基于空间注意力的 CNN 特征增强方法

许 畅,王朝辉
(武汉科技大学 计算机科学与技术学院,湖北 武汉 430065)

摘 要:卷积神经网络一般被用于特征提取,它通过提取图像底层的点、线、面的几何特征,进而映射到高层的语义特征,然而传统的卷积网络只对输入的样本进行宽泛的特征提取,而不会刻意去区分图像的前景和后景,这使得模型提取到的特征包含大量的背景噪声,降低了模型的表征能力。在空间注意力的基础上,提出了一种名为特征增强网络(FA-block)的卷积网络分支,这种网络结构从样本的掩膜中学习目标的空间分布,为原始特征图上的每一个像素点训练得到代表重要程度的权重,然后通过加权的方式突出特征图中的目标部位。此方法旨在抑制背景噪声,增强待学习的目标特征,让主干网络提取到的特征更加纯净。在 PASCAL VOC 数据集上的实验证明了 FA-block 的有效性,最后经过 MS COCO 数据集的验证,FA-block 使得 Faster Rcnn 基线的性能提高了 5.5%。

关键词:计算机视觉;卷积神经网络;空间注意力;特征增强;高频噪声抑制
中图分类号:TP391.41 **文献标识码**:A **文章编号**:1673-629X(2022)06-0074-05
doi:10.3969/j.issn.1673-629X.2022.06.013

Feature Augment of Convolutional Neural Network Based on Spatial Attention

XU Chang, WANG Zhao-hui
(School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430065, China)

Abstract: Convolutional neural network is generally used for feature extraction. It extracts the geometric features of points, lines and surfaces at the bottom of the image, and then maps them to high-level semantic features. However, the traditional convolution network only extracts general features from the input samples, instead of deliberately distinguishing the foreground and background, which makes the features extracted by the model contain a lot of background noise and weakens its representation ability. On the basis of spatial attention, a convolution branch called feature augment block(FA-block) is proposed. This network structure learns the spatial distribution of the target from the mask of the sample and acquires a weight representing the importance degree for each pixel, then highlights the target part by weighting. This method aims to suppress background noise and augment the target features to be learned, make the features extracted from the backbone network more pure. The experiment on Pascal VOC dataset proves the effectiveness of FA-block. Through the validation of MS COCO dataset, FA-block improves the performance of a group of baselines of Faster Rcnn by 5.5%.

Key words: computer vision; convolution neural network; spatial attention; feature augment; high frequency noise suppression

0 引 言

有研究证明在尺度和通道维度方面增加注意力机制,能够有效地提高卷积网络的表达能力[1-2]。SE-net[3]对样本图像的所有通道做全局池化,将每一个通道的二维特征压缩为一个实数,该文认为这个实数具有单通道上全局的感受野,将这个通道向量经过两层全连接分析特征通道之间的相关性,得到代表每一个通道重要性程度的权重,这些权值可以让神经网络重点关注某些通道。Inception 系列网络[4-5]在网络模型中加入了多尺度,它让并联的卷积层拥有不同的权重,这相当于在尺度层面上增加了注意力机制。

该文在像素级的空间层面运用了注意力机制,提出了一种对样本特征图进行处理的卷积网络分支(feature augment block,以下简写为 FA-block),这种结构的输入来自样本图片的一张掩膜,通过学习掩膜前后景的数据分布模式,为特征图中的每一个像素训练得到一个得分值权重,以此将样本中的前后景区分开来,以突出特征图中的目标,让主干网络学习到更加

收稿日期:2021-06-25 **修回日期**:2021-10-26
基金项目:国家自然科学基金资助项目(61806150)
作者简介:许 畅(1997-),男,硕士,研究方向为人工智能与深度学习、嵌入式图像处理;王朝辉,博士,教授,研究方向为计算机先进控制技术、生物医学信息处理技术等。

健壮的特征表达。FA-block 作为主干网络的分支,同样可以被梯度下降算法优化,可以实现端到端的训练。将目标检测模型 Faster Rcnn[6] 作为基线,数据集选择微软的 MS COCO[7] 数据集,在没有添加额外的数据增强手段和训练技巧的情况下,增添了 FA-block 的网络模型相较于基线取得了 5.5% 的准确率提升。

1　卷积神经网络概述

1.1　卷积神经网络用于特征提取

卷积神经网络具有表征学习的能力,它是当前深度学习领域最常用的特征提取方法之一。为了减少计算量、提高计算机资源的利用率,大多数卷积网络习惯使用 3×3 的卷积核作为特征提取器,这增加了卷积网络的非线性表达能力,同时也一定程度上增加了网络的层数[8]。VGGNets[9] 和 Inception 家族证明了卷积网络越深,特征表达能力越好,但是样本数据经过太深层次的卷积可能导致数据失真。ResNets[10] 提出了层间残差跳连,将底层的特征映射到上层特征图上,这使得深层神经网络成为可能。FA-block 中多次用到 1×1 的卷积,1×1 卷积在 Network in Network[11] 中被提到,后来也被应用在 GoogLeNets 的 Inception 结构中。在 FA-block 中,掩膜的数据分布是十分简单而干净的,经过 1×1 卷积之后的特征图获得了非线性特征,经过一次单层卷积,一次三层卷积之后,三层通道的信息得到交互和整合,具有极高性价比。

1.2　数据增强方法

数据增强是神经网络训练之前的预处理手段[12],很多常见的数据增强手段,例如翻转、旋转、缩放、位移、裁减等通过对原图像作简单变换来生成新的样本数据,通常用来弥补样本数量的不足,还有一些特殊的数据增强方法用于完成特定的任务要求。文献[13]针对小目标的识别,对数据集里的每一个小目标过采样,然后通过预先的复制、粘贴手段在原始图像上大量生成小目标,以此提高小目标的检测精度。PGAN[14] 将生成对抗网络引入目标检测来生成具有超分辨率的图像,这使得模型能够学习到的特征数量大大提高。该文提出的 FA-block 也属于数据增强方法,但是它并不提高图像的数量和多样性,而更关注图像的质量,它在主干网络的基础上额外增添了一个网络分支用于强化样本中目标的特征,和传统的数据增强手段不同,它的优化过程是自主性的,网络分支上的权重可以被梯度下降算法优化并且可以实现端到端的训练。

1.3　注意力和门机制

生物的视觉系统倾向于倾注更多的注意力在重要目标上,通过借鉴这种生物学观点,注意力机制开始被用于计算机领域,在计算资源有限的情况下,注意力机制允许对更为重要的信息投入更多的算力,从而实现资源的高效利用[15]。相比于硬注意力,该文更加专注于软注意力,软注意力专注于空间区域或者是通道,它可以通过网络直接生成并且是可微的,这意味着它能通过神经网络计算出梯度,然后由反向传播来优化。Inception-net 想要消除尺度对于识别结果的影响,它一次使用多个不同大小的过滤器来捕获特征,它让并联的多个卷积层拥有不同的权重大小,从而实现尺度维度的注意力机制。STN[16] 认为之前的 pooling 方法太过暴力,直接将信息合并可能导致重要的信息被遗漏,于是提出了一种叫空间转换器的模块,这种模块是基于注意力机制实现的,训练出来的空间转换器能识别出图片中需要被关注的信息。通道维度的注意力机制也可以用于增强模型的鲁棒性,SE-net 为每一个通道的特征图作最大池化,它认为最大池化之后获得的数值点具有全局的感受野,它通过两层全连接层来分析不同通道之间的关联性,最后为特征图的每一个通道分配一个权重。文献[17]不仅仅是对空间域或者通道域添加注意力,它提出的注意力 mask 可以看作是给每一个特征元素赋予一个权重,这就同时形成了空间和通道域的注意力机制。FA-net 作为一个相对轻量级的空间门机制,它从掩膜的数据分布来训练得到样本空间域上的重要性程度,以此来削减背景噪声的印象,对目标倾注更多的关注。

2　特征增强模块

卷积神经网络学习的是数据的分布模式,一个直观的想法是如果一个滤波模板能够过滤掉样本中的背景杂质,进而突出样本中的目标,那么后续的神经网络层就能提取到更加健壮的目标特征,FA-block 扮演的就是这种滤波模板的角色。FA-block 是一种相对简单的结构,它由一个网络分支和添加到主干网络的几层 1×1 卷积组成,并不要求大量的参数或者权重,所以它能很方便地用于现有的许多网络结构中。

FA-block 的结构如图 1 所示,它分成两个部分,上分支的输入是样本掩膜的特征图,经过两层卷积核大小为 1×1 的卷积,首先将通道压缩,然后重新拉伸,这个过程是为了交换掩膜通道间的信息,同时在保证尺度不变的情况下为特征图提供较大的非线性特性。下分支的输入是样本特征图,对它进行同样的 1×1 卷积,为其添加非线性特性。上分支的输出经过了两次 ReLU 激活函数,可以把它看作是对下分支特征图的每一个像素点给出的得分值,把这个得分值作为权重与下分支的输出作点乘运算,得到的 FA-block 的输出,认为这种结构区别了前景和后景,突出了代表目标的像素,FA-block 的输出可以作为输入被喂入后续的

神经网络层。

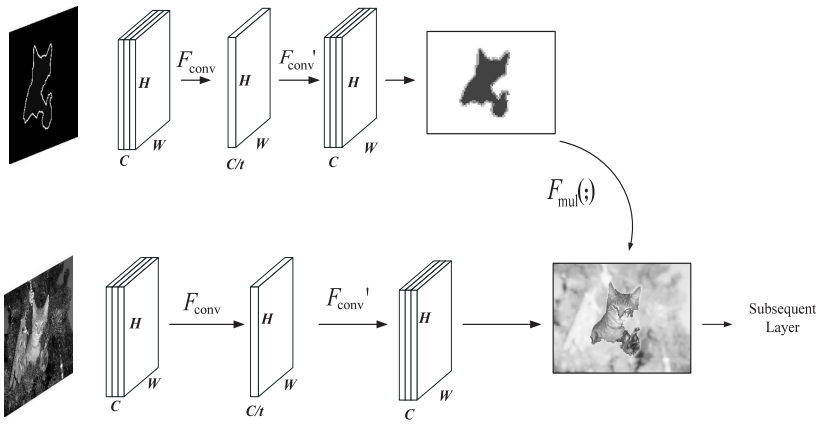

图 1 特征增强模块

2.1 从样本掩膜中提取滤波模板

对于从实例掩膜输入的一张 $X \in R^{H\times W\times C}$ 特征图，通过 F_{conv} 变换将其转换为特征图 $U^{H\times W\times C'}$，其中 $C' = C / t$（t 是一个超参数，一般默认它为 3，超参数的选择在第 3 节有实验论证），F_{conv} 指代一次卷积操作，$V = [v_1, v_2, \cdots, v_{c'}]$ 代表的是卷积过程中会用到的 c' 个过滤器，计算过程表示如下：

$$U_{c'} = V_{c'} * X = \sum_{s=1}^{C} v_{c'}^{s} * x^{s} \tag{1}$$

其中，$*$ 代表卷积运算，$V_{c'} = [v_{c'}^{1}, v_{c'}^{2}, \cdots, v_{c'}^{C}]$，$X = [x^1, x^2, \cdots, x^C]$，$X \in R^{H\times W\times C}$，$U \in R^{H\times W}$，卷积核尺寸设定为 1×1，不设置 padding 以保证输出具有相同的尺寸。经过了通道降维之后再通过 $F_{conv'}$ 将通道数从 c' 转化回原来的通道数 c，计算过程类似：

$$S_c = W_c * X = \sum_{s=1}^{C'} w_c^{s} * x^{s} \tag{2}$$

1×1 的卷积并不能提取特征图空间上隐含的信息，它的作用是整合通道上的信息并为掩膜的滤波模板提供更多的非线性表达，使其作为主干网络的权重能够更好的拟合。两层 1×1 卷积首先将通道数降维，然后再将通道升维成原本的维数，这样能够在不影响主干网络的情况下添加 FA-block，因此滤波模板分支的输出可以被写作如下公式（σ 代表 ReLU 激活函数）：

$$S_c = F_{trans}(X) = \sigma(f_{conv'}(U_c)) = \sigma(f_{conv'}(\sigma f_{conv}(X))) \tag{3}$$

2.2 为主干特征图赋予权重

为了保证数据分布的一致性，对主干网络的特征图同样进行两次 1×1 卷积来整合通道间的特征，然后将滤波模板的每一个像素点作为主干特征图的权重映射到特征图上，得到 FA-block 的输出：

$$\tilde{x}_c = F_{mul}(S_c T_c) \tag{4}$$

其中，$\tilde{x}_c = [\tilde{x}_1, \tilde{x}_2, \cdots, \tilde{x}_c]$ 表示输出是一个 c 维通道的特征图，$T_c = [t_1, t_2, \cdots, t_c]$ 表示经过 1×1 卷积的样本特征图，$F_{mul}(S_c T_c)$ 表示对于特征图 T_c 的每一个像素，滤波模型 S_c 都有一个相应的权重值与它相乘。样本特征经过 FA-block 的处理，认为它的重要特征得到了突出，高频的背景噪声得到了抑制，它没有对特征图的尺度和通道数做出改变，所以能够直接插入到现有的各种模型如 ResNets 中，后续的网络将能提取到更加具有表征力的特征。

3 实验结果与数据分析

本节设计了一系列的实验来证明 FA-block 对卷积网络表征能力的增强作用，本节在不同深度的主干网络上添加了 FA-block 模块，测试 FA-block 嵌入在网络不同的层数时的图像分类效果，将效果最好的嵌入方式用于目标检测数据集 MS COCO 做对比实验，确保 FA-block 在不同计算机视觉任务和数据集上的泛化能力，最后讨论了超参数 t 的选择。

3.1 FA-block 嵌入不同层对比

FA-block 不会对主干网络特征图的尺寸、通道数产生影响，所以它可以被添加到网络的任意位置来强化网络的特征提取。分别用 ResNet-50、ResNet-101、ResNet-152 作为横向对比的基线，将 FA-net 作为网络分支加入到基线的不同层数：底层（Conv1 层）、中间层（Conv3_x 层）、顶层（Conv5_x 层）前面来测试模型准确率的提升。

FA-block 中需要用到掩膜，所以实验选择在图像分类数据集 PASCAL VOC2012 Augmented Dataset[18] 上进行。该数据集由两个数据集合二为一制作而成：PASCAL VOC2012 和 Semantic Boundaries Dataset，数据集包含有 10 582 张图像用于训练，1 449 张图像用于验证，1 456 张图像用于测试，提供 20 种不同种类的物体用于分类。

实验中，分别将 FA-block 添加到三种不同基线的底部、中层和顶层。对于将 FA-block 嵌入 Conv1 层

前,无需再对掩膜进行额外的操作,将掩膜的 png 格式数据转化为 RGB 三通道之后作为输入送入 FA-block 的滤波模型分支;Conv3_1 要求输入是 56×56×256 的尺寸,掩膜输入到 FA-block 前需要通过卷积对其尺寸、通道数做调整,出于计算量的考虑,只对掩膜做单次 Conv1 和 Conv2_1 卷积;Conv5_1 要求的输入是 14×14×1 024 的尺寸,于是对掩膜进行单次 Conv1、Conv2_1、Conv3_1、Conv4_1 卷积,然后送入到 FA-block 中。

三种基线和与之对应的 FA-block 改进版本均采用相同的优化策略,遵循了实践标准对样本进行 224×224 的随机裁减,随后进行随机水平翻转来增强数据,选择了 momentum 为 0.9 的同步 SGD 优化方法,每一个 minibatch 设置为 128,学习率设置为 0.01,基线和其对照网络的错误率变化趋势如图 2 所示。

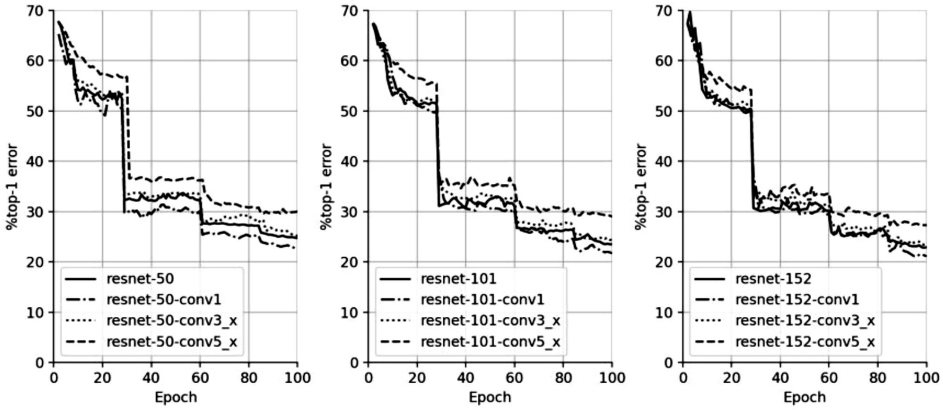

图 2　基线的不同层加入 FA-block 后的 top-1 error 变化趋势

表 1 显示的是三种不同嵌入方式下的 top-1 error 对比,对于三种不同深度的网络,在底层嵌入 FA-block 都取得了良好的效果,实际上,在 Conv1 层之前嵌入 FA-block 使得 ResNet-50 的 top-1 error 从 24.7% 减少到 22.6%,相比于 ResNet-101 的 23.5% 的错误率更小,但是 ResNet-101 需要的计算量几乎是 FA-ResNet-50 的两倍。

表 1　不同层加入 FA-block 后的 top-1 error 对比　　%

	Origin	Embeding on conv1	Embeding on conv3_x	Embeding on conv5_x
ResNet-50	24.7	22.6	25.2	30.0
ResNet-101	23.5	21.7	24.3	29.0
ResNet-152	22.9	21.1	23.5	27.3

将 FA-block 嵌入到 ResNet101 和 ResNet152 的底层也降低了错误率,虽然不如 FA-ResNet-50 明显,但是它们也用极小的计算代价换取了相当可观的精确度提升,这证明 FA-block 这种数据增强结构抑制了图像中的噪声部分并增强了目标特征,帮助网络学习到了更关键性的特征。

注意到将 FA-block 插入到 Conv3_x 前和 Conv5_x 前的实验并没有取得理想的结果,嵌入到 Conv3_x 前相较原始版本准确率有些微的下降,嵌入到 Conv5_x 下降幅度则更大,猜测这是因为掩膜的数据分布经过多层卷积的尺寸、维度调整,它的数据分布模式和原始数据的特征图出现了较大的偏差,因此不能有效地对样本特征图的特征增强做出指导,想要缓解这种偏差可以让掩膜和样本做相同的卷积变换,但是这样大量增加了网络的计算量,并不是一种性价比高的、值得用于实践的方法。

3.2　将 FA-block 用于 MS COCO 目标检测

为了更进一步地验证 FA-block 的有效性,将这种结构用于目标检测的权威数据集 MS COCO 来确认精准度的提升。MS COCO 数据集是微软于 2014 年出资开发和维护的大型图像数据集,和 PASCAL VOC 数据集相比,COCO 数据集中的图片包含了自然图片以及生活中常见的目标图片,背景更加复杂,单张图片中目标的数量更多,目标的尺寸更小,因此 MS COCO 数据集上的目标检测任务更难。对于检测任务来说,衡量一个模型的好坏更加倾向于使用 MS COCO 数据集上的检测结果,使用的 MS COCO 数据集包含有 80k 的训练数据,35k 的验证数据和 5k 的测试数据。

实验中用到了文献[6]中经典的目标检测模型 Faster rcnn,其中使用不同层数的 ResNet 作为模型的 backbone,实验中要做的就是用对应的 FA-ResNet 替换原始的 ResNet,这样就能通过模型准确率的变化知道 FA-block 在检测模型中发挥的作用。在训练过程中使用文献[6]中的设置学习超参数的方法,训练周期相较文献[6]来说要短,使用基础学习率 0.01 在四块 GPU 上迭代了 2.1k 次,使用 SGD 作为优化方法,momentum 设置为 0.9,权重衰减为 0.000 1,训练时分别在 0.5k、1.5k 的时候将学习率下降 0.1 倍。权重由 ImageNet 的预训练模型初始化,其他参数都参照 Faster Rcnn 的 baseline 进行设置。

实验结果如表 2 所示。

表 2 不同层数的 FA-ResNet 和相应基线的目标检测指标对比 %

	AP@ IoU=0.5	AP
ResNet-50	57.8	37.8
FA-ResNet-50	60.8	39.9
ResNet-101	59.9	39.6
FA-ResNet-101	62.4	41.3
ResNet-152	62.1	41.2
FA-ResNet-152	63.5	42.7

FA-ResNet-50 相较于其基线 ResNet-50 在 MS COCO 上提升了 2.1% 的标准 AP 指标(达到 5.5% 的相对精进),并且在 AP@ IoU=0.5 上提升了 3.0%,已经超过了 ResNet-101 在 AP 上的 59.9% 和 AP@ IoU=0.5 上的 39.6%。FA-ResNet-101 和 FA-ResNet-152 在 AP 上也分别取得了 1.7% 和 1.5% 的提升。总而言之,这一系列的实验证明了 FA-block 在目标检测领域也能起到正向的收益,对于提高模型的表达能力而言确实是一种行之有效并且极具性价比的方法。

3.3 超参数 t 的选择对 FA-block 效果的影响

第 2 节中提到的通道削减倍率 t 是一个超参数,t 设置不同的数值,FA-block 的表征能力和计算量都会不同,从众多备选方案中找到一个合适的值能让模型在准确率和计算量的增加之间达到一个平衡。采取不同的 t 值在 ResNet-50 上设计了实验,数据集采用的是 PASCAL VOC2012 Augmented Dataset,超参数的设置和第 3.1 节中相同,实验结果如表 3 所示。

表 3 超参数 t 的不同取值对模型准确率和参数量的影响

Ratio t	Top-1 error/%	Top-5 error/%	Params/M
1	23.51	6.19	50.7
2	23.34	6.12	40.6
3	22.62	6.03	36.1
4	22.76	6.13	32.1
5	23.13	6.30	30.9

随着通道削减倍率 t 的增加,FA-block 附加的计算量也在减小,模型的准确率增加,t 的取值达到 3 以后在准确率和计算量之间达到了相对的平衡。然而在实际的使用中,超参数 t 的选择必须根据自身模型的实际情况做出取舍,该文给出的是 FA-block 运用于 ResNet 时候的参考。

4 结束语

传统的卷积神经网络对样本图像的特征进行笼统的提取,没有明确区分图像的前后景,这就导致高频的背景噪音对模型的拟合产生影响,降低了神经网络的准确率。因此,提出了 FA-block,一种基于空间注意力的神经网络分支用作样本特征图的滤波模板,旨在突出特征图中的目标特征,适当抑制特征图中的背景噪声,使得卷积网络能够提取到更能代表目标特征分布模式的、更具表现力的模型。通过不同数据集上的实验证明了 FA-block 结构确实能够在图像分类和目标检测任务中以极低的计算代价提高模型的表达能力,希望这种结构也能在更复杂、对模型表现力要求更高的任务中发挥作用。

参考文献:

[1] 马世拓,班一杰,戴陈至力. 卷积神经网络综述[J]. 现代信息科技,2021,5(2):11-15.

[2] 卢泓宇,张 敏,刘奕群,等. 卷积神经网络特征重要性分析及增强特征选择模型[J]. 软件学报,2017,28(11):2879-2890.

[3] HU Jie,SHEN Li,ALBANIE S,et al. Squeeze-and-excitation networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2020,42(8):2011-2023.

[4] SZEGEDY C,LIU Wei,JIA Yangqing,et al. Going deeper with convolutions[J]. arXiv:1409.4842,2014.

[5] SZEGEDY C,IOFFE S,VANHOUCKE V,et al. Inception-v4,inception-ResNet and the impact of residual connections on learning[J]. arXiv:1602.07261,2016.

[6] REN Shaoqing,HE Kaiming,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.

[7] LIN Tsung-yi,MAIRE M,BELONGIE S,et al. Microsoft COCO:common objects in context[J]. arXiv:1405.0312,2014.

[8] 王浩滢. 深度学习及其发展趋势研究综述[J]. 电子制作,2021(10):92-95.

[9] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv:1409.1556,2014.

[10] HE K,ZHANG X,REN S,et al. Deep residual learning for image recognition[J]//IEEE conference on computer vision and pattern recognition (CVPR). Las Vegas, NV, USA: IEEE,2016:770-778.

[11] LIN M,CHEN Q,YAN S. Network in network[C]//International conference on learning respresentations. Banff,Canada:OpenReview.net,2014.

[12] 朱晓慧,钱丽萍,傅 伟. 图像数据增强技术研究综述[J]. 软件导刊,2021,20(5):230-236.

[13] KISANTAL M,WOJNA Z,MURAWSKI J,et al. Augmentation for small object detection[J]. arXiv:1902.07296,2019.

[14] LI Jianan, LIANG Xiaodan,WEI Yunchao,et al. Perceptual

(下转第 111 页)