基于改进 YOLOv8 的无人机视角下小目标检测模型

谢明宏 [1]
XIE Minghong

摘 要 随着无人机技术的快速发展，无人机在城市管理和灾害监测等领域的应用愈加广泛。无人机拍摄的航拍图像具备广泛视角和高覆盖范围的优势，但也面临小目标检测难度大、背景复杂等挑战。针对这些问题，文章提出了一种基于改进 YOLOv8 的目标检测算法——DAE-YOLO。该模型通过引入 Deformable Attention 机制，实现了动态调整注意力的关键点位置，提升了对小目标的检测能力；采用创新的 Inner Wise-IoU 损失函数，通过引入自适应的辅助边界框优化 IoU 损失计算，提高了边框回归的精度；同时设计了轻量化的 Detect_Efficient 检测头，在保证检测精度的同时提升了模型效率。实验结果表明，DAE-YOLO 在 VisDrone2019 数据集上相较于原始模型有显著性能提升：精确率提升 7.2%，召回率提升 6.9%，mAP50 提升 9.8%，mAP50-95 提升 10.7%。在夜间和白天的复杂场景测试中，DAE-YOLO 都表现出了优异的小目标检测能力。

关键词 改进 YOLOv8；小目标检测；DAE-YOLO；Deformable Attention；Inner Wise-IoU

doi：10.3969/j.issn.1672-9528.2025.03.002

0 引言

随着无人机技术的日新月异，其应用已渗透到农业、交通和城市规划等诸多领域 [1]。无人机凭借卓越的机动性和操控灵活度，能够高效采集大规模实时数据，为各行业带来显著的效率提升。其中，航拍成像以其独特的俯瞰视角和广阔的覆盖范围，在目标检测领域扮演着不可或缺的“角色”。相较于传统的地面拍摄，航拍视角不仅能够获取更宽广的场景信息，还能提供更为直观的空间分布特征，这一优势在城市管理和灾害监测等领域尤为突出。

然而，航拍图像的目标检测也面临着独特的技术挑战。首要问题是小目标识别难度大：航拍图像中的目标往往体积微小，特征不够明显，且分辨率受限。其次，航拍场景常常伴随着建筑物、道路和植被等复杂背景元素，这些因素会干扰目标识别的准确性。另外，由于垂直观测视角的特性，目标之间的上下文关联信息相对匮乏，这进一步增加了检测任务的复杂度。

近年来，目标检测领域已经取得了显著进展，特别是以 YOLO[2] 系列为代表的单阶段检测算法展现出了优异的性能。其中，YOLOv8 作为最新一代的目标检测算法，通过改进的骨干网络结构和特征融合策略，在通用目标检测任务上取得了显著成果。然而，在面对无人机航拍场景中的小目标检测时，YOLOv8 的原始模型仍存在一些局限性。

针对上述问题，本文提出了一种改进的目标检测算法 DAE-YOLO。该算法通过引入 Deformable Attention 机制 [3] 增强对小目标的特征提取能力，采用创新的 Inner Wise-IoU 损失函数优化边界框回归精度，并设计了轻量化的检测头结构提升模型效率。这些改进使得模型能够更好地适应无人机视角下的目标检测任务需求，为实际应用提供了更可靠的技术支持。

1 算法模型

1.1 YOLOv8

YOLOv8 是 Ultralytics 公司在 2023 年 1 月推出的新一代目标检测算法。根据模型结构的宽度和深度，YOLOv8 提供了从轻量到重量级的 5 个版本：YOLOv8n（nano）、YOLOv8s（small）、YOLOv8m（medium）、YOLOv8l（large）和 YOLOv8x（xlarge），计算复杂度依次递增，以适应不同的应用场景需求。其骨干网络的设计采用基于 CSPDarknet-53 的架构，包含 53 层卷积层和多个残差块结构。这种设计能够有效提取图像的多层次特征，增强模型对图像细节和上下文信息的理解能力。值得注意的是，YOLOv8 创新性地将传统的 C3 模块升级为 C2f 模块，这一改进不仅增强了梯度流动，还通过多尺度特征融合提升了特征提取的丰富性。颈部网络采用改进的 PAN-FPN（特征金字塔网络）结构。在 FPN [4] 的基础上融入路径聚合网络（PAN）[5]，实现了深浅

1. 贵州财经大学信息学院 贵州贵阳 550025

层特征的双向互补增强。通过简化 PAN 部分的上采样后卷积层，在保证性能的同时实现了结构轻量化，提高了模型效率。此外，检测头放弃了 YOLOv5 的耦合头，采用的是解耦头。

1.2 DAE-YOLO 算法模型

由于无人机视角下的目标存在背景复杂、目标小、多尺度等问题，导致 YOLOv8 的原始模型结构并不能 满足对无人机拍摄图像的目标检测任务。为有效解决以上问题，本文基于 YOLOv8 基础模型提出了目标检测算法 DAE-YOLO，DAE-YOLO 的结构如图 1 所示。首先，通过将可变形卷积 [6] 与 Transformer[7] 相结合，实现动态调整注意力的关键点位置。这种数据驱动的稀疏注意力机制不仅降低了计算复杂度，还能自适应地关注图像中的重要区域，提升对小目标的检测能力。结合 Inner-IoU 与 Wise-IoU 的优势，通过引入辅助边界框来优化 IoU 损失计算。对于不同 IoU 值的样本采用不同大小的辅助边界框，有效提升了边框回归的精度，特别是对小目标的定位效果。针对 YOLOv8 检测头计算开销大的问题，提出了 Detect_Efficient 轻量化检测头设计，在保证检测精度的同时提升了模型效率。

1.2.1 Deformable Attention

传统注意力机制在视觉任务中面临着诸多挑战。Vision Transformer 采用的全局注意力机制不仅计算开销巨大，还容易受到图像中无关区域的干扰，导致特征表示效率低下；而 PVT[8] 等模型通过下采样特征图虽然降低了计算量，却造成了严重的信息损失；Swin Transformer[9] 引入的窗口注意力机制虽然在一定程度上缓解了计算压力，但其固定的注意力模式限制了对大尺度目标的建模能力。

Deformable Attention[10] 巧妙地将可变形卷积的思想与 Transformer 相结合，通过学习共享的偏移量来动态调整注意力的关键点位置。这种数据驱动的稀疏注意力机制不仅大幅降低了计算复杂度，更重要的是能够自适应地关注图像中的重要区域。

具体流程如下：

给定特征值图$x \in \mathbb{R}^{H\times W\times C}$，在特征图上均匀分布网格作为参考点$p \in \mathbb{R}^{H_G\times W_G\times 2}$。其中$H_G=\frac{H}{r}, W_G=\frac{W}{r}$，$r$ 是下采样率，将参考点进行归一化。特征图首先通过线性投影生成查询向量$q=xW_q$。轻量级的偏移网络θ_{offset}用于预测$\Delta p=\theta_{\text{offset}}(q)$。在变形点的位置对特征进行采样，生成键（key）和值（value）$\tilde{k}=\tilde{x}W_k$，$\tilde{v}=\tilde{x}W_v$。变形特征图 $\tilde{x}$ 是通过采样函数 ϕ 获得的，该函数执行双线性插值具体公式为：

$$\tilde{x}=\phi(x;p+\Delta p) \tag{1}$$

其中 ϕ 定义为：

$$\phi(z;(p_x,p_y))=\sum_{(r_x,r_y)} g(p_x,r_x)g(p_y,r_y)z[r_y,r_x,:] \tag{2}$$

式中：$g(a,b)=\max(0,1-|a-b|)$；(r_x, r_y) 表示特征图上的所有位置。

对查询 q、变形键 $\tilde{k}$ 和变形值 $\tilde{v}$ 应用多头变形注意力。第 m 个注意力头的输出。具体为：

$$z^{(m)}=\sigma\left(\frac{q^{(m)}(\tilde{k}^{(m)})^{\top}}{\sqrt{d}}+\phi(\hat{B};R)\right)\tilde{v}^{(m)} \tag{3}$$

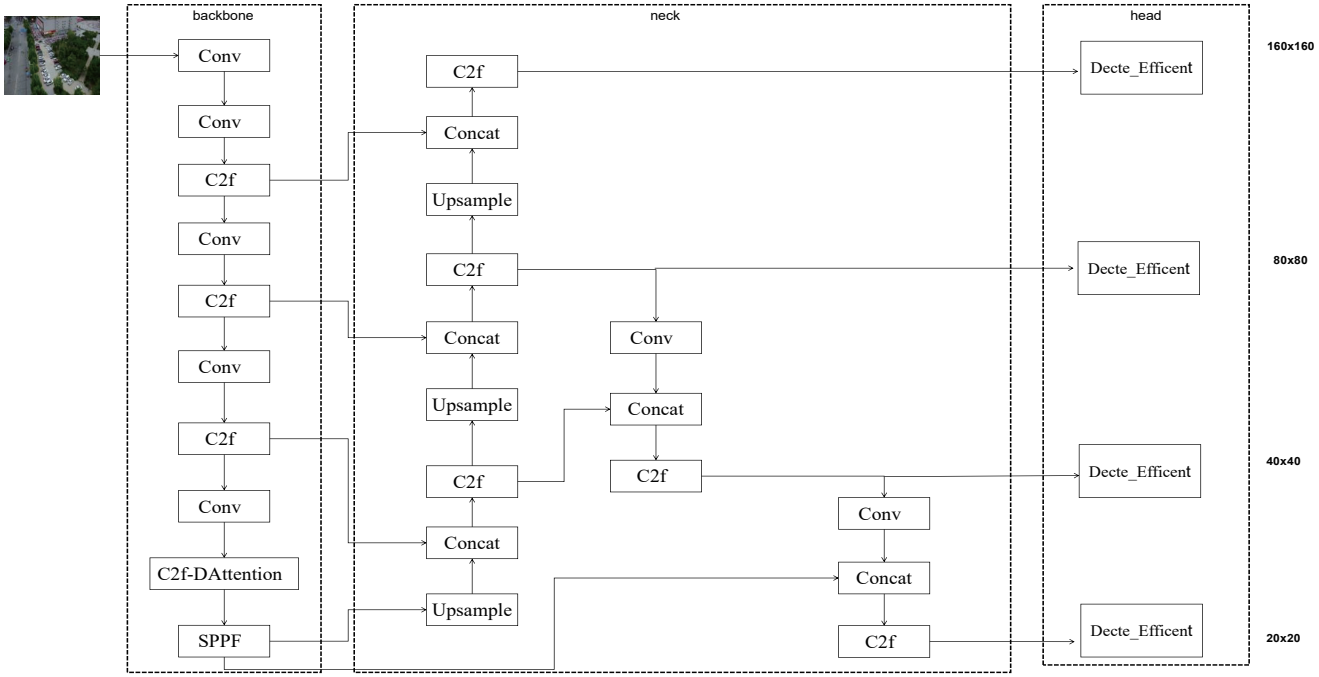


图 1 DAE-YOLO 模型结构

式中：$\phi(\hat{B};R)\in\mathbb{R}^{HW\times H_G W_G}$表示位置嵌入矩阵，结合相对位置偏移 R；σ 表示 softmax。在所有注意力头处理完成后将其拼接在一起通过输出投影层生成最终的特征图 Z。

1.2.2 Inner Wise-IoU

随着边界框回归（bounding box regression, BBR）损失函数不断更新和优化，IoU 表示预测框和目标框的交并比，然而现有基于 IoU 的 BBR 存在部分局限性，无法根据不同检测器和检测任务自适应调整。在小目标检测的中，目标较小导致样本质量参差不齐，由于检测器过度依赖高质量样本，导致检测效果不佳，本文通过采用 Inner-IoU[10] 与 Wise-IoU[11] 相结合来解决上述问题。

Inner-IoU 通过引入辅助边界框（auxiliary bounding box）来计算 IoU 损失，对于 IoU 值高的样本使用较小的辅助边界框获得大于实际边框的梯度值。对于 IoU 较小的样本使用较大的辅助边界框，扩大边框回归的有效范围。

辅助边界框的计算公式分别为：

$$b_l^{\mathrm{gt}}=x_c^{\mathrm{gt}}-\frac{w^{\mathrm{gt}}\cdot r}{2} \tag{4}$$

$$b_r^{gt}=x_c^{gt}+\frac{w^{gt}\cdot r}{2} \tag{5}$$

$$b_t^{\mathrm{gt}}=y_c^{\mathrm{gt}}-\frac{h^{\mathrm{gt}}\cdot r}{2} \tag{6}$$

$$b_b^{\mathrm{gt}}=y_c^{\mathrm{gt}}+\frac{h^{\mathrm{gt}}\cdot r}{2} \tag{7}$$

$$b_l=x_c-\frac{w\cdot r}{2} \tag{8}$$

$$b_r=x_c+\frac{w\cdot r}{2} \tag{9}$$

$$b_t=y_c-\frac{h\cdot r}{2} \tag{10}$$

$$b_b=y_c+\frac{h\cdot r}{2} \tag{11}$$

式中：r 是缩放系数。Inner-IoU 的计算公式为：

$$\begin{aligned}\mathrm{inter}=&\left(\min(b_r^{\mathrm{gt}},b_r)-\max(b_l^{\mathrm{gt}},b_l)\right)\\&\times\left(\min(b_b^{\mathrm{gt}},b_b)-\max(b_t^{\mathrm{gt}},b_t)\right)\end{aligned} \tag{12}$$

$$\mathrm{union}=(w^{\mathrm{gt}}\times h^{\mathrm{gt}}+w\times h)\cdot r^2-\mathrm{inter} \tag{13}$$

$$\mathrm{IoU}_{\mathrm{inner}}=\frac{\mathrm{inter}}{\mathrm{union}} \tag{14}$$

Inner-IoU 损失为：

$$L_{\text{Inner-IoU}}=1-\mathrm{IoU}_{\mathrm{inner}} \tag{15}$$

在 YOLOv8 模型中采用的是 CIoU，其计算公式为：

$$L_{\mathrm{CIoU}}=1-\mathrm{IoU}+\frac{\rho^2(b,b^{\mathrm{gt}})}{c^2}+\alpha v \tag{16}$$

式中：L_{CIoU} 表示 CIoU 损失；$\rho^2(b,b^{\mathrm{gt}})$ 表示预测框 b 和真实框 b^{gt} 中心点的欧氏距离的平方 c 是包围预测框和真实框的最小闭包区域的对角线距离；α 表示权重参数；v 表示衡量长宽比相似度的参数。该方法并未考虑宽度和高度与置信度之间的真实差异。因此，在某些情况下，可能会影响模型对目标形状的优化。WIoU 分为 v1、v2、v3 三个版本，本文所使用的是 v3 版本，计算公式为：

$$L_{\mathrm{WIoUv1}}=R_{\mathrm{WIoU}}\cdot L_{\mathrm{IoU}} \tag{17}$$

$$R_{\mathrm{WIoU}}=\exp\left(\frac{(x-x_{\mathrm{gt}})^2+(y-y_{\mathrm{gt}})^2}{(W_{\mathrm{g}}^2+H_{\mathrm{g}}^2)}\right) \tag{18}$$

$$L_{\mathrm{WIoUv3}}=r\cdot L_{\mathrm{WIoUv1}} \tag{19}$$

$$r=\frac{\beta}{\delta\alpha^{\beta-\delta}} \tag{20}$$

$$\beta=\frac{L_{\mathrm{IoU}}^{*}}{L_{\mathrm{IoU}}} \tag{21}$$

式中：α、δ 是控制聚焦机制的超参数；L^{*}_{IoU} 是当前 batch 计算得到的 IoU 损失值。

1.2.3 轻量化检测头

目标检测中，检测头的设计对模型的性能和效率有着重要影响。YOLOv8 检测头采用复杂的三层结构，导致计算开销的增加。为了解决这个问题，本文提出了 Detect_Efficient，一个专注于效率的检测头设计。

传统检测头通常包含三层及更多卷积，用于增强特征表达能力。然而，多余的卷积层会带来计算复杂度增加和参数冗余问题。通过观察发现，深度卷积网络的性能收益在堆叠层数达一定程度后趋于饱和。因此，本文采用双层卷积设计，以最小的计算代价获取足够的感受野和特征提取能力。

在目标检测头中，不同特征层的输入通道数不同，因此本文提出使用固定通道策略，所有卷积操作均保持输入和输出通道一致，避免通道扩张或压缩所带来的额外开销。

2 实验与分析

2.1 数据集

在本文中，使用了 VisDrone2019 进行实验验证。VisDrone2019 数据集由天津大学机器学习与数据挖掘实验室收集，是最广泛使用的无人机航拍图像数据集之一。该数据集由多种无人机型号在多个城市中拍摄，涵盖了多样的环境、目标密度、天气和光照条件。数据集包含 8 599 张图像，其中训练集有 6 471 张，验证集有 548 张，测试集有 1 610 张。图像标注的目标类别包括 10 个类别：行人、成人、自行车、汽车、面包车、卡车、三轮车、有篷三轮车、公交车和摩托车，图像分辨率范围为 480 px×360 px ~ 2 000 px×1 500 px。

2.2 评价指标

本文采用 4 种评价指标来衡量模型性能，分别是精确率 P（precision）、召回率 R（recall），以及平均精度 mAP50 和 mAP50-95。精确率 P 和召回率 R 公式为：

$$P=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{22}$$

$$R=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{23}$$

式中：TP 表示在被分类为正确类别的正样本；FP 表示被误认为正确的负样本；FN 表示被错误分类的正样本。

$$\mathrm{AP}=\int_0^1 P(r)\mathrm{d}r \tag{24}$$

$$\mathrm{mAP}=\frac{\sum_{i=1}^{C}\mathrm{AP}_i}{C} \tag{25}$$

式中：单个类别的 P-R 曲线下的面即 AP（average precision）；mAP 是多个类别平均准度的综合度量；C 表示类别数。

2.3 实验环境以及参数设置

本实验平台为 Linux，基于深度学习框架 PyTorch-GPU 2.3.1。编译环境为：Python 为 3.8.19，CPU 为 16 vCPU Intel(R) Xeon(R) Platinum 8352V CPU @ 2.10 GHz，GPU 为 RTX 4090（24 GB）。参数设置如表 1 所示。

表 1 训练参数

参数名称	设置
批量大小（batchsize）	16
输入图片尺寸（imagesize）	640
训练轮数（epoch）	300
初始学习率（lr0）	0.01
最终学习率（lrf）	0.01
线程数（works）	8
学习率动量（momentum）	0.937
优化器（optimizer）	SGD

2.4 消融实验

由表 2 实验结果可得，加入的 InnerWIoU、加入的注意力机制以及新增并且轻量化的检测头在 VisDrone2019 数据集上均表现出色。

表 2 总体消融实验

DAttention	P2	InnerWIoU	P/%	R/%	mAP50/%	mAP50-95/%	Para/10^6	GFLOPs
			44.6	33.5	33.5	19.5	3.0	8.1
✔			45.0	36.3	36.4	21.6	3.2	8.8
	✔		46.7	36.4	37.2	22.3	3.9	10.1
		✔	45.0	34.0	34.0	19.9	3.0	8.1
✔	✔	✔	**47.8**	**35.8**	**36.8**	**22.6**	3.9	10.3

相比于 YOLOv8n 的框架，精确率 P 和召回率 R 在指标上分别提升了 7.2% 和 6.9%，在 mAP50/% 和 mAP50-95/% 在指标上分别提升 9.8% 和 10.7%。意味着改进的 DAE-YOLO 模型在检测小目标方面具有更好的性能。

2.5 对比实验

为了进一步验证 DAE-YOLO 模型的检测性能，选用目前较为流行的优秀目标检测算法 DAE-YOLO，在 VisDrone2019 数据集上进行对比。

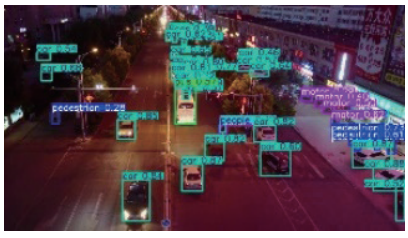
由表 3 实验结果可得，DAE-YOLO 在对比较为流行的目标检测算法 YOLOv8n、YOLOv9 和 YOLOv10n。精确率 P、召回率 R、mAP50 和 mAP50-95 指标都均有领先，对比专门用于检测小目标的 YOLOv8n-P2 和 TPH-YOLOv5 [12]，DAE-YOLO 在计算量 GFLOPs 明显小于该两个算法的情况下部分指标高于该两个算法。

表 3 基于 VisDrone 的实验对比

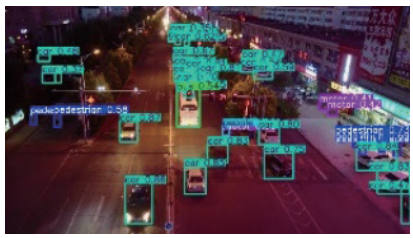
	P/%	R/%	mAP50/%	mAP50-95/%	Para/10^6	GFLOPs
YOLOv8n	44.6	33.5	33.5	19.5	3.0	8.1
YOLOv9t	46.5	33.3	34.4	20.2	1.9	7.6
YOLOv10n	43.6	34.0	34.0	19.7	2.2	6.5
YOLOv8n-p2	46.7	36.4	37.2	22.3	2.9	12.2
DAE-YOLOv8n	47.8	35.8	36.8	22.6	3.9	10.3
TPH-YOLOv5	48.3	39.0	39.0	22.3	41.5	160.1

2.6 可视化结果分析

为了展现本文改进算法在无人机视角下的检测效果，选取了夜晚和白天场景的可视化对比，如图 2、图 3 所示，在夜晚模糊的场景下 DAE-YOLO 对比 YOlOv8 检测到更多的摩托车，在白天密集场景 DAE-YOLO 对比 YOLOv8 检测到更多的行人，YOLOv8 在右下角将自行车检测为摩托车，而 DAE-YOLO 正确检测出自行车。

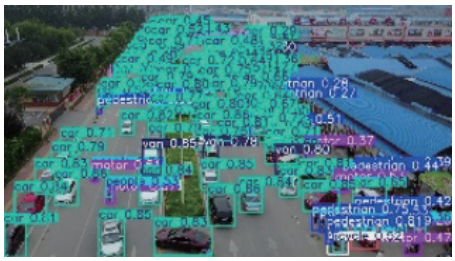


（a）YOLOv8



（b）DAE-YOLO

图 2 夜晚场景图



（a）YOLOv8

（b）DAE-YOLO

图 3 白天场景图

3 结论

本文针对无人机视角下目标检测面临的背景复杂、目标小、多尺度等挑战，提出了基于 YOLOv8 改进的 DAE-YOLO 算法模型。实验结果表明，在 VisDrone2019 数据集上，相比原始 YOLOv8 模型，DAE-YOLO 在各项性能指标上均取得显著提升：精确率提升 7.2%，召回率提升 6.9%，mAP50 提升 9.8%，mAP50-95 提升 10.7%。同时，与其他先进的目标检测算法相比，DAE-YOLO 在计算量较小的情况下仍保持了较好的检测性能。在实际场景测试中，无论是夜间还是白天的复杂场景，DAE-YOLO 都展现出了优异的小目标检测能力。

参考文献：

[1] 孙佳宇，徐民俊，张俊鹏，等．优化改进 YOLOv8 无人机视角下目标检测算法 [J]. 计算机工程与应用，2025, 61(1): 109-120.

[2] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016:779-788.

[3] ZHU X Z, SU W J, LU L W, et al. Deformable DETR: deformable transformers for end-to-end object detection[C/OL]// ICLR 2021.Washington DC: ICLR, 2021[2024-08-21]. https://openreview.net/pdf?id=gZ9hCDWe6ke.

[4] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2017:936-944.

[5] WANG C Y, LIAO H Y M, WU Y H, et al. CSPNet: a new backbone that can enhance learning capability of CNN[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Piscataway: IEEE, 2020: 1571-1581.

[6] DAI J F, QI H Z, XIONG Y W, et al. Deformable convolutional networks[C]//2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2017:764-773.

[7] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: transformers for image recognition at scale[DB/OL].(2021-06-03)[2024-05-11].https://doi.org/10.48550/arXiv.2010.11929.

[8] 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). Piscataway: IEEE, 2021:548-558.

[9] LIU Z, LIN Y, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 10012-10022.

[10] ZHANG H, XU C, ZHANG S J. Inner-IoU: more effective intersection over union loss with auxiliary bounding box[DB/OL]. (2023-11-14)[2024-08-10].https://doi.org/10.48550/arXiv.2311.02877.

[11] TONG Z, CHEN Y, XU Z, et al. Wise-IoU: bounding box regression loss with dynamic focusing mechanism[DB/OL]. (2023-04-08)[2024-08-13].https://doi.org/10.48550/arXiv.2301.10051.

[12] ZHU X K, ZHAO Q, WANG X, et al. TPH-YOLOv5: improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios[DB/OL].(2021-08-26)[2024-06-01].https://doi.org/10.48550/arXiv.2108.11539.

【作者简介】

谢明宏（1998—），男，贵州六盘水人，硕士研究生，研究方向：计算机应用技术。

（收稿日期：2024-12-18）