丰富的时序信息的特点,从其他帧中获取对目标帧具有支持作用的信息,而非孤立地检测每一帧,是提升视频目标检测精度的关键。YOLOV 方法将区域选择模块置于一阶段检测器预测头之后,从而在保证精度的同时提高检测速度。然而,该方法在高遮挡场景下仍面临性能下降的问题。目标跟踪通过构建目标轨迹,能够反映目标在时序上的一致性,并预测其可能出现的位置。因此,本文在检测框架中引入目标跟踪,利用跟踪轨迹及其预测信息进一步挖掘视频的时序信息。

本文整体框架如图 1 所示。首先,视频帧依次输入 YOLOX 检测器,通过回归分支与分类分支生成大量候选区

域,并提取相应的回归特征与分类特征。与此同时,基于检测结果,跟踪器对目标进行时序跟踪,并利用预测轨迹补充检测器未能识别的目标区域(如被遮挡目标或低置信度目标)。对于跟踪轨迹上的目标区域,提取并存储其分类特征与回归特征,构建 Memory 模块,以供后续的时序特征聚合使用。在候选区域生成阶段,候选区域生成与选择模块(Candidate Region Generation and Selection Module,CRGSM)融合 YOLOX 检测器输出的候选框与跟踪器补充的预测框,从而提升定位精度,并输出融合后更鲁棒的候选区域。随后,将这些候选区域输入帧间特征聚合模块(Inter-Frame Feature Aggregation Module,IFFAM),该模块利用多头注意力机制分别对回归特征与分类特征建模时间上下文关系。聚合过程中,特征通过矩阵乘法与拼接操作进行整合,得到新的聚合特征,最后将经过聚合后得到的聚合特征输入线性层,生成最终的检测结果。

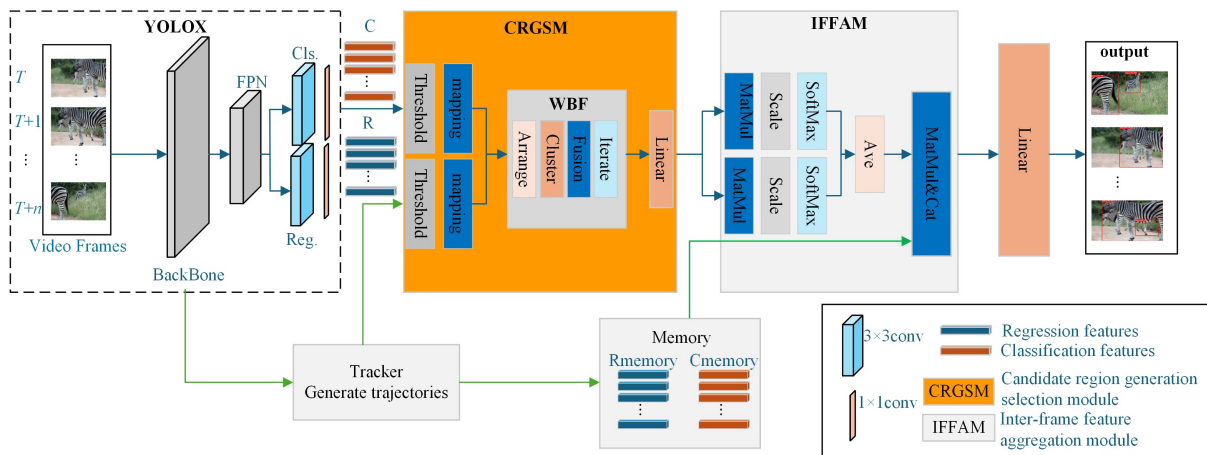


图 1 网络整体架构

Fig. 1 Overall network architecture

3.2 候选区域生成模块

视频帧中的单帧检测易受遮挡、运动模糊等图像退化因素的影响,导致出现漏检现象。此时,检测器生成的候选区域往往置信度较低,难以提供准确的候选框。如果通过降低置信度阈值来保留更多候选框,虽然能够获取部分漏检目标,但会引入大量低置信度的干扰区域,影响后续检测精度。为此,本文设计了候选区域生成与选择模块,结合检测器与跟踪器的优势,生成鲁棒且高质量的候选区域,以供后续特征聚合模块使用。

在候选区域提取阶段,首先利用 YOLOX 检测器获取每帧的初步检测框,得到目标可能存在的区域。经过 FPN^[31] 处理之后,根据置信度得分对候选区域进行降序排序,并通过设定阈值剔除置信度过低的候选框,从而在降低计算量的同时,保留更多高置信度候选框的空间分布信息。与此同时,引入轻量级目标跟踪器 ByteTrack,利用其在前几帧构建的轨迹预测当前帧中目标的潜在位置,生成一组预测框作为候选区域的补充。随后,将检测器输出的初步候选框与跟踪器预测框通过加权框融合算法^[32](Weighted Boxes Fusion,WBF)进行融合,实现候选框融合,提升对遮挡与低质量候选区域的选取能力。

在 WBF 融合过程中,首先将不同来源的候选框按置信

度从高到低进行排序,存入列表 B 。然后建立两个空列表: L 用于存储每个目标的所有候选框, F 用于存储每个目标的融合结果。遍历列表 B 中的候选框,在 F 中查找与其匹配的融合框。若未找到匹配项,则将该框作为新目标框添加至 L 与 F 的末尾;若找到匹配项,则将该框添加至 L 中对应位置。遍历完成后,对 L 中每个目标对应的所有候选框,按式(1) 计算融合后的置信度,并根据式(2)与式(3)计算融合框的坐标。加权融合过程如式(1)一式(3)所示:

$$C = \sum_{i=1}^T C_i / T \quad (1)$$

$$X = \sum_{i=1}^T C_i X_i / \sum_{i=1}^T C_i \quad (2)$$

$$Y = \sum_{i=1}^T C_i Y_i / \sum_{i=1}^T C_i \quad (3)$$

其中, C 代表融合之后新框的置信度, C_i 表示参与融合的 i 个框的置信度, T 表示参与融合的框的数量, X 和 Y 表示融合后融合框的坐标, X_i 和 Y_i 表示原始候选框的坐标。融合时,使用原始框的置信度 C_i 对原始框的坐标 X_i 进行加权,加权得到的结果对置信度 C_i 进行归一化运算后,得到新框的坐标 X, Y 。

对于候选检测框,将其初步检测所得的置信度作为融合时的权重。对于跟踪预测框,由于跟踪器输出的预测结果本身并不具备置信度,在基于检测的跟踪框架中,一个跟踪轨迹

由一系列成功匹配的检测框构成,因此可以考虑跟踪预测框的置信度继承自该轨迹中最后一个成功关联的检测框的置信度。因此在融合计算中,跟踪预测框的置信度取自该关联检测框的置信度。

3.3 帧间特征聚合模块

为提升检测精度,本文设计了帧间特征聚合模块,将当前帧候选区域的特征与其他帧及 Memory 模块中存储的目标轨迹特征进行融合。该模块基于改进的多头自注意力机制,计算当前帧特征与参考帧特征之间的相似度。注意力的基本计算方式如式(4)~式(6)所示:

$$SA(F) = \text{softmax}(A)V \quad (4)$$

$$A = QK^T / \sqrt{d} \quad (5)$$

$$MSA(Z) = \text{concat}(SA_1(Z), SA_2(Z), \dots, SA_n(Z)) \quad (6)$$

通过式(4)对每个位置的 V 进行加权求和,得到融合后的特征表示 $SA(F)$ 。最后如式(6)所示,将所有注意力头的输出通过 concat 拼接,形成多头注意力。该机制能够聚焦于更具判别力的区域特征,从而提升检测精度。

在特征聚合阶段,首先从 Memory 中获取历史帧的分类特征 C 与回归特征 R ,并通过线性映射将其统一到维度 d 。当前帧的回归与分类特征用于构建 Q 矩阵,历史帧特征用于构建 K 与 V 矩阵。与传统注意力机制不同,本文在计算过程中引入置信度调制,以缓解目标退化下余弦相似性引发的同质化问题。

改进后的注意力计算过程如图 2 所示,回归特征注意力的计算方式与之相同,将输入从分类特征换为回归特征,即可得到回归特征的注意力。

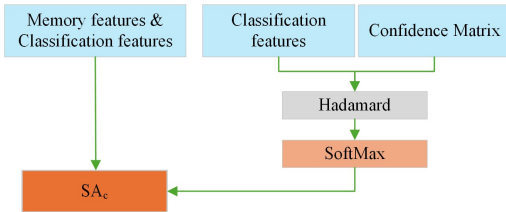


图 2 改进后的注意力计算

Fig. 2 Improved attention calculation

从原始检测器中提取每一个提案的置信度分数、分类分数(Classification Score)和 IoU 分数(IoU Score)。这些分数共同构成了置信度矩阵 P 。基于 P 中的分数,构建两个新的矩阵 S_c 和 S_r 。 S_c (用于分类分支)将每个提案的分类分数沿着列的方向重复 n 次(n 是总提案数量)。按照同样的方法构建回归分支的置信度分数矩阵 S_r 。通过 Hadamard 乘积调制注意力权重,计算时,针对分类特征和回归特征,形成改进后的注意力计算过程,如式(7)和式(8)所示:

$$SA_c(F_c) = \text{softmax}(S_c \odot A_c)V_f \quad (7)$$

$$SA_r(F_r) = \text{softmax}(S_r \odot A_r)V_f \quad (8)$$

其中, F_c 表示分类特征, A_c 是分类特征计算得到的注意力权重矩阵, S_c 是分类置信度矩阵。分类分支的特征和记忆模块中存储的轨迹上目标的分类特征相加得到 V_f 。Hadamard 乘积用于将置信度调制到注意力权重上。通过融合分类置信度信息,抑制低可信候选区域的注意力影响,使注意力聚焦于更可

靠的目标区域。回归分支注意力的计算过程和分类注意力的计算过程相同。

最后,将两个分类和回归分支的注意力输出与原始特征拼接,形成增强后的目标区域特征,如式(9)所示:

$$SA(F) = \text{concat}(SA_c(F_c) + SA_r(F_r), V_c) \quad (9)$$

其中, $SA_c(F_c)$ 表示分类注意力输出; $SA_r(F_r)$ 表示回归注意力输出; V_c 由原始分类特征构成。将分类与回归两个注意力输出相加(加权聚合),与原始特征 V_c 拼接,融合后的目标区域增强特征 $SA(F)$ 保留原始表征的同时引入时序增强信息,最终形成更鲁棒的目标表示,送入后续检测头进行检测。

通过 IFFAM 模块,不仅融合了跨帧的上下文信息,还结合了检测置信度调节聚合权重,相比之前的方法提升了遮挡、模糊等退化情况下的目标识别能力。

4 实验结果分析

本章首先介绍实验中使用的数据集,其次具体描述模型训练的实验设置,最后将提出的方法和常用的视频目标检测方法进行比较。

4.1 数据集

ImageNet DET (Detection) 数据集来自于 ILSVRC^[33] (ImageNet Large Scale Visual Recognition Challenge) 挑战赛,用于图像目标检测任务。该数据集的数据主要从 ImageNet 中挑选并标注,涵盖 200 个常见物体类别,如动物、交通工具、家用器具等。ImageNet VID (Video Object Detection) 数据集是 ImageNet 基于视频构建的目标检测子集,首次出现在 ILSVRC 2015 视频任务 (ILSVRC 2015 VID) 中。该数据集专门用于视频目标检测任务,旨在推动模型利用时序信息提高检测稳定性与精度。

ImageNet VID^[34] 包含共计 3862 个训练视频片段、555 个验证视频片段,以及 937 个测试视频片段,帧率为 25 FPS。数据集中的每段视频均标注了视频中出现目标的边界框位置和物体类别标签。

OVIS^[35] (Occluded Video Instance Segmentation) 数据集最初旨在针对视频中普遍存在的遮挡问题进行建模与评测。该数据集包含 5223 段视频、约 296000 帧图像,涵盖 25 个类别,其中超过 70% 的实例存在不同程度的遮挡,且严重遮挡比例显著高于传统视频数据集(如 ImageNet VID 和 YouTube-VIS)。尽管 OVIS 最初面向视频实例分割任务,但其高遮挡特性同样适用于视频目标检测任务,在近几年的工作中也常被用于检验模型在复杂遮挡场景下的鲁棒性。

在实验中,使用了 ImageNet VID 数据集和 ImageNet DET 数据集的数据参与训练。ImageNet VID 数据集的 30 个目标类别是 ImageNet DET 数据集中 200 个基本级别类别的子集。在训练时,将 ImageNet VID 和 ImageNet DET 数据集中类别相同的部分数据进行整合,使用整合后的数据参与训练,使用 ImageNet VID 提供的测试集的视频片段进行测试。

4.2 实验设置和评价指标

首先,使用 YOLOX 提供的 COCO 预训练模型初始化基础检测器权重。将输入视频图像的大小统一调整到 $576 \times$

576,使用 16 个 batch 对基础检测器进行 8 个 epoch 的训练。对于学习速率,采用 YOLOX 中使用的余弦学习速率模式,其他参数和基线方法保持一致。随后,在 3090 GPU 上对特征聚合模块进行了 10 万次迭代的微调,批处理大小为 16,仅对 YOLOX 预测头中的线性层、新增的视频目标分类分支和多头注意力进行微调。对于初步的检测框,将 NMS 的阈值设置为 0.5,以保留更多的可信候选。最后,在 3090 GPU 上测试了模型精度。

评价指标为 AP50。在此基础上,根据同一目标在前后两帧的交并比大小,设定不同的阈值,将目标分为快速移动的目标、中速移动的目标和慢速移动的目标。本文方法和基线

方法保持一致,设置 IoU 阈值为 0.7 和 0.4 来区分快速、中速、慢速目标。前后两帧的 IoU 在[0,0.4)内的目标定义为快速目标,前后两帧的 IoU 在[0.4,0.7) 内的目标定义为中速目标,前后两帧的 IoU 在[0.7,1] 内的目标定义为慢速目标,分别用 AP50_f,AP50_m,AP50_s评价快、中、慢速目标的精度。

4.3 实验结果与分析

4.3.1 对比实验

首先,在 ImageNet VID 数据集上分别测试了视频目标检测方法,如 FGFA, SELSA, MEGA, MAMBA, TFVOD, ClipVID^[27,36],D&T,TransVOD^[18],ODDVOD^[37],EVOD^[38],YOLOV,FAIM 的性能指标,测试结果如表 1 所列。

表 1 常用的视频目标检测方法在 ImageNet VID 数据集上的定量比较

Table 1 Quantitative comparison of commonly used video object detection methods on the ImageNet VID dataset (%)

Model	Backbone	Base Detector	AP50	AP50 _s	AP50 _m	AP50 _f	Params
FGFA	ResNet-101	R-FCN	76.4	83.2	78.7	64.0	6.133×10 ⁷
SELSA	ResNet-101	Faster R-CNN	80.7	87.5	83.4	68.1	6.212×10 ⁷
MEGA	ResNet-101	Faster R-CNN	82.5	89.3	84.7	70.1	6.481×10 ⁷
MAMBA	ResNet-101	Faster R-CNN	84.3	91.1	86.7	72.0	5.843×10 ⁷
TFVOD	ResNet-101	Faster R-CNN	81.1	87.9	83.4	69.1	6.347×10 ⁷
ClipVID	ResNet-101	Faster R-CNN	84.4	91.2	85.3	72.1	6.293×10 ⁷
D&T	ResNet-101	Faster R-CNN	79.8	86.6	82.1	67.5	5.871×10 ⁷
TranVOD	ResNet-101	DeformDETR	82.1	89.1	84.2	69.9	7.986×10 ⁷
ETVOD	ResNet-101	DeformDETR	82.0	89.1	84.1	69.7	8.015×10 ⁷
STFVOD	MCSP v5	YOLOX-L	81.2	88.1	84.3	68.7	6.051×10 ⁷
ODDVOD	MCSP v5	YOLOX-L	84.4	91.2	87.1	71.3	6.013×10 ⁷
EVOD	MCSP v5	YOLOX-L	80.3	87.1	82.6	67.9	6.073×10 ⁷
YOLOV	MCSP v5	YOLOX-L	83.6	90.1	86.4	72.2	5.945×10 ⁷
FAIM	MCSP v5	YOLOX-L	84.6	91.4	86.5	72.3	6.159×10 ⁷
Ours	MCSP v5	YOLOX-L	85.1	92.3	87.1	72.1	5.951×10 ⁷

表 1 列出了各方法在视频目标检测数据集 ImageNet VID 上的定量对比结果。由表中数据可见,FGFA, SELSA, MEGA, MAMBA, TFVOD, ClipVID 和 D&T 方法均基于传统的两阶段目标检测器 Faster R-CNN 架构,其中 ClipVID 与 MAMBA 分别取得了 84.4% 和 84.3% 的检测精度,而本文方法的检测精度达到了 85.1%,优于上述方法。ODDVOD,ETVOD 与 STFVOD 展现出了较优的平衡,但在快速运动目标下的性能仍显不足。YOLOV 采用一阶段目标检测器作为基础模型,并通过特征聚合实现了 83.6% 的检测精度;FAIM 在特征聚合过程中引入目标分割思想,以进一步减少背景干扰,使检测精度提升至 84.6%。

OVIS 数据集是用于目标分割任务的视频遮挡数据集,也适用于视频目标检测任务。为了证明本文方法的泛化性和通用性,补充了本文方法、YOLOV 方法和 FAIM 方法在 OVIS 数据集上的实验结果并加以分析,结果如表 2 所列。

表 2 在 OVIS 数据集上的定量比较结果

Table 2 Quantitative comparison results on the OVIS dataset (%)

Methods	AP50	Slow	Medium	Fast
FAIM	76.9	81.7	76.4	64.3
YOLOV	75.1	80.4	74.8	63.9
Ours	77.1	82.3	76.1	64.2

本文方法的整体 AP50 为 77.1%,相比 YOLOV 提升了 2 个百分点,相比 FAIM 也有小幅提升,证明了该方法在跨数

据集上的良好泛化能力。从不同运动速率来看,本文方法在 Slow 场景下表现最优,说明其特征聚合机制能够有效利用时序信息;在 Medium 场景中与 FAIM 相当,整体保持稳定;而在 Fast 场景下略低于 FAIM,显示出在快速运动和强遮挡下仍有改进空间。在 OVIS 数据集上的实验结果表明,本文方法适用于不同场景的数据,具有较强的鲁棒性。

本文方法在 YOLOV 的基础上引入跟踪辅助候选增强机制与置信度引导特征聚合机制,在处理遮挡帧与模糊帧时具备更强的目标提取能力,从而使整体精度提升至 85.1%。具体而言,跟踪器预测框用于增强候选区域,有效缓解了在遮挡或低置信度场景下候选区域缺失的问题;帧间特征聚合模块则在置信度引导下实现了更具针对性的特征融合,从而进一步提升了检测性能。

参数量(Parameter Count)是衡量模型大小和复杂度的核心指标之一。这些模型可以大致分为 3 类进行分析。1)基于传统两阶段方法的模型,包括 FGFA 到 D&T 等多个模型。由于两阶段检测器的特性,将目标检测分为两个阶段检测,复杂网络特性决定了这类模型往往参数量较大。2)基于 Transformer 的模型,如 TranVOD,其基础检测器替换为 DeformDETR,这是一种基于 Transformer 的检测器。Transformer 架构的参数规模大于传统的 CNN 检测器架构,尽管性能优异,但模型复杂度也更高。3)使用单阶段检测器的视频目标检测方法,以基于 YOLOX 的 EVOD 和 YOLOV 为

例,这类方法的参数量比上述方法的参数量少,随之而来的是精度的挑战。

本文方法以 YOLOV 为基线,引入追踪信息进一步提升检测精度,同时本文方法的参数量只是小幅提升。这是因为 ByteTrack 本质上是一个轻量级的跟踪框架,主要是基于卡尔曼滤波和匈牙利算法,不包含大量深度网络参数。在本文模型中,ByteTrack 和基线方法的模型共用同一个基础检测器 YOLOX,而它的核心跟踪模块几乎不增加深度卷积层权重,更多的是计算逻辑和简单线性滤波器。

为进一步衡量本文方法的模型复杂度,对基线方法和一些常用方法的 GFLOPs 和推理时间进行定量比较。实验结果如表 3 所列。

表 3 推理时间定量分析

Table 3 Quantitative analysis of inference time

	AP50/%	GFLOPs	Times/ms
ETranVOD	82.0	156.8	20.1
FAIM	84.6	151.1	18.5
YOLOV	83.6	143.1	16.3
Ours	85.1	147.2	17.6

在计算开销方面,本文方法的 GFLOPs 介于 YOLOV 与 FAIM 之间,表明其在保证较高精度的同时,计算复杂度控制在合理水平;本文方法的推理时间为 17.6 ms,略高于 YOLOV,但低于 ETranVOD(20.1 ms),依旧保持了较好的实时性。整体而言,本文方法在精度与效率之间实现了更优的平衡。

图 3 展示了不同模型对不同速度运动目标的检测精度。绿色、蓝色、红色折线分别表示不同模型对慢速、中速、快速目标的检测精度。其中最右侧是本文方法的检测结果。从细分的目标速度精度表现来看,本文方法在慢速与中速目标的检测中表现更为突出。这一优势源于跟踪器在此类场景中能够更稳定地提取时序信息和目标轨迹,为特征聚合提供精准指引。而对于快速移动目标,运动模糊与位移较大可能干扰跟踪器的关联与预测,从而使特征聚合的优势未能充分发挥。

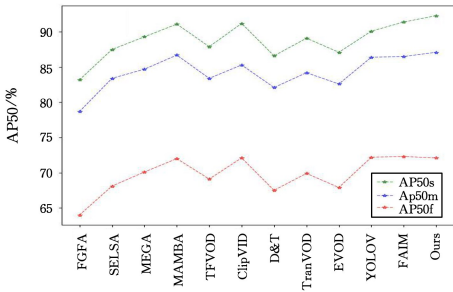


图 3 不同速度目标的检测精度比较(电子版为彩图)

Fig. 3 Comparison of detection accuracy of targets with different speeds

4.3.2 可视化实验结果

本节对 ImageNet VID 数据集中具有代表性的视频帧进行可视化对比分析。图 4 展示了本文方法与基线 YOLOV 方法在遮挡、低分辨率及背景干扰等典型场景下的检测结果。在图 4 中适当放大了目标的检测标签并标于图中。结果表明,本文方法在复杂场景中能够更准确地定位并识别目标,

尤其在遮挡的情况下,优势明显。

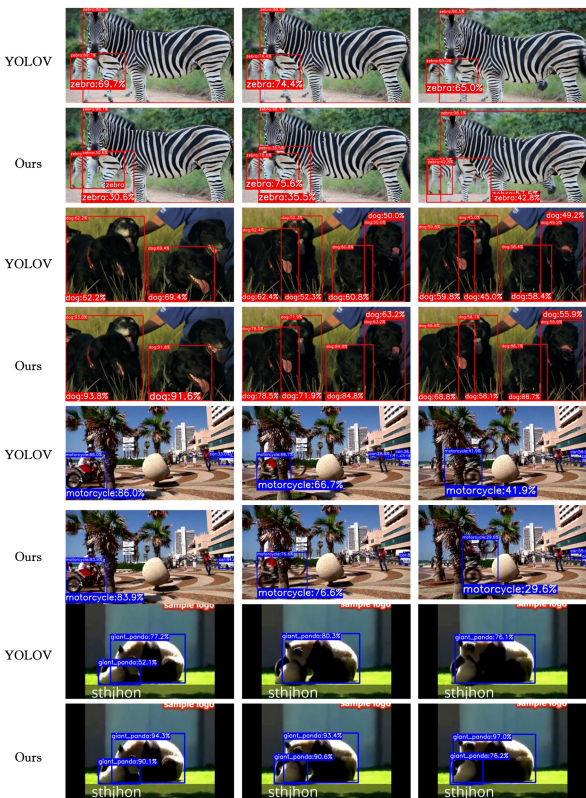


图 4 在多种场景下的可视化结果分析

Fig. 4 Visual results analysis in a variety of scenarios

图 5(a)列为待检测的视频帧,其中绿色框标注了被遮挡目标的位置与大小;图 5(b)列为基线 YOLOV 方法的检测结果;图 5(c)列为本文方法的检测结果。对比结果表明,基线方法在处理被遮挡目标时,存在漏检和检测框覆盖不完整的问题。例如,漏检被遮挡的小熊猫,或在摩托车检测中虽检测到目标,但检测框未能完全覆盖目标。而本文方法则在这种遮挡的情况下能够稳定、准确地检测出被遮挡目标,并保持较高的检测一致性与检测精度。



图 5 遮挡目标检测结果可视化结果分析(电子版为彩图)

Fig. 5 Analysis of visualization results of occluded object detection

图 6 展示了基线方法和本文方法在同一视频中的目标检测结果。第一行为基线 YOLOV 方法的可视化检测结果,第二行为本文方法的可视化检测结果,纵向自上而下依次为按时间采样得到的不同视频帧。对比可见,在检测被遮挡目标时,基线方法存在漏检现象,图中以黄色矩形框标注;而本文方法能够成功检测出该类目标。在面对遮挡情况时,基线方法由于遮挡等因素造成的漏检问题,通过本文方法可以得到有效缓解,展现出本文方法更强的鲁棒性。

结合可视化的结果分析,进一步证明了本文方法在面對遮挡问题时的有效性,相较于基线方法,本文方法能更准确、稳定地检测被遮挡的目标。

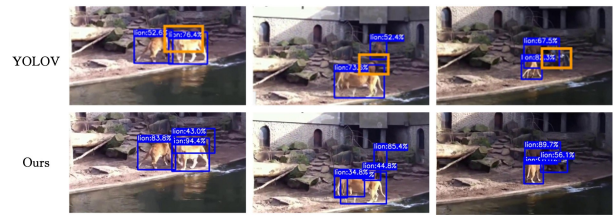


图 6 目标检测结果可视化分析(电子版为彩图)

Fig. 6 Analysis of visualization results of object detection

4.3.3 消融实验

为验证本文提出的帧间特征聚合模块(IFFAM)与候选区域生成模块(CRGSM)的有效性,本节开展消融实验,分别评估两个模块在检测精度提升中的作用。

在 ImageNet VID 验证集上分别对 CRGSM 模块和 IF-FAM 进行消融实验,比较不同模块组合对检测精度的影响,从而验证本文设计模块的有效性。消融实验的实验结果如表 4 所列。

表 4 消融实验结果比较

Table 4 Comparison of ablation experimental results (%)

	AP50	Slow	Medium	Fast
实验 1	84.2	90.1	86.4	72.2
实验 2	84.7	91.2	86.7	71.9
实验 3	85.1	92.3	87.1	72.1

实验 1 仅测试基线 YOLOV 的检测精度;实验 2 在基线基础上引入 CRGSM 模块后,AP50 提升至 84.7%,说明通过跟踪预测区域补偿,并结合 WBF 融合不同来源的候选区域,可有效提升候选区域质量,为后续特征聚合提供更优的输入;实验 3 在实验 2 基础上进一步加入 IFFAM 模块,AP50 提升至 85.1%。该模块引入历史轨迹特征参与聚合,显著提升了聚合精度。最终两模块的结合实现了从候选区域生成到特征聚合的协同优化,提高了模型在遮挡、模糊等干扰场景下的鲁棒性。

消融实验证明了 IFFAM 和 CRGSM 模块的有效性,在此基础上,本文分析了 CRGSM 保留候选区域数量对实验结果的影响,实验结果如表 5 所列。

表 5 候选区域数量对实验结果的影响分析

Table 5 Analysis of the impact of candidate area quantity on experimental results

<i>a</i>	AP50/%	Times/ms
10	82.1	17.1
20	83.7	17.9
30	85.1	18.6
50	85.1	20.1
70	85.2	24.3
100	85.2	30.7

其中,*a* 为保存的候选区域数量。从表 5 中结果可以看出,随着 CRGSM 模块中保留候选区域数量的增加,模型的 AP50 逐渐提升,在数量为 30 时达到 85.1%,此后继续增加

候选区域数量对精度提升有限甚至趋于饱和,同时,推理时间显著增加。因此,保留 30 个候选区域在精度与效率之间取得了较优的平衡。

此外,本文对不同的候选区域筛选方法,如非极大值抑制(NMS)、置信度高低排列(TopK)和目标框加权融合(WBF)方法进行了分析,实验结果如表 6 所列。

表 6 不同候选区域筛选方法实验结果比较

Table 6 Comparison of experimental results of different candidate area selection methods (%)

Methods	AP50	Slow	Medium	Fast
NMS	84.2	90.1	86.4	72.2
TopK	84.3	90.8	86.3	72.2
WBF	85.1	92.3	87.1	72.1

对比 3 种候选区域筛选方法可以看出,WBF 融合算法在整体 AP50 上达到 85.1%,优于 NMS(84.2%)和 TopK(84.3%),且在 Slow 与 Medium 场景下均取得最优结果,显示其能更有效地整合多候选框信息,提升了检测的稳定性与鲁棒性。整体来看,WBF 能够带来更明显的性能提升。

为了验证记忆模块中轨迹长度 *f* 对检测性能的影响,即参与聚合的目标轨迹上的帧数对检测性能的影响,还对比了不同轨迹长度对视频目标检测 AP50 的影响,结果如表 7 所列。

表 7 记忆模块存储轨迹长度对比分析

Table 7 Comparative analysis of trajectory lengths stored in memory module

<i>f</i>	AP50/%
15	83.8
20	84.4
25	84.7
30	85.1
35	85.1

随着轨迹长度的增加,AP50 稳步上升,从 83.8%(*f*=15)逐步提升至 85.1%(*f*=30 和 *f*=35)。这一结果表明,适当扩展轨迹长度有助于模型捕捉更完整的时序上下文特征,从而在目标遮挡、运动模糊等复杂场景中保持更稳定的检测结果。然而,轨迹长度从 30 增加到 35 时,检测精度保持在 85.1%。当轨迹过长时,模型可能引入过多冗余或噪声信息,甚至对当前检测产生干扰。

通过对比实验和消融实验的定量分析,验证了本文方法的有效性,它在检测精度上优于基线方法和常用的视频目标检测方法。

结束语 相较于静态图像检测,视频目标检测可利用视频帧间的时序信息辅助当前帧的目标检测。然而,现有方法在严重遮挡情况下仍存在性能下降的问题。针对这一问题,本文提出了一种结合目标跟踪的视频目标检测方法,在传统视频目标检测框架基础上引入跟踪器,并设计了可联合利用跟踪预测信息与检测结果的特征选择模块,同时利用目标轨迹特征参与特征聚合。在 ImageNet VID 数据集上的实验结果表明,该方法较现有方法取得了更高的检测精度。然而,本文方法在引入额外的特征聚合与跟踪网络后,增加了模型的

复杂度,对所有视频帧均进行聚合处理,限制了推理速度。同时,跟踪网络对目标速度变化较为敏感。未来的研究可围绕如何实现更灵活、精准的特征聚合机制展开,在保证精度的前提下进一步提升运行效率。

参 考 文 献

[1] ZOU Z, CHEN K, SHI Z, et al. Object detection in 20 years: A survey[J]. *Proceedings of the IEEE*, 2023, 111(3): 257-276.

[2] ZHU X, DAI J, YUAN L, et al. Towards high performance video object detection[C]// *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018: 7210-7218.

[3] QI Q, WANG X, HOU T, et al. FastVOD-Net: A real-time and high-accuracy video object detector[J]. *IEEE Transactions on Intelligent Transportation Systems*, 2022, 23(11): 20926-20942.

[4] KANG K, LI H, XIAO T, et al. Object detection in videos with tubelet proposal networks[C]// *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017: 727-735.

[5] GUO M H, XU T X, LIU J, et al. Attention mechanisms in computer vision: A survey[J]. *Computational Visual Media*, 2022, 8(3): 331-368.

[6] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]// *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2018: 7132-7141.

[7] WOO S, PARK J, LEE J Y, et al. CBAM: Convolutional block attention module[C]// *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2018: 3-19.

[8] ZHU X, SU W, LU L, et al. Deformable detr: Deformable transformers for end-to-end object detection[J]. *arXiv*, 2010. 04159, 2020.

[9] GIRSHICK R. Fast R-CNN[C]// *Proceedings of the IEEE International Conference on Computer Vision*. 2015: 1440-1448.

[10] ZHU X, XIONG Y, DAI J, et al. Flow-guided feature aggregation for video object detection[C]// *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017: 408-417.

[11] SUN D, ROTH S, BLACK M J. Secrets of optical flow estimation and their principles[C]// 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, 2010: 2432-2439.

[12] CHEN K, SUN X, LI J, et al. Memory enhanced global-local aggregation for video object detection[C]// *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020: 10326-10335.

[13] GONG T, WANG C, ZHANG Q, et al. TROIA: A temporal region-based object interaction aware approach for video object detection[C]// *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI, 2021: 1424-1432.

[14] HE K, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]// *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017: 2961-2969.

[15] HE F, GAO N, JIA J, et al. QueryProp: Object Query Propagation for High-Performance Video Object Detection[C]// *Pro-*

ceedings of the AAAI Conference on Artificial Intelligence. 2022: 834-842.

[16] WU H, CHEN Y, WANG N, et al. Sequence level semantics aggregation for video object detection[C]// *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2021: 3026-3035.

[17] SHI W, RAJKUMAR R. Point-GNN: Graph neural network for 3d object detection in a point cloud[C]// *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020: 1711-1719.

[18] ZHOU Q, LI X, HE L, et al. TransVOD: End-to-end video object detection with spatial-temporal transformers[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, 45(6): 7853-7869.

[19] DUTSON M, LI Y, GUPTA M. Eventful transformers: Leveraging temporal redundancy in vision transformers[C]// *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023: 16911-16923.

[20] ANWAR N, BILODEAU G A, BOUACHIR W. STF: Spatio-Temporal Fusion Module for Improving Video Object Detection[J]. *arXiv*: 2402. 10752, 2024.

[21] ROH S D, CHUNG K S. DiffusionVID: Denoising Object Boxes With Spatio-Temporal Conditioning for Video Object Detection[J]. *IEEE Access*, 2023, 11: 121434-121444.

[22] BEWLEY A, GE Z, OTT L, et al. Simple online and real time tracking[C]// *Proceedings of the IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016: 3464-3468.

[23] ZHANG Y, WANG C, WANG X, et al. ByteTrack: Multi-Object Tracking by Associating Every Detection[C]// *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2022: 1-17.

[24] ZHOU X, WANG D, KRÄHENBÜHL P. CenterTrack: Tracking objects as points[C]// *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2021: 9974-9983.

[25] ZHANG Y, WANG C, WANG X, et al. FairMOT: On the fairness of detection and re-identification in multiple object tracking[J]. *International Journal of Computer Vision*, 2022, 130(2): 306-317.

[26] FEICHTENHOFER C, PINZ A, ZISSERMAN A. Detect to track and track to detect[C]// *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017: 3038-3046.

[27] LYU Y, YANG M Y, VOSSELMAN G, et al. Plug & Play Convolutional Regression Tracker for Video Object Detection[J]. *arXiv*: 2003. 00981, 2020.

[28] HAN L, YIN Z. Global memory and local continuity for video object detection[J]. *IEEE Transactions on Multimedia*, 2022, 25: 3681-3693.

[29] SHI Y, WANG N, GUO X. YOLOV: Making still image object detectors great at video object detection[C]// *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*. 2023: 2254-2262.

[30] HASHMI K A, SHEIKH T U, STRICKER D, et al. Beyond

boxes:Mask-guided spatio-temporal feature aggregation for video object detection[C]//Proceedings of the Winter Conference on Applications of Computer Vision (WACV). IEEE, 2025: 8111-8122.

[31] 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, 2017,39(6):1137-1149.

[32] ZHOU Z,CUI Y,ZHANG J,et al. Weighted boxesfusion:Ensembling boxes for object detection models[J]. arXiv: 1910.13302,2019.

[33] RUSSAKOVSKY O,DENG J,SU H,et al. ImageNet large scale visual recognition challenge[J]. International Journal of Computer Vision,2015,115(3):211-252.

[34] ZHU H,WEI H,LI B,et al. A review of video object detection: Datasets, metrics and methods [J]. Applied Sciences, 2020, 10(21):7834.

[35] QI J,GAO Y,HU Y,et al. Occluded video instance segmentation:A benchmark[J]. International Journal of Computer Vision,2022,130(8):2022-2039.

[36] DENG C,CHEN D,WU Q. Identity-consistent aggregation for video object detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision(ICCV). 2023.

[37] ZHANG B,WANG S,LIU Y,et al. Object detection difficulty: Suppressing over-aggregation for faster and better video object detection[C]//Proceedings of the 31st ACM International Conference on Multimedia. 2023:1768-1778.

[38] SUN G,HUA Y,HU G,et al. Efficient one-stage video object detection by exploiting temporal consistency[C]// Proceedings of the European Conference on Computer Vision (ECCV). Springer,2022.



WANG Jiahui, born in 2001, postgraduate. His main research interest is video object detection.



WANG Hongyu, born in 1968, Ph. D, professor. His main research interests include object detection,object tracking and video image processing.

(责任编辑:何杨)