SCFSA-YOLO：遮挡下精准客流统计的实时头肩检测器

王 浩[1]
WANG Hao

摘 要 密集场景下的客流统计面临小目标信息量匮乏、目标遮挡严重的现实挑战，导致现有检测方法难以满足实际应用需求。为此，文章提出一种改进的头肩检测模型 SCFSA-YOLO，该模型以 YOLOv12 为基础框架，通过针对性优化小目标检测性能，有效提升复杂密集环境下客流统计的准确性。首先，选用头肩作为检测目标以缓解遮挡问题，并在 VSCrowd 数据集上完成模型训练；其次，通过增设小目标专用检测层、加深主干网络结构，以及将输入分辨率从 640 提升至 960，显著增强小目标的特征表征能力；此外，设计空间多尺度 - 通道频率协同注意力机制（SCFSA），提升模型在遮挡环境下的空间定位精度；同时，提出增强型空间 - 光谱下采样模块（ESSD），替代传统步长卷积，确保下采样过程中原始特征信息的最大化保留。实验结果表明，所提方法在 VSCrowd 测试集上的 mAP@0.5 达到 90.0%，较基准 YOLOv12 模型提升 8.6%，在小目标检测和遮挡场景中展现出明显优势。综上所述，SCFSA-YOLO 模型通过架构优化与创新的注意力机制、下采样，有效解决了密集场景中小目标表征不足和严重遮挡的核心难题，为高精度、实时化的客流统计应用提供了切实可行的技术方案。

关键词 头肩检测；密集场景；YOLOv12；小目标检测；注意力机制；客流统计

doi：10.3969/j.issn.1672-9528.2026.01.010

0 引言

随着社会的发展，出行旅游、购物越来越普遍，公共场所如车站、商场、景区等人员聚集现象增多，密集人群带来的安全隐患不容忽视。据统计，近年来全球范围内因人群拥挤导致的踩踏事故屡见不鲜，造成大量人员伤亡。因此，实时精准的客流统计与密度分析对于公共安全管理和突发事件预警具有至关重要的意义。

传统的客流统计方法主要依赖人工观察或简单的传感器

1. 北京数原数字化城市研究中心 北京 100083

参考文献：

[1] 向彬, 陈昂 . 我国中小学阶段书法教育的现状 [J]. 中国艺术, 2017(3): 16-21.

[2] 刘宏烽 , 张路遥 . 人工智能背景下书法教学数字化转型：动因、挑战与进路 [J]. 宿州职业技术学院学报 , 2025(4): 56-62.

[3] 温佩芝 , 姚航 , 沈嘉炜 . 基于卷积神经网络的石刻书法字识别方法 [J]. 计算机工程与设计 , 2018, 39(3):867-872.

[4] 郭佳霖 , 智敏 , 殷雁君 , 等 . 图像处理中 CNN 与视觉 Transformer 混合模型研究综述 [J]. 计算机科学与探索 , 2025, 19(1): 30-44.

[5] 胡正南 , 胡立坤 . 基于 Vision Transformer 多模型融合的视觉闭环检测算法 [J]. 激光杂志 ,2024,45(6):75-81.

[6] 李玉斌 , 宋金玉 , 姚巧红 . 游戏化学习方式对学生学习效果的影响研究 : 基于 35 项实验和准实验研究的元分析 [J]. 电化教育研究 ,2019,40(11):56-62.

[7] 胡智丹 , 王志军 , 王萌 , 等 . 基于深度学习的汉字硬笔楷书智能评测系统的设计与应用 [J]. 数字教育 , 2025, 11(3): 30-36.

[8] 肖政文 , 刘丹 . 基于优化 U-net 的单色图像分割算法 [J]. 信息记录材料 , 2025,26(1):204-207.

[9] 汪嘉珮 . 基于微服务架构的在线教育系统设计与实现 [D]. 武汉 : 华中科技大学 , 2020.

[10] 刘子扬 , 王朝坤 , 章衡 . 图对比学习方法综述 [J]. 软件学报 , 2026(1): 180-199.

【作者简介】

刘宏烽（1981—），男，广东韶关人，硕士，副高，研究方向：信息技术、书法教学。

张路遥（1983—），男，辽宁辽阳人，硕士，工程师，研究方向：软件技术、人工智能。

（收稿日期：2025-11-06 修回日期：2026-01-12）

技术，难以在复杂场景中实现准确计数。近年来，基于计算机视觉的检测方法逐渐成为研究热点，尤其是在深度学习技术的推动下，目标检测算法在精度和效率方面取得了显著进展。然而，密集场景下的客流统计仍面临两大技术挑战：一是小目标（如远处的人头）在图像中占据像素少，特征表征能力不足；二是严重遮挡导致目标信息不完整，检测难度大。

针对上述问题，当前研究主要分为两种思路：一是基于整体人体的检测，但在密集场景中，人体相互遮挡严重，效果受限；二是基于人头检测，虽然减轻了遮挡问题，但小目标特性使得检测依然困难。作为折中，头肩检测在近年来受到关注，因为头肩部位在密集场景中仍保持较高的可见性，同时相比单独的人头具有更丰富的特征信息。

在检测模型方面，YOLO 系列因其在速度与精度上的平衡而广泛应用。Tian 等 [1] 最新提出的 YOLOv12 通过引入区域注意力机制和残差高效聚合网络，在保持推理速度的同时提升了检测精度。然而，在面对密集小目标场景时，YOLOv12 仍有提升空间，特别是对小目标的特征提取和遮挡情况的处理能力。

本文基于 YOLOv12 模型进行改进，主要贡献如下：

（1）提出一种专门针对密集场景的头肩检测方法，在 VSCrowd 数据集上进行验证；

（2）通过增加小目标检测层、加深主干网络和提高输入分辨率增强小目标特征表征；

（3）设计一种新型的空间多尺度 - 通道频率协同注意力机制 SCFSA，提升模型在遮挡环境下的空间定位能力；

（4）设计一种新型的下采样模块，在下采样的同时尽量保留原始信息；

（5）在密集场景客流统计任务上验证了方法的有效性，为实际应用提供技术支撑。

1 相关工作

1.1 密集人群检测

密集人群检测是计算机视觉领域的一项挑战性任务。早期方法主要基于整体人体检测，如 R-CNN 系列和 YOLO 系列算法。然而，在高密度场景中，人体之间的严重遮挡使得这些方法难以准确检测每一个个体。为此，研究人员转向基于头部检测的方法，因为头部在密集人群中通常保持可见性。但头部作为小目标，检测难度较大，特别是在低分辨率监控视频中。

近年来，头肩检测作为一种折中方案被提出。头肩区域比头部包含更多特征信息，同时相比整个人体受遮挡的影响较小。VSCrowd[2] 数据集推动了这一领域的发展，该数据集包含超过 60 000 帧高清监控视频，覆盖多种场景，如商场、街道和景点，推动视频人群分析技术的发展，特别是在复杂场景下的人群定位。

1.2 小目标检测技术

小目标检测（small object detection, SOD）是计算机视觉中一项长期存在的挑战，尤其在密集人群、遥感图像和无人机航拍等场景中尤为突出。近年来，研究者围绕多尺度特征融合、信息保留型下采样和注意力机制等方向，提出了大量有效方法。

1.2.1 多尺度特征融合

由于小目标在深层网络中容易因多次下采样而丢失细节信息，多尺度特征融合成为提升检测性能的核心策略。代表性工作包括改进的特征金字塔网络（如 HRFPN[3]）以及跨阶段特征融合结构（如 MCFN[4]）。例如，Hu 等 [5] 提出的 MFF-YOLOv8 通过引入多尺度特征融合模块 [6]，在无人机航拍图像的小目标检测任务中显著提升了召回率与定位精度。此外，一些方法结合边缘信息与特征校准机制，进一步增强小目标的语义表达能力。

1.2.2 信息保留型下采样

在小目标检测领域，尤其是面向无人机航拍图像的场景中，如何在特征提取与下采样过程中有效保留原始空间细节信息，已成为提升检测性能的关键挑战。传统基于特征金字塔网络（FPN）的方法通常依赖高层语义特征与浅层空间特征的融合，但在下采样过程中，常规的步长卷积或池化操作易导致小目标的细粒度信息严重丢失，进而引发漏检与误检。为缓解这一问题，近期研究开始聚焦于信息保留机制的设计。例如，PRNet[7] 通过引入增强型 SliceSamp（ESSamp）模块，在下采样阶段对浅层特征进行重排与卷积优化，以最大限度保留原始空间信息，并结合渐进优化颈部（PRN）实现空间细节与语义特征的对齐。类似地，IF-YOLO[8] 提出了信息保留特征聚合（IPFA）模块，旨在构建语义表示的同时避免小目标固有特征的退化，并通过冲突信息抑制特征融合模块（CSFM）与细粒度聚合特征金字塔网络（FGAFPN）优化多尺度融合过程，减少因直接特征拼接引入的语义冲突。这些工作共同表明，面向小目标检测的下采样策略不应仅追求计算效率，更需兼顾原始信息的完整性与特征融合的语义一致性，从而为后续检测头提供更具判别力的特征表示。

1.2.3 注意力机制

注意力机制通过动态调整特征权重，使模型聚焦于关键区域和通道，在目标检测中发挥着日益重要的作用。

通道注意力（如 SE 模块）通过建模通道间依赖关系，增强重要通道的响应；空间注意力则强调特征图中的关键位置。近年来，全局注意力机制（global attention mechanism, GAM[9]）被广泛应用于小目标检测。GAM 通过捕获全局上下文并自适应调整特征权重，在火焰、无人机等小目标场景

中验证了其有效性，有研究将 GAM 嵌入 YOLOv8，显著提升了小目标检测率。此外，并行补丁感知注意力（parallelized patch-aware attention, PPA）模块通过动态调整感受野，在红外小目标检测中有效防止下采样过程中的信息丢失。该模块已被集成至 SPDC-YOLO[10]、HCF-Net[11] 等网络中，用于增强多尺度特征提取能力。

针对复杂遮挡环境，DLGADet[12] 提出可变形局部与全局注意力融合机制，在交通场景中显著提升空间定位鲁棒性。同时，尺度选择策略也受到关注：部分工作仅在特定网络层部署注意力模块，以避免冗余计算并提升小目标召回率。

综上，当前小目标检测技术正朝着多尺度特征融合、信息保留型下采样和注意力机制的方向持续演进，为密集人群分析等实际应用奠定了坚实基础。

2 方法

YOLOv12 是 YOLO 系列的最新演进版本，其主要创新在于引入了区域注意力机制（Area Attention）和残差高效聚合网络（R-ELAN）。

区域注意力机制：将特征图按水平或垂直方向分块，然后在每个块内计算自注意力。这种方法将计算复杂度从 $O(n^2 \cdot d)$ 降低到 $O(0.5n^2 \cdot d)$，其中 n 为序列长度；d 为特征维度。对于 640 px×640 px 的输入，这一改进使得注意力机制能够实时计算，同时保持较大的感受野。

残差高效聚合网络（R-ELAN）：在 YOLOv7[13] 的 ELAN 结构基础上引入残差连接和块级别缩放，解决了深层网络中的梯度阻塞问题，同时提高了特征多样性。

YOLOv12 的检测头采用多尺度预测，在 3 个不同尺度的特征图上进行检测：80 px×80 px、40 px×40 px 和 20 px×20 px，分别负责小、中、大目标的检测。

然而，在密集头肩检测任务中，发现 YOLOv12 存在以下局限性：

（1）即使是最浅层的 80 px×80 px 特征图，对于极端小目标（小于 16 px×16 px）的特征表征仍然不足；

（2）区域注意力机制在严重遮挡情况下的空间定位能力有限。

针对密集场景头肩检测的特殊需求，从两个方面对 YOLOv12 进行了改进：一是增强小目标特征表征能力，二是提升空间表达能力以应对遮挡问题。整体架构图如图 1 所示。

2.1 小目标检测层与主干网络加深

原有 YOLOv12 采用三检测头结构，对应特征图大小为 20 px×20 px、40 px×40 px 和 80 px×80 px。为了提升小目标检测能力，增加了一个更高分辨率的检测层，并加深了网络的深度，将输入尺寸由原来的 640 提高到 960，这样输出的特征图大小变成了 120 px×120 px、60 px×60 px、30 px×30 px、15 px×15 px。120 px×120 px 检测层由主干网络中较浅层的特征图生成，保留更多空间细节信息，专门用于检测 16 px×16 px 像素以下的极小头肩目标。新增检测层的实现方式如下：

（1）从主干网络的第 4 层（较浅层）提取特征图；

（2）通过一组卷积层减少通道数至 256，同时保持空间分辨率；

（3）使用最近邻上采样将特征图分辨率提高至 120 px×120 px；

（4）通过加入小目标层，同时降低其他层的特征图尺寸，参数量增加了 15% 左右（以 l 模型的参数计算），由于其他特征图的尺寸降低，训练和推理的速度并没有产生显著的影响。

2.2 空间多尺度 - 通道频率协同注意力机制（SCFSA）

为有效增强小目标在复杂背景下的特征表达能力，本文

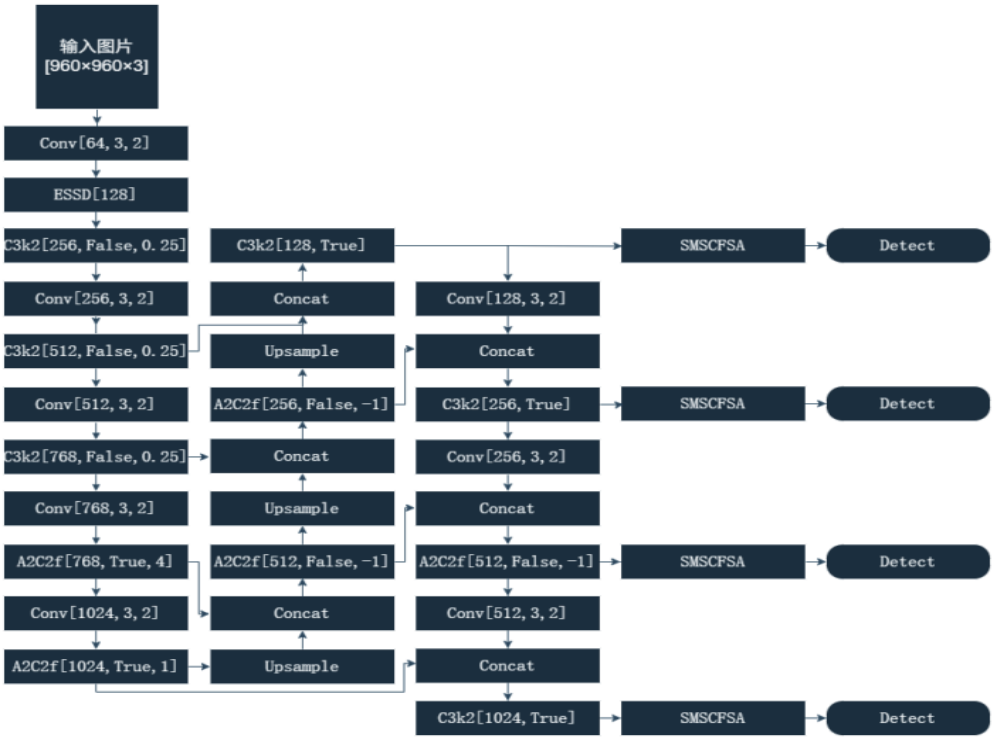

图 1 SCFSA-YOLO 的整体架构图

提出一种轻量级但高效的 空间 - 通道协同注意力机制（spatial multi-scale and channel frequency synergistic attention, SCFSA）。该机制通过联合建模空间维度上的多尺度上下文依赖与通道维度上的频率感知响应，实现对关键区域与重要通道的协同聚焦。

如图 2 所示，给定输入特征图 $\boldsymbol{X}\in\mathbb{R}^{B\times C\times H\times W}$，首先引入可学习的二维位置编码以保留空间结构信息：

$$\boldsymbol{X}'=\boldsymbol{X}+P(\boldsymbol{X}) \tag{1}$$

式中：$P(\cdot)$ 是由两个线性投影构成的位置编码模块，将归一化坐标映射至特征维度。

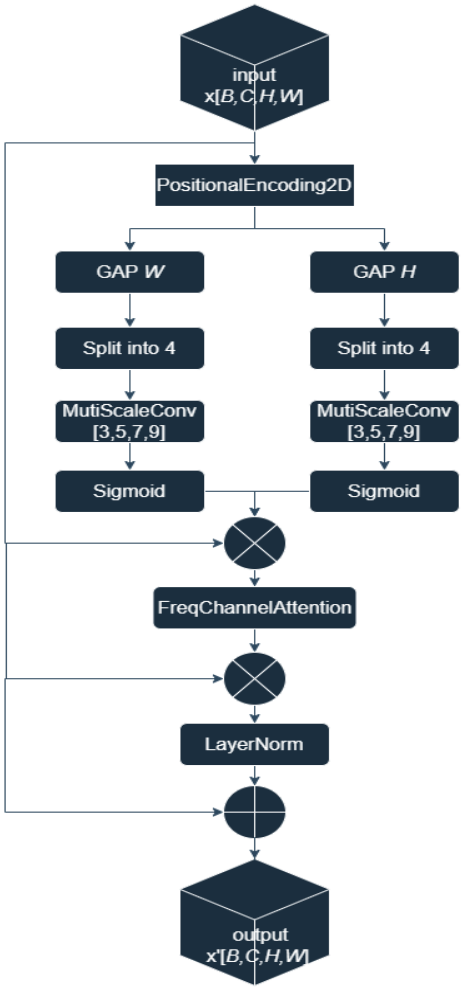

图 2 SCFSA 注意力机制的结构图

随后，沿高度和宽度方向分别进行全局平均池化，得到一维空间统计量：

$$\boldsymbol{X}_h=\frac{1}{W}\sum_{w=1}^{W}\boldsymbol{X}'_{:,:,:,w}\in\mathbb{R}^{B\times C\times H}$$
$$\boldsymbol{X}_w=\frac{1}{H}\sum_{h=1}^{H}\boldsymbol{X}'_{:,:,h,:}\in\mathbb{R}^{B\times C\times W} \tag{2}$$

为提升多尺度感知能力，将通道维度划分为四组（要求 C 可被 4 整除），每组独立通过 多尺度深度卷积模块（MultiScaleConv）：

$$\text{MultiScaleConv}(z)=\text{LayerNorm}\left(\frac{1}{K}\sum_{k\in K}\text{DWConv}_k(z)\right) \tag{3}$$

式中：K={3, 5, 7, 9} 表示卷积核尺寸集合；DWConv 表示深度可分离卷积，以低计算代价捕获不同感受野下的空间模式。

经分组处理后，拼接各组输出并施加门控函数（默认为 Sigmoid）生成空间注意力权重：

$$\boldsymbol{A}_h=\sigma(\text{Cat}([\text{MSConv}_i(\boldsymbol{X}_h^{(i)})]_{i=1}^{4}))\in\mathbb{R}^{B\times C\times H} \tag{4}$$

$$\boldsymbol{A}_w=\sigma(\text{Cat}([\text{MSConv}_i(\boldsymbol{X}_w^{(i)})]_{i=1}^{4}))\in\mathbb{R}^{B\times C\times W} \tag{5}$$

式中：$\sigma(\cdot)$ 为激活函数；$\boldsymbol{X}_h^{(i)}$为第 i 组通道切片。

最终空间注意力图通过外积形式融合：

$$M_s=\boldsymbol{A}_h\otimes\boldsymbol{A}_w\in\mathbb{R}^{B\times C\times H\times W} \tag{6}$$

与此同时，引入 频域启发的通道注意力模块（FreqChannelAttention），基于全局平均池化与瓶颈结构建模通道间依赖：

$$M_c=\sigma\left(\boldsymbol{W}_2\,\phi\left(\boldsymbol{W}_1\,\text{GAP}(\boldsymbol{X}')\right)\right) \tag{7}$$

式中：ϕ 为 SiLU 激活函数；$\boldsymbol{W}_1\in\mathbb{R}^{(C/r)\times C}$、$\boldsymbol{W}_2\in\mathbb{R}^{C\times(C/r)}$ 为降维与升维卷积核（r=4）。

最终，特征图经空间与通道注意力协同调制，并通过 LayerNorm 与残差连接稳定训练：

$$Y=\text{LayerNorm}\left((\boldsymbol{X}'\odot M_s\odot M_c)^{\top}\right)^{\top}+\boldsymbol{X} \tag{8}$$

该设计在保持低参数量的同时，实现了多尺度空间建模、位置感知增强与通道频率响应优化的有机统一，特别适用于小目标检测中因分辨率受限导致的语义模糊问题。

2.3 信息保留型下采样模块（ESSD）

在密集场景的小目标检测中，特征图在主干网络和颈部（Neck）的多次下采样过程是导致小目标信息丢失的关键因素。传统的步长为 2 的卷积或池化操作虽然能有效降低计算复杂度，但会不可逆地丢弃大量高频细节信息，这对于本身就缺乏像素信息的小目标而言是灾难性的。

为了解决这一问题，提出了一种信息保留型下采样模块（enhanced spatial-spectral downsampling, ESSD），用于替代 YOLOv12 中原有的步长卷积下采样操作。如图 3 所示，ESSD 通过并行融合空间路径和光谱路径，从两个互补的维度对输入特征进行下采样，从而最大限度地保留原始信息。

具体而言，给定输入特征图$\boldsymbol{x}\in\mathbb{R}^{B\times C_1\times H\times W}$，ESSD 模块首先通过一个残差连接对其进行处理：

$$\boldsymbol{x}_{\text{res}}=\text{Conv}_{1\times1,s=2}(\boldsymbol{x}) \tag{9}$$

该残差路径确保了下采样后的特征图尺寸为 [B, C_2, H/2, W/2]，为后续的特征融合提供基准。

ESSD 的核心在于其双路径设计：

（1）空间下采样路径：该路径采用 PixelUnshuffle(2) 操

作，将输入特征图的空间维度 $H\times W$ 重排为通道维度，得到一个形状为 $[B, 4C_1, H/2, W/2]$ 的张量。这种操作是完全可逆的，不会丢失任何原始像素信息。随后，该张量通过一个增强型深度可分离卷积（EnhancedDSConv）进行通道压缩和特征提炼。EnhancedDSConv 在标准深度可分离卷积的基础上引入了 SE 注意力机制，能够自适应地校准通道权重，进一步强化关键特征。

（2）光谱下采样路径：该路径采用传统的步长为 2 的卷积进行下采样，但同样使用 EnhancedDSConv 模块。为了增强该路径对不同尺度上下文的感知能力，在其后接入了一个多尺度融合（MultiScaleFusion）模块。该模块并行使用 3×3 和 5×5 的深度卷积核提取特征，并通过可学习的权重进行自适应融合，从而捕获更丰富的频谱信息。

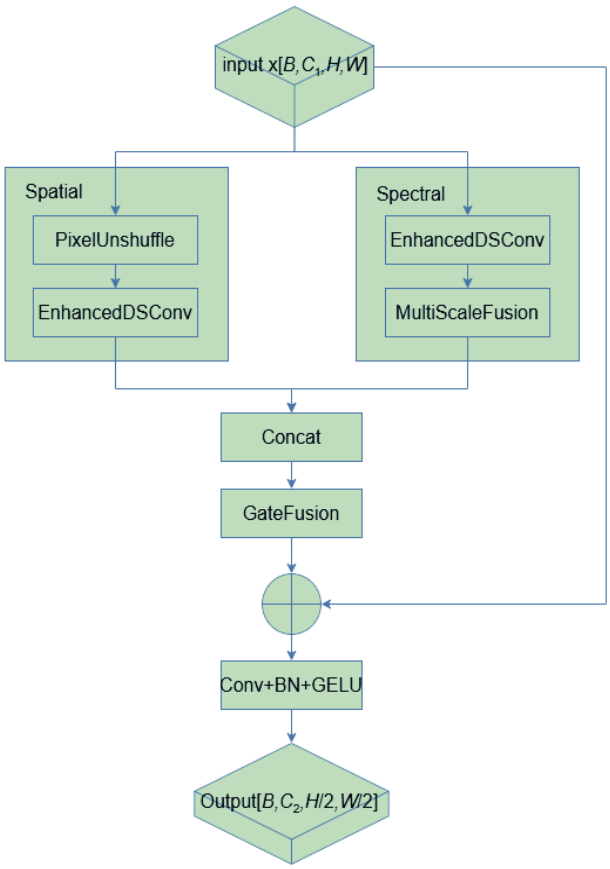

图 3 ESSD 模块结构图

两条路径的输出特征图在通道维度上被拼接（或通过门控机制融合），形成最终的下采样特征。该特征与残差连接 $\boldsymbol{x}_{\text{rec}}$ 相加，并通过一个 3×3 卷积层进行最终的特征融合与精炼。

通过这种设计，ESSD 模块巧妙地结合了无损空间重排和多尺度光谱感知的优势。空间路径确保了原始像素信息的完整性，而光谱路径则提供了必要的语义抽象和上下文信息。两者协同工作，有效缓解了传统下采样操作带来的信息损失问题，为后续的检测头提供了更具判别力和细节信息的特征表示，尤其有利于小目标和遮挡目标的精准定位。

3 实验

3.1 实验设置

3.1.1 数据集与评估指标

使用 VSCrowd 数据集进行实验，按照官方划分使用训练集和验证集。为全面评估模型性能，采用以下评估指标：

mAP@0.5：交并比（IoU）阈值为 0.5 时的平均精度；

mAP@0.5-0.95：IoU 阈值从 0.5 到 0.95（步长 0.05）的平均精度。

3.1.2 实现细节

实验基于 PyTorch 框架，使用 NVIDIA RTX L40 GPU 进行训练。优化器选择 SGD，初始学习率设为 0.01，训练周期为 150 轮，批次大小为 128。数据增强包括 Mosaic、随机色彩抖动、随机旋转（-15°~15°）和随机遮挡。

对比模型包括 YOLOv12 基准版本以及当前主流检测模型：YOLOv8 和 YOLO11。所有对比模型均使用相同数据集和训练策略进行微调，确保公平比较。

3.2 消融实验

为系统评估所提出各组件的有效性，设计了一系列消融实验，结果如表 1 所示。

首先，在 640 px×640 px 输入分辨率下对原始 YOLOv12 系列模型（包括 n、s、m、l 四个尺度）进行基准测试。实验结果表明，随着模型规模的增大，mAP@0.5 的提升逐渐趋于饱和；尤其值得注意的是，YOLOv12l 的 mAP@0.5（81.4%）反而略低于 YOLOv12m（81.6%），下降了 0.2%。这一现象说明，单纯增加模型容量已难以带来性能增益。

随后，将输入分辨率提升至 960 px×960 px。以 YOLOv12l 为例，其 mAP@0.5 显著提升了 5.0%（从 81.4% 增至 86.4%），验证了高分辨率输入对检测性能，尤其是小目标检测能力的积极影响。因此，在此基础上对网络架构进行了针对性优化：一方面加深主干网络，另一方面新增专用于小目标检测的特征层。经此改进后，YOLOv12l 在 960 px×960 px 分辨率下的 mAP@0.5 进一步提升 1.1%（达到 87.5%），表明所提出的架构优化策略有效增强了模型对小目标的感知能力。

为进一步提升性能，引入了自研的 SCFSA 注意力机制。实验结果显示，在架构优化的基础上加入 SCFSA 后，YOLOv12l 的 mAP@0.5 再次提升 1.4%（达到 88.9%），验证了该注意力机制在增强关键特征表达方面的有效性。

最后，采用 ESSD 模块替代传统步长为 2 的卷积下采样操作。ESSD 通过保留更多原始空间信息，有效缓解了深层网络和 Neck 部分的信息损失。引入 ESSD 后，YOLOv12l 的 mAP@0.5 显著提升至 90.0%，mAP@0.5-0.95 也达到 51.9%，分别较前一版本提升 1.1% 和 1.0%。这一结果充分证明，信息保留型下采样策略对于提升整体检测精度，特别是小目标检测性能具有重要作用。

表 1 消融实验结果对比

实验	分辨率	mAP50/%	mAP50-95/%
YOLOv12n	640	73.4	36.1
YOLOv12s	640	77.4	39.7
YOLOv12m	640	81.6	43.7
YOLOv12l	640	81.4	44.0
YOLOv12n	960	80.3	41.5
YOLOv12s	960	83.7	45.3
YOLOv12m	960	85.4	47.9
YOLOv12l	960	86.4	48.9
YOLOv12n + 架构优化	960	81.4	42.6
YOLOv12s + 架构优化	960	83.7	46.2
YOLOv12m + 架构优化	960	86.6	48.9
YOLOv12l + 架构优化	960	87.5	50.0
YOLOv12n + 架构优化 +SCFSA	960	82.9	43.3
YOLOv12s + 架构优化 +SCFSA	960	86.4	47.8
YOLOv12m + 架构优化 +SCFSA	960	88.4	50.1
YOLOv12l + 架构优化 +SCFSA	960	88.9	50.9
YOLOv12n + 架构优化 +SCFSA+ESSD	960	83.8	44.6
YOLOv12s + 架构优化 +SCFSA+ESSD	960	86.6	48.1
YOLOv12m + 架构优化 +SCFSA+ESSD	960	88.7	51.0
YOLOv12l + 架构优化 +SCFSA+ESSD	960	90.0	51.9

综上所述，各项改进组件均对模型性能产生了积极贡献，且具有良好的协同效应。

3.3 对比实验

本文将完整模型与当前主流检测方法进行对比，结果如表 2~3 所示。

表 2 本文方法和其他模型的比较

方法	mAP50/%
FasterR	66.2
FCOS	40.1
GLoss	60.5
RAZNet	77.4
STANet	73.4
P2PNet	58.6
GNANet	80.5
Ours	90.0

（注：数据来自文献 [5]）

在所有对比模型中，本文方法在 mAP@0.5 和 mAP@0.5-0.95 上均达到最高水平，分别达 90.0% 和 51.9%。表 2 是本文方法的结果和文献 [5] 比较，选用每个方法的最高值。

表 3 在同等条件下测试的结果，测试 YOLOv8 和 YOLO11 等主流的 YOLO 系列检测模型。

表 3 本文方法和 YOLO 系列模型的比较

实验	分辨率	mAP@50/%	mAP@50-95/%
YOLOv8l	640	82.5	44.8
	960	86.6	49.1
YOLO11l	640	82.2	45.0
	960	86.5	49.7
Ours	640	85.1	46.9
	960	90.0	51.9

值得注意的是，本文方法在 VSCrowd 数据集上 mAP50 达到 90.0%，显著高于其他对比方法，这证明所提出的空间 - 通道协同注意力机制和信息保留型下采样模块在解决遮挡问题以及密集小目标上的有效性。同时，尽管输入分辨率提高且模型加深，通过结构优化，本文方法仍保持了可接受的推理速度，满足大多数客流统计场景的实时性要求。

3.4 可视化分析

为了更直观地展示改进效果，对检测结果进行了可视化分析。在简单场景中，所有方法均能较好地检测头肩目标，但在复杂密集场景中，差异明显。

如图 4 所示，基线 YOLO11 在密集区域出现大量漏检，尤其是小目标和严重遮挡目标。增加小目标检测层，引入 SCSA 注意力机制后，模型在遮挡区域的检测精度显著提升，能够检测到更多部分可见的头肩目标。SCFSA-YOLO 模型在各类场景中均表现稳定，特别是在背景复杂、目标密集、严重遮挡的挑战性场景中，仍能保持较高的检测率。

（a）YOLOv8l 的测试结果

（b）YOLO11l 的测试结果

（c）本文 SCFSA-YOLO 的测试结果

图 4 SCFSA-YOLO 与 YOLOv8l、YOLO11l 测试结果对比

4 结论与未来工作

本文针对密集场景客流统计的需求，提出了一种基于 YOLOv12 改进的头肩检测方法。通过增加小目标检测层、加深主干网络、设计空间 - 通道协同注意力机制和信息保留型下采样模块，有效解决了小目标表征不足和遮挡严重的问题。

在 VSCrowd 数据集上的实验证明，本文方法在保持实时性的同时，显著提升了检测精度，特别是在小目标和遮挡场景下的表现。消融实验验证了各改进组件的有效性，对比实验表明本文方法优于当前主流检测模型。

从技术层面看，本研究的主要创新点在于：（1）针对密集场景特点，系统性地优化了 YOLOv12 架构，特别关注小目标检测能力；（2）提出了一种新颖的空间 - 通道协同注意力机制，通过门控单元自适应平衡两种注意力，提升了遮挡目标的检测能力；（3）验证了头肩检测在密集场景客流统计中的实用价值，为相关应用提供了技术参考。在实际应用方面，本研究的方法可以部署于车站、商场等重点区域的监控系统，实现实时精准的客流统计与密度分析，为公共安全管理和商业运营提供数据支持。

未来工作将从以下几个方向展开：（1）探索更多模态数据，如深度信息，进一步提升遮挡场景下的检测精度；（2）研究自适应计算机制，根据场景复杂度动态调整模型，平衡精度与效率；（3）扩展头肩检测的时序建模，利用视频连续帧提升检测稳定性；（4）探索跨摄像头跟踪技术，实现大范围区域的客流运动分析。通过持续优化，密集场景下的客流统计技术将在智慧城市、公共安全等领域发挥更大价值。

参考文献：

[1] TIAN Y J, YE Q X, DOERMANN D. YOLOv12: attention-centric real-time object detectors[EB/OL].(2025-08-18)[2025-12-22].https://www.semanticscholar.org/paper/YOLOv12%3A-Attention-Centric-Real-Time-Object-Tian-Ye/ae1d5360f2f556139cffd10d6e9d2e0241c937e0.

[2] LI H P, LIU L B, YANG K L,et al.Video crowd localization with multifocus gaussian neighborhood attention and a large-scale benchmark[EB/OL]. (2022-08-08)[2025-12-30].https://doi.org/10.48550/arXiv.2107.08645.

[3] JIANG X Y, ZHANG X H, GAO N,et al. When fast fourier transform meets transformer for image restoration[C]//Computer Vision–ECCV 2024. Berlin:Springer,2024:381-402.

[4] WU Y F, WANG B W, WANG G L,et al. MCFN: multi-scale crossover feed-forward network for high performance watermarking[J].Neurocomputing, 2025,622:129282.

[5] HU K, LU J Z, ZHENG C Q, et al. MFF-YOLOv8: small object detection based on multi-scale feature fusion for UAV remote sensing images[J]. IET image processing, 2025,19(1):70066 .

[6] GIRMA A, HOMAIFAR A, MAHMOUD M N. ADP-Net: adaptive point network with multi-scale attention mechanism for small object detection[J].Neurocomputing, 2026,659:131381.

[7] ZHENG P H, ZHAO Y L, CUI Z, et al. PRNet: original information is all you have[EB/OL].(2025-10-10)[2025-12-22]. https://doi.org/10.48550/arXiv.2510.09531.

[8] LI Y, CHU C F, ZHANG Q, et al.IF-YOLO: an efficient and accurate detection algorithm for insulator faults in transmission lines[J]. IEEE access,2021,12:167388-167403.

[9] ZHENG Y C, JING Y X, ZHAO J F,et al.LAM-YOLO: drones-based small object detection on lighting-occlusion attention mechanism YOLO[J]. Computer vision and image understanding, 2025,261:104489.

[10] BI J X, LI K D, ZHENG X Y,et al.SPDC-YOLO: an efficient small target detection network based on improved YOLOv8 for drone aerial image[J].Remote sensing, 2025, 17(4), 685.

[11] XU S B, ZHENG S C, XU W H, et al. HCF-Net: Hierarchical context fusion network for infrared small object detection [EB/OL].(2024-03-16)[2025-12-30]. https://doi.org/10.48550/arXiv.2403.10778.

[12] ZHANG Y F, YU J Y, WANG Y Y,et al.Small object detection based on hierarchical attention mechanism and multi-scale separable detection[J].IET image processing, 2023,17(4):3986-3999.

[13] WANG C-Y, BOCHKOVSKIY A, MARK LIAO H-Y. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C/OL]// 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).Piscataway:IEEE,2022[2025-12-23].https://ieeexplore.ieee.org/document/10204762.DOI: 10.1109/CVPR52729.2023.00721.

【作者简介】

王浩（1989—），男，河北邢台人，本科，高级算法工程师，研究方向：计算机视觉。

（收稿日期：2025-11-06 修回日期：2026-01-06）