文章编号：2095-2163(2023)10-0083-05 中图分类号：TP391 文献标志码：A

基于 YOLOv5 的高分辨率遥感图像目标检测算法

李在瑞，郑永果，东野长磊
（山东科技大学 计算机科学与工程学院，山东 青岛 266590）

摘 要：针对高分辨率遥感图像中物体排布密集、尺度变化较大等特性，提出一种目标检测算法 R-YOLOv5。算法在 YOLOv5 模型基础上首先将跨阶段局部扩张结构作用于主干网络，采用一种加强的特征提取方式，通过整合空洞卷积和密集连接，来缓解模型对密集分布目标的漏检问题；其次，在主干网络的瓶颈部分结合 Transformer 模块来增强特征的表达，突出目标区域；最后，引入多尺度特征融合模块，解决多尺度特征融合时存在的不一致性问题，以提高模型的检测效果。在公开的遥感图像检测数据集 DIOR 的实验结果表明，R-YOLOv5 算法平均精度均值（*mAP*）达到 80.6%，具有良好的检测性能。

关键词：遥感图像；目标检测；分布密集；YOLO；空洞卷积

Object detection algorithm for high resolution remote sensing image based on YOLOv5

LI Zairui, ZHENG Yongguo, DONGYE Changlei
(College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao Shandong 266590, China)

【Abstract】 Aiming at the characteristics of dense distribution and large scale variation of objects in high-resolution remote sensing images, an object detection algorithm R-YOLOv5 is proposed. On the basis of YOLOv5 model, the algorithm firstly introduces Cross Stage Partial Dilated Network in the backbone network, which adopts an enhanced feature extraction method to alleviate the problem of undetected dense distributed targets by integrating dilated convolution and dense connection. Secondly, in the bottleneck part of the backbone network, the Transformer module is combined to enhance the expression of features and highlight the target area. Finally, multi-scale feature fusion module is introduced to solve the inconsistency problem in multi-scale feature fusion to improve the detection effect of the model. The experimental results on public remote sensing image detection dataset DIOR show that the MAP of R-YOLOv5 reaches 80.6%, which has good detection performance.

【Key words】 remote sensing image; object detection; dense distribution; YOLO; dilated convolution

0 引 言

近些年，随着卫星及遥感技术的发展，遥感图像的目标检测在城市规划、灾情救援、车辆监控等各种实际应用中起到了至关重要的作用[1]。深度学习技术的迅速发展，使得目标检测有了重大突破，许多高性能的神经网络算法被提出[2]。目前，基于深度学习的目标检测算法可以大致分为二阶段算法和一阶段算法两类，二阶段算法专注于提升模型对目标的检测精度，一阶段方法则在追求精度的基础上又兼顾了检测速度。

二阶段算法的经典模型是 Fast R-CNN[3]，其使用 Region Proposal Network（RPN）来选择对象的候选边界框，随后又进一步筛选出较为准确的目标区域。特征金字塔网络（FPN）[4]使用类似金字塔的结构来学习不同尺度的特征。Tridentnet[5]通过引入扩展卷积来改变大小最佳的感受野，并基于不同大小的感受野构造多分支结构，从而解决多尺度检测问题。一阶段模型中，SSD[6]增加了多个卷积层，以获得多尺度特征图进行预测，并设计不同大小的先验边界框以更好地检测目标。YOLOv4[7]采用了更为高效的 csp-darknet 作为主干网络并设计多尺度预测。TPH-YOLOv5[8]则将 Transformer 与网络相结合，增强模型提取特征的能力。

以上算法虽然在识别自然图像时都表现出了良好的效果，但由于遥感图像存在背景复杂、目标尺度

作者简介：李在瑞（1998-），男，硕士研究生，主要研究方向：计算机视觉；郑永果（1963- ），男，博士，教授，主要研究方向：虚拟现实与可视化、图像处理与模式识别；东野长磊（1978-），男，博士，副教授，主要研究方向：医学图像处理、计算机视觉。
通讯作者：郑永果 Email：skd991317@ sdust.edu.cn

收稿日期：2022-11-05

变化范围大、物体分布密集等检测难题[9]，通用目标检测算法对高分辨率遥感图像的检测具有很大的局限性[10]。为解决上述问题，本文基于 YOLOv5 框架，提出特征信息补充与加强以及多尺度融合的方法，以增强模型的检测能力。

1 相关工作

1.1 YOLOv5 模型

随着 YOLO 系列网络的提出，其在各种视觉检测任务中展现了出色的性能。其中，YOLOv5 主干网络是由 Focus 模块、CSP 结构以及 SPP 模块组合而成。Focus 模型会对图片进行切片操作，在宽和高两个维度上每隔一个像素取一个值，从而使特征图的通道数变为原来的 4 倍，能够在最大程度减少信息损失的同时实现两倍下采样。YOLOv5 在 CSPNet[11]的基础上重新设计 csp 结构，并在原本的 darknet 网络中大量插入该结构。spp 模块对特征图做不同大小的池化操作，从而在原特征图的基础上融合不同感受野，丰富上下文信息[12]。

YOLOv5 在 Nick 部分结构参考了 FPN 和 PAN。首先，设计自顶向下路径来融合网络中不同层次的特征，将包含丰富语义信息的深层特征向下传递与浅层结合，能够提高模型对多尺度目标的检测能力；后又增加自底向上的金字塔结构，把浅层特征映射到深层网络，补充检测目标的细节及空间信息，进一步提升模型的检测效果。同时，在 nick 部分应用 csp2_x 结构，使用 X 个卷积模块替代残差单元。

Head 部分则对图片进行预测与分类，YOLOv5 设计 3 种尺寸的特征图来检测大中小不同种类的目标，最后通过非极大值抑制来筛选预测框，实现检测过程。

1.2 Transformer 模块

Transformer 模块早先广泛应用于 NLP 领域，通过自注意力机制来捕获序列元素之间的依赖关系，在可并行性和特征提取方面展现了出色的性能[13]。近些年来，许多计算机视觉的学者开始将其作用于图像相关的研究上。Parmar 等人提出 Image Transformer[14]算法，基于 Transformer 解码器用于图像生成任务；随后 Vision Transformer[15]被提出，并首次在大型图像数据集上展现出超越卷积网络的性能，在图像分类方面具有较强的泛化能力；Swin Transformer[16]则采用移动窗口的机制来计算注意力，有效解决了传统 Transformer 模块中计算复杂度较高的问题，并通过不同窗口之间的特征交互提取到更为丰富的语义信息。

Transformer 由编码器和解码器两部分组成，基本原理是通过将图片展开成一维，得到图像特征张量，输入到编码器部分使用多头自注意力学习目标特征，增强图像中目标的语义信息，再利用解码器与解码器协同训练，学习注意力规律来强化目标和特征之间的关联关系，进而提升检测效果。

2 R-YOLOv5 遥感图像目标检测算法

R-YOLOv5 目标检测算法结构如图 1 所示。首先，在 YOLOv5 的主干网络 CSPDarkNet 中使用跨阶段局部扩张结构，替代原本的跨阶段局部网络结构；其次，在主干网络的输出特征图瓶颈部分结合 Transformer 模块中的编码器；最后，在原本的 Nick 部分嵌入多尺度特征融合模块。

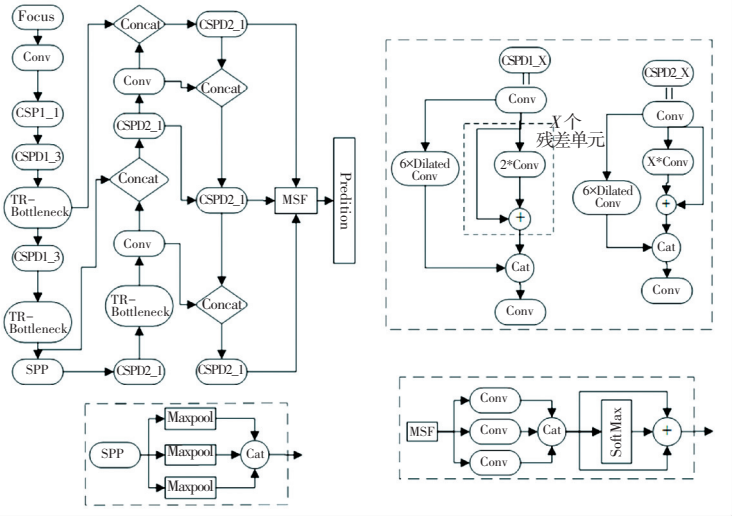


图 1 R-YOLOv5 算法结构

Fig. 1 R-YOLOv5 algorithm structure

2.1　跨阶段局部扩张结构

跨阶段局部网络结构（Cross Stage Partial Structure，CSP）被大量应用到 YOLOv4 的主干网络，YOLOv5 又在 v4 的基础上将其与 nick 部分结合。CSP 结构包括两个分支：一是将输入特征图进行 X 个残差单元的卷积操作，另一部分进行简单的 3 * 3 卷积计算特征后，与上一分支结合。CSP 结构能够增强网络的特征提取能力，使模型获取到更为丰富的语义信息。

针对遥感图像中检测目标尺度变化较大，物体分布密集的特性，对 CSP 结构进行改进，提出跨阶段局部扩张结构（Cross Stage Partial Dilated Structure，CSPD），如图 2 所示。首先，保持残差单元分支不变，在另一分支中使用 6 个连续的扩张卷积，扩张率分别为 3、6、12、18、24，来获取同一特征图的不同感受野，从而覆盖遥感图像中各种不同尺度的检测对象。

其次，当图像中目标分布较为紧密时，使用扩张卷积会丢失特征信息，为了避免检测对象的漏检现象，在连续的 6 个扩张卷积基础上采用密集连接结构，将原特征图与每层的卷积分别做逐个元素的加操作，从而加强特征的传播，丰富语义信息。

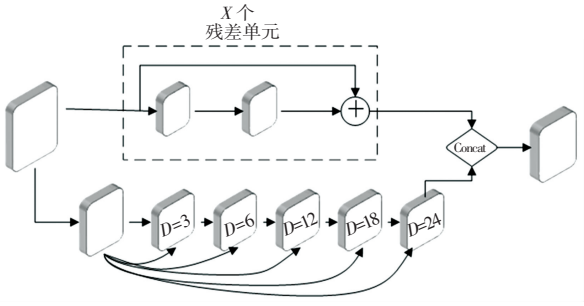


图 2　跨阶段局部扩张模块结构图

Fig. 2　Cross Stage Partial Dilated module

2.2　瓶颈 Transformer 结构

YOLOv5 主干网络分别输出 3 个不同层次大小的特征图，作为后续多尺度特征融合部分的输入。将主干网络中负责输出特征图的瓶颈（Bottleneck）部分与 Transformer 模块中的编码器相结合（如图 3 所示），提出瓶颈 Transformer 结构（TR－Bottleneck），提高模型对语义信息的提取能力，丰富图像全局信息，抑制背景对目标识别的影响。

首先，将图片做切分并降低维度，即将原本 $H * W * C$ 的图像变为 $N * (P^2 * C)$ 的 Tokens，其中 $N = \frac{H}{W} * P^2$；随后输入 Encoder 中的多头注意力机制，进一步做特征提取，如式（1）所示：

$$Atten(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \qquad (1)$$

式中：$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 分别为输入多头注意力的查询向量、键向量、值向量，d_k 代表特征维度。将查询向量与键向量相乘后，经过 softmax 激活函数并归一化处理，再与 $\boldsymbol{V}$ 相乘加权，得到输出结果。

最后输入由两个全连接层及激活函数组成的 MLP（前馈神经网络）得到整个 Transformer 模块的输出特征，并与 Bottlenck 结构的特征信息结合。

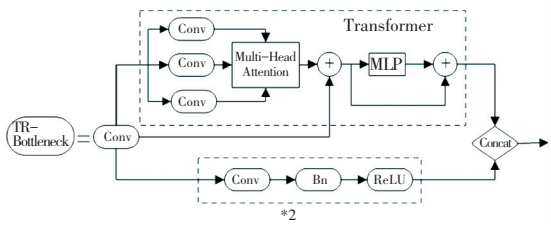


图 3　瓶颈 Transformer 模块结构图

Fig. 3　Transformer bottleneck module

2.3　多尺度特征融合模块

YOLOv5 输出的 3 种尺寸的特征图，分别对应大中小不同的检测对象，高层语义信息中检测大目标，低层语义信息中检测小目标，而遥感图像中往往既有大目标又有小目标。特征融合时，由于不同层间特征的不一致性，将会影响最后的检测结果。为了缓解上述问题，更好的让网络利用高低层语义信息，在 nick 部分的最后，嵌入多尺度特征融合模块（Multi Scale Feature Fusion Module，MSF），如图 4 所示。

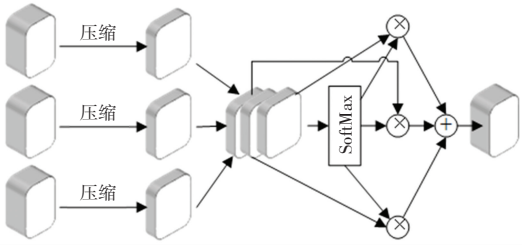


图 4　多尺度特征融合模块结构图

Fig. 4　Multi-scale feature fusion module

首先将 3 种尺寸的特征图进行采样操作，调整到同一尺寸；再根据通道维度整合并接入 SoftMax 函数生成权重参数；最后 3 层特征分别乘上各自的权重参数，得到融合后的特征，表达如式（2）所示：

$$f = \sum_{i=1}^{3} \mathrm{SoftMax}(cat(x_1\ x_2\ x_3)) \otimes x_i \qquad (2)$$

式中：x_1、x_2、x_3 分别为 3 种尺寸的特征图，cat 表示对特征图做通道维度的整合，$\otimes$ 表示点乘操作，f 则为最终的输出特征。

3 实验

3.1 实验环境与数据集

实验在 linux 系统下进行,所用 GPU 为 Tesla P100,显存 16 G,深度学习框架为 pytorch。实验所用遥感数据集为 DIOR,其中包括 23 463 张图像,训练与测试各取一半的样本。

3.2 评价指标

实验采用平均精度均值(*mAP*)、平均精确率(*AP*)作为评估指标,*AP* 和 *mAP* 是可以反映多类别目标全局检测精度的指标在文献中被广泛用于评估多类别目标检测性能表达如式(2)、(3)所示:

$$AP = \int_0^1 p(R)\,dR \tag{3}$$

$$mAP = \frac{1}{N}\sum_i AP_i \tag{4}$$

其中,平均精度 *AP* 表示的是计算单类目标 *P* - *R* 曲线下面积的结果, *p* 为精确率, *R* 为召回率;而 *mAP* 是所有类别 *AP* 的平均值; *N* 为检测目标的类别总数; AP_i 表示第 *i* 个类别的平均检测精度。

3.3 算法流程

如图 5 所示,R-YOLOv5 算法首先对输入的遥感图像进行预处理,扩展图像数据;其次,根据模型配置文件搭建网络结构,读取训练参数,并根据训练结果更新网络参数;最后,加载训练权重与测试数据集,输出模型的预测图像。

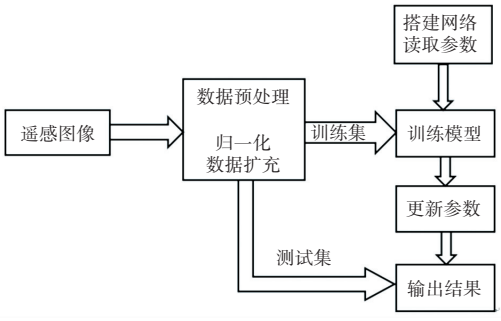


图 5 R-YOLOv5 算法流程图

Fig. 5 R-YOLOv5 algorithm flowchart

3.4 实验结果

表 1 为本文算法 R-YOLOv5 与不同目标检测模型在 DIOR 数据集下的实验结果。其中包括一阶段模型 Faster-RCNN,以 SSD、RetinaNet、YOLOv4 为代表的二阶段模型,及无锚方法 YOLOX。

表 1 DIOR 数据集下对比试验

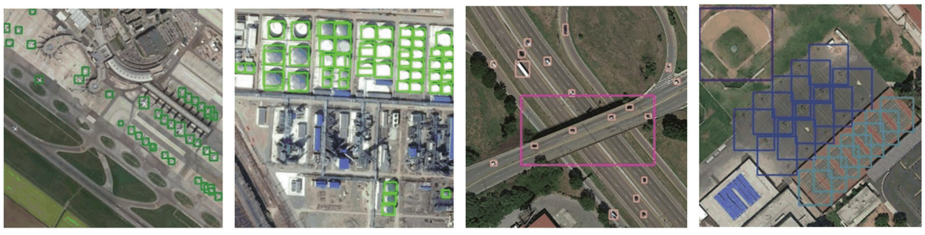
Tab. 1 Results on Dior dataset %

METHOD	Faster-RCNN	SSD	RetinaNet	YOLOv4	YOLOX	R-YOLOv5
Expressway service area	65	64	90	89	80	**93**
Basketball court	71	76	90	87	89	**92**
Tennis court	77	76	87	88	90	**92**
golffield	70	65	85	74	72	**86**
Ground track field	62	69	83	82	81	**88**
Stadium	94	61	81	70	74	**80**
Chimney	89	66	81	80	76	**82**
Airport	68	72	79	80	71	**92**
Dam	59	57	75	70	61	**81**
Baseball field	92	72	74	85	84	**81**
Wind mill	44	66	70	83	89	**92**
Airplane	91	60	68	73	85	**84**
Trainstation	40	55	61	63	48	**75**
Expressway toll station	55	53	59	71	71	**83**
Harbor	54	49	59	63	52	**67**
Overpass	51	48	57	62	61	**66**
Ship	21	59	47	85	88	**91**
bridge	22	30	37	44	44	**55**
Storagetank	73	47	34	63	70	**76**
Vehicle	30	27	21	44	49	**58**
MAP	61.58	58	66.92	72.69	71.7	**80.6**

由表 1 可知,R-YOLOv5 对飞机、机场、船、桥、车辆等密集分布、大小尺度不一目标的精度均有不同程度的提高,具有良好的表现。

图 6 所示为 R-YOLOv5 对密集分布、大小尺度不一目标的效果图。这两种情况在检测过程中都较易对目标错检或漏检,模型识别的难度较大。如图

6(a)、(b)中飞机与油罐的分布较为密集，模型对此类目标能够较为全面的做出识别；图 6(c)、(d)中车辆与桥梁、棒球场与网球场等各类物体的尺度变化给模型带来了检测难题，结果表明，R-YOLOv5 可以较为准确的检测出目标对象。



(a) 飞机场　(b) 油罐场　(c) 车辆与桥梁　(d) 棒球场与网球场

图 6　R-YOLOv5 检测结果

Fig. 6　R-YOLOv5 detection result

4　结束语

基于高分辨率遥感图像存在检测对象密集度高、大小不一等问题。本文提出 R-YOLOv5 算法，通过扩大感受野和增强特征信息以及改善特征融合来提高模型对密集物体以及多尺度目标的检测精度。实验表明，本文提出的目标检测算法在遥感数据集上具有较好的识别能力。

参考文献

[1] SCHILLING H, dULATOV D, NIESSNER R, et al. Detection of vehicles in multisensor data via multibranch convolutional neural networks[J]. IEEEE Journal of SelectedTopics in Applied Earth Observations and Remote Sensing, 2018, 11(1):4299-4316.

[2] CHEN J, YUE A, WANG C, et al. Wind turbine extraction from high spatial resolution remote sensing images based on saliency detection[J]. Journal of Applied Remote Sensing, 2018, 12(1): 016041.

[3] GIRSHICK R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.

[4] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2117-2125.

[5] LI Y, CHEN Y, WANG N, et al. Scale-aware trident networks for object detection[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 6054-6063.

[6] LIU W, ANGUELOV D, ERHAN D, et al. Ssd: Single shot multibox detector[C]//Computer Vision-ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I 14. Springer International Publishing, 2016: 21-37.

[7] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. Yolov4: Optimal speed and accuracy of object detection[J]. arXiv preprint arXiv:2004.10934, 2020.

[8] ZHU X, LYU S, WANG X, et al. TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 2778.

[9] ZHANG G, LU S, ZHANG W. CAD-Net: Acontext0aware detection network for objects in remote sensing imagery[J]. IEEE Transactions on Geoscience and Remote Sensing, 2019, 57(12): 10015-10024.

[10] ZHENG Z, LEI L, SUN H, et al. A review of remote sensing image object detection algorithms based on deep learning[C]// 2020 IEEE 5th International Conference on Image, Vision and Computing (ICIVC). IEEE, 2020: 34-43.

[11] WANG C Y, LIAO H Y M, WU Y H, et al. CSPNet: A new backbone that can enhance learning capability of CNN[C]// Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. 2020: 390-391.

[12] CAO L, ZHANG X, WANG Z, et al. Multi angle rotation object detection for remote sensing image based on modified feature pyramid networks[J]. International Journal of Remote Sensing, 2021, 42(14): 5253-5276.

[13] WANG C, BAI X, WANG S, et al. Multiscale Visual attention networks for object detection in VHR remote sensing images[J]. IEEE Geoscience and Remote Sensing Letters, 2018, 16(2): 310-314.

[14] PARMAR N, VASWANI A, USZKOREIT J, et al. Image transformer[C]//International conference on machine learning. PMLR, 2018: 4055-4064.

[15] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.

[16] 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.