©中国图象图形学报版权所有

中图法分类号:TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(2025)07-2389-19

论文引用格式:Hao C Y, Jin Y, He G Q, Zhang H, Zhu X Q and Song W R. 2025. Adaptive slicing-assisted enhanced method for small object detection. Journal of Image and Graphics, 30(7):2389-2407(郝川艳, 金怡, 何桂琴, 张昊, 祝翔祺, 宋婉茹. 2025. 自适应切片辅助增强的小物体目标检测. 中国图象图形学报, 30(7):2389-2407)[DOI:10.11834/jig.240450]

自适应切片辅助增强的小物体目标检测

郝川艳*,金怡,何桂琴,张昊,祝翔祺,宋婉茹

南京邮电大学教育科学与技术学院,南京　210023

摘　要:**目的**　在目标检测领域,深度学习模型已经取得巨大成功。但是已有的基于深度学习的目标检测算法在小物体目标检测中仍然困难重重,原因在于航拍图像多是更复杂的高分辨率场景,其中一些常见问题,如稠密度高、不固定的拍摄角度、目标物体尺寸小和高变异性等给现有目标检测方法带来巨大挑战。切片策略是近年来用于高分辨率图像小目标检测任务的众多优秀方法之一,然而现有的切片方法存在冗余计算问题,因此提出一种新的自适应切片方法,称为自适应切片辅助超推断(adaptive slicing-assisted hyper inference,ASAHI)。**方法**　该方法关注切片数量而非传统的切片大小,可以根据图像分辨率自适应调整切片数量以消除冗余计算带来的性能损耗。在推理阶段,首先根据ASAHI算法将输入图像分割为6或12个重叠的块;之后对每个图像块进行插值处理以保持长宽比;接下来考虑到小块切片在检测大物体时的明显缺陷,分别对切片图像块和完整输入图像进行目标检测前向计算;最后为了提高高密度场景下推理的准确性和检测速度,后处理阶段集成了一种更快和更高效的Cluster-NMS(non-maximum suppression)方法和DIoU(distance- intersection over union)惩罚项,即Cluster-DIoU-NMS(CDN),将ASAHI推理和全图推理结果进行合并,再转回原始图像尺寸。为了支持切片图像块的推理,本文在训练阶段构建的数据集也包括切片图像块。**结果**　在广泛的实验中,ASAHI在VisDrone(vision meets drones)和xView数据集上表现出具有竞争力的性能。结果显示,与现有切片方法相比,本文方法将IoU值为0.5时的平均精确率均值(mean average precision,mAP)mAP50提高1.7%,计算时间减少20%～25%;在VisDrone2019-DET-val(vision meets drones 2019 for detection for validation)数据集上,mAP50的结果提高到56.8%。**结论**　本文方法可以有效处理高分辨率场景下小物体稠密度高、拍摄角度不同以及变异性高等复杂的因素,实现高质量的小物体目标检测。

关键词:目标检测;小物体目标检测;切片推理;VisDrone;xView

Adaptive slicing-assisted enhanced method for small object detection

Hao Chuanyan*, Jin Yi, He Guiqin, Zhang Hao, Zhu Xiangqi, Song Wanru

College of Educational Science and Technology, Nanjing University of Posts and Telecommunications, Nanjing　210023, China

Abstract: **Objective** Object detection has attracted considerable attention because of its wide application in various fields. In recent years, the progress of deep learning technology has facilitated the development of object detection algorithms combined with deep convolutional neural networks. In natural scenes, traditional object detectors have achieved excellent results. However, current object detection algorithms still encounter difficulties in small object detection. The reason is

收稿日期:2024-08-15;**修回日期**:2025-01-10;**预印本日期**:2025-01-17

*通信作者:郝川艳　hcy@njupt.edu.cn

基金项目:国家自然科学基金项目(62477026, 62473201, 62307025);教育部人文社会科学规划项目(23YJAZH047)

Supported by: National Natural Science Foundation of China(62477026, 62473201, 62307025); The Ministry of Education of Humanities and Social Science Project (23YJAZH047)

that most aerial images refer to complex high-resolution scenes, and some common problems, such as high density, unfixed shooting angle, small size, and high variability of targets, have introduced great challenges to existing object detection methods. Thus, small object detection has emerged as a key area in the field of object detection research. Its broad applications mainly include identification of early small lesions and masses in medical imaging, remote sensing exploration in military operations, and location analysis of small defects in industrial production. Some researchers have obtained high-resolution image features through multiple up-sampling operations, while another set of approaches effectively deals with problems such as high intensity by adding a penalty item in the post-processing stage. Among them, one excellent work is the use of slicing strategy, which slices the image into smaller image blocks to enlarge the receptive field. However, the existing slicing-based methods involve redundant computation that increases the calculation cost and reduces the detection speed. **Method** Therefore, a new adaptive slicing method, which is called adaptive slicing-aided hyper inference (ASAHI), is proposed in this study. This approach focuses on the number of slices rather than the traditional slice size. The approach can adaptively adjust the number of slices according to the image resolution to reduce the performance loss caused by redundant calculation. Specifically, in the inference stage, the work first divides the input image into 6 or 12 overlapping patches using the ASAHI algorithm. Then, it interpolates each image patch to maintain the aspect ratio. Next, considering the obvious defects of the slicing strategy in detecting large objects, this method separately performs forward computation on the sliced image patches and the complete input image. Finally, the post-processing stage integrates a faster and efficient Cluster-NMS method and DIoU penalty term to improve accuracy and detection speed in high-density scenes. This method, which is called Cluster-DIoU-NMS (CDN), merges the ASAHI inference and full-image inference results and resizes them back to the original image size. Correspondingly, the dataset constructed in the training stage also includes slice image blocks to support the ASAHI inference. The dataset of slicing images and the pre-training dataset of the entire images together constitute the fine-tuning dataset for the training of this work. Notably, the slicing method used in the fine-tuning dataset can be either the ASAHI algorithm or the conventional sliding window method. In the ASAHI slicing process, this method sets a distinction threshold to control the number of slices . If the length or width of the image exceeds this threshold, then the image will be cut into total 12 slices; otherwise, it will be cut into total 6 slices. Thereafter, the width and height of the slice block are calculated according to the value of , and the coordinate position of the slice is determined. After the abovementioned calculation, the ASAHI algorithm realizes the adaptive adjustment of slice size within a limited range by controlling the number of slices. **Result** Broad experiments demonstrate that ASAHI exhibits competitive performance on the VisDrone and xView datasets. The results show that the proposed method achieves the highest mAP50 scores (45.6% and 22.7%) and fast inference speeds (4.88 images per second and 3.58 images per second) on both datasets. In addition, the mAP and mAP75 increase by 1.7% and 1.1%, respectively, on the VisDrone2019-DET-test dataset. Meanwhile, the mAP and mAP75 improve by 1.43% and 0.9%, respectively, on the xView test set. On the VisDrone2019-dt-val dataset, the mAP50 of the experiment exceeds 56.8%. Compared with state of the art, the proposed method achieves the highest mAP (36.0%), mAP75 (28.2%), and mAP50 (56.8%) values, with a highest processing speed of 5.26 images per second, which suggests a better balance performance. **Conclusion** The proposed algorithm can effectively handle complex factors, such as high density, different shooting angles, and high variability, in high-resolution scenes. It can also achieve high-quality detection of small objects.

Key words: object detection; small object detection; sliced inference; VisDrone; xView

0 引 言

随着大数据和硬件技术的发展,基于深度学习的目标检测算法取得令人瞩目的成果,如适用于自然场景的RetinaNet(Lin等,2017b)、YOLO(you only look once)系列(Redmon等,2016; Redmon和Farhadi,2017; Bochkovskiy等,2020; Jocher等,2021; Wang等,2023)和Fast R-CNN(fast region-based convolutional neural network)(Wang等,2017)等典型的目标检测器。然而,对于4K相机、无人机以及卫星等设备拍摄的高分辨率图像而言,这些检测算法的识别结果并不令人满意。以经典的两阶段目标检测算法Faster R-CNN(Ren等,2017)为例,它在多尺度

物体上的平均精确率均值(mean average precision，mAP)可以达到38.0%，但在VisDrone2019(vision meets drones 2019)(Du等，2019)的验证集中检测小物体和极小物体时，mAP仅为20.0%和6.2%。这主要由3个因素造成：1)高分辨率图像中物体数量多(稠密度高)、物体尺寸多样和物体面积占比小等特点给检测带来了巨大的挑战；2)检测小物体更困难，因为联合交并比(intersection over union，IoU)指标很难有效应对遮挡、物体之间距离过近和空中拍摄视角不同等情况；3)在较深网络层中，低分辨率特征可能无法同时匹配不同尺寸的小物体。

为此，小物体目标检测算法兴起并已成为目标检测的热点和难点研究内容，彰显着更为重要的应用价值。在医学成像中，小物体目标检测可用于对早期的微小病灶和肿块进行识别，有助于肿瘤等病症的早期发现和治疗(Mahajan等，2020)；在军事遥感勘探中，小物体目标检测可在高分辨率图像中定位地形、舰船等物体(Zhang等，2024)；在工业生产中，小目标检测可以通过定位材料表面的小缺陷进行残次品(Wang等，2018)检测；此外，小物体目标检测还广泛应用于人员识别、流量监控、交通管制、轨道损坏检测和农害检测等诸多领域，体现了其在计算机视觉任务中不可或缺的地位和挑战性。

小物体目标检测代表性的工作包括TPH(Transformer prediction heads)-YOLOv5(Zhu等，2021)、TOOD(task-aligned one-stage object detection)(Feng等，2021)、PP-YOLOE(an evolved version of YOLO)(Xu等，2022)、SOD-MTGAN(small object detection via multi-task generative adversarial network)(Bai等，2018)等。在这些工作中，一些研究人员通过多次采用上采样操作来获取高分辨率的图像特征，例如TPH-YOLOv5(Zhu等，2021)和QueryDet(query detection)(Yang等，2022)。另外，通过在后处理阶段添加惩罚项，检测算法可以有效地处理高密集度等问题。这些典型的后处理方法包括Soft-NMS(non-maximum suppression)(Bodla等，2017)、WBF(weighted boxes fusion)(Solovyev等，2021)、DIoU-NMS(distance-intersection over union nonmaximum suppression)(Zheng等，2020)和Cluster-NMS(Zheng等，2022)等。此外，研究人员提出切片方法(slicing aided hyper inference，SAHI)(Akyon等，2022)和基于移位窗口的自注意力机制(swin Transformer)(Liu等，2021)。这类方法将图像切片成更小的图像块，以放大小物体的局部特征。但是，已有切片方法(Akyon等，2022)存在冗余计算问题，增加了计算成本，降低了检测速度。因此，本文方法专注于缓解切片方法中的冗余问题，以获得更优的精度和检测速度。

立足于此，本文提出一种新的切片方法，自适应切片辅助超推理(adaptive slicing assisted hyper inference，ASAHI)。它不再使用指定的切片大小，而是通过固定切片数量，自适应地调整切片大小，从而很大程度上减少了冗余计算。同时，在后处理阶段，本文使用Cluster-DIoU-NMS(CDN)算法取代了传统的非最大抑制算法(NMS)(Bodla等，2017)，从而可以在保证检测结果的同时减少时间消耗。本文方法的提出主要由两个调研结果所启发。1) SAHI(Akyon等，2022)的切片方法固定了切片的尺寸大小，在各分辨率的图像上都产生了不同程度的冗余计算；尤其在图像的边缘部分，当重复计算超过了重叠率时，冗余计算问题更为突出。相比之下，自适应调整切片尺寸是一个更好的选择。2) 小物体目标检测常常伴随物体之间距离近、物体重叠和遮挡等挑战性问题。而主流的后处理方法NMS(Bodla等，2017)由于仅被单一指标因素IoU所影响，所以在检测高稠密度物体时表现并不理想；这种限制在小物体检测中更为明显。因此，将DIoU-NMS(Zheng等，2020)与Cluster-NMS(Zheng等，2022)结合起来不失为一个好的解决方案。

本项工作在具有挑战性的数据集VisDrone2019(Du等，2019)上验证ASAHI，该数据集包含大量不同分辨率的小物体。此外，还在遥感数据集xView(Lam等，2018)上对算法进行了测试，该数据集带有边界框的注释，包含大量的空中图像和世界各地的拍摄场景。实验表明，本文方法将推理和后处理阶段的速度提高了20%～25%，同时与基线SAHI方法(Akyon等，2022)相比，将检测性能提高了1.7%。图1显示了ASAHI获得的改进结果，图1(a)—(d)分别为YOLOv5(Jocher等，2021)、TPH-YOLOv5(Zhu等，2021)、TPH-YOLOv5 + SAHI和TPHYOLOv5 + ASAHI(本文方法)的推理结果。从粉色框中可见，使用切片方案的算法(图1(c)、图1(d))检测到更多的边界框。此外，虽然现有的切片方法SAHI(Akyon等，2022)(图1(c))在检测小物体方面表现良好，但ASAHI(图1(d))在检测小物体以及微小物体上的表现更具优势，如绿色边界框中所示，ASAHI检测到

了更多的微小物体。

综上，本文主要贡献如下：1）提出一种新的切片方法ASAHI，在SAHI（Akyon等，2022）的基础上缓解了冗余计算问题，提高了对不同分辨率图像以及多尺度目标的检测性能。2）通过将ASAHI与TPH-YOLOv5（Zhu等，2021）框架的结合，本文方法增强了对深层低分辨率图像特征的敏感度。3）针对ASAHI，设计了一种与自适应切片对应的微调方案，称为切片辅助微调SAF（slicing-assisted fine-tuning），用于数据增强；并在后处理阶段集成了更快和更高效的Cluster-NMS（Zheng等，2022）方法和DIoU（Zheng等，2020）惩罚，项以提高高密度场景下推理的准确性和检测速度。

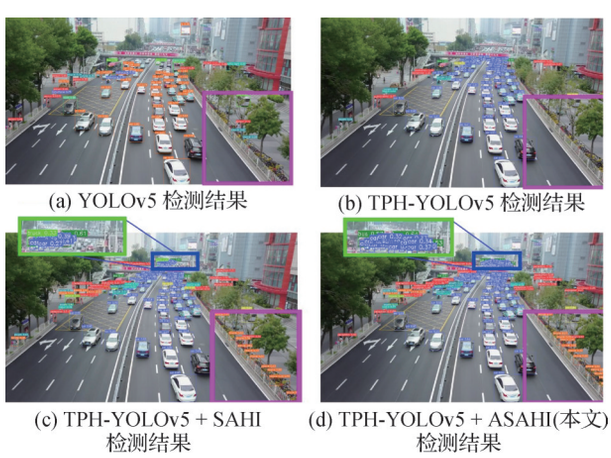

图1　不同方法的推理结果

Fig. 1　Inference results from different approaches
((a) YOLOv5; (b) TPH-YOLOv5; (c) TPH-YOLOv5 + SAHI; (d) ours)

1　相关工作

1.1　目标检测

基于卷积神经网络（convolutional neural network，CNN）的目标检测器可以分为两大类：单阶段和双阶段检测器。单阶段检测器，如YOLOv4（Bochkovskiy等，2020）、YOLOv5（Jocher等，2021）、SSD（single shot multibox detection）（Liu等，2016）、RetinaNet（Lin等，2017b）和EfficientNet（Tan等，2020）等，直接预测目标的位置，可以看做是求解一个回归问题；双阶段检测器，如Mask R-CNN（He等，2017）、R-CNN（Girshick等，2014）、Faster R-CNN（Ren等，2017）、SPPNet（spatial pyramid pooling network）（He等，2015）、DetectoRS（detecting objects with recursive feature pyramid and switchable atrous convolution）（Qiao等，2021）、特征金字塔网络（feature pyramid network，FPN）（Lin等，2017a）等，倾向于使用感兴趣区域对齐操作（region of interest align，RoIAlign）（He等，2017）来显式对齐对象的特征，在感兴趣区域中生成候选对象，并有选择地搜索图像中与目标对象高概率吻合的部分。通常情况下，单阶段检测器的计算成本更低，而双阶段检测器的精确度更高。然而，这两种方法之间的性能差距已经逐渐缩小。例如，YOLOv5（Jocher等，2021），一个经典的单阶段检测器，它使用带有SPP层的CSP-Darknet53（cross-stage partial connections-darknet 53）（Bochkovskiy等，2020）结构为骨干，PANet（path aggregation network）（Liu等，2018）为颈部。在推理过程中，YOLOv5（Jocher等，2021）首先获得大量的边界框，然后在后处理阶段通过CIoU（complete-IoU）进行聚合。这种方法已经由于其便利性和良好的性能，吸引了大量研究人员的关注，也成为后续许多工作的基础结构。

1.2　小物体目标检测

现有的小物体目标检测框架大多是在目标检测算法的基础上改进，以适应更为复杂的高分辨率场景。这些框架通常基于卷积神经网络构建（潘晓英等，2023；石争浩 等，2023），通过增加网络深度、添加提取能力更强的模块等方式应对更为复杂的小物体目标检测场景，但同时也增加了模型的参数量。另一些研究者对已有的基于自注意力机制的目标检测算法进行改进，如融入CBAM（convolutional block attention module）（Woo等，2018）、TPH-YOLOv5（Zhu等，2021）、SSA-CNN（semantic self-attention-CNN）（Zhou等，2019）架构等。其中，TPH-YOLOv5（Zhu等，2021）是一个基于YOLOv5（Jocher等，2021）的先进的小物体检测框架，集成了一个额外的预测头用于小物体检测，在颈部使用CBAM（Woo等，2018）框架，并且使用Transformer编码器（Dosovitskiy等，2021）代替原先的CSP（Bochkovskiy等，2020）。这些改进提高了针对小物体的检测精度，并保留了原有的推理速度。

还有一类方法通过使用切片策略以扩大小目标的感受野，如SAHI（Akyon等，2022）、Swin-Transformer（Liu等，2021）、CSwin-Transformer（a general vision transformer backbone with cross-shaped windows）（Dong

等，2022）等。其中，SAHI（Akyon 等，2022）在训练前的预处理阶段进行切片；而后两者的切片方案则是在网络前向传播过程中进行。然而，在复杂的前向传播中进行切片导致了大量的内存消耗，对设备资源要求较高。因此，本文聚焦轻量化的 SAHI（Akyon 等，2022）切片预处理方法，但该方法固定切片尺寸的操作带来了冗余计算。为此，本文利用 TPH-YOLOv5（Zhu 等，2021）作为骨干网络，并融合能够更灵活调整切片尺寸的 ASAHI 切片方案以达到更好的性能。

1.3 小物体目标检测中的后处理

后处理是小物体目标检测中的一个关键部分，它可以抑制在推理过程中产生的大量重复检测，例如目前最流行的 NMS（Bodla 等，2017）。另一类方法，如 DETR（deformable transformers for end-to-end object detection）（Zhu 等，2020）已经摒弃了传统的后处理过程，但这项技术的发展还不完善。相对地，本文方法保持了后处理的使用。虽然 NMS（Bodla 等，2017）的表现很有竞争力，但 IoU 是 NMS（Bodla 等，2017）中唯一的影响因素，导致它仍有一些不足之处。因此，一些研究通过增加额外的惩罚项来提高性能或减少时间消耗。例如，GIoU（generalized IoU）（Zheng 等，2020）、DIoU（Zheng 等，2020）和 CIoU（Zheng 等，2020）分别增加了非重叠区域的影响权重、边界框之间的中心点距离和边界框的长宽比作为惩罚项以提高检测性能。另一些方法与 NMS（Bodla 等，2017）相结合，在后处理阶段获得了更好的性能。这些工作包括：Focus-and-Detect（Koyun 等，2022），一种最先进的两阶段算法之一；Soft-NMS（Bodla 等，2017），一种基于高斯算法的针对 NMS（Bodla 等，2017）的平滑改进方案；以及 CascadeR-CNN + NWD（normalized Gaussian Wasserstein distance）（Wang 等，2021），一种基于归一化 Wasserstein 距离的经典改进算法。受到以上优秀工作的启发，本文采用 DIoU-NMS（Zheng 等，2020）和 Cluster-NMS（Zheng 等，2022）的组合算法实现后处理过程，达到了更好的性能和更快的推理。

2 本文方法

2.1 方法概述

切片方法 SAHI（Akyon 等，2022）的核心思想是：将输入图像划分为重叠的切片块，喂入网络进行训练和推理，这相对于一幅完整的输入图像，可以令小物体的像素区域相对较大，从而提高小物体目标检测的性能。但固定的切片大小不可避免地导致不同分辨率图像中的冗余计算。当使用固定的切片大小时，为了确保切片的完整性，边界区域上的切片会包含上一幅切片图像的部分信息。这是导致冗余计算产生的根本原因，也会给后处理中的预测合并带来问题。

因此，本文提出 ASAHI 切片方法，根据不同的图像分辨率，通过固定切片的数量来保证切片的大小可以自适应于重叠区域的比例，不仅可以改进固定切片大小下切片边界处重叠率过高导致的冗余计算问题，还可以使每个切片补丁仍然保有一定的重叠比率，从而减少边界区域的信息损失。相应地，其整体推理过程需对输入的原始图像同时使用两种策略，即完全推理（full inference，FI）和切片推理（ASAHI），其整个检测过程如图 2 所示。在 FI 中，一幅完整的图像被送入推理阶段，如图 2 中绿色箭头所示流程；而在 ASAHI 中，不同大小的图像切片块被送入推理阶段，如图 2 中青色箭头所示流程。在推理过程中，完整图像和切片图像经由 TPH-YOLOv5（Zhu 等，2021）处理以获得相应的边界框。之后，根据图像长宽比重新调整边界框的大小，并在后处理阶段使用 Cluster-DIoU-NMS 对冗余框进行合并，以去除重复识别的预选框。需注意的是，这两种策略的联合使用迫使训练阶段应做出相应的微调，其微调数据集也应包含原始图像数据和切片图像块数据；它们送入模型中进行训练的过程如图 2 中黄色箭头的流程所示。上述算法总流程步骤如下：

算法 1 训练阶段。

输入：VisDrone 2019 训练集或 xView 训练集。

输出：检测模型 M。

1）构建微调数据集。

（1）全图数据；

（2）滑窗切图法构建切片图像块数据。

2）使用预训练的 TPH-YOLOv5 模型作为基础，训练检测模型。

3）输出检测模型 M。

算法 2 推理阶段。

输入：图像 $\boldsymbol{I}_k$（$k \in \boldsymbol{N}$，$\boldsymbol{N}$ 为整数集），来自 Vis-

Drone-2019-val或xView验证集、检测模型M。

输出：检测结果。

1）使用ASAHI算法对I_k进行切片处理（详见2.3节）。

2）切片块图像送入M，得到检测边界框。

3）全图I_k送入M，得到检测边界框。

4）聚合步骤2）和步骤3）的边界框。

5）使用Cluster-DIoU-NMS后处理算法去除重叠的冗余框。

6）输出最后检测结果。

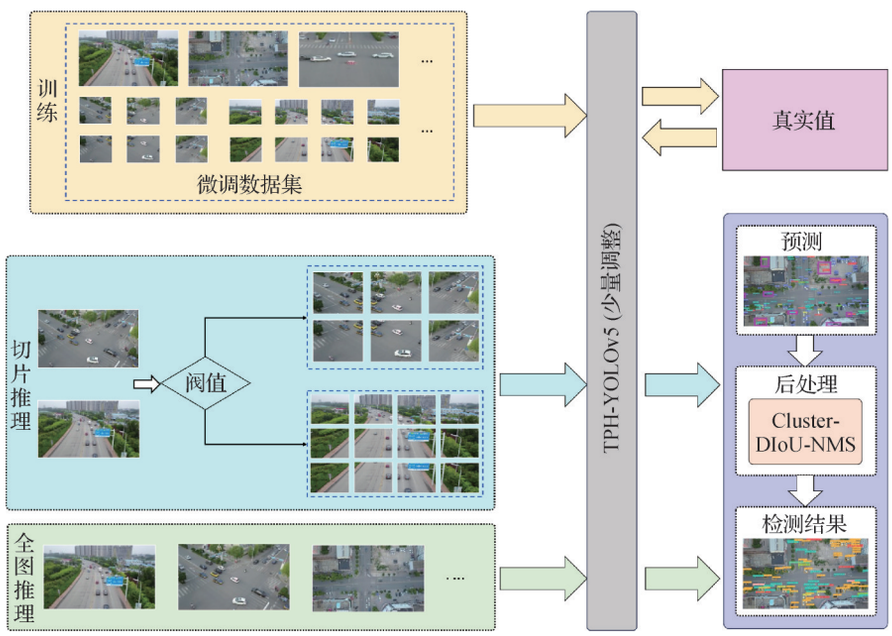

图2　基于ASAHI切片的检测过程

Fig. 2　The detection procedure with ASAHI

2.2　TPH-YOLOv5概述

TPH-YOLOv5（Zhu等，2021）提供了两个预训练模型，分别是yolov5l-xs-1和yolov5l-xs-2，它们以不同的图像尺寸进行训练。TPH-YOLOv5（Zhu等，2021）采用CSPDarknet53（Bochkovskiy等，2020）架构，以3个Transformer（Dosovitskiy等，2021）编码器块为骨干，用带CBAM（Woo等，2018）模块的PANet（Liu等，2018）为颈部，并有4个Transformer（Dosovitskiy等，2021）预测头。考虑到其在VisDrone2021竞赛（Cao等，2021）中排名第5的优异性能，本文工作选择它作为骨干架构。基于TPH-YOLOv5作者（Zhu等，2021）提供的代码，本工作在网络结构的颈部进行了适当修改，如图3所示，减少了1个CBAM模块（Woo等，2018）。

2.3　自适应切片辅助超推理（ASAHI）

由于切片图像块会被用于辅助推理，因此在推理过程中，图像会被推理两次，即FI和ASAHI。这两个过程可以同时进行或顺序进行。如图4所示，在ASAHI辅助的推理通道中，算法首先根据图像分辨率将原始图像分割为6或12个重叠的块，其效果也在后续算法推演和实验中得到验证；其次，在保持长宽比不变的条件下，通过双线性插值对切片块进行适量放大操作以统一切片块尺寸。另外，考虑到小块的切片在检测大物体时存在明显的缺陷，本文方法仍然保留FI推理通道以有效检测较大的物体。最后，通过Cluster-DIoU-NMS将ASAHI和FI的推理结果进行合并，再转回原始图像大小。

在ASAHI的切片过程中，本文方法设置一个区分阈值T以控制切片数量p。如果图像的长度或宽度超过这个阈值，那么图像将被切割成4 × 3共12个切片，反之则被切割成3 × 2共6个切片。具体为

$$T = r \times \left(4 - 3 \times \mu\right) + 1 \tag{1}$$

式中，μ定义为重叠率，r定义为限制尺寸，以保证切片块的大小在r的限定范围内浮动，在本文实现中，r设置为512。

设置阈值T值后，通过计算两种极端情况并取最大值来确保图像被精确划分为6和12个切片。即

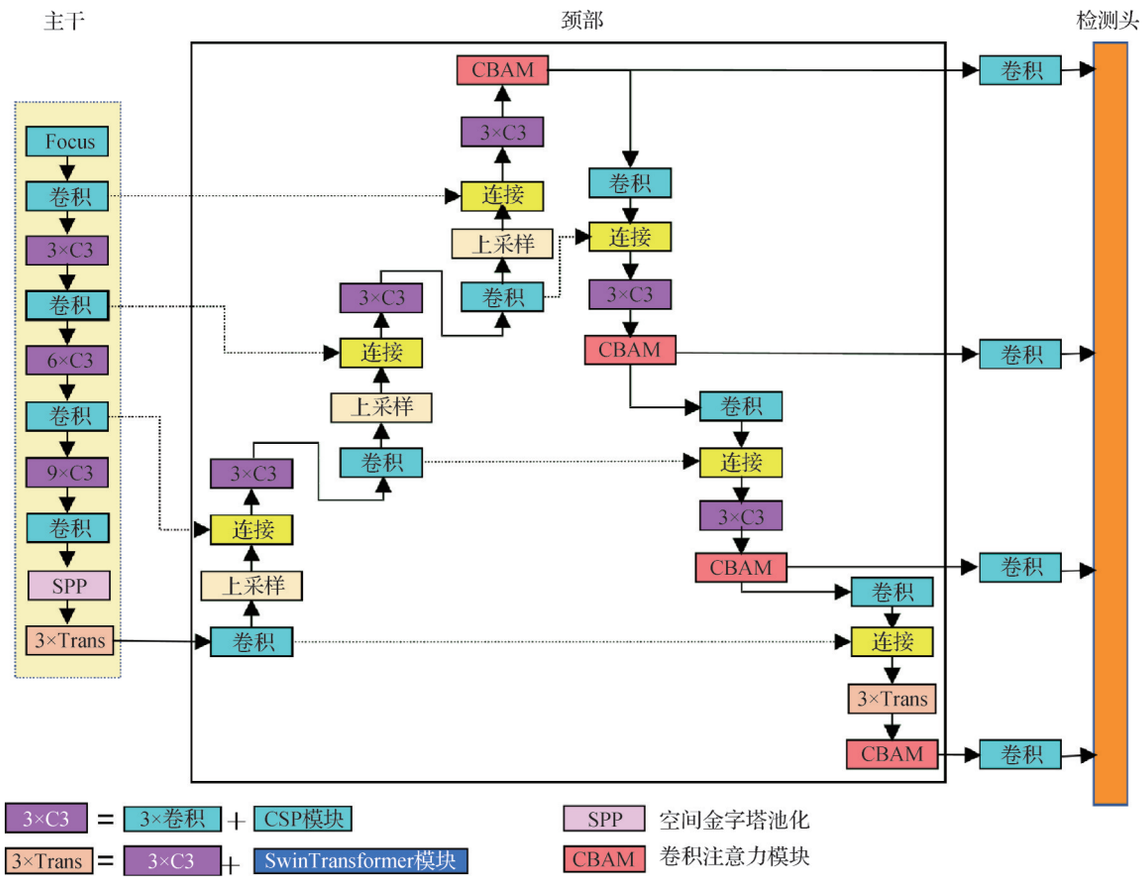

图 3　对 TPH-YOLOv5(Zhu 等,2021)架构的少量调整

Fig. 3　The architecture of TPH-YOLOv5 (Zhu et al. , 2021) with slight changes made for this study

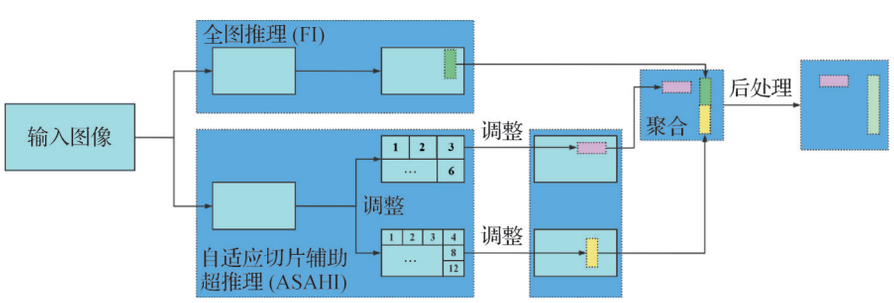

图 4　FI 和 ASAHI 辅助的推理过程

Fig. 4　FI and ASAHI assisted inference processes

当超过阈值 T 时,计算宽度 4 分以及高度 3 分的极限情况,反之计算宽度 3 分以及高度 2 分的极限情况。具体为

$$p=\begin{cases}\max\left(\dfrac{W}{3-2\times\mu}+1,\dfrac{H}{2-\mu}+1\right) & \max(W,H)\leqslant T\\ \max\left(\dfrac{W}{4-3\times\mu}+1,\dfrac{H}{3-2\times\mu}+1\right) & \max(W,H)>T\end{cases}\tag{2}$$

式中,W 和 H 分别是输入图像的长度和宽度。

之后,根据 p 的取值计算切片块的宽度和高度,并确定切片的坐标位置。经过上述计算,本文算法实现了通过控制切片数量 p 使得切片大小在限定范围内的自适应调整。具体算法步骤如下:

算法 3 ASAHI 算法切片过程。

输入:图像 $\boldsymbol{I}_k(k\in\boldsymbol{N},\boldsymbol{N}$ 为整数集),来自 VisDrone-2019-val 或 xView 验证集。

输出:切片图像块数据。

1)根据式(2)计算切片数量 p。

2)根据 p 计算切片块的长边长度 l_{long} 和短边长

度l_{short}。

if $p = 6$

$$l_{\text{long}} = \frac{W}{3 - 2 \times \mu} + 1, l_{\text{short}} = \frac{H}{2 - \mu} + 1$$

if $p = 12$

$$l_{\text{long}} = \frac{W}{4 - 3 \times \mu} + 1, l_{\text{short}} = \frac{H}{3 - 2 \times \mu} + 1$$

3)根据切片块的长边和短边计算切片块坐标位置。

令$\boldsymbol{slice}_{\text{coor}} = \left[\left(x_1, y_1\right), \left(x_2, y_2\right)\right]$表示切片块的左上角和右下角坐标，$slice_{\text{h}}$为切片块高度，$slice_{\text{w}}$为切片块宽度，则

if $H > W$

$$slice_{\text{h}} = l_{\text{long}}, slice_{\text{w}} = l_{\text{short}}$$

else

$$slice_{\text{h}} = l_{\text{short}}, slice_{\text{w}} = l_{\text{long}}$$

do

$$\boldsymbol{slice}_{\text{coor}} \cdot append\left(\left[\begin{matrix}\left(x_1 + slice_{\text{w}}, y_1 + slice_{\text{h}}\right), \\ \left(x_2 + slice_{\text{w}}, y_2 + slice_{\text{h}}\right)\end{matrix}\right]\right)$$

/附加切片宽度和高度以生成新的切片图像坐标

while

$$\max\left(x_1, x_2\right) + slice_{\text{w}} \geqslant H \text{ 或}$$
$$\max\left(y_1, y_2\right) + slice_{\text{h}} \geqslant W$$

4)输出切片块图像。

这项改进减少了在推理过程中的冗余计算，如图5所示。

设a为水平轴上的切片数，b为垂直轴上的切片数。表达式为

$$a = \left(\frac{W - p \times \mu}{p \times \left(1 - \mu\right)}\right) \tag{3}$$

$$b = \left(\frac{H - p \times \mu}{p \times \left(1 - \mu\right)}\right) \tag{4}$$

然后，根据式(2)中W、H、μ、p的4个已知值，计算出图5中的冗余面积S_r，具体为

$$R_x = p \times a - p \times \mu \times \left(a - 1\right) - W \tag{5}$$

$$R_y = p \times b - p \times \mu \times \left(b - 1\right) - H \tag{6}$$

$$S_r = R_x \times H + R_y \times W - R_x \times R_y \tag{7}$$

式中，R_x和R_y分别为冗余计算区域在长度和宽度上的冗余量。

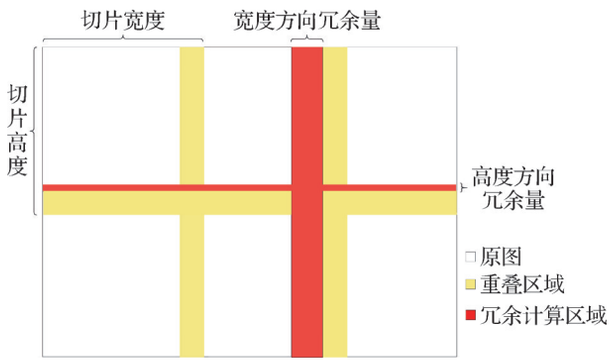

图5　减少SAHI(Akyon等，2022)中冗余计算示意图

Fig. 5　Diagram of reducing redundant computation in SAHI (Akyon et al., 2022)

在计算ASAHI方法的冗余量后，可以进一步求得与SAHI(Akyon等，2022)方法相比冗余计算面积的减少量，具体为

$$S_a = W \times H \tag{8}$$

$$S_r^{\text{ASAHI}} = \begin{cases} W \times \left(p \times \left(2 - \mu\right) - H\right) \\ \quad \left(\dfrac{W}{3 - 2 \times \mu} \geqslant \dfrac{H}{2 - \mu}\right) \text{和} \max\left(W, H\right) \leqslant T \\ H \times \left(p \times \left(3 - 2 \times \mu\right) - W\right) \\ \quad \left(\dfrac{W}{3 - 2 \times \mu} < \dfrac{H}{2 - \mu}\right) \text{和} \max\left(W, H\right) \leqslant T \\ H \times \left(p \times \left(3 - 2 \times \mu\right) - W\right) \\ \quad \left(\dfrac{W}{4 - 3 \times \mu} \geqslant \dfrac{H}{3 - 2 \times \mu}\right) \text{和} \max\left(W, H\right) > T \\ H \times \left(p \times \left(3 - 2 \times \mu\right) - W\right) \\ \quad \left(\dfrac{W}{4 - 3 \times \mu} < \dfrac{H}{3 - 2 \times \mu}\right) \text{和} \max\left(W, H\right) > T \end{cases} \tag{9}$$

$$S_{\text{redu}}^{\text{rate}} = 1 - \frac{S_r^{\text{ASAHI}} + S_a}{S_r + S_a} \tag{10}$$

式中，S_a表示被求图像总像素面积，S_r^{ASAHI}表示ASAHI切片方法下的冗余面积，$S_{\text{redu}}^{\text{rate}}$表示ASAHI相比SAHI(Akyon等，2022)减少的冗余面积。

以上推演说明本文方法确实减少了SAHI(Akyon等，2022)产生的冗余计算，这使得ASAHI能够达到更快的推理速度。表1给出了与SAHI(Akyon等，2022)相比，各种分辨率的图像减少的冗余计算的比例。图像均来自VisDrone2019和xView数据集。

表 1　不同分辨率下与 ASAHI 相比 SAHI（Akyon 等，2022）相比，减少冗余计算的比例

Table 1　Proportion of reducing the redundant calculation that ASHI made at different resolutions compared with SAHI（Akyon et al. ，2022）

图像分辨率/像素	减少冗余计算的比例/%
960 × 540	38.72
1 360 × 765	2.56
1 400 × 1 050	25.61
1 920 × 1 080	25.92
2 000 × 1 500	24.13
2 913 × 2 428	6.99

2.4　自适应切片辅助微调

广泛使用的 TPH-YOLOv5（Zhu 等，2021）目标检测框架在 VisDrone 等数据集上训练后得到预训练的权值。这使得下游任务模型可以用更小的数据集和更短的训练时间进行微调，而无需用大数据集从头开始训练。由于 ASAHI 在推理阶段输入了切片图像块，对应地，本文在训练阶段构建的数据集也包括切片图像块。该切图数据集和全图的预训练数据集共同构成本工作训练时的微调数据集，因而本文的微调过程称为切片辅助微调 SAF。

其完整过程如图 6 所示，预训练数据集中的图像 $\{\boldsymbol{I}_1, \boldsymbol{I}_2, \cdots, \boldsymbol{I}_j, j \in \boldsymbol{N}\}$，被切成重叠的切片块 $\left\{\boldsymbol{P}_1^1, \boldsymbol{P}_1^2, \cdots, \boldsymbol{P}_1^k, \cdots, \boldsymbol{P}_j^k, j \in \boldsymbol{N}, k \in \boldsymbol{N}\right\}$ 以形成切图数据集。需提及的是这里使用的切图方法与 3.2 节中所述方法无需一致，可以使用常规的滑窗切图方法。之后，与原始预训练数据集一起，SAF 可以建立完整的微调数据集。考虑到切片后的大量数据给训练过程带来的计算量，本文不再使用其他的数据增强技术，如旋转、几何失真或光度失真等。在训练时，为了保证尺度的统一性，原图和切片块会被缩放到同一尺寸，这虽然造成了原图的信息损失，但本文方法特征抽取的关键在于对切图块的学习，而原图只用于对全局信息的补充，如学习切图块的位置信息等，因而并不影响检测精度。

2.5　Cluster-DIoU-NMS（CDN）

后处理是小物体目标检测中至关重要的一部分。鉴于 Cluster-NMS（Zheng 等，2022）和 DIoU-NMS（Zheng 等，2020）的优异性能，本文在后处理阶段采

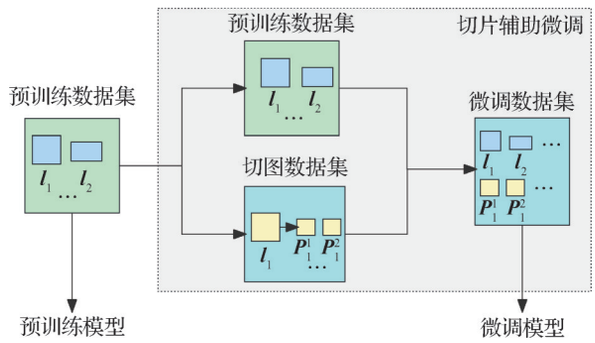

图 6　SAF 建立微调数据集的流程图

Fig. 6　Flowchart of SAF which sets up the fine-tuning dataset

用了 Cluster-NMS（Zheng 等，2020）的结构以提高检测速度，并用 DIoU 代替 IoU 度量来提高重叠场景下的性能。

1）DIoU-NMS（Zheng 等，2020）。在传统 NMS（Bodla 等，2017）算法中，IoU 是唯一考虑的因素。然而，在实际应用中，当两个物体之间的距离很小时，仅凭 IoU 很难区分物体。因此，在后处理时要同时考虑两个边界框的重叠面积和中心点的距离，而不像传统的 IoU，只考虑重叠区域，忽略了质心距离。为了解决这个问题，DIoU-NMS（Zheng 等，2020）将两个方框中心点之间的距离作为惩罚项。如果两个方框之间的 IoU 比较大，中心之间的距离比较小，则认为方框是单独的物体，不会被过滤掉。在 DIoU 中，c^2 表示覆盖在两个边界框上的最小包裹框的对角线距离，而 $\rho^2\left(x, x^{\mathrm{gt}}\right)$ 表示两个框的质心之间的欧几里得距离。其损失函数 L_{DIoU} 定义为

$$L_{\mathrm{DIoU}} = IoU - \frac{\rho^2\left(x, x^{\mathrm{gt}}\right)}{c^2} \tag{11}$$

2）Cluster-DIoU-NMS。鉴于 DIoU（Zheng 等，2020）的优势，本文算法在 Cluster-NMS（Zheng 等，2022）的后处理中采用 DIoU（Zheng 等，2020）作为度量标准。该算法将检测结果进行排序，取分数最高值并计算其余各区域的 DIoU（Zheng 等，2020）值。设定 DIoU（Zheng 等，2020）的阈值为 0.5，如果 DIoU（Zheng 等，2020）大于 0.5，则设定 Cluster-NMS（Zheng 等，2022）的值为 0，反之设为 1。这里，0 表示该检测框为冗余框，应该抑制；1 表示该检测框进入下一轮计算。之后，Cluster-NMS（Zheng 等，2022）通过左乘去除 0 项以进入下一轮计算，以此类推。当所有的值都为 0 时停止运算，整个过程如图 7 所示。需要说明的是，Cluster-NMS（Zheng 等，2022）的左乘

操作可以减少传统NMS(Bodla等,2017)的冗余计算,因为在NMS(Bodla等,2017)中,0项会被送入下一轮进行重复计算。

总的来说,Cluster-DIoU-NMS后处理算法结合了Cluster-NMS(Zheng等,2022)和DIoU-NMS(Zheng等,2020)。其中,Cluster-NMS(Zheng等,2022)解决了过度抑制问题,而使用DIoU(Zheng等,2020)代替IoU度量可以提高重叠场景下检测的准确性。

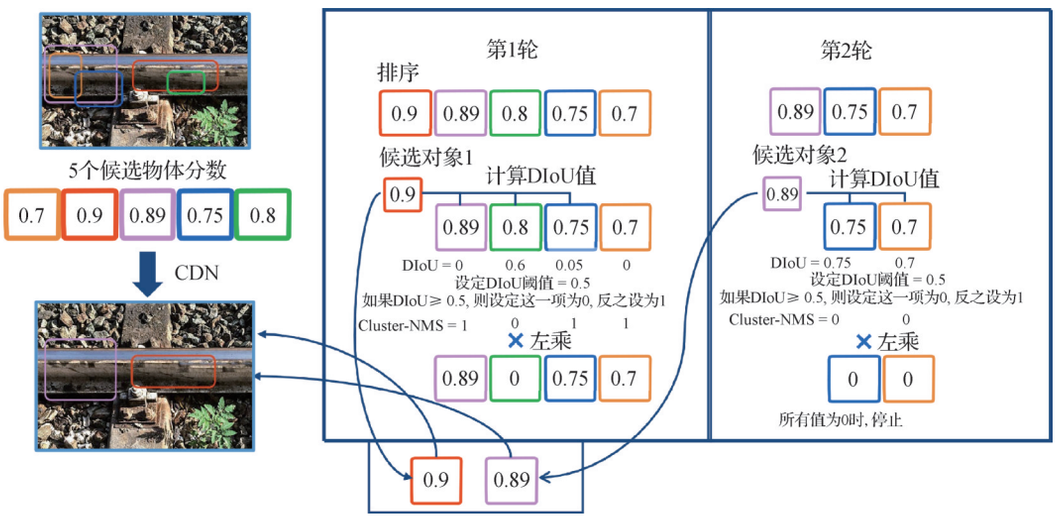

图7 Cluster-DIoU-NMS(CDN)算法流程

Fig. 7 Cluster-DIoU-NMS (CDN) algorithm flow

3 实 验

3.1 数据集和设置

本工作使用了两个数据集,VisDrone2019(Du等,2019)和xView(Lam等,2018)来评估本文的模型,使用mAP(所有类别在不同IoU阈值下的平均值,$IoU \in \{0.5, 0.55, \cdots, 0.95\}$、mAP75(所有类别在$IoU = 0.75$下的平均值)、mAP50(所有类别在$IoU = 0.5$下的平均值)、mAP50_s(面积$< 32^2$像素,即小物体)、mAP50_m($32^2$像素$\leqslant$面积$\leqslant 96^2$像素,即中物体)和mAP50_l(面积$> 96^2$像素,即大物体)的指标来报告结果。另外,VisDrone2019-DET-val数据集(Du等,2019)用于验证。

VisDrone2019-DET是一个由8 599幅不同位置和高度的无人机航拍图像组成的目标检测数据集。它包括训练集的6 471幅图像、测试集的1 580幅图像和验证集的548幅图像,图像分辨率介于1 024×960像素至1 920×1 024像素之间。这个数据集中的大多数图像具有以下特点:1)占据小区域的物体密集分布;2)某些物体被遮挡;3)存在反射、光度偏移和视点偏移等类似现象。它对超过54万物体的边界框进行了标注,包含了10个预定义类别:行人、人、自行车、汽车、货车、卡车、三轮车、遮阳篷—三轮车、公共汽车和电机。

xView是一个用于遥感图像目标检测的大型数据集。它包含了大量在世界各地拍摄的遥感图像,这些图像都具有很高的分辨率,介于2 000×2 000像素至3 000×2 500像素之间。xView包括来自60个不同类别的100多万个对象实例。在实验过程中,本工作随机选择了80%、10%和10%的图像,分别组成训练集、测试集和验证集。

3.2 实现细节

本文训练和测试使用单个NVIDIA RTX3080 GPU,其他测试在单个NVIDIA RTX2080 Ti GPU上进行。在训练阶段,本实验使用预训练的TPH-YOLOv5模型(Zhu等,2021)进行数据集微调,用于ASAHI算法的训练和评估。在使用SAF切图方案后,微调数据集的训练集包含50 708幅切片图像和6 471幅原始图像。接下来,实验将全部图像的尺寸设置为512像素,批量大小设置为32。模型共训练了120个epoch。在训练过程中使用了Adam优化器,3×10^{-3}作为初始学习率,最后一个epoch衰减到初始学习率的0.12。在推理过程中,重叠率设置为0.15,Cluster-CIoU-NMS匹配阈值设置为0.5,T由式(1)计算,为1 818。

3.3 消融实验

在 VisDrone2019-DET-test 数据集和 xView 数据集上的验证结果说明了本文提出的切片方法的普适性，详情见表 2 和表 3。考虑到常规航拍以及遥感图像的长宽比，本实验测试了当切片数量固定为 4、6、12、15 等 4 个选定值时的模型表现：切片数为 12 时，在两个数据集上都获得了最高的 mAP50 结果；但就时间耗费而言，小的切片数量更具有优势，如固定的 4 切片在两个数据集上都达到了最快的推理速度。另一方面，通过与 SAHI（Akyon 等，2022）算法的数据对比可见，固定切片尺寸（512 像素）比固定切片数量（4、6、12、15 选其一）带来了更好的性能，但固定的切片尺寸在处理不同分辨率的图像时造成的冗余计算拖慢了 SAHI（Akyon 等，2022）的推理时间。而本文提出的 ASAHI 自适应切片方法在两个数据集上均获得了最高的 mAP50 分数（45.6% 和 22.7%）和快速的推理速度（4.88 幅/s 和 3.58 幅/s）。对于大多数不同的图像分辨率，ASAHI 控制的切片数量多为 6 个或 12 个，这同时适用于检测小物体（mAP50_s 最高分 33.8% 和 17.3%）、中等大小物体（mAP50_m 最高分 59.7% 和 25.6%）。对于大物体，调整切片数量为 4 能获得最高的性能（mAP50_l 最高分 68.8% 和 20.0%）和速度（5.19 幅/s 和 3.8 幅/s）。另外，在 mAP 和 mAP75 指标上，本文方法也取得最高结果：在 VisDrone2019-DET-test 数据集上，mAP 和 mAP75 分别提升 1.7% 和 1.1%；在 xView 测试集上，mAP 和 mAP75 分别提升 1.43% 和 0.9%。

表 2　在 VisDrone2019-DET-test 数据集上的测试结果

Table 2　Test results on VisDrone2019-DET-test dataset

/%

方法	推理速度/(幅/s)	mAP	mAP75	mAP50	mAP50_s	mAP50_m	mAP50_l
ASAHI（4 切片）	**5.19**	23.9	17.1	38.7	24.9	56.5	**68.8**
ASAHI（6 切片）	4.98	28.5	22.0	41.6	28.8	58.5	65.3
ASAHI（12 切片）	2.98	29.3	22.8	41.9	30.5	56.4	64.3
ASAHI（15 切片）	2.39	27.2	21.3	40.9	29.9	55.1	62.9
SAHI（512 像素）	3.69	28.7	24.1	42.2	29.8	58.5	65.6
ASAHI（自适应）	4.88（↑1.19）	**30.4**（↑1.7）	**25.2**（↑1.1）	**45.6**（↑3.4）	**33.8**（↑4.0）	**59.7**（↑1.2）	60.7

注：加粗字体表示各列最优结果，↑表示与 SAHI（512 像素）相比的提升。

表 3　在 xView 测试集上的测试结果

Table 3　Test results on xView dataset

/%

方法	推理速度/(幅/s)	mAP	mAP75	mAP50	mAP50_s	mAP50_m	mAP50_l
ASAHI（4 切片）	**3.80**	**13.2**	7.5	18.4	11.7	22.6	**20.0**
ASAHI（6 切片）	2.78	17.3	12.0	21.3	15.3	23.1	18.2
ASAHI（12 切片）	1.13	16.2	12.4	21.9	16.5	23.5	17.9
ASAHI（15 切片）	1.09	15.3	10.8	20.6	14.8	20.7	16.8
SAHI（512 像素）	2.56	16.07	11.5	20.4	14.9	23.5	17.6
ASAHI（自适应）	3.58（↑1.02）	**17.5**（↑1.43）	**12.4**（↑0.9）	**22.7**（↑2.3）	**17.3**（↑2.4）	**25.6**（↑2.1）	15.2

注：加粗字体表示各列最优结果，↑表示与 SAHI（512 像素）相比的提升。

表 4 和表 5 给出了本文模型结构不同部分在 VisDrone2019-DET-val 和 xView 验证集上对 mAP 性能的影响。其中，TPH 表示 backbone 结构 THP-YOLOv5（Zhu 等，2021），FI 表示全图推理，ASAHI 表示本文的自适应切片辅助推理，PO 表示切图块具有重叠部分，SAF 表示对模型的微调训练，CDN 为后处

理过程。实验结果表明，所提出的方法对中小目标物体的检测性能均有改善。

表4 在VisDrone2019-DET-val数据集上验证每个组件的重要性

Table 4 The importance of each proposed component validated on VisDrone2019-DET-val dataset

/%

组件	mAP	mAP75	mAP50	mAP50_s	mAP50_m	mAP50_1
TPH+FI	15.7	10.6	34.2	21.4	53.3	74.4
TPH+ASAHI	34.1	24.9	54.4	46.6	66.1	58.9
TPH+ASAHI+FI	34.1	25.3	55.5	46.7	68.1	**76.5**
TPH+ASAHI+FI+PO	34.3	25.9	55.5	46.8	68.6	77.6
TPH+ASAHI+FI+PO+SAF+CDN	**36.0**	**28.2**	**56.8**	**48.5**	**69.6**	72.9

注：加粗字体表示各列最优结果。

表5 在xView验证集上验证每个组件的重要性

Table 5 The importance of each proposed component validated on xView dataset

/%

组件	mAP	mAP75	mAP50	mAP50_s	mAP50_m	mAP50_1
TPH+FI	13.2	7.4	14.8	9.4	16.3	12.3
TPH+ASAHI	19.9	12.3	24.4	18.6	27.7	15.2
TPH+ASAHI+FI	20.3	12.6	24.4	18.8	28.3	18.5
TPH+ASAHI+FI+PO	20.3	12.8	24.5	18.8	28.6	**19.6**
TPH+ASAHI+FI+PO+SAF+CDN	**20.5**	**14.4**	**25.7**	**19.3**	**28.6**	17.2

注：加粗字体表示各列最优结果。

3.4 对比实验

3.4.1 定量分析

实验与当前先进的小物体目标检测算法（Yang等，2019；Yang等，2022；Koyun等，2022；Deng等，2021；Suo等，2023；Li等，2020；Wang等，2020）以及几种经典的目标检测算法（Wang等，2017；Wang等，2021）在VisDrone2019-DET-val进行对比，结果如表6所示。除了Focus-and-Detect算法（Koyun等，2022），本文方法达到最高的mAP（36.0%）、mAP75（28.2%）和mAP50（56.8%）值，速度远超Focus-and-Detect算法（Koyun等，2022），达到5.26幅/s的处理速度。可见本文方法取得了更优的平衡性能。针对基线算法SAHI（Akyon等，2022），实验在VisDrone2019-DET-test上测试了不同组件条件下对应的处理速度，即每秒处理图像数量。如图8所示，相同的颜色用于表示相同的组件应用于ASAHI和SAHI（Akyon等，2022）的速度比较。在VisDrone2019-DET-test上计算5次实验的平均图像处理速度后，实验结果说明各项组件在与ASAHI结合后达到了更快的速度。

此外，实验在VisDrone2019-DET-val验证集上对不同后处理方法的结果进行对比，如表7所示。其中，Cluster-DIoU-NMS（CDN）在中小目标的检测任务中取得mAP50的最高值，分别为48.5%和69.6%，并保持了5.26幅/s的最快检测速度。最后，实验对比了各组成成分在VisDrone2019-DET-val验证集上结合SAHI（Akyon等，2022）和ASAHI的性能表现，如表8所示。可见，ASAHI在各个组件上的表现都优于SAHI（Akyon等，2022），特别是在检测小物体时，其mAP50结果为56.8%，mAP50_s结果为48.5%。

3.4.2 定性分析

本文方法对深层低分辨率图像特征的敏感度更高，从而对小物体目标具有更强的聚焦能力。图9给出了SAHI（Akyon等，2022）、TPH-YOLOv5（Zhu等，2021）和本文方法在VisDrone（Du等，2019）数据集上的热力图对比。如图9所示，SAHI（Akyon等，

表 6　在 VisDrone2019-DET-val 数据集上与其他先进的目标检测方法进行比较

Table 6　Comparison of our method with other state-of-the-art object detection methods on VisDrone2019-DET-val

方法	mAP/%	mAP75/%	mAP50/%	速度/(幅/s)
Fast R-CNN(Wang等,2017)	12.3	9.6	24.9	2.57
Cascade R-CNN+NWD(Wang等,2021)	18.7	12.9	40.2	3.20
DMNet(Li等,2020)	28.2	17.3	30.5	1.87
HIT-UAV(Suo等,2023)	34.6	21.7	55.8	3.45
CRENet(Wang等,2020)	33.7	20.9	54.3	1.10
GLSAN(Deng等,2021)	30.7	18.4	55.4	2.30
QueryDet (Yang等,2022)	33.9	19.5	56.1	2.75
ClusDet(Yang等,2019)	32.4	18.9	56.2	1.29
Focus-and-Detect(Koyun等,2022)	**42.0**	**30.1**	**66.1**	0.73
TPH + SAHI(Akyon等,2022)	34.9	27.5	55.1	4.67
TPH+ASAHI(本文)	36.0	28.2	56.8	**5.26**

注:加粗字体表示各列最优结果。

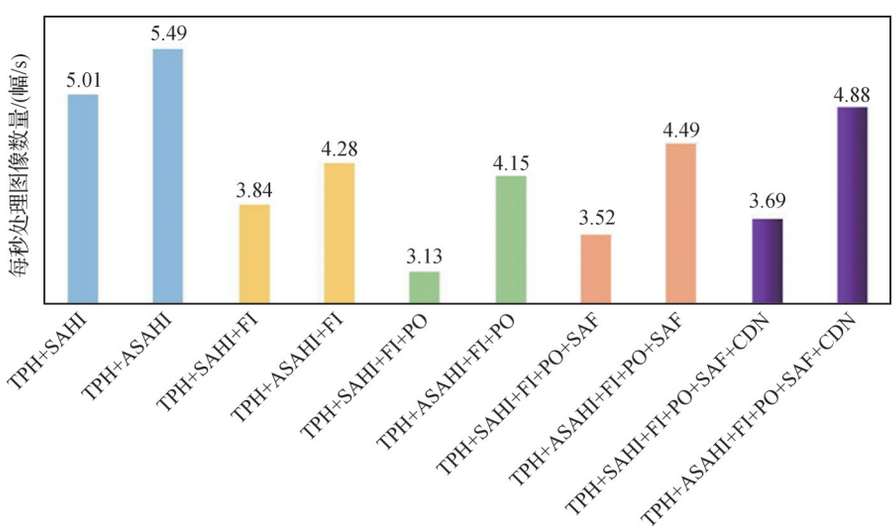

图 8　不同组件条件下 ASAHI 与 SAHI(Akyon等,2022)图像处理速度的比较

Fig. 8　Comparison of image processing speed between ASAHI and SAHI (Akyon et al. , 2022) under different components

表 7　在 VisDrone2019-DET-val 数据集上 ASAHI 使用不同后处理方法的性能比较

Table 7　The mAP and time consumption results obtained with different post-processing methods under ASAHI on VisDrone2019-DET-val

模块	mAP/%	mAP75/%	mAP50/%	mAP50_s/%	mAP50_m/%	mAP50_l/%	速度/(幅/s)
NMS	33.9	25.6	55.8	47.4	68.3	72.9	3.06
SOFT-NMS	32.3	23.3	52.6	44.3	65.8	70.0	2.72
WBF	33.4	23.9	55.1	46.7	67.5	**73.3**	2.88
Cluster-GIoU-NMS	34.5	27.3	56.3	48.0	68.5	72.9	5.17
Cluster-CIoU-NMS	35.9	27.5	**56.8**	48.4	68.6	72.9	4.87
Cluster-DIoU-NMS(CDN)	**36.0**	**28.2**	**56.8**	**48.5**	**69.6**	72.9	**5.26**

注:加粗字体表示各列最优结果。

表8　在VisDrone2019-DET-val验证集上每个组件对性能的影响

Table 8　The performance impact of each component tested on the VisDrone2019-DET-val validation set

/%

组件	mAP	mAP75	mAP50	mAP50_s	mAP50_m	mAP50_l
TPH+FI	15.7	10.6	34.2	21.4	53.3	74.4
TPH+SAHI	33.0	23.5	54.2	45.7	66.8	69.0
TPH+SAHI+FI	33.5	24.1	54.6	45.7	68.3	75.2
TPH+SAHI+FI+PO	34.9	25.7	55.1	46.1	68.8	74.6
TPH+SAHI+FI+PO+SAF+CDN	34.9	27.5	55.1	48.1	69.0	72.7
TPH+ASAHI	34.1	24.9	54.4	46.6	66.1	58.9
TPH+ASAHI+FI	34.1	25.3	55.5	46.7	68.1	76.5
TPH+ASAHI+FI+PO	34.3	25.9	55.5	46.8	68.6	**77.6**
TPH+ASAHI+FI+PO+SAF+CDN	**36.0**	**28.2**	**56.8**	**48.5**	**69.6**	72.9

注：加粗字体表示各列最优结果。

2022)的注意力关注范围较广，但对小目标的聚焦能力不强(黄色框)，检测结果中红色框内存在明显漏检情况，置信度也不高。TPH-YOLOv5(Zhu等，2021)对目标的聚焦能力更强，但在非目标区域，如“马路”上也有部分较高的注意力权重区域(黄色框)，检测结果中红色区域的漏检问题少于SAHI(Akyon等，2022)。而本文算法对目标的聚焦能力更强，红色框内检测出更多物体，且置信度更高。

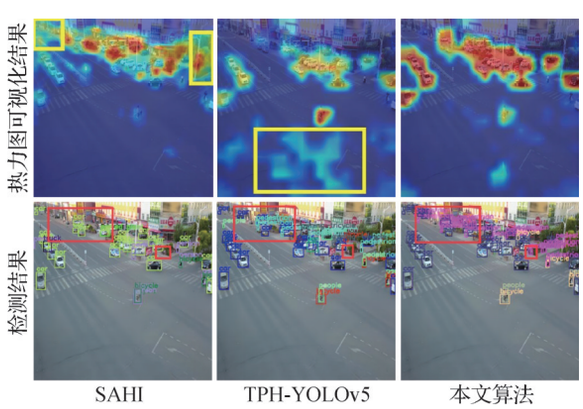

图9　VisDrone(Du等，2019)数据集上对不同模型的热力图可视化结果示例

Fig. 9　Examples of thermal map visualization results for different models on the VisDrone (Du et al., 2019) dataset

图10和图11展示了ASAHI在VisDrone和xView上对小物体目标进行检测的部分结果。从图10可以看出，ASAHI在各种具有挑战性的场景中表现良好，包括黑暗场景(如第5个示例)、有反射和不同拍摄角度的场景(如第4、9、10和11个示例)，以及高密度场景(第7个示例)等。此外，如图11所示，ASAHI在图像分辨率极高的遥感场景中的表现同样出色，包括干扰因素众多的场景(如第1个示例)、涵盖多种检测类别的场景(如第4个示例)等。由于遥感图像分辨率过高，本文随机选取了图11中红色框内的内容，做了特写显示以方便观察。

总体而言，ASAHI在VisDrone2019-DET和xView数据集上都取得了良好的性能和快速的推理速度。值得注意的是，在xView上的表现也同样出色。这可能是因为xView数据集中的图像普遍具有极高的分辨率，这给现有的特征提取算法带来了很大的困难，而本文提出的自适应切片方法(ASAHI)恰恰可以显著改善这种困境。

4　局限性讨论

本文方法仍有一些局限性。1)提高对小目标物体的检测精度可能导致对大目标物体的检测精度的损失。从表2—表5和表8可以看出，本文方法在mAP50_l的指标上并没有取得最高分数。潜在的原因在于小的切片会破坏大物体的连续性。2)仍然存在一些错误情况，如位置和类别混淆。图12是VisDrone2019-DET-val数据集中所有类别的ASAHI测试误差分析曲线，图13为xView-val数据集中车辆类别的ASAHI测试误差分析柱形图。图12说明类别混淆是导致错误的主要原因，占比为11.4，而图13

图 10　ASAHI 在 VisDrone2019-DET-val 数据集上的检测结果示例

Fig. 10　Examples of detection results by ASAHI on the VisDrone2019-DET-val dataset

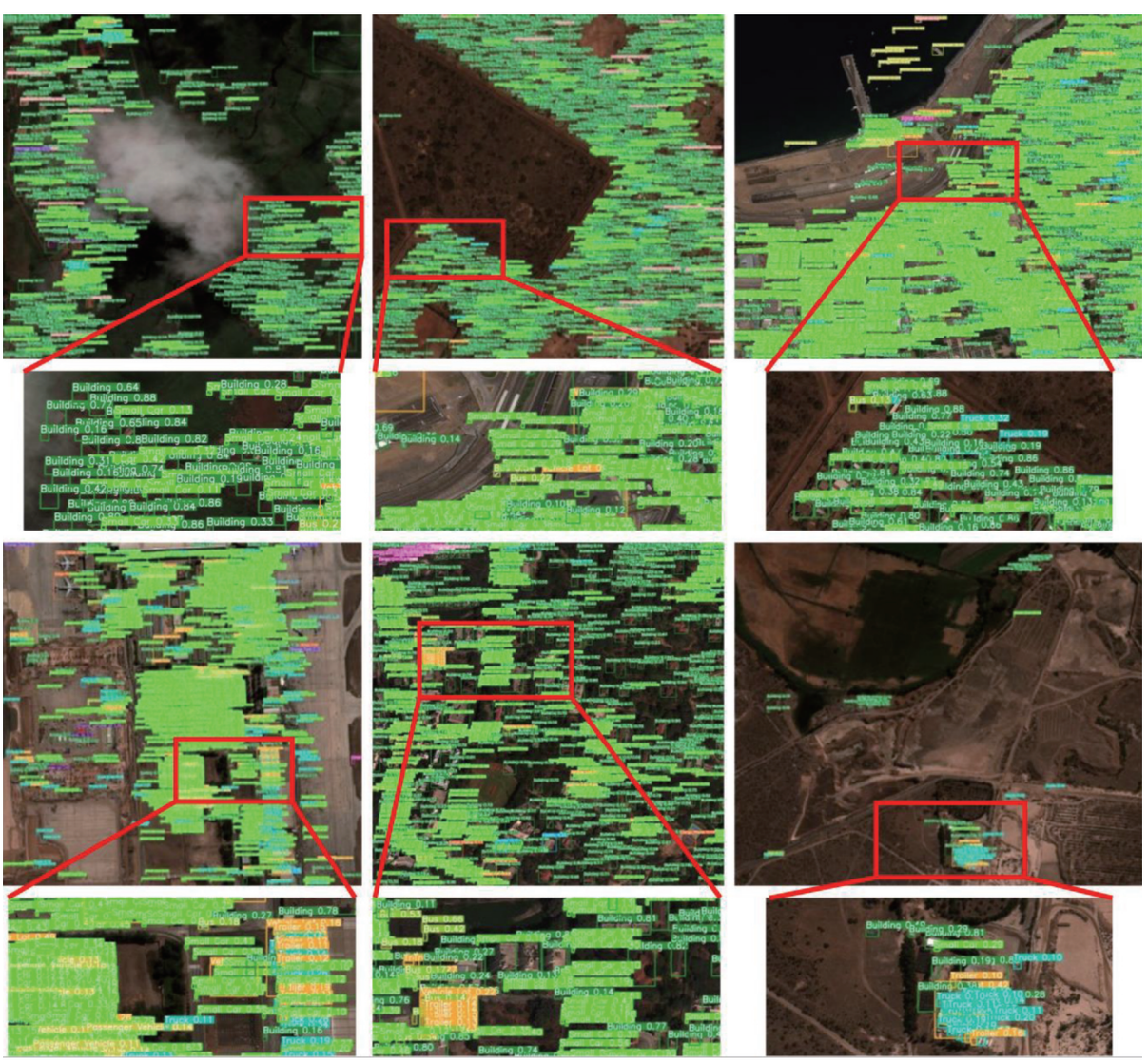

图 11　ASAHI 在 xView-val 数据集上的检测结果示例

Fig. 11　Examples of detection results by ASAHI on the xView-val dataset

呈现出位置误差是偏差的主要来源，占14%。综合观察可见，两个图中的C75（IoU阈值为0.75）、C50（IoU阈值为0.5）、Loc（忽略定位错误）、Sim（忽略超类别假阳性）、Oth（忽略类别混淆）、BG（忽略所有假阳性）和FN（忽略所有假阴性）的结果清楚地表明，在超类别假阳性、类别混淆、假阳性和假阴性方面的误差较为明显，说明算法还需提高对小物体识别的敏感性。

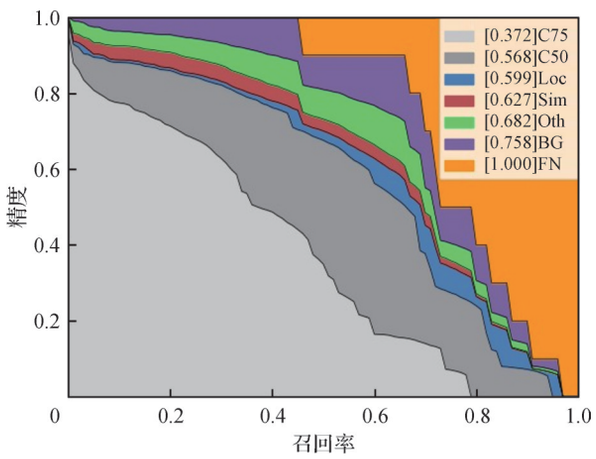

图12　VisDrone2019-DET-val数据集中所有类别的ASAHI测试误差分析曲线

Fig. 12　Error analysis curves for ASAHI testing on all classes in the VisDrone2019-DET-val set

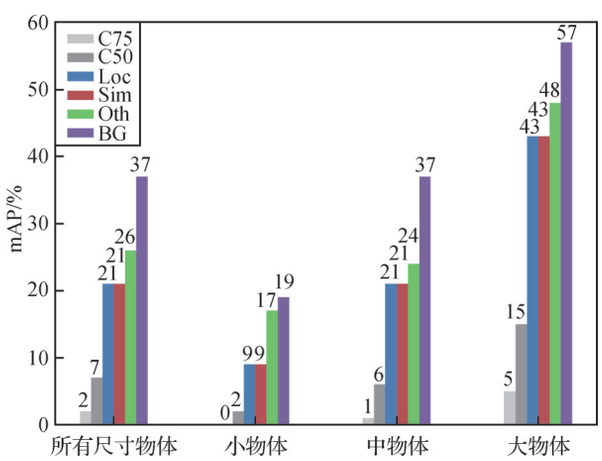

图13　xView-val数据集中车辆类别的ASAHI测试误差分析条形图

Fig. 13　Error analysis bar plot for ASAHI testing on the vehicles class in the xView-val set

5　结　论

本文工作解决了通用SAHI切片方法（Akyon等，2022）中的冗余计算问题，提出一种新的自适应切片框架：ASAHI。它可以根据图像的分辨率计算出最优的切片数量，从而进行切图操作，而不是使用固定的切片大小。这大大减轻了冗余计算带来的问题。在后处理阶段，本文使用Cluster-DIoU-NMS代替了传统的NMS（Bodla等，2017），提高了检测性能。在本文实验中，以TPH-YOLOv5（Zhu等，2021）为骨干，在VisDrone2019和xView数据集上进行了相关训练和测试。实验结果表明，本文的模型实现了mAP的改进。与SAHI（Akyon等，2022）（基线）相比，mAP50提高了1.7%，而时间消耗减少了大约25%，同时在小物体和中物体的检测性能上也达到了最优。

本文方法同样适用于其他基于YOLO系列（Redmon等，2016；Bochkovskiy等，2020；Jocher等，2021；Wang等，2023）的模型。在未来的工作中，该方法有可能通过融入坐标注意力（Hou等，2021）和CSWin Transformer2（Dong等，2022）建构等技术来提高对小目标的敏感性，并通过多尺度融合等技术改善对大物体的识别效果。

参考文献（References）

Akyon F C, Onur Altinuc S and Temizel A. 2022. Slicing aided hyper inference and fine-tuning for small object detection//Proceedings of 2022 IEEE International Conference on Image Processing. Bordeaux, France: IEEE: 966-970 [DOI: 10.1109/ICIP46576.2022.9897990]

Bai Y C, Zhang Y Q, Ding M L and Ghanem B. 2018. SOD-MTGAN: small object detection via multi-task generative adversarial network//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 210-226 [DOI: 10.1007/978-3-030-01261-8_13]

Bochkovskiy A, Wang C Y and Liao H Y M. 2020. YOLOv4: optimal speed and accuracy of object detection [EB/OL]. [2024-08-15]. https://arxiv.org/pdf/2004.10934.pdf

Bodla N, Singh B, Chellappa R and Davis L S. 2017. Soft-NMS —— improving object detection with one line of code//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 5562-5570 [DOI: 10.1109/ICCV.2017.593]

Cao Y R, He Z J, Wang L J, Wang W G, Yuan Y X, Zhang D W, Zhang J L, Zhu P F, Van Gool L, Han J W, Hoi S, Hu Q H, Liu M, Cheng C, Liu F F, Cao G J, Li G Z, Wang H K, He J Y, Wan J F, Wan Q, Zhao Q, Lyu S C, Zhao W Z, Lu X Q, Zhu X K, Liu Y J, Lv Y X, Ma Y J, Yang Y T, Wang Z, Xu Z Y, Luo Z P,

Zhang Z M, Zhang Z G, Li Z H and Zhang Z X. 2021. VisDrone-DET2021: the vision meets drone object detection challenge results//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision Workshops. Montreal, Canada: IEEE: 2847-2854 [DOI: 10.1109/ICCVW54120.2021.00319]

Deng S T, Li S, Xie K, Song W F, Liao X, Hao A M and Qin H. 2021. A global-local self-adaptive network for drone-view object detection. IEEE Transactions on Image Processing, 30: 1556-1569 [DOI: 10.1109/TIP.2020.3045636]

Dong X Y, Bao J M, Chen D D, Zhang W M, Yu N H, Yuan L, Chen D and Guo B N. 2022. CSWin transformer: a general vision transformer backbone with cross-shaped windows//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 12114-12124 [DOI: 10.1109/CVPR52688.2022.01181]

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J and Houlsby N. 2021. An image is worth 16 × 16 words: transformers for image recognition at scale//Proceedings of the 9th International Conference on Learning Representations. [s. l.]: OpenReview.net

Du D W, Zhu P F, Wen L Y, Bian X, Lin H B, Hu Q H, Peng T, Zheng J Y, Wang X Y, Zhang Y, Bo L F, Shi H L, Zhu R, Kumar A, Li A J, Zinollayev A, Askergaliyev A, Schumann A, Mao B J, Lee B, Liu C, Chen C R, Pan C H, Huo C L, Yu D, Cong D, Zeng D N, Pailla D R, Li D, Wang D, Cho D, Zhang D Y, Bai F R, Jose G, Gao G Y, Liu G Z, Xiong H T, Qi H, Wang H R, Qiu H Q, Li H L, Lu H C, Kim I, Kim J, Shen J, Lee J, Ge J, Xu J J, Zhou J K, Meier J, Choi J W, Hu J H, Zhang J Y, Huang J Y, Huang K Q, Wang K Y, Sommer L, Jin L, Zhang L, Huang L H, Sun L, Steinmann L, Jia M X, Xu N, Zhang P Y, Chen Q, Lv Q X, Liu Q, Cheng Q S, Chennamsetty S S, Chen S H, Wei S, Kruthiventi S S S, Hong S, Kang S, Wu T, Feng T, Kollerathu V A, Li W Q, Dai W, Qin W D, Wang W Y, Wang X R, Chen X Y, Chen X, Sun X, Zhang X, Zhao X, Zhang X D, Zhang X Y, Chen X K, Wei X D, Zhang X Z, Li Y, Chen Y C, Toh Y H, Zhang Y, Zhu Y, Zhong Y X, Wang Z X, Wang Z K, Song Z C and Liu Z M. 2019. VisDrone-DET2019: the vision meets drone object detection in image challenge results//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision Workshop. Seoul, Korea(South): IEEE: 213-226 [DOI: 10.1109/ICCVW.2019.00030]

Feng C J, Zhong Y J, Gao Y, Scott M R and Huang W L. 2021. TOOD: task-aligned one-stage object detection//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 3490-3499 [DOI: 10.1109/ICCV48922.2021.00349]

Girshick R, Donahue J, Darrell T and Malik J. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE: 580-587 [DOI: 10.1109/CVPR.2014.81]

He K M, Gkioxari G, Dollár P and Girshick R. 2017. Mask R-CNN//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 2980-2988 [DOI: 10.1109/ICCV.2017.322]

He K M, Zhang X Y, Ren S Q and Sun J. 2015. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(9): 1904-1916 [DOI: 10.1109/TPAMI.2015.2389824]

Hou Q B, Zhou D Q and Feng J S. 2021. Coordinate attention for efficient mobile network design//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 13708-13717 [DOI: 10.1109/CVPR46437.2021.01350]

Jocher G, Stoken A, Borovec J, NanoCode, Chaurasia A, TaoXie, Liu C Y, Abhiram V, Laughing, Tkianai, yxNONG, Hogan A, Lorenzomammana, AlexWang 1900, HajekJ, DiaconuL, Marc, KwonY, Oleg, wanghaoyang0106, DefretinY, LohiaA, ml5ah, MilankoB, FineranB, KhromovD, DingY W, Doug, Durgesh and InghamF. 2021. Ultralytics/Yolov5: v5.0- YOLOv5-P6 1280 Models, AWS, Supervise.ly and YouTube Integrations. Zenodo [DOI: 10.5281/ZENODO.4679653]

Koyun O C, Keser R K, Akkaya İ B and Töreyin B U. 2022. Focus-and-Detect: a small object detection framework for aerial images. Signal Processing: Image Communication, 104: #116675 [DOI: 10.1016/j.image.2022.116675]

Lam D, Kuzma R, McGee K Dooley S, Laielli M, Klaric M, Bulatov Y and McCord B. 2018. xView: objects in context in overhead imagery [EB/OL]. [2024-08-15]. https://arxiv.org/pdf/1802.07856.pdf

Li C L, Yang T J N, Zhu S J, Chen C and Guan S Y. 2020. Density map guided object detection in aerial images//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Seattle, USA: IEEE: 737-746 [DOI: 10.1109/CVPRW50498.2020.00103]

Lin T Y, Dollár P, Girshick R, He K M, Hariharan B and Belongie S. 2017a. Feature pyramid networks for object detection//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 936-944 [DOI: 10.1109/CVPR.2017.106]

Lin T Y, Goyal P, Girshick R, He K M and Dollár P. 2017b. Focal loss for dense object detection//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE [DOI: 10.1109/ICCV.2017.324]

Liu S, Qi L, Qin H F, Shi J P and Jia J Y. 2018. Path aggregation network for instance segmentation//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 8759-8768 [DOI: 10.1109/CVPR.2018.00913]

Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C Y and Berg A C. 2016. SSD: single shot MultiBox detector//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 21-37 [DOI: 10.1007/978-3-319-46448-0_2]

Liu Z, Lin Y T, Cao Y, Hu H, Wei Y X, Zhang Z, Lin S and Guo B N. 2021. Swin transformer: hierarchical vision transformer using shifted windows//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 10012-10022 [DOI: 10.1109/ICCV48922.2021.00986]

Mahajan K, Sharma M and Vig L. 2020. Meta-DermDiagnosis: few-shot skin disease identification using meta-learning//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 730-731 [DOI: 10.1109/CVPRW50498.2020.00373]

Pan X Y, Jia N X, Mu Y Z and Gao X R. 2023. Survey of small object detection. Journal of Image and Graphics, 28(9): 2587-2615 (潘晓英, 贾凝心, 穆元震, 高炫蓉. 2023. 小目标检测研究综述. 中国图象图形学报, 28(9): 2587-2615) [DOI: 10.11834/jig.220455]

Qiao S Y, Chen L C and Yuille A. 2021. DetectoRS: detecting objects with recursive feature pyramid and switchable atrous convolution//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 10208-10219 [DOI: 10.1109/CVPR46437.2021.01008]

Redmon J, Divvala S, Girshick R and Farhadi A. 2016. You only look once: unified, real-time object detection//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 779-788 [DOI: 10.1109/CVPR.2016.91]

Redmon J and Farhadi A. 2017. YOLO9000: better, faster, stronger//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 7263-7271 [DOI: 10.1109/CVPR.2017.690]

Ren S Q, He K M, Girshick R and Sun J. 2017. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6): 1137-1149 [DOI: 10.1109/TPAMI.2016.2577031]

Shi Z H, Wu C W, Li C J, You Z Z, Wang Q and Ma C C. 2023. Object detection techniques based on deep learning for aerial remote sensing images: a survey. Journal of Image and Graphics, 28(9): 2616-2643 (石争浩, 仵晨伟, 李成建, 尤珍臻, 王泉, 马城城. 2023. 航空遥感图像深度学习目标检测技术研究进展. 中国图象图形学报, 28(9): 2616-2643) [DOI: 10.11834/jig.221085]

Solovyev R, Wang W M and Gabruseva T. 2021. Weighted boxes fusion: ensembling boxes from different object detection models. Image and Vision Computing, 107: #104117 [DOI: 10.1016/j.imavis.2021.104117]

Suo J S, Wang T Y, Zhang X Z, Chen H Y, Zhou W and Shi W S. 2023. HIT-UAV: a high-altitude infrared thermal dataset for unmanned aerial vehicle-based object detection. Scientific Data, 10(1): #227 [DOI: 10.1038/s41597-023-02066-6]

Tan M X, Pang R M and Le Q V. 2020 EfficientDet: scalable and efficient object detection//Proceedings of 2020 IEEE/CVF conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 10778-10787 [DOI: 10.1109/CVPR42600.2020.01079]

Wang C Y, Bochkovskiy A and Liao H Y M. 2023. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 7464-7475 [DOI: 10.1109/CVPR52729.2023.00721]

Wang J W, Xu C, Yang W and Yu L. 2021. A normalized Gaussian Wasserstein distance for tiny object detection [EB/OL]. [2024-08-15]. https://arxiv.org/pdf/2110.13389.pdf

Wang T, Chen Y, Qiao M N and Snoussi H. 2018. A fast and robust convolutional neural network-based defect detection model in product quality control. The International Journal of Advanced Manufacturing Technology, 94(9): 3465-3471 [DOI: 10.1007/s00170-017-0882-0]

Wang X L, Shrivastava A and Gupta A. 2017. A-Fast-RCNN: hard positive generation via adversary for object detection//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 3039-3048 [DOI: 10.1109/CVPR.2017.324]

Wang Y, Yang Y L and Zhao X. 2020. Object detection using clustering algorithm adaptive searching regions in aerial images//Proceedings of 2020 European Conference on Computer Vision Workshops. Glasgow, UK: Springer: 651-664 [DOI: 10.1007/978-3-030-66823-5_39]

Woo S, Park J, Lee J Y and Kweon I S. 2018. CBAM: convolutional block attention module//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 3-19 [DOI: 10.1007/978-3-030-01234-2_1]

Xu S L, Wang X X, Lv W Y, Chang Q Y, Cui C, Deng K P, Wang G Z, Dang Q Q, Wei S Y, Du Y N and Lai B H. 2022. PP-YOLOE: an evolved version of YOLO [EB/OL]. [2024-08-15]. https://arxiv.org/pdf/2203.16250.pdf

Yang C H Y, Huang Z H and Wang N Y. 2022. QueryDet: cascaded sparse query for accelerating high-resolution small object detection//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 13658-13667 [DOI: 10.1109/CVPR52688.2022.01330]

Yang F, Fan H, Chu P, Blasch E and Ling H B. 2019. Clustered object detection in aerial images//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 8310-8319 [DOI: 10.1109/ICCV.2019.00840]

Zhang Y, Ye M, Zhu G Y, Liu Y, Guo P Y and Yan J H. 2024. FFCA-YOLO for small object detection in remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 62: #5611215

［DOI：10.1109/TGRS.2024.3363057］

Zheng Z H，Wang P，Liu W，Li J Z，Ye R G and Ren D W. 2020. Distance-IoU loss：faster and better learning for bounding box regression//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York，USA：AAAI Press：12993-13000［DOI：10.1609/aaai.v34i07.6999］

Zheng Z H，Wang P，Ren D W，Liu W，Ye R G，Hu Q H and Zuo W M. 2022. Enhancing geometric factors in model learning and inference for object detection and instance segmentation. IEEE Transactions on Cybernetics，52（8）：8574-8586［DOI：10.1109/TCYB.2021.3095305］

Zhou C J，Wu M Q and Lam S K. 2019. SSA-CNN：semantic self-attention CNN for pedestrian detection［EB/OL］.［2024-08-15］. https://arxiv.org/pdf/1902.09080.pdf

Zhu X K，Lyu S C，Wang X and Zhao Q. 2021. TPH-YOLOv5：improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision Workshops. Montreal，Canada：IEEE：2778-2788［DOI：10.1109/ICCVW54120.2021.00312］

Zhu X Z，Su W J，Lu L W，Li B，Wang X G and Dai J F. 2020. Deformable DETR：deformable transformers for end-to-end object detection//Proceedings of the 9th International Conference on Learning Representations.［s.l.］：OpenReview.net

作者简介

郝川艳，副教授，硕士生导师，主要研究方向为计算机视觉和深度学习。E-mail：hcy@njupt.edu.cn

金怡，硕士研究生，主要研究方向为目标检测。
E-mail：1022163106@njupt.edu.cn

何桂琴，硕士研究生，主要研究方向为文本检测与分析。
E-mail：hgqwork@qq.com

张昊，本科生，主要研究方向为小物体目标检测。
E-mail：b20150322@njupt.edu.cn

祝翔祺，本科生，主要研究方向为小样本目标检测。
E-mail：b21150223@njupt.edu.cn

宋婉茹，讲师，硕士生导师，主要研究方向为目标检测和行人重识别。E-mail：songwanru@njupt.edu.cn