Computer Science and Application 计算机科学与应用, 2025, 15(1), 28-36
Published Online January 2025 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2025.151004

Hans汉斯

融合全局特征和局部特征的行人遮挡目标检测

徐　升，苏庆华*，万开政，戚翔宇，张智超

北京物资学院信息学院，北京

收稿日期：2024年12月18日；录用日期：2025年1月15日；发布日期：2025年1月23日

摘　要

在遮挡图像中，行人目标通常被其他物体部分或完全遮挡，导致其外观特征不完整、边缘模糊，甚至与背景或遮挡物混淆。行人遮挡目标的检测需要算法能够在特征缺失的情况下，仍然准确识别和定位目标。为了解决这一挑战，本文基于YOLOv10提出一种融合多尺度自注意力机制(Efficient Multi-directional Self-Attention, EMSA)的多尺度感知能力的YOLOv10改进方法。首先在YOLOv10中的C2f中融合MSDA注意力机制，增强了模型在多尺度上的特征捕捉能力，提升了对不同尺度遮挡目标的检测能力，通过自适应地加权不同通道的特征，提高了对遮挡目标特征的关注；其次基于动态聚焦机制引入新的损失函数Focaleriou，动态调整损失焦点，提高对不同尺度目标的检测能力，同时改善边界框回归损失收敛速度，之后添加了小目标检测头，增强小遮挡目标的特征提取能力；最后使用公开数据集Citypersons进行消融实验。结果表明，该融合了MSDA注意力机制的模型平均精度(Map@0.5)达到了62.3%，相较于官方YOLOv10n提升了2.2%。实验结果表明该EMSA注意力能够有效改进行人遮挡目标的检测，满足自动驾驶、监控等应用场景下的行人遮挡场景的检测需求。

关键词

遮挡目标检测，EMSA多尺度自注意力，YOLO，动态聚焦

Detecting Occluded Pedestrian Targets by Integrating Global and Local Features

Sheng Xu, Qinghua Su*, Kaizheng Wan, Xiangyu Qi, Zhichao Zhang

School of Information, Beijing Wuzi University, Beijing

Received: Dec. 18th, 2024; accepted: Jan. 15th, 2025; published: Jan. 23rd, 2025

*通讯作者。

文章引用: 徐升, 苏庆华, 万开政, 戚翔宇, 张智超. 融合全局特征和局部特征的行人遮挡目标检测[J]. 计算机科学与应用, 2025, 15(1): 28-36. DOI: 10.12677/csa.2025.151004

Abstract

In occluded images, pedestrian targets are often partially or completely blocked by other objects, leading to incomplete appearance features, blurred edges, and even confusion with the background or occluding objects. Detecting occluded pedestrian targets requires algorithms capable of accurately recognizing and localizing targets despite missing features. To address this challenge, this paper proposes an improved YOLOv10 method with enhanced multi-scale perception by integrating an Efficient Multi-directional Self-Attention (EMSA) mechanism. Firstly, the MSDA attention mechanism is incorporated into the C2f module of YOLOv10 to enhance the model's ability to capture features at multiple scales, improving the detection of occluded targets of various sizes. By adaptively weighting features across channels, the method increases focus on occluded target features. Secondly, a novel loss function, Focaleriou, is introduced based on a dynamic focusing mechanism. This adjusts the focus of the loss dynamically, enhancing the detection of targets at different scales and improving the convergence speed of bounding box regression loss. Additionally, a small-object detection head is added to strengthen feature extraction for small occluded targets. Finally, ablation experiments are conducted on the public Citypersons dataset. Results show that the model incorporating the MSDA attention mechanism achieves a mean average precision (mAP@0.5) of 62.3%, which is 2.2% higher than the official YOLOv10n. Experimental findings demonstrate that the EMSA attention mechanism effectively improves the detection of occluded pedestrian targets, meeting the requirements for scenarios such as autonomous driving and surveillance under occluded pedestrian conditions.

Keywords

Occluded Object Detection, EMSA Multi-Directional Self-Attention, YOLO, Dynamic Focusing

1. 引言

现实生活中的复杂场景往往伴随着遮挡现象，遮挡会导致目标在视觉上信息丢失、形态模糊或边界不完整，从而显著增加目标分类与定位的难度，尤其是在拥挤场景或复杂背景中，传统目标检测方法的性能往往大幅下降，在遮挡物与目标具有相似的外观或纹理特征时，模型难以有效区分目标与遮挡物，导致误检。例如，在交通场景中，行人与背景广告牌之间的相似性可能使检测混淆。传统计算机视觉算法在遮挡目标检测中面临挑战，由于传统目标检测方法通常依赖于局部特征的提取与分析，当目标的关键特征被其他物体部分覆盖或掩蔽时，模型难以准确区分目标间的边界和关系，从而导致对遮挡目标的定位和分类准确性降低。因此，确保目标检测算法能够准确地识别和定位目标在遮挡图像中位置具有重要意义。

在目标检测领域，YOLO (You Only Look Once)系列[1]算法凭借其端到端、实时检测的优势，广泛应用于自动驾驶、安防监控等场景。Bochkovskiy 等人[2]过增强特征提取网络的多尺度能力，在 YOLOv4 中引入 FPN (Feature Pyramid Network)和 PAN (Path Aggregation Network)，显著提升了模型对小目标及部分遮挡目标的检测能力。Chen 等人[3]针对遮挡问题，提出了动态检测头(Dynamic Head)通过自适应调整不同特征层的权重，在检测复杂场景中的遮挡目标时表现出更强的鲁棒性。Sun 等人[4]在 YOLO 框架中加

入深度图或红外图像信息，有助于分离目标与背景，从而提高对遮挡目标的检测精度。

2. 相关工作

2.1. 注意力机制

注意力机制(Attention Mechanism)是一种在处理信息时动态调整对不同部分的关注程度的技术，旨在提升模型在各种任务中的表现。它最初受到人类视觉和认知注意力过程的启发，已广泛应用于自然语言处理(NLP)、计算机视觉(CV)、语音识别等领域。主要目的是根据输入信息的重要性分配权重，优化信息的处理过程。其核心思想是：在给定一个输入序列时，模型能够自动决定哪些部分的信息对当前任务最为重要，从而对这些部分赋予更高的关注度。注意力机制在计算机视觉任务中已成为提升模型性能的关键技术，尤其在目标检测、语义分割等任务中展现出重要的应用价值。

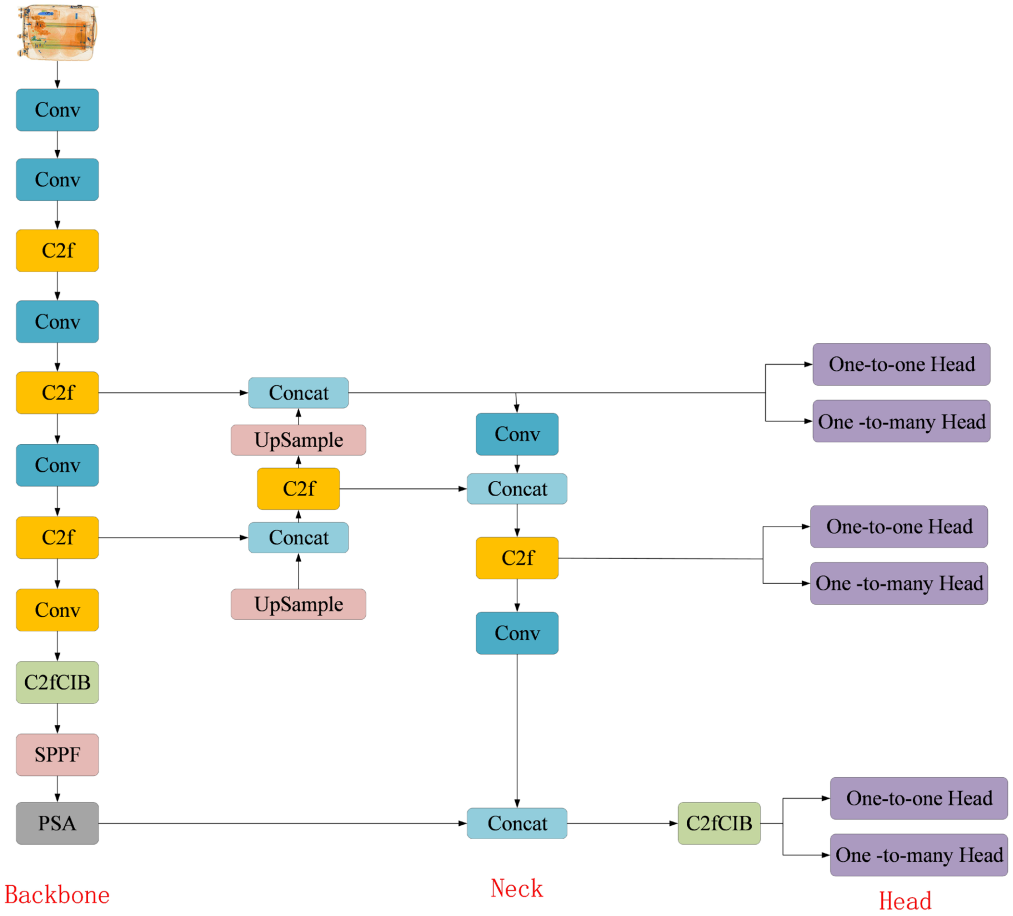


Figure 1. YOLOv10 network architecture diagram
图 1. YOLOv10 网络结构图

自注意力机制(Self-Attention)由 Vaswani 等人[5]提出，主要用于序列数据处理，如机器翻译等任务。该机制通过计算输入序列中每个元素与其他元素之间的相似度，来动态地为每个元素分配不同的注意力权重。自注意力机制不仅能够捕捉全局依赖关系，还避免了传统 RNN 和 CNN 在长序列建模中的限制。Woo 等人[6]提出的 CBAM 是一种将通道注意力和空间注意力相结合的模块。其创新之处在于通过两步注意力机制，分别增强了特征图的通道和空间维度的信息表达。通道注意力通过自适应调整每个通道的

权重，提升了特定通道特征的表达能力；而空间注意力则通过关注不同位置的特征，增强了空间局部信息的感知。Hu 等人[7]提出的 SENet 通过引入 Squeeze-and-Excitation 模块，对每个通道的特征进行自适应重标定。SENet 的创新之处在于利用全局信息来建模每个通道的重要性，并根据该重要性调整通道的响应。该方法能够提升网络的表达能力，尤其在图像分类和物体检测中表现优异。Wang 等人[8]提出的 Linformer 通过低秩近似来优化自注意力的计算过程。其创新之处在于将标准的全局自注意力机制改为稀疏注意力，利用低秩矩阵近似来减少计算和内存消耗。这一方法的作用是降低自注意力机制在处理长序列时的计算复杂度。

2.2. YOLOv10

YOLOv10 由 2024 年发布[9]，具体网络结构图如图 1 所示。YOLOv10 相较于 YOLOv8 没有较大的改变，在预测方面采用了无 NMS 的双重分配策略，包括标签分配和一致匹配度量，在保证一定精度的前提下大幅提升了网络的检测效率。

YOLOv10n 模型结构主要由输入端(Input)，主干网络(Backbone)，颈部网络(Neck)，检测头(Head)构成，结构图如图 1 所示。Input 负责将原始图像输入到模型中。Backbone 从输入图像中提取有用的特征，主要由 Conv 卷积层，C2f 残差块，SPPF 空间金字塔模块，SCDown 模块，自注意力模块 PSA 组成。Conv 卷积块主要由二维卷积层，Batchnorm2d 批归一化，SiLU 激活函数构成。SPPF 通过对输入特征图进行不同尺度的池化操作，生成固定长度的特征向量，从而使得网络能够处理任意尺寸的输入图像 C2f 主要由卷积层，BottleNeck 模块构成。SCDown 实现空间下采样(从 $H \times W$ 到 $H/2 \times W/2$)和通道变换(从 C 到 2C)，在减少计算量的同时最大限度地保留了下采样过程中的信息。Neck 部分采用了基于特征金字塔网络(FPN)的改进结构 Path Aggregation Network (PANet)，该结构采用了多尺度特征融合技术，将来自 Backbone 的不同阶段的特征图进行融合，增强了特征表示能力。Head 部分默认为 P3、P4、P5 三个检测头。

3. 遮挡目标算法

本文主要从融合注意力机制、添加检测层、改进损失函数几个方面对 YOLOv10n 进行改进。为了增强对行人遮挡目标的定位能力，基于图像特征设计多尺度自注意力机制(EMSA)，并将其融合进 YOLOv10n 的 Backbone 和 Neck 部分；为了提升对遮挡小目标行人的检测能力，在特征浅层添加小目标检测头；为了提高模型性能，使用 Focaleriou 损失函数替换 CIoU 作为算法的定位损失函数。改进后的 YOLOv10n-EMSA 网络结构如图 2 所示。

3.1. 多尺度自注意力机制融合

本文提出了一种多尺度自注意力机制 EMSA，结合了基于高效多尺度的全局注意力机制和基于多尺度扩张局部注意力的局部注意力机制。EMSA 的原理图如下图 3 所示。输入特征图经过 EMA 模块提取全局与局部信息：通道划分后的特征分别通过方向性自适应池化(水平和垂直方向)捕获全局空间信息，并通过 1×1 卷积进行信息融合与上采样恢复特征分辨率；同时，3×3 卷积用于强化局部特征，并结合组归一化(GN)优化特征分布。随后，通过自适应全局池化生成动态注意力权重，调整特征图中不同区域的重要性。与此同时，输入特征图还通过 MSDA 模块提取多尺度特征。MSDA 利用多个感受野大小的多头注意力机制，通过动态感受野权重捕获不同尺度下的特征相关性，并生成跨尺度的注意力特征。最终，EMA 模块和 MSDA 模块的输出特征在通道维度进行融合，以实现全局、局部以及多尺度信息的联合表达。融合后的特征图既保留了细粒度的空间信息，又增强了对复杂场景中多尺度目标的感知能力，为后续任务提供了更加丰富且高效的特征表示。

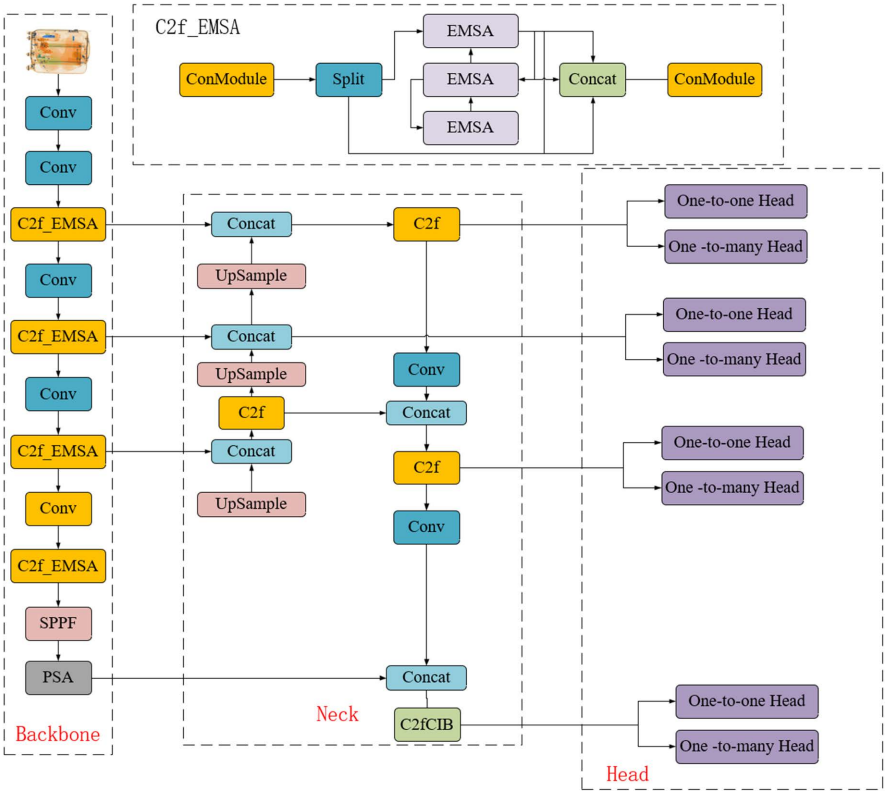


Figure 2. YOLOv10n_EMSA network architecture diagram
图 2. YOLOv10n_EMSA 网络结构图

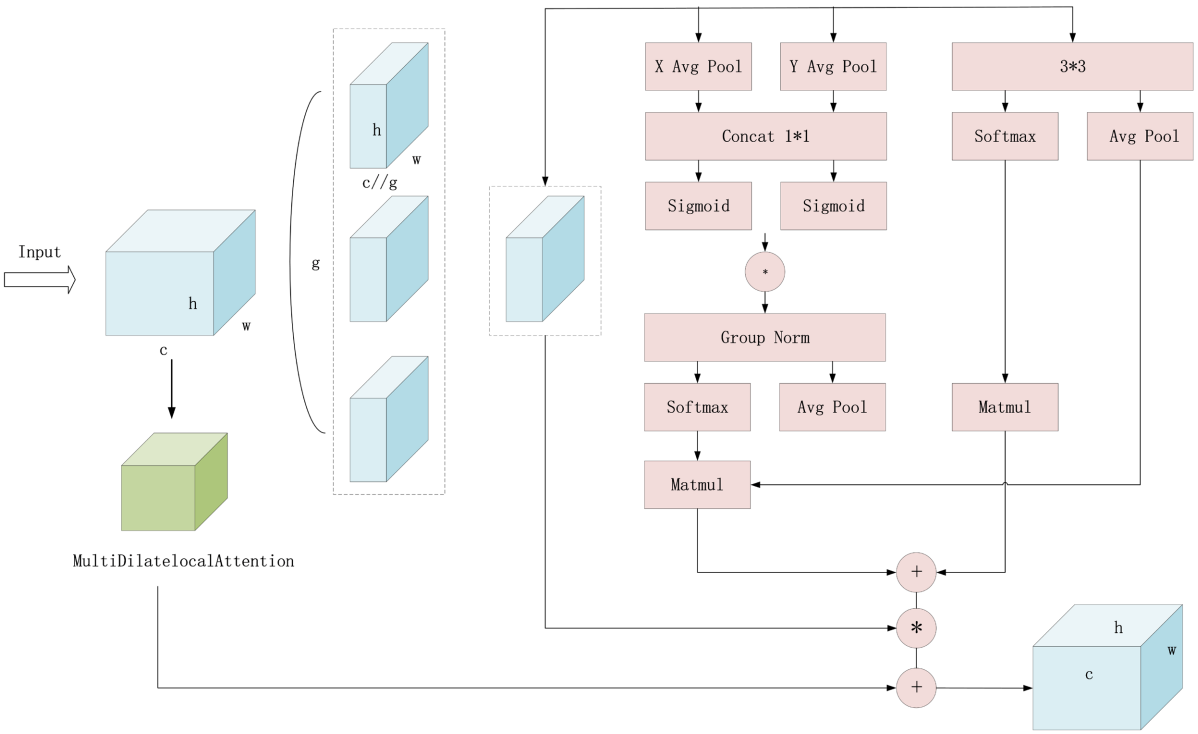


Figure 3. EMSA schematic diagram
图 3. EMSA 原理图

在该注意力结构中只有两个卷积核分别放置在并行子网络中，帮助网络避免更多的顺序处理和大深度。其中一个并行子网络是一个 1 × 1 卷积核，用于调整特征图的通道数，另一个是一个 3 × 3 卷积核，用于空间特征提取和特征细化，用于增强特征的表达能力。

多尺度自注意力(EMSA)的具体工作流程：通过自适应池化提取空间维度的特征→组归一化计算输入特征图的加权平均，增强特征的表达能力→1 × 1 卷积调整特征图的通道数，3 × 3 卷积核，用于空间特征提取和特征细化，增强特征表达→全局平均池化权重→Softmax 函数注意力权重，对输入特征图进行加权→计算加权平均，捕捉多尺度的局部特征→注意力特征融合。

C2f 模块本身具有跨阶段特征连接的特性，适合多尺度信息融合。而引入注意力机制可以自适应地调整不同尺度特征的重要性，使融合结果更加精准，因此将 EMSA 融入到 C2f 中，具体结构如下图 4 所示。

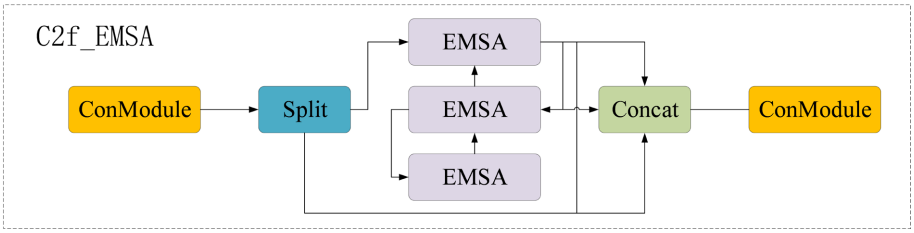


Figure 4. C2f_EMSA structure
图 4. C2f_EMSA 结构

3.2. 线性聚焦机制

在不同的检测任务中聚焦不同的回归样本，采用线性区间映射其定义为：

$$\mathrm{IoU}^{\mathrm{focaler}}=\begin{cases}0, & \mathrm{IoU}<d\\ \dfrac{\mathrm{IoU}-d}{u-d}, & d\ll\mathrm{IoU}\ll u\\ 1, & \mathrm{IoU}>u\end{cases} \tag{1}$$

其中， $\mathrm{IoU}^{\mathrm{focaler}}$ 是重构后的 Focaler-IoU 值，IoU 是原始的 IoU 值，且 $[d,u]\in[0,1]$。

当 IoU 低于阈值 d 时，$\mathrm{IoU}^{\mathrm{focaler}}$ 被设置为 0，可以忽略低 IOU 样本，避免模型浪费精力在极差的预测上。当 IOU 在 $d\ll\mathrm{IoU}\ll u$ 区间时，$\mathrm{IoU}^{\mathrm{focaler}}$ 被映射到 0 到 1 之间，使模型更关注 IoU 值在该范围内的样本，以此来重点关注特定范围的样本。当 IoU 大于 u 时，$\mathrm{IoU}^{\mathrm{focaler}}$ 为 1，表示该样本接近理想匹配。

为了使 $\mathrm{IoU}^{\mathrm{focaler}}$ 专注于不同的回归样本，可以通过调整 d 和 u 的取值范围。其损失函数定义如下：

$$L_{\text{Focaler-IoU}}=1-\mathrm{IoU}^{\mathrm{focaler}} \tag{2}$$

Focaler-IoU 损失通过引入线性区间映射机制，有效地增强了模型对不同 IoU 区间的回归样本的关注，使其可以在各种检测任务中实现更优的边界框回归效果。相比传统的 IoU 损失，它对难例和特定区间样本有更强的聚焦能力。

3.3. 小目标检测头

在特征融合的过程中，通常会使用多尺度特征图(如 P3、P4、P5)来进行目标检测。这些特征图分别对应不同分辨率的图像层，P3 层较浅，P5 层较深。P3、P4、P5 这三个层级的特征图虽然能够很好地处理小到大尺度的目标，但由于它们来自较深的网络层次，分辨率相对较低，对于微小目标来说，细节信息的保留不够充分，导致小目标难以检测，而安检 X 光图像中可能存在物体重影和遮挡，这会导致一些

原本就小的目标更难检测，通过添加一个小目标检测头 P2，可以有效提升模型的检测性能，尤其是针对小目标的检测能力。

4. 实验结构及分析

4.1. 数据集、实验环境和实验指标

实验数据集选用 Citypersons 公开数据集，广泛应用于行人检测任务的数据集，其图片中覆盖了多种城市环境，包括不同的天气条件、光照变化、以及动态的行人和车辆。本实验只选取了数据集中的行人标签进行训练。

实验硬件环境：CPU：AMD Ryzen 7 5800X 8-Core Processor，32GB 内存；GPU：Nvidia GeForce RTX 3060 12g 显存。实验软件环境：操作系统：ubuntu20.04，cuda 版本为 11.8，python 版本为 3.8，深度学习框架 PyTorch 版本为 2.0.1。实验模型：YOLOv10n。模型权重文件：YOLOv10n.pt。

超参数设置：批量大小 Batch Size 设为 16，训练轮次 Epochs 为 100 轮，学习率 0.01，输入图片尺寸为 640 × 640，优化器为 SGD，其余参数都为 YOLOv10 默认值。

评估指标：精准度(Precision, P)：预测为正样本的结果中，实际为正样本的比例；召回率(Recall, R)：实际正样本中被模型正确预测的比例；平均精度 Map@0.5：表示 IOU 为 0.5 时的平均精度。

4.2. 融合多尺度自注意实验结果

为了验证多尺度自注意力机制 EMSA 改进算法的效果，将融合了 EMSA 注意力机制的模型与原模型进行了对比实验，实验结果如下表 1 所示，在 Citypersons 数据集中该改进后的模型相较于原模型平均精度提升了 2.2%。其中 YOLOv10n 的检测结果如下图 5 左所示，改进后的模型检测结果如右所示，实验结果表明，该改进后的注意力机制的精度相较于原模型有一定提升。

Table 1. Performance comparison of multi-scale self-attention mechanisms
表 1. 多尺度自注意力机制性能对比表

model	P	R	mAP@0.5	mAP@0.5:0.95
YOLOv10n	0.944	0.79	60.1	37.2
YOLOv10n + EMSA	0.886	0.83	62.3	38.9

Figure 5. Comparison of results before and after adding the EMSA attention mechanis
图 5. 加入 EMSA 注意力机制前后效果图

4.3. 消融实验对比

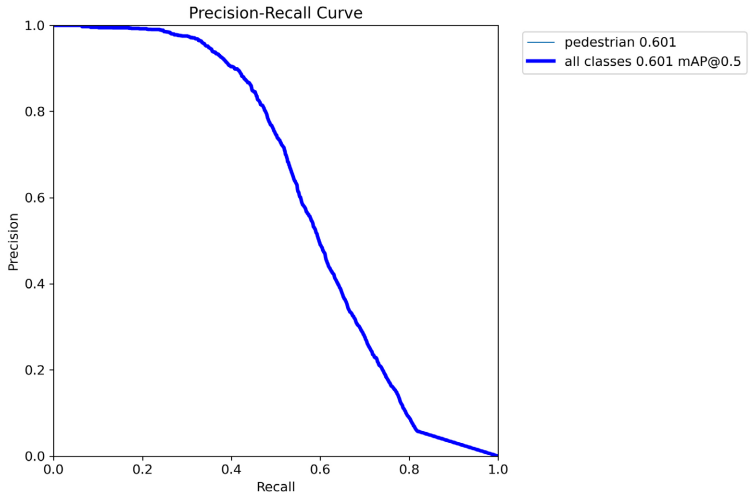
为研究在遮挡目标检测任务中注意力机制在模型结构中不同位置的有效性。通过将注意力机制嵌入

特征提取阶段(Backbone)和特征融合阶段(Neck)，并使用 Citypersons 数据集通过消融实验系统地分析不同位置对模型性能的影响，消融实验的结果如表 2 所示。

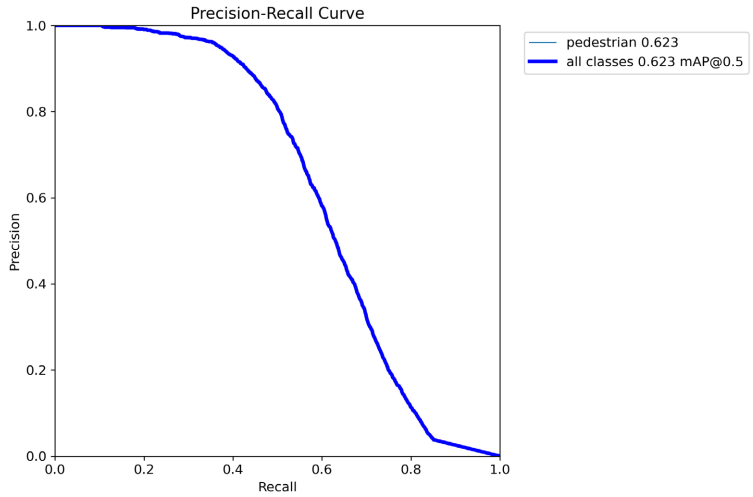
实验表明该改进是有一定提升的，实验结果如表 2 所示，由表 2 可知，该注意力机制与 neck 部分的 C2f 进行融合提升最大，相较于原始的 YOLOv10n 其 map@0.5 提升了 2.2%。图 6 是 Citypersons 数据集中算法改进前后 PR 曲线对比图，左边是原算法的 PR 曲线图，右边的是 YOLOv10n-OOD 的 PR 曲线图。

Table 2. Experimental comparison of adding EMSA to different modules
表 2. EMSA 加在各模块的实验对比表

model	Neck	Backbone	Focaleriou	map@0.5
YOLOv10n YOLOv10n_EMSA	×	√	√	61.5
	√	√	×	60.2
	√	×	√	62.3



YOLOv10n



YOLOv10n_EMSA

Figure 6. PR comparison chart after algorithm improvement on the Citypersons dataset
图 6. Citypersons 数据集中算法改进后 PR 对比图

4.4. 结论

针对遮挡图像复杂且特征模糊难以准确检测的问题，在 YOLOv10n 模型的基础上提出了融合了多尺度自注意力机制遮挡目标检测。将该机制与 YOLOv10n 中的 C2f 进行融合，并将其替换掉 neck 部分的 C2f，以此来提升网络的特征提取能力。为解决小目标占图像的像素比例小，导致特征表达弱，难以与背景分离的问题，添加了小目标检测头 P2，提升对小目标的特征提取能力。为解决正负样本之间存在的不平衡，导致忽视了困难样本的学习，从而导致模型对小目标或重叠目标的检测能力不足的问题，使用 Focaler-IoU 替换 CIOU，增强了模型对低 IoU 样本的关注。通过在 Citypersons 数据集上进行消融实验得知，该注意力机制相较于原 YOLOv10 的 map@0.5 提升了 2.2%，该模型能够更加精确地检测到遮挡目标，满足了复杂场景精确识别的要求。

参考文献

[1] Redmon, J., Divvala, S., Girshick, R. and Farhadi, A. (2016) You Only Look Once: Unified, Real-Time Object Detection. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 779-788. https://doi.org/10.1109/cvpr.2016.91

[2] Bochkovskiy, A., Wang, C.Y. and Liao, H.Y.M. (2020) YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv: 2004.10934.

[3] Chen, Z., Wang, Y. and Li, H. (2023) Dynamic Head YOLO for Detecting Occluded Objects in Complex Scenes. *Computer Vision and Image Understanding*, **237**, Article ID: 103446.

[4] Sun, X., Zhao, Y. and Gao, T. (2022) Multi-Modal YOLO for Detecting Occluded Objects in Traffic Surveillance. *Sensors*, **22**, Article 1943.

[5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.A., Kaiser, Ł. and Polosukhin, I. (2017) Attention Is All You Need. arXiv: 1706.03762.

[6] Woo, S., Park, J., Lee, J. and Kweon, I.S. (2018) CBAM: Convolutional Block Attention Module. In: Ferrari, V., Hebert, M., Sminchisescu, C. and Weiss, Y., Eds., *Computer Vision—ECCV* 2018, Springer, 3-19. https://doi.org/10.1007/978-3-030-01234-2_1

[7] Hu, J., Shen, L. and Sun, G. (2018) Squeeze-and-Excitation Networks. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 7132-7141. https://doi.org/10.1109/cvpr.2018.00745

[8] Wang, S., *et al*. (2020) Linformer: Self-Attention with Linear Complexity. arXiv: 2006.04768.

[9] Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J. and Ding, G. (2024) YOLOv10: Real-Time End-to-End Object Detection. arXiv: 2405.14458.