



融合注意力机制的半监督物体检测 在轨道交通上的应用

王俊宇¹, 陈哲^{2,3}, 孙俊勇²

(1. 国能包神铁路集团有限责任公司 机务分公司, 陕西 榆林 719000;

2. 中车株洲电力机车有限公司 重载快捷大功率电力机车全国重点实验室, 湖南 株洲 412001;

3. 西南交通大学 轨道交通运载系统全国重点实验室, 四川 成都 610031)

摘要: 轨道交通的运营和维护需要高精度的物体检测和跟踪技术, 以确保乘客安全和系统正常运行。近年来, 随着 Vision in Transformer 的提出, 融合注意力机制的大模型在物体检测领域取得了广泛的应用, 但是在实现高精度与高鲁棒性的模型训练的过程中, 注意力机制对数据的需求量是巨大的。在处理图像数据的过程中, 往往会伴随着大量人力、物力的消耗。为降低数据处理成本, 文章提出了一种融入了注意力机制的半监督物体检测策略, 以提高模型的鲁棒性。研究结果表明, 在只处理了 10% 数据的前提下, 采用 Grounding DINO 和 YOLO-World 等检测器作为算法的主干, 然后在算法的 head 层采用 CBAM、CoTAttention、SEAttention 等融合注意力机制, 在数据集上达到了 0.70 ± 0.04 的 mAP 精度, 相比于传统的半监督物体检测, 可以得到 5.14% 的 mAP 增益, 为后续轨道交通物体检测大模型的研究提供参考。

关键词: 半监督; 注意力机制; 物体检测; 轨道交通

中图分类号: U268.4; U298.1

文献标志码: A

文章编号: 1000-128X(2025)01-0035-07

0 引言

轨道交通自动驾驶系统是一个高度智能化的交通系统, 依赖各种传感器和计算机视觉技术, 以确保列车能够安全、高效运行。其中, 物体检测技术在轨道交通自动驾驶过程中具有极其重要的地位和作用。近年来, 随着边缘计算算力与算法的高速发展, 物体检测技术的精度与实时性也得到了大幅提高, 在列车高速行驶时, 物体检测技术可为车辆系统提供实时决策、车辆位置跟踪、能见度改善、自动停靠和站台对接等功能的提升。为了在实际复杂车况下完成上述任务, 多模态大模型成为当前的最优解。2023 年, 注意力机制与算法的结合使得多模态大模型逐渐走进人们的视野, 尤其是 ChatGPT4.0^[1] 的推出, 使得可以解决多种任务的大模型成为了当前的主流, 并且随着 Vision In Transformer^[2] 的提出, 注意力机制与视觉模型的结合成

为了目前计算机视觉领域的热点。百度的“Autonomous Driving Foundation Model”被称为是全球首个支持 L4 级自动驾驶业务的大模型, 是百度“萝卜快跑”出行服务平台的核心技术, 该技术的实现主要依靠一个视觉大模型和一个决策大模型来完成^[3]。但是, 要实现 L4 级自动驾驶, 对数据量的需求是巨大的, 其中百度“萝卜快跑”所使用的 ApolloScape 数据集包含 100 h 以上的立体声驾驶视频、144 000 张以上图像数据, 并且华为盘古 CV 大模型在研发初期使用了超过 100 TiB 的通用图像数据, 对这些数据的处理与使用消耗了大量的人力与物力^[4]。因此, 在模型训练时, 如何降低数据量和降低处理这些数据所需的人力物力成本, 并尽可能多地利用已有数据对大模型的训练、提升模型的鲁棒性有着重要的意义。

目前, 降低数据处理成本的方法有半监督与无监督两种方法^[5], 其中半监督学习位于传统监督学习和无监督学习之间, 是一种新型机器学习方法。半监督学习通过引入未标记样本来弥补标记样本不足的情况,

基金项目: 国家能源集团重大项目 (168204800040)

通信作者: 王俊宇, 男, 高级工程师, 研究方向为电力机车技术检修管理; E-mail: junyuwangok@163.com

引用格式: 王俊宇, 陈哲, 孙俊勇. 融合注意力机制的半监督物体检测在轨道交通上的应用[J]. 机车电传动, 2025(1): 35-41. DOI:10.13890/j.issn.1000-128X.2025.01.104.

Citation: WANG Junyu, CHEN Zhe, SUN Junyong. Application of semi-supervised object detection technology incorporating attention mechanism in rail transit[J]. Electric drive for locomotives, 2025(1): 35-41. DOI:10.13890/j.issn.1000-128X.2025.01.104.

避免在传统监督学习中由于数据不足而导致性能下降的问题。这一方法已经在多个领域取得了成功应用,可以通过应用平滑假设、聚类假设、流形假设来建立预测样本和学习目标之间的关系^[6]。基于上述假设,半监督目标检测可分为生成方法、基于图的方法、一致性正则化方法、伪标签方法、混合方法^[7]。

目前,基于伪标签的半监督物体检测策略是半监督领域的热点,其核心思想主要是通过使用预训练模型来估计未标记图像的伪标签,然后在增强后使用标记数据和未标记数据联合训练模型。因此,其检测器大多采用二阶段的形式,例如 Faster R-CNN^[8],并且在这些模型中,大多使用 1 个 Teacher 模型辅助生成伪标签。文献[9]基于此提出了一种简单、有效的半监督学习框架 STAC 和一种新的数据增强策略(增强一致性),用于更新模型,并在未标记图像中部署高度置信度的局部目标伪标签。文献[10]提出一种简单有效的 UnBiased Teacher 方法,使用类平衡损失来生成伪标签,改善了半监督目标检测中伪标签偏差的问题。但是,一直以来都没有一种端到端的半监督方案,直到文献[9]提出了 Soft-Teacher 算法^[11],才解决了这一问题,该算法为了充分利用教师模型的信息,无标记的边界框分类损失由无标记边界框产生的分类分数对教师网进行加权。此外,通过选择框回归方差小于阈值的候选框作为伪标签来更好地完成学习框回归。该算法可以在只处理 10% 数据的前提下,对所有的图像数据都加以训练和利用,大幅降低了处理数据所需的人工成本,降低了注意力机制使用在计算机视觉上的门槛。

传统的两阶段检测器(如 Faster R-CNN)尽管在半监督目标检测领域中有一定应用,但在当前实际工程应用中的效率较低,尤其是在边缘计算环境下,当计算资源有限时,这类模型的工程实用性较差。因此,本文选择使用更符合当下技术发展的先进一阶段检测器,如 Grounding DINO 和 YOLO-World,来替代原有的两阶段检测器。

Grounding DINO 具有强大的多模态对齐能力,能够结合多种数据源进行目标检测,极大提升了在复杂环境下的检测精度,非常适用于轨道交通等复杂场景的数据处理需求^[18],而 YOLO-World 则凭借其出色的实时检测性能和轻量级结构,特别适合在边缘设备上部署,能够以较少的计算资源完成高效的目标检测^[19]。

通过与传统两阶段检测器的对比试验,可发现一阶段检测器在推理速度、计算效率和精度上都表现出色,尤其是在资源受限的边缘设备部署中,这些检测器能够显著降低推理时间,同时保持较高的检测准确率。试验结果表明,这些改进有效解决了传统二阶段检测器在实际工程应用中的瓶颈,进一步增强了半监督目标检测的工程可行性。

基于上述研究,本文基于 Soft-Teacher 算法开展不

同注意力机制下物体检测效果试验,提出一种融合注意力机制的 Soft-Teacher 物体检测模型。相比于传统的半监督物体检测,该模型可大幅降低数据处理成本和提高模型鲁棒性,为后续轨道交通物体检测大模型的研究提供参考。

1 融合注意力机制的 Soft-Teacher 模型

本文提出的模型利用 Grounding DINO 和 YOLO-World 的特征提取网络对数据进行特征提取,并融合进入 RPN 网络,然后进行分类和回归生成候选框,在分类和回归之前,融入各种注意力机制模块,提升算法的泛化能力和鲁棒性,最终通过对比试验验证最具优越性的算法,模型总体流程如图 1 所示。

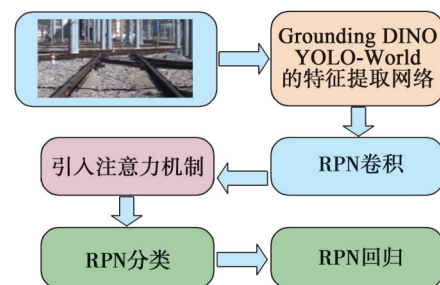


图 1 总体流程图

Fig. 1 General flow diagram

1.1 Soft-Teacher 半监督物体检测算法

在半监督学习中,数据被划分为有监督和无监督两部分。其中,有监督部分的损失记为 L_s ,无监督部分的损失记为 L_u 。在这两类损失中,还需进一步计算分类损失 L_{cls} 和回归损失 L_{reg} 。Soft-Teacher 算法的开发框架如图 2 所示,总体来说,对于有标签的数据,直接进入 Student 模型进行训练,得到损失 L_s ,Teacher 模型由前几次的 Student 模型加权计算得到。在半监督物体检测模型训练的过程中,数据增强是提高模型泛化性和鲁棒性的关键步骤,特别适用于单阶段目标检测。为了增强模型的鲁棒性,可以充分利用无标签数据的信息,并通过一致性正则化来限制扩增数据,以确保输出标签的一致性。然而,值得注意的是,由于不同数据增强方式的差异,数据增强的方法也存在很大差异,具体可分为强增强、弱增强与混合增强。在本文的开发框架中,对无标签的数据重新进行定义,分为强增强数据和弱增强数据,其中强增强数据进入 Student 模型进行检测,弱增强数据则进入 Teacher 模型生成伪标签。在数据增强之后,下一步是设计一个训练框架,用于集成来自标注和未标注图像的信息。对伪标签进行 Score Filter 筛选,选出可能性较大的候选框,并结合 Student 模型的检测结果,从而得到无标签数据的分类损失 L_{cls}^u 。不仅如此,模型还会对伪标签进行 Box-Jettering 再回归,得到其回归损失 L_{reg}^u 。因此,未标注数据损失 L_u 分为回归损失

L_u^{reg} 和分类损失 L_u^{cls} , 损失函数定义为:

$$L_u^{\text{reg}} = \frac{1}{N_b^{\text{fg}}} \sum_{i=1}^{N_b^{\text{fg}}} l_{\text{reg}}(b_i^{\text{fg}}, G_{\text{reg}}) \quad (1)$$

$$L_u = \frac{1}{N_u} \sum_{i=1}^{N_u} [L_u^{\text{cls}}(I_u^i, G_i^{\text{cls}}) + L_u^{\text{reg}}(I_u^i, G_i^{\text{reg}})] \quad (2)$$

式中: N_u 为无标签数据的总数; I_u^i 为第 i 张无标签图像; N_b^{fg} 为前景的候选框数量; b_i^{fg} 为预测为前景的候选框; $l_{\text{reg}}()$ 为当前图像的回归损失; 为了区分分类损失和回归损失, 使用 G_i^{cls} 和 G_i^{reg} 进行区别。

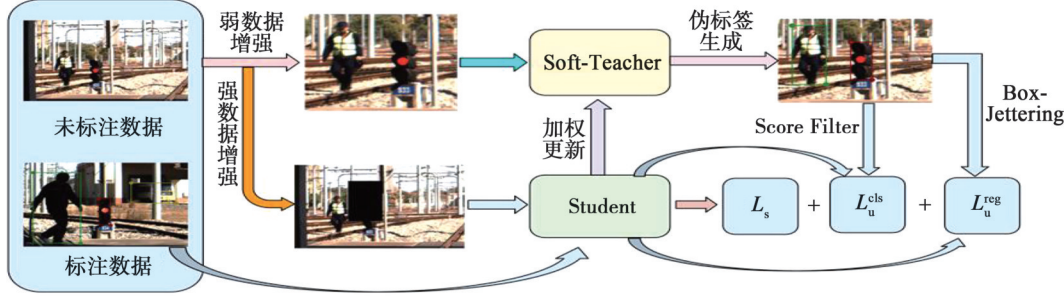


图2 Soft-Teacher训练框架结构图

Fig. 2 Structure diagram of the Soft-Teacher training framework

总体损失 L 为有标签损失和无标签损失的加权和, 具体为

$$L = L_s + \alpha L_u \quad (3)$$

式中: L_s 、 L_u 分别为根据训练图像数量不同进行归一化处理后的有监督损失和无监督损失; α 为无标签损失的权重。

归一化处理过后的损失分别为:

$$L_s = \frac{1}{N_l} \sum_{i=1}^{N_l} [L_{\text{cls}}(I_l^i) + L_{\text{reg}}(I_l^i)] \quad (4)$$

$$L_u = \frac{1}{N_u} \sum_{i=1}^{N_u} [L_{\text{cls}}(I_u^i) + L_{\text{reg}}(I_u^i)] \quad (5)$$

式中: N_l 为有标签数据的总数; N_u 为无标签数据的总数; I_l^i 为第 i 张有标签图像; I_u^i 为第 i 张无标签图像; $L_{\text{cls}}()$ 为分类损失; $L_{\text{reg}}()$ 为回归损失。

1.2 融合注意力机制的RPN模块

在Soft-Teacher原先的训练框架中, RPN模块是网络的核心组成部分, 负责生成潜在目标区域建议。该模块通常由卷积层组成, 用于对从前部得到的特征图提取特征, 并通过分类和回归头部产生目标建议。尽管传统RPN在很多情况下表现良好, 但在处理复杂背景、小目标和有遮挡等场景时, 仍然存在改进的潜力。为了进一步提高RPN模块的性能, 本研究引入了注意力机制, 具体包括通道注意力机制 (Channel-wise Attention Mechanisms, CAM) [12]、上下文注意力机制 (Contextual Transformer Attention, CoTAttention) [13] 和“挤压-激励”注意力机制 (Squeeze-and-Excitation Attention, SEAttention) [14], 用于改善区域候选网络 (Region Proposal Network, RPN) 的特征提取和建议生成过程。

卷积块注意力模块 (Convolutional Block Attention Module, CBAM) 是一种通道级别的注意力机制, 旨在通过对模型内不同通道的特征进行重点关注来提升模型性能。在具体操作中, CBAM首先对输入的特征

图进行自适应池化操作, 从而捕获各通道间的关键信息; 然后, 通过一个全连接层产生通道注意力权重, 这些权重用于调整每个通道在特征图中的贡献度。通过这种方式, CBAM帮助模型在处理复杂视觉任务时, 能更加精确地聚焦于信息丰富的区域, 显著提高了目标检测的精度和效率。

CoTAttention则采用了一种创新的自注意力机制, 不仅关注全局信息, 而且特别融入了局部上下文信息, 从而使模型能够更全面地理解图像内部的空间关系。这种结合局部和全局信息的方法提供了更深层次的图像解析能力, 特别是在处理场景中的交互复杂性和细节方面表现出色。

SEAttention则专注于通过增强特征图中各通道的表达能力, 以此提升模型的整体性能。它通过“挤压 (Squeeze)”操作来聚合全局信息, 随后再通过“激励 (Excitation)”步骤利用这些信息来加权各通道, 强化模型对关键特征的感知和响应。这种通道级重调权重的方法在许多现代视觉处理任务中, 特别是在强调特征区分度的应用场景中, 都显示出了非常好的效果。

本文通过整合CBAM、CoTAttention与SEAttention这3种有代表性的注意力机制对Soft-Teacher算法进行深入的改进。改进后的算法在RPN层中集成了这些注意力模块, 使得RPN不仅能生成高质量的目标建议, 而且能以更加高效的方式处理各种复杂的输入特征。整体算法结构的改进设计如图3所示, 这种融合注意力机制的方式, 可显著提升目标检测任务的性能, 尤其是在处理有遮挡和小目标检测方面的能力。

1.3 轨道交通系统大型数据集

目前, 虽然存在众多公开可用的数据集, 如COCO, 但是这些数据集主要是针对一般视觉对象的识别和分割, 并未专门针对轨道交通系统内部的物体和障碍物特性进行特殊设计。这类数据集通常缺少对轨道交通环境特有场景的覆盖, 也未能捕捉到系统随时间

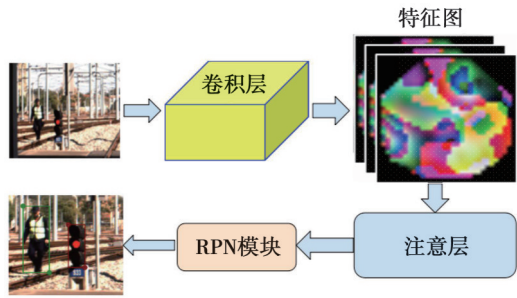


图3 融入注意力机制的RPN结构图
Fig. 3 RPN structure incorporating attention mechanisms

变化的动态特征，这些都是构建稳定、高效神经网络所需的关键因素。鉴于此，本文提出一个新型的、更为密集的、具有时间变化性的大规模数据集（见图4），旨在更全面地捕捉轨道交通系统的各种动态场景。



图4 轨道交通系统大规模数据集
Fig. 4 Large-scale datasets for rail transit systems

在此数据集中，本研究共收集了25 000张高质量的图像数据，涵盖了地铁系统全天的运行情况。这些图像详细记录了不同地铁路段的行车情况、轨道、车站等关键位置的信息，以及在各种天气和光照条件下的场景变化。为了支持半监督学习的需求，本研究对这些图像进行了严格的质量控制，对其中10%的图像进行了详细的人工标注，确保了数据的高准确性和一致性，使其能够有效地训练和验证半监督物体检测算法的性能。

数据集的多样化设置不仅使其成为测试半监督算法在真实条件下的准确性和鲁棒性的理想选择，而且与现有数据集相比，本数据集提供了更广泛的地理区域覆盖，更丰富的视角和时序变化，这些特点丰富了轨道交通物体检测的研究内容，为探索 and 开发新的算法、技术和应用程序提供了坚实的基础。同时，这些数据能够帮助研究人员深入理解轨道交通系统中物体和障碍物的动态特性，为今后轨道交通安全技术和智能监控系统的发展提供支持。

2 试验分析

2.1 实现细节

为了深入探索不同类型的注意力机制对Soft-Teacher

半监督物体检测算法的影响，本研究选取了3种具有代表性的注意力模块——CBAM、CoTAttention和SEAttention。其中，CBAM主要关注模型通道间的互动，强化关键特征的学习；CoTAttention融合了局部上下文信息，增强了对全局依赖的理解；SEAttention则通过重新调整通道权重，增强网络对有用特征的响应。本研究通过对3种机制的研究评估了他们在不同交并比（Intersection over Union, IoU）阈值下的性能，考察了其在平均精度均值（mAP）、召回率和混淆矩阵等方面的表现。

试验采用Grounding DINO和YOLO-World作为物体检测框架，基于Grounding DINO和YOLO-World的特征提取网络构建骨干网络，结合特征金字塔网络构建模型，旨在提高小目标的检测精度和模型的空间分辨率。所有试验均在MMDetection 3.00rc4框架下实施，以确保试验环境的一致性和高效性。同时，本研究采用了文献[11]的研究方法，采用相同的指数平均数（Exponential Moving Average, EMA）参数 α 和无监督损失权重 λ ，以保证不同方法之间的可比性。具体的试验参数如表1所示。

表1 模型超参数设置
Table 1 Model hyperparameter settings

超参数名称	数值或说明
批尺寸（Batch size）	5
时期（Epoch）	18 000
初始学习率	0.01
优化器	Adam
损失函数	Categorical Crossentropy
α 初始值	0.5

半监督学习的核心优势在于能够在标注数据稀缺的条件下充分利用未标注数据，从而提升学习模型的性能。因此，合理确定标注数据的比例是开展高效半监督学习试验的关键。本研究针对目标检测任务，将新提出的Soft-Teacher方法与当前领先的STAC^[17]方法进行详细比较。

首先，为了评估标注数据比例对模型性能的影响，本研究对标注比例分别为1%、5%和10%的标注数据集进行了全面的监督学习评估，并将结果与相应的无监督学习性能进行了对比，以此对比验证无监督学习在数据利用效率方面的潜在优势。随后，在相同的标注比例下，对Soft-Teacher（采用原始检测器或新型检测器）和STAC在半监督设置下的平均精度（mAP）进行比较。

在具体的模型实施方面，将CBAM、CoTAttention和SEAttention模块分别集成到RPN层的后续卷积操作中，这一策略可以对不同注意力机制进行直观比较，并探索最优的模块融合策略。此外，为了全面评估模型的性能和适应性，本研究对所有图像数据进行了180 000次迭代训练，使用3块RTX 4080 GPU来加速计算过程，

每块 GPU 同时处理 5 张图像。在训练集中, 已标注数据与未标注数据的比例为 1:4, 以强化半监督学习的学习效果。模型的初始学习率设置为 0.1, 并随迭代次数增加而逐步减小, 这有助于模型在训练过程中逐渐达到收敛, 从而不断提升性能。

2.2 试验结果

2 种检测模型在有监督、半监督学习条件下的 mAP 对比结果如表 2 所示。由表 2 可知:

- ①在完全监督和完全无监督的设置中, Soft-Teacher 与 STAC 的表现相似。在此基础上, 对 2 种方法进行系统级比较分析, 具体而言, 在 NR-L 数据集上, Soft-Teacher 方法在标注比例分别为 1%、5% 和 10% 时, 其性能分别比 STAC 方法约高出 6 个百分点、12 个百分点和 14 个百分点。
- ②Grounding DINO 和 YOLO-World 主干都是较新的模型, 比传统的二阶段模型在物体检测任务上往往有更好的表现。Grounding DINO 在检测任务中能够捕捉更

多的全局信息, YOLO-World 的轻量化设计则更加适合实时性要求高的应用。在有监督学习和小数据量 (1%, 5%, 10%) 的情况下, 2 种检测模型的表现均优于 Faster R-CNN。这是因为 2 种检测模型有更高效率的特征提取能力, 能从少量数据中提取更多有价值的信息。

由于伪标签机制在半监督学习下能够充分利用未标记的数据, Grounding DINO 和 YOLO-World 主干在无监督的情况下依然能提供良好的性能, 尤其是在 10% 数据时达到较高精度。

此外, 本研究还引入了几种新的注意力机制, 旨在进一步提升模型对少量标注数据的敏感性和利用效率, 希望通过这种方式, 更全面地评估半监督学习方法在实际应用中的有效性和适用性。

综上所述, 上述试验不仅证明了半监督学习方法在数据利用效率方面的优越性, 而且通过对比分析和系统评估, 展示了 Soft-Teacher 方法在具体试验设置下相比 STAC 方法的性能优势。这些研究可为后续的研究提供重要的试验依据和理论支持。

表2 Soft-Teacher 与 STAC 在有监督和半监督下 mAP 比较

Table 2 Comparison of mAP between Soft-Teacher and STAC models under supervised and simi-supervised conditions

方法		1%	5%	10%
有监督	STAC	0.200±0.007	0.380±0.002	0.470±0.005
	Soft-Teacher	0.210±0.003	0.400±0.005	0.490±0.006
	Soft-Teacher (Ground DINO 主干)	0.280±0.005	0.450±0.004	0.520±0.006
	Soft-Teacher (YOLO-World 主干)	0.270±0.004	0.440±0.003	0.510±0.005
半监督	STAC	0.300±0.003	0.430±0.004	0.510±0.006
	Soft-Teacher	0.360±0.002	0.550±0.002	0.650±0.004
	Soft-Teacher (Ground DINO 主干)	0.380±0.004	0.570±0.003	0.690±0.005
	Soft-Teacher (YOLO-World 主干)	0.370±0.003	0.590±0.002	0.680±0.004

为全面评估分别集成了 CBAM、CoTAttention 和 SEAttention 的 Soft-Teacher 算法相对于采用新型检测器的 Soft-Teacher 算法的性能提升, 本研究进行了一系列对比试验。这些试验主要在只使用 10% 标记数据的半监督学习环境下进行, 探索研究增强模型中的注意力机制对复杂场景目标的空间关系和关键特征识别能力的效果。

图 5 为部分检测结果。由图 5 可知, 相比于原始算

法, 融入了注意力机制的 Soft-Teacher 算法在处理被遮挡的对象和小尺寸目标时表现出更优的性能。这得益于注意力机制增强了模型对图像重要特征的聚焦能力, 尤其是在视觉信息部分缺失或复杂背景中, 能够更有效地识别和定位目标。

需要注意的是, 分别集成了 CBAM 和 SEAttention 新型检测器的 Soft-Teacher 算法 (记为 “CBAM-ST” 和 “SEAttention-ST”) 在 mAP 上的提升比使用 CoTAttention 的 Soft-Teacher 算法 (记为 “CoTAttention-ST”) 更为显著, 具体对比如表 3 所示。产生这一现象的原因在于 CBAM 和 SEAttention 的通道注意力机制优化了网络对不同通道特征的加权过程, 强化了模型对小目标和部分遮挡目标的检测能力。通道注意力机制通过重新分配特征图中通道的权重, 使得模型能够更聚焦于图像的关键信息, 从而提升了小目标的检测精度和模型的整体鲁棒性。

虽然 CoTAttention 的自注意力机制也提供了对图像序列中空间关系的深入理解, 促进了模型在全局上下文中的学习, 但其在小目标检测上的效果相对较弱。



图5 融合注意力机制后的部分检测结果

Fig. 5 Partial detection results of attention mechanism incorporated

自注意力机制可使模型能够基于序列中各位置之间的关系动态调整注意力，这对于理解目标之间的空间和语义关系非常有帮助，但不足以在所有情况下准确识别小尺寸目标。

表3 mAP精度比较

Table 3 Comparison of the mAP accuracy

模型	mAP
Soft-Teacher	0.690±0.005
CoTAttention-ST	0.680±0.030
CBAM-ST	0.710±0.030
SEAttention-ST	0.700±0.040

为了进一步分析各种注意力机制的效果，本研究展示了图6所示的混淆矩阵。由图6可知，通道注意力机制在特征提取上具有优势，尤其是对小目标（如标牌和信号灯）的检测能力显著优于自注意力机制。即使在复杂的遮挡情况下，通道注意力模型也能保持较高的检测性能，显示出优越的模型鲁棒性。

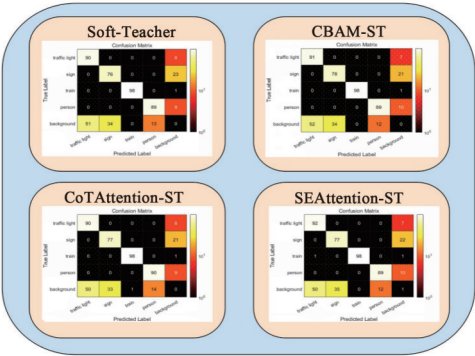


图6 Soft-Teacher检测混淆矩阵

Fig. 6 Confusion matrices of Soft-Teacher detection

综上所述，通过将注意力机制与新型检测器融入传统的半监督物体检测方法中，不仅能显著降低训练大型视觉模型的人工成本，而且为解决多模态视觉大模型中的多种问题提供了新的思路 and 策略。这些研究发现表明通过注意力机制的集成，可以进一步推动半监督学习在实践中的广泛采用，特别是在资源受限和数据标注成本高昂的场景中。

3 结论

本文研究的核心目标是探索和验证融合不同注意力机制的半监督物体检测技术在轨道交通系统中的应用潜力。

本文通过对 Soft-Teacher 和 STAC 两种半监督学习方法进行详细的比较分析，探讨了在不同标注数据比例下这些方法的表现及其在目标检测任务中的应用效果；通过在 1%、5%、10% 的标注比例下进行的半监督学习与有监督学习评估，验证了半监督学习在充分利用未标注数据方面的潜在优势，并且引入了新型一阶段检测器，以提升有监督与半监督下的检测性能。试

验结果表明，Soft-Teacher 与 STAC 在完全监督和无监督设置下展示了相似的性能，但在半监督学习的设置中，Soft-Teacher 方法在所有标注比例下均表现优于 STAC，在 NR-L 数据集上的性能提升分别为 6 个百分点、12 个百分点和 14 个百分点左右，并为后续试验确定了数据标注比例。为了提升 Soft-Teacher 算法中核心组件 RPN 模块的性能，引入了 3 种先进的注意力机制——CBAM、CoTAttention 和 SEAttention。这些机制增强了模型对关键特征的感知能力和处理复杂场景的能力，显著提高了 RPN 模块在特征提取和建议生成过程中的效率和准确性。

在经过一系列严格的试验后，其试验结果证明，集成注意力的机制在处理小型目标（如路标和信号灯）时显示出卓越的性能，能够在有遮挡情况下保持强大的鲁棒性。此外，通过优化特征的处理，这些机制显著降低了对大量标注数据的依赖，从而降低了训练大型视觉模型所需的人工成本和时间成本。此项技术的成功应用为解决轨道交通系统中多模态物体检测的难题提供了新的策略和方向。

在本文研究的基础上，未来的研究可进一步探索不同注意力机制的组合或优化，以充分利用各自的优势，进一步提升检测模型的整体性能、鲁棒性和泛化能力，尤其是在半监督物体检测模型的部署方面，未来将面临边缘计算设备的算力和存储限制问题。因此，研究如何在保持性能的同时减轻模型负担，实现模型的轻量化和高效运行，将是未来工作的重要方向。

此外，为了确保该技术的实际应用价值，将继续扩展该技术在更广泛轨道交通场景下的应用，如地铁、城市轻轨和长途铁路系统。通过在这些不同环境中进行实地测试和优化，全面评估融合注意力机制的半监督物体检测技术在实际应用中的表现和潜在的改进空间，进一步验证其实用性和有效性。这不仅将推动轨道交通安全技术的发展，而且可能为相关领域提供跨界的解决方案。

参考文献：

[1] BAHBINI A, KHAMOSHIFAR M, ABBASIMEHR H, et al. ChatGPT: applications, opportunities, and threats[C]// IEEE. 2023 Systems and Information Engineering Design Symposium. Charlottesville: IEEE, 2023: 274-279.

[2] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: transformers for image recognition at scale[EB/OL]. (2021-06-03) [2024-10-23]. <https://arxiv.org/abs/2010.11929>.

[3] GAO Haoxiang, WANG Zhongruo, LI Yaqian, et al. A survey for foundation models in autonomous driving[EB/OL]. (2024-09-05)[2024-10-23]. <https://arxiv.org/abs/2402.01105>.

[4] HUANG Xinyu, CHENG Xinjing, GENG Qichuan, et al. The apolloscape dataset for autonomous driving[C]//IEEE. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Salt Lake City: IEEE, 2018: 1067-10676.

- [5] 张震, 刘美含, 李朝, 等. 基于自监督的端到端图数据异常检测方法[J]. 中国科学(信息科学), 2023, 53(11): 2202-2213.
ZHANG Zhen, LIU Meihan, LI Zhao, et al. Self-supervised end-to-end graph level anomaly detection[J]. Science in China (information sciences), 2023, 53(11): 2202-2213.
- [6] 王健, 杨渡佳, 黄科举, 等. 认知电子战发展趋势: 从单体智能到群体智能[J]. 信息对抗技术, 2023, 2(4): 151-170.
WANG Jian, YANG Dujia, HUANG Keju, et al. Developing trend of cognitive electronic warfare: from single-agent intelligence to multi-agent intelligence[J]. Information countermeasures technology, 2023, 2(4): 151-170.
- [7] 陈理, 韩毅, 杨广, 等. 基于改进半监督模型的空间异质性农田特征提取研究[J]. 农业机械学报, 2023, 54(12): 173-185.
CHEN Li, HAN Yi, YANG Guang, et al. Spatially heterogeneous cropland characteristic extraction based on improved semi-supervised models[J]. Transactions of the Chinese society for agricultural machinery, 2023, 54(12): 173-185.
- [8] REN Shaoqing, HE Kaiming, 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.
- [9] SOHN K, ZHANG Zizhao, LI Chunliang, et al. A simple semi-supervised learning framework for object detection [EB/OL]. (2020-12-03) [2024-10-23]. <https://arxiv.org/abs/2005.04757>.
- [10] LIU Yencheng, MA C Y, HE Zijian, et al. Unbiased teacher for semi-supervised object detection[EB/OL]. (2021-02-18) [2024-10-23]. <https://arxiv.org/abs/2102.09480>.
- [11] XU Mengde, ZHANG Zheng, HU Han, et al. End-to-end semi-supervised object detection with soft teacher[C]// IEEE. 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021: 3040-3049.
- [12] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module[C]//Springer. Computer Vision - EC-CV 2018. Cham: Springer, 2018: 3-19.
- [13] LI Yehao, YAO Ting, PAN Yingwei, et al. Contextual transformer networks for visual recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2023, 45(2): 1489-1500.
- [14] HU Jie, SHEN Li, SUN Gang. Squeeze-and-excitation networks[C]//IEEE. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 7132-7141.
- [15] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//IEEE. 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 770-778.
- [16] CHEN Kai, WANG Jiaqi, PANG Jiangmiao, et al. MMDetection: open mmlab detection toolbox and benchmark[EB/OL]. (2019-06-17) [2024-10-23]. <https://arxiv.org/abs/1906.07155>.
- [17] SOHN K, ZHANG Zizhao, LI Chunliang, et al. A simple semi-supervised learning framework for object detection [EB/OL]. (2020-12-03) [2024-10-23]. <https://arxiv.org/abs/2005.04757>.
- [18] LIU Shilong, ZENG Zhaoyang, REN Tianhe, et al. Grounding dino: marrying dino with grounded pre-training for open-set object detection[EB/OL]. (2024-07-19) [2024-10-23]. <https://arxiv.org/abs/2303.05499>.
- [19] CHENG Tianheng, SONG Lin, GE Yixiao, et al. YOLO-world: real-time open-vocabulary object detection[C]// IEEE. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024: 16901-16911.

Application of semi-supervised object detection technology incorporating attention mechanism in rail transit

WANG Junyu¹, CHEN Zhe^{2,3}, SUN Junyong²

(1. Locomotive Branch, Baoshen Railway Group Co., Ltd., CHN ENERGY, Yulin, Shaanxi 719000, China; 2. State Key Laboratory of Heavy-duty and Express High-power Electric Locomotive, CRRC Zhuzhou Locomotive Co., Ltd., Zhuzhou, Hunan 412001, China; 3. State Key Laboratory of Rail Transit Vehicle System, Southwest Jiaotong University, Chengdu, Sichuan 610031, China)

Abstract: The operation and maintenance of rail transit systems require high-accuracy object detection and tracking technology to ensure the safety of passengers and the normal operation of traffic systems. In recent years, with the emergence of Vision in Transformer (ViT), large models incorporating attention mechanisms have garnered widespread application in the field of object detection. However, attention mechanisms require substantial demands of data in the process of model training to achieve high precision and robustness. Moreover, the process of image data processing is often accompanied by significant consumption of manpower and material resources. This paper proposes a semi-supervised object detection strategy that incorporates an attention mechanism, to reduce the cost of data processing and improve the robustness of the models. Detectors such as Grounding DINO and YOLO-World are employed as the backbone of the algorithm, and attention mechanisms such as CBAM, CoTAttention, and SEAttention are applied at the algorithm's HEAD layer. The study results show that, with only 10% of the data processed, a mean Average Precision (mAP) of 0.70±0.04 is achieved on the dataset, corresponding to a mAP gain of 5.14% compared to traditional semi-supervised object detection techniques. The study findings provide a theoretical reference for future research on object detection large models in the field of rail transit.

Keywords: semi-supervised; attention mechanism; object detection; rail transit