http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2024.0561

# 基于状态空间模型的航拍图像小目标检测算法

滕尚志*，龚曦，李宝安，吕学强
(北京信息科技大学 网络文化与数字传播北京市重点实验室，北京 100192)

**摘 要**：小目标检测旨在解决传统目标检测算法处理微小目标时的困难。小目标检测需要克服目标尺寸小、像素少、目标与背景之间差异小导致难以检测的问题，对目标检测算法的鲁棒性、准确性提出更高的要求。基于此，设计了一种基于状态空间模型的小目标检测算法，探索状态空间模型在目标检测领域的应用。设计了滑动窗口扫描机制以将二维图像转换为一维序列，更适用于小目标检测任务；设计了空间信息聚合模块，通过双分支结构分别计算全局和局部注意力，抑制无关背景信息；针对传统交并比损失进行改进，关注边界框形状和尺度对检测性能的影响，并使用辅助框加速模型收敛。在公开数据集 VisDrone2019 上进行了充分的实验，论证了所提算法的有效性。

**关 键 词**：小目标检测；状态空间模型；空间信息聚合；注意力机制；辅助边界框
**中图分类号**：TP391; TP751; V557
**文献标志码**：A **文章编号**：1001-5965(2026)11-3832-10

目标检测是对图像、视频中感兴趣的目标进行分类和定位的一项重要任务。由于深度卷积神经网络强大的学习能力和庞大的数据量，近年来目标检测取得了显著的成就。小目标检测(small object detection, SOD)作为通用目标检测的一个子领域，专注于对小尺寸物体的检测，在监控、无人机场景分析、行人检测、自动驾驶交通标志检测等多种场景中具有重要的理论和现实意义。从定义上理解，小目标被分为相对小目标和绝对小目标，相对小目标即相对原图目标尺寸为其 1/10 或更小，绝对小目标指目标面积小于 32 像素×32 像素。随着人工智能的发展，无人机广泛应用于航拍、城市交通、遥感测绘、电力线巡航等领域。无人机作为一种高效的图像采集系统，具有高智能、高机动性和广阔视野等优点，在大范围、远距离搜索目标这一新兴领域得到了广泛的应用。在无人机视角下检测地面物体是一个非常具有挑战性的任务，因为从无人机视角下获得的图像中大多数目标的尺度小，信噪比低，容易受到背景的干扰。由于无人机运行过程中飞行高度的不断改变，极易造成拍摄物体尺度变化剧烈、采集图像模糊等问题。此外，拍摄广度的改变可能会导致出现背景复杂、遮挡目标无法识别等问题。针对上述问题，航拍图像小目标检测具有很高的研究与应用价值。

基于深度学习的目标检测算法根据算法的流程主要分为 2 类：两阶段和单阶段。两阶段目标检测的代表性算法有基于区域的卷积神经网络(region-based convolutional neural network, R-CNN)[1]、Fast R-CNN[2]、Faster R-CNN[3] 等。两阶段模型的算法流程主要是先选择出若干目标候选区域，再对候选区域进行进一步的分类以及边界框调整，尽管能够得到较高的准确率，但是由于存在大量的冗余计算导致推理速度较慢。单阶段目标检测的代表性算法有 YOLO[4] 系列、单次多边框检测器(single shot multibox detector, SSD)[5]、RetinaNet[6] 等。单阶段模型的算法流程主要是在图像上直接进行目标的

**收稿日期**：2024-07-23；**录用日期**：2024-09-14；**网络出版时间**：2024-10-14 16:19
**网络出版地址**：link.cnki.net/urlid/11.2625.V.20241014.1513.003
**基金项目**：国家自然科学基金 (62202061)；北京市自然科学基金 (4232025)；北京市教育委员会科研计划 (KM202311232002)
* **通信作者**. E-mail：tengshangzhi@bistu.edu.cn

*引用格式：滕尚志，龚曦，李宝安，等. 基于状态空间模型的航拍图像小目标检测算法 [J]. 北京航空航天大学学报，2026，52（11）：3832-3841.*
*Teng S Z，Gong X，Li B A，et al. Aerial image small object detection algorithm based on state space model[J]. Journal of Beijing University of Aeronautics and Astronautics，2026，52（11）：3832-3841（in Chinese）.*

分类及调整边界框, 大大减少了计算量, 能够在保证一定准确率的前提下大幅度地提高检测速度, 因此, 在现实中得到了广泛应用。但在小目标检测领域, 深度卷积神经网络由于其感受野随深度增加而变大, 更着重于图像的全局特征, 往往容易忽略小目标检测所需的局部特征。Transformer[7] 模型最开始是应用在自然语言处理中, 近年来, 学者们逐渐将 Transformer 的应用扩展到计算机视觉领域。DETR[8] 是 Facebook 团队提出的基于 Transformer 的端到端目标检测器, 没有非极大值抑制(non-maximum suppression, NMS)后处理步骤、没有 anchor 等先验知识和约束, 整个由网络实现端到端的目标检测实现, 大大简化了目标检测的框架。DETR 利用 Transformer 的全局建模能力, 直接对图像进行集合预测, 其性能优于传统的基于 CNN 的目标检测器。自 DETR 提出后, Transformer 在计算机视觉领域持续发力, Deformable DETR[9] 将注意力机制视为训练收敛慢的主要原因, 提出了 Deformable Attention, 加快模型收敛速度并降低算法复杂度。SMCA-DETR[10] 提出一种空间调制的协同注意模块, 取代了 DETR 中现有的协同注意力机制, 以克服 DETR 训练收敛速度慢的缺点。REGO-DETR[11] 提出一种基于感兴趣区域的检测改进方法, 以改进 DETR 中的注意力机制。DINO[12] 考虑了正负噪声对象查询, 以使训练收敛更快, 并增强对小目标的检测性能。RT-DETR[13] 设计了一个混合编码器来解耦尺度内和尺度间的同时交互, 提出了交并比 (intersection over union, IoU) 感知查询选择机制, 通过为解码器提供更高质量的初始对象查询来进一步提升性能。这些目标检测模型受限于 Transformer 的二次计算复杂度, 对计算资源和内存消耗较大, 难以处理长序列任务。Sparse Semi-DETR[14] 集成了一个查询细化模块, 以提高对象查询的质量, 显著提高了对小目标和部分模糊目标的检测能力。Salience DETR[15] 提出了分层显著过滤细化, 只对过滤的判别查询执行 Transformer 编码, 以更好地权衡计算效率和精度。

最近, 状态空间模型(state space model, SSM)在处理长序列建模方面表现出了效率上的优势, 为解决深度学习任务中的远程依赖建模提供了另一种选择。与 Transformer 相比, SSM 效率更高, 因为其随序列长度线性扩展, 并且保留了对远程依赖关系建模的优越能力。SSM 进一步将其应用扩展到复杂的视觉任务中, 在这些任务中, 其有望比 CNN 和 ViT[16] 更具有效率上的优势。Mamba[17] 进一步完善了 SSM, 提出一种依赖于数据的 SSM 层, 并构建了一个通用语言模型主干 Mamba, 在大规模真实数据上, 其在多个方面都优于 Transformer, 并且在序列长度上保持线性计算复杂度, 展示了其优秀的性能。这一特点凸显了 Mamba 作为语言建模变革性架构的潜力, 表明 SSM 可以成为 Transformers 在深度学习应用领域的有力竞争者。S4ND[18] 是第一个将 SSM 应用到视觉任务的工作, 并展示了其性能可与 ViT 竞争的潜力, 然而 S4ND 仅仅对 S4[19] 模型进行了简单的扩展, 无法以数据依赖的方式有效地捕获图像信息。ViM[20] 将 Mamba 的成功扩展到视觉任务中, 通过整合双向选择性 SSM 和位置嵌入, 增强了视觉表示学习, 在 ImageNet 分类及密集预测任务上表现出了超越 DeiT[21] 的卓越性能。同时, VMamba[22] 专注于在视觉表示学习中提高效率和可扩展性。VMamba 引入了交叉扫描机制, 以在二维图像空间中实现一维选择性扫描, 在保持全局接受域的同时, 将计算复杂度从二次降低到线性。这一设计及架构的改进, 使得 VMamba 成为了在性能上与主流视觉模型 (如 ResNet[23]、ViT、Swin Transformer[24]) 相匹配甚至超越的有竞争力的主干模型。此外, U-Mamba[25] 解决了在生物医学图像分割中对高效距离依赖建模的需求。通过将 CNN 在局部特征提取方面的优势与 SSM 在捕捉长距离上下文方面的优势相结合, U-Mamba 展示了混合架构在改善生物医学影像任务的分割性能方面的潜力。

基于上述研究, 本文针对无人机视角下的小目标检测任务, 设计了一个基于 SSM 的目标检测算法 DE-SSM。具体工作如下:

1) 设计了基于滑动窗口扫描机制的状态空间模型 SSMFormer, 能够有效对图像全局建模的同时保留小目标的信息。

2) 设计了空间信息聚合模块(spatial information aggregation module, SIAM), 同时捕捉图像全局和局部空间信息。

3) 设计了基于辅助边界框的 Inner-Shape-IoU 损失函数, 在边界框回归 (bounding box regression, BBR) 过程中关注边界框自身的形状与尺度, 提高小目标检测的精确度。

## 1　本文算法

针对目前无人机航拍图像中的小目标检测性能低、漏检、遮挡等问题, 本文基于 RT-DETR 模型进行改进, 提出 DE-SSM, 探索 SSM 在目标检测中的应用, 并且 DE-SSM 在无人机遥感图像目标检测任务中表现出更高的性能和鲁棒性。该模型主要由特征提取主干网络, 混合高效编码器和基于可变形注意力的 Transformer 解码器组成。

模型中的混合高效编码器包括SSMFormer、SIAM和基于CNN的跨尺度特征融合模块(cross-scale convolutional fusion module, CCFM),改进后的模型结构如图1所示。

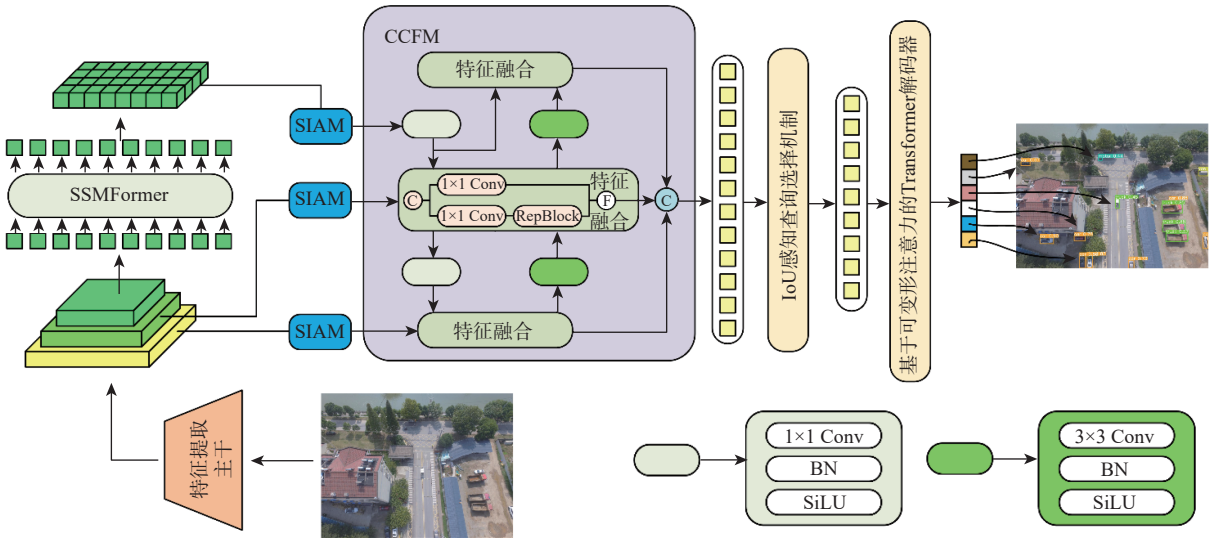


图1 DE-SSM结构

Fig. 1 DE-SSM structure

### 1.1 基于滑动窗口扫描机制的状态空间模型

Transformer在计算机视觉任务中表现出巨大的潜力,普遍的观念认为基于注意力的令牌混合器对Transformer的贡献最大,Yu等[26]从Transformer中抽象出一般架构MetaFormer,认为这种架构才是导致Transformer类视觉模型取得优越性能的关键。

MetaFormer架构如图2(a)所示,基于这种架构,提出SSMfromer,其结构如图2(b)所示,将滑动窗口扫描机制作为MetaFormer的令牌混合器。

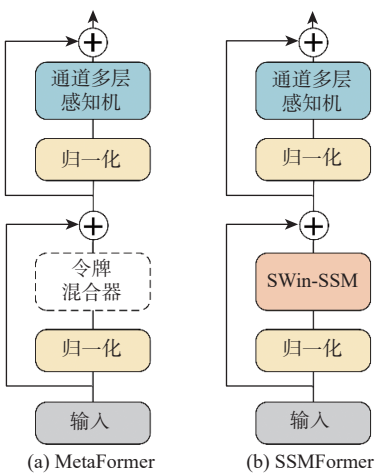


图2 SSMFormer和MetaFormer模型对比

Fig. 2 Comparison of SSMFormer and MetaFormer models

#### 1.1.1 结构化状态空间模型

结构化SSM依赖于经典的连续系统,该系统通过中间隐藏状态$\boldsymbol{h}(t)\in\mathbf{R}^N$将一维序列$\boldsymbol{x}(t)\in\mathbf{R}$映射到$\boldsymbol{y}(t)\in\mathbf{R}$,上述过程可以表示为线性常微分方程:

$$\begin{cases}\boldsymbol{h}'(t)=\boldsymbol{A}\boldsymbol{h}(t)+\boldsymbol{B}\boldsymbol{x}(t)\\ \boldsymbol{y}(t)=\boldsymbol{C}\boldsymbol{h}(t)\end{cases}\tag{1}$$

式中:$\boldsymbol{A}\in\mathbf{R}^{N\times N}$为状态矩阵;$\boldsymbol{B}\in\mathbf{R}^{N\times 1}$、$\boldsymbol{C}\in\mathbf{R}^{1\times N}$为投影参数。

S4是线性时间不变的结构化SSM,通过将该连续系统离散化,使其更适合深度学习场景。具体来说,引入时间尺度参数$\theta$将连续参数$\boldsymbol{A}$、$\boldsymbol{B}$转换为离散参数$\overline{\boldsymbol{A}}$、$\overline{\boldsymbol{B}}$,常用的变换方式是零阶保持(Zero-Order Hold, ZOH),其定义如下:

$$\begin{cases}\overline{\boldsymbol{A}}=\exp(\Delta\boldsymbol{A})\\ \overline{\boldsymbol{B}}=(\Delta\boldsymbol{A})^{-1}(\exp(\Delta\boldsymbol{A})-\boldsymbol{I})\cdot\Delta\boldsymbol{B}\end{cases}\tag{2}$$

式中:$\Delta$为步长;$\boldsymbol{I}$为单位矩阵。

在参数离散化后,式(1)可以重写为

$$\begin{cases}\boldsymbol{h}_t=\overline{\boldsymbol{A}}\boldsymbol{h}_{t-1}+\overline{\boldsymbol{B}}\boldsymbol{x}_t\\ \boldsymbol{y}_t=\boldsymbol{C}\boldsymbol{h}_t\end{cases}\tag{3}$$

模型通过SSM卷积核同时计算合成整个序列的输出,从而提高计算效率与可扩展性:

$$\begin{cases}\boldsymbol{K}=\left(\boldsymbol{C}\overline{\boldsymbol{B}},\boldsymbol{C}\overline{\boldsymbol{A}}\,\overline{\boldsymbol{B}},\cdots,\boldsymbol{C}\overline{\boldsymbol{A}}^{M-1}\overline{\boldsymbol{B}}\right)\\ \boldsymbol{y}=\boldsymbol{x}*\overline{\boldsymbol{K}}\end{cases}\tag{4}$$

式中:$M$为输入序列$\boldsymbol{x}$的长度;$\boldsymbol{K}\in\mathbf{R}^M$为长度为$M$的权重向量;*为卷积操作。

#### 1.1.2 选择性扫描机制

S4模型实现了线性时间复杂度,但其捕获序列上下文的能力也受到了静态参数化的限制。为解

决这一限制, Mamba 通过将选择性扫描机制 S6 作为核心 SSM 算子与其他 SSM 区分开来, S6 致力于保留轻量级且有用的状态信息, 兼顾性能和效率, 其通过有选择地将数据压缩到状态中来实现这一点。在计算机视觉任务中, 通过扫描将二维图形转换为一维序列后送入 S6 中进行状态选择, 其结构如图 3 所示。在 S6 中, 从输入数据 $\boldsymbol{x}\in\mathbf{R}^{B\times L\times D}$ 计算矩阵 $\boldsymbol{B}\in\mathbf{R}^{B\times L\times N}$, $\boldsymbol{C}\in\mathbf{R}^{B\times L\times N}$ 和 $\nabla\in\mathbf{R}^{B\times L\times D}$, $B$ 为批大小, $L$ 为输入序列长度, $D$ 为输入特征维度, $N$ 为隐状态维度。利用选择性扫描机制, Mamba 不仅在序列长度上实现了线性可扩展性, 而且在长序列上下文建模任务中表现出有竞争力的性能。

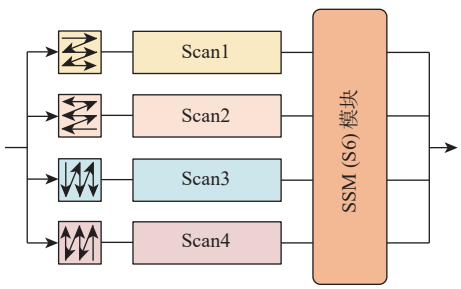


图 3　选择性扫描机制示意图

Fig. 3　Schematic diagram of selective scanning mechanism

### 1.1.3　滑动窗口扫描机制

选择性扫描机制 S6 在处理一维因果序列数据上表现出卓越的性能。这种机制有效地捕获扫描片段中的重要信息, 类似于语言建模, 其中, 理解顺序序列之间的依赖关系至关重要。然而, 将二维图像展开为一维序列会破坏局部二维依赖关系的完整性, 从而降低模型有效识别空间关系的能力。如图 3 所示, ViM[20] 中使用的序列化方式破坏了这些局部依赖, 水平扫描时显著增加了垂直相邻令牌之间的距离, 垂直扫描时又显著增加了水平相邻令牌之间的距离, 并阻碍了模型捕捉局部细微差异的能力。VMamba[22] 试图通过在水平和垂直方向上扫描图像来解决该问题, 但其仍然无法在单次扫描中全面处理空间区域。为解决该问题, 本文提出一种滑动窗口扫描机制, 将图像划分为多个不同的局部窗口, 在每个窗口中进行线性扫描, 确保图像相邻令牌排列更加紧密, 增强对局部依赖的捕获。如图 4 所示, 以窗口尺寸大小设置为 3、滑动步数设置为 2 的正向水平扫描为例, 在沿水平方向扫描时拉近了垂直方向相邻令牌之间的距离, 最终在 4 个方向上集成滑动窗口扫描机制, 在图像局部细节和全局视角之间取得平衡。

## 1.2　空间信息聚合模块

注意力机制可以强化与检测目标相关的信息, 抑制无关信息。本文设计了 SIAM以同时捕捉全局

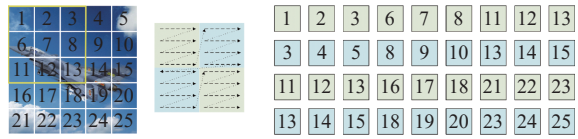


图 4　滑动窗口扫描机制示意图

Fig. 4　Schematic diagram of scanning mechanism of sliding window

与局部空间特征, 其包含 2 个独立的局部和全局注意力单元。如图 5 所示, 这种双分支路径设计有效地保留了局部细节信息和全局建模信息。此外, 使用分离的通道维度来平衡检测精度和计算资源。具体地, 将主干网络提取的特征图 $\boldsymbol{F}_i$ 沿着通道维度平均拆分成 2 个特征图 $\boldsymbol{F}_i^1$ 和 $\boldsymbol{F}_i^2$, 并分别送入全局注意力模块和局部注意力模块。最后, 将这 2 个注意力单元的输出连接起来, 再加上一个 1×1的卷积层。该过程可以公式化为

$$\begin{cases}\boldsymbol{F}_i^1,\boldsymbol{F}_i^2=\mathrm{Split}(\boldsymbol{F}_i)\\ \boldsymbol{F}_i'=\mathrm{pwConv}\left(\mathrm{Concat}\left(G_{\mathrm{sa}}\left(\boldsymbol{F}_i^1\right),L_{\mathrm{sa}}\left(\boldsymbol{F}_i^2\right)\right)\right)\end{cases}\tag{5}$$

式中: pwConv 为 1×1 逐点卷积层; $G_{\mathrm{sa}}$ 为全局空间注意力(global spatial attention, GSA); $L_{\mathrm{sa}}$ 为局部空间注意力(local spatial attention, LSA); Split 为分割成 2 个矩阵; Concat 为拼接。

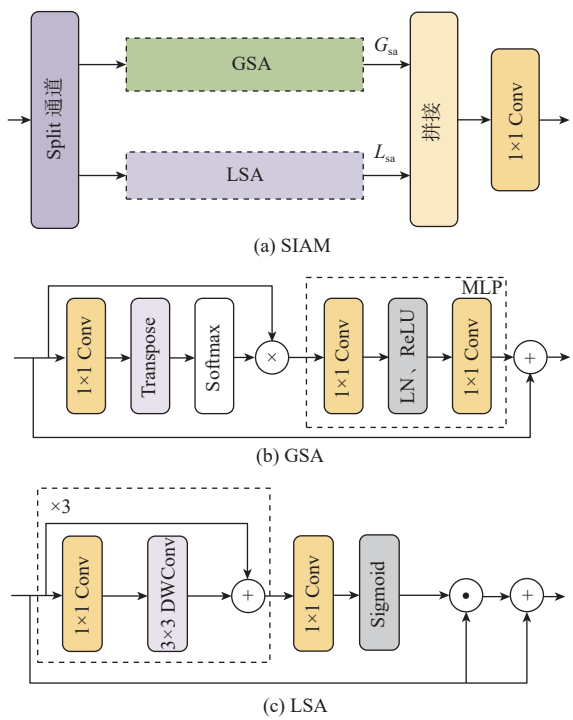


图 5　空间信息聚合模块结构

Fig. 5　Spatial information aggregation module structure

全局空间注意力关注各个像素在空间中的远程关系, 可以与局部空间注意力进行互补。长距离依赖关系可以使特征表示更强大, 使用 $\boldsymbol{F}_i^1$ 作为输入, 生成全局空间注意力过程如下:

$$\begin{cases} \text{Att-G}\left(\boldsymbol{F}_i^1\right)=\phi\left(\left(\text{pwConv}(\boldsymbol{F}_i^1)\right)^{\text{T}}\right) \\ G_{\text{sa}}\left(\boldsymbol{F}_i^1\right)=\text{MLP}\left(\text{Att-G}\left(\boldsymbol{F}_i^1\right)\otimes\boldsymbol{F}_i^1\right)+\boldsymbol{F}_i^1 \end{cases} \tag{6}$$

式中：Att-G为全局注意力算子；⊗表示矩阵乘法；$\phi(\cdot)$表示Softmax函数；MLP由2个具有ReLU非线性激活函数和归一化层的全连接层组成，MLP的第1层将输入变换到高维空间，第2层将其维度恢复为与输入相同。

局部空间注意力模块能够在给定特征图的空间维度上提取感兴趣区域的局部特征。使用$\boldsymbol{F}_i^2$作为输入，计算局部空间注意力如下：

$$\begin{cases} \text{Att-L}\left(\boldsymbol{F}_i^2\right)=\sigma\left(\text{pwConv}\left(F_{\text{c}}\left(\boldsymbol{F}_i^2\right)+\boldsymbol{F}_i^2\right)\right) \\ L_{\text{sa}}\left(\boldsymbol{F}_i^1\right)=\text{Att-L}\left(\boldsymbol{F}_i^2\right)\odot\boldsymbol{F}_i^2+\boldsymbol{F}_i^2 \end{cases} \tag{7}$$

式中：$F_{\text{c}}$表示堆叠的3个1×1卷积层和3×3深度可分离卷积层；$\sigma(\cdot)$表示Sigmoid激活函数；Att-L为局部注意力算子；⊙为矩阵逐元素乘法。这种结构设计可以用更少的参数有效地聚焦局部空间信息。

特征经过空间信息聚合后送入基于CNN的CCFM中，CCFM的作用是将相邻层特征进行融合，其结构如图1所示。融合块包含2个1×1卷积和1个重参数块，2条路径的输出通过逐元素相加进行融合，最后，将特征展开为一维序列。

### 1.3 损失函数

BBR是目标检测任务的重要分支任务，现有的BBR方法通常考虑真实框与预测框之间的几何关系，通过使用边框间的相互位置与相对形状计算损失，而忽略了边界框自身的形状与尺度等固有属性对BBR的影响。Shape-IoU[27]可以通过关注边界框自身的形状和尺度来计算损失，从而使边界框预测更加准确。Inner-IoU[28]通过辅助边界框计算损失加速BBR过程，与IoU损失相比，当比例小于1且辅助边界框尺寸小于实际边界框时，回归的有效范围小于IoU损失，但梯度的绝对值大于从IoU损失获得的梯度，可以加速高IoU样本的收敛。相反，当比例大于1时，较大尺度的辅助边界框扩大了回归的有效范围，对低IoU样本的回归具有增强效果。

引入Inner-IoU思想使用辅助边界框加速模型收敛，结合Shape-IoU，本文提出基于辅助边界框的Inner-Shape-IoU，在考虑边界框自身形状的同时使用辅助边界框加速模型收敛。如图6所示，虚线框与点线框均为辅助边界框，Inner-Shape-IoU的计算过程如下：

$$I^{\text{inner}}=\frac{(\alpha Z)\cap(\alpha Z^{\text{gt}})}{(\alpha Z)\cup(\alpha Z^{\text{gt}})} \tag{8}$$

$$w_w=\frac{2(w^{\text{gt}})^\beta}{(w^{\text{gt}})^\beta+(h^{\text{gt}})^\beta} \tag{9}$$

$$h_h=\frac{2(h^{\text{gt}})^\beta}{(w^{\text{gt}})^\beta+(h^{\text{gt}})^\beta} \tag{10}$$

$$d^{\text{shape}}=h_h(x_c-x_c^{\text{gt}})^2/c^2+w_w(y_c-y_c^{\text{gt}})^2/c^2 \tag{11}$$

$$\Omega^{\text{shape}}=\sum_{t=w,h}(1-\text{e}^{-w_t})^\tau \qquad \tau=4 \tag{12}$$

$$\begin{cases} w_w=\dfrac{h_h|w-w^{\text{gt}}|}{\max(w,w^{\text{gt}})} \\ h_h=\dfrac{w_w|h-h^{\text{gt}}|}{\max(h,h^{\text{gt}})} \end{cases} \tag{13}$$

式中：$I^{\text{inner}}$为使用辅助框计算出的IoU；$\alpha$为放缩因子，与预测框和真实框的IoU有关；$Z$为边界框；上标gt表示真实值；$\beta$为尺度因子，与数据集中目标的尺度有关；$w_w$和$h_h$分别为水平方向和垂直方向的权重系数，其值与真实框的形状相关；$c$为归一化因子；$w$和$h$分别为预测框的宽度和高度；$d^{\text{shape}}$为形状距离；$\Omega^{\text{shape}}$为形状损失，用$\tau$控制对形状损失的关注程度。其对应的BBR损失如下：

$$L_{\text{inner-shape-iou}}=1-I^{\text{inner}}+d^{\text{shape}}+0.5\Omega^{\text{shape}} \tag{14}$$

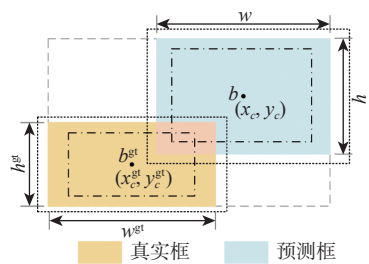


图6 Inner-Shape-IoU计算示意图

Fig. 6 Inner-Shape-IoU calculation diagram

## 2 实 验

### 2.1 数据集与评价指标

本文在VisDrone2019数据集上进行实验。VisDrone2019数据集是由天津大学AISKYEYE团队获取的一个大规模无人机视角的数据集，涵盖多种场景、不同无人机平台、天气和光照条件，共有8 629张图像，其中，训练集6 471张，验证集548张，测试集1 610张，包含10个类别：行人、人、汽车、面包车、巴士、卡车、摩托车、自行车、三轮车和带棚三轮车。

在目标检测任务中，评价模型性能的指标主要有精度$P$、召回率$R$、平均精度$A_{\text{P}}$、均值平均精度$m_{\text{AP}}$，各指标的具体计算式如下：

$$P=\frac{T_{\text{P}}}{T_{\text{P}}+F_{\text{P}}} \tag{15}$$

$$R = \frac{T_P}{T_P + F_N} \tag{16}$$

$$A_P = \int_0^1 P(R)\mathrm{d}R \tag{17}$$

$$m_{AP} = \frac{1}{N}\sum_{i=0}^{1} A_{Pi} \tag{18}$$

式中: $T_P$为真正例; $F_P$为假正例; $F_N$为假负例;

在轻量化与性能平衡的考虑中, 模型的参数量和浮点运算量是至关重要的考量因素。参数量指模型中所有参数的总数量, 包括权重和偏置等。浮点运算量的大小用来衡量模型在前向传播时的计算负荷。

本文实验采用 mAP0.5 和 mAP0.5∶0.95、参数量和浮点运算量作为模型性能评价指标。mAP0.5 表示 IoU 阈值为 0.5 时的 $m_{AP}$, mAP0.5∶0.95 则为 IoU 从 50% 开始, 以步长为 0.05 上升到 95% 时的平均 $m_{AP}$, mAP0.5 和 mAP0.5∶0.95 越大, 则模型的整体精度就越高。

### 2.2　实验环境

本文实验在 Linux 操作系统上进行, 操作系统为 Ubuntu 18.04-LTS, CPU 为 IntelGold 5118, 运行内存为 32 GB, GPU 为 2 块 NVIDIATeslaV100-16 GB, Python 解释器版本为 3.9, CUDA 版本为 11.7, Pytorch 版本为 2.0.1。网络未使用预训练模型参数权重, 设置训练轮次为 150, 批处理大小为 4, 网络初始学习率为 0.01, 输入图像尺寸大小统一缩放为 640 像素×640 像素。

### 2.3　消融实验

模型结构框架的改进基于 RT-DETR 模型, 为验证各改进算法的有效性, 在 VisDrone2019 数据集上进行消融实验, 在测试集上的实验结果如表 1 所示, 表中加黑数值为最优结果。图 7 为训练过程中模型在验证集上的各评价指标变化曲线。可以看出, DE-SSM 算法相较于基线模型其各项性能指标均有提升。表 1 依次对各个模块改进前、后的实验效果进行对比。可以看出, 加入 SSMFormer 后, 在仅增加少量参数量和浮点运算量的情况下, 各个评价指标均有提升, mAP0.5 和 mAP0.5∶0.95 分别提升了 0.84% 和 0.45%, 同时精度提高了 2.34%。这是因为基于滑动窗口扫描机制的 SSMFormer 有效拉近了不同扫描方式下相邻令牌之间的距离, 以获得更聚焦的局部特征表示, 有利于后续定位与分类任务。加入 SIAM 后, 参数量增加了 $0.208\times10^7$, 浮点运算量增加了 $6.7\times10^9$ 次, 性能也同样得到提升, mAP0.5 和 mAP0.5∶0.95 分别提升了 0.78% 和 1.23%。原因在于 SIAM 可以过滤空间位置信息, 抑制无关背景噪声的干扰。Inner-Shape-IoU 在不增加参数量和浮点运算量的情况下, mAP0.5 和 mAP0.5∶0.95 分别提升了 1.08% 和 1.36%, 可以看出, Inner-Shape-IoU 是比 IoU 更适合小目标检测任务的衡量标准。

**表 1　消融实验结果**

**Table 1　Ablation experiment results**

| SSMFormer | SIAM | Inner-Shape-IoU | 输入图像尺寸/(像素×像素) | *P*/% | *R*/% | mAP0.5/% | mAP0.5∶0.95/% | 参数量 | 浮点运算量/$10^9$次 |
|---|---|---|---|---|---|---|---|---|---|
| | | | 640×640 | 59.55 | 44.30 | 46.11 | 27.68 | $1.988\times10^7$ | **57.0** |
| √ | | | 640×640 | **61.89** | 44.70 | 46.95 | 28.13 | $2.018\times10^7$ | 57.3 |
| | √ | | 640×640 | 61.15 | 44.77 | 46.89 | 28.91 | $2.196\times10^7$ | 63.7 |
| | | √ | 640×640 | 61.63 | 45.08 | 47.19 | 29.04 | **$1.988\times10^7$** | 57.0 |
| √ | √ | | 640×640 | 61.42 | 45.66 | 47.69 | 29.27 | $2.227\times10^7$ | 64.0 |
| √ | | √ | 640×640 | 61.10 | **45.67** | 47.31 | 29.14 | $2.018\times10^7$ | 57.3 |
| | √ | √ | 640×640 | 60.68 | 45.27 | 47.50 | 29.36 | $2.196\times10^7$ | 63.7 |
| √ | √ | √ | 640×640 | 61.07 | 45.37 | **47.71** | **29.57** | $2.227\times10^7$ | 64.0 |
| √ | √ | √ | 960×960 | 66.60 | 52.66 | 54.71 | 34.81 | $2.227\times10^7$ | 64.0 |

将各模块进行组合以验证其兼容性。在仅加入 SSMFormer 的情况下, 精度最高, 达到了 61.89%。在同时加入 SSMFormer 和 Inner-Shape-IoU 的情况下, 召回率最高达到了 45.67%。将所有改进加入模型中得到的最终改进算法的 mAP0.5和 mAP0.5∶0.95 均达到了最高, 分别为 47.71% 和 29.57%。最后, 使用不同的输入尺寸进行对比, 使用 960 像素×960 像素大小的图像作为输入, 相较于 640 像素×640 像素的图像输入, mAP0.5 和 mAP0.5∶0.95 分别提升了 7.0% 和 5.24%。通过实验结果可以看出, 本文算法在含有众多小目标及密集目标的 VisDrone2019 数据集上有良好的检测效果。

### 2.4　对比实验

为证明 DE-SSM 算法的先进性, 本节在

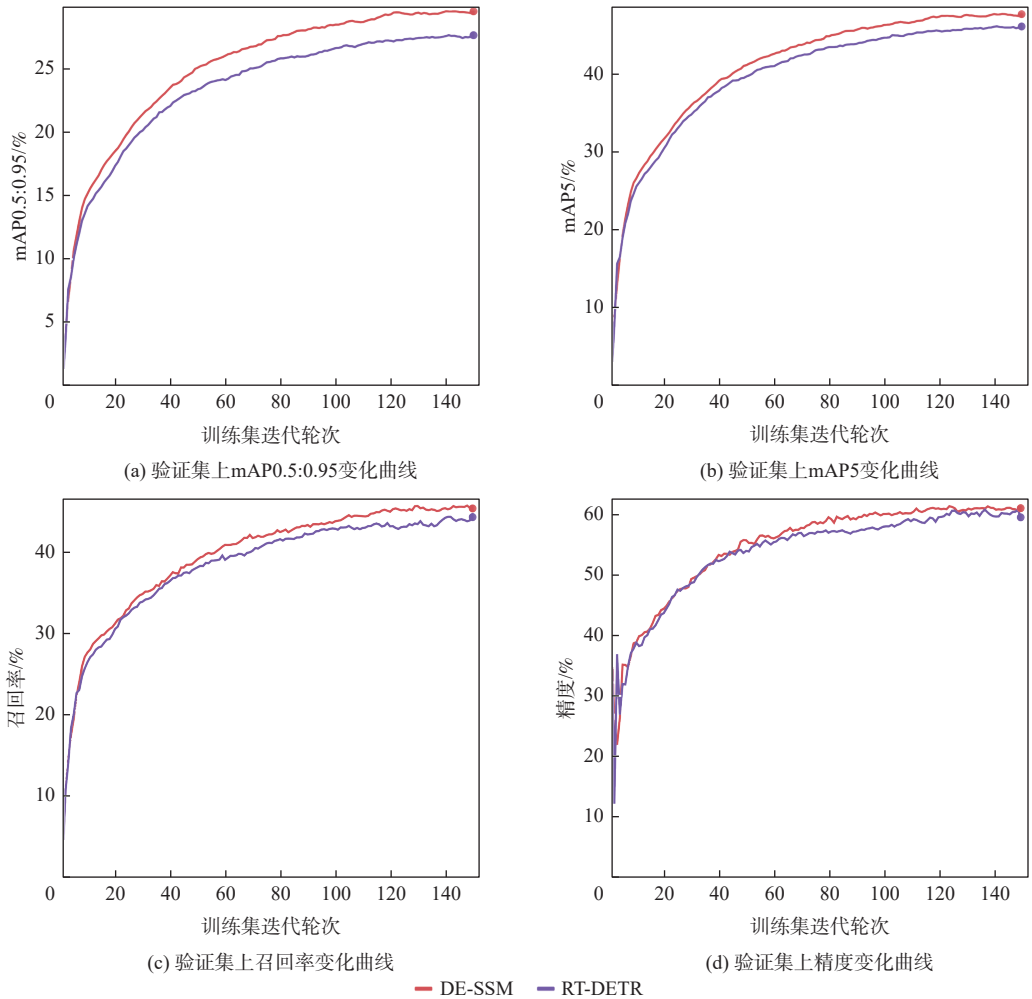


图 7 训练过程中评价指标变化曲线

Fig. 7 Change curves of evaluation index during training process

VisDrone2019测试集上与各种目标检测算法进行对比分析。表2为近几年较为流行的算法在VisDrone2019测试集中的检测结果，选用mAP0.5和mAP0.5∶0.95中的检测作为评价指标，表中加黑数值为最优结果。可以看出，DE-SSM模型在这2个指标上均优于其他算法。

**表 2 对比实验结果**

**Table 2 Contrast experiment results**

| 算法 | mAP0.5∶0.95/% | mAP0.5/% | 浮点运算量/$10^9$次 |
|---|---|---|---|
| IterDet[29] | 20.40 | 36.80 | 173 |
| Swin Transformer[24] | 23.10 | 42.50 | 226 |
| Cascade R-CNN[30] | 16.10 | 31.90 | 236 |
| Deformable DETR[9] | 24.80 | 42.70 | 193 |
| Light-headR-CNN[31] | 16.50 | 32.80 | 94 |
| PP-YOLOE[32] | 24.60 | 39.60 | 110 |
| YOLOX-X[33] | 25.80 | 43.20 | 281 |
| RT-DETR[13] | 27.68 | 46.11 | **57** |
| TPH-YOLOv5[34] | 26.10 | 44.10 | 261 |
| 本文算法 | **29.57** | **47.71** | 64 |

DE-SSM通过SSMFormer建模全局信息，使用SIAM过滤无关信息，加入Inner-Shape-IoU针对边界框自身的形状和尺度来计算损失，显著提高了检测性能。

VisDrone2019数据集中的图像具有检测目标密集、部分目标过小、数据分布不同和目标遮挡严

重的特点。根据 DE-SSM 模型在 VisDrone2019 数据集上的性能表现可以得出, 相较于其他模型, DE-SSM 更注重图像中的局部细节信息, 在小目标检测任务上具有较大的优势。

### 2.5　检测结果可视化

为更直观展示改进模型的检测效果, 对检测结果进行可视化, 如图 8 所示。在原图中对差异部分进行了标注, 图 8(a) 中的黄色边框是对检测差异部分进行等比例放大后的图像。如图 8(a)所示, 1 处是 1 个行人被重复框选; 2 处是被错误识别为 1 个行人; 3 处是把 1 把椅子识别成 1 辆卡车; 4 处是 1 个拐角的阶梯被错误识别为行人; 5 处是 1 辆巴士和 1 个补给站被识别成了 2 辆面包车。而这些物体均被 DE-SSM 算法正确识别。



(a) 原图　(b) RT-DETR检测结果　(c) DE-SSM检测结果

图 8　检测结果可视化效果

Fig. 8　Visualization of the detection result

## 3　结　论

针对无人机视角下的航拍图像小目标检测任务, 本文提出一种基于状态空间模型的小目标检测模型 DE-SSM。

1) 为减轻 Transformer 序列长度的二次计算复杂度, 引入状态空间模型, 同时设计了滑动窗口扫描机制以将二维图像转换为一维序列, 滑动窗口扫描机制可以更好地聚合图像周围的特征, 更适合于小目标检测任务。

2) 设计了空间信息聚合模块, 通过 2 个独立的注意力单元分别捕捉全局特征和局部特征, 可以强化与检测目标相关的信息。

3) 对传统交并比损失函数进行改进, 通过关注边界框的形状和尺度来计算损失, 从而使边界框预测更加精确, 并且加入辅助框来帮助模型更好地收敛。

在 VisDrone2019 数据集上进行充分的实验, mAP0.5 和 mAP0.5:0.95 分别达到了 47.71% 和 29.57%, 优于现有主流模型。DE-SSM 能够有效检测出无人机航拍图像中的小目标, 为小目标检测提供了一种切实可行的解决方案。

### 参考文献 (References)


[1] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2014: 580-587.

[2] Girshick R. Fast R-CNN[C]//Proceedings of the IEEE International Conference on Computer Vision. New York: ACM, 2015: 1440-1448.

[3] 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.

[4] Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 779-788.

[5] Liu W, Anguelov D, Erhan D, et al. SSD: single shot multibox detector[C]//Proceedings of the Computer Vision-ECCV 2016. Berlin: Springer, 2016: 21-37.

[6] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE International Conference on

Computer Vision. Piscataway: IEEE Press, 2017: 2999-3007.

[7] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C]//Proceedings of the Advances in Neural Information Processing Systems 30. Red Hook: Curran Associates, Inc. , 2017: 6000-6010.

[8] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with Transformers[C]//Proceedings of the Computer Vision-ECCV 2020. Berlin: Springer, 2020: 213-229.

[9] Zhu X Z, Su W J, Lu L W, et al. Deformable DETR: deformable Transformers for end-to-end object detection[EB/OL]. (2021-03-18)[2024-06-15]. https://arxiv.org/abs/2010.04159.

[10] Gao P, Zheng M H, Wang X G, et al. Fast convergence of DETR with spatially modulated co-attention[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2022: 3601-3610.

[11] Chen Z, Zhang J, Tao D C. Recurrent glimpse-based decoder for detection with Transformer[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 5250-5259.

[12] Zhang H, Li F, Liu S L, et al. DINO: DETR with improved denoising anchor boxes for end-to-end object detection[EB/OL]. (2022-07-11)[2024-06-15]. https://arxiv.org/abs/2203.03605.

[13] Zhao Y A, Lv W Y, Xu S L, et al. DETRs beat YOLOs on real-time object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2024: 16965-16974.

[14] Kim J, Cho H, Kim J, et al. SDDGR: stable diffusion-based deep generative replay for class incremental object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2024: 28772-28781.

[15] Hou X Q, Liu M Q, Zhang S L, et al. Salience DETR: enhancing detection Transformer with hierarchical salience filtering refinement[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2024: 17574-17583.

[16] 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)[2024-06-15]. https://arxiv.org/abs/2010.11929.

[17] Gu A, Dao T. Mamba: linear-time sequence modeling with selective state spaces[EB/OL]. (2024-05-31)[2024-06-15]. https://arxiv.org/abs/2312.00752.

[18] Baccus S, Dao T, Downs G, et al. S4ND: modeling images and videos as multidimensional signals with state spaces[C]//Proceedings of the Advances in Neural Information Processing Systems 35. Red Hook: Curran Associates, Inc. , 2022: 2846-2861.

[19] Smith J T H, Warrington A, Linderman S W. Simplified state space layers for sequence modeling[EB/OL]. (2023-03-03)[2024-06-15]. https://arxiv.org/abs/2208.04933.

[20] Zhu L H, Liao B C, Zhang Q, et al. Vision Mamba: efficient visual representation learning with bidirectional state space model[EB/OL]. (2024-01-17)[2024-06-15]. https://arxiv.org/abs/2401.09417.

[21] Touvron H, Cord M, Douze M, et al. Training data-efficient image Transformers & distillation through attention[EB/OL]. (2021-01-15)[2024-06-15]. https://arxiv.org/abs/2012.12877.

[22] Liu Y, Tian Y J, Zhao Y Z, et al. VMamba: visual state space model [EB/OL]. (2024-01-18)[2024-06-15]. https://arxiv.org/abs/2401.10166.

[23] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 770-778.

[24] Liu Z, Lin Y T, Cao Y, et al. Swin Transformer: hierarchical vision Transformer using shifted windows[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2022: 9992-10002.

[25] Ma J, Li F F, Wang B. U-Mamba: enhancing long-range dependency for biomedical image segmentation[EB/OL]. (2024-01-09)[2024-06-15]. https://arxiv.org/abs/2401.04722.

[26] Yu W H, Luo M, Zhou P, et al. MetaFormer is actually what you need for vision[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 10809-10819.

[27] Zhang H, Zhang S J. Shape-IoU: more accurate metric considering bounding box shape and scale[EB/OL]. (2024-01-12)[2024-06-15]. https://arxiv.org/abs/2312.17663.

[28] Zhang H, Xu C, Zhang S J. Inner-IoU: more effective intersection over union loss with auxiliary bounding box[EB/OL]. (2023-11-14)[2024-06-15]. https://arxiv.org/abs/2311.02877.

[29] Rukhovich D, Sofiiuk K, Galeev D, et al. IterDet: iterative scheme for object detection in crowded environments[C]//Proceedings of the Structural, Syntactic, and Statistical Pattern Recognition. Berlin: Springer, 2021: 344-354.

[30] Cai Z W, Vasconcelos N. Cascade R-CNN: delving into high quality object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 6154-6162.

[31] Li Z M, Peng C, Yu G, et al. Light-head R-CNN: in defense of two-stage object detector[EB/OL]. (2017-11-23)[2024-06-15]. https://arxiv.org/abs/1711.07264.

[32] Xu S L, Wang X X, Lv W Y, et al. PP-YOLOE: an evolved version of YOLO[EB/OL]. (2022-12-12)[2024-06-15]. https://arxiv.org/abs/2203.16250.

[33] Ge Z, Liu S T, Wang F, et al. YOLOX: exceeding YOLO series in 2021[EB/OL]. (2021-08-06)[2024-06-15]. https://arxiv.org/abs/2107.08430.

[34] Zhu X K, Lyu S C, Wang X, et al. TPH-YOLOv5: improved YOLOv5 based on Transformer prediction head for object detection on drone-captured scenarios[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. Piscataway: IEEE Press, 2021: 2778-2788.

# Aerial image small object detection algorithm based on state space model

TENG Shangzhi*，GONG Xi，LI Baoan，LYU Xueqiang

(Key Laboratory of Internet Culture and Digital Dissemination Research，Beijing Information Science and Technology University，Beijing 100192，China)

**Abstract**：Small object detection aims to solve the difficulties of traditional target detection algorithms in dealing with tiny objects. Small object detection needs to overcome the problems of small object size, few pixels, and a small difference between the target and background, which makes it difficult to detect. It puts higher requirements on the robustness and accuracy of object detection algorithms. A small object detection algorithm based on a state space model is designed to explore the application of the state space model in the field of object detection. A sliding window scanning mechanism is designed to convert two-dimensional images into one-dimensional sequences, which is more suitable for small object detection tasks. In order to suppress irrelevant background information, a dual-branch structure is used in a spatial information aggregation module to calculate global and local attention, respectively. Improvements are made to the traditional IoU loss, focusing on the influence of bounding box shape and scale on detection performance, and auxiliary boxes are used to accelerate model convergence. Sufficient experiments are carried out on the public dataset VisDrone2019 to demonstrate the effectiveness of the proposed method.

**Keywords**：small object detection; state space model; spatial information aggregation; attention mechanism; auxiliary bounding box

**Received**：2024-07-23；**Accepted**：2024-09-14；**Published Online**：2024-10-14 16：19

**URL**：link.cnki.net/urlid/11.2625.V.20241014.1513.003

**Foundation items**：National Natural Science Foundation of China (62202061); Beijing Municipal Natural Science Foundation (4232025) ; Scientific Research Program of Beijing Municipal Commission of Education (KM202311232002)

* **Corresponding author**. E-mail：tengshangzhi@bistu.edu.cn