第 58 卷 第 12 期/2021 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

基于上下文自校准双重注意力的目标检测算法

罗浚铠，张宝华*，张艳月，谷宇，王月明，刘新，任彦，李建军，张明

内蒙古科技大学信息工程学院，内蒙古 包头 014010

摘要 基于先验框设计(anchor-based)的多类目标检测算法存在超参数多、泛化能力差、正负样本不平衡的问题。针对这些问题，提出一种基于改进无锚(anchor-free)方法的目标检测算法。首先，针对传统算法在多类目标检测任务中难以获得鲁棒的特征表达的问题，构建基于上下文结合的自校准双重注意力模块，通过混合空洞卷积组获取多感受野信息；然后以低维空间嵌入的方式进行自校准获取上下文空间信息；最后将空间信息与通道信息结合，增强算法特征表达能力。针对在多类目标检测任务中由于目标尺度变化大、外观不规则而易引入背景噪声的问题，利用改进的可变卷积，对目标进行自适应采样。在目标检测数据集 MSCOCO 上的实验结果表明，所提算法能有效提升目标检测精度，优于对比检测算法。

关键词 图像处理；目标检测；上下文自校准；双重注意力机制；可变卷积；anchor-free

中图分类号 TP391.4 **文献标志码** A **doi**：10.3788/LOP202158.1210013

An Object Detection Algorithm Based on Contextual Self-Calibration And Dual-Attention Mechanism

Luo Junkai, Zhang Baohua*, Zhang Yanyue, Gu Yu, Wang Yueming, Liu Xin, Ren Yan, Li Jianjun, Zhang Ming

College of Information Engineering, Inner Mongolia University of Science & Technology, Baotou, Inner Mongolia 014010, *China*

Abstract To prevent numerous hyperparameters and to overcome poor generalization ability and imbalance between positive and negative samples in anchor-based multiclass object detection algorithms, an object detection algorithm based on an improved anchor-free method is proposed herein. To address the difficulty faced by traditional algorithms in obtaining robust feature representations in multiclass object detection tasks, a self-calibration dual-attention module based on contextual combination is first constructed herein. It obtains the multireceptive field information through a mixed dilated convolution group. Then, a low-dimensional spatial embedding method is self-calibrated to obtain the contextual spatial information. Finally, the spatial information and channel information are combined to enhance the feature representation ability of the proposed algorithm. To prevent the usual introduction of background noise owing to large changes of object scale and irregular appearance in multiclass object detection tasks, the improved deformable convolution is used to adaptively sample the target position. Experimental results obtained using the large multiclass object detection data set MSCOCO show that the proposed algorithm can effectively improve the detection accuracy of multiclass object and outperforms the existing detection algorithms.

Key words image processing; object detection; contextual self-calibration; dual-attention mechanism; deformable convolution; anchor-free

OCIS codes 100.4996; 100.3008; 040.1880; 150.1135

收稿日期：2020-09-09；**修回日期**：2020-09-23；**录用日期**：2020-09-30

基金项目：国家自然科学基金(61962046，61663036，61841204)、内蒙古杰青培育项目(2018JQ02)、内蒙古科技计划(202001)、内蒙古青年科技创新人才项目(第一层次)、内蒙古自治区自然科学基金(2015MS0604)、内蒙古自治区高等学校科学技术研究项目(NJZY145)

* **E-mail**：zbh_wj2004@imust.cn

1 引 言

近年来，深度学习凭借能自适应提取目标不同层级特征的特性，被广泛应用于多类目标检测领域。基于深度学习的目标检测可分为基于锚的(anchor-based)和无锚的(anchor-free)。其中基于锚的方法有 Faster R-CNN[1]、SSD[2]、YOLOv2[3]、YOLOv3[4]等，这些方法依赖于一系列预先设定的锚框，且一组成功锚框的设置对于提升多类目标检测性能有至关重要的作用。但是基于锚的方法有 3 个缺点：1)引入大量额外的超参数，即锚框的尺度、长宽比、数量设置，这些超参数的设置依赖于先验经验；2)泛化能力差，由于锚框是预先设定的，所以当算法在不同应用中切换时，由于目标形状的变化较大而无法获得良好的检测效果，甚至无法检测；3)存在大量的冗余框，即在大量预设的锚框中，只有很小一部分被标记为正样本，这会导致样本间的不平衡。针对上述问题，基于无锚的方法被学者提出。Law 等[5]提出了 CornerNet，不再以 anchor 机制预先设置锚框，而是通过目标物体左上角与右下角进行预测，通过角点插值算法组合两个角点并修正，从而完成多类目标的定位，即通过关键点进行检测定位。Zhou 等[6]针对 CornerNet 中角点语义信息不足的缺点，提出 ExtremeNet，引入上、下、左、右 4 个方位的极值点协助预测，以获得更明确的特征表达。不同于上述通过关键点进行多类目标检测的方法，Huang 等[7]提出 Densebox，在特征图上，直接通过空间点到物体边框的四个边界距离进行框的预测回归，而 Tian 等[8]提出的 FCOS 算法在此基础上通过引入特征金字塔[9]在不同特征层处理不同尺度目标，通过多尺度检测提升检测性能，并利用 Center-ness 模块进一步过滤多余的目标框，以缓解正负样本不平衡的问题。

由于目标尺度跨度大，且部分目标具有不规则的轮廓，会增加有用特征提取的难度，如何获得更为鲁棒的特征表达成为多类目标检测研究的重点和难点。文献[10]通过分别整合特征图中不同通道间的分类特征信息、空间中的位置信息，提升特征表达能力。文献[11]通过使用不同方法进行高层特征与低层特征的融合，增强输出特征中的语义信息与边缘信息。文献[12]通过构建预测优化模块，整合感兴趣区域的上下文信息，增强网络特征表达能力。文献[13]通过构建多感受野模块，增强特征尺度不变性。受到上述方法的启发，本文以 FCOS 算法为框架，针对 FCOS 算法在多类目标检测任务中特征表达能力不足的问题，提出上下文自校准双重注意力模块(CSDAM)，以获得更为鲁棒的特征表达能力；同时针对算法对不规则轮廓物体的特征提取能力差，易引入背景噪声的缺点，利用改进的可变卷积算法(DCNV2)[14]更好地提取具有不规则轮廓的物体的特征，削弱背景噪声影响。实验结果表明，所提算法能够有效提升多类目标检测效果，在大型多类目标检测数据集 MSCOCO 上检测精度高。

2 算法原理

2.1 上下文自校准双重注意力模块

将上下文自校准双重注意力模块分为多核通道注意力和上下文自校准空间注意力两个部分，分别获取用于提升算法目标分类能力的通道信息与位置定位能力的空间信息。如图 1 所示，通过整合通道注意力和空间注意力，获取目标有效特征的权重，对输入特征与权重进行逐像素乘法运算，将获得的校正后的特征图作为输出用于目标的定位与分类。

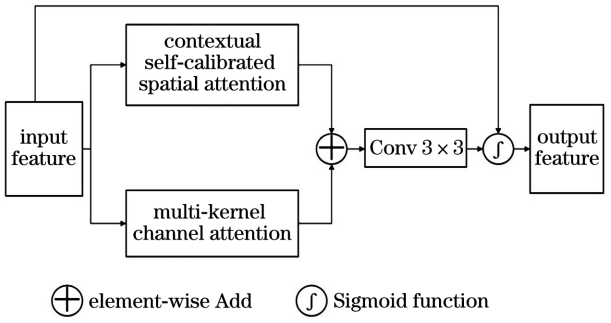

图 1 上下文结合的自校准双重注意力模块

Fig. 1 Contextual self-calibration dual-attention module

2.1.1 上下文自校准空间注意力

采用多支路卷积层结构获取不同感受野的信息，通过不同尺度的感受野信息的结合，模拟人类视觉皮层对不同尺寸感受野信息的敏感度，增强算法在检测任务中的尺度不变性。为了避免池化操作导致内部数据结构丢失和小尺度物体信息无法重建的缺点，采用 Deeplab[15]中提出的空洞卷积获取多尺度感受野信息。针对采样点间隔设置过大(空洞率过大)，获取的信息不连续而丢失大量细节的网格现象，本文设置了两个含有特定空洞率组合的空洞卷积组 HDCONV3 和 HDCONV5[16]，其中数字为最后一个卷积的空洞数，而之前的空洞数则统一设置为 1 和 2。

受文献[17]启发，构建上下文自校准空间注意

力模块，获取上下文空间信息，增强算法的空间定位能力。具体地，设置具有不同功能的路径$\{f_i\}_{i=1}^{4}$，其中f_i表示第i个核大小为C的卷积操作。将经过多感受野结合处理的特征图X分为$\{X_1,X_2\}$两部分，分别输入到这些路径中，再收集不同类型的特征信息。对于第一部分，利用路径$\{f_2,f_3,f_4\}$，在两个不同比例的空间尺度中对特征图进行卷积特征变换，即与输入共享相同分辨率的原始空间尺度与经过下采样后的较小潜在空间尺度。在较小潜在空间尺度中，进行变换的特征图具有较大的空间信息接收场，所以可以用于指导原始特征空间中的特征变换过程，突出空间位置信息。且自校准操作仅考虑局部的上下文联系，所以能够过滤来自无关联区域的噪声干扰，增强空间位置信息的鲁棒性。同时自校准操作会对多尺度信息进行编码，所以对检测任务中目标尺度的变化具有一定鲁棒性。自校准空间注意力结构如图 2 所示。具体过程为

$$X'_1=\mathrm{Up}\left[f_2*\mathrm{AvgPool}_r(X_1)\right],\tag{1}$$

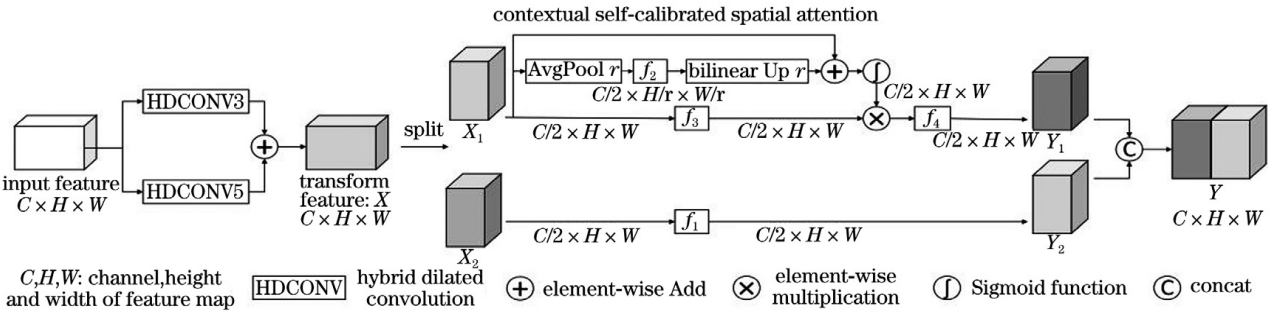

图 2　上下文自校准空间注意力

Fig. 2　Contextual self-calibrated spatial attention

式中：$*$为卷积操作符号；$\mathrm{AvgPool}_r(\cdot)$是步长为$r$、核大小为$r\times r$的自适应平均池化操作。在用$f_2$对下采样后的特征图进行卷积变换后，通过双线性上采样运算$\mathrm{Up}(\cdot)$，图像恢复到原始尺寸大小。自校准运算公式为

$$Y'_1=(f_3*X_1)\cdot\sigma(X_1+X'_1),\tag{2}$$

式中：σ为 Sigmoid 函数。而经自校准后，最终的输出特征图为

$$Y_1=f_4*Y'_1。\tag{3}$$

对于第二部分，通过简单的卷积采样操作$Y_2=f_1*X_2$保留原始特征图信息，然后通过级联操作对$\{Y_1,Y_2\}$两个特征图进行合并，得到有丰富空间位置信息的最终输出特征Y。

2.1.2　多核通道注意力

受文献[18]启发，通过构建多核通道注意力模块提升算法在检测任务中的分类能力。令给定特征图X分别经过卷积核大小为 3×3 和 5×5 的卷积核，获得变换后的特征U_1和U_2，并通过逐像素相加获得特征图U；对所得特征图U进行空间压缩，将二维特征通道压缩为 1×1 的特征单元，获取全局的特征通道响应分布；其次通过通道扩大和激活操作获得每个特征通道的权重，并分别对特征U_1和U_2进行加权，得到对通道信息有不同贡献的特征图M_1和M_2；最后合并得到多核通道特征图M。多核通道注意力结构如图 3 所示。

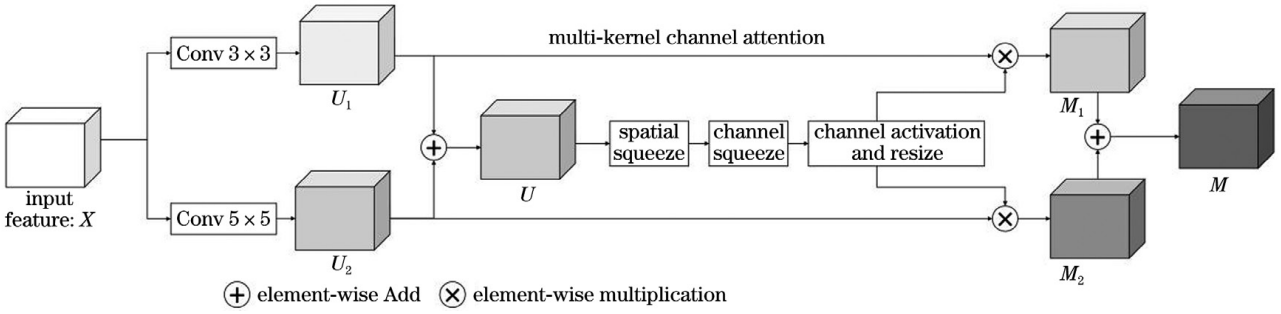

图 3　多核通道注意力

Fig. 3　Multi-kernel channel attention

2.2　改进的可变卷积

传统的矩形卷积核对近似矩形的目标有较好的特征提取效果，但对于尺度和形态变化较大的物体或轮廓不规则的物体易引入背景噪声，从而影响目

标检测效果。如对于汽车，传统卷积核的采样点能够准确落在目标区域，而对于有着不规则的轮廓或者特异姿势的目标，卷积核的一部分采样点必然会落在背景中，从而引入噪声，影响提取目标特征的质量，降低检测精度。为此，针对 FCOS 算法中传统矩形卷积运算的局限性，使其更好地获取多类目标的特征，减小背景噪声，引入改进的可变形卷积(DCNV2)，表达式为

$$y(p)=\sum_{1}^{K}w_k\cdot x(p+p_k+\Delta p_k)\cdot\Delta m_k,\tag{4}$$

式中：x 为输入特征图；y 为输出特征图；p 为特征图上像素的位置；k 为卷积核的元素数量；w_k 为第 k 个位置的权重；p_k 为预定义的采样偏移。若 $K=9$，则有 $p_k\in\{(-1,-1),(-1,0),\cdots,(1,1)\}$。通过增加一个由卷积操作获得的可学习偏移量 Δp_k，卷积核的采样点能进行一定程度的偏移，更好地契合目标的轮廓和尺寸。简单的偏移会导致采样点偏向目标特征薄弱的位置，原本在含有丰富特征的目标中心及周围的采样点会偏移到目标的边缘，影响目标检测效果。所以设置可学习调制因子 $\Delta m_k\in[0,1]$，使偏移指向目标特征丰富的位置。

2.3　算法整体框架

所提算法选用基于无锚的(anchor-free)FCOS 算法为基础框架，在特征提取网络输出上添加改进的可变卷积模块(DCNV2)，通过自适应位置采样，增强网络特征的表达能力，减少背景噪声干扰；同时构建 CSDAM 提取特征金字塔输出特征图的上下文空间信息与全局通道信息，并将其与多感受野信息融合，增强算法的尺度不变性、空间定位能力以及目标分类能力，进一步提升算法的检测性能。改进的可变卷积模块和 CSDAM 对算法运行效率影响小。整体算法的网络结构如图 4 所示。

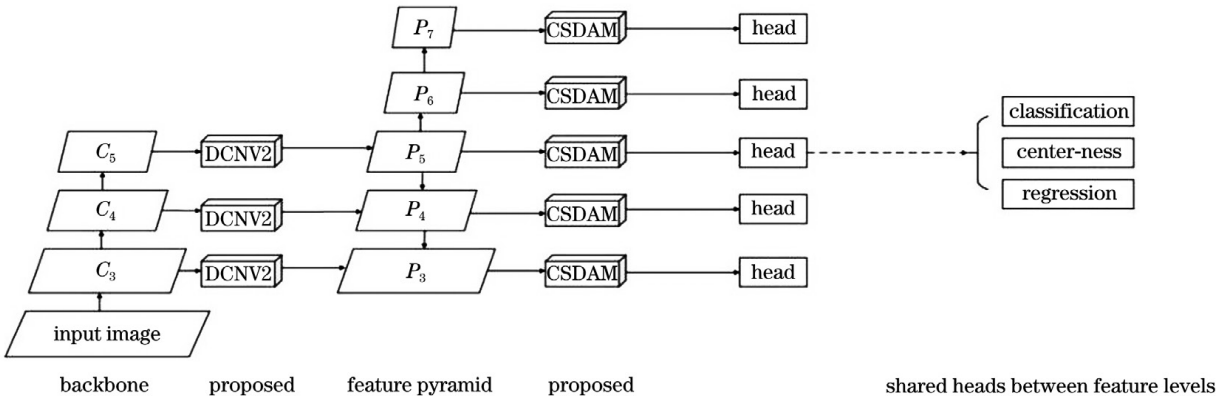

图 4　所提算法的网络结构

Fig. 4　Network structure of proposed algorithm

3　实验结果分析

3.1 实验细节

实验模型基于 Pytorch 框架，实验操作系统为 Ubuntu16.04，使用 Tesla V100GPU 的 NVIDIA DGX Station 服务器的 Linux 系统，用 4 块 GPU 进行训练和测试。测试与训练数据集为大型公开数据集 MSCOCO。以 FCOS 算法为框架，以 ResNet-101 作为特征提取的骨架网络，利用 ImageNet 数据集预训练该网络进行迁移学习。整个训练过程共进行 165000 次迭代。初始学习率设置为 0.01，衰减系数设为 0.0001，批量大小设置为 16，最终学习率为 0.0001。

3.2　评价指标

采用综合精度评价指标 mean Average Precision(mAP)。mAP 是所有类别平均精度的均值，而每一个类的平均精度为 AP。AP 的表达式为

$$P_{\mathrm{A}}=\int P(R)\mathrm{d}R,\tag{5}$$

式中：R 为召回率；P 为准确率。AP 的数学含义为 P-R 曲线与坐标轴包围得到的面积。准确率的含义为在预测的目标中，预测正确的目标数量占总预测数量的比值，即总预测目标数量中预测正确目标的占比，因此准确率又称为查准率。召回率的含义是预测正确的目标占目标总数量的比值，即正确检测目标在总的待检测目标中的占比，因此召回率又称为查全率。

3.3　数据集介绍

实验选取的 Microsoft COCO：Common Objects in Context(MSCOCO)数据集是微软于 2014 年构建的大型多类目标检测数据集。数据集包含 80 种类别数据，并针对不同尺度的目标提出了

评估标准。其中训练集共有图像 82783 张，测试集共有图像 40775 张，验证集共有图像 40504 张，平均每张图片包含 3.5 个类别和 7.7 个实例目标。每张图片中较多的目标数量、较大的尺度差异、严格的评估标准使 MSCOCO 数据集成为多类目标检测任务的主流数据集。

3.4　实验结果对比

在 MSCOCO 数据集下对所提算法与主流的目标检测算法进行对比，其中 P_{AP} 为步长在 0.50 至 0.95 共 10 个 T_{IoU} 阈值下的 mAP 值；P_{AP50} 和 P_{AP75} 分别为在阈值为 0.50 和 0.75 下的 mAP 值；P_{APs}、P_{APm}、P_{APl} 分别为小、中、大不同尺寸目标的 mAP 值，如表 1 所示。可以看出：所提算法的 mAP 达到了 45.2%，相比 CornerNet 算法提升 4.7 个百分点，相比 ExtremeNet 算法提升 1.5 个百分点，证明相比主流目标检测算法，所提算法在检测任务中拥有更强的特征表达能力；在小，中，大三个类型的目标尺度检测任务中获得了优于主流算法的检测结果，证明所提算法在检测任务中鲁棒性更好。

表 1　各种算法在 MSCOCO 数据集上的测试结果

Table 1　Test results of various algorithms on MSCOCO dataset

Method	Backbone	P_{AP}/%	P_{AP50}/%	P_{AP75}/%	P_{APs}/%	P_{APm}/%	P_{APl}/%
YOLOv2[3]	DarkNet-19	21.6	44.0	19.2	5.0	22.4	35.5
YOLOv3[4]	DarkNet-53	30.0	57.9	44.1	24.1	44.2	51.2
Faster R-CNN[1]	ResNet-101	36.2	59.1	39.0	18.2	39.0	48.2
SSD513[2]	ResNet-101	31.2	50.4	33.3	10.2	34.5	49.8
CornerNet[5]	Hourglass-104	40.5	63.8	46.3	19.4	42.7	53.9
ExtremeNet[6]	Hourglass-104	43.7	60.5	47.0	24.1	46.9	57.6
Proposed algorithm	ResNet-101	45.2	64.3	48.9	28.3	49.5	58.9

3.5　消融实验

以 FCOS 算法为框架，ResNet-101 为骨架网络提取特征，在 MSCOCO 数据集上进行消融实验，以验证新添加模块的有效性，其中 FCOS 算法的检测结果由实验得到。表 2 为消融实验结果，在不使用新添加模块的情况下，FCOS 算法的 mAP 为 42.3%，每张图的测试时间为 57 ms。只添加 CSDAM，mAP 提升 0.7 个百分点，变为 43.0%，而测试时间仅提升 4 ms，证明 CSDAM 能够通过获取多感受野的尺度信息、基于上下文的空间位置信息和不同通道的分类信息有效地提升检测性能且几乎不降低算法的运行效率。只添加 DCNV2 模块，mAP 提升 2.7 个百分点，变为 45.0%，算法测试时间增加 16 ms，变为 73 ms，证明 DCNV2 能够通过改进的可变卷积获取外形不规则的目标特征，有效抑制原方法中由矩形卷积核带来的背景噪声，仅略微降低运行效率。当同时添加两个模块时，mAP 提升 2.9 个百分点，变为 45.2%，比只添加 CSDAM 提升 2.2 个百分点，比只添加 DCNV2 模块提升 0.2 个百分点，测试时间变为 78 ms，相比原算法，仅提升 21 ms。结果证明，两个方法有效结合后，不会出现特征冗余现象和降低检测效果，并对算法运行效率影响小。

表 2　消融实验结果

Table 2　Ablation experiment results

CSDAM	DCNV2	P_{AP}/%	Testing time /ms
		42.3	57
√		43.0	61
	√	45.0	73
√	√	45.2	78

3.6　算法检测效果图

在 MSCOCO 验证集上的检测效果如图 5 所示，在检测任务中，所提算法有良好的尺度不变性与目标分类能力，且位置定位能力较好，具有良好的检测效果。

4　结　　论

针对 FCOS 算法目标尺度跨度大、轮廓不规则导致难以提取有用特征的问题，构建了上下文自校准双重注意力模块来获取鲁棒的目标特征，并运用改进的可变卷积 DCNV2 解决算法中矩形卷积核在检测外形不规则目标时引入背景噪声的问题，进一步提升特征的表达能力。在MSCOCO大型多类目

图 5 多类目标检测结果

Fig. 5 Multi-class object detection results

标数据集上进行训练与测试，相较于 FCOS 算法，所提算法加入新模块后的 mAP 提升了 2.9 个百分点，优于近年来的主流目标检测算法。

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] 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.

[3] 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.

[4] Redmon J, Farhadi A. YOLOv3: an incremental improvement[EB/OL]. (2018-04-08)[2019-11-01]. https://arxiv.org/abs/1804.02767.

[5] Law H, Deng J. CornerNet: detecting objects as paired keypoints [J]. International Journal of Computer Vision, 2020, 128(3): 642-656.

[6] Zhou X Y, Zhuo J C, Krähenbühl P, et al. Bottom-up object detection by grouping extreme and center points[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: 850-859.

[7] Huang L C, Yang Y, Deng Y F, et al. DenseBox: unifying landmark localization with end to end object detection [EB/OL]. (2015-11-19)[2019-11-01]. http://export.arxiv.org/abs/1509.04874.

[8] Tian Z, Shen C H, Chen H, et al. FCOS: fully convolutional one-stage object detection[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2019: 9626-9635.

[9] 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 (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 936-944.

[10] Woo S, Park J, Lee J Y, et al. CBAM: convolutional block attention module[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11211: 3-19.

[11] 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.

[12] Chen J M, Jin J, Wang W F, et al. Improved algorithm based on feature pyramid networks [J]. Laser & Optoelectronics Progress, 2019, 56(21): 211505.
陈景明, 金杰, 王伟锋, 等. 基于特征金字塔网络的改进算法[J]. 激光与光电子学进展, 2019, 56(21): 211505.

[13] Wang W F, Jin J, Chen J M, et al. Rapid detection algorithm for small objects based on receptive field block[J]. Laser & Optoelectronics Progress, 2020, 57(2): 021501.
王伟锋, 金杰, 陈景明, 等. 基于感受野的快速小目标检测算法[J]. 激光与光电子学进展, 2020, 57(2): 021501.

[14] Zhu X Z, Hu H, Lin S, et al. Deformable ConvNets V2: more deformable, better results[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: 9300-9308.

[15] 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.

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

[17] Liu J J, Hou Q B, Cheng M M, et al. Improving convolutional networks with self-calibrated convolutions[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 10093-10102.

[18] Li X, Wang W H, Hu X L, et al. Selective kernel networks[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: 510-519.