文章编号：1007-2780(2026)02-0222-18

融合隐藏状态压缩与空间依赖感知的 Mamba-Transformer雾天目标检测方法

陈　悦，许　锋*，宋京昊
（中国刑事警察学院 公安信息技术与情报学院，辽宁 沈阳 110854）

摘要：在复杂天气与动态场景下，小目标检测易受到雾霾干扰和特征弱化的影响，现有方法在实时性和鲁棒性方面仍存在不足。为此，本文旨在构建一种兼具高效性与强鲁棒性的检测模型，以提升雾天小目标检测性能。本文提出基于Mamba-Transformer的混合状态空间检测网络HS-MambaDet。首先设计融合隐藏状态压缩与状态空间建模的HSM-SSD骨干结构以降低计算复杂度；随后引入包含高频感知模块与空间依赖感知模块的多尺度频空融合器（MFSF），强化小目标的细节表达与空间上下文建模。整体框架采用CNN、Transformer与Mamba的混合架构，分别实现局部特征提取、全局上下文建模与线性递归结构的高效结合。在RTTS与CityScapes数据集上的大量实验表明，HS-MambaDet在精度与效率上均优于当前主流模型。完整模型在RTTS上的预测精度为87.3%、召回率为73.1%、mAP@0.5为81.2%、mAP@(0.5～0.95)为51.0%，分别较对比模型最高提升3.8%、3.9%、3.6%和3.3%；同时推理时间为0.26 s，保持了良好的实时性。此外，在小目标场景中，本模型的mAP@0.5提升最多达4.4%，并在跨雾强度测试中表现出更强的泛化能力。HS-MambaDet通过引入高效的隐藏状态压缩机制与多尺度频空融合结构，有效增强了雾天条件下对小目标的细节感知和空间建模能力，在检测精度、鲁棒性与推理效率上均取得显著优势，为动态恶劣环境中的实时目标检测提供了一种可行且高效的解决方案。

关　键　词：目标检测；Mamba；高频感知；融合隐藏状态压缩
中图分类号：TP391.4　**文献标识码**：A　**doi**：10.37188/CJLCD.2025-0245　**CSTR**：32172.14.CJLCD.2025-0245

Mamba-Transformer foggy object detection method integrating hidden state compression and spatial dependency perception

CHEN Yue，XU Feng*，SONG Jinghao

（*School of Public Security Information Technology and Intelligence*，*Criminal Investigation Police University of China*，*Shenyang* 110854，*China*）

Abstract：Object detection in foggy and dynamic environments remains challenging，especially for small objects whose features are easily degraded by low visibility. Existing methods often suffer from insufficient

收稿日期：2025-12-02；**修订日期**：2026-01-17.
基金项目：十三五国家重点研发技术项目（No.2017YFC0821004）；2022辽宁省教育厅基本科研重大攻关项目（No.LJKZZ20220007）；中央高校基本科研业务重大培育项目（No.3242022004）
Supported by National Key R&D Program of China (13th Five-Year Plan) (No.2017YFC0821004); 2022 Major Fundamental Research Project of the Liaoning Provincial Department of Education (No.LJKZZ20220007); Major Cultivation Project for Fundamental Research in Central Universities (No.3242022004)
*通信联系人，E-mail：7677759@qq.com

robustness and limited real-time performance. This study aims to develop a high-efficiency and highly robust detection framework tailored for foggy small-object detection. We propose HS-MambaDet, a hybrid state-space detection network built upon the Mamba-Transformer architecture. A hidden-state compression-enhanced backbone(HSM-SSD) is designed to reduce computational overhead while preserving global dependency modeling. Additionally, a Multi-scale Frequency-Spatial Fusion(MFSF) module—consisting of a high-frequency perception branch and a spatial dependency modeling branch—is introduced to enhance fine-grained details and long-range spatial relations for small objects. The overall framework integrates CNN, Transformer, and Mamba components to jointly capture local textures, global context, and efficient sequential dependencies. Extensive experiments on the RTTS and CityScapes datasets demonstrate that HS-MambaDet outperforms mainstream SOTA models in both accuracy and efficiency. On RTTS, the complete model achieves 87.3% precision, 73.1% recall, 81.2% mAP@0.5, and 51.0% mAP@(0.5～0.95), exceeding the best comparison baselines by up to 3.8%, 3.9%, 3.6%, and 3.3%, respectively. The inference time remains low at 0.26 s, ensuring real-time capability. In small-object scenarios, HS-MambaDet improves mAP@0.5 by up to 4.4%, and cross-domain tests under varying fog intensities further verify its superior generalization and robustness. By combining hidden state compression with multi-scale frequency-spatial fusion, HS-MambaDet effectively enhances fine-detail perception and spatial dependency modeling under foggy conditions. The proposed framework achieves strong performance in accuracy, robustness, and inference speed, offering a practical and efficient solution for real-time object detection in adverse weather and dynamic environments.

Key words: object detection; Mamba; high-frequency perception; fusion hidden state compression

1　引　　言

目标检测(Object Detection, OD)是计算机视觉领域的核心任务，旨在对图像或视频中感兴趣的目标进行识别与定位[1]。在实际应用中，如自动驾驶、交通监控、无人机巡检和安防系统[2]，模型往往需要在复杂天气[3](如雾、雨、低照度等)和动态道路环境下实现实时且高精度的检测。此外，在小目标检测[4]中，由于目标尺寸较小，其特征表达十分有限，在恶劣天气干扰下更容易被掩盖或模糊，导致检测性能显著下降[5]。因此，如何在复杂天气道路场景下实现对小目标的实时、高效、鲁棒检测仍是一个亟待解决的问题。

近年来，基于深度学习的检测框架在常规目标检测任务中已取得显著进展。例如，FPN(Feature Pyramid Network)[6]通过多尺度特征融合提升了对不同尺寸目标的感知能力，但其对小目标的细节增强不足，同时在下采样与跨层对齐时容易引入误差。卷积神经网络(CNN)[7]虽具备较强的局部感知能力，但在长距离依赖建模与全局语义捕获方面存在先天不足。

Transformer架构[8]凭借全局自注意力机制在长距离依赖建模上表现出色，但其计算复杂度随序列长度呈二次方增长($O(N^2)$)，导致训练和推理时的高内存消耗与计算延迟。这一固有缺陷使Transformer在处理高分辨率图像或要求实时性的检测任务时面临严峻挑战，限制了其在资源受限场景(如移动端、嵌入式设备)中的部署。相比之下，近期提出的Mamba状态空间模型(State Space Model, SSM)凭借其线性复杂度和递归结构，在语言建模和视觉任务中展现出优越的效率与表现。

Mamba的递归结构使其能够自适应地聚焦于重要信息，过滤无关噪声，这一特性对于雾天等退化场景中的特征选择尤为有利。然而，标准Mamba模型采用因果(单向)建模方式，其序列递归机制在建模全局语义信息时存在一定局限，难以在单次前向传播中实现对全局上下文的同步感知。而视觉任务(尤其是目标检测)往往需要同时整合局部细节与全局语义信息才能做出准确预测。尽管相比自注意力机制，Mamba在推理速度和内存利用方面具有显著优势，但其全局信

息交互能力仍有待进一步加强。尽管已有工作尝试引入双向SSM或结合卷积、Transformer等混合结构，但这些方法仍存在延迟增加、训练不稳定或过拟合风险等问题。

为此，本文提出一种基于Mamba-Transformer的新型混合状态空间检测网络HS-MambaDet，主要有以下3方面的贡献：

(1)提出一种基于Mamba-Transformer的混合架构，显著增强了模型在捕获全局特征以及空间依赖建模的能力，弥补了传统视觉Mamba模型在图像处理方面的劣势；

(2)提出一种基于融合隐藏状态压缩与状态空间建模的骨干结构HSM-SSD，通过降低计算过程中的状态数来降低算法的计算复杂度；

(3)提出包含高频感知模块与空间依赖感知模块的多尺度频空融合器(MFSF)，分别针对小目标特征弱、空间感知不足的问题，显著提升了模型在复杂天气下的鲁棒性；

(4)构建完整的HS-MambaDet框架，并在RTTS和CityScapes数据集上进行了系统实验，结果表明，该方法在准确性和效率上均优于现有SOTA模型。

2 相关工作

在计算机视觉发展历程中，基于卷积的模型自AlexNet[9]提出以来长期占据主导地位。为提升模型表达能力，研究者尝试将Transformer的设计理念融入卷积架构，例如ConvNeXt[10]通过增加网络宽度、采用大卷积核和层归一化对ResNet进行重构，实现了与Transformer相媲美的性能；RegNet[11]通过系统性空间分析推进网络设计；EfficientNet[12]则通过神经架构搜索与渐进式学习优化效率权衡。然而，这类模型受卷积操作的局部性限制，难以建立有效的长距离依赖关系。

基于Transformer的模型通过自注意力机制突破了这一限制，显著扩大了感受野范围。但ViT[13]模型不仅缺乏CNN固有的归纳偏差，还需要大规模数据支撑。为此，DeiT[14]引入知识蒸馏策略提升小数据集上的训练效果，Swin Transformer[15]通过分层结构和移位窗口机制平衡局部与全局表征，PVT[16]则通过空间可分离注意力与金字塔结构提升计算效率。尽管如此，自注意力机制的二次计算复杂度仍制约着其在计算敏感场景中的应用。

为兼顾CNN的局部特征提取能力与Transformer的全局建模优势，研究者提出了多种混合架构。CoAT[17]与CrossViT[18]通过卷积与注意力的协同设计增强特征学习；NextViT[19]将类CNN处理系统化融入Transformer结构；EfficientFormer[20]与FasterViT[21]则致力于优化效率与精度的平衡，通过精心设计的混合架构实现优异性能。

近年来，状态空间模型Mamba因其线性计算复杂度在视觉任务中展现出潜力。Vim[22]提出双向SSM公式以增强空间理解，但双向处理增加了计算开销；VMamba[23]通过交叉扫描模块实现四向选择性扫描，并结合深度卷积与多分辨率结构优化视觉表征，然而其感受野仍受扫描路径限制；EfficientVMamba[24]采用多分辨率混合设计，在高分辨率阶段使用CNN，低分辨率阶段使用SSM，显著提升准确率与吞吐量；SiMBA[25]虽通过EinFFT通道建模改善训练稳定性，但对空间关系的理解仍存在局限。

在复杂感知环境下，空间结构建模不足往往会进一步放大弱目标与背景之间的混淆问题，尤其是在小目标检测任务中表现得尤为明显。针对红外小目标检测场景，王伯霄[26]等提出了一种难点注意力感知网络，通过引导模型关注难以区分的关键区域，有效增强了弱目标特征的表达能力。针对水下声呐目标检测问题，何梦云[27]等从通道调控角度出发，提出了一种弱特征共焦通道调控方法，在复杂噪声背景下提升了模型对弱特征的感知与辨别能力。上述研究表明，在复杂环境条件下，通过引入针对弱特征和关键区域的建模机制，有助于缓解空间信息不足带来的检测性能退化问题，但其在复杂天气视觉场景中的适用性与统一建模方式仍有待进一步研究。这些探索为视觉Mamba模型的发展奠定了基础，但如何实现全面准确的空间理解仍需进一步探索。

3 HS-MambaDet方法

3.1 整体框架

如图1所示，HS-MambaDet主干网络有一个由4个不同阶段组成的层次结构，其中每个阶段

的输出特征记作$\{C_2, C_3, C_4, C_5\}$最终会被送入目标检测网络用于最终的目标检测任务。前两个阶段包括基于CNN的层，用于在更高的输入分辨率下快速提取特征，而第3和第4阶段包括提出的Mixer和自注意力块。具体而言，给定尺寸为$H\times W\times 3$的图像，首先将输入转换为尺寸为$H/4\times W/4\times 3$的重叠块，并由两个连续的3×3的CNN层组成的特征提取前端（Stem）模块投影到C维嵌入空间中。两个阶段之间的下采样器由一个3×3的步长为2的CNN层组成，该层将图像分辨率降低了1/2。此外，阶段1和阶段2的CNN块遵循式(1)和式(2)所示的通用残差块公式：

$$\hat{z}=\mathrm{GELU}\left(\mathrm{BN}\left(\mathrm{Conv}_{3\times 3}(z)\right)\right), \tag{1}$$

$$z=\mathrm{BN}\left(\mathrm{Conv}_{3\times 3}(\hat{z})\right)+z, \tag{2}$$

其中，GELU和BN分别表示高斯误差线性单元激活函数和批归一化。

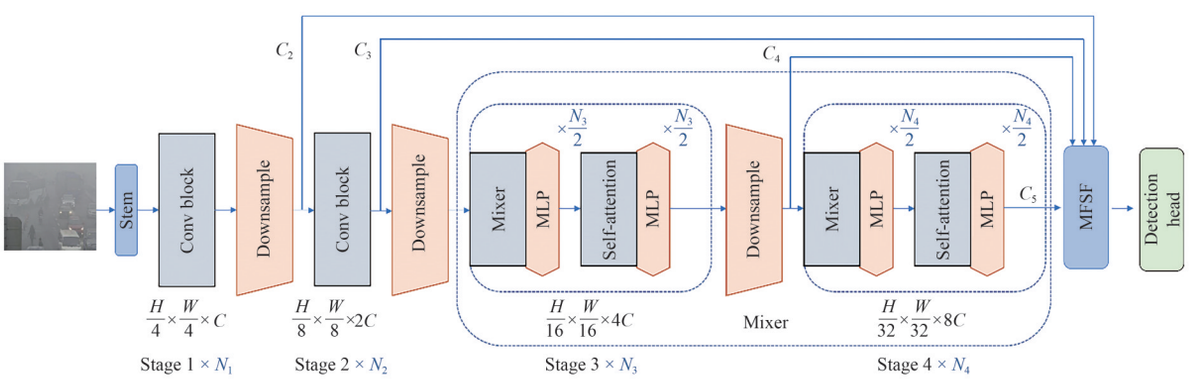


图1　HS-MambaDet模型框图

Fig. 1　Block diagram of HS-MambaDet model

如图2所示，在经过前两个阶段的快速特征提取之后将提取的特征送入两个阶段的Mamba混合器进行后续处理。每个混合器包含一个基于HSM-SSD的Mixer模块和一个基于注意力机制的多头注意力模块。这个过程可以用式(3)所示的过程表示：

$$z_{\mathrm{out}}=\mathrm{Mixer}(z). \tag{3}$$

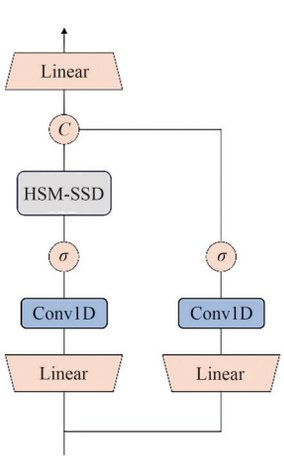


图2　混合器模块

Fig. 2　Mixer module

此外，要注意的是，本文用正则卷积代替因果卷积，因为因果卷积将影响限制在一个方向上，这对于视觉任务来说是不必要的和限制性的。在Mamba模块中设计了一个基于隐藏状态的压缩的HSM-SSD模块，此外，本文添加了一个没有SSD的对称分支，由额外的卷积和Sigmoid Linear Unit（SiLU）激活组成，以补偿由于SSD的顺序约束而丢失的任何内容。然后将两个分支的输出连接起来，并通过最后的线性层进行投影。这个过程可以表示为：

$$X_1=\text{HSM-SSD}\left(\sigma\left(\mathrm{Conv}\left(\mathrm{Linear}(X_{\mathrm{in}})\right)\right)\right), \tag{4}$$

$$X_2=\sigma\left(\mathrm{Conv}\left(\mathrm{Linear}(X_{\mathrm{in}})\right)\right), \tag{5}$$

$$X_{\mathrm{out}}=\mathrm{Linear}\left(\mathrm{Concat}(X_1, X_2)\right), \tag{6}$$

其中：X_{in}是HSM-SSD模块的输入，X_{out}是该模块的最终输出。这种组合确保最后的特征表示结合了顺序信息和空间信息，利用了这两个分支的优势。

在经过HSM-SSD处理后，本文使用一个通用的多头自注意机制来捕获全局注意力关系，以此弥补Mamba机制中对于上下文关系的缺失：

$$\text{Attention}(\boldsymbol{q},\boldsymbol{k},\boldsymbol{v})=\text{Softmax}\left(\frac{\boldsymbol{q}\boldsymbol{k}^{\top}}{\sqrt{d_{\text{h}}}}\right)\boldsymbol{v}\ ,\quad(7)$$

其中：$\boldsymbol{q}$、$\boldsymbol{k}$、$\boldsymbol{v}$分别表示查询、键和值；d_{h}代表缩放因子，取值为键向量$\boldsymbol{k}$的特征维度。此外，本文所提框架允许以类似于先前工作的窗口方式计算注意力。

在经过HS-MambaDet主干网络4个阶段的处理后，每个阶段的中间特征$\{C_2,C_3,C_4,C_5\}$都将被保存并送入高频与空间感知检测网络中进行最终的目标检测。

3.2 基于HSM-SSD的Mamba混合器

HSM-SSD模块不仅通过隐藏状态压缩显著降低了计算复杂度，更重要的是，通过在低维隐藏状态空间中进行特征混合与门控，有效缓解了深层网络中因特征传递导致的小目标细节湮没问题，从而为后续检测模块提供了更丰富、更鲁棒的特征表示。本文简要讨论已有工作中NC-SSD层的计算成本。NC-SSD层的整个过程可以概括为：

$$\hat{B},\ C,\ \Delta,\ x,\ z=\text{Linear}(x_{\text{in}})\ ,\quad(8)$$

$$a,\ B=\text{Discretization}\left(\hat{a},\hat{B},\Delta\right),\quad(9)$$

$$B,\ C,\ x:=\text{DWConv}(B,C,x)\ ,\quad(10)$$

$$y=\text{NC-SSD}(x,a,B,C)\ ,\quad(11)$$

$$x_{\text{out}}=\text{Linear}\left(y\odot\sigma(z)\right),\quad(12)$$

其中：x_{in}，x_{out}，x，$z\in R^{L\times D}$，σ为激活函数。在核大小不变的情况下，计算代价为$O(LD^2+LND)$。考虑到状态N的数量通常比通道D的数量小得多（即$N\ll D$），总体的复杂性主要是由产生x_{out}、x、z所涉及的线性投影所驱动的，从而导致$O(LD^2)$。因此，优化SSD块中的线性投影对于可伸缩性至关重要。

为此进一步深入研究优化这些计算以实现有效的层设计。NC-SSD可分解为两步。首先，利用重要性权重$a\in R^{L}$，通过输入状态$B\in R^{N\times D}$的加权线性组合，得到共享的全局隐藏状态$h\in R^{N\times D}$。其次，每个输入对应的输出是通过用其对应的$C\in R^{L\times N}$投影一个隐藏状态生成的。在这里，如果将投影的输入x表示为x_{in}，W_{in}去除了DWConv，那么等式(13)成立：

$$h=h_{\text{in}}W_{\text{in}}\ ,\quad(13)$$

其中：$h_{\text{in}}\in R^{N\times D}$，$W_{\text{in}}\in R^{D\times D}$。通过计算$h_{\text{in}}$可以在隐藏状态空间上执行一个线性投影。这种方法将依赖于状态数的成本从$O(LD^2)$降低到$O(ND^2)$。换句话说，可以通过调整状态的数量，使$N\ll L$来降低该层的主要成本。

为此，本文采用隐藏状态混合的SSD（Hidden State Mixer-based SSD，HSM-SSD）模块。其核心思想是：将通道方向上的混合与门控操作从输入特征空间迁移到低维的隐藏状态空间中进行，从而降低整体计算开销。

本文关注的是一个共享的全局隐藏状态h。请注意，隐藏状态h本身是一个简化的潜在数组，它用一个显著更小的序列长度n压缩输入数据。基于这一观察，本文提出了一个隐藏状态混合器，它直接在简化的潜在数组h上执行信道混合，包括门控制和输出投影，如图3所示。为此，本文将NC-SSD层的输出近似如下：

$$\begin{aligned}x_{\text{out}}&=f(y)=\text{Linear}\left(y\odot\sigma(z)\right)=\\&\left(Ch\odot\sigma(x_{\text{in}}W_z)\right)W_{\text{out}}\approx\\&C\left(h\odot\sigma(h_{\text{in}}W_z)\right)W_{\text{out}}=Cf(h),\end{aligned}\quad(14)$$

式中：$y=Ch$，f表示门控函数的通道混合，然后与可学习矩阵$\boldsymbol{W}_{\text{out}}$（$\boldsymbol{W}_{\text{out}}\in R^{D\times D}$）进行线性投影。与原始NC-SSD层（先计算$Ch$然后馈送到$f$）相反，本文使用HSM直接将门控和投影应用于隐

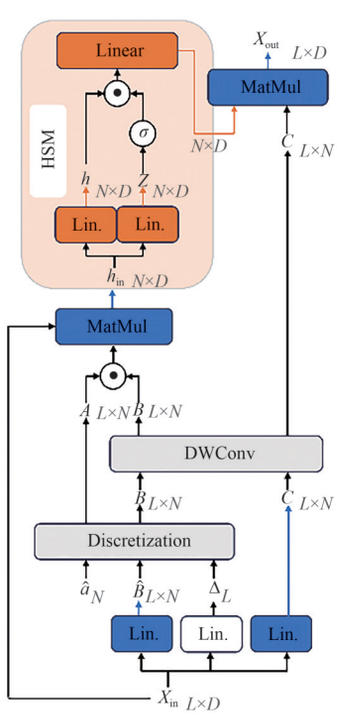


图3 HSM-SSD模块

Fig. 3 HSM-SSD module

藏状态。然后，通过用投影更新的隐藏状态来生成最终输出x_{out}。因此，HSM-SSD模块的复杂度从$O(LD^2+LND)$变为$O(ND^2+LND)$，而N远小于L，随着N的减小这项计算开销可以忽略不计。

3.3　基于高频感知与空间依赖的目标检测模块

尽管基于Mamba的状态空间模型在视觉任务中展现出处理长序列依赖和线性计算复杂度的优势，但在应对雾天场景下的小目标检测任务时仍存在明显局限。首先，雾霾导致的图像退化主要表现为低频成分的扩散与高频细节的衰减，而Mamba模型的递归机制更倾向于捕获序列中的主导模式，对高频细节的敏感度不足，难以有效恢复因雾霾而弱化的边缘与纹理特征。其次，小目标在图像中占据像素有限，其特征表达极易在深层网络中湮没，Mamba模型虽能建模长程依赖，但其隐藏状态在序列传递过程中可能过度平滑化，不利于微小目标的精确定位与判别。此外，传统特征金字塔网络（FPN）通过简单的多尺度特征融合，虽在一定程度上提升了对不同尺寸目标的感知能力，但其在融合过程中缺乏对跨尺度特征间显式的空间依赖关系建模，限制了模型在复杂雾天背景下对模糊小目标的上下文推理能力。

为解决上述问题，本文提出了一种新颖的多尺度频空融合器（Multi-scale Frequency-Spatial Fusion，MFSF）。该模块通过协同作用的高频感知模块（HFP）与空间依赖感知模块（SDP），分别针对性地强化雾天退化图像中的细节信息与空间上下文关系，从而有效弥补了纯Mamba模型在细节感知与复杂空间关系建模方面的不足，通过MFSF处理后的多尺度特征再送入后续的检测头中，显著提升了模型在恶劣天气条件下的检测鲁棒性。具体而言，HFP模块通过引入高通滤波操作，显式提取并增强特征图中的高频成分（如边缘、纹理等细节信息），从而抵抗雾霾对图像细节的掩盖效应；SDP模块则通过跨尺度交叉注意力机制，建模不同层级和位置之间的空间依赖关系，捕获远程空间上下文信息，提升特征对齐能力和语义一致性，为小目标提供更丰富的上下文支撑。

3.3.1　高频感知模块

本文提出的高频感知模块如图4所示。在雾天条件下，图像的高频成分（如边缘、纹理等细节信息）往往受到严重衰减，而Mamba机制虽然擅长建模序列依赖，但对高频细节的敏感度相对不足。HFP模块的设计初衷正是为了弥补这一缺陷，通过与HSM-SSD模块形成功能互补，共同提升模型对雾天小目标的感知能力。HFP首先对输入特征C_i应用高通滤波器以获得高频响应F_i，其尺寸与C_i相同。随后，F_i分别作为两通道的输入，用于生成通道注意力权重u^{cp}与空间注意力权重u^{sp}，这个过程可以表示为：

$$Y_1=\mathrm{Conv}_{1\times1}\left(\mathrm{sigmoid}\left(\mathrm{GAP}(F_i)\right)\right),\quad(15)$$

$$Y_2=\mathrm{Conv}_{1\times1}\left(\mathrm{sigmoid}\left(\mathrm{GMP}(F_i)\right)\right),\quad(16)$$

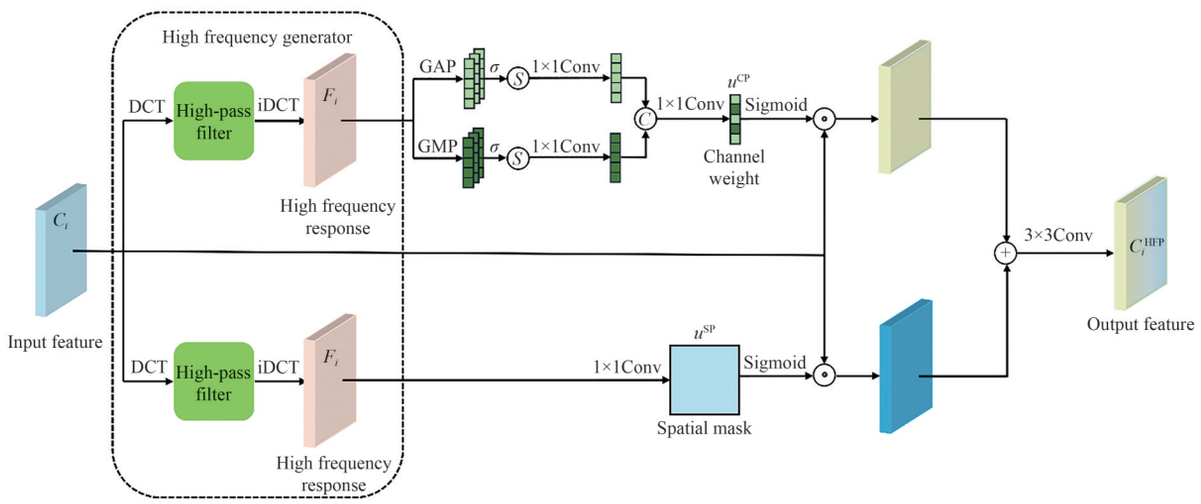


图4　高频感知模块

Fig. 4　High-frequency perception module

$$u^{\mathrm{cp}}=\mathrm{sigmoid}\left(\mathrm{Conv}_{1\times1}\left(\mathrm{concat}\left(Y_1,Y_2\right)\right)\right),\quad(17)$$

$$u^{\mathrm{sp}}=\mathrm{sigmoid}\left(\mathrm{Conv}_{1\times1}\left(F_i\right)\right),\quad(18)$$

其中，GAP 和 GMP 分别是全局平均池化和全局最大池化。

这些权重分别在空间维度与通道维度上进行广播，并与 C_i 做逐元素相乘。最后，来自两通道的结果相加，并通过一个 3×3 卷积，得到 HFP 的输出，记作 C_i^{HFP}。

3.3.2 空间依赖感知模块

HS-MambaDet 主干网络通过 HSM-SSD 模块和注意力机制已经具备了较强的序列建模能力，但在复杂的雾天场景中，单纯依靠序列建模难以充分捕获图像中物体间的空间结构关系。Mamba 模型虽然在线性复杂度下实现了长序列建模，但其扫描机制在处理二维图像空间关系时仍存在固有局限。为此，本文设计了空间依赖感知模块（SDP），旨在建立跨尺度的空间依赖关系，增强模型对雾天环境中空间上下文的感知能力。本文设计的空间依赖感知模块用来捕获 $\{C_i,P_i\}$ 中像素之间的映射，其中 P_i 是特征金字塔中用于最后检测任务的特征。注意机制精通于捕获远程依赖关系，使得来自较低层次特征 C_i 的像素不仅能够与 P_i 中相同空间位置的像素交互，而且能够与更大范围的像素交互。

如图 5 所示，本文首先对 P_i 进行上采样以匹配 C_i 的大小，并将上采样的特征命名为 P_{i+1}^{u}。随后，本文通过 3 个独立的 1×1 卷积计算查询（$\boldsymbol{q}$）、键（$\boldsymbol{k}$）和值（$\boldsymbol{v}$）矩阵，其中 $\boldsymbol{q}$ 来自 C_i 和 $\boldsymbol{k},\boldsymbol{v}$ 来自 P_{i+1}^{u}。接下来将 $\{\boldsymbol{q},\boldsymbol{k},\boldsymbol{v}\}$ 矩阵划分为多个特征块。然后从 $\{\boldsymbol{q},\boldsymbol{k}\}$ 中提取相应的特征块 q_j 和 k_j（$1\leqslant j\leqslant n$，q_j，$k_j\in R^{H_5\times W_5\times C}$），计算它们的相似度，得到相似矩阵 $\boldsymbol{a}_j$。最后，对 $\boldsymbol{v}$ 对应块 v_j 中的每个特征点进行 $\boldsymbol{a}_j$ 加权，生成具有丰富空间依赖性和高级语义的特征块。对 $\boldsymbol{q}$ 的每个特征块应用交叉注意机制后，将这些特征块根据其空间位置进行聚合，形成一个具有丰富空间依赖性的新特征，并将其添加到 C_i 中。这一过程有效地丰富了微小对象的特征内容，并为其注入了高级特征的丰富语义。

最后融合后的特征 C_i^{SDP} 被送入具体的检测头中用于目标检测任务。

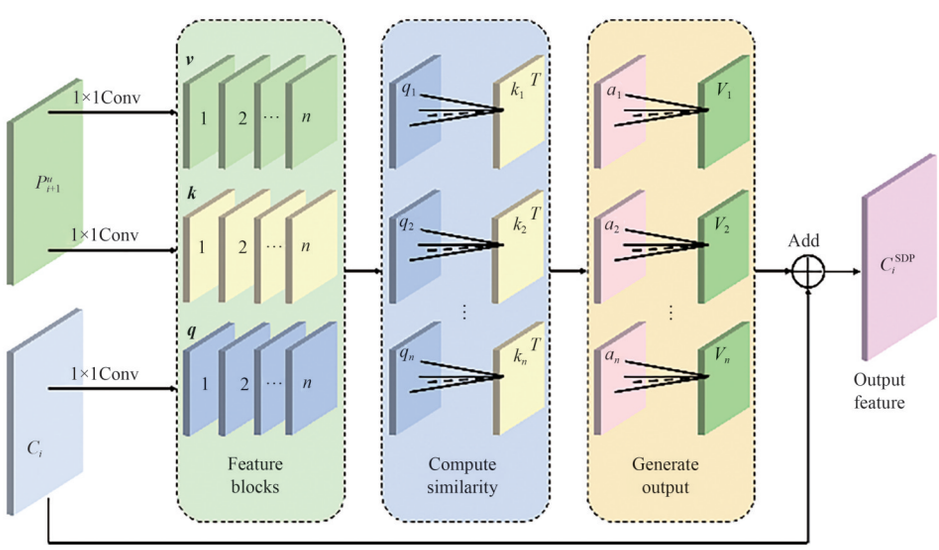


图 5 空间依赖感知模块

Fig. 5 High-frequency perception module

4 实验结果与讨论

4.1 实验准备

在硬件方面，本文使用 NVIDIA RTX A6000 GPU 进行模型训练和推理。在软件方面，使用 PyTorch 深度学习环境。所有模型使用 AdamW 优化器，批大小为 16，学习率为 4e−3。模型的阶段 3 和阶段 4 中的自注意力使用的窗口大小分别为 14 和 7。本文使用 RTTS 数据集与 CityScapes 数据集来评估模型的性能。

4.2　实验结果与对比分析

4.2.1　消融实验结果

首先验证本文所提出方法中各个模块的有效性，依次测试了去除掉主干网络中的注意力模块、高频感知检测网络和基于隐藏状态的HSM-SSD模块时的性能。从表1可以看出，去掉这些模块后模型性能有不同程度的下降，其中去掉高频感知检测网络和注意力模块后的预测精度、召回率、mAP@0.5和mAP@(0.5～0.95)值分别下降了3.4%、4.2%、3.6%和4.5%与2.3%、3.5%、3.1%和3.8%，而使用普通SSM模块替换HSM-SSD模块后性能下降并不显著，这说明注意力模块和高频感知检测网络对图像特征感知更加精细，对性能影响最大。

表1　消融实验结果

Tab. 1　Ablation experiment results

模块	预测精度/%	召回率/%	mAP@0.5/%	推理时间/s	mAP@(0.5～0.95)/%	FLOPs/G
去除注意力模块	83.9	69.5	77.7	0.22	47.2	39.8
去除高频感知检测网络	83.1	70.8	78.1	0.25	46.5	41.3
去除HSD模型	86.2	71.9	80.1	0.28	49.5	44.9
完整模型	87.3	73.1	81.2	0.26	51.0	45.6

虽然HSM-SSD模块对性能影响较小，但是相比于普通SSM模块，其在推理时间上的优势较为明显，注意力模块则会对模型推理时间有较大影响，而使用普通的FPN网络进行目标检测和本文的高频感知检测网络对推理时间影响不大，这也说明本文的高频感知检测网络不会对实际部署中的实时性产生较大影响。

4.2.2　目标检测结果对比分析

为了验证所提方法的有效性，本文选取了CNN模型ConvNeXt、Transformer模型ViT、CNN-ViT混合模型CoAT以及VMamba进行对比实验。实验结果如表2与图6所示。

表2　不同模型的对比结果

Tab. 2　Comparison results of different models

模型	预测精度/%	召回率/%	mAP@0.5/%	推理时间/s	mAP@(0.5～0.95)/%	参数量/M	FLOPs/G
HS-MambaDet	87.3	73.1	81.2	0.26	52.4	45.2	45.6
ConvNeXt	83.5	69.1	77.6	0.25	48.1	28.9	78.2
ViT	85.7	71.8	79.3	0.42	49.6	86.4	190.5
CoAT	86.7	72.2	80.1	0.51	50.8	75.8	120.3
VMamba	85.1	70.4	78.5	0.35	48.9	32.6	35.8

实验结果表明，本文所提模型在所有评价指标上均达到最优性能。具体而言，在预测精度方面，本文模型较表现次优的CoAT高出0.6%；在召回率方面，领先ViT模型1.3%；在mAP@0.5指标上，比混合模型提高了1.1%。尤其值得注意的是，尽管基于Transformer的模型在性能上接近本文方法，但其计算复杂度和推理时间显著高于本文模型，限制了其在实时检测任务中的应用可行性。而VMamba与ViT性能相近，但是在推理时间上有着显著优势，证明了其在实时目标检测任务中的巨大潜力。HS-MambaDet兼顾ViT与VMamba的优势，既在性能上取得较大优势又具有较短的推理时间。此外，尤其需要注意的是，本文所提模型在mAP@(0.5～0.95)指标上相较于其他模型有着更大的优势，这正是因为我们针对雾天特点专门设计的模型结构带来的增益。

在模型复杂度方面，本文模型展现出优异的效率与可部署潜力。具体而言，本模型参数量为45.2M，虽略高于ConvNeXt (28.9M)与VMamba (32.6M)，但显著低于ViT (86.4M)与CoAT (75.8M)，表明本结构通过引入适度增加的参

数，有效融合了多架构优势。更为关键的是，在衡量实际计算负载的FLOPs指标上，本文模型（45.6G）不仅远低于ViT与CoAT等重型模型，与计算高效的VMamba（35.8G）亦处于同一量级，却带来了显著的性能提升。这一结果印证了所提出的HSM-SSD骨干结构在压缩隐藏状态、降低计算复杂度方面的有效性，使本文模型以可控的参数增长为代价，实现了更高的计算效率与更优的性能表现，奠定了其在资源敏感场景下的应用基础。

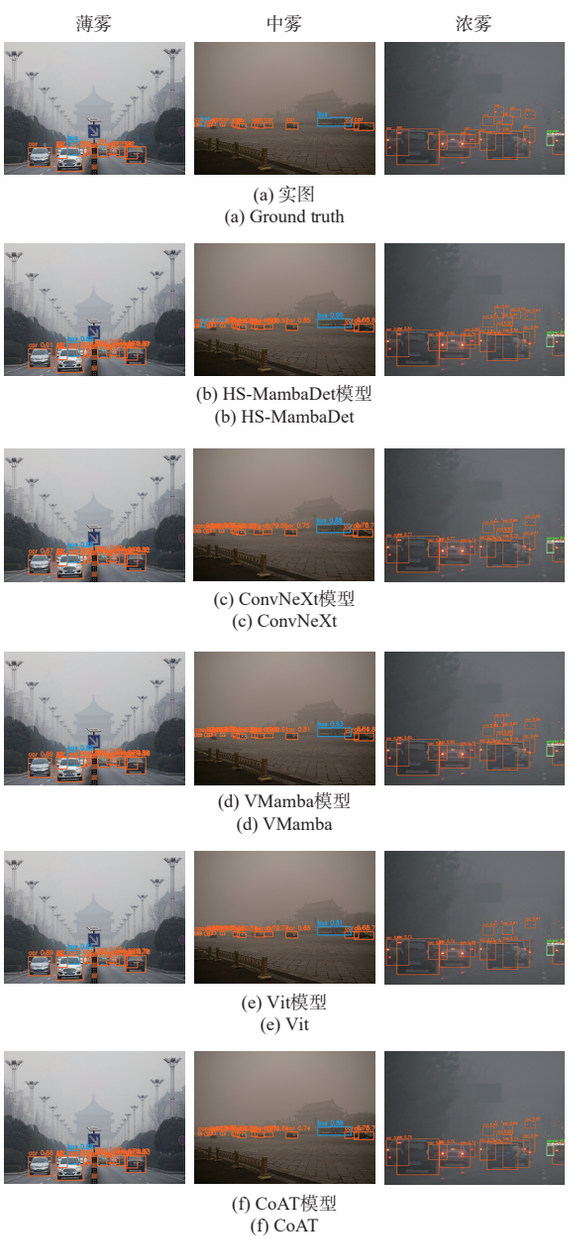


图6 雾天检测可视化结果

Fig. 6 Visualization results of object detection in foggy scenes

综上所述，本文方法不仅在检测精度、召回率，mAP@0.5和mAP@（0.5～0.95）四个核心指标上都优于现有主流模型，而且保持了较低的计算开销，显示出在实时和高精度需求场景下的综合优势与实用价值。

4.2.3 可视化结果

为了展示本文算法对不同程度雾天目标的识别效果以及验证其鲁棒性，将所提方法与主流模型的检测效果进行对比，选取3种场景：薄雾（Light Fog）、中雾（Moderate Fog）和浓雾（Dense Fog）。

如图6所示，对于薄雾场景各主流模型均能有效检测出车辆目标，这是因为天气原因对图像干扰较少，这种情况下噪声较少，模型提取特征更加凝练，即使如此，本文所提方法依然取得了更高的置信度。对于中雾场景，ConvNeXt由于主要捕获局部纹理信息，在受到干扰之后已经出现漏检，可以看出对于两个远处的车辆以及公交车共计3处出现了漏检。ViT模型虽然有更强大的全局上下文建模能力，但是和VMamba一样均出现了两处漏检。CoAT兼顾全局与局部特征，除了远处的公共汽车之外能够全部检测出来，本文所提出的HS-MambaDet能够很好地处理高频信息，对于远处公共汽车也能准确检测出；对于浓雾场景，由于雾天能见度过低及车辆目标被电线遮挡，边缘信息等局部特征的检测受到干扰，目标识别的难度大幅提高。如图6所示，ConvNeXt对于浓雾中的小目标和受遮挡目标已经出现多处漏检。而其他方法由于其同样无法有效分离图像中的高低频成分，对于模糊小目标已经出现数量不等的漏检。本文所提方法通过综合全局注意力和局部细节以及高频分量提取，在较大的干扰下仍然成功检测出小目标以及被遮挡车辆，有效减少了雾天场景的漏检情况，表现出优越的检测准确性和强大的鲁棒性，展现出较高的实际应用潜力。

4.2.4 复杂场景检测结果

为评估模型在复杂多目标场景下的泛化性能，本文对行人、汽车、公共汽车、自行车和摩托车5类目标进行了分类别测试。定量结果（表3～表6）表明，HS-MambaDet在所有目标类别上均取得了最优或接近最优的综合性能。具体来看，在核

表3　预测精度结果

Tab. 3　Prediction accuracy results

模型	预测精度/%				
	行人	汽车	公共汽车	自行车	摩托车
HS-MambaDet	92.2	96.8	72.7	76.6	74.0
ConvNeXt	87.2	93.5	66.8	71.5	61.6
ViT	89.2	94.5	69.5	74.3	68.5
CoAT	90.8	96.0	71.2	75.5	69.0
VMamba	88.5	94.5	68.5	73.0	64.0

表4　召回率结果

Tab. 4　Recall results

模型	召回率/%				
	行人	汽车	公共汽车	自行车	摩托车
HS-MambaDet	77.4	79.8	56.7	59.2	65.4
ConvNeXt	74.1	75.2	52.5	54.0	60.5
ViT	74.9	77.8	54.1	56.5	62.4
CoAT	76.2	78.5	55.5	57.8	63.8
VMamba	75.6	77.4	54.4	55.7	61.8

表5　mAP@0.5结果

Tab. 5　mAP@0.5 results

模型	mAP@0.5/%				
	行人	汽车	公共汽车	自行车	摩托车
HS-MambaDet	85.9	91.0	66.6	69.9	74.3
ConvNeXt	81.3	87.5	61.5	65.6	70.7
ViT	83.8	89.4	64.0	67.8	72.6
CoAT	84.5	90.2	65.2	68.5	73.5
VMamba	83.6	89.2	63.3	67.1	72.4

表6　mAP@(0.5~0.95)结果

Tab. 6　mAP@(0.5～0.95) results

模型	mAP@(0.5～0.95)/%				
	行人	汽车	公共汽车	自行车	摩托车
HS-MambaDet	51.3	56.6	39.8	42.9	46.9
ConvNeXt	46.0	52.2	36.1	38.7	43.0
ViT	48.3	54.1	37.1	40.4	45.1
CoAT	48.7	54.5	37.7	41.1	45.7
VMamba	48.0	53.7	36.7	39.9	44.3

心指标mAP@(0.5～0.95)上，本文模型相较于表现最佳的基线模型，在行人、汽车、公共汽车、自行车和摩托车类别上分别实现了2.6%、2.1%、2.1%、1.8%和1.2%的性能提升，表现出精准检测的强大能力。同时，在预测精度、召回率与mAP@0.5上，HS-MambaDet也展现出全面优势，特别是在“公共汽车”和“自行车”等难以识别的类别上提升更为显著，充分证明本文所提方法生成的特征表示具备更强的判别性，能够有效应对复杂场景下多类别目标精确识别与定位的挑战。

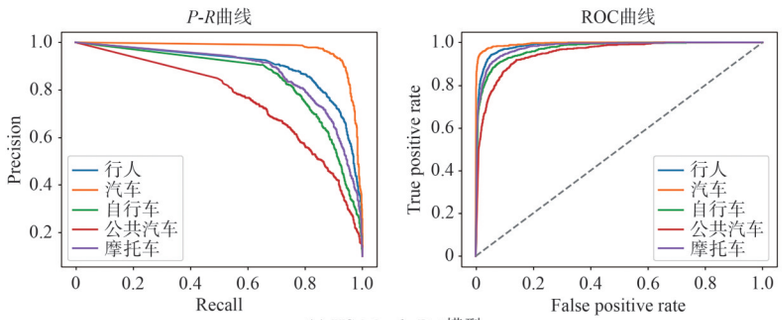
为从分类器判别能力角度进一步评估各模型在多类别目标上的性能，本文绘制了各类别的精确率-召回率（*P-R*）曲线与受试者工作特征（ROC）曲线，如图7所示。

在*P-R*曲线中，曲线下的面积（AUC）直观反映了模型在精度与召回率之间的均衡能力。可以观察到，本文所提HS-MambaDet模型的*P-R*曲线最为饱满，覆盖面积最大，尤其在召回率提升时其精度衰减最为缓慢，这表明该模型在保持高查全率的同时，依然能维持优异的查准性能，有效降低了误检风险。相比之下，CoAT模型表现次之，而ConvNeXt模型的曲线在低召回率后便迅速下降，表明其难以在复杂场景中同时兼顾高精度与高召回率。

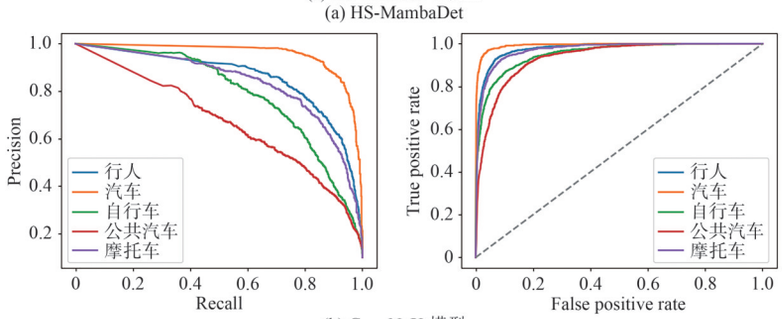
在ROC曲线中，HS-MambaDet的曲线最逼近左上角，其AUC值最高，证明该模型在所有类别上均拥有最优的二分类判别能力，能最好地区分前景目标与背景干扰。ViT与VMamba模型的ROC曲线表现相近且处于中游水平，这与表2中的整体性能排名相符。

曲线分析结果与前述定量指标高度一致，共同验证了HS-MambaDet所融合的全局上下文与局部细节特征能够显著提升模型的特征分辨力，从而在复杂多目标场景下实现更精准的类别识别。

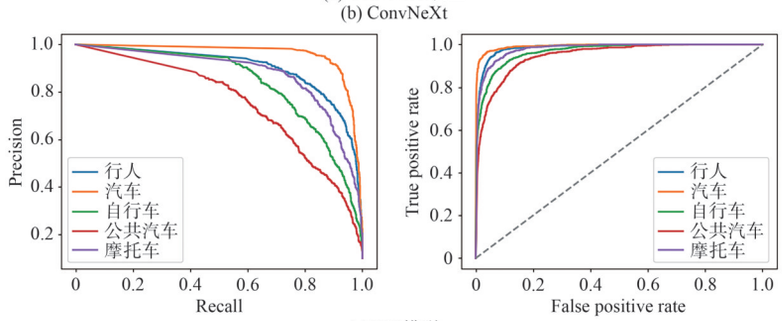
为进一步从类别维度剖析模型的判别能力，本文提供了归一化混淆矩阵的可视化结果（图8～图12）。对比分析可知，HS-MambaDet的混淆矩阵呈现出最集中、最显著的对角线分布，表明其在不同类别上的分类置信度最高，识别最准确。具体而言，本模型在“汽车”与“行人”等主要类别上保持了极高的查全率，同时显著降低了类别间的误判，例如“自行车”与“摩托车”这两个形态相似类别间的混淆程度为所有模型最低。相比之下，ConvNeXt与VMamba等模型则在对角线外



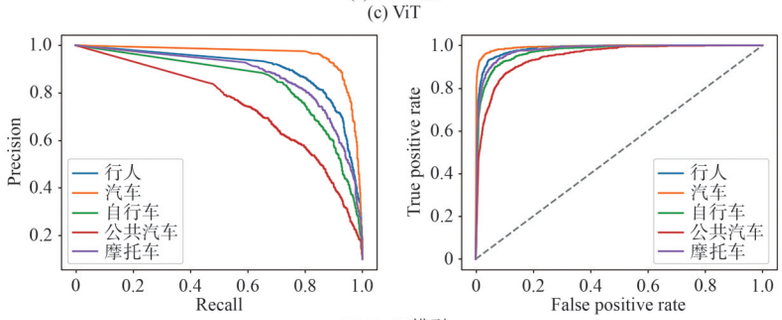
(a) HS-MambaDet模型
(a) HS-MambaDet



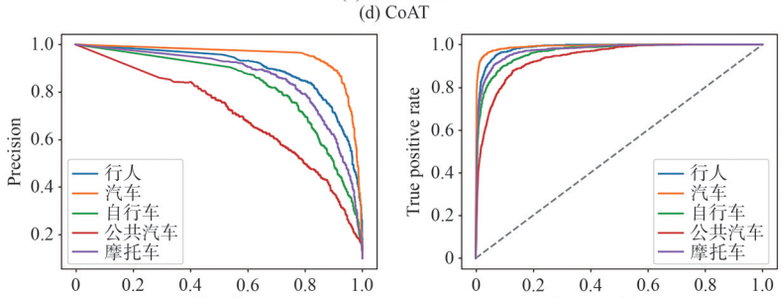
(b) ConvNeXt模型
(b) ConvNeXt



(c) ViT模型
(c) ViT



(d) CoAT模型
(d) CoAT



(e) VMamba模型
(e) VMamba

图 7 各模型的 *P-R* 曲线与 ROC 曲线

Fig. 7 *P-R* curves and ROC curves of the models

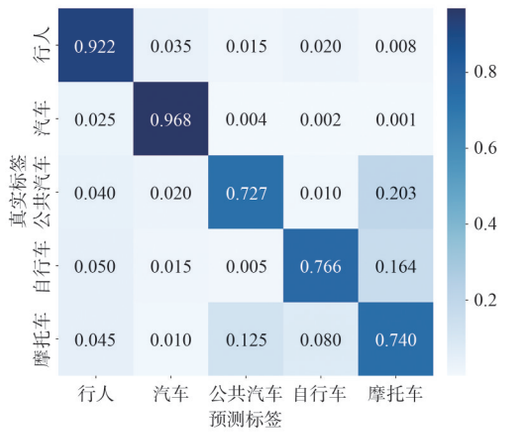


图 8　混淆矩阵(HS-MambaDet)

Fig. 8　Confusion matrix (HS-MambaDet)

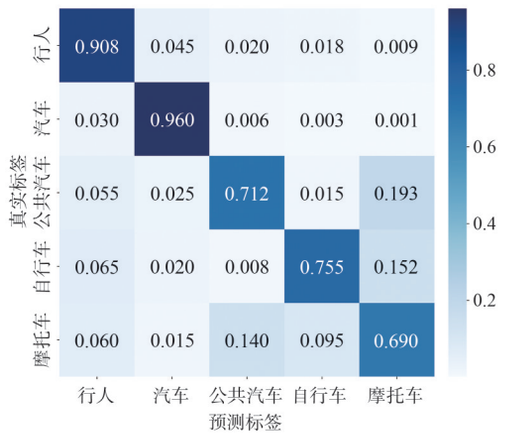


图 11　混淆矩阵(CoAT)

Fig. 11　Confusion matrix (CoAT)

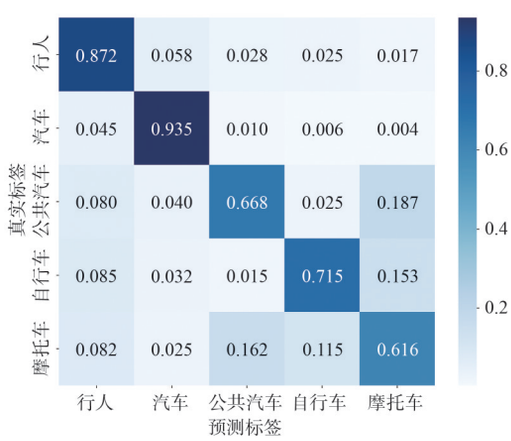


图 9　混淆矩阵(ConvNeXt)

Fig. 9　Confusion matrix (ConvNeXt)

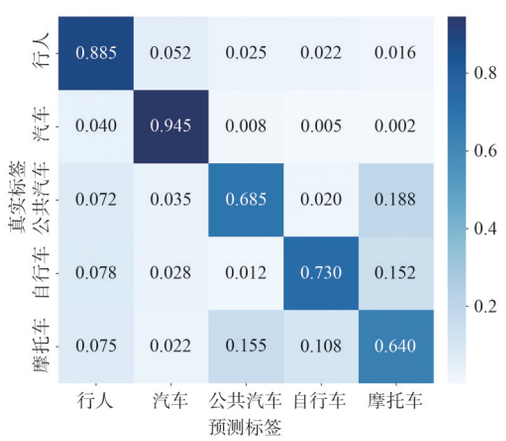


图 12　混淆矩阵(VMamba)

Fig. 12　Confusion matrix (VMamba)

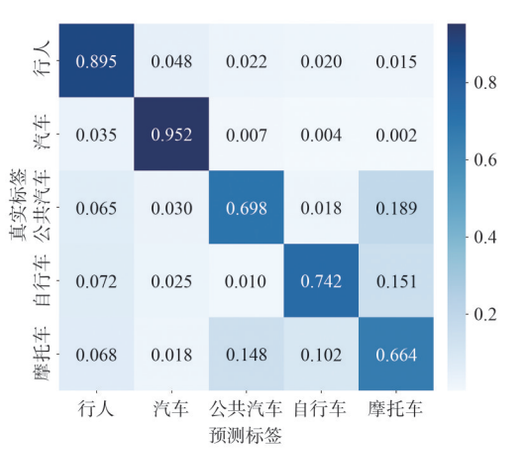


图 10　混淆矩阵(ViT)

Fig. 10　Confusion matrix (ViT)

出现更多显性响应，表明其存在更多的误识别情况。这一可视化证据与前述定量指标相互印证，共同证明了HS-MambaDet所融合的全局上下文与局部细节特征能够有效提升模型的特征分辨力，从而在复杂多目标场景下实现更精准的类别识别。

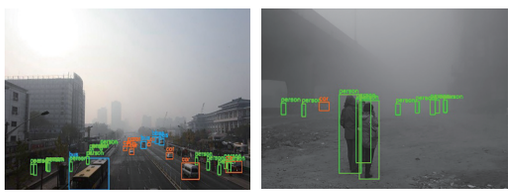
特别地，在小目标检测场景中，本文所提方法表现出更强的鲁棒性和细节感知能力。本文选取了存在较多小目标的场景图进行测试，从表7中可以看出，在小目标场景下HS-MambaDet的基于高频感知的检测方法相比于其他方法能够展现

表 7　小目标检测结果

Tab. 7　Small target detection results

模型	预测精度/%	召回率/%	mAP@0.5/%
HS-MambaDet	81.2	68.6	76.7
ConvNeXt	76.5	64.1	72.3
ViT	79.2	66.7	74.1
CoAT	80.5	67.8	75.0
VMamba	78.9	66.3	73.7

出巨大的优势。此外，本文将小目标场景下的测试样例进行部分可视化以便更加直观地对比不同算法的效果。如图 13 所示，对于远处的车辆、公共汽车以及行人，HS-MambaDet 以其独特的高频与空间感知能力能够有效地检测出较多目标，而其他模型则表现较差。综上所述，本文算法不仅显著提升了对多类别目标的综合检测精度，更在小目标场景中表现出优越的识别性能与稳定性，充分证明了其在实际应用中的潜力与价值。



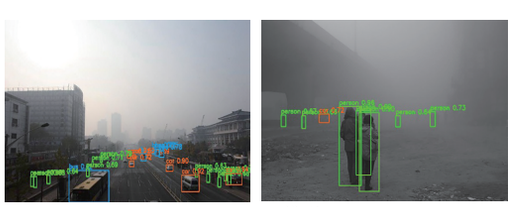
(a) 实图
(a) Ground truth



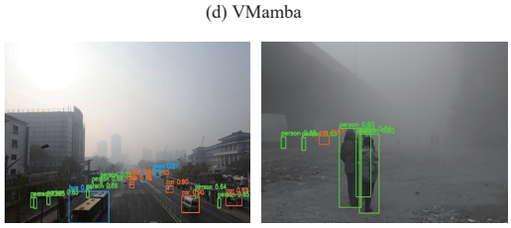
(b) HS-MambaDet模型
(b) HS-MambaDet



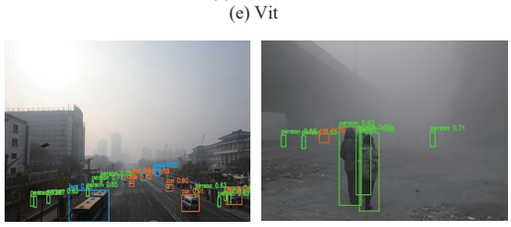
(c) ConvNeXt模型
(c) ConvNeXt



(d) VMamba模型
(d) VMamba



(e) Vit模型
(e) Vit



(f) CoAT模型
(f) CoAT

图 13 小目标检测可视化结果

Fig. 13 Small target detection visualization results

4.2.5 泛化性实验

为了验证所提 HS-MambaDet 模型在不同雾强度场景下的泛化能力，本文将训练完成的模型直接迁移至 CityScapes-foggy-low 和 CityScapes-foggy-high 数据集进行测试，并与 ConvNeXt、ViT、CoAT 与 VMamba 进行对比，实验数据如表 8～表 13 所示。

表 8 预测精度结果（CityScapes-foggy-low）

Tab. 8 Prediction accuracy results (CityScapes-foggy-low)

模型	预测精度/%				
	行人	汽车	公共汽车	自行车	摩托车
HS-MambaDet	78.3	81.4	61.9	65.4	71.5
ConvNeXt	70.8	75.0	55.1	60.2	60.9
ViT	73.0	77.4	57.1	62.4	65.1
CoAT	74.4	78.3	58.1	63.2	66.8
VMamba	73.3	77.1	56.0	62.0	63.2

表 9 召回率结果（CityScapes-foggy-low）

Tab. 9 Recall results (CityScapes-foggy-low)

模型	召回率/%				
	行人	汽车	公共汽车	自行车	摩托车
HS-MambaDet	60.1	64.4	45.4	47.6	54.1
ConvNeXt	56.0	58.0	40.6	42.8	49.0
ViT	57.9	61.6	42.3	44.8	51.8
CoAT	59.0	62.6	43.1	45.2	52.1
VMamba	58.3	61.2	42.1	44.4	50.2

表 10 mAP@0.5 结果（CityScapes-foggy-low）

Tab. 10 mAP@0.5 results (CityScapes-foggy-low)

模型	mAP@0.5/%				
	行人	汽车	公共汽车	自行车	摩托车
HS-MambaDet	65.1	69.6	44.5	49.2	52.9
ConvNeXt	60.1	64.0	40.0	44.0	47.0
ViT	62.0	65.8	41.7	45.5	49.2
CoAT	63.2	67.0	42.3	46.7	50.1
VMamba	61.5	65.6	41.2	45.1	48.7

表11　预测精度结果（CityScapes-foggy-high）

Tab. 11　Prediction accuracy results (CityScapes-foggy-high)

模型	预测精度/%				
	行人	汽车	公共汽车	自行车	摩托车
HS-MambaDet	75.5	79.1	58.2	62.3	68.9
ConvNeXt	64.3	71.2	52.3	58.1	58.8
ViT	68.2	76.1	54.0	59.4	63.2
CoAT	70.1	76.9	55.2	60.0	64.8
VMamba	66.3	76.4	53.7	58.6	62.7

表12　召回率结果（CityScapes-foggy-high）

Tab. 12　Recall results (CityScapes-foggy-high)

模型	召回率/%				
	行人	汽车	公共汽车	自行车	摩托车
HS-MambaDet	58.4	62.1	43.6	46.5	52.7
ConvNeXt	53.2	55.2	37.9	39.9	46.0
ViT	55.5	58.1	40.7	41.2	48.2
CoAT	56.4	59.3	41.2	42.3	49.1
VMamba	55.0	57.9	40.2	40.9	47.3

表13　mAP@0.5结果（CityScapes-foggy-high）

Tab. 13　mAP@0.5 results (CityScapes-foggy-high)

模型	mAP@0.5/%				
	行人	汽车	公共汽车	自行车	摩托车
HS-MambaDet	63.3	67.4	43.3	47.2	50.6
ConvNeXt	55.2	61.3	37.2	40.1	44.0
ViT	58.1	62.3	39.2	40.6	45.5
CoAT	59.3	64.1	41.3	43.1	47.7
VMamba	57.5	63.0	39.2	41.3	45.6

在低雾强度数据集（CityScapes-foggy-low）上，HS-MambaDet在5类目标上均取得最佳结果，尤其在公共汽车、自行车等小目标类别中表现更为突出，说明高频增强与空间依赖建模能够有效提升模型对细节弱化目标的识别能力。

在高雾强度数据集（CityScapes-foggy-high）上，各模型性能均进一步下降，尤其是卷积主导的ConvNeXt，其对纹理退化更为敏感。相比之下，HS-MambaDet仍保持明显优势，表明该模型在强雾与跨域条件下具有更强的鲁棒性。

此外，相较于其他模型，HS-MambaDet在CityScapes数据集上性能下降更少，浓雾场景相较于最优对比方法其精度指标最高提升5.4%，展现出的卓越泛化能力源于其多层次、结构化的特征提取与融合机制。该方法的核心优势在于其构建的域不变性特征表示：高频感知模块提取边缘、纹理等几何结构特征，这些特征受雾霾等外观变化的影响相对较小，从而在跨域场景中保持了稳定的判别能力；空间依赖感知模块则通过建模像素间的上下文关系，使其不依赖于容易变化的局部纹理，转而依赖更为鲁棒的全局空间结构进行推理。

此外，HSM-SSD骨干网络的选择性状态空间机制能够动态调整并关注关键信息，自适应地忽略域特异性噪声。最终，CNN-Transformer-Mamba构成的混合架构形成了功能冗余，当面对未知数据分布时，不同组件能够互补协作，确保了系统性能的优雅降级而非急剧恶化。因此，即使在不同雾强度的未知数据集上进行测试，该模型也能通过其内在的鲁棒性设计，最大限度地维持检测精度，展现出强大的实际部署适应潜力。

4.3　与主流及新兴SOTA模型的对比分析

为全面、公正地评估HS-MambaDet在复杂天气小目标检测任务上的性能，本节设计了系统的对比实验。所选对比模型旨在构建一个多维度的评估基准，主要包括：(1)广泛应用的工业级主流检测器，涵盖单阶段检测的典范YOLO系列（YOLOv5, YOLOv8）与两阶段检测的奠基模型Faster R-CNN；(2)新兴架构范式下的代表性模型，即近期将状态空间模型（Mamba）与检测框架融合的最新工作Mamba-YOLO。通过与此类模型的综合比较，本研究旨在从实际任务性能、架构创新性与模型效率等多个核心维度，充分验证所提HS-MambaDet方法的有效性与先进性。

4.3.1　综合检测性能分析

不同模型在RTTS数据集上的综合性能对比如表14所示。HS-MambaDet在所有核心检测精度指标上均展现出显著优势。

分析表14数据可知，HS-MambaDet的mAP@0.5达到了81.2%，相较于目前表现最优的主流检测器YOLOv8(80.9%)提升了0.3%。在要求更为严格的mAP@(0.5～0.95)指标上，HS-MambaDet取得了52.4%的结果，较YOLOv8(48.6%)领先3.8%（相对提升约7.8%）。这一优势表明，本模型在不同IoU阈值下的平均定位精度和鲁棒性更出色，这对于受雾霾干扰导致边界模糊的小目标检测至关重要。

表14 主流目标检测及视觉Mamba模型在RTTS数据集上的综合性能对比

Tab. 14 Comprehensive performance comparison with mainstream object detection and visual Mamba models on the RTTS dateset

模型	预测精度/%	召回率/%	mAP@0.5/%	推理时间/s	mAP@(0.5~0.95)/%	参数量/M	FLOPs/G
HS-MambaDet	87.3	73.1	81.2	0.26	52.4	45.2	45.6
YOLOv5	83.2	70.1	78.4	0.23	42.5	24.0	64.2
YOLOv8	85.7	71.8	80.9	0.22	48.6	28.6	78.9
Faster R-CNN	81.5	68.7	77.3	0.56	40.2	42.0	155.4
Mamba-YOLO	84.5	69.9	79.8	0.28	46.8	18.5	50.0

与同为Mamba架构新近工作的Mamba-YOLO相比，HS-MambaDet在mAP@0.5和mAP@(0.5~0.95)上分别领先1.4%和5.6%，证明本文所提出的CNN-Transformer-Mamba混合架构以及多尺度频空融合器(MFSF)相较于单纯的Mamba骨架替换策略，能更有效地融合局部细节、全局上下文与序列建模的优势，从而在复杂退化场景中生成判别力更强的特征表示。

在效率方面，HS-MambaDet保持了良好的实时性，其0.26 s的单帧推理时间与以速度见长的YOLO系列(0.22~0.23 s)处于同一水平，远快于Faster R-CNN (0.56 s)。尽管参数量(45.2M)高于部分轻量模型，但其计算量(FLOPs为45.6G)显著低于YOLOv8 (78.9G)和Faster R-CNN (155.4G)，甚至低于参数量更小的YOLOv5 (FLOPs为64.2G)。这验证了HSM-SSD骨干网络通过隐藏状态压缩有效降低了计算复杂度，实现了精度与效率的更好平衡。

4.3.2 多类别目标详细性能对比

为进一步剖析模型对不同类型、不同尺度目标的检测能力，表15~表17分别列出了行人、汽车、公共汽车、自行车和摩托车5类目标的预测精度、召回率及AP50详细结果。

表15 各类别预测精度对比

Tab. 15 Per-category prediction accuracy comparison

模型	预测精度/%				
	行人	汽车	公共汽车	自行车	摩托车
HS-MambaDet	92.2	96.8	72.7	76.6	74.0
YOLOv8	89.5	95.1	70.2	74.8	71.5
Mamba-YOLO	88.1	94.2	69.0	73.6	70.0
YOLOv5	87.0	93.0	67.8	72.4	68.9
Faster R-CNN	85.9	92.3	66.4	70.1	67.2

表16 各类别召回率对比

Tab. 16 Per-category recall comparison

模型	召回率/%				
	行人	汽车	公共汽车	自行车	摩托车
HS-MambaDet	77.4	79.8	56.7	59.2	65.4
YOLOv8	75.5	78.6	55.0	57.9	63.0
Mamba-YOLO	74.9	77.8	54.3	57.1	62.0
YOLOv5	74.0	77.1	53.1	55.8	61.2
Faster R-CNN	72.8	75.9	51.9	54.3	59.5

表17 各类别mAP@0.5对比

Tab. 17 Per-category mAP@0.5 comparison

模型	mAP@0.5/%				
	行人	汽车	公共汽车	自行车	摩托车
HS-MambaDet	85.9	91.0	66.6	69.9	74.3
YOLOv8	83.6	89.4	64.8	67.5	71.2
Mamba-YOLO	82.8	88.5	63.9	66.7	70.4
YOLOv5	81.7	87.6	62.8	65.4	69.3
Faster R-CNN	80.3	86.9	61.1	64.0	68.2

从分类结果可以得出：首先，HS-MambaDet在所有5个目标类别、所有3项评估指标上均取得了一致性的最优性能，证明其提升是普遍且稳健的，而非依赖于特定类别的特性。其次，对于“汽车”和“行人”这类尺寸相对较大、特征较明显的目标，所有先进模型都能达到较高水平，但HS-MambaDet仍能保持1%~2%的领先优势。更为重要的是，在“公共汽车”(尤其是远处小尺度公交车)、“自行车”和“摩托车”等尺度较小、形状狭长或易受遮挡的目标上，HS-MambaDet的优势被进一步放大。例如，在mAP@0.5指标上，相较于YOLOv8，HS-MambaDet在“公共汽车”和“自行车”类别上分别领先1.8%和2.4%。这直接归功于本文设计的高频感知模块(HFP)对边缘纹理等细节特征的增强，以及空间依赖感知模

块（SDP）所提供的跨尺度上下文信息，二者共同作用，有效缓解了小目标及弱纹理目标在雾天环境下的特征退化与湮没问题。

综上所述，本节通过与涵盖主流框架与新兴范式的多种SOTA模型进行系统性对比，充分证实了HS-MambaDet的卓越性能。其在综合检测精度、多类别目标识别能力以及精度-效率权衡方面均展现出显著优势，论证了本文所提方法在解决复杂天气小目标检测实际挑战中的有效性与先进性。

5　结　　论

本文提出了一种面向复杂天气与动态环境下的小目标检查的混合状态空间模型HS-MambaDet，该方法构建了一种高效的CNN-Transformer-Mamba混合架构：首先，设计融合隐藏状态压缩的HSM-SSD骨干结构，在保持全局依赖建模能力的同时显著降低计算复杂度；其次，引入包含高频感知模块与空间依赖感知模块的多尺度频空融合器（MFSF），以此缓解小目标特征弱与空间感知不足的问题，通过高通滤波和双通路注意力增强小目标的细节特征表达。实验结果表明，在RTTS和CityScapes数据集上，所提方法在检测精度与推理效率方面均优于现有主流模型，有效解决了动态恶劣环境下小目标检测鲁棒性差与实时性不足的难题，为实际应用提供了可靠解决方案。

参　考　文　献：

[1] 李士刚，王维佳，朱圣杰，等. 基于时空三维卷积网络的复杂背景下红外弱小目标检测方法[J]. 液晶与显示，2025，40(11)：1688-1699.
LI S G, WANG W J, ZHU S J, *et al*. Detection method for infrared dim and small targets under complex backgrounds via a spatio-temporal three-dimensional convolutional network [J]. *Chinese Journal of Liquid Crystals and Displays*, 2025, 40(11): 1688-1699. (in Chinese)

[2] 张洋，孙海江，张笑闻，等. 基于局部上下文增强的快速3D弱特征目标检测方法[J]. 液晶与显示，2025，40(11)：1675-1687.
ZHANG Y, SUN H J, ZHANG X W, *et al*. Fast 3D weak feature object detection method based on local context enhancement [J]. *Chinese Journal of Liquid Crystals and Displays*, 2025, 40(11): 1675-1687. (in Chinese)

[3] 牛海鹏，颜昌翔，王一霖，等. 基于快速局部对比度和目标特征的星图弱小目标检测算法[J]. 液晶与显示，2024，39(1)：69-78.
NIU H P, YAN C X, WANG Y L, *et al*. Star map spatial target detection method based on fast local contrast and target features[J]. *Chinese Journal of Liquid Crystals and Displays*, 2024, 39(1): 69-78. (in Chinese)

[4] 张志豪，厉小润，陈淑涵. 基于改进YOLO11的无人机航拍图像小目标检测算法[J]. 液晶与显示，2025，40(6)：915-930.
ZHANG Z H, LI X R, CHEN S H. Small object detection algorithm in UAV aerial images based on improved YOLO11 [J]. *Chinese Journal of Liquid Crystals and Displays*, 2025, 40(6): 915-930. (in Chinese)

[5] 王栋，罗亚波，张峰，等. CORM-YOLO：低照度环境下交通行人检测算法[J]. 液晶与显示，2025，40(9)：1369-1380.
WANG D, LUO Y B, ZHANG F, *et al*. CORM-YOLO: pedestrian detection algorithm in low-light conditions for traffic environments [J]. *Chinese Journal of Liquid Crystals and Displays*, 2025, 40(9): 1369-1380. (in Chinese)

[6] LUO Y H, CAO X, ZHANG J T, *et al*. CE-FPN: enhancing channel information for object detection [J]. *Multimedia Tools and Applications*, 2022, 81(21): 30685-30704.

[7] REN S Q, HE K M, GIRSHICK R, *et al*. Faster R-CNN: towards real-time object detection with region proposal networks [C]//*Proceedings of the 29th International Conference on Neural Information Processing Systems*. Montreal: MIT Press, 2015: 91-99.

[8] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, *et al*. An image is worth 16×16 words: Transformers for image

recognition at scale [C]//*Proceedings of the 9th International Conference on Learning Representations*. Online: OpenReview.net, 2021.

[9] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks [C]//*Proceedings of the 26th International Conference on Neural Information Processing Systems*. Lake Tahoe: Curran Associates Inc., 2012: 1097-1105.

[10] LIU Z, MAO H Z, WU C Y, *et al*. A ConvNet for the 2020s [C]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans: IEEE, 2022: 11966-11976.

[11] RADOSAVOVIC I, KOSARAJU R P, GIRSHICK R B, *et al*. Designing network design spaces [C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020: 10425-10433.

[12] TAN M X, LE Q V. EfficientNet: rethinking model scaling for convolutional neural networks [C]//*Proceedings of the 36th International Conference on Machine Learning*. Long Beach: PMLR, 2019: 6105-6114.

[13] KIM J, KONG J, SON J. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech [C]//*Proceedings of the 38th International Conference on Machine Learning*. Online: PMLR, 2021: 5530-5540.

[14] TOUVRON H, CORD M, JÉGOU H. DeiT Ⅲ: revenge of the ViT [C]. 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer, 2022: 516-533.

[15] 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*. Montreal: IEEE, 2021: 9992-10002.

[16] WANG W H, XIE E Z, LI X, *et al*. Pyramid vision Transformer: a versatile backbone for dense prediction without convolutions [C]. 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal: IEEE, 2021: 548-558.

[17] XU W J, XU Y F, CHANG T, *et al*. Co-scale conv-attentional image Transformers [C]//*Proceedings of the IEEE/CVF International Conference on Computer Vision*. Montreal: IEEE, 2021: 9961-9970.

[18] CHEN C F R, FAN Q F, PANDA R. CrossViT: cross-attention multi-scale vision Transformer for image classification [C]//*Proceedings of the IEEE/CVF International Conference on Computer Vision*. Montreal: IEEE, 2021: 347-356.

[19] LI J S, XIA X, LI W, *et al*. Next-ViT: next generation vision Transformer for efficient deployment in realistic industrial scenarios [J/OL]. *arXiv*, 2022: 2207.05501.

[20] LI Y Y, YUAN G, WEN Y, *et al*. EfficientFormer: vision Transformers at MobileNet speed [C]//*Proceedings of the 36th International Conference on Neural Information Processing Systems*. New Orleans: Curran Associates Inc., 2022: 940.

[21] HATAMIZADEH A, HEINRICH G, YIN H X, *et al*. FasterViT: fast vision Transformers with hierarchical attention [C]//*Proceedings of the 12th International Conference on Learning Representations*. Vienna: OpenReview.net, 2024.

[22] ZHU L H, LIAO B C, ZHANG Q, *et al*. Vision Mamba: efficient visual representation learning with bidirectional state space model [C]//*Proceedings of the 41st International Conference on Machine Learning*. Vienna: OpenReview.net, 2024: 62429-62442.

[23] LIU Y, TIAN Y J, ZHAO Y Z, *et al*. VMamba: visual state space model [C]//*Proceedings of the 38th International Conference on Neural Information Processing Systems*. Vancouver: Curran Associates Inc., 2024: 3273.

[24] PEI X H, HUANG T, XU C. EfficientVMamba: atrous selective scan for light weight visual Mamba [C]//*Proceedings of the 39th AAAI Conference on Artificial Intelligence*. Philadelphia: AAAI, 2025: 6443-6451.

[25] PATRO B N, AGNEESWARAN V S. SiMBA: simplified Mamba-based architecture for vision and multivariate time series [J/OL]. *arXiv*, 2024: 2403.15360.

[26] 王伯霄,宋延嵩,董小娜. 难点注意力感知红外小目标检测网络[J]. 中国光学(中英文),2024,17(3):538-547.

WANG B X, SONG Y S, DONG X N. Indistinguishable points attention-aware network for infrared small object detection [J]. *Chinese Optics*, 2024, 17(3): 538-547. (in Chinese)

[27] 何梦云，何自芬，张印辉，等. 用于水下声呐目标检测的弱特征共焦通道调控方法[J]. 中国光学(中英文)，2024，17(6)：1281-1296.
HE M Y, HE Z F, ZHANG Y H, *et al*. Weak feature confocal channel regulation for underwater sonar target detection [J]. *Chinese Optics*, 2024, 17(6): 1281-1296. (in Chinese)

作者简介：



陈　悦，男，硕士研究生，2022年于江苏科技大学获得学士学位，主要研究方向为计算机视觉、目标检测等。E-mail：13338939052@163.com



许　锋，男，博士，教授，三级警监，2006年于东北大学获得博士学位，主要研究方向为视频侦查、刑事影像等。E-mail：7677759@qq.com