第 60 卷 第 12 期/2023 年 6 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

基于跨层注意力增强的遥感小目标检测

韩兴勃[1,2]，李凡[1,2*]

[1]昆明理工大学信息工程与自动化学院，云南 昆明 650504；

[2]云南省人工智能重点实验室，云南 昆明 650504

摘要 针对遥感图像中小目标对象存在像素少、信息有限、检测困难和失准等实际问题，对YOLOv5进行改进，提出增加残差连接与跨层注意力的方法来提升模型对遥感图像中小目标的检测能力。该方法采用对特征图进行残差连接并增加检测头的方式，有效提高了YOLOv5在遥感图像中对小目标的检测能力。此外，还通过跨层注意力为不同网络层的特征附加语义信息，从而提高模型对遥感图像中复杂背景信息的抑制能力。在Detection in Optical Remote(DIOR)遥感数据集的实验中，所提方法取得了86.4%的平均精度均值(mAP)，小目标检测精度评价指标(APs)达23.4%，比基准网络高出5.9个百分点。实验结果表明，所提方法在遥感图像小目标检测问题上具有较好性能，同时也验证了特征金字塔中底层特征图与注意力机制对提升小目标检测性能具有十分重要的作用。

关键词 小目标检测；遥感图像；跨层注意力；特征金字塔

中图分类号 TP391.4 **文献标志码** A **DOI**：10.3788/LOP221744

Remote Sensing Small Object Detection Based on Cross-Layer Attention Enhancement

Han Xingbo[1,2], Li Fan[1,2*]

[1]*Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming* 650504, *Yunnan, China*;

[2]*Yunnan Key Laboratory of Artificial Intelligence, Kunming* 650504, *Yunnan, China*

Abstract To address the practical issues of few pixels, limited information, detection difficulties, and misalignment of small objects in remote sensing images, this paper aims to improve the YOLOv5 and proposes a technique of boosting residual connections and cross-layer attention to improve the model's detection capability for small objects in remote sensing images. To effectively improve the detection capability of YOLOv5 for small objects in remote sensing images, the method employs residual linking for feature maps and the addition of detection heads. Furthermore, using cross-layer attention, this paper attaches semantic informations to the features of different network layers, improving the model's ability to suppress complex background informations in remote sensing images. In the experiments on the Detection in Optical Remote (DIOR) remote sensing dataset, the proposed approach achieves a mean accuracy precision (mAP) of 86.4% and a small object detection accuracy evaluation metrics (APs) of 23.4%, which is 5.9 percentage points higher than the benchmark network. The experimental results show that the method proposed in this research performs well in small object detection problems in remote sensing images, and it also confirms that the bottom feature map and attention mechanism in the feature pyramid are critical for improving small object detection performance.

Key words small object detection; remote sensing image; cross-layer attention; feature pyramid

1 引 言

目标检测旨在通过特定算法在数字图像中寻找若干特定目标对象，针对遥感图像时，可应用于环境监管、植物保护、城市监测等实际问题。近年，基于深度学习的目标检测技术取得了显著进展，涌现出大量性能优越的目标检测模型[1-5]。虽然这些主流方法在遥感图像中的大尺度与中尺度目标检测方面有着不错的

收稿日期：2022-05-31；**修回日期**：2022-07-01；**录用日期**：2022-07-14；**网络首发日期**：2022-07-24

基金项目：云南省重大科技专项(202002AD080001)、云南省科技厅科技计划项目(基础研究专项)(202101AT070136)、国家自然科学基金(62161015)

通信作者：*478263823@qq.com

效果，但是在船只调度、车辆管控等有关遥感小目标检测任务中由于它们存在像素少、信息有限等问题仍面临着挑战（COCO数据集[6]中将像素数小于32×32的物体定义为小目标）。为解决小目标检测困难的问题，研究者们在该领域进行了深入探索，并先后提出了大量优秀的小目标检测方法。FPN模型[7]分别通过两个自底向上的分支与自顶向下的分支，产生多尺度的特征，有利于对小目标进行检测。文献[8]与文献[9]对多尺度特征图进行融合并取得不错的效果。在PANet[10]中，Lin等[7]对FPN模型中不同尺寸的特征图进行再次融合，使得FPN[7]的高层特征图也拥有丰富的底层特征与多层特征。Libra R-CNN[11]则提取了4个不同尺度的特征图，将其集成为顶层特征图后再与各不同尺寸特征图进行融合。文献[12]与文献[13]也通过在SSD网络[3]中引入注意力机制与上下文信息提升SSD网络[3]的小目标检测能力。虽然这几种方法都利用不同的特征融合方式分别在COCO数据集[6]中提升了小目标检测性能，但它们都并未在遥感图像中的小目标上进行测试，忽略了遥感图像小目标更容易丢失细节纹理特征的问题与遥感图像中复杂背景信息对小目标检测的干扰问题。

鉴于YOLOv5技术路线较为成熟且具有高检测精度与高检测速度的优势，本文以YOLOv5为基础进行改进来解决现有方法在遥感图像小目标检测中存在的问题。提出ResCatPAN结构，通过ResCat模块将主干网络中提取的更为精准的目标位置信息逐步补充到高层特征图中，从而使每一层特征图的位置信息更加准确；同时额外增加了一层特征图来适应遥感数据集中小目标的特征分布并缓解数据集中大目标与小目标之间剧烈的尺度变化。此外，本文提出了跨层注意力（cross-layer attention）模块，通过跨层注意力模块对特征图金字塔中的高层特征图进行通道权重计算，逐步为低层特征图赋予通道权重来加强其语义信息，抑制特征图中复杂背景信息的负面影响，进而提升模型对遥感图像背景信息的抑制能力。在Detection in Optical Remote（DIOR）遥感数据集[14]进行实验的结果表明，所提方法的小目标检测精度评价指标（APs）达23.4%，比基准网络高出5.9个百分点。

2　YOLO基本原理

随着深度学习的不断发展，不断涌现出大量的性能优越的目标检测模型，其中单阶段检测算法YOLO系列因为实时性与准确度都能满足人们的需要，不断成为实际应用中的首选。

YOLO系列中最具有代表性的算法则为YOLOv3[15]。其网络的基本结构主要由输入端、Backbone、Neck与Head四部分组成。输入图像经输入端进行变换后输入到Backbone中；对图像进行特征提取，提取出三个不同尺寸的特征层；然后将三个特征层送入Neck部分与上采样后的其他特征层进行堆叠拼接；最后将结果送入Head部分，对三个不同尺寸特征层进行检测，依次获得目标的位置和类别。YOLOv5则是

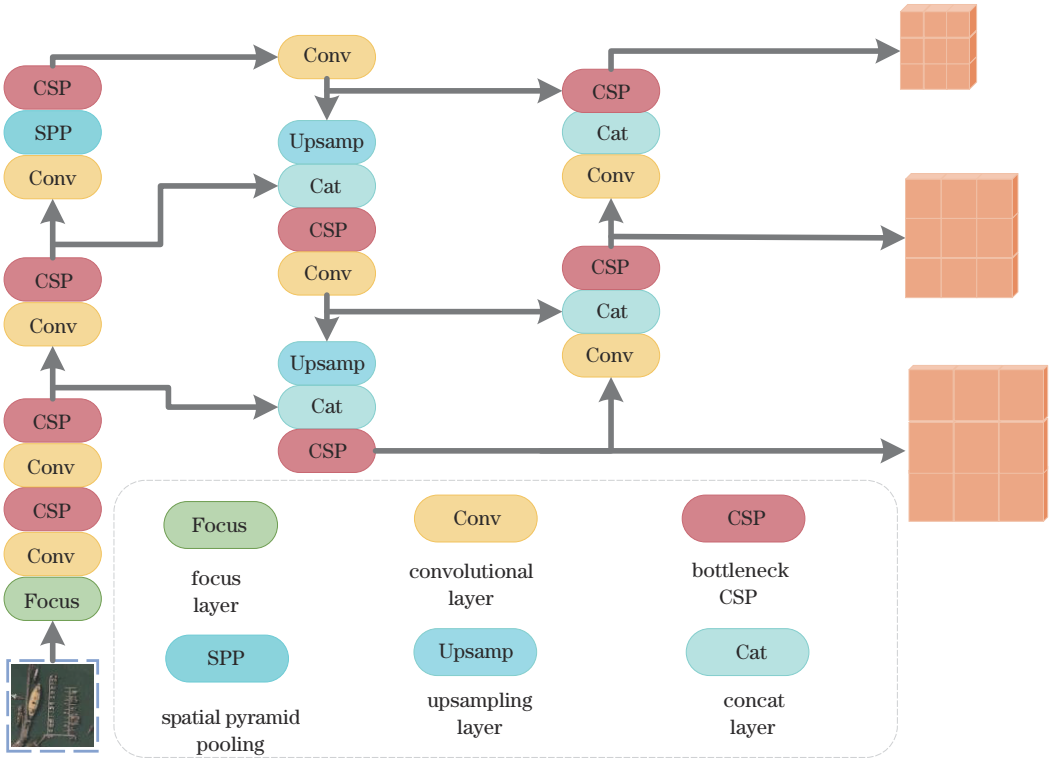

图1　YOLOv5网络细节

Fig. 1　Details of YOLOv5 network

YOLO 系列中最新的模型，网络细节如图 1 所示，其检测速度与精度与 YOLOv3[15] 相比都有很大提升，其输入部分相较 YOLOv3[15] 增加了数据增强与自适应锚框两部分。数据增强为 Mosaic 数据增强，通过对 4 张输入图像以随机缩放、随机裁剪、随机排布的方式进行拼接，以达到增广样本数量的目的。自适应锚框计算首先计算默认 anchor 的最大可能召回率，如果该值小于 98%，就利用 K-means 算法和遗传算法更新 anchor。在 Backbone 部分中，YOLOv5 较 YOLOv3[15] 新增了 Focus 结构与 CSP 结构。其中，Focus 结构用来对输入特征图进行切片处理，可有效减少网络的参数量，从而增加网络的推理速度。CSP 结构则优化了网络的梯度传递，解决了梯度信息重复的问题，从而优化网络的计算量。在 Neck 部分，YOLOv5 采用 path aggregation network（PAN）结构进行多尺度特征融合，具有自顶向下、自底向上的特点，卷积部分采用了与主干网络不同的 CSP 结构，具有更好的多样性和鲁棒性。对所有特征图在 Neck 部分进行融合后将结果输入到 Head 部分，分别检测并依次获得目标的位置和类别。

3　所提方法内容

3.1　整体框架

所提网络主要由输入端、Backbone、Neck 与 Head 四部分组成。首先对图像在输入端进行数据增强后，从 Backbone 中通过 Focus 与 CSP（bottleneck CSP）等结构提取出不同尺寸的特征图；然后将这些不同尺寸的特征图送入到 Neck 部分并进行融合，使各尺寸特征图拥有较强的语义信息及位置信息后再送入后续输出端 Head 部分，对各尺寸特征图分别进行检测。通过对 YOLOv5 的 Neck 部分和 Head 部分进行改进，整体结构如图 2 所示，从而提升了它在 DIOR 数据集[14] 上的性能，使其在遥感小目标上有着更好的表现。

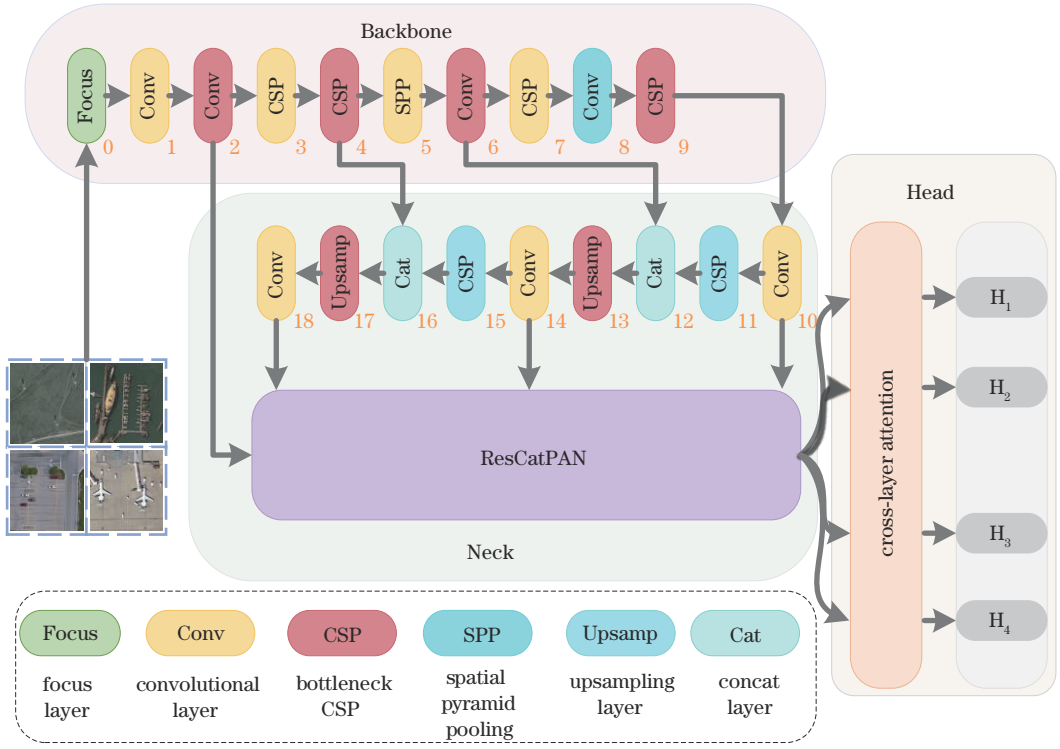

图 2　所提模型的整体结构

Fig. 2　Overall structure of the proposed model

3.2　ResCatPAN 结构

在 YOLOv5 中，输入图像经 Backbone 提取特征后，提取出的（$\boldsymbol{c}_4$, $\boldsymbol{c}_6$, $\boldsymbol{c}_9$）三层特征图经融合后输出到三个检测头。虽然这三层特征图在大目标和中目标检测中能够提供充足的位置信息及语义信息，但对于小目标检测来说，较粗粒度的特征图反而会丢失精确的位置信息。而且，由于 DIOR 数据集[14] 中目标大小存在着显著差异，三层特征图与三个检测头并不能很充分地适应 DIOR 数据集[14] 中目标的特征分布。

为解决该问题，提出 ResCatPAN 结构来获取更精确的小目标的位置信息并在特征融合中加以充分利用。在该结构的输入中增加从 Backbone 模块中提取出的 $\boldsymbol{c}_2$ 特征图，并在 PAN 结构中利用融合方式对（$\boldsymbol{c}_2$, $\boldsymbol{c}_4$, $\boldsymbol{c}_6$, $\boldsymbol{c}_9$）四层特征图进行改进，使 ResCatPAN 在小目标检测上拥有更好的性能。在 ResCatPAN 结构中加入一个新的 ResCat 模块，通过在结构中连用该模块，可以将底层特征图的位置信息逐步补充到高层特征图中，从而使每一层特征图的位置信息更加准确。此外，为了充分适应 DIOR 数据集[14] 中小目标的特征分布并缓解数据集中大目标与小目标之间剧烈的尺度

变化，在 YOLOv5 三个检测头（H_1, H_2, H_3）的基础上又额外增加了一个检测头 H_4，将在 ResCat 模块中生成的四层特征图（$\boldsymbol{x}_0$, $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, $\boldsymbol{x}_3$）分别送往这四个检测头进行检测。ResCatPAN 结构如图 3 所示，将 YOLOv5 中的 FPN 结构[7]输出的三层特征图（$\boldsymbol{c}_{10}$, $\boldsymbol{c}_{14}$, $\boldsymbol{c}_{18}$）与从 Backbone 中提取的 $\boldsymbol{c}_2$ 一起送入 ResCatPAN 结构中。

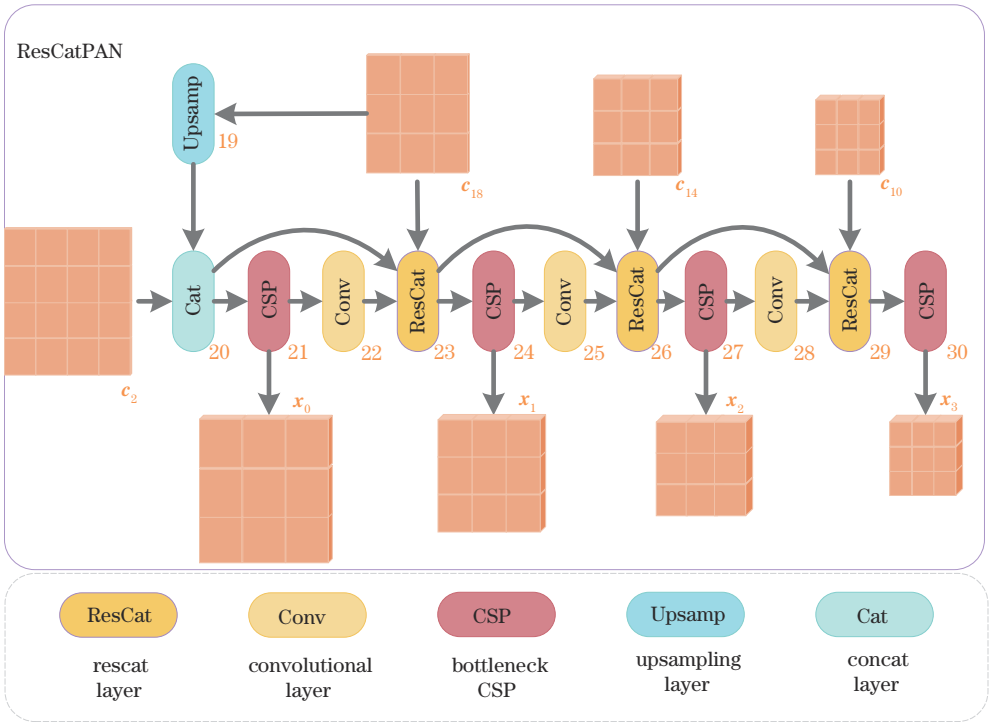

图 3　ResCatPAN 结构

Fig. 3　ResCatPAN structure

为了使 H_4 检测头能够更好地检测小目标，对 $\boldsymbol{c}_{18}$ 特征图进行了一次上采样，扩充了 $\boldsymbol{c}_{18}$ 特征图的尺寸，然后对其与具有丰富小目标位置信息的 $\boldsymbol{c}_2$ 特征图进行 Cat 后生成具有语义信息的 $\boldsymbol{c}_{20}$ 特征图，再经后续 CSP 模块对特征进行聚合细化后生成特征图 $\boldsymbol{x}_0 \in (320, 32, 32)$，整个过程表示为

$$\boldsymbol{x}_0 = \mathrm{CSP}\left\{\mathrm{Cat}\left[\boldsymbol{c}_2, \mathrm{Upsamp}(\boldsymbol{c}_{18})\right]\right\}。\tag{1}$$

随后对 $\boldsymbol{x}_0$ 进行卷积后生成 $\boldsymbol{c}_{22}$，$\boldsymbol{c}_{22}$ 与 $\boldsymbol{c}_{20}$、$\boldsymbol{c}_{18}$ 特征图一块输入 ResCat 模块，ResCat 模块结构如图 4 所示；在该模块中，使用一层 3×3 的卷积层来使 $\boldsymbol{c}_{20}$ 的通道数与尺寸减半，减少它不准确的语义信息表达；接着对 $\boldsymbol{c}_{20}$ 特征图进行一次归一化与激活函数处理，通过该操作，可以保证 $\boldsymbol{c}_{20}$ 特征图的位置信息的准确性并减少不准确语义信息的干扰；然后，再对其与具有准确语义信息的 $\boldsymbol{c}_{18}$ 进行 Cat 操作，来补充 $\boldsymbol{c}_{20}$ 的通道语义信息并用 1×1 卷积进行特征融合；最后用融合后的特征图与 $\boldsymbol{c}_{22}$ 进行 Cat 与 1×1 卷积操作，将信息融入到 $\boldsymbol{c}_{22}$ 中，生成 $\boldsymbol{x}_{\mathrm{down}} \in (640, 16, 16)$。

将所有特征融合后，将其分别作为下一个 CSP 模块与 ResCat 模块的输入，经 CSP 模块后生成 $\boldsymbol{x}_1$：

$$\boldsymbol{x}_{\mathrm{down}} = \mathrm{Conv}\left\{\mathrm{Cat}\left\{\boldsymbol{c}_{20}, \mathrm{Conv}\left\{\mathrm{Cat}\left\{\mathrm{ReLU}\left\{\mathrm{BN}\left[\mathrm{Conv}(\boldsymbol{c}_{20})\right]\right\}, \boldsymbol{c}_{18}\right\}\right\}\right\}\right\},\tag{2}$$

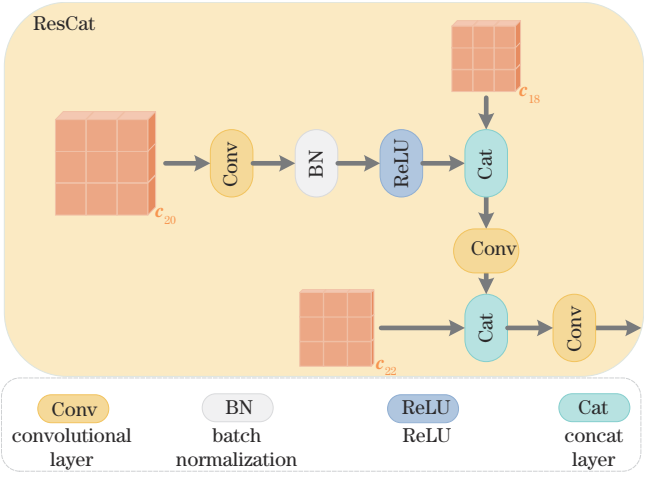

图 4　ResCat 结构

Fig. 4　ResCat structure

$$\boldsymbol{x}_1 = \mathrm{CSP}(\boldsymbol{x}_{\mathrm{down}}),\tag{3}$$

式中：$\boldsymbol{x}_1 \in (640, 16, 16)$。并且在后续 ResCat 模块中依次对 $\boldsymbol{c}_{14}$、$\boldsymbol{c}_{10}$ 进行相同处理，生成 $\boldsymbol{x}_2$、$\boldsymbol{x}_3$ 特征图：

$$\boldsymbol{x}_{\mathrm{mid}} = \mathrm{Conv}\left\{\mathrm{Cat}\left\{\mathrm{Conv}(\boldsymbol{x}_1), \mathrm{Conv}\left\{\mathrm{Cat}\left\{\mathrm{ReLU}\left\{\mathrm{BN}\left[\mathrm{Conv}(\boldsymbol{x}_{\mathrm{down}})\right]\right\}, \boldsymbol{c}_{14}\right\}\right\}\right\}\right\},\tag{4}$$

$$\boldsymbol{x}_2 = \mathrm{CSP}(\boldsymbol{x}_{\mathrm{mid}}),\tag{5}$$

$$\boldsymbol{x}_3=\mathrm{CSP}\left\{\mathrm{Conv}\left\{\mathrm{Cat}\left\{\mathrm{Conv}(\boldsymbol{x}_2),\right.\right.\right.$$
$$\left.\left.\left.\mathrm{Conv}\left\{\mathrm{Cat}\left\{\mathrm{ReLU}\left\{\mathrm{BN}\left[\mathrm{Conv}(\boldsymbol{x}_{\mathrm{mid}})\right]\right\},\boldsymbol{c}_{10}\right\}\right\}\right\}\right\},\quad(6)$$

式中：$\boldsymbol{x}_{\mathrm{mid}}\in(1280,8,8)$，$\boldsymbol{x}_2\in(960,8,8)$，$\boldsymbol{x}_3\in(1280,4,4)$。经过该结构的处理，可以使最终输出的每一层特征图$(\boldsymbol{x}_0,\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3)$都具有更加准确的目标位置信息与语义信息，提升对每一层特征图的检测能力。

3.3　跨层注意力模块（cross-layer attention）

近年来，注意力模型被广泛使用在图像识别任务中并表现出优秀的性能，因此本文也尝试在YOLOv5模型中引入注意力机制，通过对Neck部分输出的四层特征图$(\boldsymbol{x}_0,\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3)$分别使用不同的注意力机制，为每层特征图增强语义信息或位置信息，进而达到提高小目标检测性能的目的。由于特征金字塔的高层特征具有丰富的语义信息，通过使用注意力机制，将高层特征的语义信息依次映射到底层特征中，对每层特征进行再次增强，如图5所示。

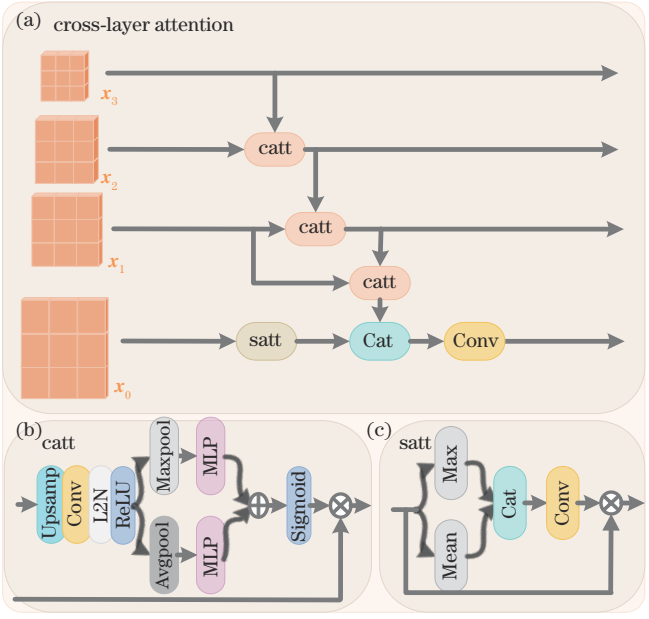

图5　cross-layer attention整体结构。(a) cross-layer attention完整流程；(b) cross-layer attention结构中的catt模块；(c) cross-layer attention结构中的satt模块

Fig. 5　Overall structure of the cross-layer attention. (a) Complete flow of the cross-layer attention; (b) catt module of the cross-layer attention; (c) satt module of the cross-layer attention

在该部分中，通过在$(\boldsymbol{x}_0,\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3)$之间进行跨层注意力计算，高层特征的语义信息逐层传递到底层，为提高小目标检测性能起到了一定的积极作用。在特征金字塔中，如果简单地对不同尺寸特征图进行拼接，则会不可避免地将高层特征图中不精确的空间位置信息引入到底层特征图中，降低底层特征图中空间位置信息的准确性，进而对小目标检测起到不利影响。为了避免高层粗粒度位置信息干扰底层细粒度位置信息，设计了一个专注于语义信息的注意力模块，即跨层注意力模块。

跨层注意力模块输入两个不同层的特征图，对高层粗粒度特征图依次进行上采样与卷积等操作，得到与低层细粒度特征图尺寸相匹配的$\boldsymbol{x}_{3_\mathrm{up}}$，如图5(b)所示。

$$\boldsymbol{x}_{3\_\mathrm{up}}=\mathrm{ReLU}\left\{\mathrm{L2N}\left\{\mathrm{Conv}\left[\mathrm{Upsamp}(\boldsymbol{x}_3)\right]\right\}\right\},\quad(7)$$

式中：$\boldsymbol{x}_{3\_\mathrm{up}}\in(960,8,8)$。接着沿$W$和$H$方向对$\boldsymbol{x}_{3_\mathrm{up}}$分别进行一次自适应最大池化与自适应平均池化来对特征图进行压缩，将其分别输入一个权重共享的全连接层后对二者进行相加，最后将结果输入映射函数，得到最终的通道注意力，将通道注意力乘以$\boldsymbol{x}_2$，来为$\boldsymbol{x}_2$的各个通道赋予不同的权重，生成$\boldsymbol{x}_{2_\mathrm{final}}$：

$$\boldsymbol{x}_{2\_\mathrm{final}}=\boldsymbol{x}_2\times\left\{\mathrm{Sigmoid}\left\{\mathrm{MLP}\left[\mathrm{Maxpool}\left(\boldsymbol{x}_{3\_\mathrm{up}}\right)\right]+\mathrm{MLP}\left[\mathrm{Avgpool}\left(\boldsymbol{x}_{3\_\mathrm{up}}\right)\right]\right\}\right\},\quad(8)$$

式中：$\boldsymbol{x}_{2\_\mathrm{final}}\in(960,8,8)$。再对$\boldsymbol{x}_1$与$\boldsymbol{x}_{2_\mathrm{final}}$进行相同操作来生成$\boldsymbol{x}_{1_\mathrm{final}}$：

$$\boldsymbol{x}_{2\_\mathrm{up}}=\mathrm{ReLU}\left\{\mathrm{L2N}\left\{\mathrm{Conv}\left[\mathrm{Upsamp}\left(\boldsymbol{x}_{2\_\mathrm{final}}\right)\right]\right\}\right\},\quad(9)$$

$$\boldsymbol{x}_{1\_\mathrm{final}}=\boldsymbol{x}_1\times\left\{\mathrm{Sigmoid}\left\{\mathrm{MLP}\left[\mathrm{Maxpool}\left(\boldsymbol{x}_{2\_\mathrm{up}}\right)\right]+\mathrm{MLP}\left[\mathrm{Avgpool}\left(\boldsymbol{x}_{2\_\mathrm{up}}\right)\right]\right\}\right\},\quad(10)$$

式中：$\boldsymbol{x}_{2\_\mathrm{up}}\in(640,16,16)$，$\boldsymbol{x}_{1\_\mathrm{final}}\in(640,16,16)$。

在底层特征$\boldsymbol{x}_0$中，再次在$\boldsymbol{x}_0$的通道方向上进行一次空间注意力的计算，以再次强化小目标的位置信息，得到特征图$\boldsymbol{x}_{0_\mathrm{satt}}$，如图5(c)所示，然后对其与计算$\boldsymbol{x}_{1_\mathrm{final}}$通道注意力后得到的特征图$\boldsymbol{x}_{1_\mathrm{catt}}$进行Cat操作后，再对通道进行压缩融合，得到最终的$\boldsymbol{x}_{0_\mathrm{final}}\in(320,32,32)$：

$$\boldsymbol{x}_{1\_\mathrm{up}}=\mathrm{ReLU}\left\{\mathrm{L2N}\left\{\mathrm{Conv}\left[\mathrm{Upsamp}\left(\boldsymbol{x}_{1\_\mathrm{final}}\right)\right]\right\}\right\},\quad(11)$$

$$\boldsymbol{x}_{1\_\mathrm{catt}}=\boldsymbol{x}_{1\_\mathrm{up}}\times\left\{\mathrm{Sigmoid}\left\{\mathrm{MLP}\left[\mathrm{Maxpool}\left(\boldsymbol{x}_{1\_\mathrm{up}}\right)\right]+\mathrm{MLP}\left[\mathrm{Avgpool}\left(\boldsymbol{x}_{1\_\mathrm{up}}\right)\right]\right\}\right\},\quad(12)$$

$$\boldsymbol{x}_{0\_\mathrm{satt}}=\boldsymbol{x}_0\times\left\{\mathrm{Conv}\left\{\mathrm{Cat}\left[\mathrm{Max}(\boldsymbol{x}_0),\mathrm{Mean}(\boldsymbol{x}_0)\right]\right\}\right\},\quad(13)$$

$$\boldsymbol{x}_{0\_\mathrm{final}}=\mathrm{Conv}\left[\mathrm{Cat}\left(\boldsymbol{x}_{1\_\mathrm{catt}},\boldsymbol{x}_{0\_\mathrm{satt}}\right)\right],\quad(14)$$

式中：$\boldsymbol{x}_{1\_\mathrm{up}}\in(320,32,32)$，$\boldsymbol{x}_{1\_\mathrm{catt}}\in(320,32,32)$，$\boldsymbol{x}_{0\_\mathrm{satt}}\in(320,32,32)$。至此，将得到的$(\boldsymbol{x}_{0_\mathrm{final}},\boldsymbol{x}_{1_\mathrm{final}},\boldsymbol{x}_{2_\mathrm{final}},\boldsymbol{x}_3)$分别送入4个检测头中进行检测。

3.4　损失函数

在模型中利用焦点损失函数[4]与CIOU Loss[16]来减小模型中的置信度损失、分类损失与回归损失，进而达到优化模型的效果。总损失的公式为

$$L_{\mathrm{total}}=\lambda L_{\mathrm{obj}}+\eta L_{\mathrm{class}}+\theta L_{\mathrm{box}},\quad(15)$$

式中：λ、η与θ分别为置信度损失、分类损失与回归损

失的超参数，遵循 YOLOv5，分别将其设置为 0.7、0.3 与 0.05。由于 DIOR 数据集[14]中存在着类别不平衡问题，因此采用焦点损失函数[4]作为置信度与分类的损失函数，可表示为

$$L_{\text{obj, class}} = -\alpha\left(1-p_{\text{t}}\right)^{\gamma}\log p_{\text{t}}, \tag{16}$$

式中：p_{t} 表示输出为每一类的概率值；α 为每一类别样本的权重值，γ 为衰减参数，通过对所提整体模型进行参数分析，最终分别将其设置为 0.25 与 1.5。消融实验的 baseline 的实验结果也是在相同参数设置下得出的，因此在对比过程中具有普适性。

由于 CIOU Loss[16]在回归损失的计算中考虑到了预测框与标签框的宽高比、交叉面积与中心点的距离，可以提高训练过程中的稳定性与收敛速度，因此选用 CIOU Loss[16]作为回归损失函数，可表示为

$$L_{\text{box}} = 1-\left[\text{IOU}(A,B)-\frac{\rho^2}{c^2}-\varepsilon v\right], \tag{17}$$

$$v = \frac{4}{\pi^2}\left(\arctan\frac{w_{\text{l}}}{h_{\text{l}}}-\arctan\frac{w_{\text{p}}}{h_{\text{p}}}\right)^2, \tag{18}$$

$$\varepsilon = \frac{v}{1-\text{IOU}(A,B)+v}, \tag{19}$$

式中：IOU(A,B)为预测框与标签框的交并比；ρ 为中心点距离；ε 为影响因子；v 为宽高比的相似度；c 为包含预测框与标签框的矩形的对角线长度；w_{p} 和 h_{p} 为预测框的宽和高，w_{l} 和 h_{l} 为标签框的宽和高。

4　实验与分析

4.1　数据集及评估指标

基于深度学习的目标检测所使用的数据集已经被研究者们相继提出。但大部分数据集中的大、中目标的数量远远高于小目标的数量，而且还存在着小目标标注质量不高的问题。为了促进小目标检测技术的发展，有很多专门用来针对小目标的数据集也被提出与发布，比如 COCO 数据集[6]、Tiny Person 数据集[17]、DOTA 数据集[18]、NWPU VHR-10 数据集[19]及 DIOR 数据集[14]等这些较常使用的数据集。出于类别丰富度、标注质量及小目标数量等方面的考量，选用 DIOR 数据集[14]作为基础数据集来训练网络模型。该数据集共计 23463 张图像，其中 18770 张图片用于训练，4693 张图片用于测试，每张图片的尺寸大小为 800×800，包含 ship、vehicle 等 20 个类别，并含有大量优质的小目标标注。如图 6 所示，该数据集中标签框的尺寸大部分都小于图像的 10% 且都集中分布在图像的中心位置，可证明该数据集的可靠性。

在目标检测中通常用平均精度均值（mAP）来衡量检测器整体性能的好坏。COCO 数据集[6]的评价指标中的 average precision of small（APs）也是衡量小目标检测器精度的一项重要指标。average precision of medium（APm）、average precision of large（APl）则分别为中、大尺度目标检测精度。

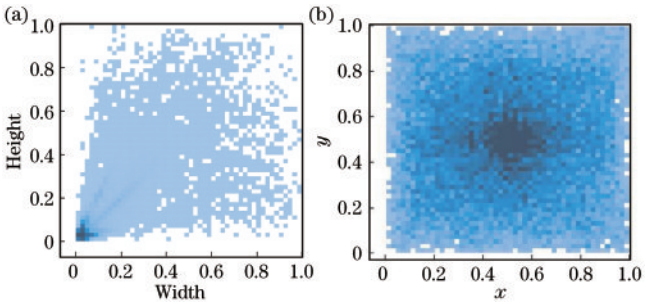

图 6　数据集中标签框分布统计。(a) 标签框尺寸分布；(b)标签框中心点分布

Fig. 6　Distribution statistics of label boxes in the data set. (a) Distribution of the size of the label box; (b) distribution of the center points of the label box

4.2　参数设置

在训练阶段，为了减少对单张图片的计算量，将图像尺寸大小重新设置为 640×640。遵循 YOLOv5 的设置，通过随机裁剪、随机缩放、随机排布的方式来实现数据增强。实验中，batchsize 设置为 12，整个网络采用 SGD 优化器，权重衰减设为 0.0005，学习率为 0.015。总共训练 300 代，在 0～3 代通过 warm-up 策略线性调节学习率，随后学习率逐渐线性递减。所有实验均是在单张 3090 GPU 上，以 YOLOv5 为 baseline，基于 PyTorch 框架实现的。

4.3　对比实验

在 DIOR 数据集[14]上对所提方法与近年的一些经典目标检测算法进行比较，主要包括 SSD[3]、YOLOv3[15]、Faster R-CNN[1]、Mask R-CNN[20]、Libra R-CNN[11]、YOLOx[21]、Dynamic R-CNN[22]、YOLOv5。对比过程中，将图像尺寸大小统一定义为 640×640，并在同一张 3909 GPU 显卡上，使用 CUDA 11.0、CUDNN 11.3、Python 3.7 与 PyTorch 1.7 进行测试。测试结果如表 1 所示。所提方法具有最优的 APs；mAP 仅次于 YOLOv5x6，相差 0.2 个百分点，APs 则较之高出 3.5 个百分点。由此可看出所提方法在小目标检测上具有优越性。

4.4　消融实验

所提小目标检测方法主要包含 ResCatPAN 结构（RCP）与跨层注意力模块（CLAT）两个部分。为了证明这两个部分的有效性，分别对这两个部分进行消融实验分析，实验结果如表 2 所示。

为了证明 ResCatPAN 结构的有效性，在 baseline 结构中只添加了该模块进行实验，并用 CIOU Loss[16]和焦点损失函数[4]对模型进行优化。如表 2 所示，该模块可以有效地提高对小目标的检测精度。与基准方法相比，ResCatPAN 结构的 APs 在 DIOR 数据集[14]上比 baseline 高出 4.1 个百分点，mAP 则降低了 0.1 个百分

表 1 对比实验

Table 1 Contrast experiment

Method	Backbone	APs /%	mAP /%
SSD	VGG16		58.6
YOLOv3	Darknet-53	11.6	57.1
Faster R-CNN with FPN	ResNet-101		65.1
Mask R-CNN with FPN	ResNet-101		65.2
Libra R-CNN	ResNet-101	14.9	79.7
Dynamic R-CNN	ResNet50	12.1	77.3
YOLOx	Darknet-53	17.3	85.7
YOLOv5x6	CSPDark-53	19.9	86.6
Proposed method	CSPDark-53	23.4	86.4

表 2 消融实验

Table 2 Ablation experiment

Baseline	RCP	CLAT	mAP /%	APs /%	APm /%	APl /%
√			85.5	17.5	52.0	75.6
√	√		85.4	21.6	51.7	75.4
√		√	86.2	19.2	52.2	76.1
√	√	√	86.4	23.4	52.3	75.9

点，APm 与 APl 也分别降低了 0.3 个百分点与 0.2 个百分点。这意味着，使用更底层特征图信息和增加小目标检测头的 ResCatPAN 结构虽然在整体性能上稍微低于基准方法，但该结构能更好地提升 YOLOv5 的小目标检测性能，证明了 ResCatPAN 结构的有效性。

为了证明跨层注意力模块的有效性，在 baseline 结构中只添加了该模块，在三层金字塔特征图与三个检测头上进行实验，并用 CIOU Loss[16]和焦点损失函数[4]对模型进行优化。如表 2 所示，跨层注意力模块的 APs 在 DIOR 数据集[14]上比 baseline 高出 1.7 个百分点，mAP 则提升了 0.7 个百分点，APm 与 APl 也分别提升了 0.2 个百分点与 0.5 个百分点。这表明了利用该模块在各层不同尺寸的特征图之间计算注意力来进行语义信息传递可以有效地提升对小尺度目标的检测性能，并在一定程度上提升对大、中尺度目标的检测性能，进而提升整体性能，证明了跨层注意力模块的不可缺失性。

为了证明整体结构的有效性，将 ResCatPAN 结构与跨层注意力模块共同加入到 baseline 中进行实验。如表 2 所示，在两个结构的共同作用下，mAP 在 DIOR 数据集[14]上比 baseline 高出 0.9 个百分点，APs 则高出 5.9 个百分点，APm 与 APl 也较基准方法高出 0.3 个百分点。这意味着所提出的两个结构可以在增强小目标位置信息与抑制背景信息两方面提升小目标检测性能，且跨层注意力模块可以在一定程度上弥补 ResCatPAN 结构对大、中目标检测精度的负面影响，证明了两结构的互补性与不可分割性。

4.5 参数分析

在开始实验前，首先对 DIOR 数据集[14]训练集中的数据分布进行了查看与统计。发现该数据集存在着较严重的样本分布不均匀问题，如表 3 所示，这意味着使用该数据集训练的模型容易对样本数目少的类别不敏感而损害该类别的检测精度。

由于该数据集中的 ship 与 vehicle 等小目标数量与 golf-field 等大目标样本数量不平衡，容易造成 golf-field 等类别被网络损失函数忽略，进而对大尺度目标的检测精度造成一定影响。为了缓解该问题，在所提模型中使用了焦点损失函数[4]来对该模型进行优化并对焦点损失函数中的参数 γ 进行了实验分析，实验结果如图 7 所示。

表 3 DIOR 数据集样本分布

Table 3 Sample distribution of the DIOR dataset

Class	Number of samples	Class	Number of samples	Class	Number of samples	Class	Number of samples
golf-field	881	dam	824	stadium	1003	airport	1071
train-station	811	chimney	1340	harbor	4447	bridge	3161
expressway-service-area	1743	basketball-court	2658	overpass	2478	airplane	8100
expressway-toll-station	1028	ground-track-field	2390	windmill	4371	vehicle	32180
baseball-field	4674	tennis-court	9621	storage-tank	20717	ship	50207

参数 α 解决了训练样本中正负样本不平衡的问题。通过调节该参数，可以降低正样本或者负样本的权重，在实验中，使用默认值 0.25 来训练模型。

参数 γ 解决了训练样本中难易样本不平衡的问题。通过调节该参数，可以给一个分类正确的样本一个较低的损失权重，给一个分类错误的样本一个较高的损失权重，进而让模型更重视难样本，忽视易样本。由图 7 可知：当 γ=1.3 时，中尺度目标检测精度可达 53.4%，但小尺度目标检测精度与总体检测精度却有着一定的下降且大尺度目标检测精度未达到最高；当 γ=1.7 时，大尺度目标检测精度可达 76.6%，但小尺度目标检测精度与中尺度检测精度却有着一定的下降；当 γ=1.5 时，小尺度目标检测精度可达最高。经综合考虑，将 γ 设为可使小尺度目标检测精度最高的值，即 1.5，此时可使得 APs 取得最高值 23.4%，mAP 可达 86.4%。

4.6 可视化与时间复杂度分析

采用 Grad-CAM 方法生成热力图，来直观展示 CLAT 模块抑制背景信息的效果，具体效果如图 8 所

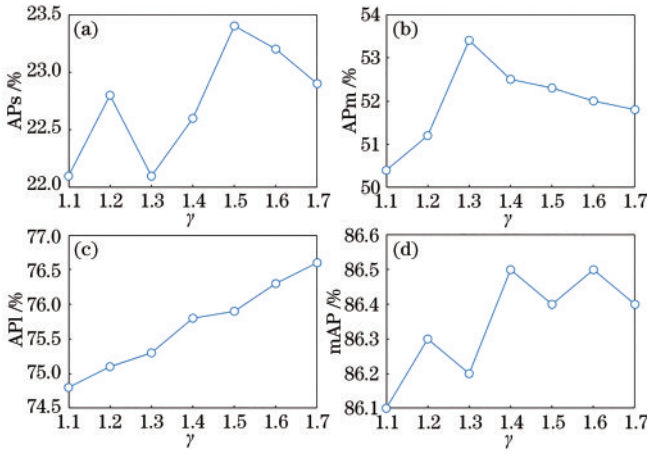

图 7　超参数 γ 的有效性分析。(a)超参数 γ 对小尺度目标检测性能的影响;(b)超参数 γ 对中尺度目标检测性能的影响;(c)超参数 γ 对大尺度目标检测性能的影响;(d)超参数 γ 对整体目标检测性能的影响

Fig. 7　Effect analysis on hyperparameter γ. (a) Influence of hyperparameter γ on detection performance for small object; (b) influence of hyperparameter γ on detection performance for medium object; (c) influence of hyperparameter γ on detection performance for large object; (d) influence of hyperparameter γ on detection performance for object

示。通过该处理效果图可看出,CLAT 模块可通过增强特征图语义信息的方式使模型有效地将检测重点集中在目标区域,进而达到抑制背景信息的效果。

图 8　CLAT 模块通过 Grad-CAM 方法生成的热力图

Fig. 8　Heat map generated by the proposed CLAT module by using the Grad-CAM method

对所提模型与 baseline 进行可视化对比,对比效果如图 9 所示,所提模型可改善 baseline 中存在的误检、漏检问题。

所提模型对单张图片的检测时间较 baseline 稍有延长,如表 4 所示,baseline 对每张图片的平均处理时间为 0.06 s,而所提模型则仅高出 0.03 s。

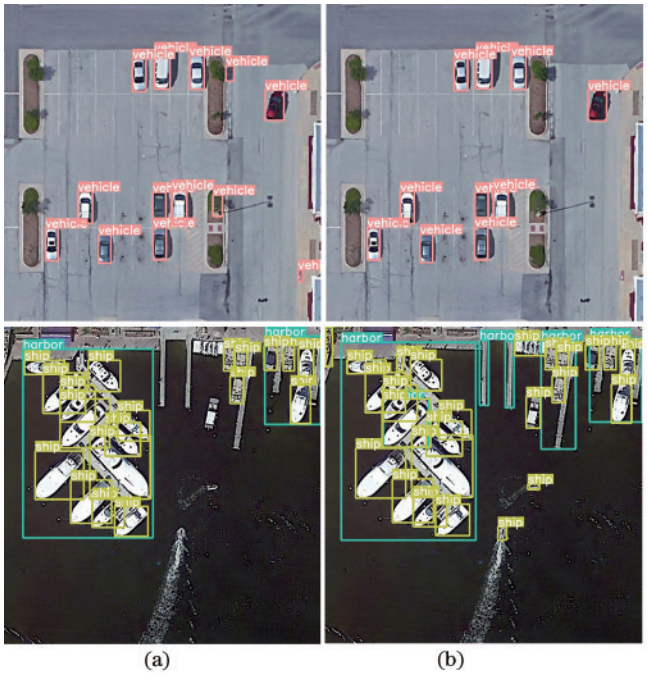

图 9　所提模型与 baseline 在测试集上的效果对比。(a) baseline;(b) 所提模型

Fig. 9　Effect contrast of the proposed model and baseline on the test set. (a) baseline; (b) proposed model

表 4　耗时比较

Table 4　Time-consuming comparison

Model	Backbone	Time /s
baseline	CSPDark-53	0.06
Proposed model	CSPDark-53	0.09

5　结　　论

提出了一种新的基于 YOLOv5 的小目标检测方法。该方法主要由 ResCatPAN 结构和 CLAT 模块两部分组成。ResCatPAN 结构通过采用将更底层的特征图的位置信息传递给高层特征图并增加检测头的方式来提升对遥感图像中的小目标检测精度。CLAT 模块通过在各层不同尺寸的特征图之间计算注意力的方式来进行由高到低的语义信息传递,可以有效抑制图像中的复杂背景信息,进而在提升对小目标的检测性能的同时在一定程度上提升对大、中尺度目标的检测性能。在 DIOR 公共基准数据集[14]上的实验验证了所提方法在小目标检测任务上的有效性和相对于其他优秀算法的优越性。尽管所提方法有较好的性能,但检测耗时却有所增加。为此,后续工作是将在减少计算量方向继续研究来提高检测速度。消融实验部分可充分证明:合理利用高层语义信息与底层位置信息、消除复杂背景信息对遥感小目标的不良影响仍是提高遥感

目标检测精度的有效手段。

参 考 文 献

[1] 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.

[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, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 779-788.

[3] Liu W, Anguelov D, Erhan D, et al. SSD: single shot MultiBox detector[M]//Leibe B, Matas J, Sebe N, et al. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9905: 21-37.

[4] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//2017 IEEE International Conference on Computer Vision, October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 2999-3007.

[5] Zhang H Y, Wang Y, Dayoub F, et al. VarifocalNet: an IoU-aware dense object detector[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 8510-8519.

[6] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: common objects in context[M]//Fleet D, Pajdla T, Schiele B, et al. Computer vision-ECCV 2014. Lecture notes in computer science. Cham: Springer, 2014, 8693: 740-755.

[7] 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, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 936-944.

[8] 刘峰, 郭猛, 王向军. 基于跨尺度融合的卷积神经网络小目标检测[J]. 激光与光电子学进展, 2021, 58(6): 0610012.
Liu F, Guo M, Wang X J. Small target detection based on cross-scale fusion convolution neural network[J]. Laser & Optoelectronics Progress, 2021, 58(6): 0610012.

[9] 刘鑫, 陈思溢, 陈小龙, 等. 基于深度学习的深层次多尺度特征融合目标检测算法[J]. 激光与光电子学进展, 2021, 58(12): 1210029.
Liu X, Chen S Y, Chen X L, et al. Deep multi-scale feature fusion target detection algorithm based on deep learning[J]. Laser & Optoelectronics Progress, 2021, 58(12): 1210029.

[10] Liu S, Qi L, Qin H F, et al. Path aggregation network for instance segmentation[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 8759-8768.

[11] Pang J M, Chen K, Shi J P, et al. Libra R-CNN: towards balanced learning for object detection[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 821-830.

[12] Lim J S, Astrid M, Yoon H J, et al. Small object detection using context and attention[C]//2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), April 13-16, 2021, Jeju Island, Korea (South). New York: IEEE Press, 2021: 181-186.

[13] 汪亚妮, 汪西莉. 基于注意力和特征融合的遥感图像目标检测模型[J]. 激光与光电子学进展, 2021, 58(2): 0228003.
Wang Y N, Wang X L. Remote sensing image target detection model based on attention and feature fusion[J]. Laser & Optoelectronics Progress, 2021, 58(2): 0228003.

[14] Li K, Wan G, Cheng G, et al. Object detection in optical remote sensing images: a survey and a new benchmark[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2020, 159: 296-307.

[15] Redmon J, Farhadi A. YOLOv3: an incremental improvement[EB/OL]. (2018-04-08) [2022-05-30]. https://arxiv.org/abs/1804.02767.

[16] 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.

[17] Yu X H, Gong Y Q, Jiang N, et al. Scale match for tiny person detection[C]//2020 IEEE Winter Conference on Applications of Computer Vision, March 1-5, 2020, Snowmass, CO, USA. New York: IEEE Press, 2020: 1246-1254.

[18] Xia G S, Bai X, Ding J, et al. DOTA: a large-scale dataset for object detection in aerial images[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 3974-3983.

[19] Cheng G, Han J W, Zhou P C, et al. Multi-class geospatial object detection and geographic image classification based on the collection of part detectors[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2014, 98: 119-132.

[20] He K M, Gkioxari G, Dollár P, et al. Mask R-CNN [C]//2017 IEEE International Conference on Computer Vision, October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 2980-2988.

[21] Ge Z, Liu S T, Wang F, et al. YOLOX: exceeding YOLO series in 2021[EB/OL]. (2021-08-06) [2022-05-30]. https://arxiv.org/abs/2107.08430.

[22] Zhang H K, Chang H, Ma B P, et al. Dynamic R-CNN: towards high-quality object detection via dynamic training [M]//Vedaldi A, Bischof H, Brox T, et al. Computer vision-ECCV 2020. Lecture notes in computer science. Cham: Springer, 2020, 12360: 260-275.