文章编号　1004-924X(2024)05-0727-13

位置敏感 Transformer 航拍图像目标检测模型

李大湘，辛嘉妮*，刘　颖
(西安邮电大学 通信与信息工程学院，陕西 西安 710121)

摘要：针对无人机视角下航拍图像小目标多且检测困难的问题，提出了一个位置敏感 Transformer 目标检测(PS-TOD)模型。设计了一个基于位置通道嵌入三维注意力(PCE3DA)的多尺度特征融合(MSFF)模块，即 PCE3DA 利用空间与通道信息的相互依赖关系生成三维注意力，用于加强模型对兴趣区域的特征表达能力，且基于它构造了一个自底向上的跨层 MSFF 方案，使得融合后的特征语义信息更加丰富；然后，设计了一种新的位置敏感自注意力(PSSA)机制，且以此构造位置敏感 Transformer 编-解码器，使模型在捕获图像全局上下文信息的长期依赖关系时，也可提高模型对目标的位置敏感能力。基于无人机航拍数据集 VisDrone 的对比实验结果表明，提出模型的 AP 达到 28.8%，与基线模型(DETR)相比提高了 4.1%。该模型在复杂背景下能对无人机航拍图像进行精确的目标检测，且改善小目标的检测效果。

关　键　词：目标检测；无人机图像；位置敏感 Transformer；多尺度特征融合；注意力机制
中图分类号：V279；TP394.1　　**文献标识码**：A　　**doi**：10.37188/OPE.20243205.0727

Position-sensitive Transformer aerial image object detection model

LI Daxiang，XIN Jiani*，LIU Ying

(*College of communication and information engineering*，*Xi'an University of Posts and Telecommunication*，*Xi'an* 710121，*China*)
* *Corresponding author*，*E-mail*：*xjn_*2000@163.*com*

Abstract：Addressing the challenge of detecting numerous small objects in UAV-captured aerial images，this paper introduces the Position-Sensitive Transformer Target Detection (PS-TOD) model. Initially，it presents a multi-scale feature fusion (MSFF) module incorporating a Positional Channel Embedded 3D Attention (PCE3DA) mechanism. PCE3DA leverages the interplay between spatial and channel data to generate 3D attention，enhancing feature representation in areas of interest. This foundation supports a bottom-up，cross-layer MSFF approach，augmenting the semantic richness of combined features. Subsequently，it proposes a novel Position-Sensitive Self-Attention (PSSA) mechanism，leading to the development of a position-sensitive Transformer encoder-decoder. This innovation heightens the model's sensitivity to target positioning，facilitating the capture of long-term dependencies within the image's global context. Comparative tests using the VisDrone dataset reveal that the PS-TOD model attains an Average Precision (AP) of 28.8%，marking a 4.1% enhancement over the baseline model (DETR). Furthermore，it demonstrates precise object detection in UAV aerial imagery against complex backdrops，significantly boosting the detection accuracy of small targets.

收稿日期：2023-05-30；**修订日期**：2023-07-15.
基金项目：国家自然科学基金资助项目(No. 62071379)；陕西省自然科学基金资助项目(No. 2019JM-604)；西安邮电大学研究生创新基金资助项目(No. CXJJZL2022003)

Key words: object detection; unmanned aerial vehicle image; position sensitive Transformer; multi-scale feature fusion; attention mechanism

1 引　言

随着飞行器与通信技术的快速发展,无人机作为一种新型的拍摄工具,凭借独特的拍摄视角,以及携带方便与成本低的特点,在民用和军事方面得到了广泛的应用[1]。面向无人机航拍影像,为了提高用户对航拍内容的观看效率,基于机器学习技术设计无人机航拍图像目标检测算法已经成为当今计算视觉领域中的一个新兴研究分支[2]。

近年来,深度学习作为无人机航拍图像目标检测的主流方法,根据是否使用锚框相关算法可分为两大类。基于锚框(Anchor-based)的代表性算法有Faster R-CNN[3],Cascade R-CNN[4],SSD[5]与YOLOv4[6]等。针对航拍图像目标检测的应用需求,Yang等[7]提出了用于小目标检测的QueryDet网络,设计了一种简单有效的级联稀疏查询机制,有效地利用航拍图像高分辨率特征,提高对小目标的检测性能。Li等[8]提出了一种Oriented RepPoints空中目标检测方法,通过引入灵活的自适应点,能够捕捉任意方向实例的几何信息。Liang等[9]提出了一个称之为DEA-Net的动态锚点增强网络,该网络实现了基于锚的单元和无锚单元之间的交互式样本筛选,以生成合格样本,提高检测小目标的性能。这类基于锚框的方法虽然在航拍图像目标检测中取得了较好的性能,但在检测过程中要依赖于人工预先设置的锚框信息,不仅会增加模型超参的数量(如:锚框的数量、尺寸与高宽比等),还会增大参数调试的复杂性,即无法通过反向传播进行端到端训练,通常需要人为仔细地调整锚框参数才能获得最佳的检测性能。

在无锚框方法的研究上,Law等[10]提出的CornerNet算法先预测目标左上角和右下角点,再对角点分类组合形成检测框。Tian等[11]提出的FCOS算法针对每个图像像素进行预测,得到该像素到检测框的4个边框的距离,最终输出整体目标的检测框。Dai等[12]提出了ACE空中旋转目标检测方法,使用四边形边界框来定位任意方向对象和动态采样方法,有助于关键点的准确定位。除了这些方法之外,近两年来,由于Transformer在计算机视觉领域的广泛应用,Carion等[13]将它整合到目标检测基线中,设计了一种DETR的目标检测算法,该算法不需要任何的人工干预,可以用端到端的方式进行训练。Zhu等[14]提出了一种Deformable DETR的目标检测算法,设计了可变形注意力模块,该模块只注意参考点周围的某些采样点,减少了计算量。Li等[15]通过引入带有噪声的真实边界框作为查询向量,通过去噪技术解决二分图匹配的不稳定性问题,加速模型训练。基于Transformer的方法框架简洁,不用手工设置锚框及非极大值抑制(Non-Maximum Suppression, NMS),泛化能力强,建模图像的全局依赖关系,有效利用上下文信息,减少由于锚框设置不合理导致的问题,但需要一些特殊的损失函数提高算法稳定性,小目标的检测性能相对较差。

综上所述,Transformer框架下的DETR虽然具有思想简洁、结构清晰与无NMS操作等优点,但因无人机拍摄距离过远,小目标过多,现有模型很难取得理想的检测效果。所以,本文设计了一种位置敏感Transformer目标检测(Position Sensitive Transformer Object Detection, PSTOD)模型。该模型在DETR的基础上,设计了一个基于位置通道嵌入三维注意力(Position Channel Embedding 3D Attention, PCE3DA)的多尺度特征融合(Multi-Scale Feature Fusion, MSFF)模块,且将该模块连接在骨干网络和Transformer之间,让网络更好地获取具有多层级上下文信息的特征,以增强模型对小目标的检测能力;此外,设计了位置敏感自注意力(Position Sensitive Self-Attention, PSSA)机制,用它替代原模型中的自注意力(Self Attention, SA),即使用可学习的相对位置敏感编码信息,帮助Transformer模型中的编-解器获得更准确的目标位置信息,以提高无人机航拍图像目标的定位能力及检测精度。

2　模型设计

2.1　PS-TOD 模型架构

图 1 是本文设计的 PS-TOD 模型示意图，它主要由 CNN 主干网络、MSFF 模块、位置敏感 Transformer 编-解码器与集合匹配预测模块 4 个组件构成。对于待检测图像，首先使用 CNN 主干网络与 MSFF 模块，获得图像的跨层融合多尺度特征；然后，采用带有 PSSA 机制的 Transformer 编码器，对图像的多尺度特征连同其相对位置信息一起进行学习，获得图像的位置敏感编码特征；其次，在 Transformer 解码器中再通过多头 SA 及交叉注意力将对象查询向量转换为解码输出；最后，利用两个不同的 FFN 对解码器输出的每个特征进行预测，分别得到它们所对应的框坐标和类标签，以获得最终的目标预测集合。

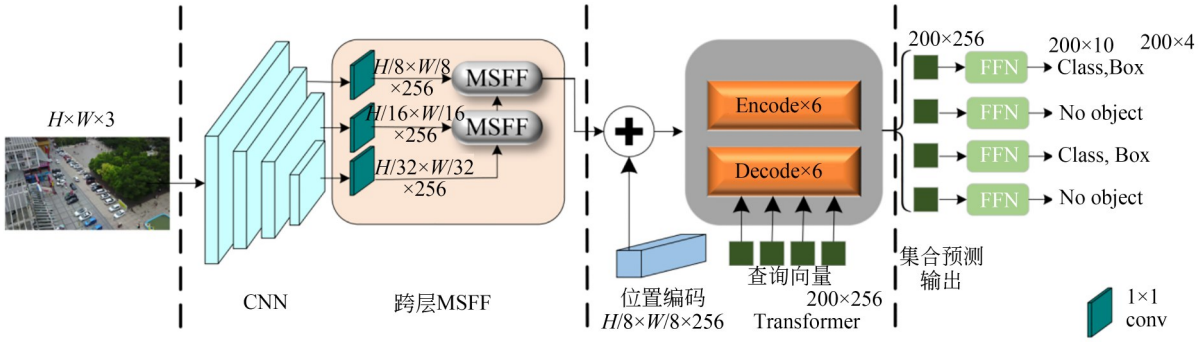


图 1　PS-TOD 模型示意图

Fig. 1　Schematic diagram of PS-TOD model

2.2　跨层多尺度特征融合模块

小目标数量多作为无人机航拍图像目标检测的主要挑战。DETR 算法[13]因只使用 ResNet 最后一个卷积模块 conv5_x 的输出作为特征表示，即特征图谱经 32 倍下采样后，导致原图中的小目标消失在特征图中从而造成漏检。所以，本文设计了 PCE3DA，且基于它构造了一个自底向上的跨层 MSFF 模块，在提高小目标检测精度的同时还可兼顾整个算法对多尺度目标的检测能力。

设 IMG 表示任意一幅训练图像，将它送入主干网络 ResNet-50，conv3_x，conv4_x 与 conv5_x 输出的特征图谱分别记作 F_3，F_4 与 F_5，且使用 1×1 卷积将它们的通道数均调整为 256，分别记为 $\bar{F}_3$，$\bar{F}_4$ 与 $\bar{F}_5$。为了将它们的信息融合起来而得到图像的多尺度特征表示，设计了一个自下而上的跨层特征融合方案，即图 1 中的 MSFF 模块。

2.2.1　多尺度特征融合

以特征 $\bar{F}_4$ 与 $\bar{F}_5$ 为例，为了将这两种不同分辨率的特征图谱融合起来，如图 2 所示，以 $\bar{F}_4$ 与 $\bar{F}_5$ 作为输入，对于 $\bar{F}_5$ 先采用转置卷积完成上采样得到 $\bar{F}_{5u}$，此时 $\bar{F}_{5u}$ 和 $\bar{F}_4$ 具有相同的分辨率。与传统的上采样方法相比，转置卷积具有可学习的参数，可通过训练学习来获取最优的上采样结果，缓解传统方法特征细节丢失的问题，进而提升特征融合效果。然后，采用相加操作融合 $\bar{F}_{5u}$ 和 $\bar{F}_4$，记为：

$$F_a = \bar{F}_{5u} + \bar{F}_4. \tag{1}$$

将融合后的 F_a 通过设计的 PCE3DA 进行加权得到加权特征 $\bar{F}_a$，即：

$$\bar{F}_a = \mathrm{PCE3DA}(F_a). \tag{2}$$

为了保留特征的初始信息，使用残差连接将自适应增强的特征与其原始特征分别相加。因此，获得增强特征 $\bar{F}_4^{\mathrm{en}}$ 和 $\bar{F}_5^{\mathrm{en}}$，分别为：

$$\bar{F}_4^{\mathrm{en}} = \bar{F}_a + \bar{F}_4, \tag{3}$$

$$\bar{F}_5^{\mathrm{en}} = \bar{F}_a + \bar{F}_5. \tag{4}$$

最后，分别通过 3×3 卷积层后再进行特征相加融合，获得跨层融合特征 $\bar{F}_{45}$，即：

$$\bar{F}_{45} = \mathrm{Conv3}(\bar{F}_4^{\mathrm{en}}) + \mathrm{Conv3}(\bar{F}_5^{\mathrm{en}}). \tag{5}$$

在得到 $\bar{F}_{45}$ 后，同理可进一步将 $\bar{F}_{45}$ 和 $\bar{F}_3$ 进行融合，从而获得最终的多尺度特征 F_{ms}。

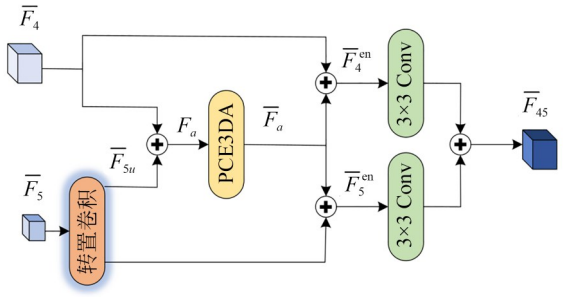


图 2　PCE3DA 跨层特征图谱融合方案示意图

Fig. 2　Fusion scheme of PCE3DA cross layer feature map

2.2.2　PCE3DA 原理

为了更好地提取无人机图像的特征信息，传统方法是分别对特征图谱实施空间与通道注意力，这类方法导致参数与计算量大，且不能同时考虑空间维度和通道维度之间的相互关系，导致空间和通道信息相互孤立。如图 3 所示，在坐标注意力[16]的启发下设计了 PCE3DA，式(2)中采用 PCE3DA 进行注意力加权，即：将空间位置信息嵌入到通道注意力中，这样可以同时利用空间和通道维度的相互依赖信息，得到三维注意力权值，用于加强感兴趣区域的特征表示，以帮助模型聚焦有助于目标精准定位的局部细节信息。

设 $F\in\mathbf{R}^{C\times H\times W}$ 表示任意输入 PCE3DA 的特征图谱，其中 C，H 与 W 分别表示 F 的通道数、高度与宽度。首先，使用一个 X 轴的 1×1 卷积对 F 中的数据沿水平方向进行聚合，在捕获 X 轴长距离依赖关系的同时，也可以保留垂直方向的位置信息，该过程可表示为：

$$z^X=Conv1\_X(F), \tag{6}$$

其中 $z^X\in\mathbf{R}^{C\times H\times 1}$ 表示卷积结果。然后，将 z^X 送入 1×1 卷积，且经过归一化与激活函数处理，得到：

$$f^X=\sigma(\mathrm{BN}(\mathrm{Conv1}(z^X))), \tag{7}$$

其中：σ 表示 Swish 非线性激活函数，BN( )表示批量归一化，$f^X\in\mathbf{R}^{C/r\times H\times 1}$ 表示垂直方向上对空间信息进行编码的中间特征图。这里，r 表示压缩通道比例(实验中 r=4)；随后，利用另外一个 1×1 卷积，将 f^X 变换并与输入特征图 F 的通道数相同，记为：

$$g^X=\mathrm{Conv1}(f^X), \tag{8}$$

其中 $g^X\in\mathbf{R}^{C\times H\times 1}$ 表示通道扩充结果。

同理，采用另一个 Y 轴的 1×1 卷积对 F 中的数据沿垂直方向进行聚合，在捕获 Y 轴长距离依赖关系的同时，也可以保留水平方向的位置信息，该过程可表示为：

$$\begin{cases} z^Y=\mathrm{Conv1\_Y}(F) \\ f^Y=\sigma(\mathrm{BN}(\mathrm{Conv1}(z^Y))) \\ g^Y=\mathrm{Conv1}(f^Y) \end{cases}. \tag{9}$$

综上所述，将 g^X 与 g^Y 作广播机制加法 $\oplus$，再经 Sigmoid 函数处理之后，记为：

$$\beta=\mathrm{Sigmoid}(g^X\oplus g^Y), \tag{10}$$

其中 $\beta\in\mathbf{R}^{C\times H\times W}$，表示三维注意力权值。最后，将权值 β 与输入 F 点乘 $\otimes$，从而得到经 PCE3DA 加权之后特征 $\bar{F}$，记为：

$$\bar{F}=\beta\otimes F. \tag{11}$$

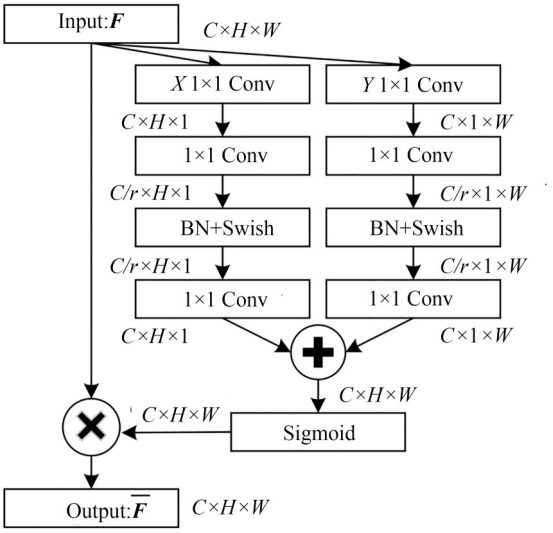


图 3　位置通道嵌入三维注意力流程

Fig. 3　Flow chart of position channel embedding 3D attention

2.3　位置敏感 Transformer 编-解码器

对于目标检测任务，位置信息极为重要。在 DETR 算法中，采用绝对位置编码感知图像的全局上下文信息，但在目标检测中图像的分辨率通常很高，目标特征更多依赖图像的局部信息。因此，本文设计了一种 PSSA 机制，且以此构造位置敏感 Transformer 编-解码器，以提高模型对位置信息的敏感能力，从而提升目标检测精度。

2.3.1　PSSA 机制

为了利用每个元素在序列中的位置信息，提高它在计算机视觉任务中的表达能力，传统的做法是将绝对位置编码 AP=[p_1; p_2; …; p_N]嵌入

到序列 X 的每个元素 $\boldsymbol{x}_i$（如 ViT[17]）中，即：

$$\bar{\boldsymbol{x}}_i = \boldsymbol{x}_i + \boldsymbol{p}_i, \tag{12}$$

其中 $\boldsymbol{p}_i \in \mathbf{R}^{d_x}$ 表示第 i 个元素的绝对位置编码向量，通常可采用正余弦函数计算得到[17]。最后，绝对位置编码 SA 可表示为：

$$\begin{aligned}\mathrm{asSA}(\boldsymbol{X}) &= [Q;\ K;\ V] = \\ &[(\boldsymbol{X}+\boldsymbol{P})\boldsymbol{W}^Q;(\boldsymbol{X}+\boldsymbol{P})\boldsymbol{W}^K;(\boldsymbol{X}+\boldsymbol{P})\boldsymbol{W}^V].\end{aligned} \tag{13}$$

在目标检测任务中，像素之间的相对位置信息对于提高模型对目标的定位能力尤其重要。如图 4 所示，这里利用序列各元素之间的相对位置信息，设计了一种 PSSA 机制，即通过嵌入可学习的相对位置编码向量到 SA 机制中，利用图像中各特征之间的相对位置关系，提高模型的位置敏感能力，从而实现目标的精确定位。

设 $\boldsymbol{F}_{\mathrm{ms}} \in \mathbf{R}^{C\times H\times W}$ 表示经 MSFF 模块得到的多尺度特征图谱，其中 C，H 与 W 分别表示通道数、高度与宽度。首先，对 F_{ms} 中每个位置 (h,w) 沿通道维度的 C 个数据抽取出来，由此可将 F_{ms} 转化成一个由 N（这里 $N=W\times H$）个元素组成序列，记为 $S=\{s_n(h,w)|n=1,2,\cdots,N\}$，其中 $s_n(h,w)\in\mathbf{R}^{1\times C}$ 表示第 n 个元素，$h\in[1,H]$ 与 $w\in[1,W]$ 分别表示它在 F_{ms} 中对应的空间位置坐标；然后，为了建模 $s_n(h,w)$ 相对于 S 中任意其他元素 $s_m(h,w)$ 之间的相对位置关系，定义一个索引函数 $E(n,m)$ 与 3 个相对位置编码向量，记为：

$$\begin{cases} p^Q_{nm} = \alpha^Q_{E(n,m)} \\ p^K_{nm} = \alpha^K_{E(n,m)} \\ p^V_{nm} = \alpha^V_{E(n,m)} \\ E(n,m) = \min\left(\left|s^h_n - s^h_m\right| + \left|s^w_n - s^w_m\right|, T\right) \end{cases}, \tag{14}$$

其中：s^h_n，s^h_m，s^w_n 与 s^w_m 分别表示这两个元素在高度与宽度方向的坐标值，p^Q_{nm}，p^K_{nm} 与 p^V_{nm} 分别表示对应于查询（Query，Q）、键（Key，K）、值（Value，V）的相对位置编码，T 表示邻域阈值，即如果序列中两个元素之间的城区距离超过 T，这两元素之间的位置信息就没有意义。T 取图像高与宽度之和的 20%，即 $T=\mathrm{Round}[(W+H)/5]$。在模型训练过程中，自动学习 $\boldsymbol{P}^Q=[\alpha^Q_0,\ldots,\alpha^Q_T]$、$\boldsymbol{P}^K=[\alpha^K_0,\ldots,\alpha^K_T]$ 与 $\boldsymbol{P}^V=[\alpha^V_0,\ldots,\alpha^V_T]$ 这 3 组相对位置编码向量，其中 $\alpha^Q_i,\alpha^K_i,\alpha^V_i\in\mathbf{R}^C$，$i=0,1,\ldots,T$。

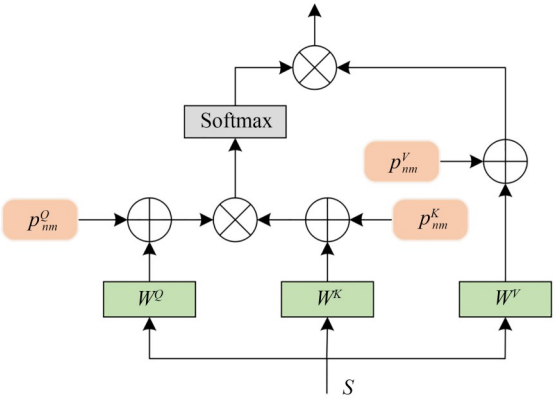


图 4　位置敏感自注意力机制

Fig. 4　Position sensitive self-attention mechanism

综上所述，在输入序列 S 中，根据两个元素 $s_n(h,w)$ 与 $s_m(h,w)$ 之间的城区距离，为了学习它们之间的相对位置依赖关系，需额外考虑 3 个与位置相关的向量，即在 Query，Key 与 Value 上分别加入相对位置编码 p^Q_{nm}，p^K_{nm} 与 p^V_{nm} 构成 PSSA，记为：

$$\begin{aligned}\mathrm{PSSA}(\mathrm{S}) &= [Q;\ K;\ V] = \\ &[(S+P^Q)W^Q;(S+P^K)W^K;(S+P^V)W^V],\end{aligned} \tag{15}$$

其中：W^Q，W^K，$W^V\in\mathbf{R}^{C\times C'}$ 分别表示与 Q，K，V 相对应的且可学习的变换矩阵。C 与 C' 分别表示输入、输出特征的维度，则对于 S 中的任意一个元素 $s_n(h,w)\in\mathbf{R}^{1\times C}$，其 PSSA 编码过程可表示为：

$$\omega_{nm} = \frac{\exp\left(\frac{1}{\sqrt{d}}(\boldsymbol{Q}_n\boldsymbol{K}_m^{\mathrm{T}})\right)}{\sum_{m=1}^{N}\exp\left(\frac{1}{\sqrt{d}}(\boldsymbol{Q}_n\boldsymbol{K}_m^{\mathrm{T}})\right)}, \tag{16}$$

$$z_n = \sum_{m=1}^{N}\omega_{nm}V_m, \tag{17}$$

其中：$z_n\in\mathbf{R}^{1\times C'}$ 表示 PSSA 编码输出，ω_{nm} 表示使用缩放点积与 SoftMax 计算的归一化权重[18]。

2.3.2　位置敏感 Transformer 编-解码器

基于 PSSA 机制，在 DETR 算法[13]的启发下，设计的位置敏感 Transformer 编-解器如图 5 所示，它主要由编码器与解码器两部分组成。为了使模型在性能与参数量之间得到一个很好的平衡，如图 5 左侧所示，编码器由 6 个相同的层构

成，且每个层主要由多头 PSSA 与 MLP 组成。对于输入序列 S，将其写成矩阵形式 $S_0 \in \mathbf{R}^{N\times C}$，记为：

$$S_0=[s_1(h,w);\ s_2(h,w);\ \cdots;\ s_N(h,w)]. \tag{18}$$

编码器重构特征的过程可表示为：

$$\begin{cases} Z'_L=(S_{L-1}+\mathrm{mhPSSA}(S_{L-1})) \\ Z_L=\mathrm{LN}(Z'_L+\mathrm{MLP}(Z'_L)),\quad L=1\cdots 6 \\ Y=Z_6 \end{cases} \tag{19}$$

其中：LN( )，MLP( )与 mhPSSA( )分别表示层归一化、多层感知机与多头 PSSA 等操作，$Y\in\mathbf{R}^{N\times C}$ 表示第 6 层编码器的输出，即对序列 S 的最终编码结果。mhPSSA 作为 PSSA 的扩展，即并行地运行 K 个不同的 PSSA 操作，每个注意力头将分别关注输入信息的不同部分，并将它们的输出串联起来作为最终的编码结果：

$$\mathrm{mhPSSA}(z)=[\mathrm{PSSA}_1(z);\ \cdots;\ \mathrm{PSSA}_K(z)]. \tag{20}$$

为了使 S 经 mhPSSA 编码之后，其输入 S 与输出 Y 保持相同的维度，每个 PSSA 输出的维度 C' 设置为输入元素维度的 K 分之一，即 $C'=C/K$。为了与 DERT 模型进行公平比较，本文模型中的 C 也与其一样也设置为 256，且为了保证 K 能整除 C，K 只能取 2，4，8，16 等整数。随着注意力头数的增加，模型计算复杂度会增加，所以本文后续实验中 K 取 4，一则可以在计算效率和性能之间达到折中；二则由于设计的模型面向无人机航拍图像目标检测，mhPSSA 机制中的每个头将从不同的角度感知目标的不同部分。这些目标按 4 个角度观察也可满足要求，例如车的车头和车尾、人体的头部和身体等，mhPSSA 将从 4 个角度感知这些目标，且捕捉它们之间的语义关系而提取图像的全局特征，从而能够提高目标检测的准确率。

MLP 包括两个 FC 层，FC1 层将输入扩大为原来的 4 倍，由于残差连接的存在，FC2 输出层再恢复原始维度，相应的计算过程为：

$$\mathrm{MLP}(x)=\mathrm{ReLU}(x\boldsymbol{W}_1+\boldsymbol{b}_1)\boldsymbol{W}_2+\boldsymbol{b}_2, \tag{21}$$

其中：$\boldsymbol{W}_1$ 表示将特征从 256 维投影到 1 024 维的变换矩阵，$\boldsymbol{W}_2$ 表示从 1 024 维投影回 256 维变换矩阵，$\boldsymbol{b}_1$ 与 $\boldsymbol{b}_2$ 均表示偏置向量。

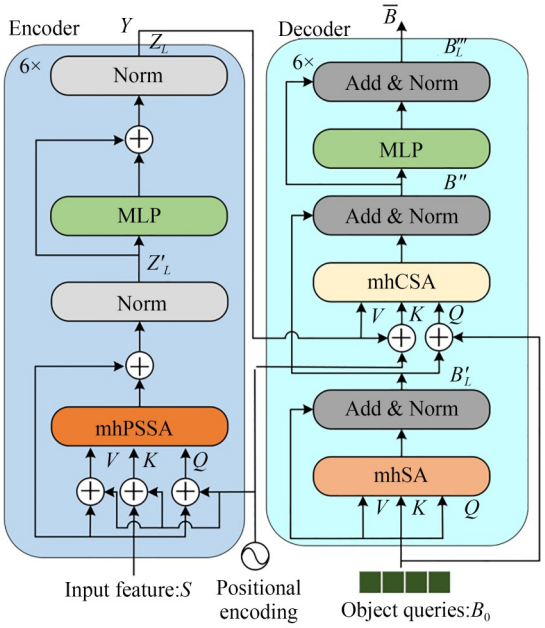


图 5　编-解码器结构

Fig. 5　Encoder-decoder structure

如图 5 右侧所示，解码器类似于 Transformer 的标准结构，由 6 个完全相同的层构成，每个层主要由多头 SA、多头交叉 SA 与 MLP 组成。设 $\boldsymbol{B}_0=[\boldsymbol{b}_1;\ \boldsymbol{b}_2;\ \cdots;\ \boldsymbol{b}_M]$ 表示由 M 个元素组成的目标查询(object query)序列，其中 $\boldsymbol{b}_i\in\mathbf{R}^{1\times C}$ 表示 B 中的第 i 个元素，对应的是图像中第 i 个预测目标的特征向量。编码器的第一个阶段是先采用多头 SA 对进行编码，然后，将输出与编码器的输出 Y 相结合，再采用多头交叉 SA 进行编码；最后，经类似于编码器的 MLP 处理，得到最终的解码特征 $\bar{B}$。该过程描述为：

$$\begin{cases} B'_L=\mathrm{LN}(B_{L-1}+\mathrm{mhSA}(B_{L-1})) \\ B''_L=\mathrm{LN}(B'_L+\mathrm{mhCSA}(Y;B'_L;P^K)) \\ B'''_L=\mathrm{LN}(B''_L+\mathrm{MLP}(B''_L)),L=1\cdots 6 \\ \bar{B}=B'''_6 \end{cases}, \tag{22}$$

其中 mhCSA( )表示由 K 个交叉自注意力 CSA( )组成的多头交叉自注意力，即：

$$\begin{cases} mhCSA(Y;B'_L;P^K)=[\mathrm{CSA}_1;\ \cdots;\ \mathrm{CSA}_K] \\ \mathrm{CSA}=[K;Q;V]= \\ \left[(Y+P^K)W^K;(B'_L+B_{L-1})W^Q;YW^V\right] \end{cases}. \tag{23}$$

2.4　集合预测与损失函数

对于目标查询序列 $B_0=[b_1;\ b_2;\ \cdots;\ b_M]$，经

解码器输出得到 $\bar{B}=[\bar{b}_1;\bar{b}_2;\cdots;\bar{b}_M]$，再将它们输入两个不同的 FFN，以分别预测每个解码特征所对应目标的类别标签 $\overline{\mathrm{cls}}_i$ 与边框 $\overline{\mathrm{box}}_i$，得到预测结果记为 $\bar{U}=\left\{\bar{u}_i=(\overline{\mathrm{cls}}_i,\overline{\mathrm{box}}_i)\right\}_{i=1}^{M}$，相应训练图像所有真实目标的类别 cls_i 与边框 box_i 的 Ground Truth 集合记为 $U=\left\{u_i=(\mathrm{cls}_i,\mathrm{box}_i)\right\}_{i=1}^{J}$，实验中 M 设置为 200，通常远远大于图像中真实目标的数量 J。在 Transformer 这种端到端的目标检测框架中，因不需要 NMS 后处理，训练时就得在 U 与 $\bar{U}$ 二个集合之间寻找最佳匹配[13]。为了便于用匈牙利算法[19]在集合 U 与 $\bar{U}$ 中找到最佳匹配，首先，将集合 U 填充 $M-J$ 个 $\varnothing$（表示无目标），使它与 $\bar{U}$ 元素数量相等，对于 U 中的每个 $u_i=(\mathrm{cls}_i,\mathrm{box}_i)$，其中 cls_i 是目标类标签（可能是 $\varnothing$），$\mathrm{box}_i\in[0,1]^4$ 是其相对于图像尺寸的中心坐标及高度与宽度；然后，要在集合 U 与 $\bar{U}$ 之间寻找最佳匹配，就是要寻找 $\bar{U}$ 中 M 个元素的最佳置换 $\sigma\in\xi_M$，使式(25)所示的匹配损失最小，即：

$$\hat{\sigma}=\arg\min_{\sigma\in\xi_M}\sum_{i}^{M}L_{\mathrm{match}}(u_i,\bar{u}_{\sigma(i)}),\tag{24}$$

其中：$L_{\mathrm{match}}(u_i,\bar{u}_{\sigma(i)})$ 表示真实 u_i 与索引为 $\sigma(i)$ 预测 $\bar{u}_{\sigma(i)}$ 之间的匹配损失。在预测结果 $\bar{U}$ 中对于索引为 $\sigma(i)$ 的 $\bar{u}_{\sigma(i)}$，定义它属于 cls_i 类别的概率为 $\bar{p}_{\sigma(i)}(\mathrm{cls}_i)$，预测框为 $\overline{box}_{\sigma(i)}$。类似于常见的目标检测器，该匹配损失由分类损失 L_{cls} 和边框回归损失 L_{box} 线性组成，即：

$$L_{\mathrm{match}}(u_i,\bar{u}_{\sigma(i)})=\begin{cases}L_{\mathrm{cls}}(\bar{p}_{\sigma(i)}(\mathrm{cls}_i))+L_{\mathrm{box}}(box_i,\overline{box}_{\sigma(i)}), & \mathrm{cls}_i\neq\varnothing\\ 0, & \mathrm{cls}_i=\varnothing\end{cases}.\tag{25}$$

对于分类损失 $L_{\mathrm{cls}}(\bar{p}_{\sigma(i)}(\mathrm{cls}_i))$，由于实验用到的 VisDrone 数据集[20]存在各类样本数量比例不平衡，标签分布不均匀的问题，本文使用 Focal Loss[21]作为分类损失，即：

$$L_{\mathrm{cls}}(\bar{p}_{\sigma(i)}(\mathrm{cls}_i))=-\alpha_t(1-\bar{p}_{\sigma(i)}(\mathrm{cls}_i))^{\gamma}\log(\bar{p}_{\sigma(i)}(\mathrm{cls}_i)),\tag{26}$$

其中 α_t 与 γ 为超参数，分别表示第 cls_i 类的权重与衰减参数，$\alpha_t=0.25$，$\gamma=2$。

对于回归损失函数 $L_{\mathrm{box}}(box_i,\overline{box}_{\sigma(i)})$，采用 DIOU 损失[22]，即在 IOU 基础上加入度量真实框和预测框之间中心点的距离的惩罚项，定义如下：

$$L_{\mathrm{box}}(\mathrm{box}_i,\overline{\mathrm{box}}_{\sigma(i)})=1-\mathrm{IOU}+\frac{\rho^2(\mathrm{box}_i^c,\overline{\mathrm{box}}_{\sigma(i)}^c)}{c^2},\tag{27}$$

其中：box_i^c 和 $\overline{\mathrm{box}}_{\sigma(i)}^c$ 分别表示真实框与预测框的中心点的坐标，ρ 表示计算两个框中心点之间的欧氏距离，c 表示能够同时包含预测框和真实框的最小封闭框的对角线距离。

3　实验结果及分析

3.1　数据集与实验方法

为了验证本文提出的 PS-TOD 模型的有效性，本文使用公开的 VisDrone 数据集[20]进行对比实验。该数据集由天津大学 AISKYEYE 团队使用无人机在不同条件下低空拍摄获得，包含 1 360×765 和 960×540 像素两种图像尺寸，涵盖各种天气和光照条件下日常生活中的各种场景，其中训练集 6 471 张图像、测试集 3 190 张图像和验证集 548 张图像。数据集的图像中包括行人、人、汽车、公交车、自行车、卡车、三轮车、雨棚三轮车、面包车以及摩托车等十类目标。

本文采用 COCO 数据集中的评价指标来评价模型性能[23]，主要比较 AP，AP_{50}，AP_{75}，AP_{S}，AP_{M} 与 AP_{L}，其中 AP 表示在 0.5 至 0.95 步长 0.05 共 10 个交并比阈值下的平均检测精度的平均值，AP_{50} 与 AP_{75} 分别表示交并比阈值为 0.5 和 0.75 时的平均检测精度，AP_{S}，AP_{M} 与 AP_{L} 分别表示对测试集中的小目标（像素数量$<32^2$）、中等目标（$32^2<$像素数量$<96^2$）与大目标（像素数量$>96^2$）的平均检测精度。实验平台采用 Ubuntu18.04 操作系统，GPU 为 NVIDIA TITANX×4，CPU 为 Intel(R)Core(TM) Xeon E5-2640，内存为 128 GB，编程语言为 Python3.8，torch 版本为 1.7.0。模型训练过程中使用 AdamW 优化器来优化模型，批大小（Batch_size）为 16，初始学习率为 2×10^{-4}，权值衰减为 1×10^{-4}，整个模型训练 500 个 Epoch，为了加快训练收敛速度，在初始训练时使用官方提供的 Transformer 预训练模型。所有实验均以 VisDrone 的训练集与验证集来完成模型的训练，然后对测试集中的所有图像进行目标检测，统计相应评价指标。

3.2 消融实验

3.2.1 模块消融实验

为了验证PS-TOD模型中两个关键模块(即基于PCE3DA的MSFF模块与基于PSSA的Transformer编-解码模块)以及修改损失函数在无人机航拍图像目标检测中的有效性,基于VisDrone数据集进行了消融实验,且在相同实验条件下,再与基线模型DETR[13]进行对比,消融实验结果如表1所示。其中"Param"表示模型的参数量,单位取"兆(M)",即当不同模块被嵌入到"基线"模型之后,以对比改进模型参数量的变化。

表1 VisDrone测试集上的消融实验结果

Tab. 1 Ablation experiment results on VisDrone test set (%)

方法	MSFF	PSSA	Loss	AP_S	AP_M	AP_L	AP	Param/M
基线	—	—	—	13.8	36.8	47.5	24.7	41.30
	√	—	—	16.4	38.9	49.4	26.4	42.36
	—	√	—	15.0	37.6	48.7	25.8	41.45
	—	—	√	15.6	39.1	48.9	26.0	41.30
	√	√	—	17.1	39.7	49.8	27.2	42.51
	—	√	√	16.5	40.0	49.1	26.9	41.45
	√	—	√	18.5	39.6	50.1	28.1	42.36
Ours	√	√	√	19.4	40.1	50.9	28.8	42.51

由表1实验结果可见,在基线模型的基础上,分别只应用MSFF,PSSA的Transformer编-解码或修改损失函数等部件,其AP分别提高了1.7%,1.1%或1.3%,这说明本文所设计的两个模块与修改损失函数在无人机图像目标检测任务中是有效的;若同时使用其中任意二个模块,较之只使用一个模块检测精度可得到进一步提高,当同时使用三个部件时,AP达到最高28.8%。通过对各类目标的检测结果分析可知,MSFF模块通过类似于残差连接的方式进行多尺度特征融合,且在PCE3DA的驱动下,模型在具备多尺度特征提取能力的基础上,还可更好地保留小目标的特征信息;设计的PSSA机制,较之原始的自注意力更能获取像素之间的相对位置关系,在位置敏感的作用下,模型可以更好地关注图像中的重点区域,并且在修改损失函数的约束下,不仅缓解了数据集类别以及正负样本不平衡带来的问题,同时使损失函数更加关注边界框的位置,更能优化模型的训练而提高无人机图像中目标的检测精度。虽然设计的模块可提高目标检测精度,但是会带来参数量的增加,例如:当MSFF或PSSA模块分别被引入之后,较之"基线"模型,会带来2.4M或3.3M参数量的增加,同时引入MSFF与PSSA模块时,模型参数量达到42.51M。

3.2.2 PCE3DA机制消融实验

为了验证设计的PCE3DA机制在MSFF模块中的有效性,设计了7组消融实验,即在A组(Baseline DETR[13])的基础上,B,C,D,E与F组分别表示基于-SE(SENet[24]的SE通道注意力),-SA(BAM[25]的空间注意力),-CA(文献[16]的坐标注意力),-CBAM(文献[26]的通道和空间注意力)与-PCE3DA(本文设计的)等5种不同的注意力机制,对骨干网络的最后一层特征图谱进行注意力加权;G组表示在F组的基础上还采用MSFF进行多尺度特征融合,再结合Baseline模型中编-解码器与检测头。消融实验结果如表2所示。

表2 不同注意力机制及使用多尺度特征的实验结果

Tab. 2 Experimental results for different attention mechanisms and using multi-scale features (%)

组别	方　法	AP_S	AP_M	AP_L	AP
A	Baseline	13.8	36.8	47.5	24.7
B	Baseline-SE	13.9	37.0	47.5	24.9
C	Baseline-SA	14.5	38.1	47.7	25.2
D	Baseline-CA	14.3	37.7	48.3	25.4
E	Baseline-CBAM	14.6	37.5	48.1	25.2
F	Baseline-PCE3DA	15.2	38.4	48.7	25.7
G	F+MSFF	16.4	38.9	49.4	26.4

由表2可知,骨干网络的特征图谱只要经注意力加权之后,不同尺寸目标的检测精度均可得到提高,且空间注意力要优于通道注意力。总体上,本文设计的PCE3DA(即F组)优于其他4种注意力,并且经MSFF模块对多层级特征图谱进行融合,检测效果达到最优(即G组)。这主要得

益于 PCE3DA 能将特征更好地聚焦在感兴趣区域，抑制无关信息，同时增强了特征表达与空间位置结构信息，融合后的特征图具有更丰富的语义信息和几何细节信息。

3.2.3　PSSA 机制消融实验

在 Transformer 编-解码中，为了验证设计的 PSSA 机制的性能，与文献[27]及[28]计算相对位置编码的方法进行了消融实验，实验结果如表 3 所示。可以看出，在计算注意力得分时考虑两个元素之间的相对位置，即引入相对位置编码是必要的。本文所提相对位置计算方法最大程度提升了模型的 AP 值，其主要原因是 PSSA 通过定义的索引函数映射相对位置，使得到的相对位置编码信息更加准确，模型能够获得一定的平移不变性，更加符合目标检测任务的需求。

表 3　不同相对位置计算方法的实验结果

Tab. 3　Experimental results of different relative position calculation methods　　(%)

方　法	AP_S	AP_M	AP_L	AP
基线模型	13.8	36.8	47.5	24.7
文献[27]	14.3	37.0	48.3	25.0
文献[28]	14.6	37.4	48.1	25.1
PSSA	15.0	37.6	48.7	25.8

3.3　综合对比实验

为了进一步验证本文提出的 PS-TOD 模型在无人机航拍图像目标检测任务中的性能，在 VisDrone 数据集上与经典及先进的目标检测模型进行实验对比，包括 Cascade R-CNN[4]、YOLOv8[32]与 PVTv2[33]等方法。为了对比的公平性，每种算法除了其专门参数沿用原文之外，学习率、批大小与 Epoches 等超参设置均与 3.1 节相同，实验结果如表 4 所示。

根据表 4 的数据，本文设计的 PS-TOD 模型在无人机航拍图像目标检测中表现良好，其 AP_{50}，AP_{75}与 AP 值分别达到了 51.8%，28.3%与 28.8%。与 YOLOv8(速度最快)相比，虽然 FPS 有所下降，但 YOLOv8 识别物体位置的精准性差，而 PS-TOD 的检测精度获得了 2.3% 的提升；与具有相近检测精度的 QueryDet 模型相比，PS-TOD 的准确率 AP 和检测速度 FPS 都高于该模型。但 AP_{75}较之低了 0.5，原因是 AP_{75}指标对于目标检测框的重合率要求更高，PS-TOD 模型作为一种无锚框引导的检测方法，在目标定位精确方面可能稍弱于专门针对小目标优化的 QueryDet 模型，但与其他模型相比，PS-TOD 在 AP_{75}方面仍然具有明显的优势，即 PS-TOD 能较好地平衡检测精度与检测速度。综上所述，通过对比实验结果可知，在设计的 PS-TOD 模型中，首先基于 PCE3DA 机制构造自底向上的跨层 MSFF 模块，可让网络更好地获取图像的上下文多尺度特征，在提高小目标检测精度的同时，还可兼顾多尺度目标的检测能力；然后，基于 PSSA 机制设计的 Transformer 编码器，可使用像素之间的相对位置信息，增强模型的位置敏感能力，提高了无人机航拍图像目标的定位能力及检测精度。

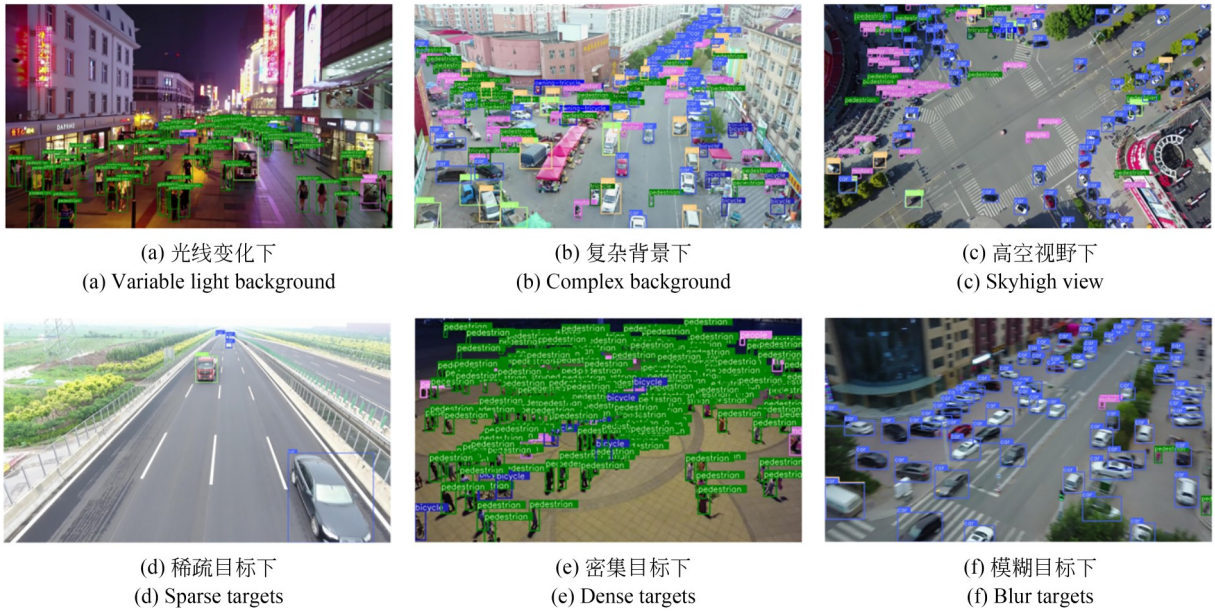
表 4　不同算法在 VisDrone 测试集上的性能对比

Tab. 4　Performance comparison of different algorithms on VisDrone test set　　(%)

方　法	AP_{50}	AP_{75}	AP	FPS
Faster R-CNN[3]	21.7	/	/	15.9
Cascade R-CNN[4]	38.6	25.0	23.5	9.0
YOLOv4[6]	31.2	16.7	16.8	28.8
QueryDet[7]	48.1	**28.8**	28.3	2.8
CornerNet[10]	34.1	15.8	17.4	15.5
RetinaNet[20]	28.4	12.3	11.3	16
Double-Head RCNN[29]	38.3	24.8	23.8	6.5
IterDet[30]	36.8	20.3	20.4	11.4
RSOD[31]	43.3	27.1	25.4	28
YOLOv8[32]	46.4	27.5	26.5	**30.1**
PVTv2[33]	34.1	21.4	20.6	10.9
PS-TOD(Ours)	**51.8**	28.3	**28.8**	22.7

为了观察 PS-TOD 模型在无人机航拍图像目标检测中的具体表现，图 6 为可视化 VisDrone 测试集中各种情况下具有代表性的图像检测结果。可以看出，本文模型在光照变化、复杂背景、高空拍摄视角、目标稀疏、目标密集与运动模糊等 6 种不同的环境下，均能够检测出大多数的目标，说明设计的 PS-TOD 模型对无人航拍机图像在各种情况下都具有非常优秀的检测能力，足以应对生活中发生的各类实际情况。

除此之外，为了进一步观察PS-TOD对每类目标的检测性能，分别统计了它与基线模型DETR[13]对VisDrone测试集中每类目标的具体检测性能，如表5所示。对比结果表明，PS-TOD总体上改善了单类目标的平均检测精度，尤其是对于小目标，提升效果非常明显。在小目标比例较多的行人、人、自行车与摩托车这四类目标中，相较于基线模型检测精度分别提升了4.2%，3.7%，2.6%与3.5%；另外，在目标尺寸相对较大的类别（如汽车类和卡车类）中同样也有明显优势，如汽车类别别的AP高达64.3%。综合各种尺寸目标的检测效果，充分验证了本文提出的PS-TOD模型在提高小目标检测精度的同时，还可兼顾其他尺度的目标检测能力。



(a) 光线变化下
(a) Variable light background

(b) 复杂背景下
(b) Complex background

(c) 高空视野下
(c) Skyhigh view

(d) 稀疏目标下
(d) Sparse targets

(e) 密集目标下
(e) Dense targets

(f) 模糊目标下
(f) Blur targets

图6　PS-TOD在VisDrone测试集上的部分检测结果

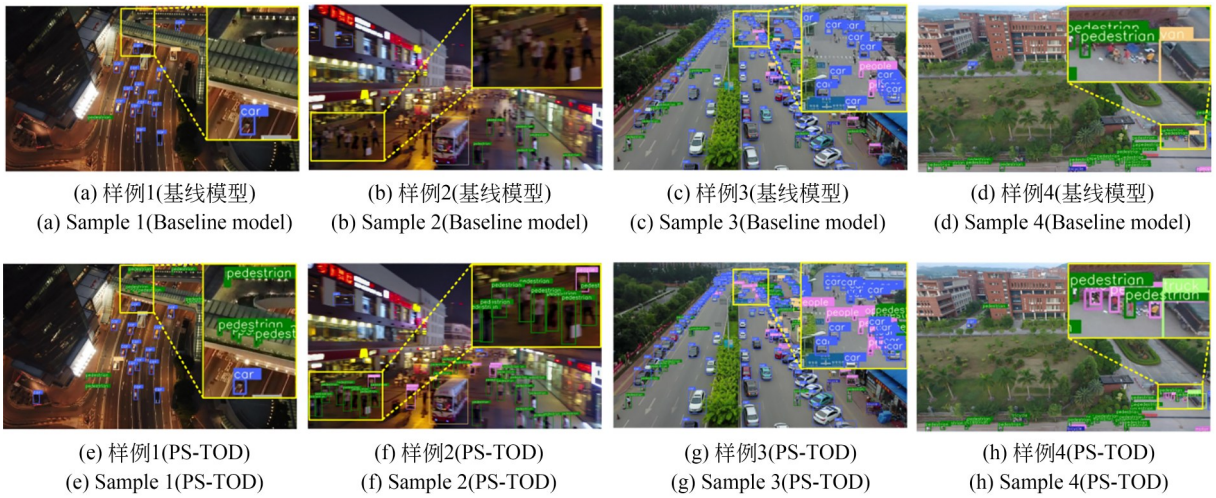
Fig. 6　Partial detection results of PS-TOD on VisDrone test set

表5　VisDrone测试集中不同类别实验结果

Tab. 5　Experimental results of different categories on VisDrone test set　(%)

目标类别	行人	人	汽车	公交车	自行车	卡车	三轮车	雨棚三轮车	面包车	摩托车
基线模型	24.8	18.7	61.6	35.2	12.1	23.3	15.2	4.6	28.6	24.9
PS-TOD	29.0	22.4	64.3	45.9	14.7	27.1	21.4	9.0	31.7	28.4

为了更深入观察基线DETR[13]模型与PS-TOD在小目标检测中的性能优劣，在VisDrone测试集中选取小目标存在的夜晚和白天等两种场景，如图7所示，可视化得到4组检测效果的对比图。通过对比图7(a)与图7(e)、图7(b)与图7(f)，在夜间较低照明的状态下，基线模型由于背景噪声信号的影响漏检了图7(a)中站立在高架桥上的行人与图7(b)中大量行人，而PS-TOD通过注意力机制，减少背景影响、增加感兴趣目标的特征信息，成功检测到基线漏检的行人；对比图7(c)与图7(g)、图7(d)与图7(h)可以发现，基线模型漏检了图7(c)大量远处的人及汽车与图7(d)中路口处的人及行人等小目标，而PS-TOD通过融合多尺度信息与添加位置信息，使得模型得到更好的目标特征信息，强化模型对小目标的定位能力，可精确检测部分漏检的小目标。总之，PS-TOD相对于基线模型具有更为优越的检测性能，尤其针对较小尺寸目标具有更强的检测辨别能力，有效降低小目标漏检与误检的概率。



(a) 样例1(基线模型)　(b) 样例2(基线模型)　(c) 样例3(基线模型)　(d) 样例4(基线模型)
(a) Sample 1(Baseline model)　(b) Sample 2(Baseline model)　(c) Sample 3(Baseline model)　(d) Sample 4(Baseline model)

(e) 样例1(PS-TOD)　(f) 样例2(PS-TOD)　(g) 样例3(PS-TOD)　(h) 样例4(PS-TOD)
(e) Sample 1(PS-TOD)　(f) Sample 2(PS-TOD)　(g) Sample 3(PS-TOD)　(h) Sample 4(PS-TOD)

图 7　小目标检测效果对比

Fig. 7　Comparison of small object detection result

4　结　论

针对无人机航拍图像小目标多且检测困难的问题，本文在 Transformer 框架下提出了一个 PS-TOD 模型。首先，设计了基于 PCE3DA 的多尺度特征融合模块，即通过融合不同层级的特征图谱，有效地利用它们在空间与通道二个维度中的上下文信息，以增加骨干网络的多尺度特征提取能力；然后，结合相对位置编码，设计了 PSSA 机制，且以此构造了一个 Transformer 编-解码器，以帮助模型在捕获图像全局上下文信息的长期依赖关系时，也可提高模型对位置信息的敏感能力，从而提升模型对小目标的检测精度。基于 VisDrone 数据集的实验结果表明，所提 PS-TOD 作为一种端到端的目标检测模型，其检测过程不需要事先锚框设置与事后 NMS 处理，在复杂背景下能精确地对无人机航拍图像进行目标检测，且有效地改善了小目标的检测效果。在后续工作中，除了进一步优化 PSSA 机制，以降低模型的参数量，提高检测速度之外，还需要将研究成果应用到其他数据集中，进一步验证所提模型的检测精度与泛化能力。

参考文献：

[1] 朱威，王立凯，靳作宝，等. 引入注意力机制的轻量级小目标检测网络[J]. 光学 精密工程，2022，30(8)：998-1010.
ZHU W，WANG L K，JIN Z B，*et al*. Lightweight small object detection network with attention mechanism[J]. *Optics and Precision Engineering*，2022，30(8)：998-1010.(in Chinese)

[2] 范丽丽，赵宏伟，赵浩宇，等. 基于深度卷积神经网络的目标检测研究综述[J]. 光学 精密工程，2020，28(5)：1152-1164.
FAN L L，ZHAO H W，ZHAO H Y，*et al*. Survey of target detection based on deep convolutional neural networks[J]. *Optics and Precision Engineering*，2020，28(5)：1152-1164.(in Chinese)

[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] CAI Z W，VASCONCELOS N. Cascade R-CNN：delving into high quality object detection[C]. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition. June* 18-23，2018. *Salt Lake City*，*UT*. IEEE，2018：6154-6162.

[5] LIU W，ANGUELOV D，ERHAN D，*et al*. *SSD*：*Single Shot Multibox Detector*[M]. Computer Vision-ECCV 2016. Cham：Springer International Publishing，2016：21-37.

[6] BOCHKOVSKIY A，WANG C Y，LIAO H. *YOLOv4*：*Optimal Speed and Accuracy of Object Detection* [EB/OL]. *arXiv preprint* arXiv：

2004.10934, 2020.

[7] YANG C, HUANG Z H, WANG N Y. QueryDet: cascaded sparse query for accelerating high-resolution small object detection[C]. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 18-24, 2022. *New Orleans*, *LA*, *USA*. IEEE, 2022: 13658-13667.

[8] LI W T, CHEN Y J, HU K X, *et al*. Oriented RepPoints for aerial object detection[C]. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 18-24, 2022. *New Orleans*, *LA*, *USA*. IEEE, 2022: 1829-1838.

[9] LIANG D, GENG Q X, WEI Z Q, *et al*. Anchor retouching via model interaction for robust object detection in aerial images[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2022, 60: 1-13.

[10] LAW H, DENG J. CornerNet: detecting objects as paired keypoints[J]. *International Journal of Computer Vision*, 2020, 128(3): 642-656.

[11] TIAN Z, SHEN C H, CHEN H, *et al*. FCOS: fully convolutional one-stage object detection[C]. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). *October* 27-*November* 2, 2019. *Seoul*, *Korea* (*South*). IEEE, 2019: 9626-9635.

[12] DAI P W, YAO S Y, LI Z K, *et al*. ACE: anchor-free corner evolution for real-time arbitrarily-oriented object detection[J]. *IEEE Transactions on Image Processing*, 2022, 31: 4076-4089.

[13] CARION N, MASSA F, SYNNAEVE G, *et al*. *End-to-end Object Detection with Transformers*[M]. Computer Vision-ECCV 2020. Cham: Springer International Publishing, 2020: 213-229.

[14] ZHU X Z, SU W J, LU L W, *et al*. Deformable DETR: deformable transformers for end-to-end object detection[C]. *Proceedings of the International Conference on Learning Representations* (*ICLR*), 2021: 1-14.

[15] LI F, ZHANG H, LIU S L, *et al*. DN-DETR: accelerate DETR training by introducing query DeNoising[C]. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 18-24, 2022. *New Orleans*, *LA*, *USA*. IEEE, 2022: 13609-13617.

[16] HOU Q B, ZHOU D Q, FENG J S. Coordinate attention for efficient mobile network design[C]. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), 2021: 13713-13722.

[17] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, *et al*. An image is worth 16x16 words: transformers for image recognition at scale[C]. *Proceedings of the International Conference on Learning Representations* (*ICLR*), 2021: 15-35.

[18] VASWANI A, SHAZEER N, PARMAR N, *et al*. Attention is all you need[C]. *Proceedings of the Advances in neural information processing systems* (*NeurIPS*), 2017: 6000-6010.

[19] KUHN H W. The Hungarian method for the assignment problem[J]. *Naval Research Logistics Quarterly*, 1955, 2(1/2): 83-97.

[20] ZHU P F, WEN L Y, DU D W, *et al*. *Vision Meets Drones: Past, Present and Future*[EB/OL]. arXiv preprint: arXiv: 2001.06303, 2020.

[21] LIN T Y, GOYAL P, GIRSHICK R, *et al*. Focal loss for dense object detection[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020, 42(2): 318-327.

[22] ZHENG Z H, WANG P, LIU W, *et al*. Distance-IoU loss: faster and better learning for bounding box regression[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, 34(7): 12993-13000.

[23] LIN T Y, MAIRE M, BELONGIE S, *et al*. *Microsoft COCO: Common Objects in Context*[M]. Computer Vision-ECCV 2014. Cham: Springer International Publishing, 2014: 740-755.

[24] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. *June* 18-23, 2018. *Salt Lake City*, *UT*. IEEE, 2018: 7132-7141.

[25] PARK J, WOO S, LEE J Y, *et al*. Bam: Bottleneck attention module[EB/OL]. *arXiv preprint* arXiv:1807.06514, 2018.

[26] WOO S, PARK J, LEE J Y, *et al*. Cbam: Convolutional block attention module[C]. *Proceedings of the European Conference on Computer Vision* (*ECCV*), 2018: 3-19.

[27] DAI Z H, YANG Z L, YANG Y M, *et al*. Transformer-XL: attentive language models beyond a fixed-length context[C]. *Proceedings of the 57th Annual Meeting of the Association for Compu-*

tational Linguistics. Florence, Italy. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019: 2978-2988.

[28] HUANG Z H, LIANG D, XU P, *et al*. Improve transformer models with better relative position embeddings[C]. *Findings of the Association for Computational Linguistics: EMNLP* 2020. *Online. Stroudsburg, PA, USA: Association for Computational Linguistics*, 2020: 3327-3335.

[29] WU Y, CHEN Y P, YUAN L, *et al*. Rethinking classification and localization for object detection [C]. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 13-19, 2020. *Seattle, WA, USA*. IEEE, 2020: 10186-10195.

[30] RUKHOVICH D, SOFIIUK K, GALEEV D, *et al*. *IterDet: Iterative Scheme for Object Detection in Crowded Environments*[M]. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2021: 344-354.

[31] SUN W, DAI L, ZHANG X R, *et al*. RSOD: real-time small object detection algorithm in UAV-based traffic monitoring[J]. *Applied Intelligence*, 2022, 52(8): 8448-8463.

[32] LI Y T, FAN Q S, HUANG H S, *et al*. A modified YOLOv8 detection network for UAV aerial image recognition[J]. *Drones*, 2023, 7(5): 304.

[33] WANG W H, XIE E Z, LI X, *et al*. PVT v2: improved baselines with pyramid vision transformer [J]. *Computational Visual Media*, 2022, 8(3): 415-424.

作者简介：

李大湘(1974－)，男，湖南怀化人，博士，硕士生导师，2005 年、2011 年于西北大学分别获得硕士和博士学位，主要从事遥感图像分类、目标检测与跟踪、医学图像识别和深度学习等方面的研究。E-mail: www_ldx@163.com

通讯作者：

辛嘉妮(2000－)，女，陕西渭南人，硕士研究生，2021 年于西安邮电大学获得学士学位，主要从事深度学习、无人机航拍图像目标检测的研究。E-mail: xjn_2000@163.com