Software Engineering and Applications 软件工程与应用, 2023, 12(5), 697-704
Published Online October 2023 in Hans. https://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2023.125067

Hans汉斯

基于RFB模块与注意力机制的目标检测算法

王志青

南京邮电大学计算机学院，江苏 南京

收稿日期：2023年8月15日；录用日期：2023年10月7日；发布日期：2023年10月18日

摘 要

针对目标检测算法中深度卷积网络提取特征图关联性不足导致的检测精度下降问题，提出一种基于群体感受野模块(Receptive Field Block, RFB)与坐标注意力(Coordinate Attention, CA)的改进SSD目标检测算法。使用深层特征提取网络ResNet50作为主干网络，并在卷积层结构中添加坐标注意力模块，捕获方向和位置感知的信息；为充分利用不同特征图之间的关联信息，在特征提取与预测中采用反卷积与上采样等方式，融合低层位置特征和高层语义信息。同时在网络结构中引入多尺度卷积核与空洞卷积的RFB模块，以提高感受野的方式提高网络的特征提取能力。实验表明：该算法在PASCAL VOC 2007数据集上的mAP为78.08%，相较于传统的SSD算法检测能力得到了显著提升。

关键词

目标检测，单阶多层检测器，RFB模块，坐标注意力

Object Detection Algorithm Based on RFB Module and Attention Mechanism

Zhiqing Wang

School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing Jiangsu

Received: Aug. 15th, 2023; accepted: Oct. 7th, 2023; published: Oct. 18th, 2023

Abstract

Aiming at the problem of insufficient correlation of feature map extracted by deep convolutional network in object detection algorithm, an improved SSD object detection algorithm based on Receptive Field Block and Coordinate Attention is proposed. The deep feature extraction network ResNet50 is used as the backbone network, and a coordinate attention module is added to the convolutional layer structure to capture the information of direction and location awareness. In

文章引用: 王志青. 基于 RFB 模块与注意力机制的目标检测算法[J]. 软件工程与应用, 2023, 12(5): 697-704.
DOI: 10.12677/sea.2023.125067

order to make full use of the association information between different feature maps, deconvolution and upsampling are used in feature extraction and prediction to integrate low-level location features and high-level semantic information. At the same time, the RFB module of multi-scale convolution kernel and hole convolution is introduced in the network structure to improve the feature extraction ability of the network by improving the receptive field. Experiments show that the mAP of the algorithm on the PASCAL VOC 2007 dataset is 78.08%, which is significantly improved compared with the traditional SSD algorithm.

Keywords

Object Detection, Single-Stage Multilayer Detector, RFB Module, Coordinate Attention

1. 引言

目标检测作为计算机视觉的一项基本任务，也是计算机视觉领域中研究热点之一，在无人驾驶、智能视频监控、生物检测等方面有着广泛的应用。随着基于深度学习目标检测算法的研究与发展，其主要检测算法可以划分成两类：基于候选区域的双阶段目标检测算法和基于回归分析的单阶段目标检测算法[1]，前者第一阶段生成大量的候选区域(region proposal)，第二阶段把候选区域放入分类器中，执行分类定位任务，该类算法检测精度高但速度慢[2]，主要算法有：R-CNN [3]系列和 SPPNet 等；后者直接进行端到端检测，生成目标的类别概率和位置信息，大幅度降低了计算资源消耗，该类算法简单高效、检测速度快，但准确度较低[4]，主要算法有：SSD [5]和 YOLO [6]系列等。

SSD 算法对小目标检测性能不佳，主要原因有：特征信息少，正负样本不平衡、数据集不完备以及锚框设计难等[7] [8]。针对以上不足，研究人员提出很多改进算法。DSSD [9]针对浅层特征图表征能力不强，引入反卷积模块，通过反卷积层学习得到上采样特征图与浅层特征图融合，充分利用上下文信息。M2Det [10]针对传统 FPN 网络的特征层对于目标检测任务不够代表性且特征图仅包含单层信息，提出了多级特征金字塔网络，构建多尺度多层级的特征金字塔。STDN [11]引入骨干网络 DenseNet [12]和尺度转换层(Scale-transfer layer)，通过 reshape 函数将通道数转化为特征图的宽和高生存大尺度预测特征图，极大减少了算法的参数量和计算量。文献[13]针对传统交并比框回归效果差及收敛速度慢，提出基于关键点距离交并比算法，降低定位损失，减少目标漏检情况。

本文基于传统 SSD 算法的不足，设计了一种基于群体感受野模块与坐标注意力机制的改进 SSD 目标检测算法。在前两层较浅预测特征图后添加 RFB 模块，增加感受野。并在 ResNet50 网络中引入坐标注意力机制，不仅获取通道间信息，还考虑了方向相关的位置信息，有效地提升模型的准确率。比较实验结果，改进后的 SSD 算法检测精度有明显提升。

2. 本文模型

2.1. 网络结构

基于群体感受野模块与坐标注意力的改进 SSD 目标检测算法结构如图 1 所示。网络架构设计遵循的主要原则是在主干网络中降低特征信息损失，采用 ResNet50 [14]网络作为特征提取的骨干网络，其残差

连接方式可以有效的抑制神经网络因深度增加而带来的梯度衰减问题，提高模型对图像特征的表达能力。

在主干网络提取出来的(38, 38)、(19, 19)和(10, 10)特征层后，考虑通道之间远程依赖关系，经过坐标注意力模块进行编码。为增强信息表达能力，将深层特征层与浅层特征层进行有效信息融合，(19, 19)和(10, 10)尺寸特征层采用线性插值方式将其放大到(38, 38)尺寸，并进行 Concat 方式融合。同时为改善网络的梯度，防止梯度爆炸，实现归一化，采用 BN 层处理。在生成预测特征层上，采用反卷积方式再次进行特征融合[15]，并为增加感受野，获得更大的上下文信息，添加了 RFB 模块。从可以得出，低层特征层分辨率较高，包含更多图像细节。

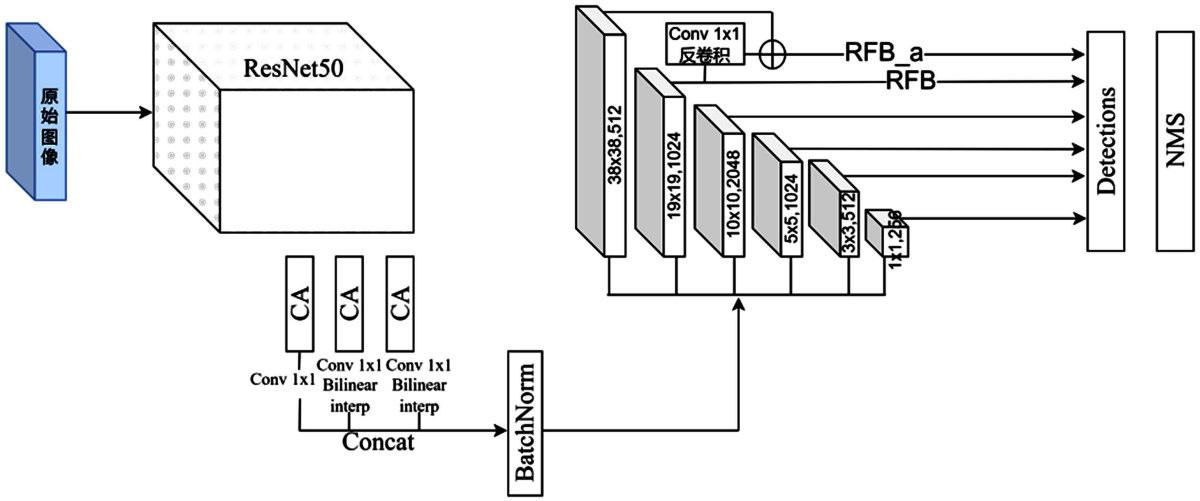

Figure 1. SSD object detection algorithm based on RFB and attention mechanism
图 1. 基于 RFB 与注意力机制的 SSD 目标检测算法

2.2. 坐标注意力模块

坐标注意力模块[16]如图 2 所示。为了缓解二维全局池化导致的位置信息损失，坐标注意力将通道注意力分解为两个并行的一维特征编码过程，分别沿两个空间方向聚合特征，有效地将整合空间坐标信息输入到生成的注意力特征图中。

具体来说，输入的特征图为 X，大小为 $C\times H\times W$ 的特征图，其高和宽为 H 和 W，高度为h的第 c 个通道的输出 $z_c^h(h)$ 为：

$$z_c^h(h)=\frac{1}{W}\sum_{i=0}^{W-1}x_c(h,i) \tag{1}$$

式中 $x_c(h,i)$ 为第 i 行特征向量。

宽度 w 的第 c 个通道的输出 $z_c^w(w)$ 为：

$$z_c^w(w)=\frac{1}{H}\sum_{j=0}^{H-1}x_c(j,w) \tag{2}$$

式中 $x_c(j,w)$ 为第 j 列特征向量。

通过上述操作可以获得全局感受野和位置信息。拼接两个输出结果后，再使用 1 × 1 卷积操作可以得到空间信息经过编码后的中间特征图 F。

$$F=\delta\left(F_1\left(\left[z^h,z^w\right]\right)\right) \tag{3}$$

式中 z^h 为所有高度为 h 的通道的输出； z^w 为所有宽度为 w 的通道的输出； $\left[z^h, z^w\right]$ 为特征图在垂直和水平方向上的拼接； F_1 为卷积操作； δ 为非线性激活函数。

然后将 F 切分为 2 个单独的张量，再分别利用 1 × 1 卷积变换成 X 相同的通道数，最后利用 Sigmoid 激活函数得到注意力权重 g^h 和 g^w 。

输入特征图 X 的第 c 通道上高度为 i 宽度为 j 的特征 $x_c(i,j)$ 经过坐标注意力模块后的输出 $y_c(i,j)$ 为：

$$y_c(i,j) = x_c(i,j) \times g_c^h(i) \times g_c^w(j) \tag{4}$$

式中： $g_c^h(i)$ 表示第 c 通道上高度为 i 的水平注意力权重， $g_c^w(j)$ 表示第 c 通道上宽度为 j 的水平注意力权重。

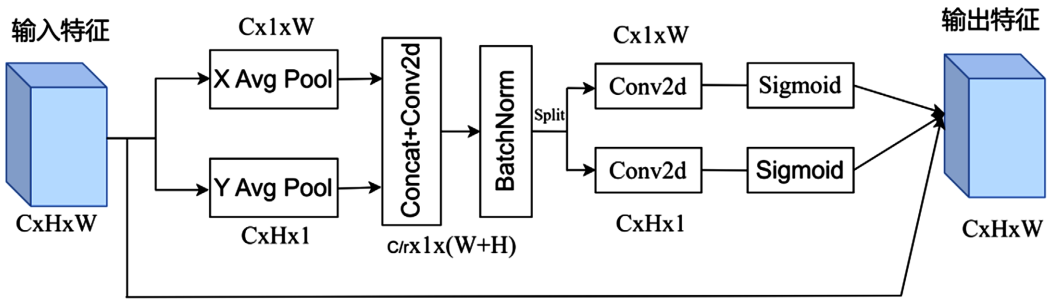

Figure 2. Coordinate attention module
图 2. 坐标注意力模块

2.3. RFB 模块

RFB 模块[17]是通过多分支结构是由于不同尺寸的卷积核卷积同一张特征图可以得到不同大小的感受野，模拟人类视觉感受野的不同范围。网络结构借鉴 Incetion 思想，核心结构包括多个分支，分别对应不同的感受野，最终通过 Concat 函数合并在一起。感受野随着卷积神经网络层数的增加而变小，影响了网络对特征的提取能力和检测效果。

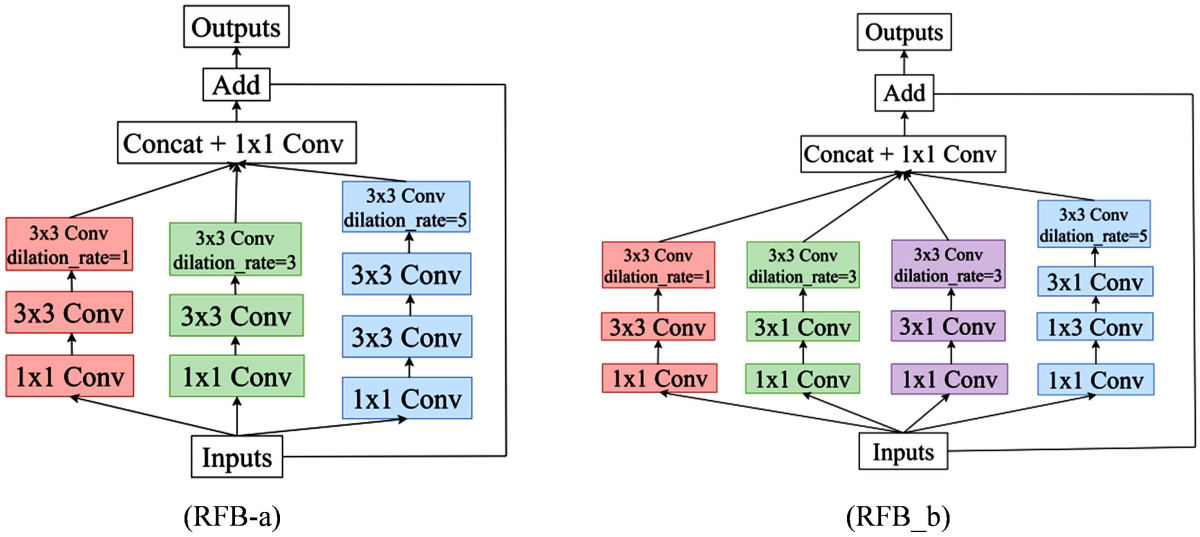

Figure 3. RFB module
图 3. RFB 模块

本文采用的 RFB 模块的两种结构如图 3 所示，由多分支卷积结构和空洞卷积构成。先用一个 1 × 1 的卷积减少通道数，再加上一个 3 × 3 的卷积，RFB_a 模块第 1 个分支为一个空洞率为 3 的 3 × 3 空洞卷积，第 2 个分支先经过一个 3 × 3 的普通卷积，再经过一个空洞率为 3 的 3 × 3 空洞卷积，第 3 个分支首先经过两个 3 × 3 的卷积，其效果等于一个 5 × 5 的卷积，之后经过空洞率为 5 的 3 × 3 空洞卷积，最后通过 Concat 通道级联。RFB-b 和 RFB_a 相比主要采用 1 × 3 和 3 × 1 卷积层代替 3 × 3 卷积层，主要增加了模型的非线性特征并减少计算量。

3. 结果与分析

3.1. 数据集和评价指标

数据集：本文采用公开数据集 PASCAL VOC 2007 + 2012 进行实验，该数据集总共有 20 种常见类别，此次实验为扩大训练集数据，合并了 VOC 2007 和 2012 的训练集，测试数据为 VOC 2007 测试集。

评价指标：*mAP* 作为目标检测领域重要评估指标之一，综合考虑了所有的类别以及定位精度等问题，其值越大模型的检测性能越好。计算方法为式 1~式 3。

$$P=\frac{TP}{TP+FP},R=\frac{TP}{TP+FN} \tag{5}$$

$$AP=\int_0^1 P(R)\mathrm{d}R\approx\frac{1}{m}\sum_{i=1}^{m}P(R_i) \tag{6}$$

$$mAP=\frac{1}{C}\sum_{i=1}^{C}AP_i \tag{7}$$

实验平台配置：本实验操作系统为 Windows11，具体配置如表 1 所示。

Table 1. Experiment configuration parameters
表 1. 实验配置参数

配置项	配置参数
内存	16GB
显卡(显存)	NVIDIA RTX3060(8GB)
处理器(CPU)	Intel i5-12490F
深度学习框架	Pytorch
CUDA 版本	11.7
输入图片尺寸	300 × 300
权重衰减参数	0.1
batchsize	16
初始学习率	0.001
MAX_ITER	120000

3.2. 结果与分析

为了验证改进后模型的有效性，本文设置了对照实验：① 标准 SSD 模型作为实验对照组；② 使用 ResNet50 作为骨干网络，在特征提取阶段融合不同深度特征图；③ 在实验 2 的基础上，增加坐标注意力模块；④ 在实验 3 的基础上，增加 RFB 模块，同时将 19 尺寸特征图反卷积操作与上层特征图融合。

4 种模型的实验结果平均精度值 *mAP* 如表 2 所示，分析实验结果，在原 SSD 模型基础上每改进后，

mAP 值均有提升，其中引入坐标注意力模块和 RFB 模块后，*mAP* 值分别提升了 0.92%和 0.85%。实验表明，坐标注意力模块和 RFB 模块对于传统 SSD 算法性能的提升有着显著的效果。

Table 2. Comparison of PASCAL VOC 2007 test sets
表 2. PASCAL VOC 2007 测试集对比

模型	VGG-16	ResNet50	CA	RFB	*mAP*/%
SSD	√				76.17
SSD		√			76.31
SSD_CA		√	√		77.23
SSD_CA_RFB		√	√	√	78.08

20 种目标类别检测准确率 *AP* 如表 3 所示，绝大部分类别的准确率均有显著提升，其中 aeroplane、bird、diningtable 和 tvmonitor 尤为突出，分别提升了 4.21%、3.62%、4.45%和 4.28%。

Table 3. 20 categories of different algorithm test results
表 3. 20 种类别不同算法测试结果

种类	SSD_VGG	SSD_ResNet	SSD_CA	SSD_CA_RFB
aeroplane	0.7982	0.8222	0.8270	0.8403
bicycle	0.8386	0.8649	0.8522	0.8552
bird	0.7438	0.7885	0.7827	0.7800
boat	0.7000	0.6732	0.6987	0.6924
bottle	0.5122	0.4871	0.5008	0.5087
bus	0.8443	0.8534	0.8295	0.8573
car	0.8579	0.8574	0.8613	0.8597
cat	0.8693	0.8752	0.8713	0.8773
chair	0.6062	0.6095	0.6019	0.5936
cow	0.8097	0.7964	0.8372	0.8237
diningtable	0.7282	0.7489	0.7665	0.7727
dog	0.8349	0.8627	0.8686	0.8602
horse	0.8596	0.8529	0.8741	0.8765
motorbike	0.8252	0.8579	0.8535	0.8567
person	0.7871	0.7774	0.7839	0.7847
pottedplant	0.4920	0.5217	0.5177	0.5224
sheep	0.7316	0.7656	0.8153	0.7852
sofa	0.7979	0.7781	0.8029	0.8117
train	0.8503	0.8672	0.8524	0.8672
tvmonitor	0.7480	0.7942	0.7816	0.7908

VOC2007 测试集数据有 4952 张，测试阶段通过单张批量测试得到分类损失、定位损失及总损失，以传统 SSD 算法作为基线模型，对比每个组件对模型的检测性能的作用，本文模型的总损失值最低且收敛速度略优于原模型，表明了本文模型能有效降低分类定位损失，提升算法的检测性能。分类损失和回归损失的总损失效果对比如图 4 所示。

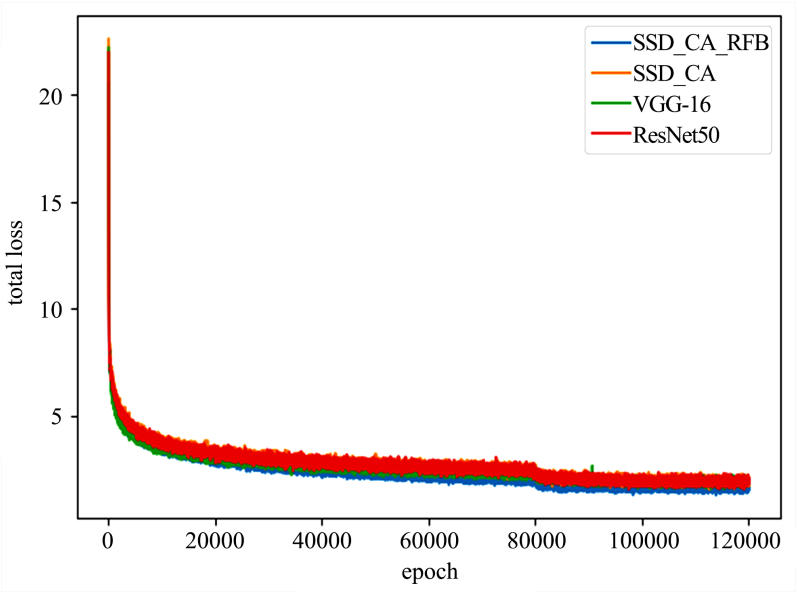

Figure 4. Comparison chart of different algorithms
图 4. 不同算法 loss 对比图

4. 结论

本文设计了一种基于群体感受野模块与坐标注意力机制的改进 SSD 目标检测算法，采用 ResNet50 网络作为特征提取的骨干网络，并引入轻量级坐标注意力机制，能够同时考虑通道间关系以及长距离的位置信息，利于模型更准确定位目标信息，增强识别能力，且坐标注意力模块轻量灵活仅带来少量的计算消耗。在特征提取阶段通道拼接融合不同深度卷积层输出，丰富预测特征图上下文信息，同时在预测过程中加入 RFB 模块，通过不同尺寸卷积核的多分支结构和空洞卷积来提高感受野，增强特征提取能力。通过实验验证，叠加各个模块后，算法的检测精确度均有提升。实验表明，改进后目标检测算法在 PASCAL VOC 数据集上各类别的检测准确率较传统 SSD 算法有着显著提高，mAP 比传统 SSD 算法提升了 1.91%。

基金项目

国家自然科学基金(61872190)；江苏省博士后科研资助计划项目(2020Z058)。

参考文献

[1] 罗会兰, 陈鸿坤. 基于深度学习的目标检测研究综述[J]. 电子学报, 2020, 48(6): 1230-1239.

[2] 张阳婷, 黄德启, 王东伟, 贺佳佳. 基于深度学习的目标检测算法研究与应用综述[J/OL]. 计算机工程与应用. https://kns.cnki.net/kcms/detail/11.2127.TP.20230620.1746.002.html, 2023-01-13.

[3] Girshick, R. (2015) Fast R-CNN. 2015 *IEEE International Conference on Computer Vision* (*ICCV*), Santiago, 7-13 December 2015, 1440-1448. https://doi.org/10.1109/ICCV.2015.169

[4] 王琳毅, 白静, 李文静, 等. YOLO 系列目标检测算法研究进展[J]. 计算机工程与应用, 2023, 59(14): 15-29.

[5] Liu, W., Anguelov, D., Erhan, D., *et al*. (2016) SSD: Single Shot MultiBox Detector. In: Leibe, B., Matas, J., Sebe, N. and Welling, M., Eds., *Computer Vision—ECCV* 2016, Springer, Cham.

[6] Redmon, J., Divvala, S.K., Girshick, R.B. and Farhadi, A. (2016) You Only Look Once: Unified, Real-Time Object Detection. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 779-788. https://doi.org/10.1109/CVPR.2016.91

[7] 许德刚, 王露, 李凡. 深度学习的典型目标检测算法研究综述[J]. 计算机工程与应用, 2021, 57(8): 10-25.

[8] 杜紫薇, 周恒, 李承阳, 等. 面向深度卷积神经网络的小目标检测算法综述[J]. 计算机科学, 2022, 49(12):

205-218.

[9] Fu, C.Y., Liu, W., Ranga, A., *et al*. (2017) DSSD: Deconvolutional Single Shot Detector. *Computer Vision and Pattern Recognition*, 1-11. https://doi.org/10.48550/arXiv.1701.06659

[10] Zhao, Q., Sheng, T., Wang, Y., *et al*. (2019) M2Det: A Single-Shot Object Detector Based on Multi-Level Feature Pyramid Network. *Proceedings of the AAAI Conference on Artificial Intelligence*, Washington DC, 7-14 February 2023, Vol. 33, 9259-9266. https://doi.org/10.1609/aaai.v33i01.33019259

[11] Zhou, P., Ni, B.B., Geng, C., Hu, J.G. and Xu, Y. (2018) Scale-Transferrable Object Detection. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 528-537. https://doi.org/10.1109/CVPR.2018.00062

[12] Huang, G., Liu, Z. and Weinberger, Q.K. (2016) Densely Connected Convolutional Networks. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 2261-2269. https://doi.org/10.1109/CVPR.2017.243

[13] 聂志勇, 阴宇薇, 汤佳欣, 等. 一种基于边界框关键点距离的框回归算法[J]. 计算机工程, 2023, 49(7): 65-75.

[14] He, K.M., Zhang, X.Y., Ren, S.Q. and Sun, J. (2015) Deep Residual Learning for Image Recognition. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 770-778. https://doi.org/10.1109/CVPR.2016.90

[15] Cao, G., Xie, X., Yang, W., *et al*. (2018) Feature-Fused SSD: Fast Detection for Small Objects. *International Conference on Graphic and Image Processing*, Qingdao, 14-16 October 2017, 106151E. https://doi.org/10.1117/12.2304811

[16] Hou, Q., Zhou, D. and Feng, J. (2021) Coordinate Attention for Efficient Mobile Network Design. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 20-25 June 2021, 13713-13722. https://doi.org/10.1109/CVPR46437.2021.01350

[17] Liu, S. and Huang, D. (2018) Receptive Field Block Net for Accurate and Fast Object Detection. In: Ferrari, V., Hebert, M., Sminchisescu, C. and Weiss, Y., Eds., *Computer Vision—ECCV* 2018, Springer, Cham, 404-419. https://doi.org/10.1007/978-3-030-01252-6_24