计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.201200196

基于相邻特征融合的目标检测

李亚泽　刘宏哲
北京联合大学北京市信息服务工程重点实验室　北京 100101
北京联合大学机器人学院　北京 100101
(yaze_li@126.com)

摘　要　随着智能驾驶领域的发展,人们对目标检测的精度要求越来越高,尤其是针对高速行驶时对距离较远的小目标的检测和低速行驶时对密集目标的检测。在当前的两阶段检测框架的特征融合部分,使用 bottom-up 的双向融合方法虽然能够更有效地对大目标进行语义信息和位置信息的特征融合,但会给几个或几十个像素的小目标造成很大的信息损失。当检测网络特征融合部分使用 top-down 的单向融合方法时,则对大目标检测的效果欠佳。为此,文中提出了相邻特征融合(Neighbour Feature Pyramid Network,NFPN)方法、Double RoI(Region of Interest)方法和递归特征金字塔(Recursive Feature Pyramid,RFP)的方法。以 Faster RCNN 50 为基准,同时使用提出的 NFPN,Double RoI 和 RFP 后,在 Lisa 交通数据集中平均精度(mAP)提升了 2.6 个百分点。在 VOC2007 数据集上,以 VOC07+12 train 数据集为训练集,VOC2007 test 为测试集,以 Faster RCNN101 为基准,同时使用提出的 3 个模型,*mAP* 提升了 6 个百分点,同时小、中、大目标的精度也得到提高。

关键词:深度学习;目标检测;计算机视觉;特征融合;智能驾驶
中图法分类号　TP183

Object Detection Based on Neighbour Feature Fusion

LI Ya-ze and LIU Hong-zhe
Beijing Key Laboratory of Information Service Engineering,Beijing Union University,Beijing 100101,China
College of Robotics,Beijing Union University,Beijing 100101,China

Abstract　With the development of intelligent driving,the precision requirements for target detection are getting higher and higher,especially for small targets that are far away. In the neck of two-stage object detection network,although the feature fusion of semantic information and location information is more effective for large targets if the bottom-up fusion method is used,it will cause big information loss to small targets. To address this problem,we propose neighbor feature pyramid networks(NFPN) method of feature fusion of neighbor layers,the Double RoI(Region of Interest) method to fuse the FPN and NFPN features,and the recursive feature pyramicl(RFP) method. Using Faster RCNN 50 as the benchmark,the mean average precision(mAP) of our model in the Lisa data set has increased by 2.6% while using NFPN,Double RoI and RFP. On the VOC2007 data set,using the VOC07+12 train data set for training,VOC2007 test as the test set,and Faster RCNN101 as the baseline,the mAP of our model both used NFPN,Double RoIE and RFP has increased by 6%,and the object detect accuracy of large,medium and small targets is improved at the same time.

Keywords　Deep learning,Object detection,Computer vision,Feature fusion,Autonomous driving

1　引言

随着智能驾驶技术的发展,人们对检测的精度要求越来越高,尤其是对远处比较模糊的交通标志等小目标。当前的目标检测框架对近处比较清晰的目标的检测精度普遍比较高,但是对于远处比较小或者模糊的目标,检测精度则相对较低。本文使用的交通标志数据集 Lisa Traffic Sign[1] 中,定义分辨率小于 32*32 的目标为小目标,介于 32*32 到 96*96

到稿日期:2020-12-22　返修日期:2021-06-08　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(61871039,61906017,61802019);北京市教委项目(KM202111417001,KM201911417001;视觉智能协同创新中心项目(CYXC2011);北京联合大学学术项目(ZK80202001,202011417004,202011417005)
This work was supported by the National Natural Science Foundation of China(61871039,61906017,61802019),Beijing Municipal Commission of Education Project(KM202111417001,KM201911417001),Collaborative Innovation Center for Visual Intelligence(CYXC2011) and Academic Research Projects of Beijing Union University(ZK80202001,202011417004,202011417005).
通信作者:刘宏哲(liuhongzhe@buu.edu.cn)

之间的目标为中目标，大于 96 * 96 的目标为大目标。我们使用 Lisa 交通数据集，以 ResNet[2] 50 作为骨架网络（Backbone，用于提取特征）的 Faster RCNN[3] 为基础网络（baseline）进行实验，发现当前的双向特征融合方法（如 PANet[4] 等）虽然会提高大目标的精度（AP_L），但同时会降低小目标的精度（AP_S），从而使平均精度（mean Average Precision，mAP）降低。对此，我们研究了在两阶段检测框架中，在保持大、中目标检测精度的前提下，提高小目标的检测精度的方法，并提出了相邻尺度特征融合金字塔网络（Neighbor Feature Pyramid Networks，NFPN）、递归 FPN（FPN[5]（Feature Pyramid Networks）为特征融合金字塔）和 Double RoIE（RoIE（Region of Interest Extractor）为感兴趣区域提取）的方法。在 Lisa 数据集上，相比 baseline（Faster RCNN），*mAP* 提升了 1.9 个百分点。在 VOC2007[6] 的 test 测试集上，分别以 VOC2007train 数据集，以及 VOC2007train 数据集和 VOC2012train 数据集的合集作为训练集，在 Backbone 为 ResNet50 的 Faster RCNN 网络上，*mAP* 分别提高了 4.2 和 1.8 个百分点；在以 ResNet[2] 101 网络为 Backbone 的检测网络上，使用相邻特征融合的 *mAP* 分别提高了 2.0 和 6.0 个百分点。本文方法的代码已公开在 https://github.com/dream-in-night/NFPN-GitHub 上。

当前的特征融合方法可分为无融合、单向融合以及双向融合。无融合方法主要应用于单阶段目标检测，如 SSD[7] 等网络。单向融合的主要代表方法为 FPN[1]。双向融合的代表方法有 PANet[4]，BiFPN[8]，以及融合各层多尺度信息的特征融合网络如 Libra RCNN[9]。我们通过实验发现，在 Lisa 交通标志数据集中，单向或双向特征融合方式虽然能够提高对大目标的检测精度，但是对小目标的检测精度比 FPN 更低。我们统计数据集中所有目标的像素并进行排序，发现很多小目标的宽和高都在 100 以下，有些甚至是个位数。这种小目标特征在经过 Backbone 的卷积过程后，相比原图，特征的宽高都缩小为原来的 1/4，有些宽高甚至降到了 10 个像素以下。在经过完整的 Backbone 后，缩小为原图的 1/32，在 FPN 的 extra_layer（FPN 中对最深层特征再次缩放分辨率的卷积操作）后宽高更是缩小为原图的 1/64。经过提取后的特征中，大目标同时具有很强的语义和位置信息，但是小目标的位置信息损失严重。而这时其余的像素都是冗余的噪声信息，在当前的网络结构中，无法利用这些冗余信息。如果使用特征金字塔中最深层的特征进行上采样，当特征传播到特征金字塔的第一层时，由于小目标在深层缺失位置信息，同时还有很强的背景像素产生的噪声干扰，对小目标的检测精度大幅下降。大目标则相反，在最深层即宽高缩小至原图的 1/64 后，特征依然具有很强的语义和位置信息，因此上采样传播到最顶层时，顶层特征同时具备了位置信息和语义信息。Lisa 数据集上的实验表明，相邻特征融合的方法能够提高小目标的精度，但是对大目标的检测精度会下降。

为此，本文参考 ResNet 的思想，在 FasterRCNN 网络的单层感兴趣区域特征提取（Single RoI Extractor）[3] 部分，将 FPN 的特征和相邻特征融合后的特征同时进行卷积，将得到特征相加，以保留大目标和小目标的信息，在提高小目标检测精度的同时提高大目标的检测精度。

为进一步去除小目标在相邻特征融合阶段由背景信息产生的噪音，我们使用 Faster RCNN 中 Backbone 的 4 个卷积块，对相邻层特征融合后的特征进行卷积并提取信息。在交通标志数据集 Lisa 上的实验也表明了该方法的有效性。为验证相邻特征融合方法的通用性，本文在实验部分使用了 VOC2007 数据集进行对比实验，结果表明使用相邻特征融合方法检测的 *mAP* 高于 Faster RCNN 的 *mAP*。

2 研究背景

常见的两阶段检测网络一般都是基于骨架网络生成抽象的语义特征，再根据特定的任务设计检测回归方法，对 Backbone 各层输出的特征进行处理。

按照两阶段 Backbone 的结构（见图 1），将原始图像分辨率修改为（1 333，800，3），并且经过预处理后的原始特征，其层特征分别记作 P_1，P_2，P_3，P_4，P_5，P_6，特征的宽高分别为原图的 1/2，1/4，1/4，1/8，1/16，1/32。基于 Faster RCNN 检测系列网络的 Backbone 的输出特征是 P_3，P_4，P_5，P_6，其使用这些特征进行特征融合，而单阶段目标检测则直接使用基础网络的输出进行检测。

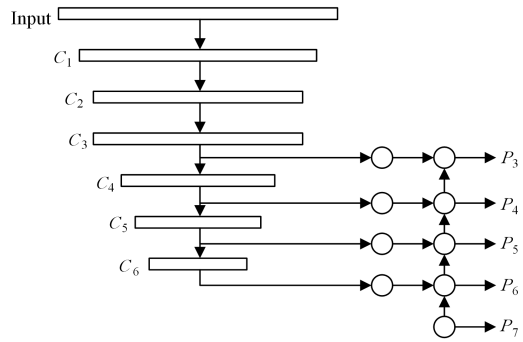


图 1 Backbone 和 FPN 模型

Fig. 1 Backbone and feature pyramid networks

在两阶段检测网络中，当前常见的特征融合方法有 top-down 的单向融合，如 Faster RCNN，Mask R CNN[10]，Yolov3[11]，RetinaNet[12]，Cascade RCNN[13] 等；双向融合和其他的融合方式一般是基于单向融合的，如 PANet（见图 2(b)）和 BiFPN（见图 2(c)）则在单向融合的基础上增加了自下向上的融合方法，在 FPN（见图 2(a)）融合了 Backbone 的 4 个卷积块的输出特征后，使用这些特征作为输入特征。Libra RCNN 针对多尺度下的特征不平衡问题，提出了 BFPN[9]（Balanced FPN，见图 2(d)）方法。在这些方法中，最底层的特征如果经过 4 次上采样，对于小目标来说，其周边的背景像素也经过 4 次上采样，对小目标区域的特征值产生了巨大影响，此时小目标的浅层特征位置信息损失比较严重。因此，本文提出相邻特征融合的方法（Neighbour Feature Fusion Network，NFPN），以减少上采样的次数，从而降低背景像素对浅层小目标特征的影响。

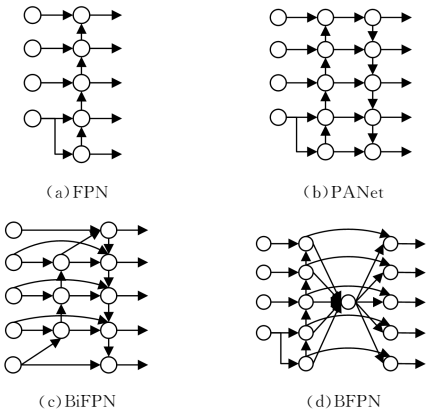


图 2 自顶向下和自底向上的特征融合方式

Fig. 2 Top-down and bottom-up feature fusion method

3 特征融合

本文提出的方法 double RoI，首先在特征融合(Neck)阶段进行相邻特征融合，然后将上一步融合后的各层特征再次输入 Backbone 进行特征提取，最后在特征提取阶段同时对 FPN 和 NFPN 进行感兴趣区域特征提取。

3.1 相邻特征融合

相邻特征融合的方法作用于特征融合部分，在相邻层的特征间进行融合，降低了检测时小目标的信息损失。NFPN(见图 3)的融合方式为融合相邻两个层次的特征，下层特征上采样，上层特征下采样，从而获得与本层特征相同的尺度。下层特征上采样后，使用卷积方式优化插值后的特征，上层的特征下采样时使用步长为 2 的卷积，或者池化后步长为 1 的卷积，然后再与本层特征相加进行融合，随后我们将融合后的特征与原来输入的特征相加，以减小上采样时造成的小目标信息损失，同时保持上采样对大目标的信息增益。

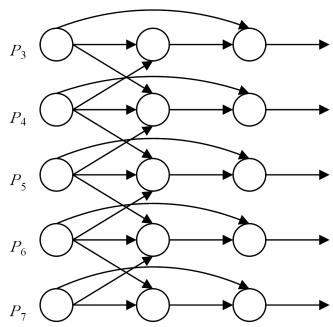


图 3 相邻特征融合网络

Fig. 3 Neighbor feature pyramid network

相邻特征融合中，设 O_i 为 FPN 的输出，P_7 为 FPN 对 P_6 进行的额外特征提取后的特征，Resize 为上采样或下采样操作：

$$O_3 = C_3 + P_3 + \text{Resize}(P_4)$$

$$O_4 = C_4 + P_4 + \text{Resize}(P_3) + \text{Resize}(P_5)$$

$$O_5 = C_5 + P_5 + \text{Resize}(P_4) + \text{Resize}(P_6)$$

$$O_6 = C_6 + P_6 + \text{Resize}(P_5) + \text{Resize}(P_7)$$

$$O_7 = C_7 + P_7 + \text{Resize}(P_6)$$

在实验中，将 NFPN 叠加后可以发现，检测的 mAP 和小目标的检测精度有所下降(见表 1)，而大目标的检测精度则会提升，这进一步说明了 bottom-up 方法对小目标会造成不利影响，即前文描述的连续上采样时背景像素会对小目标造成干扰，从而影响其位置信息。

表 1 在 Lisa 数据集上的实验结果

Table 1 Experiment results on Lisa traffic sign dataset

(单位：%)

Method	mAP	AP_{50}	AP_{75}	AP_s	AP_M	AP_L	Backbone
Faster RCNN	65.2	76.1	74.8	66.8	71.0	70.5	Res50-FPN
BFP	63.9	75.9	74.0	63.4	71.2	77.9	ResN50-FPN
BiFPN	58.4	70.9	69.8	55.4	65.1	83.5	Res50-FPN
BiFPN * 2	49.2	59.9	58.4	49.1	56.6	78.5	Res50-FPN
NFPN (ours)	66.1	77.5	75.8	65.9	72.1	75.7	Res50-FPN
DoubleRoI+ NFPN(ours)	**67.4**	**78.8**	**77.5**	**68.0**	**74.3**	**76.0**	Res50-FPN
NFPN * 2+ DoubleRoI(ours)	66.0	78.1	76.4	66.2	72.5	80.7	Res50-FPN
NFPN+递归+ DoubleRoI(ours)	**67.8**	**79.0**	**77.7**	**68.4**	**74.0**	**81.1**	Res50-FPN

3.2 递归特征金字塔

受到 DetectoRS[15] 中 think twice 的思路以及 CBNet[16] 中双 Backbone 和 OHEM[17] 在进行困难样本采样时重新打分排序的启发，本文设计了递归特征金字塔(见图 4)。其将 NFPN 融合后的特征再次输入 Backbone 中对应的卷积块进行卷积，这样不仅可以过滤特征融合部分上采样时由背景信息造成的冗余噪声信息，还可以增强上采样时大目标由语义信息产生的位置特征。

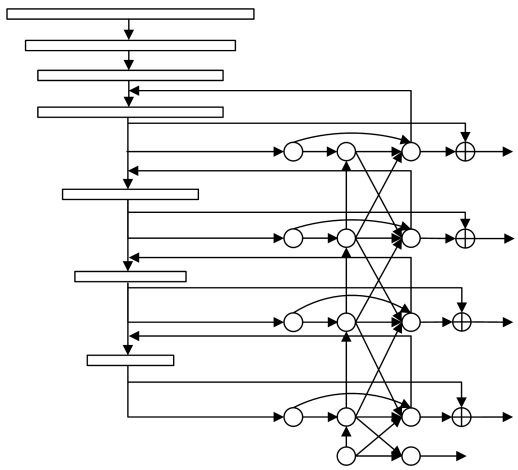


图 4 NFPN 递归金字塔结构

Fig. 4 Recursive NFPN

3.3 多级感兴趣区域提取

两阶段检测网络在 neck 阶段输出单个特征，即 top-down 或 bottom-up 方式融合的特征。在 RoIE 阶段，根据预测框的面积，映射到特征金字塔的对应层，并对该层进行卷积得到 7 * 7 的特征，进一步对其分类和回归。

由此得到的特征中只包含一种融合方法得到的信息，比如使用 bottom-up 方式融合的特征，可能有利于大目标的检测，但会降低小目标的检测精度；而只使用 top-down 方式融合特征则无法利用双向融合时充分融合特征的优点来对大目标产生有利影响。

因此，本文设计了 Double RoI 结构（见图 5），在 neck 阶段融合相邻特征后，同时输出 FPN 的特征，RoIE（RoI Extractor）将这两个特征卷积后得到的特征相加，进行位置分类和回归。

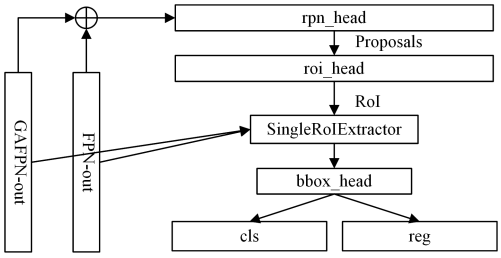


图 5 Double RoI 网络结构

Fig. 5 Double RoI Net

4 实验

我们以 Faster RCNN 为 baseline，在 Lisa 交通标志数据集上，将本文方法与 Libra-RCNN，PANet，BiFPN 和 BFP 进行了对比实验。此外，为了验证 neck 阶段上采样时对小目标造成了信息损失，我们将最近邻插值替换为双线性插值法，结果表明，使用双线性插值的上采样比最近邻法插值的上采样的检测精度高，从而验证了 neck 阶段上采样对小目标会造成信息损失的假设。我们叠加了相邻特征融合模块，实验表明叠加后对小目标的检测精度会下降，这是因为叠加后会使得插值次数增多，插值次数增多会使得背景冗余信息造成更大的误差，从而降低小目标的检测精度。

本实验的实验平台为 Ubuntu18.04，使用 GTX1080TI 显卡，cuda10.1，pytorch 1.3.1，mmdetection[18] 2.4。

4.1 Lisa 交通标志数据集

Lisa 交通标志数据集是一组包含美国交通标志的视频和注释帧，共包含 47 个类，6 610 张图像，共有从 6×6 像素到 167×168 像素范围的 7 855 个目标。

首先，我们修改 mmdetection 中的 coco 数据集，将其中的 80 个类别替换为 Lisa 数据集的 47 个类别。通过修改 mmdetection 框架源码，向该框架中添加相邻特征融合网络模块、Double RoI 模块和递归模块，并分别进行了实验。

实验数据表明，虽然 BiFPN 和 BFP 对大目标的检测精度提升较大，相比 baseline 精度分别提高了 7.4 和 13 个百分点，但是由于小目标检测精度下降，导致 BiFPN 和 BFP 的 *mAP* 下降；而本文提出的 NFPN 方法相比 BiFPN 和 BFP，在提高大目标检测精度的同时，也提高了小目标的检测精度和 *mAP*。同时使用 Double RoI 和递归 NFPN，可以提高各尺度目标的精度。本文使用了双层 NFPN（即前一个 NFPN 的各层输出特征作为后一个 NFPN 网络的各层输入特征），相比 NFPN，其对大目标的检测精度提高了 4.7 个百分点，对中目标的检测精度下降了 1.8 个百分点，对小目标的检测精度下降了 1.8 个百分点，*mAP* 下降 1.4 个百分点，从而验证了特征融合时，上采样对小目标特征会造成很大的信息损失而对大目标特征具有信息增强作用。

4.2 VOC 数据集

为了检测相邻特征融合方法对于检测网络的性能提升，我们在 VOC 数据集上进行了测试，训练集为 VOC2007train 和 VOC2012 train 的合集，测试集为 VOC2007，在以 ResNet50 为 Backbone 的 Faster RCNN 上，相比 baseline，使用 NFPN 检测的 *mAP* 提高了 1.8 个百分点，以 Faster RCNN 101 作为 baseline，使用 NFPN 检测的 *mAP* 提高了 6%，如表 2 所列。

表 2 在 VOC07+12 上的实验结果

Table 2 Experiment results on VOC07+12 dataset

Method	*mAP*/%	Backbone
Faster RCNN	73.2	VGG-16
Faster RCNN	79.5	Res-50
NFPN(ours)	**81.3**	Res-50
Faster RCNN	76.4	Res-101
NFPN(ours)	**82.4**	**Res**-101

我们以 VOC2007 train 作为训练集，以 Faster RCNN 为 baseline，在 VOC2007 test 上进行对比实验，结果如表 3 所列。

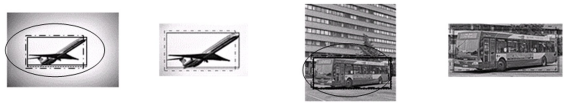
表 3 在 VOC2007 上的实验

Table 3 Experiments on the VOC2007 dataset

Method	*mAP*/%	Backbone
Faster RCNN	69.9	VGG-16
Faster RCNN	72.1	Res-50
NFPN(ours)	**76.3**	Res-50
Faster RCNN	74.4	Res-101
Faster RCNN	74.4	Res2Net-50
Faster RCNN	72.0	Res50-I[19]
NFPN(ours)	**76.4**	ResNet-101

4.3 各尺度检测可视化对比

我们选择了 COCO 数据集中的图片，以 Faster RCNN 和 NFPN 模型的检测框与实际目标框展示可视化的直观效果，如图 6 所示，其中实线是真实目标框，虚线是 NFPN 预测框，点横线是 Faster RCNN 预测框，实线椭圆用于标注目标，在原图的右侧展示放大图，以便于观察差异。



(a)大目标，分辨率大于 96 * 96

(b)中目标，分辨率介于 32 * 32 到 96 * 96 之间

(c)小目标，分辨率小于 32 * 32

图 6 可视化检测结果对比

Fig. 6 Comparison of visual test results

结束语 在交通标志识别中，人们对远方小目标的检测要求越来越高。针对当前两阶段检测网络 neck 结构中大目标和小目标特征不平衡的问题，本文提出了相邻特征融合、Double RoI 和 FRP 方法。相比 Faster RCNN 50，使用 NFPN 方法检测的 *mAP* 提高了 2.6 个百分点，小目标的检测精度提高了 1.6 个百分点，大目标的检测精度提高了 10.6 个百分

点。在 VOC 数据集上使用相同的 Backbone，检测的 *mAP* 最高提高了 6 个百分点。本文提出的 NFPN 和 Double RoI 方法，不仅能够提高 Faster RCNN 目标检测框架下交通标志的检测精度，还能够提高 VOC 数据集中其他物体检测的精度。

参 考 文 献

[1] MOGELMOSE A, TRIVEDI M M, MOESLUND T B. Vision-based traffic sign detection and analysis for intelligent driver assistance systems: Perspectives and survey[J]. IEEE Transactions on Intelligent Transportation Systems, 2012, 13(4): 1484-1497.

[2] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.

[3] REN S, HE K, GIRSHICK R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[C]// Advances in Neural Information Processing Systems. 2015: 91-99.

[4] LIU S, QI L, QIN H, et al. Path aggregation network for instance segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 8759-8768.

[5] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2117-2125.

[6] JOHN M E. The PASCAL Visual Object Classes Challenge 2007(VOC2007) Development Kit[J]. International Journal of Computer Vision, 2006, 111(1): 98-136.

[7] LIU W, ANGUELOV D, ERHAN D, et al. Ssd: Single shot multibox detector[C]// European Conference on Computer Vision. 2016: 21-37.

[8] TAN M, PANG R, LE Q V. Efficientdet: Scalable and efficient object detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 10781-10790.

[9] PANG J, CHEN K, SHI J, et al. Libra r-cnn: Towards balanced learning for object detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 821-830.

[10] HE K, GKIOXARI G, DOLLÁR P, et al. Mask r-cnn[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017: 2961-2969.

[11] REDMON J, FARHADI A. Yolov3: An incremental improvement[J]. arXiv: 1804. 02767, 2018.

[12] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017: 2980-2988.

[13] CAI Z, VASCONCELOS N. Cascade r-cnn: Delving into high quality object detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 6154-6162.

[14] QIAO S, CHEN L C, YUILLE A. DetectoRS: Detecting Objects with Recursive Feature Pyramid and Switchable Atrous Convolution[J]. arXiv: 2006. 02334, 2020.

[15] LIU Y, WANG Y, WANG S, et al. CBNet: A Novel Composite Backbone Network Architecture for Object Detection[C]// AAAI. 2020: 11653-11660.

[16] SHRIVASTAVA A, GUPTA A, GIRSHICK R. Training region-based object detectors with online hard example mining [C]// IEEE Conference on Computer Vision & Pattern Recognition. IEEE Computer Society, 2016: 761-769.

[17] CHEN K, WANG J, PANG J, et al. Mmdetection: Open mmlab detection toolbox and benchmark[J]. arXiv: 1906. 07155, 2019.

[18] GAO S, CHENG M M, ZHAO K, et al. Res2Net: A New Multi-scale Backbone Architecture[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(2): 652-662, 1.

[19] WANG T, YUAN L, ZHANG X, et al. Distilling object detectors with fine-grained feature imitation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 4933-4942.

LI Ya-ze, born in 1991, postgraduate. His main research interests include computer vision and object detection.

LIU Hong-zhe, born in 1971, Ph.D. Her main research interests include computer vision, deep learning, media semantic computing, etc.