第 58 卷 第 20 期/2021 年 10 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

# 融合场景上下文的轻量级目标检测网络

刘婷婷，苗华*，李琳，向阳，李琦，孟奇

长春理工大学光电工程学院，吉林 长春 130022

**摘要** 提出了一种融合场景上下文的轻量级目标检测网络，有效地解决了现有检测算法在无人机领域应用效果较差的问题。在该网络的设计中，首先用 MobileNetV3 替换 YOLOv3 的主干网络，并通过 1×1 卷积层提取场景信息。同时，构建场景上下文模块以筛选物体的细粒度特征。再采用完全交并比(CIOU)损失对损失函数中的边界框位置误差项进行优化。最后，在新建无人机航拍数据集上对所提算法进行训练与测试。实验结果表明，相较于 YOLOv3 算法，所提算法的平均检测精度提高了 8.4 个百分点，检测速度提高了 5.8 frame/s。

**关键词** 图像处理；目标检测；场景上下文；YOLOv3；简单循环单元

**中图分类号** P407.8 **文献标志码** A **doi**：10.3788/LOP202158.2010005

# Lightweight Target Detection Network Integrating Scene Context

Liu Tingting, Miao Hua*, Li Lin, Xiang Yang, Li Qi, Meng Qi

*School of Opto-Electronic Engineering, Changchun University of Science and Technology, Changchun, Jilin* 130022, *China*

**Abstract** A lightweight target detection network that integrates scene context is proposed, effectively solving the problem of poor application of existing detection algorithms in the field of unmanned aerial vehicles. In the design of the network, first, the backbone network of YOLOv3 is replaced with MobileNetV3, and scene information is extracted through the 1×1 convolutional layer. Simultaneously, a scene context module is constructed to filter fine-grained object features. Then, complete intersection over union (CIOU) loss is used to optimize the position error of the bounding box in the loss function. Finally, the algorithm is trained and tested on the newly constructed unmanned aerial vehicle aerial photography data set. Experimental results show that compared with the YOLOv3 algorithm, the average detection accuracy of the proposed algorithm increased by 8.4 percent and the detection speed increased by 5.8 frame/s.

**Key words** image processing; target detection; scene context; YOLOv3; simple recurrent unit

**OCIS codes** 100.3008; 100.2000; 100.4996

## 1 引 言

近年来，无人机被广泛应用到航空摄影[1]、安全监控[2]、基础设施检查[3]等领域。基于航拍图像的视觉对象检测也成为人们越来越感兴趣的领域。但是基于传统机器学习和以人工设计特征为主的目标检测方法泛化能力较弱，存在一定的局限性。而随着深度神经网络的发展，尤其是 Krizhevsky 等[4]提出的 Alexnet 在 ImageNet 图像分类比赛中取得冠军，深度学习在目标检测中取得了巨大的进展。

目前，基于深度学习的目标检测算法可以大致分为两类：基于候选区域的两阶段算法和基于直接回归的一阶段算法。两阶段算法以 Fast R-CNN[5]、Faster R-CNN[6]和 Mask-RCNN[7]等为

**收稿日期**：2020-11-12；**修回日期**：2020-12-15；**录用日期**：2021-01-02

**基金项目**：吉林省科技发展计划(20200404157YY)

**通信作者**：*ilev24@163.com

主，依靠区域建议模块生成的高质量区域建议来获得良好的检测精度。一阶段算法以 YOLO 系列[8-10]、SSD 系列[11-12]和 RetinaNet[13]等为主，通过输入图像直接获取类别概率和边界框坐标来提高检测速度。这些检测算法在自然场景下取得了巨大的成功。而无人机嵌入式设备的计算能力有限，且航拍图像背景复杂、小目标居多、成像视角与自然场景图像不同，因此直接将现有检测算法应用于无人机领域效果较差。为此，国内外许多学者提出一些效果显著的算法来解决这一挑战性问题。文献[14]通过深度可分离卷积降低了网络计算成本和参数的数量，在计算能力有限的设备上实现了高效的检测性能。文献[15]采用不同分辨率的多层特征和尺度转换层生成大尺寸特征图，提升了整体网络表达的精确性。文献[16]提出一种多尺度自适应候选区域生成网络，该网络按照通道维度加权融合空间尺寸一致的特征图，增强了特征的表达能力。文献[17]利用语义分割指导区域候选网络模块来抑制航拍图像的背景杂波，从而得到更准确的回归结果，但获取的特征与上下文联系不够紧密，容易造成信息的丢失。文献[18]将图像金字塔模型与特征金字塔网络结合，提出一种图像级联网络来增强对小目标的定位能力。文献[19]提出一种解决复杂背景下小目标信息丢失的检测算法，该算法利用全局注意力模块来整合深层特征和浅层特征。文献[20]在特征金字塔网络原有基础上，通过并行结构加强小目标特征的表达能力，通过级联结果加强小目标的定位能力。文献[21]提出一种基于 YOLOv2[9]、YOLOv3[10]网络改进的实时道路目标检测模型。文献[22]在 YOLOv3 网络的基础上，通过深度残差网络和多尺度预测优化网络提高航拍检测的精度。虽然上述算法在一定程度上提高了航拍图像的精度和效率，但仍未达到预期效果。

为解决无人机航拍图像目标检测所面临的问题，本文提出了一种基于场景上下文信息的轻量级目标检测(SS-YOLOv3)网络，实现了实时高效的航拍图像检测。首先用 MobileNetV3[14]替代 YOLOv3 的主干网络，减少计算量，在保证检测精度的前提下优化模型延迟。然后构建场景上下文模块，有选择地增强或忽略某些物体状态，提高目标的检测精度。最后通过改进损失函数，优化边界框回归位置，进一步提高网络的性能。

## 2　SS-YOLOv3 网络

在 YOLOv3 网络结构的基础上，通过改进主干网络和引入场景上下文模块来构建 SS-YOLOv3，结构如图 1 所示。

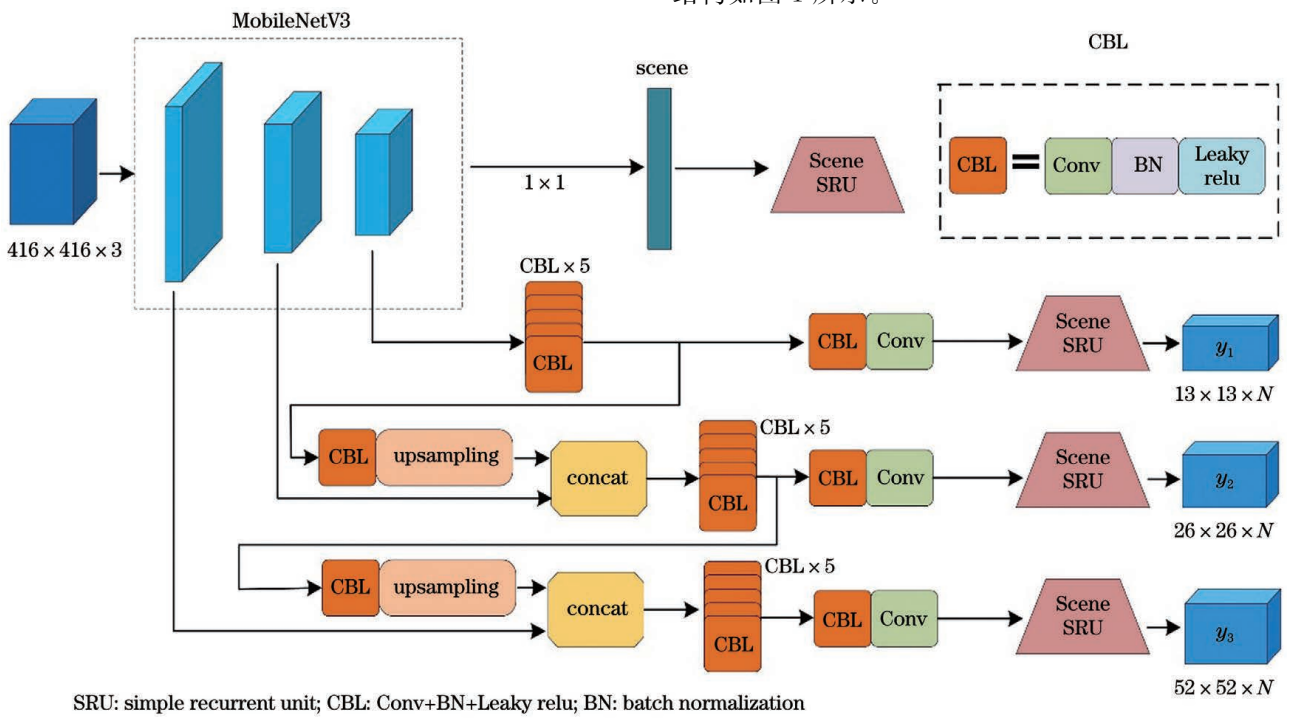


图 1　SS-YOLOv3 结构图

Fig. 1　Structure diagram of SS-YOLOv3

### 2.1　MobileNet-YOLOv3 网络

MobileNetV3 使用深度可分离卷积将标准卷积分解成深度卷积和逐点卷积两个部分，使参数量减少为标准卷积的 1/9，减小了模型的空间复杂度，同时提高了计算效率，降低了时间复杂度。它还利用线性瓶颈逆残差结构在输入和输出处保持紧凑的表示，同时去除前面的纺锤形 3×3 卷积和 1×1 卷积，进一步减少了参数量，从而进一步降低了空间和时间的计算复杂度；并在未增加时间消耗的前提下，通过轻量级注意力模型提高有效特征图的权重。它集现有轻量级模型思想于一体，在检测速度和精度上达到了较好的平衡。因此，所提算法选取轻量级卷积神经网络 MobileNetV3 作为特征提取网络。

YOLOv3 采用了类似特征金字塔的上采样和特征融合的方法共融合了 3 个尺度，分别为 13×13、26×26 和 52×52，并在 3 个尺度的特征图上进行位置和类别的预测。因此在 MobileNetV3 中找到对应的 13×13×160、26×26×112 和 52×52×40 部分，使用 YOLOv3 多尺度特征融合方法进行融合，同时去除 MobileNetV3 最后 4 层卷积和全局池化网络，得到初始轻量级检测模型。MobileNet-YOLOv3 结构如图 2 所示。

| type | filters | size | input | output |
|---|---|---|---|---|
| Conv2d | 16 | 3×3/2 | 416×416×3 | 208×208×16 |
| bneck | 16 | 3×3 | | 208×208×16 |
| bneck | 24 | 3×3/2 | 208×208×16 | 104×104×24 |
| bneck | 24 | 3×3 | | 104×104×24 |
| bneck | 40 | 5×5/2 | 104×104×24 | 52×52×40 |
| bneck | 40 | 5×5 | | |
| bneck | 40 | 5×5 | | 52×52×40 |
| bneck | 80 | 3×3/2 | 52×52×40 | 26×26×80 |
| bneck | 80 | 3×3 | | |
| bneck | 80 | 3×3 | | |
| bneck | 80 | 3×3 | | 26×26×80 |
| bneck | 112 | 3×3 | 26×26×80 | 26×26×112 |
| bneck | 112 | 3×3 | | 26×26×112 |
| bneck | 160 | 5×5/2 | 26×26×112 | 13×13×60 |
| bneck | 160 | 5×5 | | |
| bneck | 160 | 5×5 | | 13×13×60 |

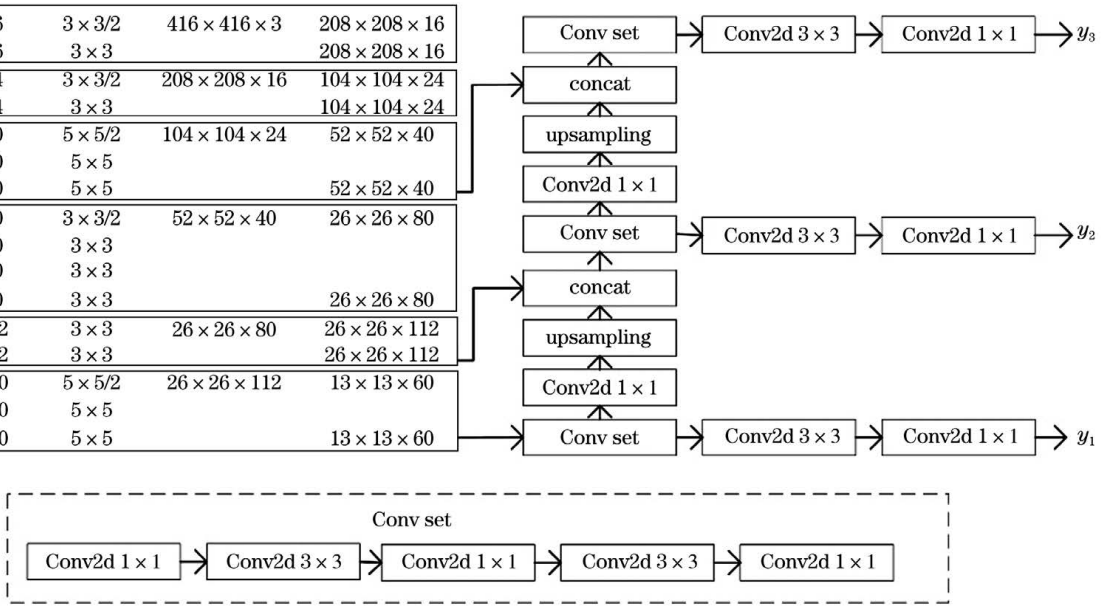


图 2　MobileNet-YOLOv3 结构图

Fig. 2　Structure diagram of MobileNet-YOLOv3

### 2.2　场景上下文模块

通常，图像包含丰富的场景上下文信息，但大多数检测算法只考虑图像内感兴趣对象区域附近的信息，忽略上下文信息，不可避免地会降低检测物体的准确性。由于只关注物体的视觉外观，河流里的某些船只被错误识别为汽车，如图3所示。若考虑该

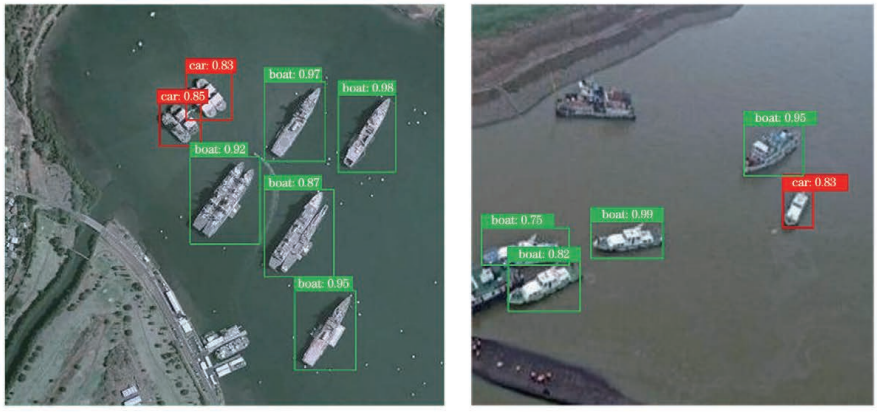


图 3　YOLOv3 检测结果

Fig. 3　Test results of YOLOv3

图片的场景信息，则很容易避免这种错误。因此，所提算法通过构建场景上下文模块在对象与整个场景之间迭代地传播消息来改进检测模型。具体来说，一个物体从场景中接受消息，这样物体的状态不仅仅由其细粒度的外观细节决定，还受场景上下文信息的影响。

所提算法在特征提取网络 MobileNetV3 后增加 1×1 卷积层来提取整个图像的视觉特征信息，并将其作为场景信息 $\boldsymbol{f}^{\mathrm{s}}$，使用轻量级的简单循环单元(SRU)[23]作为存储单元，SRU 的体系结构分为轻度循环和公路网络两个部分，可以表示为

$$\boldsymbol{f}_t=\sigma(\boldsymbol{W}_{\mathrm{f}}\boldsymbol{x}_t+\boldsymbol{v}_{\mathrm{f}}\odot\boldsymbol{c}_{t-1}+\boldsymbol{b}_{\mathrm{f}}),\tag{1}$$

$$\boldsymbol{c}_t=\boldsymbol{f}_t\odot\boldsymbol{c}_{t-1}+(\boldsymbol{1}-\boldsymbol{f}_t)\odot(\boldsymbol{W}\boldsymbol{x}_t),\tag{2}$$

$$\boldsymbol{r}_t=\sigma(\boldsymbol{W}_{\mathrm{r}}\boldsymbol{x}_t+\boldsymbol{v}_{\mathrm{r}}\odot\boldsymbol{c}_{t-1}+\boldsymbol{b}_{\mathrm{r}}),\tag{3}$$

$$\boldsymbol{h}_t=\boldsymbol{r}_t\odot\boldsymbol{c}_t+(\boldsymbol{1}-\boldsymbol{r}_t)\odot\boldsymbol{x}_t,\tag{4}$$

式中：$\boldsymbol{W}_{\mathrm{f}}$、$\boldsymbol{W}$ 和 $\boldsymbol{W}_{\mathrm{r}}$ 均是参数矩阵；$\boldsymbol{v}_{\mathrm{f}}$、$\boldsymbol{v}_{\mathrm{r}}$、$\boldsymbol{b}_{\mathrm{f}}$ 和 $\boldsymbol{b}_{\mathrm{r}}$ 均是训练学习到的参数向量；$\sigma$ 为激活函数；$\odot$ 为逐点乘法；$\boldsymbol{x}_t$ 是输入向量；$\boldsymbol{c}_t$ 是状态向量；$\boldsymbol{f}_t$ 是遗忘门；$\boldsymbol{W}\boldsymbol{x}_t$ 是当前观察值；$\boldsymbol{r}_t$ 是复位门；$\boldsymbol{h}_t$ 为输出向量；$(\boldsymbol{1}-\boldsymbol{r}_t)\odot\boldsymbol{x}_t$ 是一个跳跃连接。轻度循环依次读取输入向量 $\boldsymbol{x}_t$ 并计算捕获顺序信息的状态向量 $\boldsymbol{c}_t$。遗忘门 $\boldsymbol{f}_t$ 控制信息流和状态向量 $\boldsymbol{c}_t$。状态向量由 $\boldsymbol{f}_t$ 自适应地平均先前状态 $\boldsymbol{c}_{t-1}$ 和当前观察值 $\boldsymbol{W}\boldsymbol{x}_t$ 来确定。公路网络促进了基于梯度的深度网络的训练。复位门 $\boldsymbol{r}_t$ 自适应地组合输入向量 $\boldsymbol{x}_t$ 和从轻度循环中产生的状态向量 $\boldsymbol{c}_t$。$(\boldsymbol{1}-\boldsymbol{r}_t)\odot\boldsymbol{x}_t$ 允许将梯度直接传播到上一层。由于跳跃连接去除了对前一时刻计算结果的依赖，降低了模型的计算复杂度，并实现了并行化处理，很大程度上提高了训练速度。

所提算法将特征图作为节点 $\boldsymbol{v}_i$，同样通过 1×1 卷积层来获取物体的细粒度信息 $\boldsymbol{f}_i^{\mathrm{v}}$，将 $\boldsymbol{f}_i^{\mathrm{v}}$ 固定为 SRU 的初始状态，然后将场景信息 $\boldsymbol{f}^{\mathrm{s}}$ 作为 SRU 输入，以此构建场景上下文(Scene SRU)模块来达到更新物体状态信息的目的。Scene SRU 模块可以有选择地忽略与该场景上下文不相关的物体状态的某些部分，或利用场景上下文来增强物体状态的某些部分。Scene SRU 模块如图 4 所示。

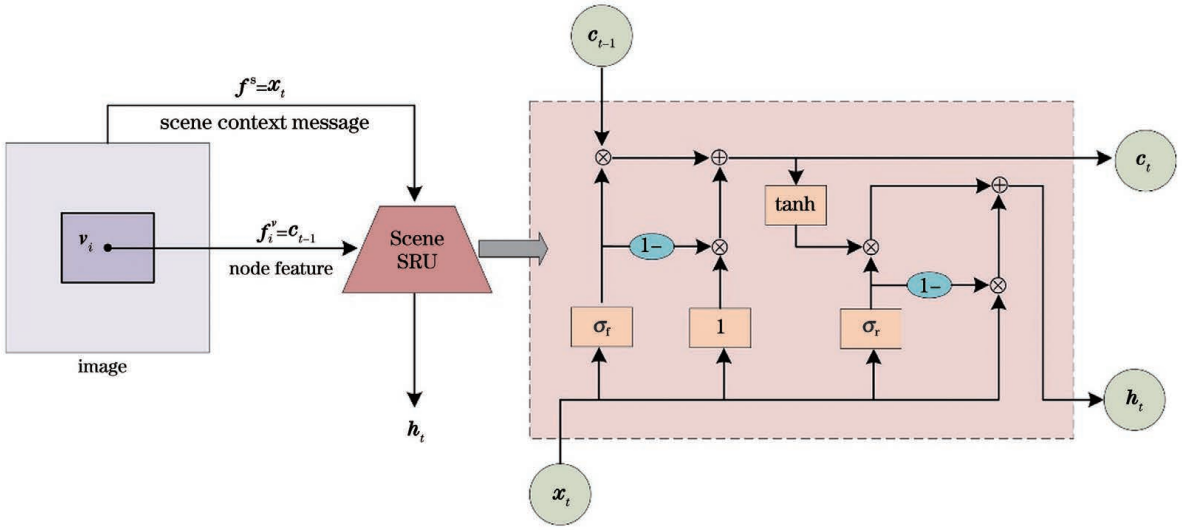


图 4　场景上下文模块

Fig. 4　Scene context module

## 2.3　改进损失函数

目标检测中常用交并比(IOU)来衡量预测框和目标框之间的重合度，但当两个框不重叠时，IOU 损失[24]无法进行度量评估，同时广义交并比(GIOU)损失[25]的收敛速度缓慢。因此，所提算法用完全交并比(CIOU)损失[26]对损失函数中的边界框位置误差项进行优化。CIOU 损失综合考虑了边界框的重叠区域、中心点距离和边框的纵横比。CIOU 损失函数为

$$L_{\mathrm{CIOU}}=1-L_{\mathrm{IOU}}+R_{\mathrm{CIOU}},\tag{5}$$

$$R_{\mathrm{CIOU}}=\frac{\rho^2(\boldsymbol{b},\boldsymbol{b}^{\mathrm{gt}})}{c^2}+\alpha\nu,\tag{6}$$

$$\alpha=\frac{\nu}{(1-L_{\mathrm{IOU}})+\nu},\tag{7}$$

$$\nu=\frac{4}{\pi^2}\left(\arctan\frac{w^{\mathrm{gt}}}{h^{\mathrm{gt}}}-\arctan\frac{w}{h}\right)^2,\tag{8}$$

式中：$R_{\mathrm{CIOU}}$ 为预测框和目标框的惩罚项；$\frac{\rho^2(\boldsymbol{b},\boldsymbol{b}^{\mathrm{gt}})}{c^2}$ 为两个边界框中心点之间的最小标准化距离，其中 $\boldsymbol{b}$ 为预测框的中心点值，$\boldsymbol{b}^{\mathrm{gt}}$ 为目标框的中心点值，$\rho$

为欧几里得距离，$c$ 为覆盖两个边框的最小封闭边框的对角线长度；$\alpha$ 为一个正的权衡参数，$\nu$ 为衡量长宽比的一致性；$w^{gt}$ 和 $h^{gt}$ 为目标框的宽和高；$w$ 和 $h$ 为预测框的宽和高。

可以看出，重叠面积因子在回归上具有更高的优先级，尤其是对于非重叠情况。CIOU 可以更好地描述边框的回归，而且训练时收敛速度更快。因此，所提算法用 CIOU 来代替原损失函数中的位置误差项 IOU。

## 3　实验方法

所提算法选用 keras 开源框架，并使用 TensorFlow 作为后端进行目标检测网络训练。操作系统为 Windows 10，实验平台处理器为 Inter i7-8750H，运行内存为 12 GB，显卡为 GeForce GTX 1080 Ti。依靠手动调参的方式对超参数进行选取，根据经验设置权重衰减系数和动量系数的建议值；通过实验发现，较小的学习率更适合 SS-YOLOv3 网络，为了进一步精调模型，采用分段常数衰减的方法来设置学习率；由于内存资源的限制，选取较少的批样本数量；同时通过计算测试错误率和训练错误率的差值来选取最大迭代次数。预训练模型的超参数设置如表 1 所示，其中 $M$ 表示迭代次数。

表 1　预训练模型的超参数

Table 1　Hyperparameters of pre-trained model

| Weight decay | Batch size | Learning rate | Momentum | Maximum iteration |
|---|---|---|---|---|
| 0.0005 | 16 | 0.001 ($M$<1000)<br>0.0001($M$>1000) | 0.9 | 10000 |

### 3.1　数据集的制作

针对所研究的问题，从互联网上选取了大量无人机航拍图片，通过旋转、裁剪、随机平移等数据增强方法生成数据集。该数据集共 4000 张图片，包括 7 类目标，分别为行人、汽车、船、自行车、飞机、房屋和电动车。使用标注工具 LabelImg 按 Visual Object Classes(VOC)数据集格式进行手工标记，并按 75%、25%的比例分割为训练集、验证集。测试集由未参与数据增强的原始图片制作而成。图 5 为几种典型航拍图像。


图 5　典型航拍图像。(a)尺度变化大；(b)背景复杂；(c)小目标居多

Fig. 5　Typical aerial images. (a) Large scale variation; (b) complex background; (c) most small goals

### 3.2　实验结果与分析

为了测试 SS-YOLOv3 在无人航拍图片检测中的表现，对 SS-YOLOv3 和经典目标检测算法在新建数据集上进行对比实验，并选用平均精准度均值(mAP)和检测速度(FPS)作为评价指标来衡量不同算法的综合性能。SS-YOLOv3 与其他算法对比结果如表 2 所示。

由表 2 可以看出，SS-YOLOv3 的 mAP 值高于其他检测算法，达到了 84.3%，比 YOLOv3 网络提高了 8.4 个百分点。虽然 SS-YOLOv3 的检测速度

表 2　不同目标检测算法结果

Table 2　Results of different target detection algorithms

| Algorithm | Backbone | mAP /% | FPS /(frame·s$^{-1}$) |
|---|---|---|---|
| Fast R-CNN | VGG16 | 70.0 | 6 |
| Faster R-CNN | VGG16 | 75.8 | 8 |
| YOLO | | 64.3 | 18 |
| $SSD_{300}$ | VGG16 | 73.4 | 39 |
| $SSD_{500}$ | VGG16 | 74.9 | 22 |
| YOLOv2 | Darknet-19 | 74.3 | 34 |
| YOLOv3 | Darknet-53 | 75.9 | 27.6 |
| SS-YOLOv3 | MobileNetV3 | 84.3 | 33.4 |

低于 $SSD_{300}$ 和 YOLOv2，但仍比 YOLOv3 提高了 5.8 frame/s，达到了 33.4 frame/s，满足实时检测的要求。实验结果证明，SS-YOLOv3 在满足实时性要求的前提下，提高了对无人机航拍图像的检测速度。

为了更加直观地体现模型的准确性，通过平均精准度（AP）来比较 SS-YOLOv3 和 YOLOv3 网络的每一类目标的检测精度。各类别的 AP 值如表 3 所示。可以看出，SS-YOLOv3 相较于 YOLOv3，每一类目标的检测精度都有所提高，具有更好的检测性能。

图 6 是 SS-YOLOv3 与 YOLOv3 对比的可视化

表 3　YOLOv3 与 SS-YOLOv3 检测结果

Table 3　Detection results of YOLOv3 and SS-YOLOv3　　unit：%

| Class | YOLOV3 | SS-YOLOv3 | Subtract |
|---|---|---|---|
| Aeroplane | 74.6 | 84.9 | 10.3 |
| Car | 75.9 | 84.0 | 8.1 |
| Boat | 75.7 | 85.6 | 9.9 |
| Cycle | 76.3 | 82.4 | 6.1 |
| House | 74.3 | 84.6 | 10.3 |
| Electrocar | 76.9 | 82.7 | 5.8 |
| Person | 77.6 | 85.9 | 8.3 |

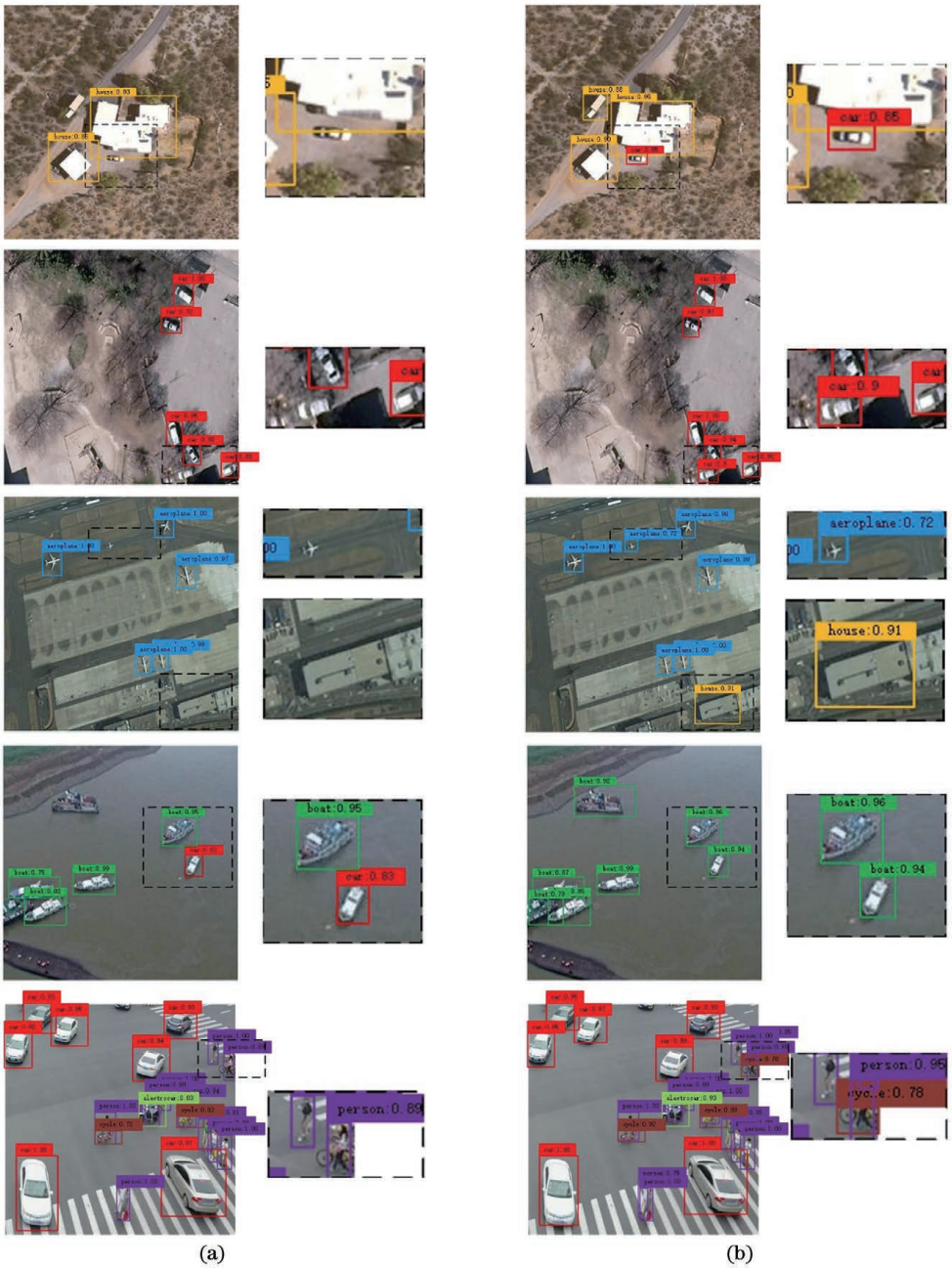


图 6　可视化结果及部分区域放大图对比。（a）YOLOv3 检测结果；（b）SS-YOLOv3 检测结果

Fig. 6　Comparison of visual test results and enlarged view of some areas. (a) YOLOv3 test results; (b) SS-YOLOv3 test results

检测结果及部分区域的放大图。其中图 6(a)为 YOLOv3 的检测结果，图 6(b)为 SS-YOLOv3 的检测结果。可以看出，在背景复杂、目标尺度变化过大或小目标居多时，所提算法比 YOLOv3 有更好的适应性，在一定程度上降低了误检率和漏检率，提高了检测精度。

为了更好地理解 SS-YOLOv3 中各个改进模块对检测效果的影响，对各个模块进行消融学习，结果如表 4 所示。可以看出，MobileNetV3 作为 YOLOv3 的主干网络后，检测速度提高了 8.8 frame/s，mAP 值降低了 1.7 个百分点；引入场景上下文模块后，mAP 值提高了 9.3 个百分点，虽然检测速度下降了 3 frame/s，但仍超出 YOLOv3 网络 5.8 frame/s；用 CIOU 改进回归框损失函数后，mAP 值提高了 0.8 个百分点。结果表明，各个改进模块可以有效提高检测网络对航拍图像的检测性能。

表 4　消融学习结果

Table 4　Results of ablation learning

| YOLOv3 | MobileNetV3 | Scene SRU | CIOU | mAP /% | FPS /(frame·s$^{-1}$) |
|---|---|---|---|---|---|
| √ | | | | 75.9 | 27.6 |
| √ | √ | | | 74.2 | 36.4 |
| √ | √ | √ | | 83.5 | 33.4 |
| √ | √ | √ | √ | 84.3 | 33.4 |

为了进一步验证 SS-YOLOv3 算法的通用性，本实验组在公开无人机目标检测数据集 VisDrone[27]上进行验证。该数据集包含行人、货车、小汽车、自行车等 10 类目标，共 10209 张静态图片，其中训练集包含 6471 张，验证集包含 548 张，测试集包含 3190 张。输入图片尺寸大小不等，需将其归一化为 608×608。SS-YOLOv3 与 YOLOv3 对每个类别的检测结果如表 5 所示。可以看出，除了人和带蓬车外，SS-YOLOv3 的 AP 值较 YOLOv3 都有所提高。表明 SS-YOLOv3 可以提高航拍图像的检测精度，有一定的通用性。图 7 为 SS-YOLOv3 部分可视化结果。

表 5　YOLOv3 与 SS-YOLOv3 检测结果

Table 5　Detect results of YOLOv3 and SS-YOLOv3　　unit：%

| Algorithm | Pedestrian | Person | Bicycle | Car | Van | Truck | Tricyle | Awn | Bus | Motor |
|---|---|---|---|---|---|---|---|---|---|---|
| YOLOv3 | 20.75 | 8.04 | 7.43 | 45.82 | 30.56 | 22.89 | 13.04 | 9.27 | 33.84 | 11.46 |
| SS-YOLOv3 | 22.45 | 7.93 | 9.65 | 51.03 | 34.54 | 24.13 | 15.02 | 9.13 | 36.89 | 13.73 |

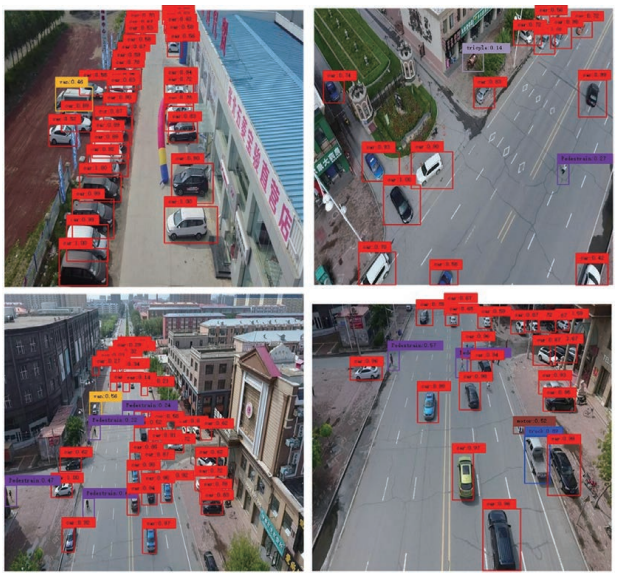

图 7　SS-YOLOv3 可视化结果

Fig. 7　Visualization results of SS-YOLOv3

## 4　结　　论

提出了一种融合场景上下文信息的轻量级目标检测模型。Scene SRU 模块利用场景信息来增强或忽略部分物体特征信息，以降低检测模型的漏检率和错检率。利用 MobileNetV3 改进 YOLOv3 的主干网络，并将 Scene SRU 模块与 MobileNet-YOLOv3 相结合构建轻量级检测模型 SS-YOLOv3。为进一步提高 SS-YOLOv3 的性能，利用 CIOU 代替 IOU 来改进损失函数。在自建的数据集上与其他经典算法进行对比实验，并在公开数据集上进行验证。结果表明，对于无人机航拍图像来说，SS-YOLOv3 具有更高的检测精度和更快的检测速度。但引入 Scene SRU 后，给检测模型增加了一定的复杂度。因此，研究模型压缩，通过剪裁卷积通道等方法来减小模型的大小是进一步的研究方向。

## 参 考 文 献


[1] Bhaskaranand M, Gibson J D. Low-complexity video encoding for UAV reconnaissance and surveillance [C]//2011 Military Communications Conference, November 7-10, 2011, Baltimore, MD, USA. New York: IEEE Press, 2011: 1633-1638.

[2] Aguilar W G, Luna M A, Moya J F, et al. Pedestrian detection for UAVs using cascade classifiers with meanshift [C] // 2017 IEEE 11th International Conference on Semantic Computing (ICSC), January 30-February 1, 2017, San Diego, CA, USA. New York: IEEE Press, 2017: 509-514.

[3] Sa I, Hrabar S, Corke P. Outdoor flight testing of a pole inspection UAV incorporating high-speed vision [M] // Mejias L, Corke P, Roberts J. Field and service robotics. Springer tracts in advanced robotics. Cham: Springer, 2015, 105: 107-121.

[4] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [J]. Communications of the ACM, 2017, 60(6): 84-90.

[5] Girshick R. Fast R-CNN[C]//2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE Press, 2015: 1440-1448.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] Redmon J, Farhadi A. YOLOv3: anincremental improvement[EB/OL]. (2018-04-08) [2019-09-19]. https://arxiv.org/abs/1804.02767.

[11] 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.

[12] Fu C Y, Liu W, Ranga A, et al. DSSD: deconvolutional single shot detector[EB/OL]. (2017-01-23) [2019-12-15]. https://arxiv.org/abs/1701.06659.

[13] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(2): 318-327.

[14] Howard A, Sandler M, Chen B, et al. Searching for MobileNetV3 [C] // 2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2019: 1314-1324.

[15] Tang Y X, Wang J, Wang X F, et al. Visual and semantic knowledge transfer for large scale semi-supervised object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(12): 3045-3058.

[16] Liu F, Wu Z W, Yang A Z, et al. Multi-scale feature fusion based adaptive object detection for UAV[J]. Acta Optica Sinica, 2020, 40(10): 1015002.
刘芳, 吴志威, 杨安喆, 等. 基于多尺度特征融合的自适应无人机目标检测[J]. 光学学报, 2020, 40(10): 1015002.

[17] Li C, Xu C, Cui Z, et al. Learning object-wise semantic representation for detection in remote sensing imagery[C]//Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 16-20, Long Beach, USA. New York: IEEE, 2019: 20-27.

[18] Azimi S M, Vig E, Bahmanyar R, et al. Towards multi-class object detection in unconstrained remote sensing imagery[M]//Jawahar C V, Li H D, Mori G, et al. Computer vision-ACCV 2018. Lecture notes in computer science. Cham: Springer, 2019, 11363: 150-165.

[19] Ying X, Wang Q, Li X W, et al. Multi-attention object detection model in remote sensing images based on multi-scale [J]. IEEE Access, 2019, 7: 94508-94519.

[20] Liu Y J, Yang F B, Hu P. Parallel FPN algorithm based on Cascade R-CNN for object detection from UAV aerial images [J]. Laser & Optoelectronics Progress, 2020, 57(20): 201505.
刘英杰, 杨风暴, 胡鹏. 基于 Cascade R-CNN 的并行特征金字塔网络无人机航拍图像目标检测算法[J].

激光与光电子学进展, 2020, 57(20): 201505.

[21] Huang T, Zhao S F, Bai Y R, et al. Method of real-time road target depth neural network detection for UAV flight control platform[J]. Laser & Optoelectronics Progress, 2020, 57(4): 041509.
黄涛, 赵栓峰, 拜云瑞, 等. 面向无人机飞控平台的实时道路目标深度神经网络检测方法[J]. 激光与光电子学进展, 2020, 57(4):041509.

[22] Ma Q, Zhu B, Zhang H W, et al. Low-altitude UAV detection and recognition method based on optimized YOLOv3[J]. Laser & Optoelectronics Progress, 2019, 56(20): 201006.
马旗, 朱斌, 张宏伟, 等. 基于优化 YOLOv3 的低空无人机检测识别方法[J]. 激光与光电子学进展, 2019, 56(20): 201006.

[23] Lei T, Zhang Y, Wang S I, et al. Simple recurrent units for highly parallelizable recurrence[EB/OL]. (2017-09-08)[2019-12-15]. https://arxiv.org/abs/1709.02755v5.

[24] Yu J H, Jiang Y N, Wang Z Y, et al. Unitbox: an advanced object detection network[EB/OL]. (2016-08-04)[2019-12-15]. https://arxiv.org/abs/1608.01471v1.

[25] Rezatofighi H, Tsoi N, Gwak J, et al. Generalized intersection over union: a metric and a loss for bounding box regression[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: 658-666.

[26] 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.

[27] Zhu P F, Wen L Y, Bian X, et al. Vision meets drones: a challenge[EB/OL]. (2018-04-20)[2019-12-15]. https://arxiv.org/abs/1804.07437.