ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 38,No. 2,Mar. 2023,pp. 364－374
DOI:10. 16337/j. 1004-9037. 2023. 02. 012

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于非局部融合的多尺度目标检测研究

马　倩[1,2]，曾　凯[1,2]，吴家文[1,2]，沈　韬[1,2]

(1. 昆明理工大学云南省计算机重点实验室,昆明 650500;2. 昆明理工大学信息工程与自动化学院,昆明 650500)

摘　要: 针对现有的多尺度目标检测模型在面对尺度变换和遮挡场景时所使用的融合方法融合不充分,且没有捕捉长距离依赖关系的问题,本文设计了通道融合增强模块和非局部特征交互模块,用于学习不同通道特征之间的相关性和捕捉特征图之间的长距离依赖关系。此外,针对当前检测架构都是基于单金字塔检测结构,存在信息丢失的情况,设计了双金字塔结构,并将提出的融合方法与双金字塔结构结合,在保留原始特征信息的基础上,补充融合后的特征信息。实验结果表明,提出的方法在公共数据集KITTI与PASCAL VOC上与其他先进工作相比具有更高的检测精度,证明了该方法在目标检测任务中的有效性。

关键词: 机器视觉;多尺度目标检测;尺度变换;特征融合;双金字塔

中图分类号: TP391.4　　**文献标志码:** A

Multi-scale Object Detection Based on Non-local Feature Fusion

MA Qian[1,2], ZENG Kai[1,2], WU Jiawen[1,2], SHEN Tao[1,2]

(1. Computer Technology Application Key Lab of Yunnan Province, Kunming University of Science and Technology, Kunming 650500, China;2. Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China)

Abstract: Aiming at the problem that the fusion method used by the existing multi-scale object detection model in the face of scale variation and occlusion scene is not sufficient, and does not capture the long-distance dependency relationship, channel feature fusion aggregation module and non-local feature interaction module are designed to learn the correlation between different channel features and capture the long-distance dependence between feature maps. In addition, the current detection architecture is based on single pyramid detection structure, which exists information loss. In this paper, a double pyramid structure is designed, and the proposed fusion method is combined with the double feature pyramid structure to supplement the fusion feature information on the basis of preserving the original feature information. Experimental results on public datasets KITTI and PASCAL VOC show that the proposed method has higher detection accuracy than other advanced work, proving its effectiveness in object detection task.

Key words: machine vision; multi-scale object detection; scale variation; feature fusion; double feature pyramid

---

基金项目: 国家自然科学基金(61971208, 61702128);云南省重大科技专项基金(202002AB080001-8);国网信通产业集团两级协同研发项目(SGIT0000XTJS1900433)。
收稿日期: 2021-11-17;**修订日期:** 2021-12-28

引　言

得益于深度学习的不断发展,目标检测这一计算机视觉分支任务也取得了不断的进步。近年来,目标检测在自动驾驶、无人机巡检等新兴领域应用越来广泛,但同时也遇到很多挑战。这类应用包含的视觉场景往往极其复杂,易出现目标尺度小、易受遮挡和姿态变化快等情况[1],因此需要目标检测算法具有较强的检测能力,能够正确识别不同尺度物体和被遮挡目标[2]。

当前基于深度神经网络的目标检测算法主流分为一阶段检测(One-stage)和两阶段检测(Two-stage)两大类[3]。一阶段检测算法将整个检测流程统一到一个过程中,即对输入图像直接进行分类和回归,代表性的算法有YOLO系列[4]和单次多边框检测器(Single shot multibox detector,SSD)[5]系列,这类算法检测速度快、实时性出色,但是检测精度有所不足。两阶段检测类算法将检测流程分为两个过程,首先提取出感兴趣的区域处理得到候选框,然后再对候选框的物体进行识别,代表性算法有R-CNN[6]、Fast R-CNN[7]和Faster R-CNN[8]等。这类算法的优缺点跟一阶段检测的正好相反,表现为检测精度高但检测速度慢。因此,在辅助驾驶场景这样对实时性要求较高的检测场景下,所使用的目标检测算法几乎都是基于一阶段算法。

一阶段检测算法SSD[5]通过对不同层次的特征图进行检测,一定程度上对各个尺度的目标都有所兼顾,但是所提取出来的特征图仅是简单的经过多层卷积后得到的,深层特征图所包含的小目标信息缺失较多,不利于小目标物体的检测。反卷积SSD(Deconvolutional SSD,DSSD)使用残差网络作为特征提取网络替代VGG16,并引入非对称的反卷积模块与原提取的特征图融合以增加上下文信息,在小目标检测性能上提升明显[9]。特征融合SSD(Feature fusion SSD,FSSD)将不同层次的特征图通过双线性插值规整为同一尺寸进行融合,利用融合后的特征图再生成新的特征金字塔进行检测,在保证增加少量运算开销的同时很好地提升了性能[10]。感受野模块(Receptive field block,RFB)从人类视觉的感受野结构中获得启发,提出了一种感受野模块,增加了低层次特征图中每个单元的感受范围,使得特征图中包含更多的上下文信息,对最后的检测提供了更加具有高判别性的特征,获得了接近两阶段检测算法的精度[11]。尽管上述方法在目标检测性能上有了一定的提升,但还存在以下不足:

(1)常见的特征图融合方式主要有两种:通道级联和同位置元素相加[12]。通道级联法增加了原有的通道数,但是每一特征下的信息没有增加;同位置元素相加法将对应的特征图相加,虽增加了每一维下的信息量,但是没有很好地解决长距离依赖问题[13]。

(2)上述的一阶段检测器的基础架构都是基于单金字塔结构,所用于预测的特征层包含信息过于单一。这些检测器有的只使用原始特征层进行预测,使模型缺乏融合后的上下文信息;有的只对特征层融合后的信息进行预测,忽视了原始信息。

针对该领域提出的两个问题,本文的主要贡献为:

(1)提出了通道特征增强融合模块和非局部特征交互模块两个特征融合模块。首先采用通道特征融合增强(Channel feature fusion aggregation,CFFA)模块学习每一维的通道信息并重新分配,接下来采用非局部特征交互(Non-local feature interaction, NLFI)模块实现对不同尺度特征图之间空间信息的充分融合,捕捉不同层不同像素位置之间的长距离依赖关系。

(2)提出了双金字塔检测结构,并将该结构与RFB模型结合,设计了NL-RFB目标检测模型。该模型不仅有效地保存了原有的每一层特征图的基础信息,还补充了不同尺度特征图融合后的全局上下文信息,有效提升了模型的检测性能。

1　基于非局部融合的多尺度目标检测模型

在特征金字塔结构中,随着层数的深度增加,语义信息也会越来越丰富,但是也会丢失细节信息。RFB网络模型以特征金字塔结构为基础,采用VGG16作为特征提取骨干网络,并添加额外的卷积层在

该网络中。该网络主要借鉴了Inception思想[14]，并加入膨胀卷积，从而有效增大了特征层的感受野。但是RFB算法不能有效地将包含几何细节丰富的低层特征图与语义信息丰富的高层特征图进行融合，因此其检测精度仍需进一步提升。尤其对于小目标物体，一般在图像中会占据较小的像素，要想正确地识别它需要充分结合其外观细节信息和语义信息[15]。所以实现不同层特征图的充分融合是十分必要的。当前特征图的融合方式主要有两种，通道级联和同位置元素相加。通道级联是指增加特征图本身的通道数，具体结构如图1(a)所示；同位置元素相加则是通过增加特征图下每一维的信息量来进行融合，具体结构如图1(b)所示。

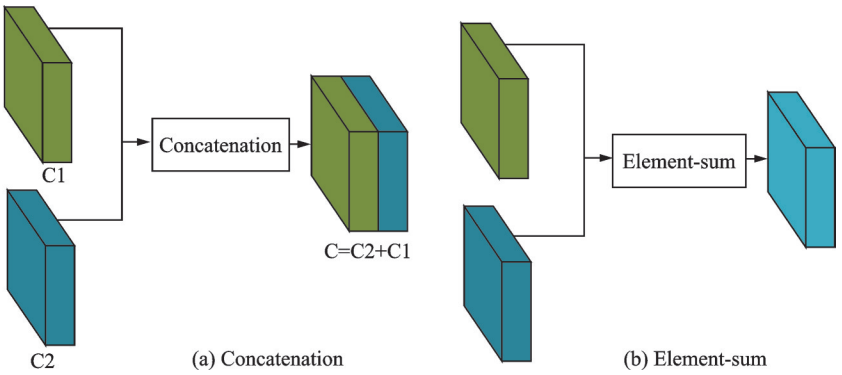


图1 通道级联和同位置元素相加过程

Fig.1 Process of concatenation and element-sum

这两种融合方法只是简单地实现同位置语义信息的交互，仅仅增加了空间细节特征，并没有捕捉不同位置之间的长距离依赖关系。因此，需要设计一种能够充分融合通道信息与空间信息的方法。

本文改进了传统的RFB算法，并命名为NL-RFB。NL-RFB的整体结构如图2所示，主要包括CFFA模块以及NLFI模块，并在此基础上设计双金字塔检测结构，如图3所示。这种结构使模型在保留原始特征信息的基础上，进一步补充融合后的特征信息。文献[16]证明，一般浅层特征图具有丰富的几何信息，而深层特征图则包含了大量的语义信息，这两种特征图之间存在着较大的差异性，并还认为分辨率小于10像素×10像素的特征图中缺乏足够的信息去学习。因此本文遵循文献[16]的建议，选择Conv 4_3、Conv 5_3以及FC7这3个特征层作为CFFA模块的输入，利用CFFA模块建模各个通道之间的重要程度获得更多通道间的信息。NLFI模块的输入则是CFFA的输出，该模块用来建模像素点之

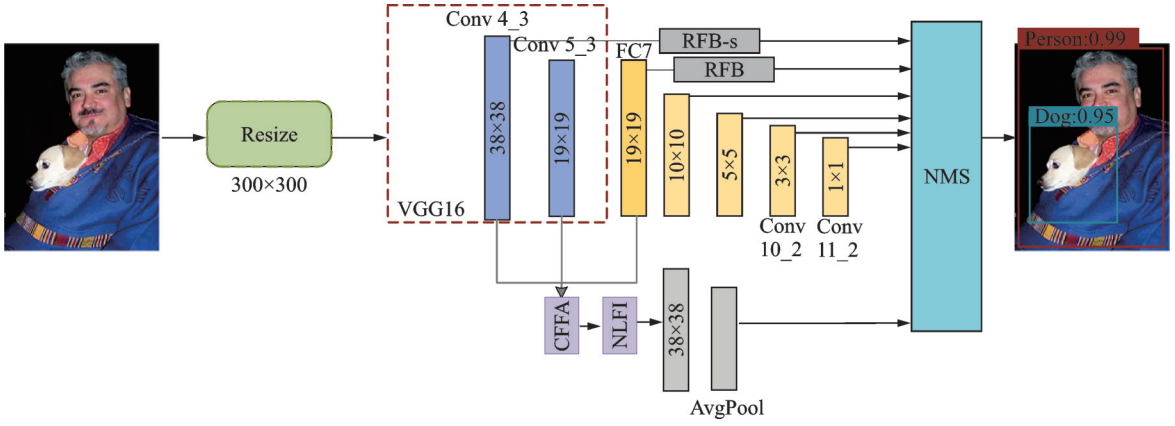


图2 NL-RFB算法整体架构图

Fig.2 Overall framework of NL-RFB

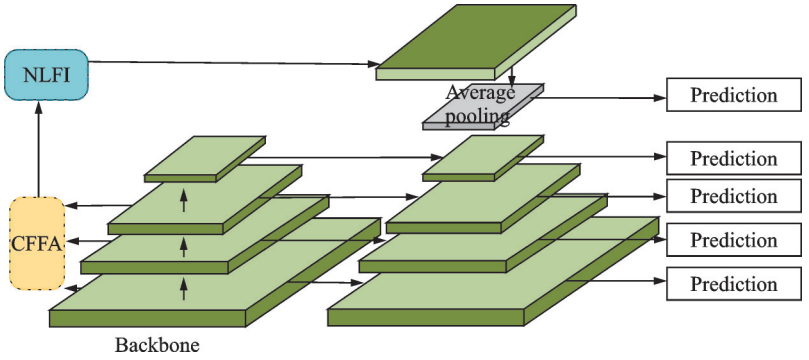


图3　双层金字塔结构

Fig.3　Double feature pyramid structure

间的长距离依赖关系,从而获取更多的全局上下文语义信息,实现空间信息的充分融合。两种模块的设计有效提升了模型对不同尺度目标的检测能力,尤其是小目标物体。最后,模型在NLFI模块的输出层后增加全局平均池化层,旨在降低模型出现过拟合的可能性。

1.1　CFFA模块

CFFA模块旨在使网络可学习通道间特征的相关性与重要性,获得更多通道的有用信息,其设计如图4所示。首先将高层特征图$\boldsymbol{x}_{\mathrm{h}}$以及中层特征图$\boldsymbol{x}_{\mathrm{m}}$使用双线性插值进行上采样到与低层特征图$\boldsymbol{x}_{\mathrm{l}}$同样尺度,然后进行通道级联。通道级联的过程可表示为

$$\boldsymbol{X}_i=\mathrm{Concat}(\boldsymbol{x}_{\mathrm{l}},\boldsymbol{x}_{\mathrm{m}},\boldsymbol{x}_{\mathrm{h}}) \tag{1}$$

式中$\boldsymbol{X}_i$代表经过通道级联后的特征图。

接下来引入通道注意力机制,使网络进一步学习通道间的相关性与重要性。首先将经过通道级联后的特征图$\boldsymbol{X}_i$作为通道注意力机制的输入,然后分别采用平均池化层和最大池化层对特征图进行压缩。最大池化可以保留更多的纹理信息,平均池化可以保留更多的图像背景信息。文献[17]证明融合最大池化和平均池化产生的特征图对建立特征图通道之间的关系有重要意义;然后利用两个全连接层和Relu函数以及sigmoid函数实现注意力机制,可表示为

$$s=F_{\mathrm{ex}}(z,\boldsymbol{W})=\varphi(\boldsymbol{W}_2\sigma(\boldsymbol{W}_1 z)) \tag{2}$$

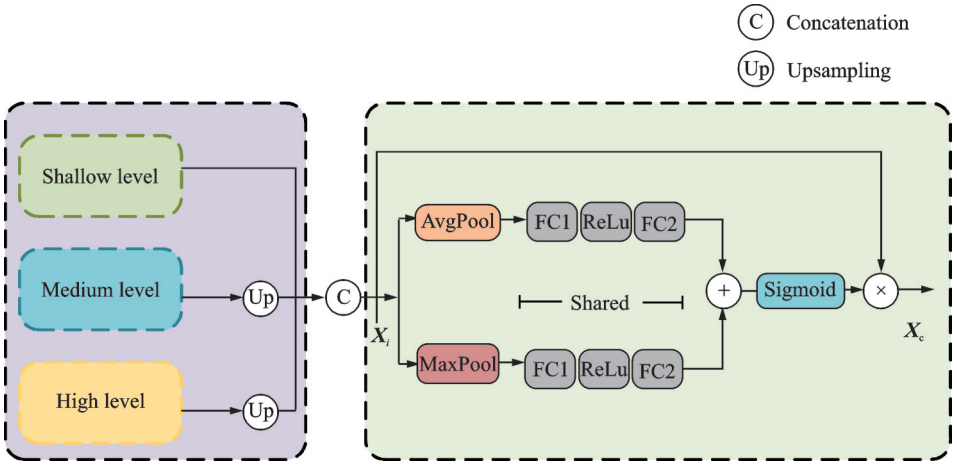


图4　通道融合增强模块

Fig.4　Channel feature fusion aggregation module

式中：φ代表ReLu激活函数，σ代表sigmoid函数，$\boldsymbol{W}_1\in\mathbf{R}^{\frac{C}{r}\times C}$，$\boldsymbol{W}_2\in\mathbf{R}^{C\times\frac{C}{r}}$，$r$代表还原比（本文设置为$r=16$）。

CFFA模块的最后一步则是对各通道的特征进行权值的重新分配，其输出定义为X_c。X_c包含不同层的特征图经过通道融合增强模块后的最有用的通道信息。

1.2 NLFI模块

为了使融合后的特征图能够包含更多的空间信息，捕捉特征图之间的长距离依赖关系以及更多的全局上下文语义信息。本文受Transformer[18-19]中的多头注意力机制的启发，设计了轻量化的NLFI模块，用来捕捉不同层像素之间的长距离依赖关系，得到融合信息更加充分的特征图，具体设计如图5所示。

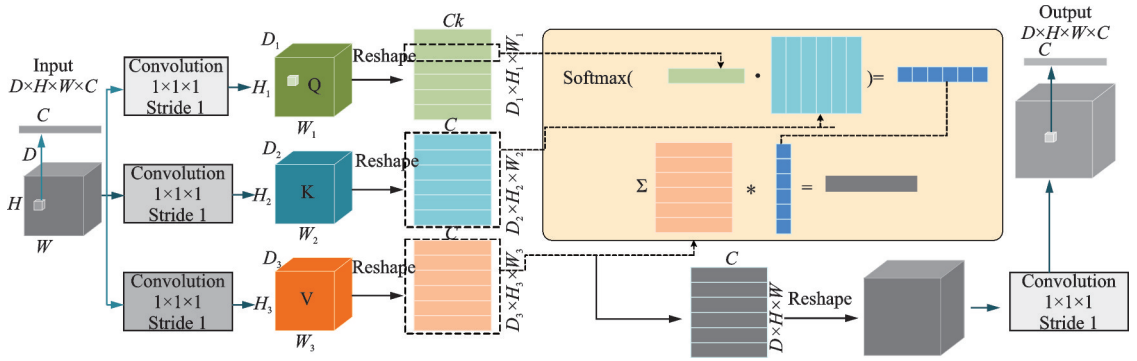


图5 非局部特征交互模块

Fig.5 Non-local feature interaction module

本模块的输入来自上一节设计的通道增强模块的输出$\boldsymbol{X}_c$。整个模块设计的第一步主要生成查询（$\boldsymbol{Q}$）、键（$\boldsymbol{K}$）以及值（V）的3个可学习权重矩阵，分别可表示为

$$\boldsymbol{Q}=\mathrm{Reshape}(\mathrm{QueryTransform}c_k(\boldsymbol{X})) \tag{3}$$

$$\boldsymbol{K}=\mathrm{Re\,shape}(\mathrm{Conv1}\times 1_{C_k}(\boldsymbol{X})) \tag{4}$$

$$\boldsymbol{V}=\mathrm{Re\,shape}(\mathrm{Conv1}\times 1_{C_v}(\boldsymbol{X})) \tag{5}$$

式中：$\boldsymbol{X}$代表输入的特征图，尺寸大小为$B\times H\times W\times C$，其中$B$代表batch size，$H$代表特征图的高，$W$代表特征图的宽，$C$代表通道数；Reshape(•)代表将一个$B\times H\times W\times C$的张量变成一个$(B\times H\times W)\times C$的矩阵；QueryTransform$c_k$，Conv1$\times 1_{C_k}$和Conv1$\times 1_{C_v}$代表1×1的卷积。

第二步则是通过缩放点积，得到新的权重矩阵$\boldsymbol{A}$，用于自注意力机制的信息加权，如式(6)所示，然后将权重矩阵$\boldsymbol{A}$与$\boldsymbol{V}$进行点乘，得到包含长距离依赖关系的全局特征结果$\boldsymbol{Y}$，如式(7)所示。

$$\boldsymbol{A}=\mathrm{Soft\,max}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{C_k}}\right) \tag{6}$$

$$\boldsymbol{Y}=\boldsymbol{A}\boldsymbol{V} \tag{7}$$

1.3 损失函数

损失函数是用来评估模型的计算值与真实值差异性程度的评价指标。本文采用多任务损失函数，可同时实现置信度判别与位置回归。训练的总损失函数L由预测框和目标框的回归误差和分类误差的加权组成，即

$$L(x,c,l,g)=\frac{1}{N}\left[L_{\mathrm{conf}}(x,c)+\lambda L_{\mathrm{loc}}(x,l,g)\right] \tag{8}$$

式中：x为预测框中目标的真实类别；c为预测框所判定的类别信息；l为预测框的坐标信息；g为真实框

的坐标信息;λ为权重参数;N为所有真实框的个数;L_{conf}和L_{loc}分别为分类损失和回归损失函数,可分别表示为

$$L_{\text{conf}}(x,c)=-\left(\sum_{i\in\text{pos}}^{N}x_{ij}^{P}\lg(\hat{c}_i^{p})+\sum_{i\in Neg}^{N}\lg(\hat{c}_i^{N})\right) \tag{9}$$

$$L_{\text{loc}}(x,l,g)=\sum_{i\in\text{Pos}}^{N}\sum_{m\in(cx,cy,w,h)}x_{ij}^{P}\text{smooth}_{L_1}(l_i^m-\hat{g}_j^m) \tag{10}$$

2　实验结果与分析

为了验证NL-RFB目标检测算法的性能,首先在KITTI数据集上设计了大量的实验进行验证[20],然后在PASCAL VOC[21]上验证模型的泛化能力。

实验环境:操作系统为Ubuntu20.04,深度学习框架为pytorch1.5.1, CPU为 Intel Xeon Gold512, 运行内存为512 G,GPU为Nvidia Tesla V100 GPU, CUDA10.2。

实验数据集:采用两个大型公共通用数据集KITTI与PASCAL VOC。对于KITTI数据集,采用KITTI数据集中的2012 2D object detection left color images of object data作为训练集和测试集,并选择KITTI数据集中的7类进行实验,分别为Car、Pedestrian、Cyclist、Person_sitting、Van、Tram和Truck;对于PASCAL VOC数据集,训练集设置为VOC2007与VOC2012联合训练,测试集为VOC2007。该数据集中包含了日常生活常见的物体种类,共20个, 其中Chair、Bottle和Plant是典型的小目标物体。训练集和验证集有16 551张图片,测试集有4 952张图片。

训练细节:训练过程中图片输入尺寸的大小固定为300×300。采用SGD算法进行优化,学习率为1e−5, 动量为0.9,批次大小设置为8。训练的最大Epoch数设置为300。

评价指标:目标检测任务最常用的评价指标是多类别平均精度(mean average precision, mAP)。对于每一类的平均精准度可以用AP(Average precision)来表示,它是通过召回率和准确率两个角度来衡量的。模型复杂度则通过模型权重文件的大小和参数量来体现,模型计算复杂度可以用每秒浮点运算次数(Floating-point operation per second, FLOPs)和检测帧率(Frames per second, FPS)进行评估。

2.1　在KITTI数据集上的对比实验

为验证所提出的NL-RFB目标检测算法在辅助驾驶任务中的有效性,首先在KITTI数据集上将提出的NL-RFB模型与其他先进的目标检测算法进行对比。为了保证实验的公平性,在KITTI数据集上,对所有的对比实验都重新训练300个Epoch。NL-RFB目标检测模型与其他经典的先进检测器在KITTI数据集上的对比结果如表1所示。实验结果证明,NL-RFB具有更高的识别精度,与RFB模型相比提高了0.8个百分点。对于模型复杂度,NL-RFB相较于RFB模型略微增加了参数量以及模型大小,并且模型的计算复杂度也稍有提高, 但是目标检测模型的帧率可达到54,能够满足实时目标检测的需求,因此模型复杂度稍

表1　在KITTI数据集上不同目标检测算法的结果对比

Table 1　Comparison of detection performance of different target detection methods on KITTI dataset

方法	骨干网络	输入图像尺寸/(像素×像素)	模型尺寸/MB	参数量/MB	mAP/%	FPS	FLOPs
SSD	VGG-16	300×300	187.3	206.52	72.23	92	30.79
FSSD	VGG-16	300×300	125	260.32	68.18	52	34.87
DSSD	ResNet-101	320×320	932	399.81	75.26	32	35.22
RFB	VGG-16	300×300	132.4	257.38	76.72	58	35.31
NL-RFB	VGG-16	300×300	150	272.45	**77.52**	54	41.08

微增加也是可以接受的。表2展示了NL-RFB模型与其他方法在KITTI数据集上每一类的检测精度，可以看出本文提出的模型对于其中的大多数类别都有着不错的检测效果，比如Van、Truck和Cyclist。

表2 在KITTI数据集上不同目标检测算法每一类的结果对比

Table 2 Per-class comparison of detection performance of different target detection methods on KITTI dataset %

方法	Car	Van	Truck	Cyclist	Pedestrian	Person_sitting	Tram
SSD	86.59	86.73	89.31	59.70	45.52	46.98	90.77
FSSD	80.79	78.46	89.13	51.64	37.19	52.89	87.17
DSSD	**88.22**	85.38	93.23	69.09	44.01	**57.48**	89.41
RFB	87.48	88.23	89.75	64.62	**56.21**	56.20	**94.54**
NL-RFB	87.68	**88.33**	**97.64**	**70.40**	54.28	57.13	87.19

为了更直观地表现出本文所提模型的优势，图6展示了可视化RFB与NL-RFB的检测结果，其中图6(a)为基模型RFB的检测效果，图6(b)为NL-RFB模型的检测效果。从图6可以看出，本文提出的NL-RFB模型对于辅助驾驶场景下出现的遮挡与尺度变换等复杂情况，都有着不错的检测效果。这是因为NL-RFB模型不仅通过CFFA以及NLFI模块充分融合了不同尺度之间的特征信息，还通过双金字塔结构保留了原始特征图的信息。

(a) RFB (b) NL-RFB

图6 KITTI测试集上的检测结果

Fig.6 Detection results comparison on KITTI test dataset

2.2 在PASCAL VOC数据集上的对比实验

为进一步验证提出模型的通用性与有效性，在PASCAL VOC上进行了大量对比实验。表3展示了NL-RFB模型与其他6种先进的目标检测算法的检测精度以及模型大小，参数量和计算量以及检测帧率。从表3可以看出，本文提出的NL-RFB模型，在网络的输入尺度为300×300时，达到了80.74%的检测精度，相较于经典的SSD算法提高了4.68%；与基于SSD算法改进的模型DSSD、FSSD和RFB相比，分别提高了3.7%，3.81%和1.46%；相对其他性能出色的一阶段流行的检测算法MDFNI1和MDFNI2，也分别提高了1.44%和2.44%。虽然在模型复杂度上与RFB相比模型的大小增加了12 MB，模型的参数量增加了14.03 MB，计算复杂度上增加了4.98 G，但在检测速度上，NL-RFB的帧率几乎没有下降。

表3　在PASCAL VOC 2007数据集上不同目标检测算法的结果对比

Table 3　Comparison of detection performance of different target detection methods on PASCAL VOC 2007 dataset

方法	骨干网络	输入图像尺寸/(像素×像素)	模型大小/MB	参数量/MB	mAP/%	FPS	FLOPs
SSD	VGG-16	300×300	200.6	206.89	76.06	88	31.4
FSSD	VGG-16	300×300	130.2	263,84	76.93	47	34.96
DSSD	ResNet-101	320×320	937.2	400.12	77.04	28	35.97
RFB	VGG-16	300×300	139.5	258.23	79.28	52	35.51
MDFNI1	VGG-16	320×320	130.2	250.49	79.3	68	43.95
MDFNI2	VGG-16	320×320	140.8	252.82	78.3	56	44.23
NL-RFB	VGG-16	300×300	151.5	272.26	**80.74**	50	40.49

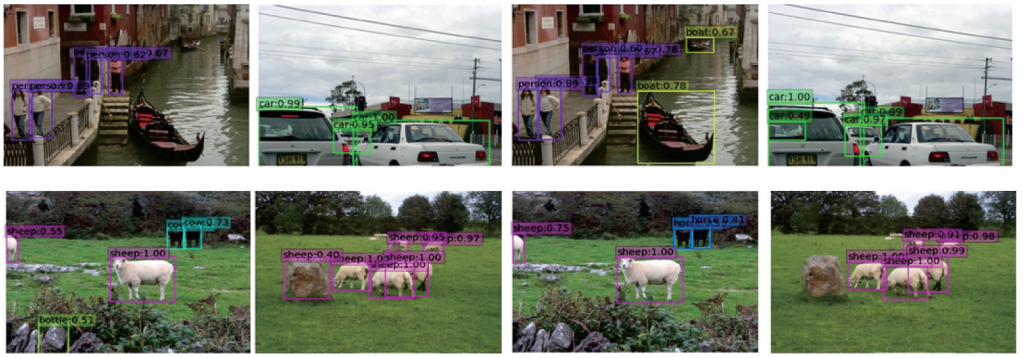
表4展示了在PASCAL VOC 2007测试集下每一类的检测精度的对比。从表4分析可知,本文提出的NL-RFB模型在大多数类别上的检测精度都有着出色的表现,尤其是对于Bottle、Plant等小目标,相较于之前的先进方法都有着明显的提升。

表4　在PASCAL VOC 2007数据集上每一类的检测精度对比

Table 4　Per-class comparison of detection AP on PASCAL VOC 2007 dataset　%

类别	NL-RFB	SSD	FSSD	DSSD	RFB	MDFN I1	MDFNI2
Aeroplan	**84.33**	80.62	79.41	81.94	82.82	81.2	82.5
Bike	**87.81**	82.74	83.35	85.74	87.56	87.0	85.9
Bird	**80.83**	74.31	77.97	78.57	78.37	79.2	78.0
Boat	75.35	70.10	69.48	67.98	**78.37**	72.3	70.5
Bottle	**60.67**	49.31	50.36	46.01	55.82	57.0	54.0
Bus	**88.51**	84.09	85.51	85.44	87.80	87.3	87.9
Car	87.58	85.39	86.53	85.66	87.81	87.1	**87.9**
Cat	87.84	87.48	87.33	87.61	88.19	87.5	**88.6**
Chair	64.13	58.89	56.28	60.81	**64.46**	63.1	60.3
Cow	**85.48**	80.33	84.85	80.14	84.12	84.2	82.6
Table	76.58	73.55	75.84	74.38	74.23	**76.7**	73.7
Dog	85.62	83.23	85.38	85.74	84.36	**87.6**	86.7
Horse	88.47	85.46	86.48	87.29	88.20	**88.8**	87.5
Motorbike	**87.71**	82.84	84.88	84.74	86.02	85.6	85.0
Person	**82.45**	78.45	78.7	77.65	81.62	81.0	80.6
Plant	**56.66**	49.78	50.85	50.26	55.26	56.0	52.3
Sheep	**83.32**	75.21	75.66	79.65	78.73	80.4	77.9
Sofa	**81.59**	79.16	75.64	80.67	79.89	79.9	80.6
Train	88.05	85.44	86.90	84.2	88.03	88.0	**88.1**
Tv	**82.11**	74.79	77.25	76.34	79.32	77.0	76.6

图7是NL-RFB与基模型RFB在PASCAL VOC上的可视化检测结果,可以看出在一些目标检测复杂的情况下,RFB存在一定程度的漏检和错检,比如漏检了第一幅图中的船、错误地认为靠近羊群的石头是羊,而NL-RFB能够正确地检测出复杂场景下的目标,它可以成功识别出第一幅图上远处的船,



(a) RFB (b) NL-RFB

图7 PASCAL VOC2007测试集上的检测结果

Fig.7 Detection results on PASCAL VOC2007 test dataset

并对一些遮挡的目标也能全部检测出来。

为了更好地证明本文提出的改进模块对检测结果都有着积极的影响，本文在PASCAL VOC数据集上设计了消融实验，具体如表5所示。从表5可知，加入CFFA模块和NLFI模块后，模型的mAP都有着一定程度的提升。表6为消融实验后每个类别的AP值，可以看出加入CAFF模块和NLFI模块可以改善不同尺度目标的检测性

表5 消融实验

Table 5 Ablation study

RFB基线	加入CFFA	加入NLFI	NL-RFB	mAP/%
√				79.28
√	√			80.03(+0.75)
√		√		80.12(+0.84)
√	√	√	√	80.74(+1.45)

表6 消融实验后每个类别的AP值

Table 6 AP values of each category after ablation study %

类别	RFB	加入CFFA	加入NLFI	NL-RFB
Aeroplan	82.82	82.48	84.08	84.33
Bike	87.56	86.90	86.90	87.81
Bird	78.37	79.62	78.73	80.83
Boat	78.37	73.53	74.49	75.35
Bottle	55.82	59.06	60.02	60.67
Bus	87.80	87.72	88.67	88.51
Car	87.81	87.73	87.43	87.58
Cat	88.19	86.81	87.99	87.84
Chair	64.46	64.78	64.14	64.13
Cow	84.12	83.86	85.31	85.48
Table	74.23	80.45	78.11	76.58
Dog	84.36	85.81	85.20	85.62
Horse	88.20	88.03	87.34	88.47
Motorbike	86.02	87.02	87.41	87.71
Person	81.62	81.85	81.85	82.45
Plant	55.26	55.86	55.48	56.66
Sheep	78.73	79.47	81.54	83.32
Sofa	79.89	80.40	80.20	81.59
Train	88.03	88.20	87.73	88.05
Tv	79.32	80.98	79.75	82.11

能,如加入CFFA模块后Bottle的AP值从55.82%提升到59.06%,而加入NLFI模块后Bottle的AP值从55.82%提升到60.02%。

3　结束语

当前的特征融合主要采用通道级联和同位置元素相加方法,但是这两种方法大多只关注局部信息,使得模型在融合过程中丢失了很多的有效信息。针对该缺陷,本文设计了通道融合增强模块以及非局部特征交互模块,改善了不同层间特征图的融合效果,成功捕捉不同层不同像素位置之间的长距离依赖关系;并设计双金字塔检测结构,使得模型在保存原始特征信息的基础上,补充了更加丰富的全局上下文语义信息,提升了模型的整体效果;最后,在KITTI与PASCAL VOC数据集上进行了大量的实验,证明了所提方法的有效性。

参考文献:

[1] MA W, WU Y, CEN F, et al. MDFN: Multi-scale deep feature learning network for object detection[J]. Pattern Recognition, 2020. DOI:10.48550/arXiv.1912.04514.

[2] 高新波, 莫梦竟成, 汪海涛,等. 小目标检测研究进展[J]. 数据采集与处理, 2021, 36(3): 391-417.
GAO Xinbo, MO Mengjingcheng, WANG Haitao, et al. Recent advances in small object detection[J]. Journal of Data Acquisition and Processing, 2021, 36(3): 391-417.

[3] ZHAO Z Q, ZHENG P, XU S T, et al. Object detection with deep learning: A review[J]. IEEE Transactions on Neural Networks Learning Systems, 2019, 30(11): 3212-3232.

[4] BOCHKOVSKIY A, WANG C Y, LIAO H Y, et al. Yolov4: Optimal speed and accuracy of object detection[J]. (2020-04-26).https://arXiv.org/abs/2004.10934.

[5] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single shot multibox detector[C]//Proceedings of European Conference on Computer Vision. [S.l.]: Springer, 2016: 21-37.

[6] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014: 580-587.

[7] GIRSHICK R. Fast R-CNN[C]//Proceedings of the IEEE International Conference on Computer Vision. New York: IEEE, 2015: 1440-1448.

[8] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. Advances in Neural Information Processing Systems, 2015, 28: 91-99.

[9] FU C Y, LIU W, RANGA A, et al. DSSD: Deconvolutional single shot detector[EB/OL]. (2017-01-23). https://doi.org/10.48550/arXiv.1701.06659.

[10] LI Z, ZHOU F. FSSD: Feature fusion single shot multibox detector[EB/OL]. (2018-05-17). https://doi. org/10.48550/arXiv.1712.00960.

[11] LIU S, HUANG D. Receptive field block net for accurate and fast object detection[C]//Proceedings of the European Conference on Computer Vision (ECCV). [S.l.]: Springer, 2018: 385-400.

[12] DAI Y, GIESEKE F, OEHMCKE S, et al. Attentional feature fusion[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa: IEEE, 2021: 3560-3569.

[13] 王红梅, 王晓鸽, 王晓燕. 基于深度学习的复杂背景下目标检测研究[J]. 控制与决策, 2021(S): 1-8.
WANG Hongmei, WANG Xiaoge, WANG Xiaoyan. Research on target detection under complex background based on deep learning[J]. Control and Decision, 2021(S): 1-8.

[14] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions[C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston:IEEE, 2015: 1-9.

[15] 王燕妮, 余丽仙. 注意力与多尺度有效融合的SSD目标检测算法[J]. 计算机科学与探索, 2022,16(2): 438-447.

WANG Yanni, YU Lixian. SSD object detection algorithm with effective fusion of attention and multi-scale[J]. Journal of Frontiers of Computer Science and Technology, 2022, 16(2): 438-447.

[16] LIANG X, ZHANG J, ZHUO L, et al. Small object detection in unmanned aerial vehicle images using feature fusion and scaling-based single shot detector with spatial context analysis[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30: 1758-1770.

[17] WOO S, PARK J, LEE J Y, et al. CBAM: Convolutional block attention module[C]//Proceedings of the European Conference on Computer Vision (ECCV). [S.l.]: Springer, 2018: 3-19.

[18] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach:IEEE, 2017: 6000-6010.

[19] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies(NAACL-HLT). Minneapolis: Association for Computational Linguistics, 2019: 4171-4186.

[20] GEIGER A, LENZ P, URTASUN R. Are we ready for autonomous driving? The KITTI vision benchmark suite[C]//Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence: IEEE, 2012: 3354-3361.

[21] EVERINGHAM M, ESLAMI S M A, VAN GOOL L, et al. The pascal visual object classes challenge: A retrospective[J]. International Journal of Computer Vision, 2015, 111(1): 98-136.

作者简介:

马倩(1997-),女,硕士研究生,研究方向:计算机视觉、目标检测,E-mail: 2211031@tongji.edu.cn。

曾凯(1985-),**通信作者**,男,副教授,硕士生导师,研究方向:计算机视觉、分布式计算,E-mail:zengkailink@sina.com。

吴家文(1996-),男,硕士研究生,研究方向:计算机视觉、目标检测等。

沈韬(1984-),男,教授,博士生导师,研究方向:计算机视觉、太赫兹光谱等。

(编辑:王静)