第 58 卷 第 6 期/2021 年 3 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

基于跨尺度融合的卷积神经网络小目标检测

刘峰[1,2**]，郭猛[1,2*]，王向军[1,2]

[1] 天津大学精密测试技术及仪器国家重点实验室，天津 300072；

[2] 天津大学微光机电系统技术教育部重点实验室，天津 300072

摘要 针对小目标(像素占比小于 0.02)检测存在的目标特征容易丢失、分辨率低的问题，提出了一种基于改进 YOLOv3(You only look once)卷积神经网络的检测方法。首先，对数据集中的小目标进行复制变换增强，以提升训练过程中网络对小目标的注意力。其次，针对浅层视觉信息与深层语义信息的尺度融合，提出了跨尺度检测层的网络结构，提高了网络对小目标的适应能力。最后，针对高分辨率图像的检测效果，提出了深度和广度结合的残差块组传递结构，丰富了深层特征图的感受野。实验结果表明，相比 YOLOv3 网络，改进跨级尺度预测层的网络检测小目标的精确率提升了 1.9 个百分点，召回率提升了 5.9 个百分点；优化感受野的网络检测小目标的精确率提升了 31.6 个百分点，召回率提升了 46.4 个百分点。

关键词 图像处理；卷积网络；小目标；尺度融合；高分辨率

中图分类号 TP18 **文献标志码** A **doi**：10.3788/LOP202158.0610012

Small Target Detection Based on Cross-Scale Fusion Convolution Neural Network

Liu Feng[1,2**], Guo Meng[1,2*], Wang Xiangjun[1,2]

[1] *State Key Laboratory Precision Measuring Technology and Instruments, Tianjin University, Tianjin* 300072, *China*;

[2] *Micro Optics Electronic Machine System Education Ministry Key Laboratory, Tianjin University, Tianjin* 300072, *China*

Abstract Aiming at the problem of small target (pixel ratio less than 0.02) detection that the target features are easily lost and the resolution is low, a detection method based on improved YOLOv3 (You only look once) convolutional neural network is proposed in this paper. First, the small targets in the data set are copied and transformed to enhance the network′s attention to the small targets during the training process. Second, for the scale fusion of shallow visual information and deep semantic information, a cross-scale detection layer network structure is proposed, which improves the network′s adaptability to small targets. Finally, for the detection effect of high-resolution images, a residual block transfer structure combining depth and breadth is proposed, which enriches the receptive field of deep feature maps. Experimental results show that compared with the YOLOv3 network, the precision rate of the network detection of small targets with the improved cross-scale prediction layer increased by 1.9 percentage points, and the recall rate increased by 5.9 percentage points. The precision rate of the network detection of small targets with the optimized receptive fields increased 31.6 percentage points, the recall rate increased by 46.4 percentage points.

Key words image processing; convolutional network; small target; scale fusion; high resolution

OCIS codes 100.2000; 100.4996; 120.1880

收稿日期：2020-07-17；**修回日期**：2020-08-15；**录用日期**：2020-08-31

基金项目：国家自然科学基金(51575388)

* **E-mail**：mengg@tju.edu.cn；** **E-mail**：tjuliufeng@tju.edu.cn

1 引 言

传统目标检测分为区域选择(滑窗)、特征提取(如尺度不变特征变换 SIFT、方向梯度直方图 HOG)和分类器分类(如支持向量机 SVM、Adaboost)三部分。基于滑动窗口的区域选择策略没有针对性,时间复杂度高,窗口冗余,而手工设计的特征对于多样化的目标鲁棒性较差。随着深度学习的不断成熟和发展,其在目标检测上的精度和可靠性已经远优于传统检测技术,成为计算机视觉领域的研究热点。检测的最终目的是从图像中识别目标并提取出目标的空间位置信息[1],已在行人分析、智能机器人导航[2]、辅助自动驾驶等领域[3]得到了广泛应用。目前主流的目标检测方法有 one-stage 和 two-stage 两种[4],two-stage 是候选窗加深度学习分类检测方法,如区域卷积神经网络(RCNN)[5-7]用生成的候选锚框提取特征信息,该网络检测速度较慢。one-stage 是端到端的回归目标检测网络,如 SSD(Single shot MultiBox detector)[8]、YOLO(You only look once)网络[9-12]。用 one-stage 替代 two-stage 的特征提取过程,可避免候选锚框的生成计算,从而提高检测速度,但检测精度较低。

近年来目标检测进入一个快速发展的阶段,随着检测条件的变化,图像视场中会出现小尺度目标,给检测算法的精确性、可靠性带来了挑战。Bell 等[13]提出将小目标定义为 COCO 数据库中尺寸小于等于 32 pixel×32 pixel 的目标;Mäenpää 等[14]针对像素尺寸小、分辨率低的小目标,通过分析局部二值模式(LBP)算子并结合其他纹理分析方法,发布了一个小目标检测专用数据库,并将尺寸为 512 pixel×512 pixel 的图像中尺寸约为 20 pixel×20 pixel 的目标定义为小目标。小目标检测的核心难点是分辨率低、细节特征少,进行图像特征语义信息提取的过程中目标特征细节容易丢失,导致检测识别率不高或误警率高。

为了提高小目标的检测精度,He 等[15]提出了用于视觉识别的空间金字塔池,采用多尺度 pooling 窗口,进一步强调了卷积神经网络(CNN)特征计算前移、区域处理后移的思想,完成单次特征图的计算后,再进行区域窗口信息的结合,在每个特征层进行预测,并利用多尺度特征信息,提高了多尺度目标的检测精度,但检测速度较慢。王海涛等[16]提出了基于两级上下文卷积网络宽视场图像的小目标检测方法,采用两个 Faster-RCNN 分别学习和建模小目标及其上下文背景特征,提高了小目标特征图的分辨率。周苏等[17]针对 PVANet(Performance vs accuracy network)存在的小目标检测能力不足问题,将浅层 7×7 卷积核拆解成 3 层 3×3 卷积核,将深层 3×3 卷积核进一步非对称地分解成两个 1×3 和 3×1 卷积核,以增加浅层卷积滤波器的细粒度及小目标图像特征的提取能力和非线性表达能力,提升了网络对小目标的检测能力。Ju 等[18]提出了一种基于交并比(IOU)的数学推导方法,针对 YOLOv3 网络各候选锚框的数量和纵横比尺寸进行了优化,同时将网络前 6 个卷积层转换为 2 个残差单元,以避免梯度衰落并增强特征的重用性,提升了网络检测小目标的能力。王俊强等[19]参考密集连接网络(DenseNet)设计特征提取网络,并用其替代原有的 16 层 VGG-Net(Visual geometry group-network),通过引入特征金字塔网络(FPN)进行多尺度特征融合,实现了对小目标的检测。刘力荣等[20]采用 Slim-Net 实现了全景图像上小目标的检测,通过建立点云深度图像与全景图像的映射关系,进行小目标的地理定位和矢量提取,给出了城市小目标的精确三维空间地理定位,为小目标的检测定位提供了新思路。

针对目标像素数在图像总像素数中占比小于 0.02 的小目标检测识别任务,综合考虑了速度和准确率因素,本文提出了一种基于 YOLOv3 网络的跨尺度融合检测方法。减少了特征提取过程中小目标信息的丢失问题,并通过传递模块结合更多的细粒度信息和语义信息,提高了网络对小目标的检测能力。同时提出了结合深度和广度的残差块组传递结构,丰富了深层特征图的感受野,进一步提高了网络在高分辨率图像下的小目标识别效果。

2 YOLOv3 网络的原理

2.1 特征提取网络 Darknet-53

Darknet-53 将 1 个卷积层、1 个正则化层和 1 个激活层封装成 1 个模块,在 Darknet-19 的基础上引入残差模块,进一步加深了网络结构,该结构共由 53 个卷积层组合而成,使用了一系列卷积核为 3×3、1×1 的卷积层。为了降低池化带来的梯度消失问题,直接摒弃了 pooling 操作,用步长为 2 的卷积核实现降采样。输入图像经过 5 次降采样后,用全连接层进行分类预测,其网络结构如图 1 所示,其中,res 为残差结构,DBL 为 YOLOv3 网络的基本组件,即卷积+批标准化(BN)+激活函数(Leaky

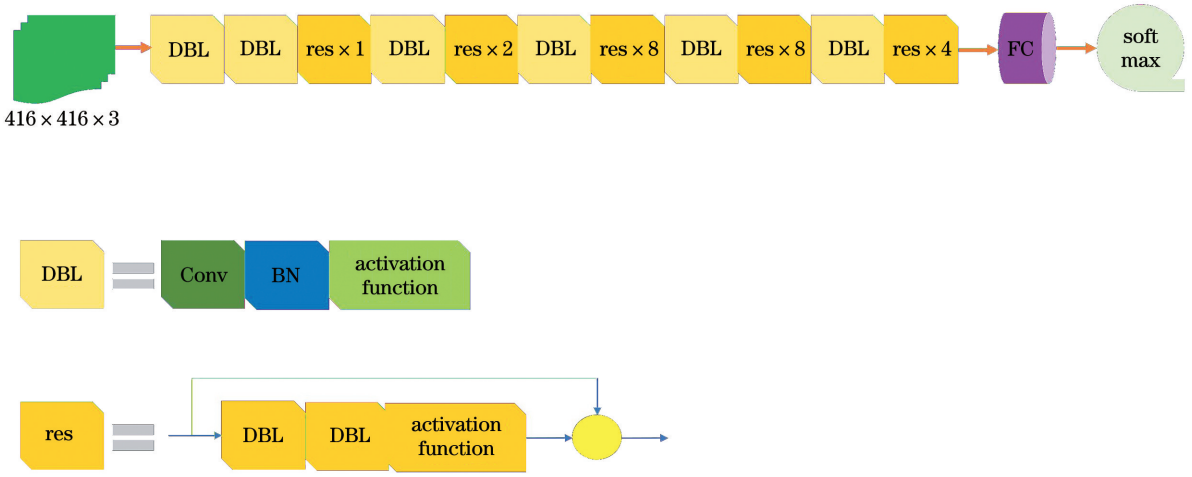

图 1　Darknet-53 的结构

Fig. 1　Structure of the Darknet-53

ReLU)，Conv 为卷积操作，FC 为全连接层。Darknet-53 架构在效果更好的前提下，速度是 ResNet-101 的 1.5 倍；在与 ResNet-152 效果相近的情况下，速度是其 2 倍，几种常见网络的性能如表1所示，其中，FPS为每秒处理图像的帧数，Top-1 与 Top-5 分别为排名第一与前五的类别与实际结果相符的准确率。

表 1　不同骨干网络的性能

Table 1　Performance of different backbone networks

Backbone	Top-1/%	Top-5/%	FPS /frame
Darknet-19	74.1	91.8	171
ResNet-101	77.1	93.7	53
ResNet-152	77.6	93.8	37
Darknet-53	77.2	93.8	78

2.2　YOLOv3 网络结构

YOLOv3 网络使用 Darknet-53 的前 52 层，用单独的逻辑分类器取代了最后的全连接层，是一个全卷积网络。通过特征图的上采样和信息融合在三个尺度特征图上检测目标，用二元交叉熵损失计算类别预测损失，用误差平方和计算位置和尺寸损失，并在计算检测框损失时用系数增加小尺寸目标的学习效果，其中 w_i 和 h_i 分别为第 i 个真实框的宽和高。相比 YOLOv2 网络，YOLOv3 网络提升了对小目标的检测性能。其损失函数可表示为

$$
\begin{aligned}
X_{\text{loss}}(o)=&l_{\text{coord}}\sum_{i=0}^{K\times K}\sum_{j=0}^{M}I_{ij}^{\text{obj}}[(x_i-\hat{x}_i)^2+(y_i-\hat{y}_i)^2]+l_{\text{coord}}\sum_{i=0}^{K\times K}\sum_{j=0}^{M}I_{ij}^{\text{obj}}(2-w_i\times h_i)[(w_i-\hat{w}_i)^2+\\
&(h_i-\hat{h}_i)^2]-\sum_{i=0}^{K\times K}\sum_{j=0}^{M}I_{ij}^{\text{obj}}[\hat{C}_i\log(C_i)+(1-\hat{C}_i)\log(1-C_i)]-l_{\text{noobj}}\sum_{i=0}^{K\times K}\sum_{j=0}^{M}I_{ij}^{\text{noobj}}[\hat{C}_i\log(C_i)+\\
&(1-\hat{C}_i)\log(1-C_i)]-\sum_{i=0}^{K\times K}I_{ij}^{\text{obj}}\sum_{c\in X_{\text{classes}}}\{\hat{p}_i(c)\log[p_i(c)]+[1-\hat{p}_i(c)]\log[1-p_i(c)]\},
\end{aligned}
\tag{1}
$$

式中，l_{coord} 和 l_{noobj} 为设定的有目标框与无目标框对损失函数的权重比例，输入图像经卷积网络后被分成 $K\times K$ 个网格，M 为每个网格产生的候选框数，I_{ij}^{obj} 表示第 i 个网格的第 j 个 anchor box 是否负责该 object(o)，负责为 1，不负责为 0；I_{ij}^{noobj} 表示第 i 个网格的第 j 个 anchor box 是否不负责该 object，x_i、y_i 分别为实际的物体中心坐标，$\hat{x}_i$、$\hat{y}_i$ 为预测的物体中心坐标，$\hat{w}_i$、$\hat{h}_i$ 为预测的 box 尺寸，$\hat{C}_i$ 表示 bounding box 是否负责预测目标，$\hat{p}_i(c)$为实际目标属于类别 c 的概率，$p_i(c)$为预测目标属于类别 c 的概率，X_{classes} 为所有类别的集合。

3　数据预处理

3.1　数据来源与扩充

实验使用的训练集是从 DOTA 数据集提取的 Plane、Small-vehicle 和 Large-vehicle 三类。由于 DOTA 数据集的图像分辨率最大为 12029 pixel×5014 pixel，直接将原始图像送入 YOLO 网络会损失过多的图像信息，训练效果不佳。因此，将 DOTA 数据集有检测目标的区域统一裁剪成尺寸

为 1000 pixel×1000 pixel 后作为新的数据集，如图 2 所示。本数据集共包含 4360 张图像，相比 PASCAL VOC、COCO 数据集，小目标样本更多。

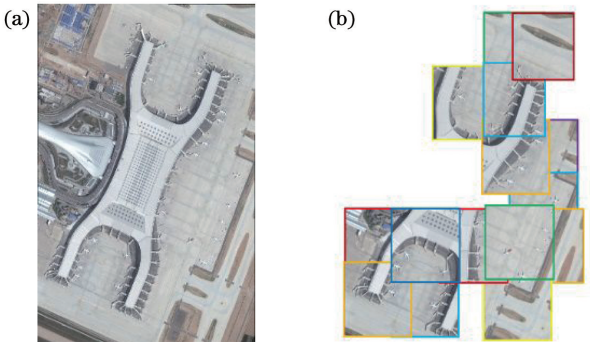

图 2　DOTA 数据集。(a)原始图像；(b)裁剪后的图像

Fig. 2　DOTA data set. (a) Original image; (b) cropped image

针对数据集中小目标比例少，对小目标训练不足导致的检测效果差问题，对原数据集中的较小目标进行了复制扩充，以增加小目标样本在网络训练过程中对参数修正的影响，结果如图 3 所示。

3.2　重新聚类 anchor

不同数据集标注样本的尺度分布不同，实验以小目标为主，与生成 YOLOv3 网络中先验框的 COCO 数据集差别较大。因此，仅通过 K-means 聚类算法对处理后的数据重新进行聚类，得到检测层的先验框尺寸。为了丰富检测尺度，将聚类中心点由 9 个增至 12 个，每个检测层分配 4 个先验框，聚类后的先验框像素尺寸分别为 9×11，22×12，13×22，24×18，18×24，25×25，18×46，51×22，39×36，30×56，51×51，86×81。

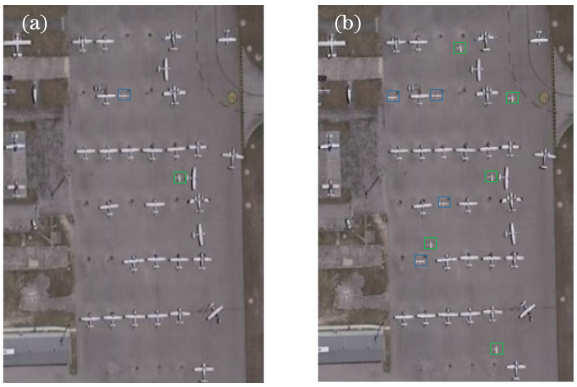

图 3　DOTA 数据集的增强。(a)原始图像；(b)增强后的图像

Fig. 3　Enhancement of the DOTA data set. (a) Original image; (b) enhanced image

4　网络结构优化

4.1　网络检测尺度的优化

加深网络后，深层特征图拥有更高的语义特征，但随着降采样深度的增加，目标部分的纹理细节会出现丢失现象，不利于小尺寸目标的检测。而浅层特征图包含更多的细粒度信息，有利于小目标的定位识别。因此，用直通模块将浅层特征图与上采样后的深层特征图进行融合。YOLOv3 网络在尺寸为 13 pixel×13 pixel、26 pixel×26 pixel、52 pixel×52 pixel 的特征图上进行目标预测，实验将输入图像的尺寸统一调整为 576 pixel×576 pixel，在更浅层特征图(尺寸为 36 pixel×36 pixel、72 pixel×72 pixel、144 pixel×144 pixel)上进行目标预测，改进后的预测结构 1 如图 4 所示。

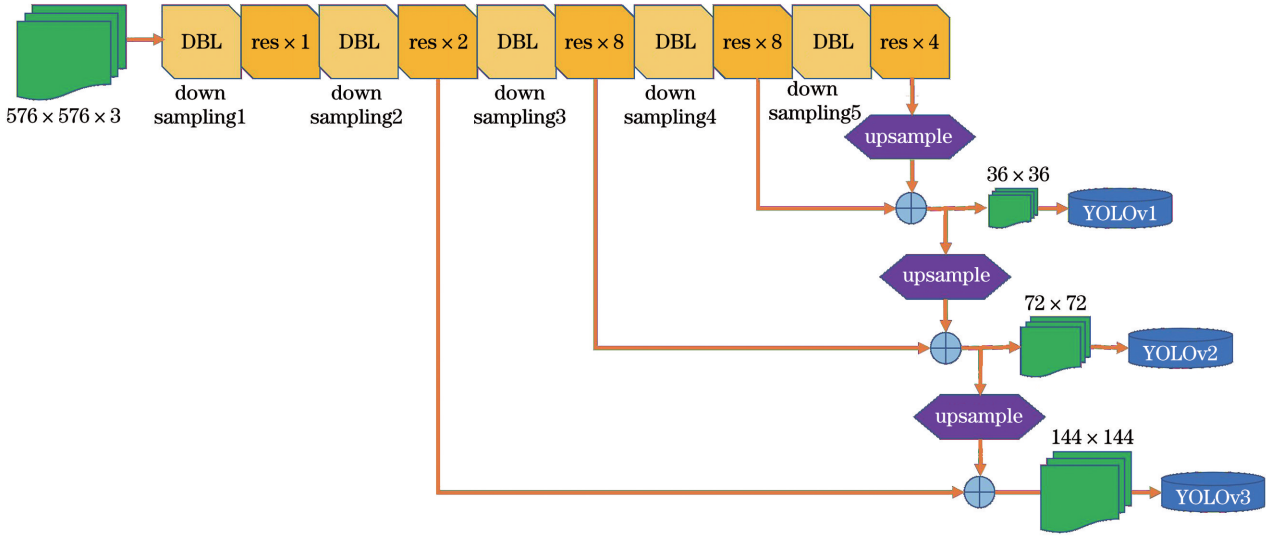

图 4　改进的网络预测结构 1

Fig. 4　Improved network prediction structure1

改进网络预测结构 1 通过上移检测层的特征图提高对小目标的检测效果，但幅射范围仍为连续的三级特征图。在此基础上，采用跨尺度预测层，在尺寸为 18 pixel × 18 pixel、72 pixel × 72 pixel、144 pixel×144 pixel 的特征图上进行了跨尺度目标预测，预测层特征图跨越的尺度更广，改进的预测结构 2 如图 5 所示。

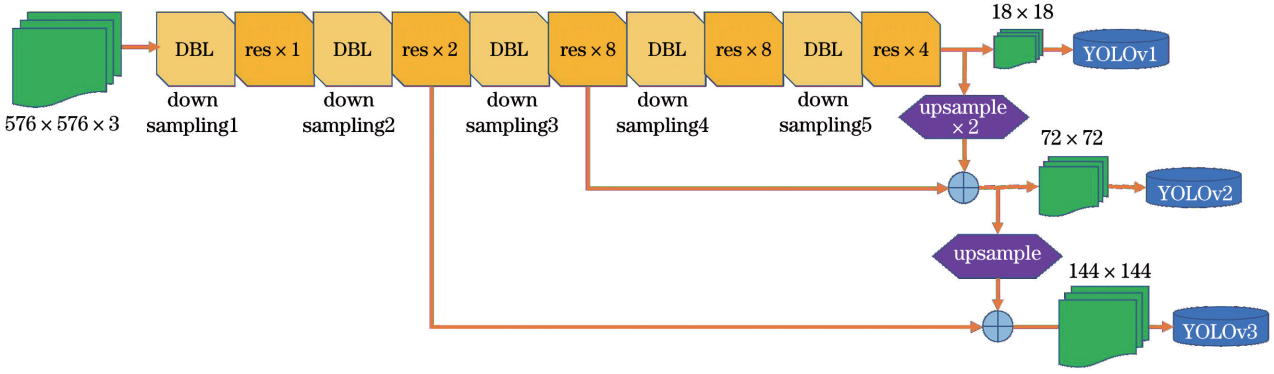

图 5　改进的网络预测结构 2

Fig. 5　Improved network prediction structure2

4.2　网络感受野的提高

YOLOv3 网络对输入图像进行了 5 次降采样卷积，并在卷积操作后加入了五组残差块实现恒等映射，解决了网络深度加深导致的退化问题，改善了更深层网络的检测效果。但引入了大量的卷积层和 shotcut 层，导致网络的检测速度下降。杜泽星等[21]采用多分支不同大小的卷积核提高感受野。为了进一步提高检测层的感受野，避免进入网络前降采样过程中的精度损失，在改进结构 2 不增加参数量的基础上提出将深度残差块组传递改为结合深度和广度传递的网络结构，如图 6 所示。

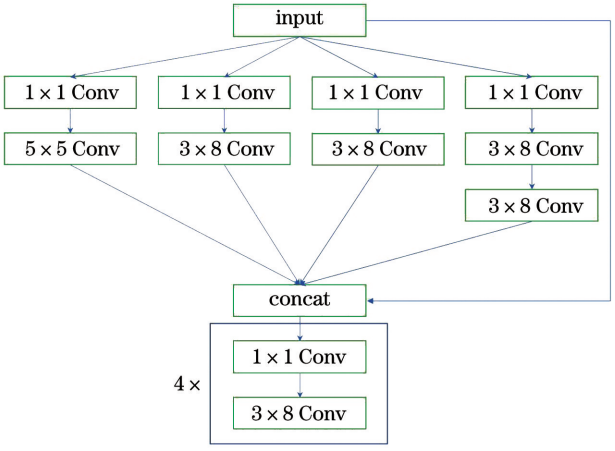

图 6　感受野的优化网络

Fig. 6　Optimized network of the receptive field

5　实验结果分析

5.1　评估指标

用精确率(Precision)和召回率(Recall)作为网络的评价指标，精确率为单类别的正确识别个数 N_{true} 与预测总数(包含正确识别为该类别数 N_{true} 和错误识别为该类别数 N_{false})的比值，可表示为

$$X_{\text{Precision}}=\frac{N_{\text{true}}}{N_{\text{true}}+N_{\text{false}}}。\tag{2}$$

召回率表示单类别的正确识别个数 N_{true} 与实际该类别总数 N_{all} 的比值，可表示为

$$X_{\text{Recall}}=\frac{N_{\text{true}}}{N_{\text{all}}}。\tag{3}$$

对于多类别数据集用平均精确率(mPre)和平均召回率(mRec)表征网络的性能水平，可表示为

$$X_{\text{mPre}}=\frac{\sum_{n}X_{\text{Precision}}}{n},\tag{4}$$

$$X_{\text{mRec}}=\frac{\sum_{n}X_{\text{Recall}}}{n},\tag{5}$$

式中，n 为类别总数。

5.2　结果分析

将 DOTA 数据集中的 Plane、Small-vehicle 和 Large-vehicle 三类图像裁剪成尺寸为 1000 pixel×1000 pixel 的图像进行训练测试，Plane 图像的目标像素占比为 0.02，Large-vehicle 图像的目标像素占比为 0.002，Small-vehicle 图像的目标像素占比为 0.0007。用 3.1 节提出的网络进行训练测试，并与原始 YOLOv3 网络进行对比，结果如表 2 和表 3 所示。可以发现，相比 YOLOv3 网络，结构 2 在尺寸为 1000 pixel×1000 pixel 的图像中检测小目标的平均精确率提升了 1.9 个百分点，召回率提升了 5.9 个百分点。结构 1 和结构 2 的平均精确率和平

均召回率相比 YOLOv3 网络均有提升，但结构 1 仍存在小目标漏检现象，结构 2 在平均精确率与结构 1 持平的情况下，平均召回率更高，达到 94.5%，对较小尺寸 Small-vehicle 图像的类别提升更明显，且改进网络的定位信息更精确，检测框的尺寸更切合实际尺寸。3 种网络对三类目标的检测效果如图 7 所示。

表 2　不同网络的召回率

Table 2　Recall rates of different networks　unit：%

Network	Plane	Large-vehicle	Small-vehicle	Average
YOLOv3	98.0	85.8	82.4	88.6
Improved structure1	98.3	90.0	85.6	91.3
Improved structure2	97.8	88.9	96.9	94.5

表 3　不同网络的精确率

Table 3　Precision rates of different networks　unit：%

Network	Plane	Large-vehicle	Small-vehicle	Average
YOLOv3	97.4	80.0	81.4	86.3
Improved structure1	97.0	83.0	85.0	88.3
Improved structure2	96.5	81.1	87.1	88.2

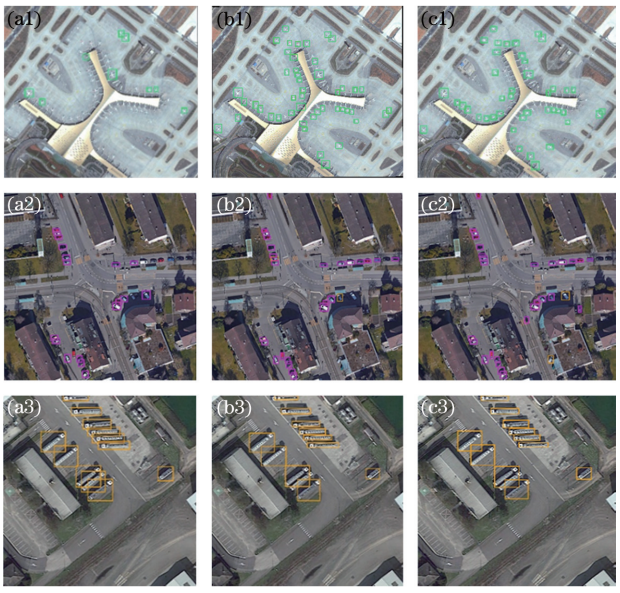

图 7　不同网络的检测效果。(a) YOLOv3；(b)结构 1；(c)结构 2

Fig. 7　Detection effect of different networks. (a) YOLOv3; (b) structure1; (c) structure2

在保证目标像素大小不变的情况下，基于尺寸大于 2500 pixel×2500 pixel 的高分辨率图像对结构 2 与 3.2 节中优化感受野后的结构进行对比。为了进一步验证优化网络的检测效果，在对比实验中加入了基于区域的全卷积网络(R-FCN)[22]，R-FCN 引入了位置敏感得分图，实现了更多参数与特征的信息共享，对小目标的检测效果更好。4 种网络的召回率、精确率和基本参数如表 4～表 6 所示，耗时为 GTX-1060 下的测试结果。可以发现，相比 YOLOv3 网络，优化感受野网络的平均精确率提升了 31.6 个百分点，召回率提升了 46.4 个百分点。

表 4　不同网络的多类别召回率

Table 4　Multi-category recall rates of different networks　unit：%

Network	Plane	Large-vehicle	Small-vehicle	Average
YOLOv3	64.3	27.5	11.6	34.5
R-FCN	72.9	31.9	14.2	39.7
Improved structure2	89.6	57.2	61.0	69.3
Receptive field optimization	93.1	76.8	72.7	80.9

表 5　不同网络的多类别精确率

Table 5　Multi-class precision rates of different networks　unit：%

Network	Plane	Large-vehicle	Small-vehicle	Average
YOLOv3	62.0	16.7	3.5	27.4
R-FCN	68.6	29.1	13.5	34.9
Improved structure2	87.9	45.8	44.8	59.5
Receptive field optimization	90.3	37.7	49.4	59.0

表 6　不同网络的基本参数

Table 6　Basic parameters of different networks

Network	Volume /Mb	Time consuming /s
YOLOv3	246.3	0.063
R-FCN	102.5	0.180
Improved structure2	242.7	0.085
Receptive field optimization	239.4	0.083

实验的训练过程基于 2 台 GTX-1080 处理器，网络训练过程每轮迭代抽取 64 张图像，再将 batch 分 4 次送入网络训练，权重衰减正则项为 0.0005，初始学习率为 0.01，随着训练进程逐渐衰减。采用 multistep 调整策略，设定调整学习率 step 的间隔

为 20000，30000，35000，40000，学习率调整比率为 0.2。为了生成更多的训练样本，训练过程中将图像随机旋转角度±45°，尺度上训练设定图像在 320～576 pixel 之间随机调整，以丰富训练样本，训练时的损失曲线如图 8 所示。

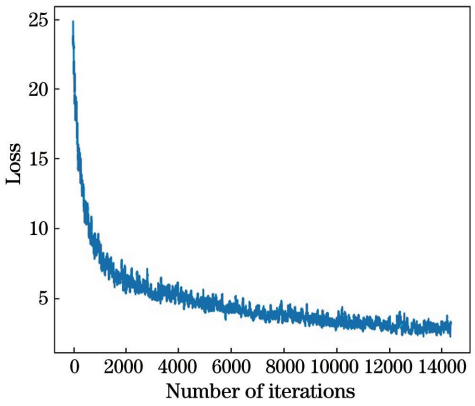

图 8　训练时的损失曲线

Fig. 8　Loss curve during training

实验结果表明，进行高分辨率图像的目标检测时，优化感受野后的网络在精确率和耗时一致的情况下，平均召回率由 69.3%提升到了 80.9%。高分辨率图像的压缩导致 Large-vehicle 和 Small-vehicle 容易被混淆，在 Large-vehicle 上的精确率有所下降。相比 R-FCN，优化感受野后的网络在精确率和召回率上均有提高，且速度更快。

结构 2 与优化感受野后的结构实际测试结果如图 9 所示，对比图 9(a1)和图 9(b1)发现，结构 2 在图像边缘容易出现漏检现象，原因是高分辨率图像下的小目标会出现过度降采样；对比图 9(a2)和图 9(b2)虚线圆圈标注区域发现，图 9(a2)将 Large-vehicle 的影子误识别为 Small-vehicle，这表明结构 2 对高分辨率图像下的小目标容易出现误检，而优化感受野后的网络可以有效准确地识别出图像及边界的小目标物体。

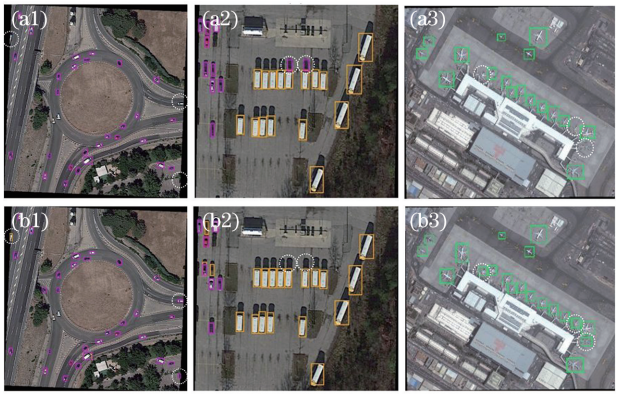

图 9　不同网络的识别效果。(a)结构 2；(b)优化感受野的网络

Fig. 9　Recognition effect of different networks. (a) Structure2; (b) optimize the network of the receptive field

为了进一步验证改进网络在其他类型图像中对小目标检测的有效性，将航拍数据集替换为 COCO 数据集后，对改进的网络和原始 YOLOv3 网络重新进行训练，并对检测结果进行对比分析，结果如表 7 和表 8 所示。其中，Small 表示目标像素面积小于 32 pixel×32 pixel；Medium 表示目标像素面积大于 32 pixel×32 pixel，小于 96 pixel×96 pixel；Large 表示目标像素面积大于 96 pixel×96 pixel。优化后网络在 COCO 数据集上的平均召回率提升了 8.9 个百分点，小目标召回率提升了 12.2 个百分点；平均精确率提升了 6.8 个百分点，小目标精确率提升了 11.0 个百分点，整体检测效果均有提升。以小目标绵羊为例，其检测效果如图 10 所示。

表 7　不同网络在 COCO 数据集下的召回率

Table 7　Recall rates of different networks under the COCO data set　　unit：%

Network	Small	Medium	Large	Average
YOLOv3	24.0	48.2	61.1	44.4
Receptive field optimization	36.2	58.2	65.5	53.3

表 8　不同网络在 COCO 数据集下的精确率

Table 8　Precision rates of different networks under the COCO data set　　unit：%

Network	Small	Medium	Large	Average
YOLOv3	14.2	34.1	46.4	31.6
Receptive field optimization	25.2	41.5	48.5	38.4

图 10 不同网络在 COCO 数据集下的检测结果。(a)YOLOv3 网络;(b)优化感受野后的网络

Fig. 10 Detection results of different networks under the COCO data set. (a) YOLOv3 network; (b) optimize the network of the receptive field

6 结 论

针对小目标检测中的问题,基于 YOLOv3 网络提出了一种跨尺度融合检测方法,在经过处理的 DOTA 数据集上进行了训练和测试实验,并进一步在 DOTA 高分辨率图像和 COCO 数据集上进行了优化感受野后网络的测试实验。实验结果表明,扩大检测层的辐射范围并提升网络的实际感受野可以提高对视场中小目标的识别检测效果。但本方法在检测速度上并没有太多提升,因此在保证精度的前提下进一步提升检测速度,是下一步需要研究的重点。

参 考 文 献

[1] Zhao Z Q, Zheng P, Xu S T, et al. Object detection with deep learning: a review[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 30(11): 3212-3232.

[2] Yue P Y, Xin J, Zhao H, et al. Experimental research on deep reinforcement learning in autonomous navigation of mobile robot[C]// 2019 14th IEEE Conference on Industrial Electronics and Applications (ICIEA), June 19-21, 2019, Xi′an, China. New York: IEEE Press, 2019: 1612-1616.

[3] Zhang D F, Liu Y H, Zhang R F. Intelligent assistant driving system based on deep leaning[J]. Electronic Science and Technology, 2018, 31(10): 60-63.
张达峰, 刘宇红, 张荣芬. 基于深度学习的智能辅助驾驶系统[J]. 电子科技, 2018, 31(10): 60-63.

[4] Duan Z J, Li S B, Hu J J, et al. Review of deep learning based object detection methods and their mainstream frameworks[J]. Laser & Optoelectronics Progress, 2020, 57(12): 120005.
段仲静, 李少波, 胡建军, 等. 深度学习目标检测方法及其主流框架综述[J]. 激光与光电子学进展, 2020, 57(12): 120005.

[5] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]// 2014 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2014, Columbus, OH, USA. New York: IEEE Press, 2014: 580-587.

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

[7] 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.

[8] Liu W, Anguelov D, Erhan D, et al. SSD: single shot MultiBox detector[EB/OL]. [2020-07-02]. https://arxiv.org/abs/1512.02325.

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

[10] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 6517-6525.

[11] Redmon J, Farhadi A. YOLOv3: an incremental improvement[EB/OL]. [2020-07-04]. https://arxiv.org/abs/1804.02767.

[12] Bochkovskiy A, Wang C Y, Liao H Y Mark. YOLOv4: optimal speed and accuracy of object detection[EB/OL]. [2020-07-02]. http://arxiv.org/

abs/2004.10934.

[13] Bell S, Zitnick C L, Bala K, et al. Inside-outside net: detecting objects in context with skip pooling and recurrent neural networks [C] // 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 2874-2883.

[14] Mäenpää T, Pietikäinen M. Texture analysis with local binary patterns [M] // Kalviainen H, Parkkinen J, Kaarna A, et al. Image Analysis. SCIA 2005. Lecture Notes in Computer Science. Cham: Springer, 2005, 3540: 115-118.

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

[16] Wang H T, Jiang W D, Cheng Y, et al. Two-stage context convolutional network for small target detection in wide-view-field images [J]. Computer Measurement & Control, 2019, 27(6): 199-204.
王海涛, 姜文东, 程远, 等. 两级上下文卷积网络宽视场图像小目标检测方法[J]. 计算机测量与控制, 2019, 27(6): 199-204.

[17] Zhou S, Zhi X L, Liu D, et al. A convolutional neural network-based method for small traffic sign detection [J]. Journal of Tongji University (Natural Science), 2019, 47(11): 1626-1632.
周苏, 支雪磊, 刘懂, 等. 基于卷积神经网络的小目标交通标志检测算法[J]. 同济大学学报(自然科学版), 2019, 47(11): 1626-1632.

[18] Ju L Y, Wang H. The application of improved YOLOv3 in multi-scale target detection [J]. Applied Sciences, 2019, 9(18): 3775.

[19] Wang J Q, Li J S, Zhou X W, et al. Improved SSD algorithm and its performance analysis of small target detection in remote sensing images [J]. Acta Optica Sinica, 2019, 39(6): 0628005.
王俊强, 李建胜, 周学文, 等. 改进的 SSD 算法及其对遥感影像小目标检测性能的分析[J]. 光学学报, 2019, 39(6): 0628005.

[20] Liu L R, Tang X M, Zhao W J, et al. Detection and geo-localization of small traffic signs based on images and laser data [J]. Chinese Journal of Lasers, 2020, 47(9): 0910002.
刘力荣, 唐新明, 赵文吉, 等. 基于影像与激光数据的小交标检测与地理定位[J]. 中国激光, 2020, 47(9): 0910002.

[21] Du Z X, Yin J Y, Yang J. Remote sensing image detection based on dense connected networks [J]. Laser & Optoelectronics Progress, 2019, 56(22): 222803.
杜泽星, 殷进勇, 杨建. 基于密集连接网络的遥感图像检测方法[J]. 激光与光电子学进展, 2019, 56(22): 222803.

[22] Dai J F, Li Y, He K M, et al. R-FCN: object detection via region-based fully convolutional networks [EB/OL]. [2020-07-04]. https://arxiv.org/abs/1605.06409.