用于空中红外目标检测的增强单发多框检测器方法

谢江荣[1,2,3]，李范鸣[1,3*]，卫红[1]，李冰[1]，邵保泰[1,2,3]

[1]中国科学院上海技术物理研究所，上海 200083；

[2]中国科学院大学，北京 100049；

[3]中国科学院红外探测与成像技术重点实验室，上海 200083

摘要 提出了一种用于空中红外目标检测的增强单发多框检测器(SSD)方法。分析了感受野与特征图层数的关系，同时采用池化和转置卷积操作的特征图双向融合机制，从整体上增强了特征的表达能力。通过引入浅层特征图的语义增强分支，并在高分辨率特征图上增加预测框，可提升小尺寸目标的定位精度。在 VOC2007 小目标和空中红外目标数据集上进行了对比测试，平均精度分别提高了 7.1%和 8.7%，此时检测速度略有下降。结果表明，增强 SSD 可在空中红外目标检测中获得较好的性能。

关键词 机器视觉；单发多框检测器；空中红外目标；目标检测；特征融合；语义分割

中图分类号 TP391 **文献标识码** A **doi**：10.3788/AOS201939.0615001

Enhancement of Single Shot Multibox Detector for Aerial Infrared Target Detection

Xie Jiangrong[1,2,3], Li Fanming[1,3*], Wei Hong[1], Li Bing[1], Shao Baotai[1,2,3]

[1] *Shanghai Institute of Technical Physics, Chinese Academy of Sciences, Shanghai* 200083, *China*;

[2] *University of Chinese Academy of Sciences, Beijing* 100049, *China*;

[3] *Key Laboratory of Infrared System Detection and Imaging Technology, Chinese Academy of Sciences, Shanghai* 200083, *China*

Abstract A method for enhancement of a single shot multibox detector (SSD) for aerial infrared target detection is proposed. Herein, the relationship between the sensing field and number of feature layers is analyzed, and a bidirectional feature map fusion mechanism that uses both pooling and deconvolution operations is proposed to enhance the feature expression ability. The semantic enhancement branch of the shallow feature map is introduced and the prediction boxes on the high-resolution feature map are increased, so that the positing accuracy of small-size targets is improved. Comparative experiments on the VOC2007 small object dataset and an aerial infrared target dataset reveal that the mean average precisions increase by 7.1% and 8.7%, respectively, accompanied by a slight decrease in detection speed. The results demonstrate that SSD enhancements can achieve good performance in aerial infrared target detection.

Key words machine vision; single shot multibox detector; aerial infrared target; target detection; feature fusion; semantic segmentation

OCIS codes 110.3080; 100.4996; 150.1135

1 引 言

目标检测一直是计算机视觉领域的研究重点和热点[1-2]，在无人驾驶、安防监控、机器人视觉、防空光电跟踪等领域具有不可替代的作用。在红外光电跟踪系统中，空域背景变化多样，并且空中目标占有的像素数较少、特征结构稀疏；另外，高机动性的飞行器难以捕捉定位，成像容易出现模糊失真的现象，加大了算法识别的难度。因此，有效的空中目标检测算法成为了提高跟踪精度和实时性的关键技术。

目标检测要求同时获得目标的类别信息和定位信息。传统的模式识别方法大多依赖先验特性来建

收稿日期：2018-12-18；**修回日期**：2019-01-25；**录用日期**：2019-02-19

基金项目：国家十三五国防预研项目(Jzx2016-0404/Y72-2)、上海市现场物证重点实验室基金(2017xcwzk08)

* **E-mail**：lfmjws@163.com

立数学模型并完成求解匹配，如帧差法[3]、光流法[4]、Hough变换法[5]等；另一种主流的方法采用特征算子[如SIFT(Scale-Invariant Feature Transform)[6]、方向梯度直方图[7]]加分类器(如支持向量机[8]、AdaBoost[9])的模式，对数据量要求较少，在某些项目中能获得不错的效果。随着深度网络在计算机视觉领域获得突破，卷积神经网络(CNN)被用于提取更高层、表达能力更强的特征，并形成两种主流的深度网络检测框架：一类是结合区域提名和CNN的基于分类的R-CNN(Regions with CNN features)系列目标检测框架；另一类是将目标检测算法转换为回归问题的单步算法。

Faster R-CNN采用RPN(Region Proposal Networks)结构，能够在一个网络框架内完成候选区域、特征提取、分类和定位修正等操作，真正实现了端到端的网络计算，检测精度达到73.2% mAP(mean average precision)/5 frames·s^{-1}(以下若无特殊说明，数据集均指VOC2007)[10]；YOLOv3(You Only Look Once)将图片划分成多个网格，在每个网格上一次性完成目标bounding box、置信度以及类别概率的预测，牺牲了精度但换取了检测速度的大幅度提升，最新版本引入多尺度特征融合的方式，对小目标检测效果提升明显，检测精度达到了57.9% mAP/20 frames·s^{-1}(COCO数据集)，基本代表了业界的最高水平[11]；单发多框检测器(SSD)结合了以上两者的优势，可以实现高准确率和实时检测[12]，在300 pixel×300 pixel分辨率时，准确率为74.3% mAP/59 frames·s^{-1}，在512 pixel×512 pixel分辨率时，SSD获得了80% mAP/19 frames·s^{-1}的结果，超过了Faster R-CNN，但仍存在小目标容易漏检、多个边界框重复检出的问题，针对此类问题，主要从改进网络结构和多尺度特征融合两个方向进行改善。全卷积网络的应用开创了语义分割的先河[13]，与之后的Mask-RCNN[14]都可以实现目标像素级别的分类，对于小目标能做出更精细的位置划分；另一方面，文献[15-16]借鉴了FPN(Feature Pyramid Networks)的思想，提出的FSSD(Feature fusion Single Shot multibox Detector)模型可以连接不同尺度的特征图，重构出一组特征金字塔，在小物体上获得了比原始SSD更高的检测精度；文献[17]中的DSSD(Deconvolutional Single Shot Detector)模型将基础网络从VGG换成了表征能力更强的ResNet，并引入反卷积层作为编解码器传递上下文信息，在小物体检测上实现了优异的效果，但是模型复杂度的增加削弱了实时性能；文献[18]提出改进的R-SSD模型，将特征图采用简单的连接和转置卷积两种方式进行融合，充分利用了特征图的方向信息，改进了小目标的检测能力；文献[19]以经典SSD模型为基础，从开辟多视窗结构出发，通过融合5个预设视窗的输出结果，在相应的小目标数据集上获得了一定的性能提升。

红外探测器的工艺水平限制了响应率和分辨率，与可见光图像相比，一定距离外的红外目标图像不具备丰富的纹理细节、边界比较模糊，只能利用热辐射几何分布和拓扑关系、飞行姿态与辐亮度的关联等浅层特征。本文基于SSD深度网络框架，针对空中大视场背景下，红外目标有效像素数少、特征稀疏，伴随云朵遮挡、大幅度的姿态变化等不利因素，提出了双向的特征图融合方法，增强了各尺度上特征图的表达能力；同时，引入语义分割支路，通过增强更浅层特征层的语义，获得更多的预测框。实验结果表明，在VOC2007的小目标数据集和空中红外数据集中，mAP分别增加了7.1%和8.7%。

2　SSD检测框架

SSD是一种能够直接预测目标类别和位置的多目标检测方法[12]。不同于Faster R-CNN等采用两阶段实现的深度检测框架，它采用了回归的思想来解决多目标的检测问题，大大地简化了网络结构和训练难度，提高了算法的实时性能；引入的anchors机制有利于提取多尺度、多比例的特征，相对于YOLO等全局特征提取的方法，获得了更高的定位精度。

2.1　SSD模型结构

SSD的检测模型主要由两部分组成：1)用于提取多层图像特征的深度CNN，位于整个结构的前端，可采用去除分类层后的图像分类网络，如VGG_16[20]、ResNet-101[21]等，其中原先骨干网络中的全连接层(FC6、FC7)替换为相应的卷积层(conv)，卷积核尺寸和输出通道数详见图1下方的数据标注；2)附加的级联网络，由几个卷积层和最后一层的均值池化层(Avg-pooling)构成，它针对前端获得的特征层，进一步提取出在不同尺寸条件下的特征信息；将多级特征同时送入检测器中，进行回归计算和极大值抑制后输出最终结果。SSD网络的结构框架如图1所示。图1中c_{lasses}为目标种类的数量。

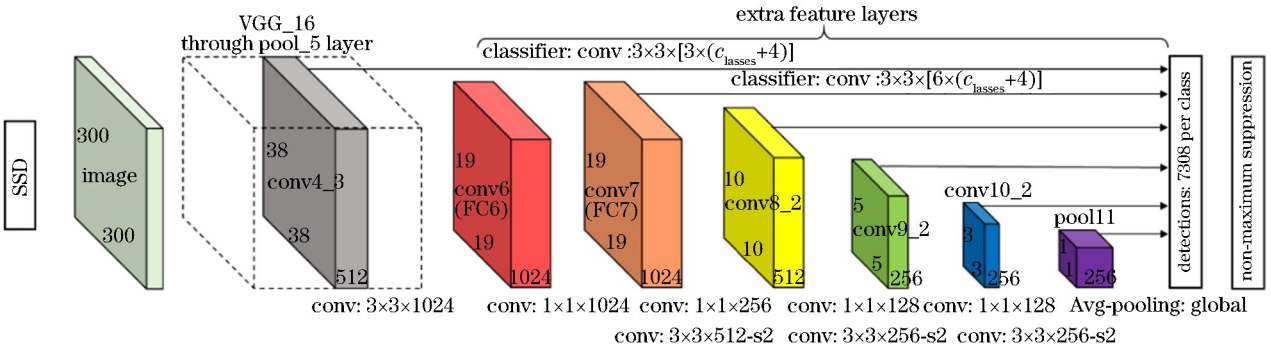

图 1　SSD 网络结构框图

Fig. 1　Structure of SSD network

2.2　特征层映射和损失函数

在 SSD 的基础网络结构之后，通过增加额外的卷积层，能够产生信息更全面的多尺度特征图，然后在各个特征图上分别进行目标预测。假设检测模型采用了 m 层特征图，则第 k 个特征图上默认框占输入图像尺寸的比例为

$$S_k=S_{\min}+\frac{S_{\max}-S_{\min}}{m-1}(k-1),\ k\in\{1,2,\cdots,m\},\tag{1}$$

式中：$S_{\min}$ 为特征层默认框占输入图像的最小比例，一般取 0.2；$S_{\max}$ 为默认特征框占输入图像的最大比例，一般取 0.9。

另外，SSD 引入了 Faster R-CNN 的 anchors 机制，在同一特征层上的默认框采用多个长宽比，能够增强默认框对不同形状物体的稳健性。通常采用 5 种高宽比：$r=\{r_n\}=\{1,2,3,1/2,1/3\}$（$n$ 为序号，$n=1,2,3,4,5$），当高宽比 $r_1=1$ 时，添加 $S_k'=\sqrt{S_kS_{k+1}}$，则所有默认框的宽 w_k^n 和高 h_k^n 可以分别表示为

$$w_k^n=S_k\sqrt{r_n},\ h_k^n=\frac{S_k}{\sqrt{r_n}},\ n=1,2,3,4,5。\tag{2}$$

设定默认框的中心坐标为 $\left(\frac{a+0.5}{|f_k|},\frac{b+0.5}{|f_k|}\right)$，其中，$|f_k|$ 为第 k 个特征图的尺寸大小，a 和 b 为锚点的横纵坐标，$a,b\in\{0,1,2,\cdots,|f_k|-1\}$，并截取默认框的坐标以保证其在 $[0,1]$ 的范围内。通过坐标转换，实现特征图上的默认框与原始图像的映射。

SSD 在训练时对边界位置和目标种类同时进行回归，其损失函数 $L(x,c,l,g)$ 定义为位置误差 L_{loc} 与置信度误差 L_{conf} 的加权和，即

$$L(x,c,l,g)=\frac{1}{N}[L_{\text{conf}}(x,c)+\alpha L_{\text{loc}}(x,l,g)],\tag{3}$$

式中：N 为先验框正样本的个数；$x\in\{0,1\}$ 表示预测框与真实框在某一类别上是否匹配 $x=0$ 表示不匹配，$x=1$ 表示匹配；c 为类别置信度的预测值；l 为先验框对应边界位置的预测值；g 为 ground truth 的位置参数；$L_{\text{conf}}(x,c)$ 为置信度误差，采用 softmax loss；$L_{\text{loc}}(x,l,g)$ 为位置误差，采用 Smooth-L1 loss；权重系数 α 通过交叉验证设置为 1。

2.3　经典 SSD 在目标检测中的不足

SSD 对不同的特征图取不同尺寸的先验框，通过回归得出目标类别的置信度和先验框与 ground truth 间的偏差。SSD 选取 IOU（Intersection Over Union）大于 0.5 的先验框作为正样本，小于 0.5 的先验框作为负样本，由于大尺寸物体覆盖 IOU 大于 0.5 的先验框多，此时正负样本数目均衡；而小尺度目标的正样本数较少，导致正负样本失衡难以训练。采用 Focal Loss 可以有效地解决上述问题[22]。

随着 CNN 层数的增加，特征图的维度越来越小，特征越来越抽象，语义特征越来越明显，而位置信息越来越模糊[23]。假设采用 SSD_300×300 的模型，即处理的图像分辨率为 300 pixel×300 pixel，其特征层为 conv4_3、conv7、conv8_2、conv9_2、conv10_2、conv11_2。小目标的检测，主要依赖于 conv4_3 的浅层特征，它具备 38 pixel×38 pixel 的高分辨率，包含的先验框尺寸也与目标尺寸较接近，但其特征表达能力仅来源于前 10 层的卷积层，不能捕捉深层次的语义信息；另一方面，随着卷积层数的增加，输入图像的感受野也迅速上升，卷积层感受野的计算公式如下：

$$S_{\text{RF}}(i)=[S_{\text{RF}}(i+1)-1]S_i+K_i,\tag{4}$$

式中：$S_{\text{RF}}(i)$ 为第 i 个卷积层的感受野大小；S_i 为累乘后的卷积步长；K_i 为卷积核的尺寸大小。

通过计算，得到各卷积层的感受野大小及各特征层默认框映射的图像区域如表 1 所示。可以发现

从特征层 conv9_2 往后，卷积感受野的大小就超过了原始图像的尺寸，即每个特征点均由整个图像作为输入产生响应，由此降低了精确定位目标的性能；另外，特征层的默认框在图像上的映射区域，在 conv9_2 中就已经超过了输入图像的一半，当该区域包含多个目标时，无法实现有效地区分。

表 1　SSD_300×300 卷积感受野、默认框映射图像区域

Table 1　Convolution receptive field and mapping image region of default boxes of SSD_300×300

Convolutional layer	Convolutional receptive field /(pixel×pixel)	Output scale of feature layer /(pixel×pixel)	Default boxes ratio	Mapping region scale /(pixel×pixel)
conv4_3	92×92	38×38	0.10	30×30
conv7	276×276	19×19	0.20	60×60
conv8_2	340×340	10×10	0.37	111×111
conv9_2	468×468	5×5	0.54	162×162
conv10_2	724×724	3×3	0.71	213×213
conv11_2	980×980	1×1	0.88	264×264

3　增强的 SSD 模型

为了提升 SSD 对小目标的检测能力，应尽量使用较高分辨率的特征图，并进一步挖掘浅层特征的表达能力，同时将浅层的细节特征和高层的语义特征结合起来，有利于为小目标提供精确的定位信息和目标类别信息。本文提出一种基于语义分割的浅层特征增强方法，以及双向的特征层融合机制，用于改进原始 SSD 的检测性能。

3.1　特征层双向融合

原始 SSD 在多层特征图上进行目标预测，虽然能够起到类似于图像金字塔的作用，但每次只利用了一层的特征图，浅层特征缺乏类别识别的语义信息，而深层特征随着感受野的增大，分辨率降低，不利于精确定位。为此，在 SSD 的主干网络的基础上，加入特征融合以提升检测性能。

在不同特征层合并的过程中，需要保持分辨率的一致。池化层结构简单，高分辨率特征层通过池化后，便获得降采样的特征；转置卷积与上采样的作用相同，但卷积核参数可在训练过程中调整，使上采样参数更加合理，对低分辨率特征进行转置卷积后，完成上采样。将统一分辨率的多层特征按通道连接，便获得融合后的多尺度特征图。

图 2 为多种特征融合方法示意图。当单独采用池化或者转置卷积操作用于特征层连接时，特征信息的传递只能从左往右或从右往左传播，方向都是单一的，无法利用其他方向上的信息[18]。本文采用双向融合的方法，在多个尺度上获得了包含基本模式信息以及高级语义信息的特征图。结合 SSD 网络结构的实际特征图分辨率情况，采用的 max-pooling 窗口为 2×2，步长为 2；转置卷积的卷积核为 3×3，步长为 2。

3.2　浅层特征增强分支

为了进一步提升 SSD 模型预测小目标的效果，将 75 pixel×75 pixel 高分辨率特征图添加进检测层，由于其拥有较少的语义信息，本文模型加入语义分割支路，用以增强浅层特征图的语义信息。分支结构如图 3 所示，模块的输入是 SSD 主干网络的 conv3_3 特征层，输出返回到主干网络替换原来的特征层，该过程通过默认框级别的 ground-truth 进行监督学习，不需要额外的标注数据[24]。

具体地，模块的输入 X 为 conv3_3 特征层，记为 $X\in\mathbf{R}^{C\times H\times W}$，其中 C 为输入特征图的通道数，H 和 W 为输入特征图的高和宽，语义分割采用边界框级的标记 ground-truth，记为 $G\in\{0,1,2,\cdots,N\}^{H\times W}$，$N$ 为类别数。语义分割网络采用 4 个空洞卷积层，得到中间结果 $g(X)$，之后用于产生两条分支结果。右侧分割支路使用 1×1 卷积之后，经过 softmax 层，将以上非线性变换记为 F，那么得到预测的语义分割结果为 $Y=F[g(X)]$，其中，$Y\in\mathbf{R}^{(N+1)\times H\times W}$ 满足：$Y\in[0,1]^{(N+1)\times H\times W}$，$\sum_{c'=0}^{N}Y_{c',h,w}=1$，其中 c' 为特征图的通道数量，h 和 w 为语义分割结果图中每个点的横纵坐标。$g(X)$ 的另一个分支用于生成带有语义信息的掩模，使用1×1 卷积，将该过程表示为 H，进而获得 $Z=H[g(X)]\in\mathbf{R}^{C\times H\times W}$，$Z$ 的尺寸和通道数与 X 相同，通过按元素相乘，得到语义信息丰富的 X'，$X'=X\odot Z$，$\odot$ 代表每个像素点的点乘运算。接下来用输出的 X' 替换原来的 X，用于后续预测。

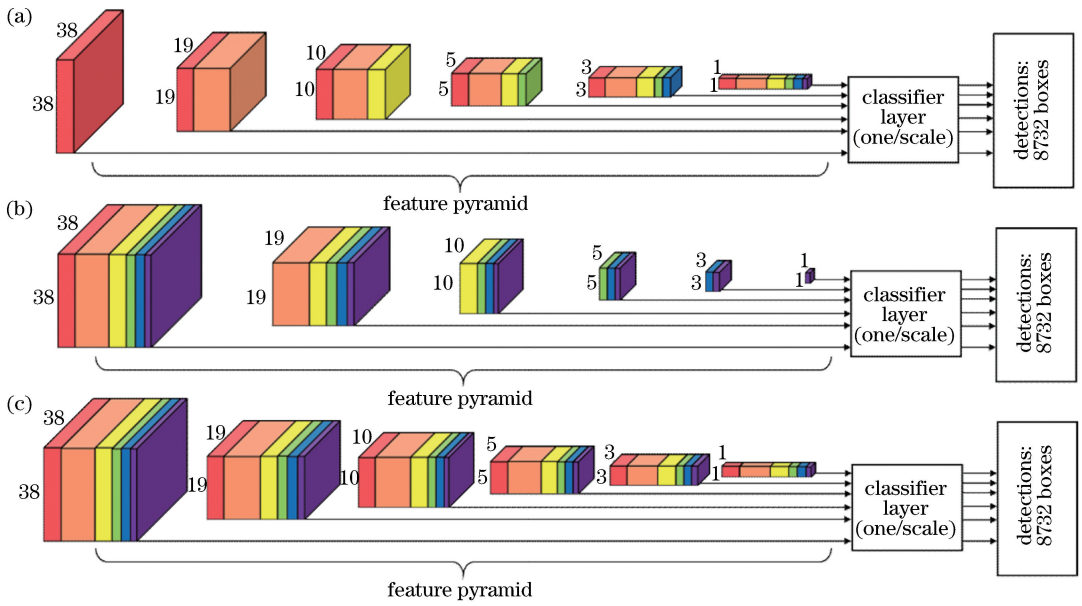

图 2　多种特征融合方法示意图。(a)池化;(b)转置卷积;(c)双向融合

Fig. 2　Schematics of multiple feature fusion methods. (a) Pooling; (b) transposed deconvolution; (c) bi-direction fusion

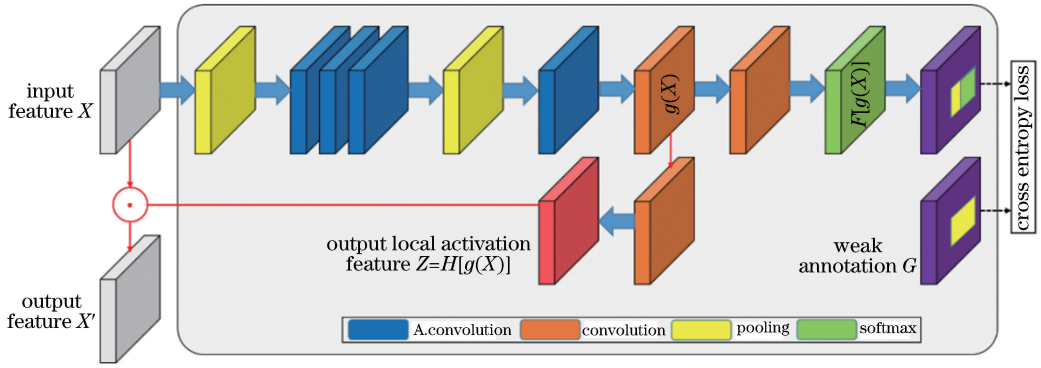

图 3　语义分割分支结构图

Fig. 3　Diagram of semantic segmentation branch

采用联合训练的策略,在目标检测损失函数 $L_{\mathrm{det}}(I,B)$ 上,加入语义分割任务的交叉熵损失函数:

$$L_{\mathrm{seg}}(I,G)=-\frac{1}{HW}\sum_{h,w}\ln[Y_{h,w}(G_{h,w})],\quad(5)$$

式中:I 为输入处理的图像。最终的损失函数定义如下:

$$L(I,B,G)=L_{\mathrm{det}}(I,B)+\beta L_{\mathrm{seg}}(I,G),\quad(6)$$

式中:B 为边界框的 ground-truth;β 为两个任务的平衡系数,本文模型取 0.1。

4　实验与分析

在实际的物体检测中,分类性能体现在预测框的置信度高低上,定位的准确性由预测框的坐标偏差衡量[25]。本文提出的改进模型,着重于改善小目标的检测,为验证本文算法的有效性,将其与原始 SSD 方法在具有代表性的小目标数据集上的 mAP 进行对比,并将置信度阈值统一设为 0.5。所采用的实验平台操作系统为 Ubuntu 16.04LTS,深度学习软件框架为 TensorFlow-GPU,CUDA-9.0,Cudnn-7.3.1,主要硬件配置如下:NVIDIA GTX1080Ti×2 GPU, Intel i7-8700k CPU,DDR4 32G Memory。

4.1　PASCAL-VOC2007 小目标数据集测试结果

为了对比本文提出的增强 SSD 与原始 SSD 模型的性能,挑选了 VOC2007 数据集中 137 张具有代表性的小目标图片,涉及的物体类别有 8 种,包括飞机(aero plane)、鸟(bird)、船(boat)、瓶子(bottle)、小汽车(car)、狗(dog)、羊(sheep)、人,经过相应处理之后,标注物体的 ground truth 共计

1164 个。分别采用原始 SSD 算法和本文方法进行目标检测实验，部分场景下的检测结果如图 4 所示，其中边界框左上角的数字分别代表类别标签和相应的置信概率。

从图 4 中可见本文方法对检测效果的改进。图 4(a)为原始 SSD_300×300 模型的检测效果，尽管其对近处物体取得较高的类别置信度，但是定位精确性有待进一步提高；图 4(b)为本文方法的检测效果，可以看出不仅大目标的定位更加精确，而且小目标的检测能力显著提升。表 2 为测试各类目标的结果，可见小尺寸物体(如：bottle)类别的平均精度(AP)提升最大，mAP 相比提升了 7.1%。

(a)　　(b)

图 4　原始 SSD 与改进 SSD 检测小目标的结果对比。(a)原始 SSD；(b)改进模型

Fig. 4　Comparison of detection results of small targets obtained by original SSD and improved SSD. (a) Original SSD; (b) improved model

表 2　VOC2007 数据集小目标检测结果

Table 2　Small object detection results of VOC2007 dataset

Method	mAP	Detection result							
		Aero plane	Bird	Boat	Bottle	Car	Dog	Sheep	Person
SSD_300×300	0.537	0.601	0.525	0.426	0.374	0.720	0.556	0.538	0.563
Proposed method	0.608	0.685	0.570	0.534	0.503	0.748	0.597	0.652	0.591

4.2　空中红外目标数据集测试结果

在空中红外目标识别中，特别是地基红外光电跟踪系统中，目标往往只占据少量的有效像素，另外云朵遮挡、大幅度的姿态变化等因素，也增加了目标跟踪识别的难度。自建的空中红外目标数据集筛选自外场实验视频流，分辨率为 320 pixel×256 pixel 和 640 pixel×512 pixel 的图像，包含了 J 型战斗机(fighter_J)、直升机(helicopter)、S 型战斗机(fighter_S)、民航客机(airliner)、飞鸟(bird)共 5 类目标，并以海天线、城市天际线、云朵背景为主。各类目标数量均衡，约为 550 张图，全部采用多目标标记，随机选取总数的 30%作为测试集，剩余的又划分为训练集和验证集。

采用 VGG_16 基础网络的卷积层参数，并在红外数据集上进行微调，部分场景检测结果如图 5 所示，可见，该方法对于大尺寸物体的定位偏差小、类别置信度普遍较高；在某些复杂场景下，如释放干扰弹时，仍能捕获目标本身的红外辐射特征，进行正确的识别。将该方法与原始 SSD、YOLOv3 进行对比，各类别的检测结果见表 3，最终 mAP 获得 8.7%的提升。

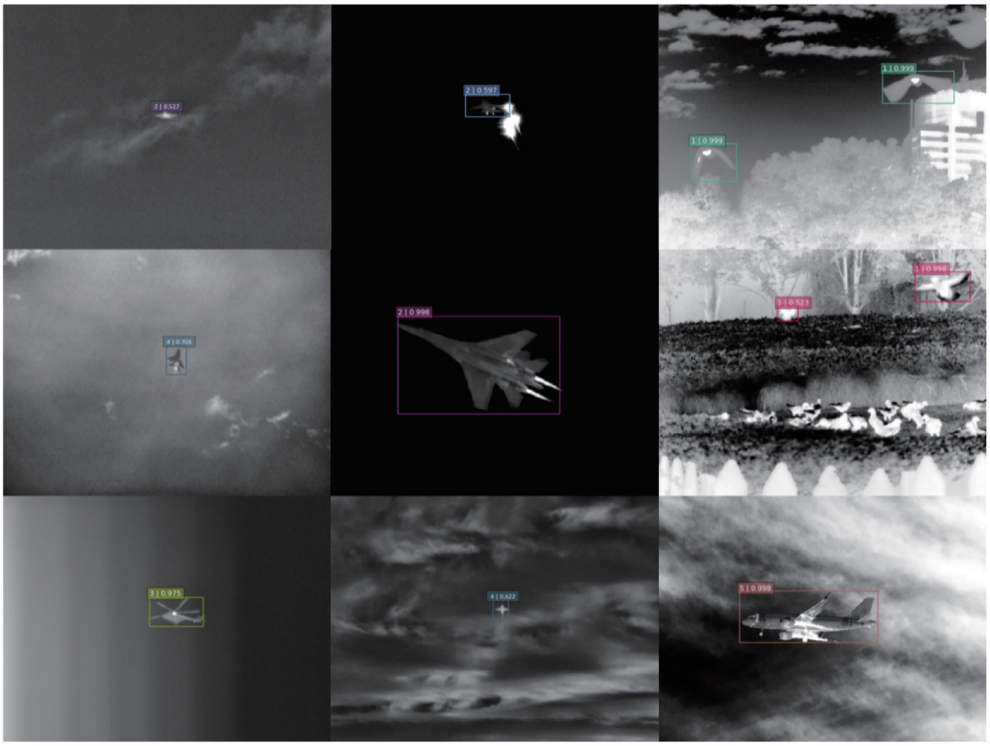

图 5　空中红外目标检测结果

Fig. 5　Detection results of infrared aerial targets

表 3　红外数据集空中目标检测结果

Table 3　Aerial target detection results of infrared dataset

Method	mAP	Detection result				
		Fighter_J	Helicopter	Fighter_S	Airliner	Bird
SSD_300×300	0.618	0.659	0.615	0.266	0.819	0.732
YOLOv3-320	0.641	0.730	0.548	0.314	0.867	0.747
Proposed method	0.705	0.784	0.636	0.485	0.822	0.796

为了进一步验证本文模型的有效性，利用召回率-准确率曲线对最终检测结果进行定量评价。准确率衡量被模型提取的目标中真实标记的比例，召回率反映了被正确提取的标记占总目标标记的比重[26]。图 6 为原始 SSD 模型、YOLOv3、本文增强 SSD 算法在空中红外目标数据集上的召回率与准确率的关系对比，其中召回率和准确率都是多个类别的平均数据。可见，相较于前面两种方法，增强 SSD 算法的召回率-准确率曲线所围成的面积更大，反映到数据上就是 AP 整体提升明显。

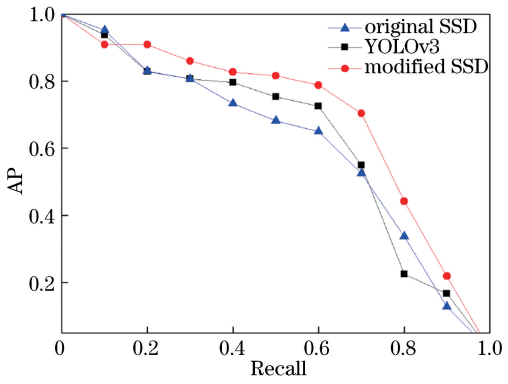

图 6　空中红外目标召回率-准确率曲线对比

Fig. 6　Comparison of recall-precision curve of infrared aerial targets

4.3　增强 SSD 的实时性分析

浅层特征的语义增强分支将更高分辨率的特征层引入检测中，不仅在语义增强过程中增加了卷积运算量，还扩大了预测框的总数，必然会导致预测速度的降低。但是，相对于简单的加深基础网络中特征层通道的方法，浅层特征的语义增强分支具有更少的网络参数。

双向特征融合采用最大池化和转置卷积，将高

低特征层归一化到相同尺寸，再通过串联获得良好的表征能力，并且没有带来太多的计算开销。通过拼接后的特征金字塔，每一层都有 3072 个通道（串联 256，512，1024，512，256，256，256 的通道），每个特征层具有相同通道数，为不同层之间共享分类网络权重提供了可能。进一步地，修改原始 SSD 模型中各分类网络中预测框数目，使得每个特征层统一，实现权重的共享。

对原始 SSD 和增强 SSD 方法的速度进行定量评估，为了测试的公平性，统一将 batchsize 设置为 8，同时去除了批归一化层，以缩短预测时间和减少内存消耗。表 4 为预测框总数和运行速度的对比，其中 FPS 代表算法运行的帧率，可见浅层预测框带来较多的时间消耗，代码实现还存在效率提升的空间。

表 4　每个分类网络的预测框数目及运行速度对比

Table 4　Number of predictive boxes for each classified network and speed comparison

Method	Number of boxes							Total boxes	FPS
	75×75	38×38	19×19	10×10	5×5	3×3	1×1		
Original SSD	0	4	6	6	6	4	4	8732	25.2
Modified SSD	4	4	4	4	4	4	4	30260	9.4

综上所述，增强后的 SSD 较原始 SSD 方法对小目标检测效果提升明显，在 VOC2007 的小目标数据集和空中红外数据集中，mAP 分别增加了 7.1%和 8.7%，在检测速率上 FPS 达到了 9.4 frames·s^{-1}，验证了本文算法能够实现精度和实时性的平衡。分析其主要原因，在于引入语义增强后的 conv3_3 层用于检测，同时增加了预测框的数量，增强了小目标的检测能力；另外，双向特征融合机制较好地结合了语义信息与定位信息，从整体上提升了物体检测性能。

5　结　论

首先阐述了经典 SSD 深度网络检测框架的结构和机理，并通过进一步分析卷积核感受野与默认特征框在输入图像上的映射关系，指出小目标检测能力不足的原因。在原始 SSD 模型基础上，引入浅层特征图语义增强分支，并提出了一种双向的特征融合机制，在 VOC2007 的小目标数据集和空中红外数据集上进行测试，得到的 mAP 分别提高了 7.1%和 8.7%，获得了较好的检测效果。未来的工作将围绕模型简化、参数压缩展开，以改善该模型的实时性能，并针对特定应用场景寻求更加合理的特征利用方式。

参　考　文　献

[1] Erhan D, Szegedy C, Toshev A, *et al*. Scalable object detection using deep neural networks[C]// 2014 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2014, Columbus, OH, USA. New York: IEEE, 2014: 2155-2162.

[2] Borji A, Cheng M M, Jiang H Z, *et al*. Salient object detection: a benchmark [J]. IEEE Transactions on Image Processing, 2015, 24(12): 5706-5722.

[3] Singla N. Motion detection based on frame difference method[J]. International Journal of Information & Computation Technology, 2014, 4(15): 1559-1565.

[4] Horn B K P, Schunck B G. Determining optical flow [J]. Artificial Intelligence, 1981, 17(1/2/3): 185-203.

[5] Barinova O, Lempitsky V, Kholi P. On detection of multiple object instances using hough transforms[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(9): 1773-1784.

[6] Lowe D G. Distinctive image features from scale-invariant keypoints [J]. International Journal of Computer Vision, 2004, 60(2): 91-110.

[7] Dalal N, Triggs B. Histograms of oriented gradients for human detection [C]//2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR′05), June 20-25, 2005, San Diego, CA, USA. New York: IEEE, 2005: 886-893.

[8] Burges C J C. A tutorial on support vector machines for pattern recognition [J]. Data Mining and Knowledge Discovery, 1998, 2(2): 121-167.

[9] Hastie T, Rosset S, Zhu J, *et al*. Multi-class AdaBoost[J]. Statistics and Its Interface, 2009, 2(3): 349-360.

[10] 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.

[11] Redmon J, Farhadi A. YOLOv3: an incremental improvement[EB/OL]. (2018-04-08)[2018-12-01]. https://arxiv.org/abs/1804.02767.

[12] Liu W, Anguelov D, Erhan D, *et al*. SSD: single shot multibox dtector[M]//Leibe B, Matas J, Sebe N, *et al*. Computer Vision: ECCV 2016. Cham: Springer, 2016, 9905: 21-37.

[13] Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 640-651.

[14] He K M, Gkioxari G, Dollár P, *et al*. Mask R-CNN [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018: 2844175.

[15] Li Z X, Zhou F Q. FSSD: feature fusion single shot multibox detector[EB/OL]. (2018-05-17)[2018-12-01]. https://arxiv.org/abs/1712.00960.

[16] Cao G M, Xie X M, Yang W Z, *et al*. Feature-fused SSD: fast detection for small objects[J]. Proceedings of SPIE, 2018, 10615: 106151E.

[17] Fu C Y, Liu W, Ranga A, *et al*. DSSD: deconvolutional single shot detector[EB/OL]. (2017-01-23)[2018-12-01]. https://arxiv. org/abs/1701.06659.

[18] Jeong J, Park H, Kwak N. Enhancement of SSD by concatenating feature maps for object detection[EB/OL]. (2017-05-26)[2018-12-01]. https://arxiv.org/abs/1705.09587.

[19] Tang C, Ling Y S, Zheng K D, *et al*. Object detection method of multi-view SSD based on deep learning[J]. Infrared and Laser Engineering, 2018, 47(1): 126003.
唐聪, 凌永顺, 郑科栋, 等. 基于深度学习的多视窗SSD目标检测方法[J]. 红外与激光工程, 2018, 47(1): 126003.

[20] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2015-04-10)[2018-12-01]. https://arxiv. org/abs/1409.1556.

[21] He K M, Zhang X Y, Ren S Q, *et al*. Deep residual learning for image recognition [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 770-778.

[22] Lin T Y, Goyal P, Girshick R, *et al*. Focal loss for dense object detection [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017: 2999-3007.

[23] Xin P, Xu Y L, Tang H, *et al*. Fast airplane detection based on multi-layer feature fusion of fully convolutional networks [J]. Acta Optica Sinica, 2018, 38(3): 0315003.
辛鹏, 许悦雷, 唐红, 等. 全卷积网络多层特征融合的飞机快速检测[J]. 光学学报, 2018, 38(3): 0315003.

[24] Zhang Z S, Qiao S Y, Xie C H, *et al*. Single-shot object detection with enriched semantics[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE, 2018: 5813-5821.

[25] Feng X Y, Mei W, Hu D S. Aerialtarget detection based on improved faster R-CNN[J]. Acta Optica Sinica, 2018, 38(6): 0615004.
冯小雨, 梅卫, 胡大帅. 基于改进Faster R-CNN的空中目标检测[J]. 光学学报, 2018, 38(6): 0615004.

[26] Wang W X, Fu Y T, Dong F, *et al*. Infrared ship target detection method based on deep convolution neural network[J]. Acta Optica Sinica, 2018, 38(7): 0712006.
王文秀, 傅雨田, 董峰, 等. 基于深度卷积神经网络的红外船只目标检测方法[J]. 光学学报, 2018, 38(7): 0712006.