计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.250200051

基于局部特征和特征融合的无人驾驶场景目标检测方法

纪　涛[1,2,3]　杨一帆[1,2]　冯亚春[2]　伍凌帆[2]　李旭亮[2]　李亚伟[2]
1 云南师范大学信息学院　昆明 650500
2 北京航空航天大学宇航学院　北京 102206
3 西南联合研究生院　昆明 650500
(jitao09@foxmail.com)

摘　要　在无人驾驶场景中,目标检测的准确性和鲁棒性对系统性能至关重要。针对现有基于深度学习的网络模型在无人驾驶场景处理小目标和遮挡目标问题时出现的误检和漏检现象,提出了一种 LSDA-YOLO 网络模型。首先,提出了 LocalSimAM(Local Simple and Effective Attention Mechanism)注意力机制,用于改善信息丢失问题,并将其应用于 Backbone;同时引入 SHSA(Single-Head Self-Attention)注意力机制,设计了一个信息聚合网络,提升对遮挡目标的检测能力。在 Neck 部分,通过动态调整上采样比例,增强模型对多尺度特征的适应性,减少小目标漏检率。在 Head 部分引入了自适应空间多尺度特征融合(Adaptive Spatial Feature Fusion,ASFF)策略,增强模型的多尺度检测能力。实验结果表明,LSDA-YOLO 网络模型在 KITTI 数据集上,$mAP_{0.5}$ 和 $mAP_{0.5:0.95}$ 分别提升了 3.1 个百分点和 3.9 个百分点,优于 YOLOv11n 基准网络模型,适用于无人驾驶场景高精度实时检测。

关键词:注意力机制;无人驾驶;车辆检测;行人检测;特征融合
中图分类号　TP391.41

Unmanned Driving Scene Object Detection Method Based on Local Features and Feature Fusion

JI Tao[1,2,3],YANG Yifang[1,2],FENG Yachun[2],WU Lingfan[2],LI Xuliang[2] and LI Yawei[2]
1 School of Information Science,Yunnan Normal University,Kunming 650500,China
2 School of Astronautics,Beihang University,Beijing 102206,China
3 Southwest United Graduate School,Kunming 650500,China

Abstract　In the context of unmanned driving,the accuracy and robustness of object detection are of vital importance to the performance of the system. Aiming at the false detection and missed detection phenomena that occur when existing deep learning-based network models deal with small objects and occluded objects in unmanned driving scenarios,an LSDA-YOLO network model is proposed. Firstly,the LocalSimAM attention mechanism is proposed to address the issue of information loss,and it is applied to the Backbone. Meanwhile,the SHSA attention mechanism is introduced,and an information aggregation network is designed to enhance the detection ability for occluded objects. In the Neck part,by dynamically adjusting the upsampling ratio,the adaptability of the model to multi-scale features is enhanced,reducing the missed detection rate of small objects. In the Head part,the ASFF strategy is introduced to enhance the model's multi-scale detection ability. Experimental results show that the LSDA-YOLO network model improves the $mAP_{0.5}$ and $mAP_{0.5:0.95}$ by 3.1 percentage points and 3.9 percentage points respectively on the KITTI dataset,outperforming the YOLOv11n baseline network model,and is suitable for high-precision real-time detection in unmanned driving scenarios.

Keywords　Attention mechanism,Unmanned driving,Vehicle detection,Pedestrian detection,Feature fusion

1　引言

无人驾驶技术的快速发展使得目标检测成为确保系统安全性和可靠性的重要组成部分。基于深度学习的目标检测算法主要分为两阶段检测方法和一阶段检测方法。两阶段方法如 Fast R-CNN[1] 和 Faster R-CNN[2],虽然在检测精度上表现出色,但其处理图像的过程较为耗时,不适用于无人驾驶场景下的实时检测需求。相比之下,一阶段方法如 SSD[3],YOLO[4] 等,通过单次前向传播完成目标的定位和分类,在保证一定检测精度的同时降低了计算成本,在无人驾驶场景下具有一定的优势。

在无人驾驶道路交通环境中,小目标和遮挡目标的误检和漏检问题尤为突出,影响系统的整体性能。为解决这些问题,已有研究提出了多种策略。Lin 等通过构建多尺度特征金字塔,结合自底向上和自顶向下路径以及横向连接,融合不同尺度的特征信息[5],一定程度上提升了小目标的检测精度,

基金项目:国家自然科学基金(62476017)
This work was supported by the National Natural Science Foundation of China(62476017).
通信作者:杨一帆(yifanyang@buaa.edu.cn)

但一定程度上增加了过拟合风险。Lim 等通过多尺度特征融合(Multi-scale Feature Fusion, MSFF)和注意力机制[6](Attention Mechanism, AM),提升了小目标在低分辨率和信息有限的情况下检测的精度,但推理速度较慢。Bai 等提出了一个端到端的多任务生成对抗网络[7](Multi-Task Generative Adversarial Network, MTGAN)。在 MTGAN 中,生成器是一个超分辨网络,可以将小的模糊图像上采样为细节丰富的图像,并恢复详细信息以进行更准确的检测,该方法计算复杂度较高。Kisantal 分析了 Mask R-CNN 在 MS-COCO 数据集上小目标检测的表现,提出了一种通过采样和复制粘贴小目标[8]来增强图像数据的方法。Li 等引入选择性内核模块[9](Selective Kernel, SK),通过动态选择不同大小的卷积核,自适应地调整特征图的响应,提高了模型对遮挡目标的检测精度,但推理速度较慢。Ju 等通过增加 2 倍上采样和残差单元,利用 K-means 聚类优化锚框[10],提升了小目标检测的召回率和平均准确率。Chen 等提出了一种结合全局和局部特性的局部补丁网络(Local Patch Network, LPNet)与全局注意力机制[11],有效提升了小目标检测的效率和精度。Li 等通过修改卷积块注意力机制(Convolutional Block Attention Module, CBAM)和采用 Lite-Hourglass 模块[12],增强了物体遮挡检测问题,但是该模型所需计算资源更多,计算复杂度高。

本文针对无人驾驶目标检测中小目标和遮挡目标的误检和漏检问题严重、待检测目标尺寸差异大等问题,提出了一种结合关注局部区域自注意力机制和特征融合的目标检测方法。主要工作如下:

1)提出一种 LocalSimAM 注意力机制,并将其引入 C3k2 模块,改善模型对细节的关注能力,减少信息丢失;

2)在 C2 层引入单头自注意力(SHSA)机制,设计了基于单头自注意力的信息聚合网络,强化模型对关键区域的关注,增强遮挡目标的检测效果;

3)摒弃传统的 Upsample 方法,通过动态调节上采样比例,优化边缘保持能力和细节复现能力,减少漏检情况;

4)引入自适应空间多尺度特征融合(ASFF)机制,根据需要动态调整各尺度特征图的权重,进一步加强模型的整体检测效能。

2 YOLOv11n 网络模型

YOLOv11n 是 Ultralytics 公司推出的 YOLOv8n 模型的改进版本,其网络结构如图 1 所示,主要由主干网络(Backbone)、颈部网络(Neck)和头部网络(Head)3 部分构成。

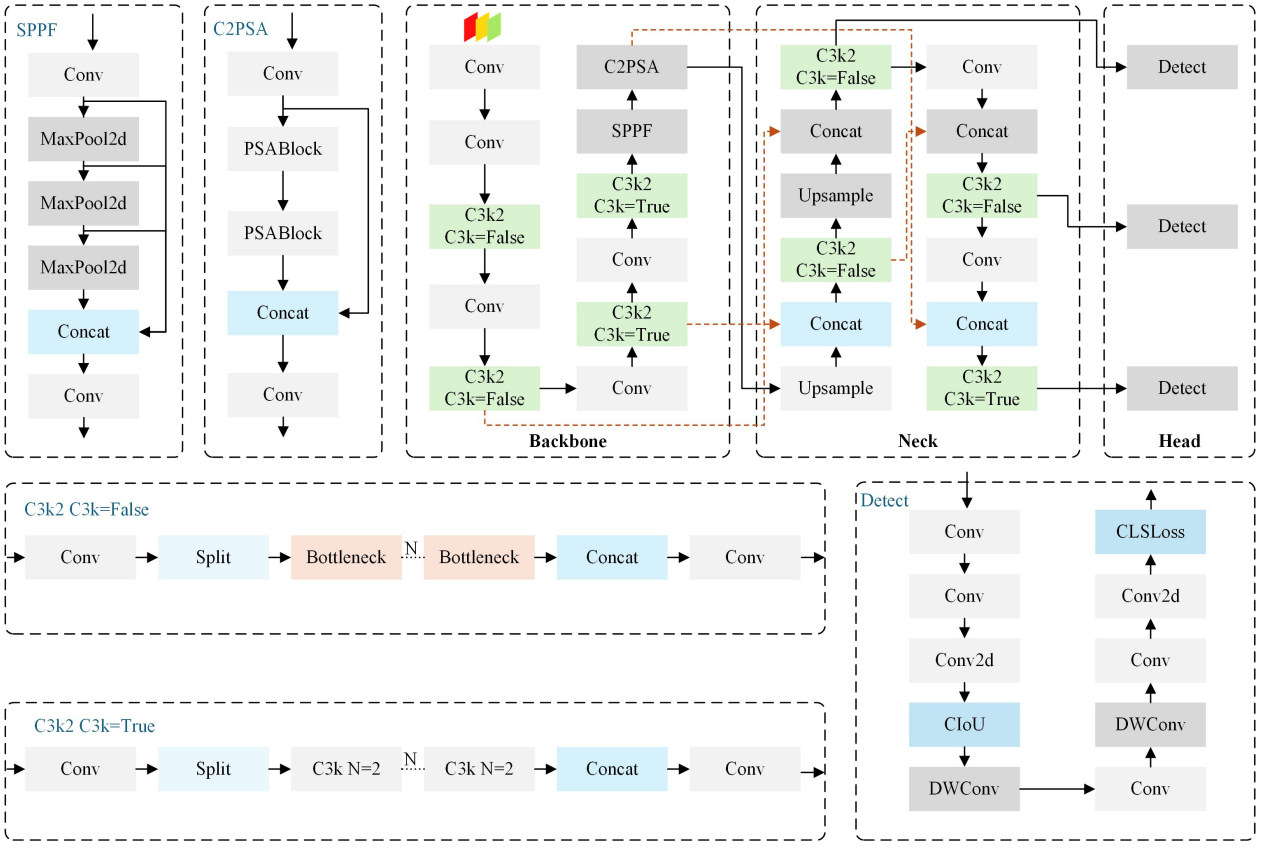


图 1 YOLOv11n 网络模型结构

Fig. 1 Structure of YOLOv11n network model

3 LSDA-YOLO 网络模型

LSDA-YOLO 网络结构如图 2 所示。本文提出一种新的 LocalSimAM 注意力机制,并将其应用于主干网络的 C3k2 模块中。LocalSimAM 通过局部区域特征捕捉的自注意力机制,增强了对小目标特征的捕捉能力,有效减少了特征信息的丢失。在 C2 层引入单头自注意力(SHSA)机制,设计了基于单头自注意力的信息聚合网络。这一改进强化了模型对关键区域的关注,提高了对遮挡目标的检测效果。传统的上采样方法在边缘保持能力和细节复现能力上存在局限性。为此,本文使用了一种动态调节上采样比例的方法来改善该问题。最后,本文引入了自适应空间特征融合(ASFF)机制。ASFF 根据需要动态调整各尺度特征图的权重,进一步提升了模型的整体检测性能。

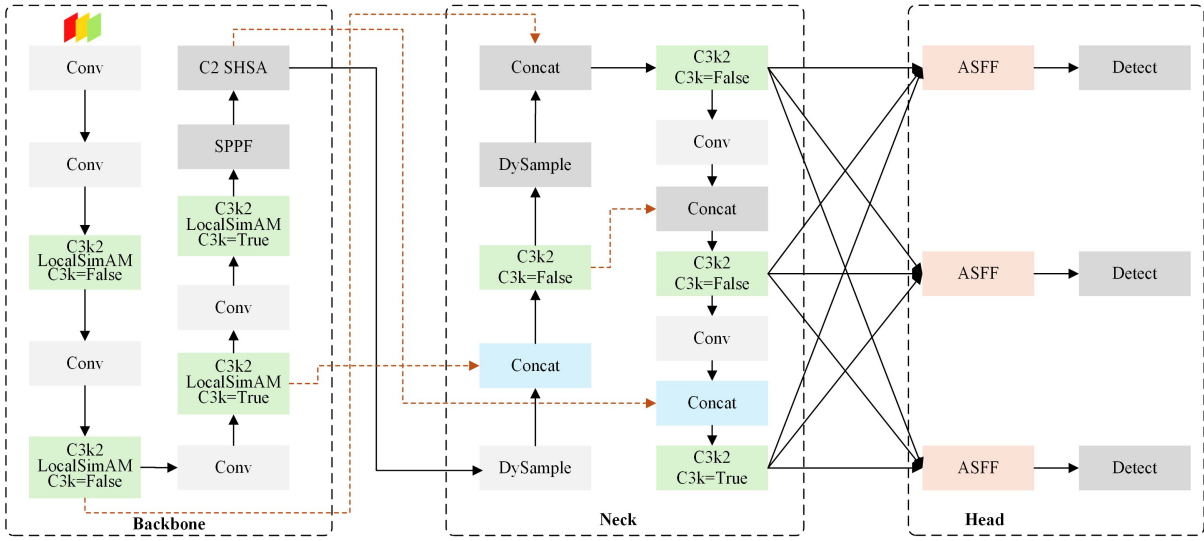


图 2　LSDA-YOLO 网络模型结构

Fig. 2　Structure of LSDA-YOLO network model

3.1　基于局部特征捕捉的注意力机制

C3k2 模块通过堆叠多个 Bottleneck 结构来提取特征。在 Bottleneck 结构中，为了降低计算量，会先进行通道数的压缩，再进行扩张。然而，这种压缩操作可能会导致部分特征信息的丢失，从而影响小目标检测效果。

尽管 SimAM[13] 在许多任务中表现出色，但在处理小目标和遮挡目标时，其全局统计信息的计算方式通常会平滑掉局部特征的差异，使得模型难以捕捉到小目标或部分被遮挡目标的独特特征，这会导致注意力权重的不准确。为了降低特征信息丢失对小目标检测的影响，本文基于 SimAM 注意力机制提出了一种改进的 LocalSimAM 注意力机制，并将其引入 C3k2 模块。Bottleneck 结构通过压缩和恢复通道数来降低计算量，而 LocalSimAM 通过局部特征捕捉来增强模型的检测能力。

为增强局部特征捕捉能力，LocalSimAM 采用滑动窗口形式计算统计量。具体而言：在输入特征图的每个空间位置，定义一个以该位置为中心的 3×3 局部窗口（步长为 1），允许窗口覆盖整个特征图并存在重叠区域。在每个窗口内独立计算特征值的均值和平方差，如式(1)所示。这种逐位置滑动的方式避免了全局统计信息对局部细节的平滑，尤其适用于小目标和遮挡目标。通过归一化处理和 Sigmoid 激活函数生成注意力权重，最终与输入特征图逐点相乘，增强局部显著区域。最后将注意力权重恢复到原始特征图的形状，并与输入特征图相乘，得到经过增强后的输出特征图。LocalSimAM 被嵌入到 C3k2 模块的每个 Bottleneck 结构内部，具体流程如图 3 所示。

图 3　嵌入 LocalSimAM 的 Bottleneck 处理流程

Fig. 3　Bottleneck processing flow embedded in LocalSimAM

LocalSimAM 不再依赖全局统计信息，而是通过局部特征捕捉来增强模型对小目标和遮挡目标的检测能力。具体来说，它会在每个局部区域内独立计算注意力权重，确保每个小目标或遮挡部分都能获得适当的注意力权重，避免被全局统计信息所忽视。这种机制使得 LocalSimAM 能够有效提升模型在密集场景下的表现，特别是针对尺寸较小的目标物体检测任务。其数学表达式如下：

$$y_{ij}=\frac{\left(x_{ij}^{w}(k)-\frac{1}{w^2}\sum_{k=1}^{w^2}x_{ij}^{w}(k)\right)^2}{4\left(\frac{1}{w^2-1}\sum_{k=1}^{w^2}\left(x_{ij}^{w}(k)-\frac{1}{w^2}\sum_{k=1}^{w^2}x_{ij}^{w}(k)\right)^2+\lambda\right)}+0.5 \quad (1)$$

其中，$\frac{1}{w^2}\sum_{k=1}^{w^2}x_{ij}^{w}(k)$ 是局部窗口 x_{ij}^{w} 内的均值，w^2 是窗口内的像素数量；$\left(x_{ij}^{w}(k)-\frac{1}{w^2}\sum_{k=1}^{w^2}x_{ij}^{w}(k)\right)^2$ 是局部窗口内每个像素值与均值的平方差；λ 是一个小的正则化项，此处为固定值 1×10^{-6}，防止除以零错误；y_{ij} 是归一化的相似性度量。为验证本文提出的 LocalSimAM 注意力机制的有效性，将 LocalSimAM 注意力机制与 SimAM 注意力机制进行了对比实验，如表 1 所列。

表 1　LocalSimAM 与 SimAM 注意力机制对比

Table 1　Comparison of attention mechanism between LocalSimAM and SimAM

（%）

Attention	P	R	$mAP_{0.5}$
SimAM	**90.4**	77.1	86.6
LocalSimAM	89.5	**78.5**	**87.2**

由表 1 可以看出，LocalSimAM 注意力机制在 R 和 $mAP_{0.5}$ 上均有明显提升，验证了改进的有效性。

3.2　基于单头自注意力机制的信息聚合网络

在 YOLOv11n 中，引入的金字塔切片注意力机制[14]（PSA）虽然在捕捉全局信息方面表现出色，但是其在早期阶段引入了较多的冗余信息。这些冗余信息可能会干扰模型对关键特征的提取，特别是在处理密集行人、遮挡车辆等时，冗余信息可能导致模型在关键区域的特征表示不够精确，从而影响检测精度。为了在保持检测精度的同时降低计算成本，本文引入单头自注意力（SHSA）机制[15]，设计了基于单头自注意力的信息聚合网络，在部分输入通道应用注意力层，其余通道保持不变，如图 4 所示。

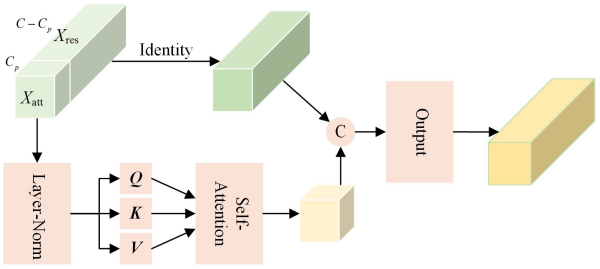


图 4 基于单头自注意力的信息聚合网络

Fig. 4 Information aggregation network based on single head self attention

该信息聚合网络在接收一个 4 维输入张量后，首先将其按照通道维度分裂成两部分 $\boldsymbol{X}_{\text{att}}$ 和 $\boldsymbol{X}_{\text{res}}$，其中 $\boldsymbol{X}_{\text{att}}$ 经过归一化处理后用于生成查询(Q)、键(K)和值(V)的表示。这些表示在空间维度上被展平以计算自注意力机制中的相似度矩阵，并通过 softmax 函数转换为注意力权重。利用这些权重对值进行加权求和，得到更新后的特征图，然后与未处理的 $\boldsymbol{X}_{\text{res}}$ 部分拼接在一起，最后经过投影层进行线性变换和非线性激活，输出同样形状的张量，完成一次单头自注意力机制的特征增强过程。SHSA 的数学定义如下：

$$\boldsymbol{X}_{\text{att}},\boldsymbol{X}_{\text{res}}=Split(\boldsymbol{X},[C_p,C-C_p]) \tag{2}$$

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\text{T}}}{\sqrt{d_{qk}}}\right)\boldsymbol{V} \tag{3}$$

$$\widetilde{\boldsymbol{X}}_{\text{att}}=Attention(\boldsymbol{X}_{\text{att}}\boldsymbol{W}_Q,\boldsymbol{X}_{\text{att}}\boldsymbol{W}_K,\boldsymbol{X}_{\text{att}}\boldsymbol{W}_V) \tag{4}$$

$$SHSA(\boldsymbol{X})=Concat(\widetilde{\boldsymbol{X}}_{\text{att}},\boldsymbol{X}_{\text{res}})\boldsymbol{W}_O \tag{5}$$

式中，将输入特征图 X 分裂为两部分：一部分用于注意力机制计算，另一部分保持不变。假设 $\boldsymbol{X}$ 的通道数为 C，将 $\boldsymbol{X}$ 分裂为 $\boldsymbol{X}_{\text{att}}$ 和 $\boldsymbol{X}_{\text{res}}$，其中 $\boldsymbol{X}_{\text{att}}$ 的通道数为 Cp，$\boldsymbol{X}_{\text{res}}$ 的通道数为 $C-Cp$。对于 $\boldsymbol{X}_{\text{att}}$，应用单头自注意力机制。首先，通过线性变换将 $\boldsymbol{X}_{\text{att}}$ 投影到查询向量 $\boldsymbol{Q}$、键向量 $\boldsymbol{K}$ 和值向量 $\boldsymbol{V}$。其中，$\boldsymbol{K}=\boldsymbol{X}_{\text{att}}\boldsymbol{W}_K$，$\boldsymbol{V}=\boldsymbol{X}_{\text{att}}\boldsymbol{W}_V$；$\boldsymbol{W}_Q$，$\boldsymbol{W}_K$，$\boldsymbol{W}_V$ 是投影权重矩阵。对注意力权重矩阵 Attention($\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$)进行归一化。式(4)中，$d_{qk}$ 是查询向量和键向量的维度。注意力特征图 $\widetilde{\boldsymbol{X}}_{\text{att}}$ 与未变化的特征图 $\boldsymbol{X}_{\text{res}}$ 拼接，并通过一个线性变换 $\boldsymbol{W}_O$ 得到最终的输出特征图。通过基于单头自注意力机制的信息聚合网络，模型能够更精准地聚焦于图像中的关键区域，减少了冗余信息的干扰。这在处理密集行人、遮挡车辆等场景时尤为重要，因为这些场景中存在大量背景噪声和部分遮挡，需要模型具备更强的特征捕捉能力。

3.3 强化边缘保持能力的动态上采样算子

在目标检测任务中，特征图的上采样是关键步骤之一。DySample[16]是基于点采样的动态上采样方法。与传统基于内核的动态上采样方法(例如 CARAFE[17]，FADE[18]及 SAPA[19])相比，DySample 具有轻量化、高效性和灵活性等优势。它不需要额外的 CUDA 包实现，参数量较少。DySample 通过学习偏移量来动态调整采样点，从而更好地适应无人驾驶场景的需求，而无需高分辨率引导特征。YOLOv11 使用的 Upsample 在处理高分辨率图像和细节丰富的场景时，其边缘保持能力和细节复现能力相对较弱。为了提高模型在道路无人驾驶场景中的表现，本文引入了 DySample 作为新的上采样方法，如图 5 所示。

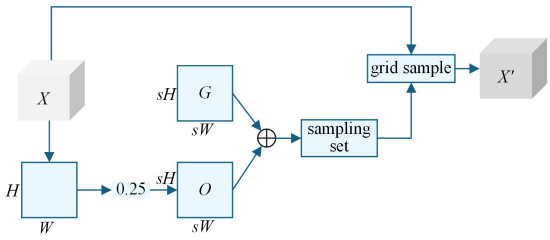


图 5 动态上采样算子处理

Fig. 5 Dynamic up sampling operator processing

处理的具体流程是将上一层输出的特征图 $\boldsymbol{X}\in\mathbb{R}^{C\times H\times W}$(其中 C 是通道数，H 和 W 分别是高度和宽度)通过一个卷积层对输入特征图 $\boldsymbol{X}$ 进行处理，得到预测初始偏移量。根据是否启用动态范围因子，调整初始偏移量的大小。在接下来的实验中本文选择了不启用动态范围因子，使用静态范围因子 0.25 来限制偏移量的幅度，以减少计算量。在预测初始偏移量与静态范围因子做乘积后，生成一个网格 G，定义每个像素点在上采样后的输出特征图上的基准位置。网格中的每个元素表示相对于每个高分辨率像素点的相对索引。网格 G 的计算式如下：

$$G(i,j)=\left(\frac{i+0.5}{s},\frac{j+0.5}{s}\right),i,j\in\left\{-\frac{s-1}{2},\cdots,\frac{s-1}{2}\right\} \tag{6}$$

其中，s 为上采样因子，sH 和 sW 表示经过上采样后的特征图尺寸，加 0.5 是为了将坐标中心化到像素中心。接下来，将经过调整的偏移量加上初始位置偏移量网格 G 得到最终偏移量。最后，基于计算出的新坐标 C 对输入特征图 X 进行双线性插值采样，得到上采样后的特征图 X'。计算新坐标 C 时，将每个像素点的坐标转换为归一化的坐标系(范围从 −1 到 1)，并加上偏移量以确定新的采样位置。通过采用动态上采样使得模型在在边缘保持能力和细节复现能力上得到了提升，减少了目标漏检的情况。

3.4 自适应空间多尺度特征融合

在无人驾驶场景中，道路环境复杂多变，包含各种大小的行人、车辆等。这些目标在图像中的尺度差异大，从占据整个画面的车辆到仅占几个像素的行人。这种尺度变化对检测器提出了较高的要求，特征金字塔网络在融合不同尺度的特征图时，采用固定的权重分配策略，这导致某些尺度的信息被忽略并且产生梯度计算中的不一致性。

如图 6 所示，为了应对多尺度特征表示带来的挑战，本文在网络头部引入了自适应空间多尺度特征融合[20](ASFF)，ASFF 通过自适应地调整不同尺度特征图的权重，可以改善某些尺度的信息被忽略和梯度计算中的不一致性，使得模型能够更好地捕捉和利用多尺度信息。

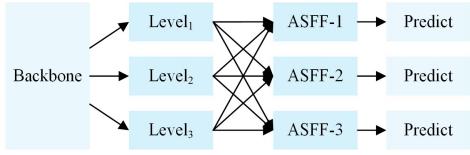


图 6 自适应空间多尺度特征融合

Fig. 6 Adaptive spatial multiscale feature fusion

首先需要将这些特征图重缩放到相同的分辨率。具体来说，设有 3 个不同尺度的特征图 X_1，X_2，X_3，分别对应于不同的感受野。本文将输入层 X_l($l\in(1,2,3)$)的特征图设为

X_1，并将其他特征层 $L_n(n\neq l)$ 的特征调整到与 X_1 相同的尺度大小。为了统一特征图的尺度，对于小于给定阈值的特征图，先使用 1×1 点卷积压缩通道数至与目标层级相同，再通过插值法扩大其尺寸；而对于大于给定阈值的特征图，则采用步长为 2 的 3×3 卷积进行下采样，将其尺寸减半并调整通道数；若需 1/4 比例下采样，则先用步长为 2 的最大池化层，再接一个步长为 2 的 3×3 卷积层，从而实现特征图在尺寸和通道数上的统一。

在特征重缩放之后，通过学习每个位置的最优融合权重，自适应地融合不同尺度的特征图。对于每个位置(i,j)，来自不同层次的特征向量根据学习到的空间重要性权重 α_{ij}^{l}，β_{ij}^{l}，γ_{ij}^{l} 进行加权求和。计算式如下：

$$y_{ij}^{l}=\alpha_{ij}^{l}\cdot x_{1\to l}^{ij}+\beta_{ij}^{l}\cdot x_{2\to l}^{ij}+\gamma_{ij}^{l}\cdot x_{3\to l}^{ij} \tag{7}$$

其中，y_{ij}^{l} 是在位置(i,j)处，第 l 层融合后的特征图；$x_{n\to l}^{ij}$ 是在位置(i,j)处，第 n 层融合后的特征图；α_{ij}^{l}，β_{ij}^{l}，γ_{ij}^{l} 是在位置(i,j)处，第 l 层融合后的特征图。它们的定义如下：

$$\alpha_{ij}^{l}=\frac{e^{\lambda_{\alpha}^{l,ij}}}{e^{\lambda_{\alpha}^{l,ij}}+e^{\lambda_{\beta}^{l,ij}}+e^{\lambda_{\gamma}^{l,ij}}} \tag{8}$$

$$\beta_{ij}^{l}=\frac{e^{\lambda_{\beta}^{l,ij}}}{e^{\lambda_{\alpha}^{l,ij}}+e^{\lambda_{\beta}^{l,ij}}+e^{\lambda_{\gamma}^{l,ij}}} \tag{9}$$

$$\gamma_{ij}^{l}=\frac{e^{\lambda_{\gamma}^{l,ij}}}{e^{\lambda_{\alpha}^{l,ij}}+e^{\lambda_{\beta}^{l,ij}}+e^{\lambda_{\gamma}^{l,ij}}} \tag{10}$$

控制参数 λ 通过 1×1 的点卷积从重缩放后的特征图中计算得到。这些权重满足：

$$\alpha_{ij}^{l}+\beta_{ij}^{l}+\gamma_{ij}^{l}=1 \tag{11}$$

最后，利用融合后的特征图进行接下来的预测。检测头为每个网格单元预测多个边界框的位置、大小、类别概率以及置信度评分。接着，使用非极大值抑制(NMS)去除冗余的重叠边界框，保留最有可能的对象检测结果。

通过引入自适应空间多尺度特征融合可以较好地改善某些尺度的信息被忽略和梯度计算中的不一致性。

4　实验设计与结果分析

4.1　实验环境与评价指标

本实验训练平台为网络服务器，操作系统是 Ubuntu 20.04，CPU 为 Intel(R) Xeon(R) Platinum 8369B CPU @ 2.90GHz，内存 30 GiB，显卡为 NVIDIA A10，显存 24 GiB，CUDA 版本 11.3，Python 版本 3.9，PyTorch 版本 1.12。实验超参数采用 YOLOv11n 默认设置，采用 SGD 优化器，初始学习率为 0.01，学习动量为 0.937，权重衰减系数为 0.0005，batchsize 设为 64，迭代 300 个 epoch。在训练过程中，本文对图像进行了 0.5～1.5 倍的随机缩放，并结合了水平翻转、平移和马赛克增强等多种数据增强操作，在训练到 290 个 epoch 后关闭数据增强。

本实验选择神经网络最常见的评价指标：精度(P)、召回率(R)、平均精度均值(mAP)、帧率(FPS)，以及参数量(Params)。计算式如下：

$$P=\frac{TP}{TP+FP} \tag{12}$$

$$R=\frac{TP}{TP+FN} \tag{13}$$

$$mAP=\frac{1}{c}\sum_{i=1}^{c}AP_i \tag{14}$$

其中，TP 为被正确预测为正类别的实例，FN 实际为正类别但被错误地预测为负类别的实例，FP 为实际上为负类别但被错误地预测为正类别的实例，c 为类别总数，AP_i 为第 i 类别的平均精度。

4.2　实验数据集与预处理

实验使用的是 KITTI 目标检测数据集[20]，这是一个广泛应用于无人驾驶领域中的基准数据集。该数据集包含了大量的真实道路场景图像，标注了多种类别目标，包括 9 个类别，分别为：汽车、面包车、货车、卡车、行人、坐着的人、骑自行车的人、杂项和不关心标签。

在实验中，本文更加注重于模型检测的能力，所以将 KITTI 数据集中的类别重新定义为 3 个类别，将汽车、面包车、货车和卡车合并为汽车(car)，行人、坐着的人合并为行人(pedestrian)，骑自行车的人(cyclist)保持不变并舍弃不关心和杂项标签。由于测试集没有真实框和标签数据，故只使用训练集。为了更好地评估模型的性能，本文将原始训练集(7481 张图片)按照 8∶2 的比例随机划分为训练集和验证集。

4.3　注意力机制消融实验

为了验证本文提出的 LocalSimAM 注意力机制和引入 SHSA 注意力机制设计的信息聚合网络对整个特征提取网络的影响，本文设计了如下消融实验。将使用了 LocalSimAM 注意力机制的模型命名为 YOLOv11n-L，将使用了信息聚合网络的模型命名为 YOLOv11n-S，同时使用 LocalSimAM 注意力机制和信息聚合网络的模型命名为 YOLOv11n-LS。实验结果如表 2 所列。从表 2 的数据可以看出，引入 LocalSimAM 和 SHSA 注意力机制对模型性能都有不同程度的提升。

表 2　注意力机制消融实验

Table 2　Ablation experiment of attentional mechanism

Model	P/%	R/%	$mAP_{0.5}$/%	$mAP_{0.5:0.95}$/%	FPS	Params
YOLOv11n	90.7	77.3	86.8	62.5	103	2.62×10^6
YOLOv11n-L	89.5	78.4	87.2	62.2	99	2.49×10^6
YOLOv11n-S	91.2	**79.7**	**87.7**	**63.3**	101	2.55×10^6
YOLOv11n-LS	**91.3**	77.6	87.4	62.6	**105**	$\mathbf{2.46\times10^6}$

YOLOv11n-L 的精确率从 90.7%降至 89.5%，但召回率提升至 78.4%，$mAP_{0.5}$ 提升了 0.4 个百分点。YOLOv11n-S 的精确率和召回率分别达到 91.2%和 79.7%，显示 SHSA 能有效增强关键区域的关注，$mAP_{0.5}$ 和 $mAP_{0.5:0.95}$ 分别提升了 0.9 个百分点和了 0.8 个百分点。同时使用两种注意力机制的 YOLOv11n-LS 在综合性能上较为均衡，在所有评估指标上保持了较高的水平，表明了这两种注意力机制结合的有效性。

4.4　动态上采样自适应空间特征融合消融实验

从 4.3 节的注意力机制消融实验中可以看出，LocalSimAM 和 SHSA 注意力机制在降低模型参数量的同时，提升了模型的特征捕捉能力。为了进一步评估动态上采样(DySample)和自适应空间特征融合(ASFF)机制的作用，本文进行了以下消融实验。基于 YOLOv11n，将使用了 DySample 的模型命名为 YOLOv11n-D，将使用了 ASFF 的模型命名为 YOLOv11n-A，而同时使用这两种机制的模型命名 YOLOv11n-DA。基于 4.3 节的 YOLOv11n-LS，将使用动态上采样的模型命名 YOLOv11n-LSD，将使用了 ASFF 的模型

命名为 YOLOv11n-LSA，而同时使用 4 种机制的模型命名 LSDA-YOLO。实验结果如表 3 所列。

表 3 消融实验结果

Table 3 Ablation experiments results

Model	P/%	R/%	$mAP_{0.5}$/%	$mAP_{0.5:0.95}$/%	FPS	Params
YOLOv11n	90.7	77.3	86.8	62.5	**103**	2.62×10^6
YOLOv11n-D	92.3	78.4	87.7	63.6	101	2.60×10^6
YOLOv11n-A	89.0	80.5	88.4	64.3	92	3.96×10^6
YOLOv11n-DA	89.5	80.9	88.7	65.1	84	3.97×10^6
YOLOv11n-LSD	91.5	78.9	88.2	63.3	95	$\mathbf{2.47\times10^6}$
YOLOv11n-LSA	90.5	80.4	89.0	64.8	86	3.83×10^6
LSDA-YOLO	**92.7**	**81.5**	**89.9**	**66.4**	98	3.85×10^6

分析表 3 可知，DySample 提升了检测精度，ASFF 增强了召回率和特征表达能力。两者结合时，模型的 $mAP_{0.5}$ 提升了 1.9 个百分点，达到 88.7%，$mAP_{0.5:0.95}$ 提升了 2.6 个百分点，达到 65.1%，但 FPS 降至 84 帧。进一步结合 LocalSimAM 和 SHSA 注意力机制后，LSDA-YOLO 的 mAP 提升了 3.1 个百分点，达到 89.9%，$mAP_{0.5:0.95}$ 提升了 3.9 个百分点，达到 66.4%。在不影响模型实时性和轻量化的前提下，LSDA-YOLO 的检测速度为每秒 98 帧，较基线模型有所降低，参数量为 3.85M，较基线模型有所增加。

4.5 主流模型对比实验

为了客观评价提出的 LSDA-YOLO 模型的优势，本文将其与其他目标检测算法（Faster R-CNN[21]，SSD[22]，DETR[23]，YOLOv5n[24]，YOLOv6[25]，YOLOv7-tiny[26]，YOLOv8n[27]，YOLOv10[28]）在 KITTI 数据集上进行了对比实验。实验结果涵盖了精度（P）、召回率（R）等关键性能指标，如表 4 所列。

表 4 对比实验

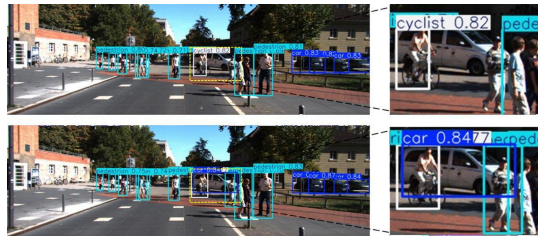
Table 4 Comparative experiments

Model	P/%	R/%	$mAP_{0.5}$/%	$mAP_{0.5:0.95}$/%	FPS	Params
Faster R-CNN	62.0	68.2	61.2	38.1	19	82.60×10^6
SSD	75.8	67.4	72.4	51.0	62	24.40×10^6
DETR	—	—	88.2	53.7	—	41.20×10^6
YOLOv5n	88.4	76.1	83.7	55.2	86	5.03×10^6
YOLOv6	87.5	79.3	87.5	59.6	91	9.00×10^6
YOLOv7-tiny	89.3	78.2	86.3	55.3	83	6.00×10^6
YOLOv8n	89.8	80.7	87.2	64.7	**108**	$\mathbf{3.15\times10^6}$
YOLOv10	91.2	79.8	89.1	65.1	102	8.10×10^6
LSDA-YOLO	**92.7**	**81.6**	**89.9**	**66.4**	98	3.85×10^6

从表中可以看出，LSDA-YOLO 在所有评估指标上均表现出色，特别是在 $mAP_{0.5}$ 和 $mAP_{0.5:0.95}$ 这两个关键指标上分别达到了 89.9% 和 66.4%，显著优于其他模型。同时，LSDA-YOLO 的检测速度为每秒 98 帧，参数量为 3.85M，在主流模型中属于高效且轻量化的模型。这表明 LSDA-YOLO 不仅在检测精度上具有明显优势，同时在实时性和计算资源需求方面也保持了良好的平衡，适合应用于复杂多变的实际场景。因此，LSDA-YOLO 是一个兼具高性能和高效率的目标检测模型。

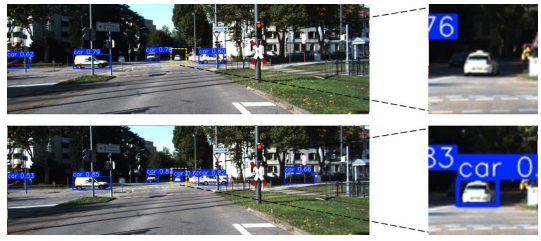
4.6 LSDA-YOLO 检测效果分析

为了直观展示模型改进前后的检测效果，本文在 KITTI 测试集对本文提出的 LSDA-YOLO 模型与 YOLOv11n 模型进行可视化对比测试，以更准确地反映它们在真实应用场景中的性能差异。具体的检测结果如图 7 所示。其中选择了 3 个具有代表性的场景：密集遮挡行人、小目标非机动车和小目标车辆。这些场景不仅涵盖了复杂的交通环境，还考验模型在处理遮挡和小目标检测方面的鲁棒性和准确性。



(a)密集遮挡行人

(b)小目标非机动车



(c)小目标车辆

图 7 检测结果对比

Fig. 7 Comparison of test results

结束语 本文针对无人驾驶场景中目标检测的准确性和鲁棒性问题，提出了一种基于局部特征捕捉和特征融合的 LSDA-YOLO 网络模型。通过引入 LocalSimAM 注意力机制和 SHSA 注意力机制，设计了信息聚合网络，增强了模型对小目标和遮挡目标的检测能力。同时，引入动态上采样算子和自适应空间多尺度特征融合（ASFF）机制，进一步提升了模型对多尺度特征的适应性和检测精度。实验结果表明，LSDA-YOLO 在 KITTI 数据集上的 $mAP_{0.5}$ 和 $mAP_{0.5:0.95}$ 分别提升了 3.1 个百分点和 3.9 个百分点，显著优于 YOLOv11n 基准模型，且在实时性和参数量方面保持了良好的平衡。可视化对比实验进一步验证了 LSDA-YOLO 在复杂场景下的性能，尤其是在密集遮挡和小目标检测任务中表现出色。本文的研究为无人驾驶场景下的高精度实时目标检测提供了有效的解决方案，具有重要的实际应用价值。未来的工作将进一步优化模型的轻量化设计，并探索其在更多复杂场景中的适用性。

参 考 文 献

[1] GIRSHICK R. Fast R-CNN[J]. arXiv:1504.08083,2015.

[2] 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,2016,39(6):1137-1149.

[3] LIU W,ANGUELOV D,ERHAN D,et al. Ssd:Single shot multibox detector[C]// Computer Vision — ECCV 2016:14th European Conference, Amsterdam, The Netherlands. Springer

International Publishing，2016：21-37.

[4] REDMON J. You only look once：Unified，real-time object detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[5] LIN T Y，DOLLÁR P，GIRSHICK R，et al. Feature pyramid networks for object detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017：2117-2125.

[6] LIM J S，ASTRID M，YOONH J，et al. Small object detection using context and attention[C]// 2021 International Conference on Artificial Intelligence in Information and Communication(ICAIIC). IEEE，2021：181-186.

[7] BAI Y，ZHANG Y，DING M，et al. Sod-mtgan：Small object detection via multi-task generative adversarial network[C]// Proceedings of the European Conference on Computer Vision(ECCV). 2018：206-221.

[8] KISANTAL M. Augmentation for Small Object Detection[J]. arXiv：1902. 07296，2019.

[9] LI X，WANG W，HU X，et al. Selective kernel networks[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019：510-519.

[10] JU M R，LUO H B，WANG Z B，et al. Improved YOLO V3 algorithm and its application in small object detection [J]. Acta Optica Sinica，2019，39(7)：0715004.

[11] CHEN F，GAO C，LIU F，et al. Local patch network with global attention for infrared small object detection[J]. IEEE Transactions on Aerospace and Electronic Systems，2022，58(5)：3979-3991.

[12] LI G，FAN W，XIE H，et al. Detection of road objects based on camera sensors for autonomous driving in various traffic situations[J]. IEEE Sensors Journal，2022，22(24)：24253-24263.

[13] YANG L，ZHANG R Y，LI L，et al. Simam：A simple，parameter-free attention module for convolutional neural networks [C]// International Conference on Machine Learning. PMLR，2021：11863-11874.

[14] ZHANG H，ZU K，LU J，et al. Epsanet：An efficient pyramid split attention block on convolutional neural network[C]// CoRR. 2021.

[15] YUN S，RO Y. Shvit：Single-head vision transformer with memory efficient macro design[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024：5756-5767.

[16] LIU W，LU H，FU H，et al. Learning to upsample by learning to sample[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023：6027-6037.

[17] WANG J，CHEN K，XU R，et al. Carafe：Content-aware reassembly of features[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019：3007-3016.

[18] LU H，LIU W，FU H，et al. FADE：A Task-Agnostic Upsampling Operator for Encoder-Decoder Architectures[J]. arXiv：2407. 13500，2024.

[19] LU H，LIU W，YE Z，et al. SAPA：Similarity-aware point affiliation for feature upsampling[J]. Advances in Neural Information Processing Systems，2022，35：20889-20901.

[20] GEIGER A，LENZ P，STILLER C，et al. KITTI Vision Benchmark Suite[EB/OL]. https：//www. cvlibs. net/datasets/kitti.

[21] REN S，HE K，GIRSHICKR，et al. Faster R-CNN：Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2016，39(6)：1137-1149.

[22] LIU W，ANGUELOV D，ERHAN D，et al. Ssd：Single shot multibox detector[C]// Computer Vision－ECCV 2016：14th European Conference，Amsterdam，The Netherlands. Springer International Publishing，2016：21-37.

[23] CARION N，MASSA F，SYNNAEVEG，et al. End-to-end object detection with transformers[C]// European Conference on Computer Vision. Cham：Springer International Publishing，2020：213-229.

[24] Ultralytics. Ultralytics/yolov5. GitHub [DB/OL]. https：//github. com/ultralytics/yolov5.

[25] LI C，LI L，JIANGH，et al. YOLOv6：A single-stage object detection framework for industrial applications[J]. arXiv：2209. 02976，2022.

[26] WANG C Y，BOCHKOVSKIY A，LIAO H Y M. YOLOv7：Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023：7464-7475.

[27] ULTRALYTICS. Ultralytics/yolov8[DB/OL]. https：//github. com/ultralytics/yolov8.

[28] WANG A，CHEN H，LIU L，et al. Yolov10：Real-time end-to-end object detection[J]. Advances in Neural Information Processing Systems，2024，37：107984-108011.

JI Tao，born in 1999，postgraduate. His main research interests include object detection and embedded system.

YANG Yifan，born in 1986，Ph.D，associate professor，master supervisor. His main research interests include embedded edge intelligent computing，image enhancement，object recognition and tracking.