基于感受野的快速小目标检测算法

王伟锋*，金杰，陈景明
天津大学电气自动化与信息工程学院，天津 300072

摘要 现有的高精度目标检测算法依赖于超深的主干网络(如 ResNet 和 Inception)，无法满足实时目标检测场景的需要，相反采用轻量级主干网络(如 VGG-16 和 MobileNet)能达到实时目标检测的目的，但会导致检测精度的损失，对小目标的检测效果变差。SSD(Single Shot Multi-Box Detector)算法具有高精度、实时检测的特点。本文以 SSD 算法的网络结构为基础，通过添加感受野模块增强轻量级主干网络的特征提取能力，同时引入特征融合模块，充分利用深层网络提取语义信息，达到实时目标检测的目的，同时提高算法整体的检测精度和对小目标的检测能力。为进一步验证引入新模块的有效性，本文算法模型在 PASCAL VOC2007 数据集上进行测试，准确率达到 80.5%，相比于原始 SSD 算法有 3.3 个百分点的提升，检测速度达到 75 frame/s，整体性能优于目前大多数目标检测算法。

关键词 机器视觉；目标检测；感受野；特征融合
中图分类号 TP391 **文献标识码** A **doi**：10.3788/LOP57.021501

Rapid Detection Algorithm for Small Objects Based on Receptive Field Block

Wang Weifeng*, Jin Jie, Chen Jingming
School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, *China*

Abstract Existing high precision object detection algorithms mostly rely on super deep backbone networks, such as ResNet and Inception, making it difficult to meet real-time detection requirements. On the contrary, some lightweight backbone networks, such as VGG-16 and MobileNet, fulfill real-time processing but their accuracies are often criticized, especially when the targets are small. In this study, we explore an alternative to build a fast and accurate detector by strengthening the feature extraction ability of lightweight backbone networks, using a new receptive field block based on a single shot multi-box detector (SSD). Simultaneously, to make full use of the semantic information extracted from deep networks, a feature fusion module is designed and added, thereby improving the overall accuracy and enhancing the detection effect of the model for small targets, while still achieving real-time detection. To further verify the validity of introducing new modules, we have tested our model on the PASCAL VOC2007 data set and achieved an accuracy of 80.5% which is 3.3 percentage points higher than that of the original SSD model. In addition, the detection speed of the proposed model reaches 75 frame/s, and its overall performance is better than that of most of the current models.

Key words machine vision; object detection; receptive field block; feature fusion
OCIS codes 150.1135; 100.3008; 040.1880

1 引 言

目标检测一直是计算机视觉领域的重要研究方向之一，随着深度学习在图像分类领域大放异彩，利用卷积神经网络(CNN)实现目标检测成为一个新的研究热点。相比于传统目标检测算法，基于深度学习的目标检测算法克服了区域选择策略差、时间复杂度高、手工提取特征稳健性差等难题，使检测效

收稿日期：2019-04-29；**修回日期**：2019-05-29；**录用日期**：2019-06-13
基金项目：国家自然科学基金(61571320)
* **E-mail**：weifeng_17@tju.edu.cn

果得到显著提升。基于不同 CNN 的目标检测算法可分为两类：一类是基于区域提取的目标检测算法，包括 R-CNN(Regions with CNN features)，Fast R-CNN 和 Faster R-CNN 等；另一类是基于回归的目标检测算法，包括 YOLO(You Only Look Once)和 SSD(Single Shot Multi-Box Detector)。

2014 年，Girshick 等[1]提出 R-CNN 算法，该算法在特征提取阶段选用 CNN 替代手动设计特征，提高了检测精度，但其他阶段仍采用传统方法，导致算法训练和测试效率低，而且全连接层需要严格保证输入的 proposal 尺度相同，在一定程度上造成图像的畸变，影响检测精度。2015 年 Girshick[2]提出 Fast R-CNN，该算法添加 ROI(Region of Interest)池化层，用 Softmax 替换支持向量机(SVM)分类器，并设计多任务损失函数，将分类任务和边框回归统一到一个框架中，这提高了检测速度，但提取 proposal 阶段仍用选择性搜索，使得时间复杂度高，无法满足实时应用。2015 年 Ren 等[3]提出 Faster R-CNN，该算法利用区域候选网络(RPN)取代耗时的选择性搜索，使检测速度大幅提高，但仍没实现真正意义上端到端的训练测试。2016 年，Redmon 等[4]提出实时的目标检测框架 YOLO，该框架将输入图片分成一定大小的网格，使用单个前馈卷积网络直接预测对象类别和位置，实现了端到端的训练和测试，YOLO 的检测速度比基于区域提取的检测算法快，但定位精度和分类的精度较差。2016 年，Liu 等[5]提出 SSD 算法，该算法融合 YOLO 的网格离散化思想和 Faster R-CNN 固定框思想，通过增加多尺度特征图上的映射能力来进行不同尺度的预测，该算法兼具 YOLO 检测速度快和 Faster R-CNN 检测精度高的优势。

改进用于特征提取的 CNN 和特征融合是提升目标检测算法准确率的两种重要方式，文献[6-7]在特征提取阶段利用更深的卷积神经网络(ResNet[8]和 Inception[9])增强 CNN 的特征提取能力，从而获得更丰富的特征信息；文献[10-11]将基础网络改为 ResNet，同时分别提出利用特征金字塔模块和反卷积模块，引入空间上下文信息，实现 CNN 中深层卷积和浅层卷积提取特征信息的融合。上述算法通过更深的 CNN 提升了算法准确率，但计算量成倍增加，无法满足实时检测场景的要求，因此进一步研究在轻量级主干网络的基础上引入新型的卷积和特征融合机制，对实现快速、准确的目标检测至关重要。本文利用 SSD 算法(主干网络采用 VGG-16[12])作为基础的网络框架，通过引入新型的感受野模块，强化主干网络的特征提取能力，另一方面，引入特征融合模块增强算法对小目标的检测能力。实验表明，本文算法的检测速度接近 SSD 算法，满足实时检测的要求，同时检测准确率有明显提升。

2 算法框架

2.1 感受野模块

感受野模块采用多支路卷积的形式，其内部结构可分为多支路卷积层和膨胀卷积层。多支路卷积层的结构和 Inception 相同，能模拟不同尺寸的感受野；膨胀卷积层利用膨胀卷积模拟不同尺寸感受野之间的关系。在 CNN 中，卷积的感受野由卷积核的尺寸大小决定，通过设置多个尺寸的卷积核能获得不同尺寸的感受野，从而更有效地利用特征信息。本文多支路卷积的设计借鉴了最新的 Inception-V4 和 Inception-ResNet[7]，每条支路采用 1×1 卷积来减少特征通道数，采用 $1\times n$ 卷积加 $n\times1$ 卷积替换 $n\times n$ 卷积来减少卷积的参数量，同时增强卷积的非线性表达能力，此外还增加一条支路用来保留更多的原始特征信息。

膨胀卷积也叫空洞卷积，最早在 Deeplab[13]中提出，旨在不增加卷积参数的前提下增大卷积感受野，被广泛应用于背景分割和目标检测领域。针对传统的 CNN 中采用池化层造成的内部数据结构丢失和小物体信息无法重建等问题，文献[14]提出混合膨胀卷积，通过叠加多个特定膨胀率的膨胀卷积以避免网格效应和平衡不同尺寸感受野之间的关系(小膨胀率对应小尺寸感受野信息，大膨胀率对应大尺寸感受野信息)，可以和多支路卷积的设计很好地结合，从而提高算法的检测效率。

本文感受野模块的网络结构如图 1 所示，该模块借鉴混合膨胀卷积和 Inception，膨胀卷积层位于不同尺寸卷积层之后，目的是模拟人类视觉皮层对不同尺寸感受野信息敏感程度的不同特点，最后不同支路提取的特征信息经过融合得到基于感受野信息的空间矩阵。

2.2 特征融合模块

SSD 算法基础主干网络采用 VGG-16，该网络添加四个小尺寸卷积层组成金字塔形式的预测结构，利用浅层卷积(Conv4_3)提取的特征预测小目标，深层卷积 (Conv7，Conv8_2 等) 提取的特征预测大目标，降低模型的预测负担。一方面 SSD 算法只利用单一浅层卷积(Conv4_3)提取的特征来预测

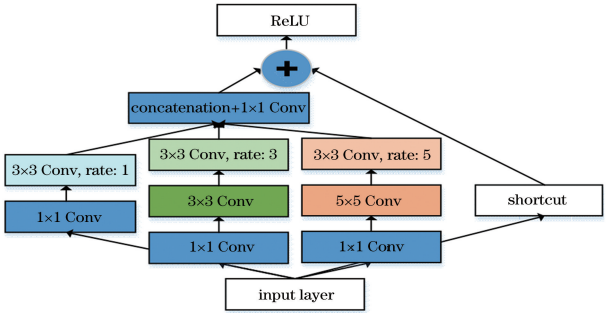

图 1 感受野模块的结构示意图

Fig. 1 Structural diagram of receptive field block

小目标，另一方面浅层卷积提取的特征信息缺少小目标相关的语义信息，导致 SSD 算法对小目标的检测效果较差，因此将深层卷积提取的语义信息回传到浅层卷积，提高算法对小目标的检测性能。文献[15]通过反卷积的方式对SSD算法中不同深度的卷积层（Conv5_3，Conv6，Conv7 等）提取的特征信息进行可视化，可以看到越深层卷积提取的特征包含语义信息越丰富，但无关的背景噪声也增多。对于图像中的小目标而言，Conv5_3 提取的特征信息相较于 Conv4_3 包含更多的上下文信息，同时相较于 Conv6，Conv7 存在的背景噪声少，因而更加适合与 Conv4_3 卷积层提取的特征信息进行融合，从而提升算法的检测效果。

特征融合模块如图 2 所示，在 Conv5_3 后引入反卷积层使得 Conv5_3 卷积层提取的特征图尺寸和 Conv4_3 卷积层提取的特征图尺寸相同。在 Conv5_3 和 Conv4_3 之后添加两个 3×3 卷积层和两个不同尺度的归一化层，以便于更好地学习融合特性，最后通过 1×1 卷积实现两个特征层之间的信息融合，同时降低特征信息的通道数。

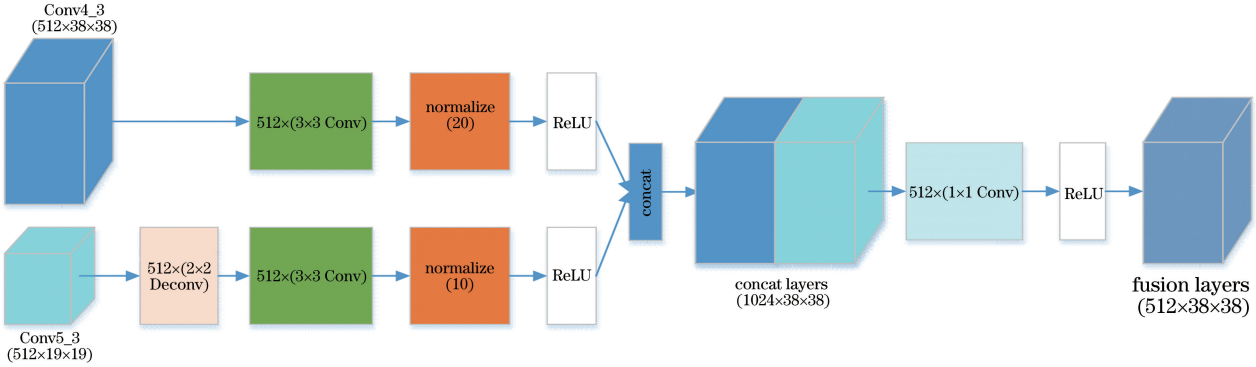

图 2 特征融合模块结构示意图

Fig. 2 Structural diagram of feature fusion module

2.3 算法整体框架

本文算法以 SSD 框架为基础，通过在特征提取网络上嵌入感受野模块来增强网络的特征提取能力，同时嵌入特征融合模块，充分利用深层卷积提取的特征信息，使提高算法整体检测精度的同时改善小目标的检测效果。感受野模块和特征融合模块结构相对简单，易嵌入到 SSD 网络框架，因此可以尽可能地保留 SSD 原有的网络结构，发挥检测速度的优势。整体算法的具体网络结构如图 3 所示，图中“RFB”代表 2.1 节介绍的感受野模块。网络的主要改进有 3 点：一是利用新添加的感受野模块对 Conv4_3 和 Conv7 层得到的高分辨率特征信息进

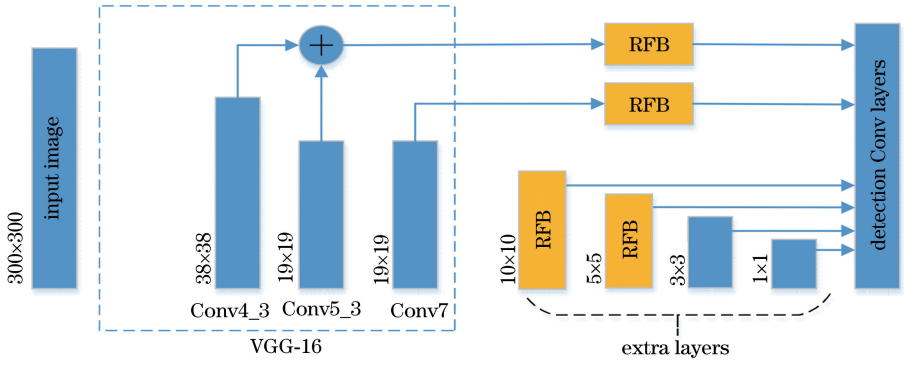

图 3 本文算法的网络结构

Fig. 3 Network structure of proposed algorithm

行预测；二是用感受野模块替换 Conv8_2 和Conv9_2 层，由于最后两个卷积层尺寸的限制，保留了原本的结构；三是为充分利用特征图信息，在 Conv5_3 和 Conv4_3 基础上添加了新的特征融合模块。

3 实验结果分析

本文实验模型基于 Pytorch-0.4.0 框架，实验操作系统为 Ubuntu 16.04.5 LTS，CPU 型号为 Intel(R)Xeon(R)E5-2678v3，主频为 1.2 GHz，核心数量为 48，GPU 型号为 GTX1080Ti，显卡型号为 NVIDIAQuadroP4000，GPU 数量为 2，系统内存为 32G，CUDA 版本为 9.0。训练的策略与 SSD 算法相近，数据集使用 PASCALVOC2007 和 PASCALVOC2012，两个数据集分别包括 9963 张和 22531 张图片，总共 20 类物体。训练过程设置输入图像大小为 300 pixel×300 pixel，训练的批次大小设置为 16，max_epoch 设置为 250，训练过程中为保证训练损失平稳下降，前 5 个 epoch 的初始学习率为 4×10^{-3}，5～150 个 epoch 为 10^{-3}，150～200 个 epoch 为 10^{-4}，200～250 个 epoch 为 10^{-5}。权重衰减为 5×10^{-4}，冲量为 0.9。

在 VOC2007 测试数据集上，将本文算法模型和主流的目标检测算法进行对比，结果如表 1 所示，本文算法的测试准确率(mAP)达到 80.5%，相比于 SSD 算法和 DSSD 算法分别有 3.3 个百分点和 1.9 个百分点的提升，达到了与 R-FCN 相同的准确率。从模型的运行速度来看，本文算法模型达到 75 frame/s，相对于 DSSD 和 R-FCN 有明显优势，满足实时检测的需求。

表 1 不同目标检测算法在 PASCAL VOC 2007 上的测试结果对比

Table 1 Comparison of detection results on PASCAL VOC 2007 test set for different object detection algorithms

Method	Backbone	Dataset	mAP /%	Frame rate /(frame·s^{-1})
R-FCN[16]	ResNet-101	VOC07+12	80.5	9
Fast YOLO[4]	Google Net-9	VOC07+12	52.7	163
YOLO[4]	VGG-16	VOC07+12	66.4	96
YOLO-V2[17]	DarkNet-19	VOC07+12	78.8	80
DSSD321[9]	ResNet-101	VOC07+12	78.6	9.5
SSD300[5]	VGG-16	VOC07+12	77.2	120
Proposed (300)	VGG-16	VOC07+12	80.5	75

进一步分析添加新模块对算法检测效果的影响，对新添加的模块进行消融实验，结果如表 2 所示，增加感受野模块可以将算法检测准确率提高 2.9 个百分点，表明增大卷积感受野能提升算法的检测性能。相应地，由于感受野采用多个支路卷积，增加了模型的复杂度，检测速度比原来降低 37 frame/s；另一方面，增加特征融合模块将算法在测试集的准确率提升 1.7 个百分点，从模型运行速度来看，这得益于特征融合模块的结构较为简单，检测速度只比原来降低 16 frame/s。本文算法将感受野模块和特征融合模块同时嵌入到 SSD 算法框架当中，在提升算法检测准确率的同时保证了算法的实时性。

表 2 不同模块对算法准确率的影响

Table 2 Influences of different modules on accuracy of algorithms

Algorithm	SSD	SSD+RFB	SSD+Fusion	Proposed model
Add RFB?	×	√	×	√
Add Fusion?	×	×	√	√
mAP /%	77.2	80.1	78.9	80.5
FPS	120	83	104	75

对比融合不同深度卷积层提取的特征信息对检测效果的影响，选择两种不同的融合方式，结果如表 3所示，融合深层卷积(Conv5_3，Conv6)提取的特征信息可以提高算法检测准确率，融合相邻卷积层提取的特征信息(Conv5_3)对算法准确率的提升更加明显，原因在于 Conv6 卷积层提取的特征信息包含更多的背景噪声。

表 3 不同融合方式的测试结果

Table 3 Detection results of different fusion ways

Layer	mAP /%
Conv4_3	80.1
Conv4_3+Conv5_3	80.5
Conv4_3+Conv6	80.2

从实际应用场景出发，小目标检测是目标检测算法应用的一项重要需求，为验证本文算法对小目标检测效果的提升，从测试集中选取 4 张包含众多小目标的图片进行测试，结果如图 4 所示，本文算法在检测到的小目标个数以及对小目标分类的准确率两个方面都要优于传统的 SSD 算法。

4 结 论

提出一种新型高效率的目标检测算法。与大多数目标检测算法依赖于超深的主干网络不同，本文算法以轻量级主干网络为基础，通过添加感受野模

图 4 小目标检测效果对比。(a) SSD 算法探测结果;(b)本文算法探测结果

Fig. 4 Comparison of detection results of small objects. (a) Detection results of SSD algorithm; (b) detection results of proposed algorithm

块,融合不同尺寸感受野的特征,充分利用卷积神经网络提取的特征信息,提高了算法整体的检测效果;另外新添加特征融合模块,进一步改善了算法对小目标的检测效果。通过对比实验可以发现,本文算法在 PASCAL VOC2007 数据集上的测试整体性能要优于传统的 SSD 算法和大多数主流目标检测算法,满足实时检测的要求,明显提升了对小目标检测的效果。

参 考 文 献

[1] 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, 2014: 580-587.

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

[3] Ren S Q, He K M, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks [C] // Advances in Neural Information Processing Systems 28 (NIPS 2015), December 7-12, 2015, Montreal, Quebec, Canada. Canada: NIPS, 2015: 91-99.

[4] 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, 2016: 779-788.

[5] 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.

[6] Chen L L, Zhang Z D, Peng L. Real-time detection based on improved single shot MultiBox detector [J]. Laser & Optoelectronics Progress, 2019, 56(1): 011002.
陈立里, 张正道, 彭力. 基于改进 SSD 的实时检测方法[J]. 激光与光电子学进展, 2019, 56(1): 011002.

[7] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, inception-ResNet and the impact of residual connections on learning [C] // Thirty-First AAAI Conference on Artificial Intelligence, February 4-10, 2017, San Francisco, California, USA. USA: AIAA, 2017: 4278-4284.

[8] 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.

[9] Szegedy C, Liu W, Jia Y Q, et al. Going deeper with convolutions [C] // 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE, 2015: 15523970.

[10] Ren Z J, Lin S Z, Li D W, et al. Mask R-CNN object detection method based on improved feature pyramid [J]. Laser & Optoelectronics Progress, 2019, 56(4): 041502.
任之俊, 蔺素珍, 李大威, 等. 基于改进特征金字塔的 Mask R-CNN 目标检测方法[J]. 激光与光电子学

进展, 2019, 56(4): 041502.

[11] Fu C Y, Liu W, Ranga A, et al. Dssd: deconvolutional single shot detector[J/OL]. (2017-01-23)[2019-04-28]. https://arxiv.gg363.site/abs/1701.06659.

[12] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J/OL]. (2015-04-10)[2019-04-28]. https://arxiv.gg363.site/abs/1409.1556.

[13] Chen L C, Papandreou G, Kokkinos I, et al. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834-848.

[14] Chen L C, Papandreou G, Schroff F, et al. Rethinking atrous convolution for semantic image segmentation[J/OL]. (2017-12-05)[2019-04-28]. https://arxiv.gg363.site/abs/1706.05587.

[15] 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.

[16] Dai J F, Li Y, He K M, et al. R-FCN: object detection via region-based fully convolutional networks[C]// Advances in Neural Information Processing Systems 29 (NIPS 2016), December 5-10, 2016, Barcelona, Spain. Canada: NIPS, 2016: 379-387.

[17] 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, 2017: 6517-6525.