

一种基于帧间运动目标框链的视频目标检测算法

李 敏^{1,2}, 李凌涵^{1,2}, 白入文^{1,2}, 姜 淼^{1,2}, 任俊星^{1,2}, 孟 博³,
杨 阳^{1,2}, 黄子豪⁴, 黄伟庆^{1,2}

¹中国科学院信息工程研究所 北京 中国 100093

²中国科学院大学网络空间安全学院 北京 中国 100049

³北京理工大学 北京 中国 100081

⁴华北电力大学 北京 中国 102206

摘要 目标检测是计算机视觉领域的基础任务,在视频监控、自动驾驶等领域有着广泛的应用。近年来,随着深度卷积神经网络的发展,YOLO、Faster R-CNN等基于单帧图像的目标检测算法已经取得显著的效果,但在实际使用中,由于视频中目标物体模糊、遮挡、变形等问题,单帧图像目标检测算法无法直接应用在视频目标检测任务中。目前主流的视频目标检测算法,如FGFA、Seq-NMS等,无法同时兼顾速度与准确性,不能满足实际应用的需求。如何在保障检测实时性的前提下,提升检测的准确性,是当前视频目标检测算法面临的挑战。本文提出一种基于帧间运动目标框链的后处理方法,以单帧检测算法为基础,从后处理部分入手,引入帧间运动信息对检测结果进行增强。我们利用基于距离的交并比(Distance Intersection over Union, DIoU)表示帧间运动信息,并提出相邻帧之间的同一个物体应该具有相似的运动信息的观点,完成帧间运动目标框链的构建。最后,引入动态置信度平均方法,来对预测结果进行增强。实验结果表明,以YOLOv5为基础,本算法达到了73.4%的平均精度均值(mean average precision, mAP),获得了6.2%(67.2%~73.4%)的mAP提升,同时检测速度达到了41每秒处理帧数(frames per second, fps)。本算法很好地兼顾速度和准确度,并为如何进行更加快速准确的视频目标检测算法提供思路。

关键词 视频目标检测; 检测框; 运动信息; 基于距离的交并比

中图分类号 TP391 DOI号 10.19363/J.cnki.cn10-1380/tn.2026.05.16

A Motion-based Seq-bbox Matching Method for Video Object Detection

LI Min^{1,2}, LI Linghan^{1,2}, BAI Ruwen^{1,2}, JIANG Miao^{1,2}, REN Junxing^{1,2}, MENG Bo³,
YANG Yang^{1,2}, HUANG Zihao⁴, HUANG Weiqing^{1,2}

¹ Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China

² School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100049, China

³ Beijing Institute of Technology, Beijing 100081, China

⁴ North China Electric Power University, Beijing 102206, China

Abstract Object detection is a fundamental task in the field of computer vision and is widely applied in video surveillance, autonomous driving, robotics and other fields. In recent years, with the development of deep convolutional neural networks, single-frame image-based object detection algorithms, such as YOLO and Faster R-CNN, have achieved remarkable results. However, in practice, single-frame image-based object detection algorithms cannot be directly applied to video object detection tasks due to the problems of blurring, occlusion, and deformation of the target objects in the video. The current mainstream video object detection algorithms, such as FGFA, Seq-NMS, etc., are unable to simultaneously take into account the detection speed and accuracy, thus cannot meet the needs of practical applications. How to improve the accuracy of detection under the premise of guaranteeing the real-time detection is the challenge faced by the current video object detection algorithms. In this paper, we propose a post-processing method based on inter-frame motion object frame chain. Based on the single-frame object detection algorithm, the video object detection results are enhanced by introducing inter-frame motion information from the post-processing part. We utilize Distance Intersection over Union (DIoU) to represent the inter-frame motion information, and propose the idea that the same object between adjacent frames should have similar motion information. We then combine the dynamic confidence averaging method to jointly complete the enhancement of the prediction results. Finally, the dynamic confidence averaging method is introduced to enhance the prediction results. The experimental results show that, based on YOLOv5, the proposed algorithm achieves a mean aver-

通讯作者: 孟博, 博士研究生, Email: bit.meng@live.com。

本课题得到国家重点研发计划(No. 2019YFB1005205)资助。

收稿日期: 2021-03-09; 修改日期: 2021-04-26; 定稿日期: 2023-08-30

age precision (mAP) of 73.4% and obtains a mAP enhancement of 6.2% (67.2%-73.4%). At the same time, the detection speed reaches 41 frames processed per second (fps). The proposed video object detection algorithm achieves excellent results in terms of balancing speed and accuracy and provides ideas on developing faster and more accurate video object detection algorithms.

Key words video object detection; bounding box; motion information; Distance Intersection-over-Union

1 引言

目标检测是计算机视觉领域的一个基本任务,在视频监控、自动驾驶、机器人等领域有着广泛的应用。近些年,伴随着诸如 YOLO^[1], Faster R-CNN^[2]等基于深度卷积神经网络的检测算法的大规模使用,基于单帧图像的目标检测算法的准确性已经获得了显著的提升。然而,在实际应用环境中,由于视频中的目标物体存在模糊、遮挡和稀有姿势等问题,单帧目标检测算法在视频目标检测任务中,仍存在准确率和召回率低等问题^[3]。

目前主要有两种思路用于提升视频目标检测的性能。第一种是构建一个针对视频流输入的专用检测模型,例如文献[4-5]。该类算法主要利用相邻视频帧之间的关联信息,在检测阶段引入新的网络结构提取特征参与决策,从而提升当前帧的检测结果。这种类型的算法在准确率上超越了基于单帧图片的目标检测算法。但其缺点是额外引入大量计算量,并且部分算法需要通过对视频进行离线处理来提取前后帧信息。上述问题导致该类算法检测速度慢,无法满足实际应用中的实时性要求。

另一种思路是从后处理的角度入手,不引入新的网络结构和特征。该类算法利用现有基于单帧图片检测网络的检测结果目标检测框(bounding box, bbox),通过跨帧链接属于同一个物体的检测结果 bbox,使用这些 bbox 链接来对当前帧的检测结果进行增强^[6]。该类算法与基于单帧的目标检测算法相比,准确率有所提升,并且与第一种类型的算法相比计算开销小,可以满足实时性的要求。但该类算法仍然存在以下三个缺陷:

- 在进行 bbox 链接时仅利用了相邻帧之间的空间位置信息,没有考虑到物体的运动信息。
- 在判断两个 bbox 是否属于同一个物体时,现有后处理方法主要依赖分类置信度。
- 在进行分类置信度增强时,错分样本也被增强。

针对现有算法的问题,本文选择从计算速度更快的后处理算法入手,以速度更加出色的基于检测框回归的目标检测算法为基础,提出了一种利用帧间运动信息的后处理方法。本方法的核心思路是认为相邻帧之间属于同一个目标物体的 bbox 具有相似

的运动信息变化。因此,在进行跨帧间 bbox 链接时,要将与前帧运动信息变化最相似的 bbox 加入到检测框链中。由于基于距离信息的交并比(Distance Intersection over Union, DIoU)^[7]对空间信息出色的表达和衡量能力,本文引入 DIoU 表达两个 bbox 之间运动信息变化。在判断两个 bbox 是否属于同一物体时,本文提出一种综合考虑 bbox 的运动信息与分类置信度的匹配方法,将这两种信息结合共同完成 bbox 链接。在检测结果增强阶段,本文提出一种动态均值增强方法。该方法有效抑制错误分类样本,提高检测结果的准确率。本算法主要包括以下几个步骤,前帧运动信息获取,当前帧目标框链接,目标框重打分。

从实际应用角度出发,在保障精度的情况下,我们使用实时性更好的 YOLOv5^[8]算法作为基础检测网络,所有实验均在 ImageNet 视频目标检测数据集(the ImageNet object detection from video, ImageNet VID)^[9]上进行。实验结果证明了我们算法的有效性,在一块英伟达 TITAN RTX 显卡上,我们的实时在线算法可以达到 73.4% mAP 的准确率和 41fps 的处理速度。本文的主要贡献如下:

- 本文提出了一种速度和准确度兼顾的后处理方法。在基于单帧图片目标检测算法的后处理部分引入运动信息对检测结果进行修正,既保留了基于单帧图片目标检测算法的速度优势,又提升了检测的准确率。以 YOLOv5 为例,在 ImageNet VID 验证集上 mAP 获得了 6.2% 的提升(67.2%~73.4%),计算速度达到了 41fps。
- 本文使用 DIoU 表达运动信息,并提出同一物体在相邻帧之间应该具有相同运动信息的思想,提高了跨帧 bbox 匹配的准确性和检测的准确率。
- 在 bbox 分类置信度增强时,本文提出了动态均值增强的思想,减少了错分样本被增强的情况,提高了检测的准确率。

本文章节安排如下,第 2 节介绍了目标检测和视频目标检测领域中的代表性算法,以及 DIoU 相关的原理内容,第 3 节为本文方法内容介绍,第 4 节是实验设计和结果分析,第 5 节是总结。

2 相关工作

本节将介绍目标检测和视频目标检测的相关工作, 以及用来表达帧间 bbox 运动信息的 DIoU 的相关原理。

2.1 目标检测

目标检测是计算机视觉领域的一个基本任务, 随着深度神经网络在目标检测领域的广泛应用, 基于深度学习的目标检测算法与传统人工设计特征的目标检测算法相比, 在速度和准确性上都获得了极大的提升。

目标检测算法主要分为以下两种类型, 分别是基于检测框回归算法(one stage)和基于感兴趣区域算法(two stage)。根据参考文献[10], 前者与后者相比, 具有更快的检测速度, 而后者则在准确性上更具有优势。

基于感兴趣区域算法通常可以概括为以下两个步骤, 首先从输入图片中提取感兴趣区域(region of interest, ROI), 之后针对感兴趣区域进行检测框的分类和位置回归, 最终输出检测结果。代表性的算法有 Faster R-CNN^[2]。该算法使用区域提议网络(region proposal network, RPN)进行 ROI 区域的提取和权重共享, 与 Fast R-CNN^[11]使用选择搜索算法(selective search, SS)相比, 在速度和精度上都获得了显著的提升。Mask R-CNN^[12]在 Faster R-CNN 的检测网络基础上, 引入了图像分割任务的分支, 通过多任务联合训练, 使准确率获得进一步提升。

基于检测框回归算法舍弃了 ROI 步骤, 将目标检测看作一个 bbox 的回归问题, 直接在卷积网络提取到的特征图上完成目标框的分类和位置回归, 检测速度与基于感兴趣区域相比有明显提升, 但准确率有一定程度的下降。代表性算法有 YOLO^[1], 该算法将图片分成 $N \times N$ 的若干区域, 对每个区域分别预测 bbox 和分类置信度。SSD^[13]算法使用不同尺度的特征图和先验框来预测物体, 一定程度上解决了小目标检测准确率低下的问题。

2.2 视频目标检测

视频目标检测是目标检测任务中的一个延伸任务, 其核心目的在于利用视频帧之间的关联信息, 解决视频中由于存在目标物体运动模糊, 物体被遮挡、场景光线变化和特殊形变等干扰, 导致检测准确率下降的问题。

针对上述问题, 目前主要分为两种解决思路。第一种是构建针对视频流输入的专用视频目标检测网络模型。这种方式主要是在基于单帧图片的目标检

测算法的基础上, 引入额外的网络结构, 提取帧间关联信息和时间上下文内容。利用相邻帧之间信息处理得到的特征, 与基于单帧图片的目标检测算法获得的特征结合, 共同完成检测任务。以 FGFA^[4]为例, 该算法在当前帧图片上使用单帧检测算法提取特征, 同时引入光流网络, 计算前后帧之间的光流特征, 并将光流特征与单帧检测算法提取到的特征合并, 共同完成检测任务。在参考文献[14]中, 针对前后帧特征变化, 提出了 STMM 模块, 该模块结合前后帧的记忆特征, 利用记忆特征, 改善当前帧的检测效果。

另一种方式是基于后处理部分的改进, 该方法不对检测网络结构进行改动, 通过对检测网络输出的结果进行处理, 利用前后帧的检测结果, 对当前帧输出 bbox 的位置结果和分类置信度进行修正。具有代表性的算法是 Seq-NMS^[6], 该算法利用检测网络输出 bbox 的分类置信度, 在视频帧之间构建 bbox 链, 一条链上的 bbox 属于同一个物体, 通过整条链的分类结果, 对部分检测效果不好的视频帧进行修正。

2.3 基于距离的交并比

交并比(Intersection over Union, IoU)的作用是衡量两个 bbox 之间的重叠度关系, 不能很好地衡量两个 bbox 之间的空间位置关系和形状相似度。如图 1 所示, DIoU^[7]是 IoU 的一个改进版本。DIoU 公式如式(1)所示, 其在 IoU 的基础上, 添加了基于尺度信息的中心点偏移量。公式中, b_1, b_2 分别代表了检测框 B_1 和检测框 B_2 的中心点, ρ 代表的是计算两个检测框中心点间的欧氏距离。 c 代表的是能够同时包含预测框和真实框的最小闭包区域的对角线距离。该式包含了两个框之间的重叠率、空间位置关系和形状信息。在进行训练时, 使用 DIoU 替代 IoU, 可以使目标框的回归更加稳定。同时, DIoU 也被应用在非极大值抑制中, 有效提升检测结果的准确性。

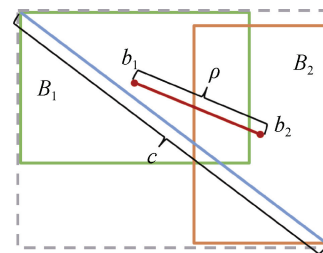


图 1 DIoU 示意图

Figure 1 Schematic diagram of DIoU

$$\text{DIOU}(B_1, B_2) = \text{IOU}(B_1, B_2) - \frac{\rho^2(b_1, b_2)}{c^2} \quad (1)$$

3 算法内容

本节将介绍本文算法的整体框架, 并分别详细讲述本文算法的三个步骤。

3.1 算法结构概述

本算法针对检测网络中的后处理部分进行改进, 属于一种实时在线目标检测算法, 可以被应用到任何现有的检测网络(YOLO, Faster R-CNN)当中。本算法包括前帧运动信息获取、当前目标框链接和目标框重打分三个部分。

算法结构如图 2 所示。以对当前时刻 T 的检测结果进行增强为例, 本文算法应用于检测网络生成候选 bbox 之后。STEP A 为前帧运动信息获取, 我们认为如果连续两帧中检测出同一个物体, 那么就认

定该检测结果是正确的。使用 DIoU 对 Frame T-2 和 Frame T-1 帧中属于同一个目标的预测框进行匹配, 并记录 DIoU 为运动信息, 为下一步寻找当前框提供信息; STEP B 为当前帧目标框链接, 我们认为属于同一个物体的 bbox 在相邻帧之间应该具有相同的运动信息。我们利用第一步提供的运动信息并结合当前帧 bbox 的分类置信度, 在 Frame T 候选 bbox 中寻找符合条件的目标框, 形成目标框链; STEP C 为当前目标框打分, 我们根据目标框链上之前时刻的分类置信度, 对当前检测框进行置信度增强。然后将增强后的 bbox 与其他候选框一起进行非极大值抑制抑制算法, 将重打分后的 bbox 与其他候选 bbox 竞争, 去除置信度过低和重叠度过高的候选框, 生成最后的检测结果。

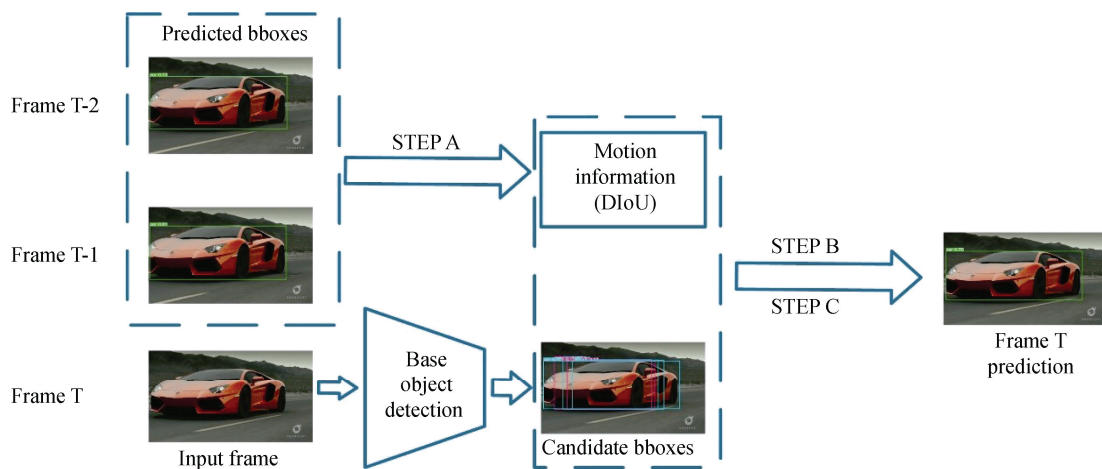


图 2 检测算法流程图

Figure 2 Flowchart of the detection algorithm

3.2 前帧运动信息获取

算法 1 前帧运动信息获取算法

输入: I bboxes detected in frame T-1 and J bboxes detected in frame T-2

DIOU_matrix = create_matrix(I, J)

- 1: FOR $i = 1$ to I do
- 2: FOR $j = 1$ to J do
- 3: set DIOU_matrix(i, j) to DIOU between i -th bbox of T-1 frame and j -th bbox of T-2 frame
- 4: END FOR
- 5: END FOR
- 6: set LAST_DIOU, LAST_BBOX to empty list
- 7: REPEAT
- 8: set I, j to line, column of maximum value in DIOU_matrix
- 9: add bbox i to LAST_BBOX
- 10: add DIOU (i, j) to LAST_DIOU

- 11: set i -th line of DIOU_matrix to 0
 - 12: set j -th column of DIOU_matrix to 0
 - 13: UNTIL maximum value of DIOU_matrix is 0
- 输出: LAST_DIOU, LAST_BBOX

对 T 时刻进行增强, 首先要获得 Frame T-1 和 T-2 之间的运动信息。我们使用 DIoU 对前两帧的预测框进行跨帧匹配。具有相同分类结果, 且 DIoU 值最大的两个框, 被认为是属于同一个物体, 记录这两个框, 并将 DIoU 作为运动信息记录。算法流程如算法 1 所示。

算法最后输出的 LAST_BBOX 和 LAST_DIOU 代表的是 Frame T-1 中与 Frame T-2 匹配成功的 bbox 和他们对应的运动信息 DIoU。

3.3 当前帧目标框链接

我们定义 S 来衡量 Frame T 候选框与 Frame T-1 检测框之间的相似程度, S 值越大代表两个 bbox 是同

一个物体的可能性越高。我们将 T 帧候选框中与 $T-1$ 帧检测框 S 值最大的框加入目标框链。

$$S(\text{bbox}_{T-1}, \text{bbox}_T) = \frac{\text{conf}_{\text{bbox}_T}^{\text{cls}=\text{bbox}_{T-1}}}{\text{motion}} \quad (2)$$

$$\text{motion} = |\text{DIOU}(\text{bbox}_{T-1}, \text{bbox}_T) - \text{LAST_DIOU}| \quad (3)$$

S 的公式如式(2)所示。分子部分的 $\text{conf}_{\text{bbox}_T}^{\text{cls}=\text{bbox}_{T-1}}$

表达了 bbox_T 在 bbox_{T-1} 的判别类别上的分类置信度, 分类效果越好, 分类置信度越大。分母部分的 motion 如式(3)所示, 表达了两个 bbox 之间的运动信息的相似程度。其中, DIOU 部分表达的是当前 Frame T 和 $T-1$ 之间的运动信息。 LAST_DIOU 表达的是上一步中获得的 Frame $T-1$ 和 $T-2$ 之间 bbox 的运动信息 DIOU 。两个 bbox 属于同一个物体的概率越高, 那么这个差值就会越小, 我们在这里加入绝对值表示关注的是两个数值之间的差值而不是大小关系。我们将运动信息相似度和分类置信度结合起来综合考虑, 对于一个 bbox 来说, 拥有越高的分类置信度和运动信息相似度都会使得 S 值变大, 我们的目的是寻找可以最好兼顾这两个数值的 bbox , 即使得 S 值最大的 bbox , 将其作为我们的目标框, 加入目标链中进行重打分, 算法流程如算法 2 所示。

算法 2 当前帧目标框链接算法

输入: I bboxes detected in LAST_BBOX,
LAST_DIOU and K bboxes detected in frame
 T . $S_matrix = \text{create_matrix}(I, K)$

- 1: **FOR** $i = 1$ to I do
- 2: **FOR** $k = 1$ to K do
- 3: set $S_matrix(i, k)$ to S between i -th bbox of
LAST_BBOX and k -th bbox of T frame
- 4: **END FOR**
- 5: **END FOR**
- 6: set LINKS to empty list
- 7: **REPEAT**
- 8: set i, k to line, column of maximum value in
 S_matrix
- 9: add (i, k) to LINKS
- 10: set i -th line of S_matrix to 0
- 11: set k -th column of S_matrix to 0
- 12: **UNTIL** maximum value of S_matrix is 0

输出: LINKS

算法最后输出的 LINKS 代表的是在 Frame T 中与 LAST_BBOX 中匹配成功的 bbox , 将匹配成功的 bbox 加入目标框链, 准备进行下一步目标框置信度重打分。

3.4 目标框重打分

我们通过对同一个目标框链上的分类置信度取平均值, 对当前框进行重打分。

$$\text{conf_re}_T^{\text{cls}=A} = \frac{\text{Avg}_{T-1}^{\text{cls}=A} * \text{Linknum}_{T-1} + \text{conf}_T^{\text{cls}=A}}{\text{Linknum}_{T-1} + 1} \quad (4)$$

$$\text{conf_re}_T^{\text{cls} \neq A} = 0 \quad (5)$$

如式(4-5)所示, 为了满足实时性的要求, 我们引入了一个动态的计数变量 Linknum_T , 该变量表示每条目标链在 T 时刻的现有长度。 $\text{Avg}_{T-1}^{\text{cls}=A}$ 表示的是 T 时刻之前的对判别类别 A 的分类置信度的平均值。利用均值和 Linknum 将当前 bbox 对 A 类别的分类置信度进行重打分, 同时将 bbox 对其他类别的分类置信度进行抑制。通过这种方法, 我们可以动态地使用整条链上的分类置信度平均值, 对当前帧的检测结果进行重打分, 利用 T 时刻之前的结果对当前时刻 bbox 的检测置信度进行修正。

将进行置信度增强后的候选框与其余候选框一起进行非极大值抑制算法(non-maximum suppression, NMS), 让经过置信度增强后的候选框与其他检测框竞争, 输出最终结果。

4 实验设计和结果

本节首先介绍实验内容设计和训练过程, 之后将本文的算法与现有的算法从精度和速度两个方面进行性能比较。最后针对本文算法中的一些模块进行消融实验, 通过控制变量来论证该算法模块的有效性。

4.1 实验设置

4.1.1 数据集介绍

本文中的实验全部在 ImageNet VID 数据集^[9]上进行, 该数据集是一个视频数据集, 包括 3862 段训练视频, 555 段验证视频和 937 段测试视频, 总图片数量超过 100 万张。该数据集一共包括 30 个类别, 是 ImageNet 数据集 200 类中的一个子类, 每一张图片都被使用 bbox 进行标注。该数据集面临的挑战包括相机虚化、部分遮挡、运动模糊、拥挤的实例、背景混乱、罕见的姿势等。由于 ImageNet VID 数据集的测试集标注没有公开, 我们在其验证集上进行性能评价。

4.1.2 训练数据筛选

虽然 ImageNet VID 数据集有 100 万张以上的总数, 但由于大量图片为相邻视频帧, 因此图片之间存在大量冗余信息, 数据缺少多样性, 根据参考文献[15], 我们从 ImageNet VID 数据集中挑选出相应类别的 10 万张图片, 从 ImageNet DET 数据集中选出对应类别的 8 万张图片, 共 18 万张图片构成训练集。

4.1.3 评价指标

本文使用 mAP 来衡量算法在 ImageNet VID 数据集上的准确性, 使用 fps 来衡量算法的计算速度。我们将与其他代表性的视频目标检测算法进行比较, 从准确性和速度两个方面综合评价我们算法的性能。同时我们将按照 FGPA^[4]论文中的分类方法, 将目标分为 fast、medium 和 slow 三个部分。将每个部分的 mAP 展示出来进行更深入的研究分析。

4.1.4 基础检测网络

基于对实时性的考虑, 我们选择了基于检测框回归算法中的 YOLO 算法, 作为我们的基础检测网络, 本文使用的是最新的 YOLOv5^[8]版本, 该版本与 YOLOv4^[16]相比, 虽然性能略有下降, 但在灵活性和速度上获得了很大提升, 更贴近实际应用需求。

4.1.5 训练细节

我们在训练 YOLOv5 时, 加载了该模型在 MS COCO 数据集^[17]上的预训练模型, 学习率为 0.0001, 批处理大小(batch size)设置为 32, 经过 60 个数据集迭代周期(epoch)后, 损失函数收敛。我们在两块 NVIDIA TITAN RTX 显卡上完成所有训练内容。

YOLOv5 可以适应多尺度输入, 在网络输入图像尺寸为 640×640 上可以达到 67.2% mAP, 在输入图像尺寸 608×608 上可以达到 66.5% mAP。

4.2 结果分析

表 1 展示了本文方法与两个著名的后处理算法 Seq-NMS^[6]和 Seq-Bbox-Matching^[18], 在 ImageNet VID 数据集上的性能对比。我们使用 YOLOv5 作为基础检测网络, 并在检测网络后端添加后处理算法。在进行推理时, 我们选择 0.55 作为非极大值抑制的阈值, 输入图片尺寸为 640×640 。

由于本文是针对 bbox 的分类置信度进行增强, 因此, 在进行推理时, 分类置信度的阈值选取将直接影响算法性能, 分类置信度低于阈值的 bbox 将被归类为背景。如图 3 所示, 该图展示了本算法检测性能 mAP 随着分类置信度阈值的变化曲线图。从图中我们可以得出, 当阈值小于 0.3 时 mAP 随阈值上升而上升, 这表示阈值的提高去除了误检的 bbox, 因此检测结果获得了提升, 而当阈值大于 0.3 时, 检测结果 mAP 下降, 漏检率提高, 因此本文选取 0.3 作为 bbox 分类置信度阈值。

表 1 与其他算法性能对比

Table 1 Comparison with other Methods

方法	mAP	mAP (fast)	mAP (medium)	mAP (slow)	检测时间+后处理时间(ms)
基础检测器(YOLOv5)	67.2	44.7	65.1	76.4	17.6
YOLOv5+本文方法	73.4	55.1	72.7	80.9	24.2(17.6+6.6)
Seq-NMS	69.7	49.2	68.5	78.3	22.1(17.6+4.5)
Seq-Bbox-Matching	72.1	52.9	71.3	79.4	22.6(17.6+5.1)

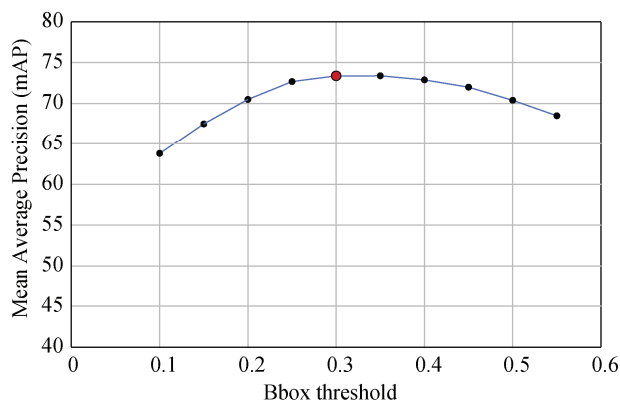


图 3 mAP 与阈值的关系

Figure 3 Relationship between mAP and threshold

对于 Seq-NMS, 按照论文中的结果我们选择筛选的 IoU 阈值为 0.5, 使用效果最好的均值增强方式, 同时为了保证实时性, 只使用前帧的信息进行置信度增强。对于 Seq-Bbox-Matching, 我们选择论文中的在线检测(online)方法进行复现并参与比较。

通过表 1 我们可以看出, 后处理方法可以在略微增加计算时间的情况下, 有效地提升单帧目标检测算法在视频目标检测任务中的性能。本文方法与 Seq-NMS 和 Seq-Bbox-Matching 相比, 由于在进行 bbox 链接时, 需要多处理一帧的信息, 因此计算时间增加了 2.1ms 和 1.5ms, 但由于引入了额外的运动信息, 我们的算法在基础检测网络的基础上获得了 6.2% 的 mAP 提升, 比 Seq-NMS 和 Seq-Bbox-Matching 分别高出了 2.7% 和 1.3%。

由表 1 可知, 与基础算法相比, 我们的算法在 fast、medium 和 slow 这三个分组上 mAP 分别增加了 10.4%、7.6% 和 4.5%。可以看出, 由于我们在后处理部分引入了运动信息, 并认为同一物体在相邻帧之间应具有相似的运动信息的思路, 更有助于检测快速运动的物体。

表 2 展示了本文提出的后处理算法, 分别在以 YOLOv5 为代表的 one stage 检测网络和以 R-FCN^[19]和 Faster R-CNN 为代表的 two stage 检测网络上的性

表 2 one stage 与 two stage 算法性能对比
Table 2 Performance comparison of one stage and two stage algorithms

方法	mAP	mAP (fast)	mAP (medium)	mAP (slow)	检测时间+后处理时间(ms)
基础检测器(YOLOv5)	67.2	44.7	65.1	76.4	17.6
YOLOv5+本文方法	73.4	55.1	72.7	80.9	24.2(17.6+6.6)
基础检测器(R-FCN)	72.9	50.9	71.1	82.3	85.7
R-FCN+本文方法	77.3	57.8	76.4	85.6	93.9(85.7+8.2)
基础检测器(Faster R-CNN)	73.6	51.6	71.8	82.6	172
Faster R-CNN+本文方法	77.8	58.1	76.7	85.7	187.1(172+15.1)

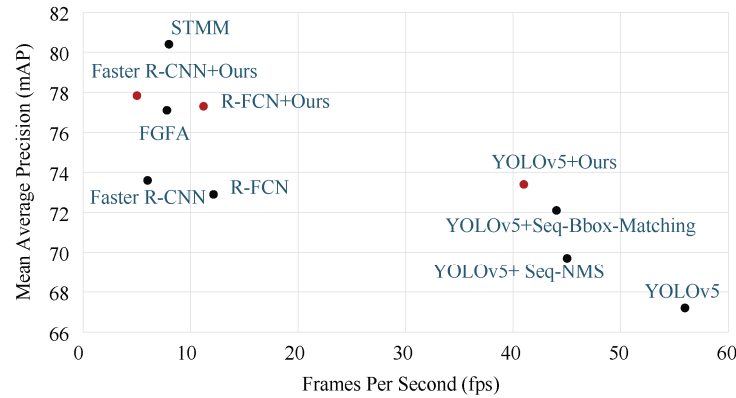


图 4 速度/准确度对比图
Figure 4 Speed/accuracy comparison chart

能比较。从表 2 中我们可以看出, 本算法在基础网络为 R-FCN 时, 获得了 4.4% 的 mAP 提升, 在 fast、medium 和 slow 这三个分组上 mAP 分别增加了 6.9%、5.3% 和 3.3%。本算法在基础网络为 Faster R-CNN 时, 获得了 4.2% 的 mAP 提升, 在 fast、medium 和 slow 这三个分组上 mAP 分别增加了 6.5%、4.9% 和 3.1%。与 one stage 检测网络相比, 由于基础检测网络性能上的差异, 基于 two stage 检测网络的算法 mAP 值高于基于 one stage 的算法。但是在计算速度上, 由于基础检测网络的区别, 总时间存在比较大的差距。从表 2 中我们还可以发现, 本文的后处理算法对于 one stage 算法的性能提升大于 two stage 算法。其原因在于 two stage 算法与 one stage 算法相比, 基础检测准确率更高, 而本算法的重点在于减少漏检率, 提高召回率从而提高 mAP 值, 对于检测性能更好的基础网络, 漏检率更低, 因此本算法的性能提升效果不明显。

图 4 展示了本文方法和主流的视频目标检测算法, 从速度与准确度这两方面进行的性能比较。我们一般认为计算速度大于 20fps 为满足实时性的要求。由图 4 所示, 基于后处理的算法与基于专用模型的算法相比, 虽然在准确率上存在一定的差距, 但在计算速度上具有明显的优势, 更加符合实际应用需

求。本文提出的算法, 在保留了基于后处理算法速度优势的前提下, 提升了检测的准确率, 获得了兼顾计算速度和准确率的性能。

4.3 消融实验

本节中, 我们将使用控制变量法, 来探究本文算法中提出的使用 DIoU 来表达运动信息和使用动态均值增强来抑制错分样本, 这两种方法的有效性。

4.3.1 DIoU 的有效性

我们将算法内, 前帧运动信息获取和当前帧目标框链接这两个步骤中, 用来表示运动信息的 DIoU 替换为原始的 IoU, 再使用 ImageNet VID 数据集的验证集进行结果评估。

由表 3 可知, 替换为 IoU 后, mAP 下降了 1.3%。在 fast、medium 和 slow 这三个分组上, mAP 分别下降了 2.3%、1.3% 和 0.8%。由此可见, 将算法中 DIoU 替换为 IoU, 会使得检测准确率下降。其中, 对快速运动的物体检测影响较大, 对缓慢运动的物体影响较小。这是由于 DIoU 与 IoU 相比, 增加了基于尺度信息的中心点偏移量, DIoU 可以更加准确地表达两个 bbox 之间的空间位置和几何形状变化, 有助于本算法中对于相邻帧的 bbox 之间运动信息的衡量和比较。

表 3 消融实验结果

Table 3 Ablation study result

方法	mAP	mAP (fast)	mAP (medium)	mAP (slow)
本文方法	73.4	55.1	72.7	80.9
IoU 替换 DIoU	72.1	52.8	71.4	80.1
最大值替换动态均值	71.1	52.4	70.2	78.8

4.3.2 动态均值重打分的有效性

我们将本文算法第三步的动态均值重打分替换为最大值重打分。如式(6)-(7)所示, $\text{Max}_{T-1}^{\text{cls}=A}$ 表示 $T-1$ 时刻之前目标链上对类别 A 的分类置信度的最大值, 通过使用目标链上分类置信度的最大值来对当前 bbox 进行重打分。从表 3 可知, 与使用动态均值相比, 总 mAP 下降了 2.3%, 在 fast、medium 和 slow 这三个分组上分别下降了 2.8%、2.5% 和 2.1%。

$$\text{conf_re}_{T}^{\text{cls}=A} = \text{Max}_{T-1}^{\text{cls}=A} \quad (6)$$

$$\text{conf_re}_{T}^{\text{cls} \neq A} = 0 \quad (7)$$

通过对比动态均值重新打分的式(4)-(5)与最大值重新打分的式(6)-(7), 我们可以发现动态均值重新打分的结果 $\text{conf_re}_{T}^{\text{cls}=A}$ 会受到当前帧检测结果的影响, 并且一定程度上反映了目标框链上分类置信度的变化, 而最大值重打分则只与目标框链上的最大值有关。

对于正确分类的样本来说, 随着时间的增加, 其分类置信度一直处于一个较高的分数, 偶尔出现的由于分类置信度低导致的漏检情况也可以通过目标框链上的前帧结果进行增强, 这种情况下两者之间性能几乎没有区别。但对于连续出现的错分样本来说, 伴随时间的增加, 其分类置信度在不断衰减, 动态均值重打分会根据这种衰减趋势来进行打分, 减少错分样本的分数, 达到去除错分样本的目的。但是对于最大值重打分来说, 其只使用最大值的方式, 会导致错分样本一直存在, 难以修正, 影响检测的效果。

样本预测分布如图 5 所示, 图 5 中横坐标为样本真实类别, 纵坐标为样本预测类别, 方框颜色越深, 代表所占比例越高。图中对角线上的方框表示分类正确的样本比例, 最底下一行方框表示样本被分类为背景的比例, 也就是漏检率, 其他方框表示错分样本比例。图 5(a)表示使用均值重新打分的预测结果图, 图 5(b)表示使用最大值重新打分的预测结果图。通过对比图 5(a)和(b), 我们可以发现, 动态均值重新打分的分布图, 对角线上的方框数量更多, 颜色更深。同时, 对角线以外的方框数量更少, 颜色更浅, 并且最底下一行的方框颜色与最大值重打分相比, 并没有明显的变化, 这表明本文方法在不增加漏检的情况下, 成功减少了错分样本的数量, 提高了检测的准确率。

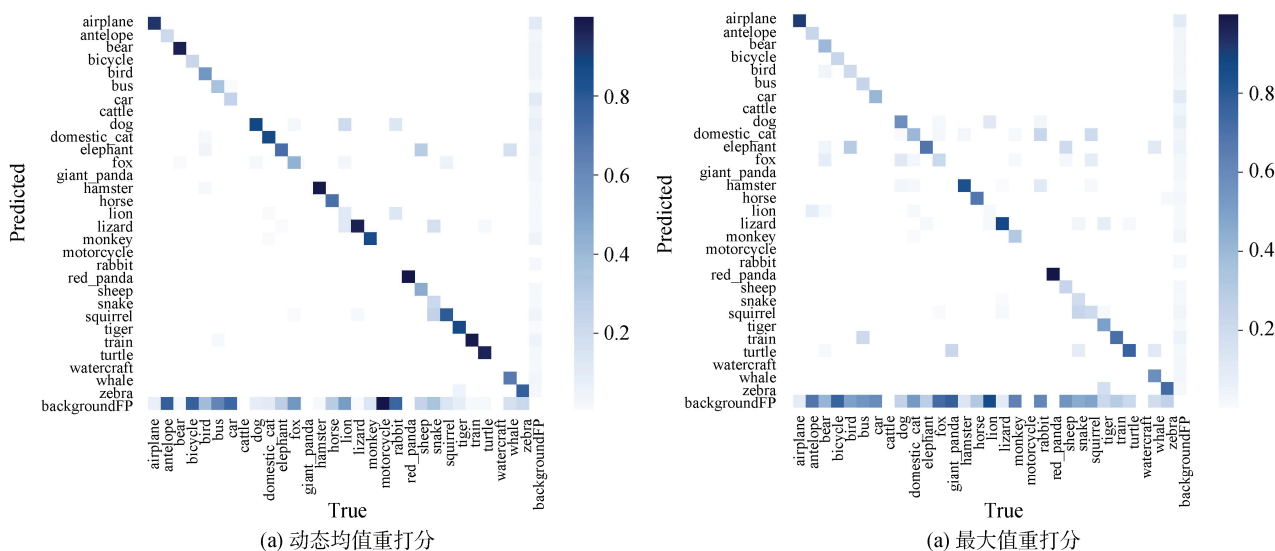


图 5 预测分类图

Figure 5 Prediction classification chart

4.4 可视化分析

本节中我们将使用 EPIC-KITCHENS-100 数据集来对我们的后处理算法进行可视化分析。通过实际的检测结果图, 来探究本文算法的性能。

EPIC-KITCHENS-100 数据集是一个从实际的厨

房工作中采集到的数据集, 该数据集从厨房工作人员视角, 详细地记录了厨房内部的实际工作状态。因为其对实际生活更加贴合的特性, 经常被用来衡量一个视频目标检测算法在现实环境中的应用性。

我们在这里使用 MS COCO 数据集预训练好的

YOLOv5x 模型作为基础检测算法, 并添加我们的后处理算法, 通过对比基础检测结果和添加后处理之后的结果, 来对我们的算法性能进行探究。

本文方法的核心思路, 利用前帧的检测结果和相邻帧之间的运动信息形成帧间目标框链, 来匹配当前帧中的目标框, 通过目标框链上的 bbox 分类置信度改善当前帧 bbox 的置信度(3.4 目标框重打分)。当前帧中 bbox 分类置信度的提升可以有效提高算法的召回率, 减少漏检率。同时, 对于一些分类置信度较低的错分样本, 也可以通过帧间目标框链上的分类置信度进行修正, 减少误检率。通过上述两点可以提高最终检测结果的 mAP。

检测结果如图 6 所示, 图 6(a1-a5)为对连续 5 帧图像使用基础检测网络的检测结果, 图 6(b1-b5)为对同样 5 帧图像添加本文后处理算法之后的检测结果。从图 6 中我们可以明显看出, 以 person 类对象为例, 添加后处理算法后的检测结果 bbox 的分类置

信度明显高于没有使用后处理算法的检测结果, 且一直保持在一个较高的置信度分数上。同时通过对比图 6(a2, b2)、(a3, b3)、(a4, b4)这三组图片我们可以发现, 以 bowl 类对象为例, 本文算法通过构成帧间目标框链, 传递分类置信度, 可以有效地减少相邻帧之间的漏检情况发生。

在本文的 3.2 和 3.3 节中, 我们的目标框链接方法的前提, 是如果连续两帧中检测出来同一个物体, 那么就认定该检测结果是正确的。该思路通常来说是正确的, 但是从图 6(a3, b3)、(a4, b4)、(a5, b5)这三组图片中我们可以发现, 以错分样本 spoon 类对象为例, 本文的后处理算法在面对连续出现的具有高分类置信度的错分样本时, 会将其当作正确分类的样本进行持续增强。因此本文使用动态均值重打分, 该方法与最大值重打分相比, 可以更加快速地削减错分样本的分类置信度, 以达到改善错分样本的目的。

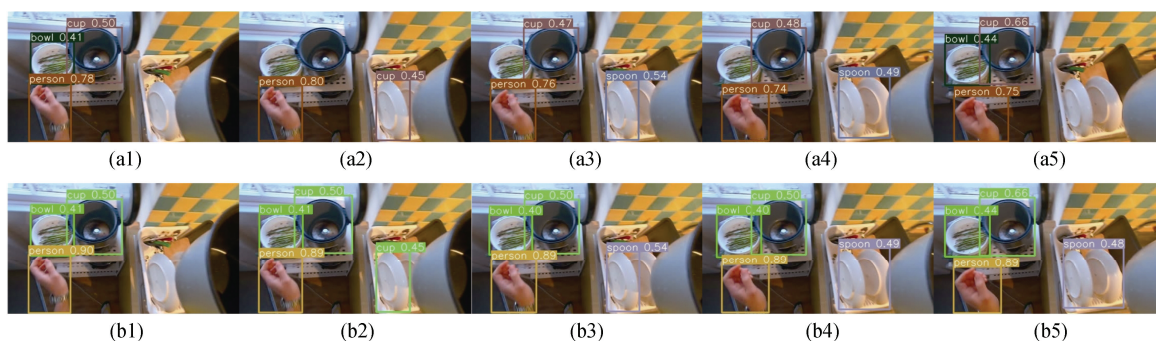


图 6 在 EPIC-KITCHENS-100 数据集上的检测结果对比图

Figure 6 Comparison chart of detection results on the EPIC-KITCHENS-100 dataset

5 总结

本文提出了一种新颖的、高速有效的后处理视频目标检测方法——基于帧间运动目标框链的后处理方法。该算法在后处理部分引入了运动信息, 利用 DIoU 对空间关系的优秀衡量性能, 表达相邻帧之间同目标物体 bbox 的运动信息, 并提出相同目标物体的 bbox 应具有相同的运动信息的思想。结合前帧运动信息和分类置信度共同完成 bbox 链接, 再使用动态均值重打分对当前帧的 bbox 的分类置信度进行修正。实验表明, 以 YOLOv5 为基础检测网络, 我们的后处理算法在保证检测速率的前提下, 与其他后处理算法相比获得了优秀的性能提升, 改善了基于单帧图像的目标检测算法在视频目标检测任务中的准确率。该算法同时兼顾速度与准确度, 可以被应用于实际生活中。

参考文献

- [1] Redmon J, Divvala S, Girshick R, et al. You only Look Once: Unified, Real-Time Object Detection[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016: 779-788.
- [2] Ren S Q, He K M, 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.
- [3] Sabater A, Montesano L, Murillo A C. Robust and Efficient Post-Processing for Video Object Detection[C]. 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2021: 10536-10542.
- [4] Zhu X Z, Wang Y J, Dai J F, et al. Flow-Guided Feature Aggregation for Video Object Detection[C]. 2017 IEEE International Conference on Computer Vision, 2017: 408-417.
- [5] Wu H P, Chen Y T, Wang N Y, et al. Sequence Level Semantics Aggregation for Video Object Detection[C]. 2019 IEEE/CVF International Conference on Computer Vision, 2020: 9216-9224.

- [6] Han W, Khorrani P, Le Paine T, et al. Seq-NMS for Video Object Detection[EB/OL]. 2016: arXiv: 1602.08465. <https://arxiv.org/abs/1602.08465>.
- [7] Zheng Z., Wang P., Liu W., et al. Distance-iou loss: Faster and better learning for bounding box regression[C]. *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence*, 2020: 12993-13000.
- [8] Jocher G., Stoken A., Borovec J., et al. ultralytics/yolov5: v3.1 <https://doi.org/10.5281/zenodo.4154370>. Oct. 2020.
- [9] 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.
- [10] Huang J, Rathod V, Sun C, et al. Speed/Accuracy Trade-Offs for Modern Convolutional Object Detectors[C]. *2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 3296-3297.
- [11] Girshick R. Fast R-CNN[C]. *2015 IEEE International Conference on Computer Vision*, 2016: 1440-1448.
- [12] He K M, Gkioxari G, Dollár P, et al. Mask R-CNN[C]. *2017 IEEE International Conference on Computer Vision*, 2017: 2980-2988.
- [13] Liu W., Anguelov D., Erhan D., et al. SSD: single shot multibox detector[C]. *14th European Conference in Computer Vision*, 2016: 21-37.
- [14] Xiao F. and Lee Y. J., Video object detection with an aligned spatialtemporal memory[C]. *15th European Conference in Computer Vision*, 2018: 494-510.
- [15] Chen K, Wang J Q, Yang S, et al. Optimizing Video Object Detection via a Scale-Time Lattice[C]. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 7814-7823.
- [16] Bochkovskiy A, Wang C Y, Liao H Y M. YOLOv4: Optimal Speed and Accuracy of Object Detection[EB/OL]. 2020: arXiv: 2004.10934. <https://arxiv.org/abs/2004.10934>.
- [17] Lin T., Maire M., Belongie S. J., et al. Microsoft COCO: common objects in context[C]. *13th European Conference in Computer Vision*, 2014: 740-755.
- [18] Belhassen H, Zhang H, Fresse V, et al. Improving Video Object Detection by Seq-Bbox Matching[C]. *The 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2019: 226-233.
- [19] Dai J F, Li Y, He K M, et al. R-FCN: Object Detection via Region-Based Fully Convolutional Networks[EB/OL]. 2016: arXiv: 1605.06409. <https://arxiv.org/abs/1605.06409>.



李敏 现任中国科学院信息工程研究所正研级高级工程师、中国科学院大学网络安全学院教授。研究领域为计算机视觉与智能分析、室内目标定位和物联网等。Email: limin@iie.ac.cn



李凌涵 于 2018 年在西安电子科技大学通信工程专业获得学士学位。现在中国科学院信息工程研究所攻读网络空间安全专业硕士学位。研究领域为基于深度学习的智能视频分析。Email: lilinghan@iie.ac.cn



白入文 于 2018 年在河北大学电子信息工程学院获得学士学位, 现在中国科学院信息工程研究所攻读博士学位。研究领域为计算机视觉、视频内容分析和人类活动识别。Email: bairuwen@iie.ac.cn



姜淼 于 2019 年在南京理工大学计算机科学与技术专业获得学士学位。现在中国科学院信息工程研究所攻读基于深度学习的图像处理技术专业硕士学位。研究领域为物体检测和动作识别。Email: jiangmiao@iie.ac.cn



任俊星 于 2014 年在北京航空航天大学通信工程专业获得硕士学位。现任中国科学院信息工程研究所中级工程师。研究领域为信息安全、电磁安全和深度学习。Email: renjunxing@iie.ac.c



孟博 于 2007 年在北京工业大学自动化工程专业获得硕士学位。目前在北京理工大学光学与光子学学院攻读博士学位。研究领域为计算机断层扫描、光谱 CT 成像和深度学习。Email: bit.meng@live.com



杨阳 于 2017 年在北京交通大学计算机科学与技术专业获得学士学位。目前在中国科学院信息工程研究所攻读基于深度学习的图像处理技术专业博士学位。研究领域为物体检测、动作识别和姿势估计。Email: yangyang1995@iie.ac.cn



黄子豪 于 2021 年在华北电力大学自动化专业获得学士学位。现在中国科学院大学计算机应用技术专业攻读硕士学位。研究领域为目标检测。研究兴趣包括图像超分辨率。Email: 120171070514@ncepu.edu.cn



黄伟庆 现任中国科学院信息工程研究所第四研究室主任、中国计算机学会信息保密专委会秘书长。研究领域为无线通信安全、电磁信号处理、网络安全保密技术、物联网安全、云计算安全技术等。Email: huangweiqing@iie.ac.cn