计算机科学
COMPUTER SCIENCE

基于改进D2Det尺度自适应目标检测算法研究

王玲，黄冠，王鹏，白燕娥，邱天衡

引用本文

王玲，黄冠，王鹏，白燕娥，邱天衡. 基于改进D2Det尺度自适应目标检测算法研究[J]. 计算机科学, 2023, 50(11A): 221100247-9.

WANG Ling, HUANG Guan, WANG Peng, BAI Yane, QIU Tianheng. Study on Scale Adaptive Target Detection Algorithm Based on Improved D2Det [J]. Computer Science, 2023, 50(11A): 221100247-9.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.221100247

基于改进 D2Det 尺度自适应目标检测算法研究

王 玲 黄 冠 王 鹏 白燕娥 邱天衡
长春理工大学计算机科学技术学院 长春 130022

摘 要 针对 D2Det(Towards High Quality Object Detection and Instance Segmentation)面对尺度变化目标和小目标的检测效果不佳并且参数量较大的问题,基于 D2Det 提出一种尺度自适应的目标检测模型 G-SAD2Det。首先在数据预处理阶段引入数据增强算法 CutOut 和 Mosaic,使模型应对复杂场景时有较好的鲁棒性;其次改进特征提取网络 ResNet,在每个残差块内构建多尺度特征提取结构,从细粒度层面上更好地提取目标特征,同时在网络结构上添加可切换的全局上下文语义特征提取模块,通过不同池化层来增强显著性特征和全局上下文语义信息;然后改进候选框生成模块,采用自主定位目标中心区域指导候选框的生成,增强算法对尺度变换目标的自适应能力;最后通过 Ghost 卷积替换普通卷积降低网络的参数量和计算量。使用 VOC 数据集和 COCO 子数据集验证算法的有效性,G-SAD2Det 比 D2Det 在两个数据集上的 mAP@0.5 分别提升了 3.6%和 4.9%;模型参数量减少了 27.42%,计算量减少了 35.96%,证明改进后的算法在提高了精度的同时也减少了计算量。

关键词:目标检测;尺度自适应;多尺度特征提取;残差块;区域指导候选框
中图法分类号 TP391.41

Study on Scale Adaptive Target Detection Algorithm Based on Improved D2Det

WANG Ling,HUANG Guan,WANG Peng,BAI Yane and QIU Tianheng
School of Computer Science and Technology,Changchun University of Science and Technology,Changchun 130022,China

Abstract Aiming at the problem that D2Det(Towards High Quality Object Detection and Instance Segmentation) has poor detection effect and large parameter quantity in the face of scale change targets and small targets,this paper proposes a scale adaptive target detection model G-SAD2Det based on D2Det. Firstly,in the data preprocessing stage,the data enhancement algorithms CutOut and Mosaic are introduced,and the model has good robustness when dealing with complex scenes. Secondly,the feature extraction network ResNet is improved,the multi-scale feature extraction structure is built in each residual block,and the target features are better extracted from the fine-grained level. At the same time,the switchable global context semantic feature extraction module is added to the network structure,and the salience features and global context semantic information are enhanced through different pooling layers. Then,the candidate frame generation module is improved,and the center area of the self-locating target is used to guide the generation of the candidate frame,so that the adaptive ability of the algorithm to the scaling target can be enhanced. Finally,replacing ordinary convolution with Ghost convolution to reduce the amount of network parameters and computation. VOC data set and COCO sub-data set are used to verify the effectiveness of the algorithm,the mAP@0.5 value of G-SAD2Det increases by 3.6% and 4.9% respectively,compared with D2Det in the two data sets. The number of model parameters reduces by 27.42% and the amount of calculation reduces by 35.96%. It is proved that the improved algorithm not only improves the accuracy,but also reduces the amount of computation.

Keywords Object detection,Scale adaptive,Multi-scale feature extraction,Residual element,Regional guidance candidate box

1 引言

目标检测作为机器视觉的一个重要研究方向,目前在各个领域都具有重要的作用,大到资源分析、天气预测、环境检测,小到车牌识别、人脸识别、工业图像识别等[1]。近年来,基于深度学习的目标检测算法逐渐展现了其优势,并提出很多优秀的算法,如 R-CNN[2],Faster R-CNN[3],Cascade[4] 等。文献[5]总结了近年来各种目标检测算法及其改进方法。在实际应用场景中,人们对目标检测的精度要求越来越高,其中尺度自适应能力又是制约检测精度的一个重要原因[6],因此文献[7-9]针对特定领域,对 Faster R-CNN 的尺度自适应能力进行改进,但几种模型均没有对一般数据集的性能进行验证,如 VOC,COCO 等。D2Det[10] 是目前二阶段目标检测性能较好的算法,它基于 Faster R-CNN 进行改进,适用于大部分场景的目标检测,通过特征金字塔网络(Feature Pyramid Networks,FPN)[11] 实现了一定的尺度自适应。Zhang 等[12] 提出通过改进传统注意力模块并引入特征融合的方式来提高浅层特征图的表征能力的目标检测器。Yu 等[13] 提出自适应

基金项目:中央引导地方科技发展基金吉林省基础研究专项(202002038JC)
This work was supported by the Jilin Provincial Basic Research Project of the Central Leading local Science and Technology Development Fund (202002038JC).
通信作者:王玲(wangling0912@cust.edu.cn)

多尺度特征(Adaptive Multiscale Features: AMF),这是一种具有双向特征融合的新多尺度特征融合方法,进一步提升了尺度自适应能力。但是上述算法仅仅在 FPN 层面进行了改进,当目标尺度频繁发生变化时,仍存在漏检和错检的情况;同时,其也受到硬件条件的制约,较难满足实时性和高精度要求的场景[14]。为了解决这个问题,本研究基于 D2Det,提出了一种针对一般数据集、且具有高精度的轻量化尺度自适应目标检测算法 G-SAD2Det(Ghost-Scale Adaptation D2Det),主要贡献如下:

(1)引入 CutOut[15] 和 Mosaic[16] 数据增强方法对输入图像做预处理,解决被检测目标发生旋转、部分特征被遮挡时检测效果不佳的问题,在不增加模型参数量的基础上增强算法的鲁棒性。

(2)改进 ResNet(Residual Network)特征提取模块并进行轻量化处理。在 ResNet 每个残差块内构造具有等级制的类似残差连接,取代 ResNet 的单个 3 * 3 卷积核,通过前后层特征融合挖掘更深层次的特征,在细粒度级别上表示多尺度的特征,增强尺度自适应能力[17]。之后使用 Ghost 卷积模块[18] 替换普通卷积,通过成本低廉的线性变换生成更多的特征图,减少网络模型参数量,使模型以更少的参数量、更快的速度获得更好的检测效果。

(3)提出可切换的全局上下文语义特征提取模块(Switchable Global Context Feature Extraction Semantic Module, SFESM)。首先在每个残差块连接处、骨干网络输出端添加 SFESM,使用全局最大池化提取目标的显著性特征,增强特征提取能力;然后在 GA-RPN(Region Proposal by Guided Anchoring)[19] 输入端添加 SFESM,使用全局平均池化获取全局上下文语义信息,更加准确地选择目标中心区域。

(4)改进候选框生成模块。GA-RPN 在特征图各个目标的中心区域利用可变形卷积指导生成更贴合目标尺寸的候选框,替换 RPN 在特征图上生成固定尺寸候选框的策略,增强目标检测的尺度适应能力,提高尺度变化目标的检测效果。

2 改进 D2Det 尺度自适应目标检测算法

2.1 算法结构

为了提高目标检测算法针对尺度变化目标和小目标的检测精度,同时降低算法对硬件的要求,本研究基于 D2Det 提出一种同时具备轻量化和高精度的尺度自适应目标检测算法 G-SAD2Det,网络结构如图 1 所示。G-SAD2Det 由数据预处理模块(CutOut、Mosaic),Res2sNet,SFESM,FPN,GA-RPN 和 D2Det Head 这 6 部分组成。首先在数据预处理阶段使用 CutOut 和 Mosaic 方法提高算法应对目标旋转、遮挡等问题的能力;Res2sNet 模块主要完成目标的特征提取,分为 5 个阶段,经过一次卷积操作后,特征图大小下降两倍,通道数上升两倍,它和 ResNet 的区别在于,Res2sNet 在残差块内部构建多尺度残差连接,在细粒度层面上提取不同尺度的目标特征,其中 GhostConv 表示该层将普通卷积替换成了 Ghost 卷积;在 Res2sNet 每一层输出端添加 SFESM,通过提高对显著性特征的关注度增强网络的特征提取能力;FPN 模块通过 1 * 1 卷积操作,将后 4 层通道数固定为 256,每一层通过上采样与下一层进行特征融合,实现不同尺度特征信息的传递,解决多尺度目标检测精度低的问题,经过 FPN 得到的特征图再通过 SFESM 获取到全局上下文的语义信息,帮助 GA-RPN 根据语义信息准确地选择目标中心区域,生成用于预测目标的候选框,经过非极大值抑制[20] 筛选后传给二阶段检测头模块(D2Det Head);D2Det Head 完成对目标的定位和分类,Dense Local Regression 首先对区域候选框进行特征池化操作,生成固定大小空间邻接局部特征,然后对内部的每个特征点做偏移计算,完成对目标的定位;Disriminative RoI pooling 使用 RoI Align 获取 RoI 特征,经过自适应加权操作获取具有更高辨识度的目标特征,完成对目标的分类。

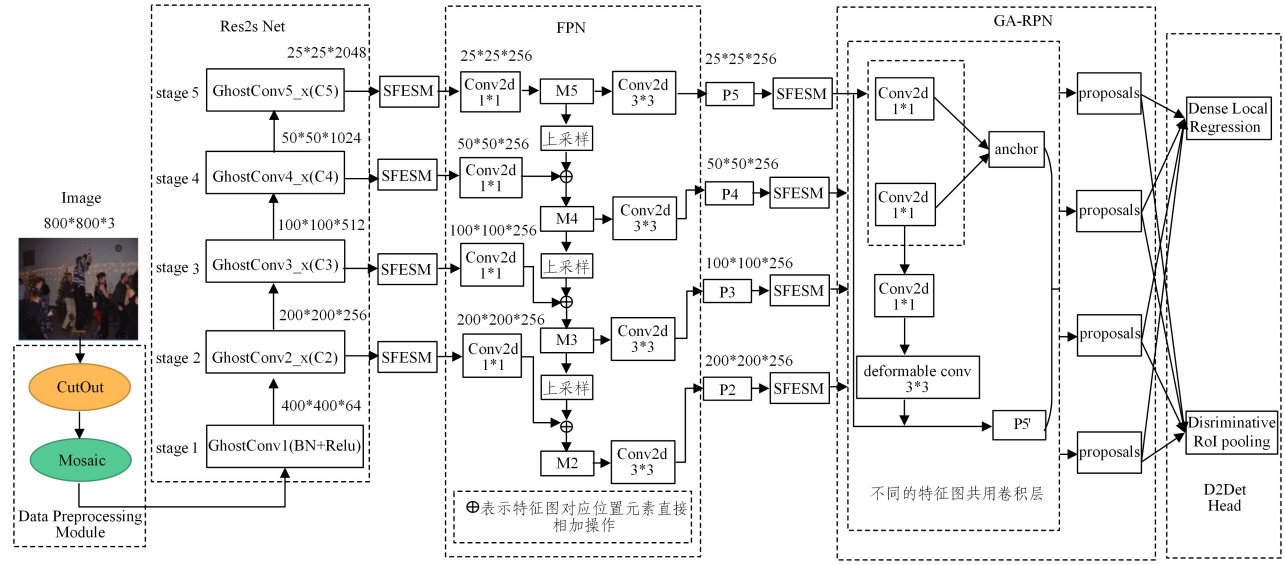


图 1 G-SAD2Det 算法整体网络框架

Fig. 1 Overall network framework of G-SAD2Det algorithm

2.2 增加数据预处理模块

使用数据预处理方法对图像进行各种变换,可以增强图像样本的多样性,增加训练集图像的数量,因此本文引入 CutOut 和 Mosaic 两种数据增强的方法,在不增加算法复杂度的基础上,提高对待检测物体旋转、部分特征被遮挡等时算法的鲁棒性。CutOut 数据增强方法的效果示例如图 2 所示。

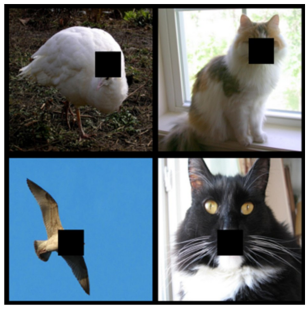


图 2　CutOut 示例图

Fig. 2　CutOut example

该方法的基本思想是通过一个固定大小的掩膜区域模仿遮挡，增加算法对目标整体的关注度，提高卷积神经网络的整体性能和鲁棒性[21]。

Mosaic 数据增强方法的过程如图 3 所示，主要思想是将 4 张图片经过随机缩放、翻转、色域调整、剪裁等操作后拼接成一张图片作为训练数据。这样做可以增加数据的多样性，丰富图片的背景，而且随机缩放可以增加很多小目标，使算法在训练阶段就可以学到更多的小目标特征，图片拼接使得 Mini-batch 并不需要很大就能达到较好的效果。

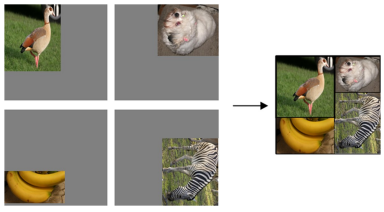


图 3　Mosaic 过程图

Fig. 3　Mosaic process diagram

为验证 Mosaic 和 CutOut 数据增强方法的有效性，本研究对主流的几种数据增强方法进行了对比实验，结果如表 1 所列。

表 1　各方法对比实验

Table 1　Comparison experiment of various methods

样本图像				
方法	Mosaic	CutOut	CutMix[23]	Mixup[22]
$AP_{0.5}$	69.90 (+0.9)	69.60 (+0.6)	69.50 (+0.5)	68.50 (−0.5)

其中样本图像是为了形象地描述操作方法而展示的数据集中的示例，表中的结果是 G-SAD2Det 算法在 COCO 子数据集上的目标检测效果，评价指标 $AP_{0.5}$ 的基础值是 69.00。通过对比实验结果可以看出，Mosaic 和 CutOut 两种数据增强方法较另外两种方法的提升效果明显，因此将 Mosaic 和 CutOut 作为本实验的数据增强方法。

2.3　提出可切换的全局上下文语义特征提取模块 SFESM

为提高特征提取模块对特征的表达和 GA-RPN 自主定位目标中心区域的能力，本文提出了可切换的全局上下文语义特征提取模块 SFESM，在较少地增加算法复杂度的同时，有效地学习目标的显著性特征和全局上下文信息。本模块实现了即插即用，可高效地嵌入到目标检测网络中，操作流程如图 4 所示。

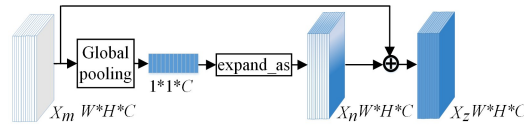


图 4　可切换的全局上下文语义特征提取模块

Fig. 4　Switchable global context semantic feature extraction module

首先接收输入 Xm 的通道数为 C，经过全局池化操作获取特征图的关键信息，然后通过上采样得到和原维度一样的特征图 Xn；最后经过特征相加融合得到 Xz。其中池化层(Pool layer)可采用全局平均池化[24]或者全局最大池化[25]。全局最大池化是对每个特征图的像素值取最大值，只找分数最高的区域，忽略低分数区域，可以增强显著性特征，因此在 Res2sNet 特征提取模块每个残差块连接处和骨干网络输出端添加的 SFESM 模块采用全局最大池化处理。全局平均池化是对每个特征图的像素值取平均值，增强全局语义信息的表达，GA-RPN 基于语义信息自主定位目标中心区域，指导锚框的生成，因此在 GA-RPN 输入端添加的 SFESM 模块采用全局平均池化处理。

2.4　改进特征提取模块

在特征提取阶段，ResNet 通过叠加浅层网络增强特征的表达能力，残差块可以在不增加网络深度的情况下较好地提取目标特征，但随着对目标检测算法的精度要求越来越高，检测场景越来越复杂，ResNet 较难应用在对精度要求较高的场景[26]。在存在多尺度目标和小目标的场景，ResNet 在特征提取过程中会逐渐忽略掉细粒度信息，导致检测精度降低，鲁棒性变差。为了解决这个问题，本研究改进了 D2Det 的骨干网络 ResNet，利用等级制的类似残差连接思想，将 ResNet 每个残差块中 1 个 n 通道 3 * 3 卷积替换成 3 个 $n/4$ 通道的 3 * 3 卷积，在每个残差块上进行分层多尺度特征提取，增强对细粒度的特征关注度，使其有更强的特征提取能力，并且在残差块连接处加入 SFESM 模块，使得网络的特征表达能力得到进一步提升，一个残差块的结构如图 5 所示，图中改进部分做了上色处理。

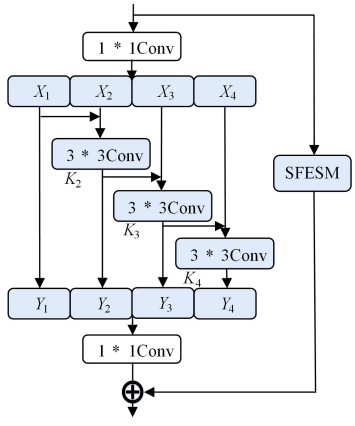


图 5　Res2sNet 网络残差块结构

Fig. 5　Res2sNet network residual block structure

特征图经过 1 * 1 卷积后，按照通道数平均分成 4 份，定义为 X_i，$i \in \{1,2,3,4\}$。替换后的 3 * 3 卷积通过类似残差网络的模式逐层连接，可以在细粒度层面上更好地提取不同尺度的目标特征。

X_i 省去了 3 * 3 卷积操作，其他特征图与前一组生成的特

征图 Y_{i-1} 经过卷积操作 $K_i(X_i+Y_{i-1})$ 进行特征提取后生成 Y_i，其中 Y_i 为每一组生成的特征图，具体计算如式(1)所示：

$$Y_i=\begin{cases}X_i, & i=1\\ K_i(X_i+Y_{i-1}), & 1<i\leqslant 4\end{cases} \tag{1}$$

再将 4 组特征图 Y_i 进行拼接，通过 1 * 1 的卷积核进行通道固定和信息融合。最后将前一个残差块的输出通过 SFESM 模块处理后和本残差块输出特征融合，生成表达能力更强的特征图。

Res2sNet 整体架构如表 2 所列，一共有 5 个阶段，第一个阶段由一个 7 * 7 卷积构成，其余 4 个阶段分别是由 3，4，23，3 个残差块构成，每个残差块都是由 2 个 1 * 1 和 3 个 3 * 3 卷积构成。

表 2 Res2sNet 整体架构

Table 2 Overall architecture of Res2sNe

层数	尺寸	步长	卷积结构
Conv1	400 * 400	2	7 * 7,64
Conv2	200 * 200	2	3 * 3 max pool,64 $\begin{pmatrix}1*1,64\\(3*3,16)*3\\1*1,256\end{pmatrix}*3$
Conv3	100 * 100	2	$\begin{pmatrix}1*1,128\\(3*3,32)*3\\1*1,512\end{pmatrix}*4$
Conv4	50 * 50	2	$\begin{pmatrix}1*1,256\\(3*3,64)*3\\1*1,1024\end{pmatrix}*23$
Conv5	25 * 25	2	$\begin{pmatrix}1*1,512\\(3*3,128)*3\\1*1,2048\end{pmatrix}*3$

2.5 改进候选框生成模块

传统候选框网络 RPN 使用滑窗在特征图的每个特征点上生成几组固定尺寸的候选框，这样整个特征图就会产生几千甚至几万个候选框，这些候选框当中绝大多数都没有目标信息，不但会降低识别精度，而且还会占用很多计算资源，并且固定大小候选框无法契合所有的目标，鲁棒性较差，不能应用在识别精度要求较高的场景[27]。为了解决上述问题，本研究使用 GA-RPN 模块替换 RPN 模块，其中一层的结构如图 6 所示。

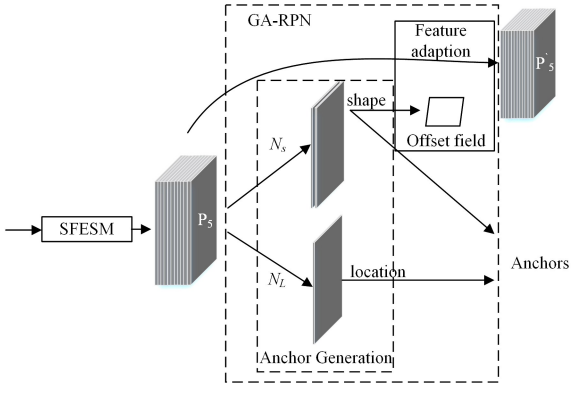


图 6 GA-RPN 网络

Fig. 6 GA-RPN network

把经过 SFESM 模块处理后得到的特征图 $P5$ 输入到 GA-RPN 模块，图中 N_L 表示位置分支，使用一个 1 * 1 的卷积获得关于目标的分数图，然后通过 Sigmoid 函数转化成概率值，产生和特征图 $P5$ 大小相同的概率映射矩阵 $\boldsymbol{p}(i,j|P5)$，再根据预先设置好的阈值(对比实验中设置为 0.05)过滤掉不可能存在目标的区域；图中 Ns 表示形状预测分支，首先对特征图 $P5$ 做一个 1 * 1 的卷积操作，产生两个通道映射 dw 和 dh。由于物体尺度变化范围很大，直接回归 w 和 h 可能导致计算不稳定，因此使用式(2)进行变换：

$$w=\partial\cdot s\cdot e^{dw},h=\partial\cdot s\cdot e^{dh} \tag{2}$$

其中，∂ 为经验尺度因子，s 为步长，e 为指数函数 exp。经过非线性变换后，可将[0,1000]之间的数值映射到[-1,1]区间，使得 Ns 的计算变得更简单、更稳定。

为了使特征具有较好的表达效果，使用 3 * 3 的可变形卷积[28]实现特征转化，公式如式(3)所示：

$$f_i'=N_T(f_i,w_i,h_i) \tag{3}$$

其中，f_i 是指第 i 个位置的特征，(w_i,h_i) 是对应的候选框的形状，$N_T()$ 表示 3 * 3 的可变形卷积操作，不同位置的卷积核采样点会根据图像内容发生自适应的变化，从而适应不同物体的形状、大小等几何形变，可以更加贴合目标的实际尺寸，进一步提高算法的尺度自适应能力。图 6 的 Feature adaption 模块根据 Ns 的输出预测出偏移量，再通过 $N_T()$ 在原始特征图上获得 f_i'。

GA-RPN 预测的每个目标中心点都只与一个逐步预测形状的候选框相对应，大大降低了候选框的生成数量，并且由于候选框的形状和目标的位置具有较大的关联性，GA-RPN 的候选框生成方法比 RPN 可实现更高的召回率。

由于 GA-RPN 比 RPN 多了位置预测分支和形状预测分支，因此需要计算候选框的位置损失 L_{loc} 和形状损失 L_{shape}，联合损失的计算如式(4)所示：

$$L=\partial_1 L_{loc}+\partial_2 L_{shape}+L_{cls}+L_{reg} \tag{4}$$

其中，∂_1 和 ∂_2 为多任务损失加权系数，L_{cls} 和 L_{reg} 为 RPN 中的分类和回归损失函数。

在训练位置预测分支时，将有效候选框设为 1，无效候选框设为 0。此外，为了使目标邻域附近生成更多候选框，首先将标注框 (xg,yg,wg,hg) 映射到特征图上，得到 $(x'g,y'g,w'g,h'g)$。然后将每个标注框都定义 3 个类型：中心区域 $(x'g,y'g,\varphi_1 w'g,\varphi_1 h'g)$ 取的是标注框的中心位置，φ_1 一般取值为 0.2；忽略区域 $(x'g,y'g,\varphi_2 w'g,\varphi_2 h'g)$ 和外部区域都被标注为 0，不参与训练，使用 Focal Loss[29] 来训练位置预测分支。

在训练形状预测分支时，使用基于 *smooth* L_1 的 *bounded iou loss* 进行训练，如式(5)所示：

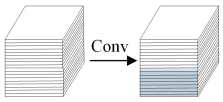
$$L_{shape}=L_1\left(1-\min\left(\frac{w}{w_g},\frac{w_g}{w}\right)\right)+L_1\left(1-\min\left(\frac{h}{h_g},\frac{h_g}{h}\right)\right) \tag{5}$$

其中，L_1 为 *smooth* L_1，w_g 和 h_g 为标注框的宽和高，w 和 h 为预测的候选框的宽和高。

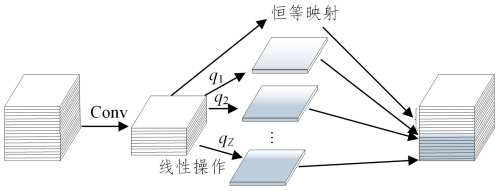
2.6 轻量化网络结构

由于特征提取模块包含上百层卷积块，是其他模块的几十倍，所以本研究通过对特征提取模块 Res2sNet 进行轻量化达到对整个网络结构轻量化的目的。Ghost 模块的基本思想是使用更少的参数来生成更多的特征图，具体来说就是根据特征图之间的联系，将深度学习中的每个卷积过程拆分为两步，首先通过普通卷积生成少量的特征图，然后将得到的特征图进行廉价线性操作生成 Ghost 特征图，最后将两组特征图

按照通道拼接，生成足够多的特征图来匹配输出通道数。图 7 给出了 Ghost 卷积和普通卷积的对比示意图：Ghost 模块由一个少量卷积、一个恒等映射和 $Z*(S-1)$ 个线性运算组成。本研究使用 Ghost 卷积对模型进行轻量化。



(a)普通卷积示意图



(b)Ghost 卷积示意图

图 7　Ghost 卷积和普通卷积示意图

Fig. 7　Ghost convolution and general convolution

本研究以此为基础，对特征提取模块 Res2sNet 的卷积层进行了轻量化改进，改进后的 Res2sNet 模块的残差块如图 8 所示，上色部分为 Ghost 卷积。特征图经过 1 * 1 卷积后，按照通道数平均分成 4 份，定义为 X_i，$i\in\{1,2,3,4\}$。首先采用普通的 1×1 卷积对 X_1 进行通道数的压缩生成 Y_{11}，然后再对 Y_{11} 进行深度可分离卷积（逐层卷积）得到 Y_{12}，然后将 Y_{12} 和 Y_{11} 特征融合生成 Z_1；其他特征图 X_i 与前一组生成的特征图 Z_i 采用普通的 1×1 卷积进行通道数的压缩生成 Y_{i1}，然后再对 Y_{i1} 进行深度可分离卷积（逐层卷积）得到 Y_{i2}，然后将 Y_{i2} 和 Y_{i1} 特征融合生成 Z_i，再将 4 组特征图 Z_i 进行拼接，通过 1 * 1 的卷积核进行通道固定和信息融合；最后将前一个残差块的输出通过 SFESM 模块处理后与本残差块的输出进行特征融合并当作下一层的输入特征图。

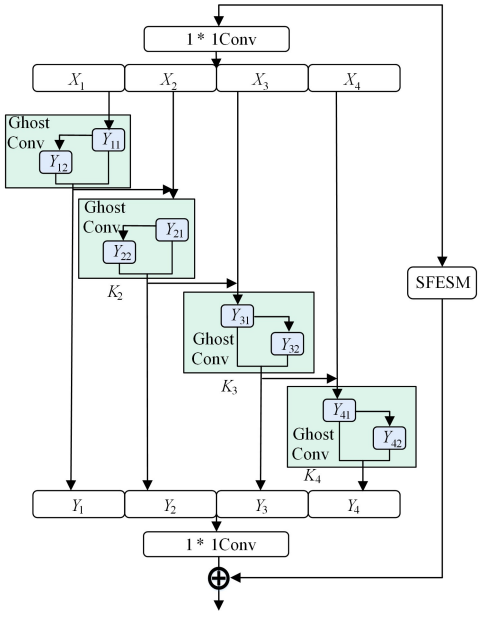


图 8　Ghost 残差结构示意图

Fig. 8　Schematic diagram of Ghost residual structure

轻量化后的模型在保证精度降低最少的条件下，大大减少了参数量和计算量，提升了网络的运行速度，表 3 为输入尺寸为 1 333 * 800 图像在轻量化后和 D2Det 模型的对比结果。

表 3　模型结果对比

Table 3　Comparison of model results

算法	参数量	GFLOPs	模型大小
D2Det	78.24×10^6	354.98	628×10^6
G-SAD2Det	$\mathbf{56.79\times10^6}$	**227.34**	$\mathbf{302\times10^6}$

3　实验结果分析

3.1　实验环境

实验的软硬件环境如表 4 所列。

表 4　实验环境

Table 4　Experimental environment

名称	标准
CPU	Intel i7-10700KF 3.8GHz
Memory	64GB
GraphicsCard	GeForce RTX^{TM} 2080Ti
操作系统	Ubuntu 20.04.3 LTS
MMdet	2.1
Python/PyTorch	3.7/1.4

实验的参数设置如表 5 所列。

表 5　实验参数

Table 5　Experimental parameter

名称	标准
神经网络优化器	采用随机梯度下降法（Stochastic Gradient Descent，SGD）
初始学习率	0.02
动量参数	0.9
权值衰减参数	0.0001
训练轮数	24

训练时采用 warm up 热身策略，先采用较小的学习率进行训练使网络适应数据，之后再恢复到设置的初始学习率。

本研究使用 VOC 数据集和 COCO 子数据集进行实验。VOC 数据集包括人、动物、交通车辆、室内用品 4 大类别，细分为 20 个小类别；COCO 子数据集是从 COCO 数据集的各大类别中分别选取几个小类别，组成的包含 12 个类别的数据集，训练集、验证集和测试集的划分如表 6 所列。

表 6　数据集

Table 6　Data sets

用途	数据集名称	图像数量/张
训练集	VOC	8218
	COCO 子数据集	20752
验证集	VOC	8333
	COCO 子数据集	6991
测试集	VOC	4952
	COCO 子数据集	6991

3.2　评价标准

VOC 数据集采用平均精度均值 mAP（mean Average Precision）、参数量（Parameters）、计算量和 FPS 对目标检测算法进行评测。mAP 指各类别 AP 的平均值，AP 是指平均精度，表示该类别 Precision-Recall 曲线下面积，面积越大表示识别的精度越高，因此，mAP 数值越大，目标检测效果越好。

mAP 使用式(6)进行计算：

$$mAP=\frac{\sum_{i=1}^{n}AP_i}{n} \tag{6}$$

其中，n 是数据集中类别的个数，本实验取值为 20。

参数量（Parameters）是指网络模型中需要训练的参数总

量，用来衡量模型大小，数值越小，代表网络模型结构越简单，对显存的要求越低。参数量计算主要包括卷积层和全连接层的参数，如式(7)所示：

$$Parameters=\begin{cases}(K\times K\times C_i+1)\times C_o, & \text{卷积层}\\(W\times H\times N+1)\times F, & \text{全连接层}\end{cases} \tag{7}$$

其中，K 为卷积核的边长，C_i 为输入通道数，C_o 为输出通道数，W 和 H 为输入特征图的宽和高，N 为卷积核的数量，F 为全连接神经元的数量，+1 表示偏置。

计算量用于衡量模型的复杂度，一般复杂度越小，对 GPU 的运算能力要求越低。

$$FLOPs=\begin{cases}(K\times K\times C_i)+(K\times K\times C_i)\times C_o\times W\times H\\[(2\times F_i-1)+1]\times F_o\end{cases} \tag{8}$$

其中，$GFLOPs=10^9$ FLOPs，F_i 表示全连接输入的神经元数，F_o 表示全连接输出的神经元数。

FPS(Frames Per Second)是指每秒钟执行的帧数，即每秒可以检测多少张图片，用来衡量模型的速度，数值越大表示检测的速度越快，实时性越强。

COCO 子数据集使用 AP 作为评价指标，包括 $AP_{0.5}$，$AP_{0.75}$，AP_s，AP_m，AP_l。和 COCO 数据集一样，$AP_{0.5}=mAP_{0.5}$，$AP_{0.75}=mAP_{0.75}$，$AP_{0.5}$ 的计算公式如式(6)所示，其中 n 取值为 12。$AP_{0.5}$ 和 $AP_{0.75}$ 中的 0.5 和 0.75 是指 IoU 的阈值，大于这个阈值表示成功检测到物体。IoU 的计算式如(9)所示：

$$IoU=\frac{region(A)\cap region(G)}{region(A)\cup region(G)} \tag{9}$$

其中，$region(A)$ 为目标检测算法预测的区域，$region(G)$ 为目标物体的真实区域。

其他 3 个指标 AP_s，AP_m，AP_l 分别指目标所占的像素面积小于 32^2、在 32^2 和 96^2 之间和大于 96^2 时的平均精度。

3.3 定量评价

3.3.1 VOC 数据集上的实验结果

本文算法在 VOC 数据集上与 SSD[30]，Cascade[4]，Faster R-CNN[3]，Grid[31]，Libra[32]，YOLOF[33]，D2Det[10]，DetectoRS[34]算法进行比较，输入图片的尺寸均为 1 333 * 800，实验结果如表 7 所列。

表 7 VOC 数据集实验结果

Table 7 VOC dataset experimental results

算法	参数量/M↓	GFLOPs↓	模型大小/M↓	mAP↓	FPS↑
Faster R-CNN	61.99	285.16	460	79.65	11.6
Cascade	87.98	285.38	672	79.08	10.4
SSD	24.15	361.63	184	76.66	48.5
Grid	83.24	396.78	636	80.21	7.5
YOLOF	42.13	103.14	322	79.88	18.3
Libra	62.25	286.26	462	80.80	8.7
D2Det	78.24	354.98	628	80.30	8.5
DetectoRS	123.26	241.26	943	81.30	6.2
G-SAD2Det	**56.79**	**227.34**	**302**	**83.90**	**11.7**

由表 6 可知，在 9 组对比实验中，算法 G-SAD2Det 与单阶段算法 YOLOF 相比，虽然模型参数量稍大，但是在 mAP 方面比其提升 4.02%，而且在 FPS 上未落后太多，GFLOPs 位列第二，仅次于单阶段检测算法 YOLOF，属于二阶段检测算法中模型复杂度最低的，模型大小排第二位。mAP 为对比实验中效果最好的，比对比算法中表现最好的 DetectoRS 提升了 2.6%。FPS 排第三位，稍次于单阶段检测算法 SSD 和 YOLOF，但在二阶段检测算法中属于最快的。G-SAD2Det 相比大部分网络不仅复杂度更小，而且在精度上具备显著优势，其相比基础算法 D2Det 无论在参数量和计算量还是检测速度上都存在较好的优势，使得模型在训练和检测时对硬件的要求降低，可以被应用在一些低成本但高精度要求的工业场景检测上。总体来看，在高精度和低设备同时兼顾的场景，算法 G-SAD2Det 在众多模型中的表现更加出色。

3.3.2 COCO 子数据集上的实验结果

本文算法在 COCO 子数据集上与 CenterNet[35]，Cascade[4]，Faster R-CNN[3]，Grid[31]，Libra[32]，YOLOF[33]，D2Det[10]，DetectoRS[34]算法进行比较，实验结果如表 8 所列。

表 7 COCO 子数据集实验结果

Table 7 Experimental results of COCO sub-dataset

Method	AP	$AP_{0.5}$	$AP_{0.75}$	AP_s	AP_m	AP_L
Faster R-CNN	43.1	64.8	46.9	20.8	41.5	56.9
YOLOF	43.7	62.7	47.4	22.4	42.6	57.6
Libra	43.8	65.6	48.3	22.6	42.3	57.3
Grid	44.8	64.8	48.7	22.3	42.7	59.2
Cascade	45.0	64.6	49.3	22.0	43.4	59.7
CenterNet	45.1	64.9	49.3	22.1	45.5	57.7
D2Det	45.4	65.5	49.5	22.4	44.6	58.1
DetectoRS	46.1	66.2	50.4	22.5	45.0	59.4
G-SAD2Det	**48.1**	**70.4**	**52.6**	**24.2**	**47.2**	**62.4**

从表 8 可以看出，G-SAD2Det 的 AP 相较于对比算法中表现最好的 DetectoRS 提升 2.0%，IoU 的阈值为 0.5 时比对比算法中表现最好的 DetectoRS 提升 4.2%，IoU 的阈值为 0.75时比对比算法中表现最好的 DetectoRS 提升 2.2%，AP_s 比对比算法中表现最好的 Libra 高出 1.6%，AP_m 比比较算法中表现最好的 CenterNet 高出 1.7%，AP_L 比比较算法中表现最好的 Cascade 高出 2.7%。

3.3.3 实验过程分析

本文对改进算法 G-SAD2Det 进行训练，得到迭代步数为 20 万步的网络训练收敛曲线，使用 MMdet 官方自带的日志可视化分析工具分析了网络的训练过程，如图 9 所示。

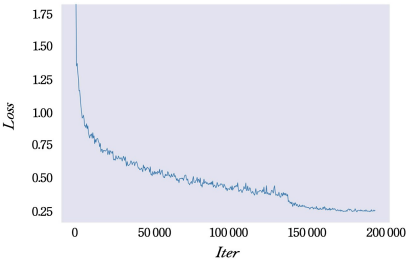


图 9 总损失值曲线图

Fig. 9 Total loss curve

大约在 13 万次迭代后模型达到收敛，平滑后的损失为 0.24。

3.3.4 实验结果分析

本文提出的算法 G-SAD2Det 在两个数据集上均获得了较好的目标检测效果，主要取决于以下 5 点改进：

(1)使用数据增强方法增强了算法对数据集中遮挡问题和旋转问题的鲁棒性；

(2)提出的 SFESM 提高了特征表达能力，使网络更容易

关注不同尺寸的目标；

(3)改进骨干网络 Res2sNet 在细粒度的多尺度特征提取中丰富了多尺度目标的特征信息，并且保留了较小目标特征；

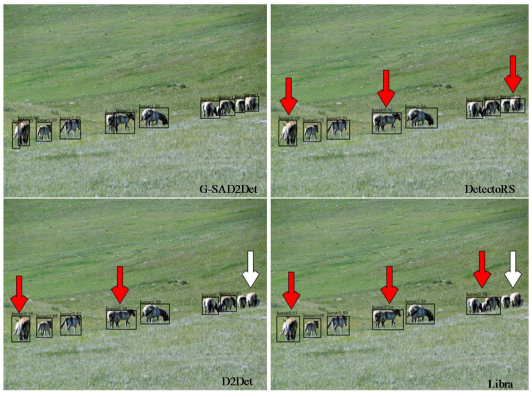
(4)使用 GA-RPN 替换 PRN，生成更加贴合目标实际尺寸的候选框；

(5)模型轻量化后，参数量和计算量下降，进而对算法的实时性能有一定的提升。

3.4　定性评价

本实验从 VOC 数据集和 COCO 子数据集中挑选出具有多尺度和小目标特点的图像进行定性实验，图 10 中的对比结果图分别是 G-SAD2Det，DetectoRS，D2Det，Libra 的检测结果。方框上方的字母和数字分别代表目标的类别和置信度，右下角是检测算法名称，为更清楚地对比，借助箭头进行指示，其中白色箭头代表漏检处，黑色箭头代表错检处。

(a)多尺度目标、复杂背景



(b)小目标

图 10　目标检测定性分析

Fig. 10　Qualitative analysis of target detection

(1)多尺度目标、复杂背景

图 10(a)4 张图片中目标物体的颜色和背景十分相似，有些目标物体被低矮植被遮挡并且目标的大小不一，复杂背景和多尺度物体对目标检测造成了一定的影响。根据箭头指示，DetectoRS 有两处被植被遮挡的目标未检测到；D2Det 也有两处未检测到，一个是较小的目标，另一个是被植被遮挡了头部特征的目标；Libra 漏检较多，较难检测到的目标均未检测到；而 G-SAD2Det 在面对多尺度且背景复杂的目标时保持了良好的检测效果，全部目标均检测成功并且无多余错检框，证明 G-SAD2Det 能提取出更全面更丰富的特征，有更好的尺度自适应能力，并且数据增强后，算法可以学习到全身特征，面对遮挡情况也能有更好的鲁棒性。

(2)小目标

图 10(b)中 4 张图片中目标物体尺度不同、小目标占比多，并且目标之间存在遮挡。根据箭头所示，Libra 和 D2Det 均未识别到最右方的两个目标，而且有多处将产生遮挡的两个目标错检成了一个目标；DetectoRS 虽然成功识别了最右方的目标，但是将两个目标错检成了一个，将其他较难检测的遮挡目标也错检成了一个；只有 G-SAD2Det 成功识别了所有的目标，证明 G-SAD2Det 在细粒度层面防止了小目标信息特征丢失，在拥有尺度自适应能力的情况下能更好地检测到小目标。

为了验证本算法的泛化能力，本文在 Objects365 数据集中选取了具有多尺度、遮挡等特点并且存在与 VOC 数据集相同目标类别的图像作为实验对象，使用在 VOC 数据集上训练的模型对其检测。如图 11 所示，模型成功检测到 Objects365 数据集上的目标，验证了本算法的泛化能力。

图 11　算法泛化性实验

Fig. 11　Algorithm generalization experiment

3.5　消融实验

为了验证 G-SAD2Det 算法的有效性，本研究在 VOC 数据集上进行了 6 组消融实验，每组实验使用相同的超参数以及训练技巧；组 1 为基础算法 D2Det，组 2 使用了数据增强算法；组 3 在组 2 基础上添加 Res2sNet 模块；同上，组 4 添加了 SFESM 模块；组 5 添加了数据增强算法、Res2sNet 模块、SFESM 模块和 GA-RPN；组 6 为本研究轻量化后的最终算法，实验结果如表 8 所列。加入两种数据增强方法后，模型复杂度并未增加，且在 FPS 未下降的情况下 mAP 提升了 1.1%，证明数据增强方法的引入使算法对目标的检测更准确，提高了鲁棒性；在添加 Res2sNet 模块后，虽然算法的参数量和计算量稍有增加且 FPS 略有下降，但是在 mAP 指标上提高了 1.5%，表明改进后的 Res2sNet 对于细粒度的多尺度特征提取起到了增强的作用；加入 SFESM 后 mAP 提高了 0.5%，并且参数量仅仅增加了 0.018，FPS 未下降，证明 SFESM 是一种有效的即插即用模块；添加 GA-RPN 后，FPS 下降了 0.3，参数量和计算量均有明显增加，且在 mAP 上提高 1.1%，表明 GA-RPN 的尺度自适应能力提高了多尺度目标检测的准确性；最后 Ghost 模块的引入，虽降低了算法的精度，但算法的计算量降低了 108.95 GFLOPs，并且参数量降低了 22.331×10^{6}，FPS 提高了 3.9，大大减小了二阶段目标检测算法的局限性，达到速度和精度两方的平衡，最终改进后的模型和 D2Det 相比，mAP 提高了 3.6%，参数量减少了

27.42%，计算量减少了 35.96%，模型对检测的能力进一步加强，且算法运行时对硬件的要求更小，可以被广泛地应用于一些对尺度自适应能力要求高且需要一定实时性的工业任务中。

表 8 消融实验结果

Table 8 Results of ablation test

组	Mosaic+CutOut	Res2sNet	SFESM	GA-RPN	Ghost	mAP	参数量/M	GFLOPs	FPS
1						80.3	78.243	354.98	8.5
2	√					81.4	78.243	354.98	8.5
3	√	√				82.9	79.002	366.04	8.1
4	√	√	√			83.4	79.020	366.05	8.1
5	√	√	√	√		**84.5**	79.123	366.29	7.8
6	√	√	√	√	√	83.9	**56.792**	**227.34**	**11.7**

结束语 在本研究针对尺度变化目标和小目标检测提出了一种轻量化的尺度自适应目标检测算法 G-SAD2Det，采用数据增强、Res2sNet、SFESM 和 GA-RPN 提高了网络的精度；使用 Ghost 模块替换普通卷积对网络进行轻量化，减轻了二阶段检测算法本身模型复杂度高的缺点。并在 VOC 数据集上 mAP 达到了 83.9%，较基础算法 D2Det 提高了 3.6%，模型复杂度方面在参数量和计算量上分别降低了 27.42%和 35.96%；在 COCO 子数据集上 AP 和 $AP_{0.5}$ 达到了 48.1%和 70.4%，较基础算法 D2Det 分别比提高了 2.7%和 4.9%。基于二阶段检测精度高的优势，后续继续在减少模型复杂度方向上进行研究，通过对网络进行剪枝、蒸馏等处理实现模型轻量化；另外，后续也将对精度提升方面进行研究，在多尺度特征融合网络(FPN)上建立多层级连接，保留更多的不同尺度信息，提高算法对特征的敏感度，使其能够适用于更复杂的检测场景。

参 考 文 献

[1] LOU G X, SHI H Z. Face image recognition based on convolutional neural network[J]. China Communications, 2020, 17(2): 117-124.

[2] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarc hies for accurate object detection and semantic segment ation[C]// Proceedings of the IEEE Conference on Com Puter Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2014: 580-587.

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

[4] CAI Z W, VASCONCELOS N. Cascade r-cnn: Delving into high quality object detecti on[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscata way, NJ: IEEE Press, 2018: 6154-6162.

[5] LUO H L, CHEN H K. A Survey oftarget detection based on deep learning[J]. Acta Electronica Sinica, 2020, 48(6): 1230-1239.

[6] ZHAO Y Q, JIA J L, GONG W J, et al. Multi scale aerial image target detection algorithm based on pro-YOLOv4[J]. Computer Engineering & Science, 2021, 38(11): 3466-3471.

[7] ZHANG R M, BI L J, WANG F B, et al. Target detection algorithm based on multi-scale feature fusion and adaptive anchor frame[J]. Laser & Optoelectronics Progress, 2022, 59(12): 420-429.

[8] LIU F, HAN X. Adaptive aerial target detection based on multi-scale depth learning[J]. Acta Aeronautica et Astronautica Sinica, 2022, 43(5): 471-482.

[9] LI Y Z, LIU H Z. Object Detection Based on Neighbour Feature Fusion[J]. Computer Science, 2021, 48(12): 264-268.

[10] CAO J L, CHOLAKKAL H, ANWER R M, et al. D2det: Towards high quality object detection and instances segmentation [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2020: 11485-11494.

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

[12] ZHANG L, ZHOU B W, WU L H. SSD Network Based on Improved Convolutional Attention Module and Residual Structure [J]. Computer Science, 2022, 49(3): 211-217.

[13] YU X Y, WU S Y, LU X Q, et al. Adaptive multiscale feature for object detection[J]. Neurocomputing, 2021, 449: 146-158.

[14] ZENG N Y, WU P S, WANG Z D, et al. A smalll-sized object detection oriented multi-scale feature fusion approach with application to defect detection[J]. IEEE Transactions on Instrumentation and Measurement, 2022, 71: 1-14.

[15] DEVRIES T, TAYLOR G W. Improved regul arization of convolutional neural networks with cutout[J]. arXiv: 1708. 04552, 2017.

[16] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. Yolov4: Optimal speed and accuracy of object detection[J]. arXiv: 2004. 10934, 2020.

[17] GAO S H, CHENG M M, ZHAO K, et al. Res2net: A new multi-scale bac kbone architecture[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 43(2): 652-662.

[18] HAN K, WANG Y, TIAN Q, et al. Ghostnet: More features from cheap operations[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 1580-1589.

[19] WANG J Q, CHEN K, YANG S, et al. Region proposal by guided anchoring[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2019: 2965-2974.

[20] HE Y H, ZHU C H, WANG J R, et al. Bounding box regression with uncertainty for accurate object detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern

Recognition. Piscataw ay，NJ：IEEE Press，2019：2888-2897.

[21] LU J，LIN W，CHEN P，et al. Research on Lightweight Citrus Flowering Rate Statistical Model Combined with Anchor Frame Clustering Optimization[J]. Sensors，2021，21(23)：7929.

[22] ZHANG H，CISSE M，DAUPHIN Y N，et al. mixup：Beyond empirical risk minimization[J]. arXiv：1710. 09412，2017.

[23] YUN S，HAN D，OH S J，et al. Cutmix：Regu larization strategy to train strong classifiers with localizable features[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019：6023-6032.

[24] HU J，SHEN L，SUN G. Squeeze-and-excitation networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway，NJ：IEEE Press，2018：7132-7141.

[25] WOO S，PARK J，LEE J Y，et al. Cbam：Con volutional block attention module[C]// Proceedings of the European Conference on Computer Vision(ECCV). Berlin：Springer，2018：3-19.

[26] BAO Y X，LU T L，DU Y H，et al. Deepfake Videos Detection Method Based on iResNet34 Model and Data Aug mentation [J]. Computer Science，2021，48(7)：77-85.

[27] NAJIBI M，SINGH B，DAVIS L S. Fa-rpn：Floating region prop osals for face detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019：7723-7732.

[28] DAI J F，QI H Z，XIONG Y W，et al. Deformable convolutional networks[C]// Proceedings of the IEEE International Conference on Computer Vision. Piscataway，NJ：IEEE Press，2017：764-773.

[29] 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. Piscataway，NJ：IEEE Press，2017：2980-2988.

[30] LIU W，ANGUELOV D，ERHAN D，et al. SSd：Single shot mul tibox detector[C]// European Conference on Computer Vision. Cham：Springer，2016：21-37.

[31] LU X，LI B Y，YUE Y X，et al. Grid r-cnn[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway，NJ：IEEE Press，2019：7363-7372.

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

[33] CHEN Q，WANG Y M，YANG T，et al. You only look one-level feature[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway，NJ：IEEE Press，2021：13039-13048.

[34] QIAO S Y，CHEN L C，YUILLE A. Detectors：Detecting objects with recursive feature pyramid and switchable atrous convolution[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway，NJ：IEEE Press，2021：10213-10224.

[35] DUAN K W，BAI S，XIE L X，et al. Centernet：Keypoint triplets for object dete ction[C]// Proceedings of the IEEE/CVF Internati ONAL Conference on Computer Vision. Piscataway，NJ：IEEE Press，2019：6569-6578.

WANG Ling，born in 1979，Ph. D. Her main research interests include machine learning and image processing.