基于注意力机制的单发多框检测器算法

赵 辉* 李志伟 张天琪
(重庆邮电大学通信与信息工程学院 重庆 400065)
(信号与信息处理重庆市重点实验室 重庆 400065)

摘 要：单发多框检测器SSD是一种在简单、快速和准确性之间有着较好平衡的目标检测器算法。SSD网络结构中检测层单一的利用方式使得特征信息利用不充分，将导致小目标检测不够鲁棒。该文提出一种基于注意力机制的单发多框检测器算法ASSD。ASSD算法首先利用提出的双向特征融合模块进行特征信息融合以获取包含丰富细节和语义信息的特征层，然后利用提出的联合注意力单元进一步挖掘重点特征信息进而指导模型优化。最后，公共数据集上进行的一系列相关实验表明ASSD算法有效提高了传统SSD算法的检测精度，尤其适用于小目标检测。

关键词：目标检测；注意力机制；特征融合；单发多框检测器
中图分类号：TN911.73; TP391.41 **文献标识码**：A **文章编号**：1009-5896(2021)07-2096-09
DOI: 10.11999/JEIT200304

Attention Based Single Shot Multibox Detector

ZHAO Hui LI Zhiwei ZHANG Tianqi
(*School of Communication and Information Engineering*, *Chongqing University of Posts and Telecommunications*, *Chongqing* 400065, *China*)
(*Chongqing Key Laboratory of Signal and Information Processing*, *Chongqing* 400065, *China*)

Abstract: Single Shot multibox Detector (SSD) is a object detection algorithm that provides the optimal trade-off among simplicity, speed and accuracy. The single use of detection layers in SSD network structure makes the feature information not fully utilized, which will lead to the small object detection are not robust enough. In this paper, an Attention based Single Shot multibox Detector (ASSD) is proposed. The ASSD algorithm first uses the proposed two-way feature fusion module to fuse the feature information to obtain the feature layer which containing rich details and semantic information. Then, the proposed joint attention unit is used to mine further the key feature information to guide the model optimization. Finally, a series of experiments on the common data set show that the ASSD algorithm effectively improves the detection accuracy of conventional SSD algorithm, especially for small object detection.

Key words: Object detection; Attention mechanism; Feature fusion; Single Shot multibox Detector (SSD)

1 引言

目标检测作为计算机视觉领域最基础的问题之一，为解决图像分割[1]、目标跟踪[2]和活动识别等更复杂或更高层次的视觉任务奠定了基础。随着深度卷积神经网络(Deep Convolutional Neural Networks, DCNNs)的发展，基于DCNNs的目标检测器算法[3–11]成为主流，根据是否需要候选区域提议可以分为两类：基于候选区域提议的目标检测算法和基于回归的目标检测算法。Girshick等人[3]提出的RCNN算法开创了基于候选区域提议的目标检测算法的先河。随后，He等人[4]提出的SPPNet通过在RCNN网络结构中的卷积层和全连接层之间引入传统的空间金字塔池(Spatial Pyramid Pooling, SPP)有效解决了RCNN网络结构中输入任意大小图片的问题，但SPPNet仍然需要为每个目标候选区域提取特征。Girshick[5]提出的Fast RCNN通过添加一个RoI(Region of Interest)池化层将目标候选区域映射到VGG-16[6]网络的特征层上，直接在特征层上提取对应区域的特征，有效解决了RCNN和SPPNet提取特征时冗余的缺点，但由于仍然需

收稿日期：2020-04-24；改回日期：2021-02-15；网络出版：2021-03-31
*通信作者： 赵辉 zhaohui@cqupt.edu.cn
基金项目：国家自然科学基金(61671095)
Foundation Item: The National Natural Science Foundation of China (61671095)

要外部的目标候选区域提议，因此它依然不是一个端到端的训练。Ren等人[7]提出的Faster RCNN通过设计的一种高效、准确的区域提议网络(Region Proposal Network, RPN)来共享网络的卷积特性和生成目标候选区域，实现了检测器端到端的训练。另外，Redmon等人[8]提出的You Only Look Once(YOLO)和Liu等人[9]提出单发多框检测器算法(Single Shot multibox Detector, SSD)则真正意义上推广了基于回归的目标检测算法。YOLO以GoogLeNet[10]为基础骨干网来提取特征，随后采用全连接层来预测目标物体的类别概率和坐标，但由于包围框位置、比例尺和高宽比的粗略划分，YOLO可能无法定位小目标对象。随后Redmon和Farhadi[11]通过更换基础骨干网和采用多尺度等策略提出了YOLO的改进版本YOLOv2。SSD以VGG-16为基础骨干网，通过人为设定的包围框并结合多尺度的思想，在实现了快速检测的同时取得了与当时最先进的基于区域提议的检测器相媲美的准确性。

SSD利用基础骨干网不同输出层的特征图来进行目标检测，它将不同层的检测结果合并起来，并采用非极大值抑制(Non-Maximum Suppression, NMS)[12]算法来生成最终的检测结果。虽然它在速度和精度上都有很好的表现，但仍存在改进提高的空间。SSD算法用于目标检测的多个特征层都是独立使用的，并没有考虑不同输出层之间的关系，这忽略了对于小目标检测很重要的上下文信息，会导致特征信息利用不够充分、小目标检测不够鲁棒的问题[13]。

目标检测任务的基础骨干网源自分类网络，随着深度神经网络的发展，网络结构越来越深，出现了如ResNet[14]和DenseNet[15]等可以获得更好特征表达能力的基础骨干网络。为了解决上述问题，提高传统SSD算法的精度，DSSD[16]用ResNet-101替换了原SSD算法的VGG16，并提出通过反褶积层来聚合上下文信息，增强浅层特征的高层语义。DSOD[17]使用参数效率高于ResNet的DenseNet作为基础骨干网络，研究如何从头开始训练目标检测器。然而复杂的网络结构意味着巨大的网络参数，这会导致检测速度的下降。

另外，研究人员发现基础骨干网络中的深层特征具有更多的语义信息，而浅层特征具有更多的内容描述[18]，浅层的特征信息和深层的特征信息在目标检测中是互补的，如FPN[19]就利用横向连接和自上而下的路径来创建特征金字塔并实现更强大的表示。近年来，RSSD[20]通过设计一种同时采用池化和反卷积操作的彩虹(rainbow)连接方式来对不同特征层进行特征融合。FSSD[21]通过将多个不同尺度的浅层特征层进行特征融合以生成一个大尺度特征，随后在这个大尺度的特征层上再通过下采样的方式构建新的用于检测的特征金字塔。虽然这些方法都有效提升了传统SSD算法的精度，但它们复杂的特征融合方式大大降低了检测速度。因此如何利用浅层和深层的特征来有效地集成特征金字塔表示决定了检测性能。

此外，人类的视觉注意力机制可以通过对全局图像进行快速扫描来获取大脑需要重点关注的目标区域，随后再对这一目标区域投入更多视觉注意力资源来获取该区域内所需关注目标的更多细节信息。视觉注意力机制是人类充分利用有限的注意力资源从海量信息中快速筛选并获取对自己有高价值信息的有效方式。近年来，Mnih等人[22]受人类视觉的注意力机制启发，通过在循环神经网络(Recurrent Neural Network, RNN)模型中使用注意力机制来提升图像分类的性能。随后Bahdanau等人[23]通过使用注意力机制使得机器翻译任务内的翻译和对齐能够同时进行，开启了注意力机制在自然语言处理(Natural Language Processing, NLP)领域中应用的先河。在传统SSD算法中，在无注意力机制下，输入特征层的每个区域或者通道对于输出特征层作用是相同的，因此受注意力机制的启发，传统SSD算法中的特征层也可以通过引入注意力机制来增强特征信息，进而提高小目标检测的性能。

本文的目标是在增加较少计算开销的情况下尽可能提高SSD的精度，同时避免现有方法的局限性。为此，本文首先基于深度卷积网络深层和浅层网络特征层的特征信息侧重点不同这一事实，提出一种多尺度的双向特征融合模块(Two-way Feature Fusion Module, TwFFM)来对传统SSD算法的检测层进行高效的特征融合，以获取包含丰富细节和语义信息的特征层，有利于提高小目标检测的效率。其次，本文进一步提出并采用一种联合注意力单元(Joint Attention Unit, JAU)来自适应地从空间和特征通道两个方向挖掘重要和有效的特征信息，通过将其嵌入到传统SSD算法检测层后，进一步强调检测层的有效信息和抑制无效信息。因此，本文在不更换基础骨干网的前提下，传统SSD算法的检测层之间的特征信息得到充分利用和增强，在提高检测精度的同时可以大大减少计算开销。最后，基于提出的TwFFM和JAU，本文提出一种基于注意力机制的SSD算法(Attention based Single Shot multibox Detector, ASSD)。

2 本文方法

ASSD算法的基本思想是在不改变基础骨干网的前提下，改变检测层单一的使用方式，利用双向特征融合模块TwFFM和联合注意力单元JAU来改善SSD对小目标检测不够鲁棒的问题，尽可能提升其检测性能。如图1所示，TwFFM用来充分挖掘传统SSD特征金字塔层之间的特征信息，使得融合后的特征层能包含丰富的几何细节和强大的语义信息，而联合注意力单元JAU则可以进一步区分特征层的不同区域和特征通道的重要性，通过为它们赋予不同的注意力权重来强调它们之间不同的重要性，以便快速从中选取更加重要的信息，进而利用最有效的信息来指导模型的优化并提高检测精度。

2.1 双向特征融合模块

双向特征融合模块TwFFM的网络结构如图2(a)所示。首先，该模块通过对传统SSD算法检测层中的深层和浅层以FPN结构进行特征融合来获取包含丰富细节和语义信息的浅层特征。其次，为了获取包含丰富特征信息的深层特征层，该模块还设计一种反向的特征融合结构，即浅层特征层经过下采样后和原始的深层特征层进行深度特征融合。由于感受的大小可以粗略地表示使用上下文信息的程度，而浅层特征层由于简单的下采样将导致深层特征层上下文信息一定程度丢失。因此，如图2(b)所示，本文设计一个有层次的全局先验结构—金字塔池化模块(Pyramid Pooling Module, PPM)来对浅层特征层进行处理，浅层特征层经过PPM后再进行下采样处理可以包含丰富的特征信息。PPM利用包含不同尺度的信息来充分挖掘全局上下文先验知识并构造全局先验表示，这种结构可以有效减少不同子区域间上下文信息的丢失，是一种有效的全局上下文先验模型。

如图2(a)所示，S1(高分辨率)和S2(低分辨率)为前后两个需要融合的特征层，P1(高分辨率)和P2(低分辨率)是经过TwFFM融合后输出的特征层。首先，利用FPN结构对S2和S1进行特征融合得到P1层，在进行特征融合之前需要对S2进行上采样得到S2_up以使其尺度大小与S1保持一致。然后，对S2和S1进行特征融合得到P2层，与上一步不同的是选择对S1进行下采样生成S1_down。在进行下采样之前S1需要经过金字塔池化模块PPM生成S1_P，然后再对S1_P下采样得到S1_down, S1_down尺度大小与S2保持一致，最后与S2融合生成P2层。整个双向特征融合模块TwFFM可以用映射F表示为

$$\mathrm{F}:\begin{cases}\mathrm{P1}=\mathrm{Concat}(\mathrm{S1},\mathrm{F\_up}(\mathrm{S2}))\\ \mathrm{P2}=\mathrm{Concat}(\mathrm{F\_down}(\mathrm{F\_PPM}(\mathrm{S1})),\mathrm{S2})\end{cases}\tag{1}$$

其中，F_up和F_down分别表示上采样和下采样操

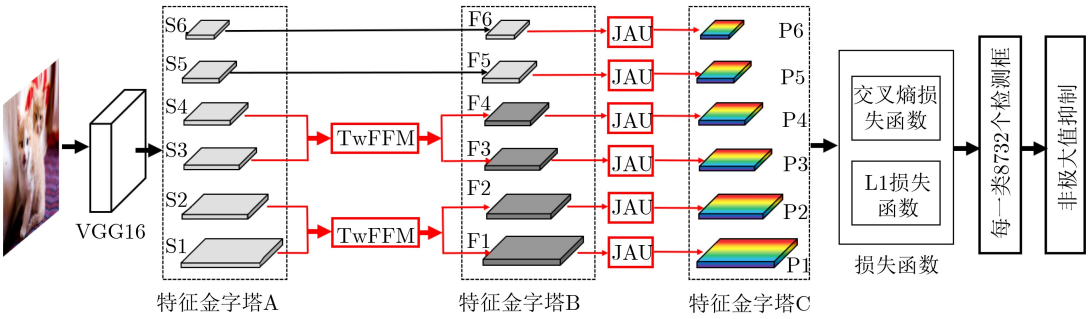
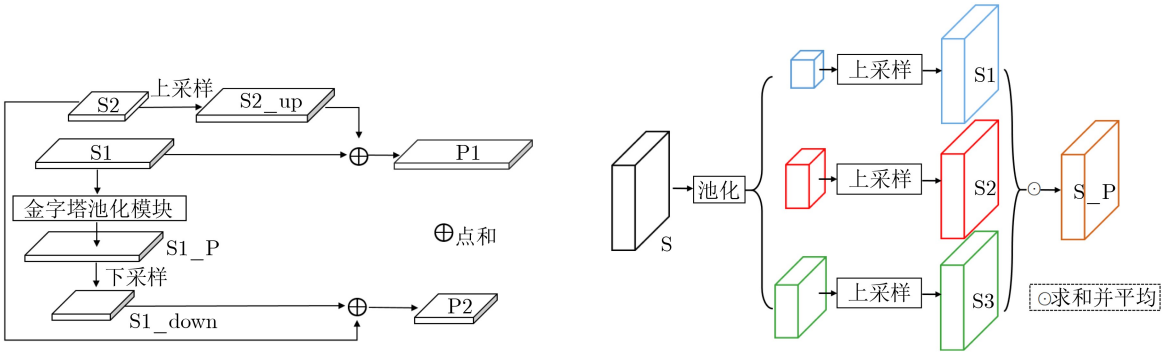


图 1 ASSD算法框架结构图



(a) 双向特征融合模块TwFFM网络结构图　　(b) 金字塔池化模块网络结构图

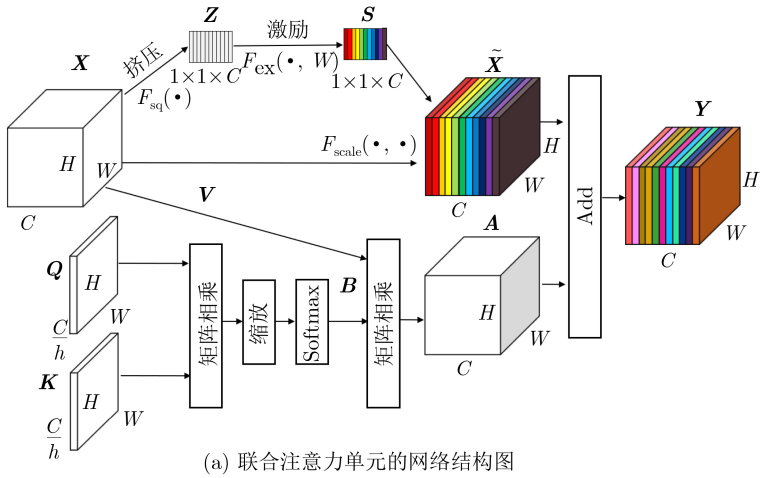
图 2 模块网络结构图

作，F_PPM表示金字塔池化映射，Concat表示点和操作。在金字塔池化模块PPM中，初始特征层S首先经过池化操作生成3种不同大小的特征映射，随后通过双线性插值的上采样方式直接对低维特征图进行上采样并得到与原始特征映射相同尺寸的特征图，最后将3个不同级别的特征求和并平均得到最终的金字塔池化全局特征层S_P。

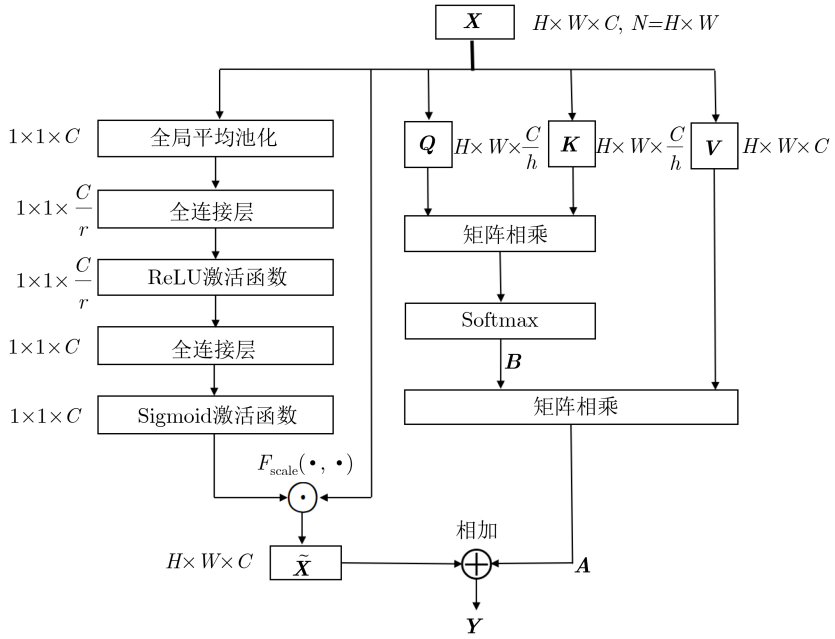
特征融合过程包括下采样、上采样和特征连接3种基本操作。其中上采样和下采样用于调整各层的尺度大小，特征连接通过元素相加实现。此外，还需要3×3卷积层来消除上采样引起的混叠效应。为了将两个特征层连接在一起，需要使用1×1卷积层来统一通道数。为了避免复杂度过高，本文分别选择最大池化(max-pooling)和双线性插值进行下采样和上采样。

2.2 联合注意力单元

联合注意力单元JAU的网络结构和算法流程框图分别如图3(a)和图3(b)所示。联合注意力单元JAU通过一种非线性变换来提取输入序列和输出序列之间的全局依赖关系，JAU由缩放点积注意力(Scaled Dot-Product Attention, SDPA)[24]和挤压激励模块(Sequeeze-and-Excitation Block, SEB)[25]组成。SDPA是一种采用归一化点乘方式计算空间分布相似性的注意力机制。而SEB则是一种通过自适应学习的方式来精确建模卷积特征层内各个通道间相关性的网络结构单元，网络旨在从全局信息出发选择有用的特征通道，抑制无用的特征通道，使提取的特征信息具有更强的指向性。本节的联合注意力单元结合了这两种网络结构单元的优点，能够从空间和通道两个角度同时挖掘输入序列



(a) 联合注意力单元的网络结构图



(b) 联合注意力单元的算法流程框图

图 3 联合注意力单元的网络结构图与算法流程框图

内部的相关性信息，进而提取对于目标任务有用的特征信息。

假定$\boldsymbol{X}=[x_1\ x_2\ \cdots\ x_c]\in\mathbb{R}^{H\times W\times C}$是联合注意力单元的输入特征空间，$H\times W$表示该输入特征层的尺度大小，$C$表示特征通道数，$x_c\in\mathbb{R}^{H\times W}$是输入特征层的第$c$个通道。SEB包含挤压(sequeeze)和激励(excitation)两个映射函数。$F_{\mathrm{sq}}(\cdot)$表示挤压映射，它可以将通道上整个空间特征编码为一个全局特征空间，在实际操作中常用全局平均池化(global average pooling)实现。首先，输入特征x_c经过$F_{\mathrm{sq}}(\cdot)$映射得到全局特征空间$\boldsymbol{Z}\in\mathbb{R}^{H\times W\times C}$的第$c$个全局特征$z_c$

$$z_c=F_{\mathrm{sq}}(x_c)=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}x_c(i,j) \tag{2}$$

在得到全局特征空间后，SEB利用激励映射$F_{\mathrm{ex}}(\cdot,\boldsymbol{W})$来自适应地学习通道之间的非线性关系。整个过程可以用式(3)表示

$$\begin{aligned}\boldsymbol{S}&=F_{\mathrm{ex}}(\boldsymbol{Z},\boldsymbol{W})=\sigma(g(\boldsymbol{Z},\boldsymbol{W}))\\&=\sigma(\boldsymbol{W}_2\ \mathrm{ReLU}(\boldsymbol{W}_1\boldsymbol{Z})),\ \boldsymbol{W}_1\in\mathbb{R}^{\frac{c}{r}\times c},\ \boldsymbol{W}_2\in\mathbb{R}^{c\times\frac{c}{r}}\end{aligned} \tag{3}$$

为了降低模型复杂度以及提升泛化能力，在实际中，本文采用包含两个全连接(Fully Connected, FC)层的瓶颈层结构来实现。第1个FC层为降维层，参数为$\boldsymbol{W}_1$, r为降维系数，g表示的ReLU激活函数。最后一个FC层的作用是恢复原始的维度，参数为$\boldsymbol{W}_2$，σ表示的sigmoid激活函数用于得到权重系数。最后，将学习到的各个通道的权重向量$\boldsymbol{S}$乘以原始输入特征空间$\boldsymbol{X}$以进行特征通道加权(F_{scale})得到SEB的输出特征空间$\tilde{\boldsymbol{X}}=[\tilde{x}_1\ \tilde{x}_2\ \cdots\ \tilde{x}_c]\in\mathbb{R}^{H\times W\times C}$, $\tilde{x}_c$表示输出特征空间$\tilde{\boldsymbol{X}}$的第c个卷积核，s_c表示第c个通道的权重

$$\tilde{x}_c=F_{\mathrm{scale}}(x_c,s_c)=s_c\times x_c \tag{4}$$

随后，对于SDPA单元，同样对于这个输入特征空间$\boldsymbol{X}$，首先将特征空间$\boldsymbol{X}$线性变换为3个不同的特征空间$\boldsymbol{Q}$, $\boldsymbol{K}$和$\boldsymbol{V}$，其公式分别为

$$\boldsymbol{Q}=\boldsymbol{W}_q^{\mathrm{T}}\boldsymbol{X},\ \ \boldsymbol{W}_q\in\mathbb{R}^{C\times C'} \tag{5}$$

$$\boldsymbol{K}=\boldsymbol{W}_k^{\mathrm{T}}\boldsymbol{X},\ \boldsymbol{W}_k\in\mathbb{R}^{C\times C'} \tag{6}$$

$$\boldsymbol{V}=\boldsymbol{W}_v^{\mathrm{T}}X,\ \ \boldsymbol{W}_v\in\mathbb{R}^{C\times C} \tag{7}$$

其中，$\boldsymbol{W}_q,\boldsymbol{W}_k\in\mathbb{R}^{C\times C'}$，$\boldsymbol{W}_v\in\mathbb{R}^{C\times C}$, $C'=C/h$, h为降维系数，空间维度的降低是为了降低计算成本。注意力得分矩阵$\boldsymbol{B}\in\mathbb{R}^{N\times N}$由$\boldsymbol{Q}$和$\boldsymbol{K}$的矩阵乘法得到

$$\boldsymbol{B}=\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{K} \tag{8}$$

注意力得分矩阵$\boldsymbol{B}$的每一行通过一个Softmax函数进行标准化

$$\bar{b}_{ij}=\frac{\exp(b_{ij})}{\sum_{j}^{N}\exp(b_{ij})},\ \ i,j=1,2,\cdots,N \tag{9}$$

其中，$\bar{\boldsymbol{B}}_i$描述了要查询的特征图其第i个位置的像素关系，对$\boldsymbol{Q}$和$\boldsymbol{K}$的矩阵乘法是为了得到相似度，并创建维度为$N\times N$的显示特征关系矩阵，即注意力得分矩阵$\boldsymbol{B}$，也就是权重系数矩阵，这种像素关系也是通过网络学习得到的。随后，对$\boldsymbol{B}$和$\boldsymbol{V}$进行矩阵乘法，为特征图上每个位置的特征进行加权求和，得到SDPA单元的输出特征空间$\boldsymbol{A}$

$$\boldsymbol{A}=\boldsymbol{B}\cdot\boldsymbol{V} \tag{10}$$

最后，联合注意力单元网络将分别从特征通道和空间两个不同角度建模，得到重点特征信息$\tilde{\boldsymbol{X}}$和$\boldsymbol{A}$，并进行相加得到整个联合注意力单元的输出$\boldsymbol{Y}$

$$\boldsymbol{Y}=\tilde{\boldsymbol{X}}+\boldsymbol{A} \tag{11}$$

SDPA单元的输出特征空间$\boldsymbol{A}$将输入特征中所有位置的长期依赖关系联系起来，进而挖掘了特征图的全局上下文信息，它突出特征图的相关部分，用细化后的信息来指导检测任务。而SEB单元的输出特征空间$\tilde{\boldsymbol{X}}$通过自适应学习的方式，显示建模特征图通道间的相互依赖关系，进而强化重要通道上有用的特征信息，弱化抑制非重要通道上无用的特征信息。

3 实验与结果分析

3.1 实验设置

(1)数据集：为了充分保证实验对比分析的公平性，本文所有实验均采用相同的公开数据集：Pascal VOC2007和Pascal VOC2012[26]。训练集由Pascal VOC2007 trainval和Pascal VOC2012 trainval组成，用来评价性能指标的是测试集Pascal VOC2007 test。ASSD算法采用与传统SSD算法相同的数据增强策略，目的是让检测器对多尺度和多颜色的目标具有鲁棒性。

(2)实验环境：本文所有实验都是以Nivida GTX1080Ti GPU和Intel i7-7800x 3.50 GHz CPU等硬件上完成的，软件环境为Ubuntu16.04系统下的Python3.5和TensorFlow1.4加CUDA8.0深度学习框架。

(3)网络结构：与传统SSD算法一致，ASSD算法选择通过ImageNet数据集上预先训练的VGG-16网络为基础骨干网来提取检测层，其中最后两个全连接层Fc6和Fc7分别转换为卷积层Conv_6和

Conv_7。如图2所示，把SSD算法特征金字塔中的Conv4_3, Conv7, Conv8_2, Conv9_2, Conv10_2和Conv11_2分别记作：S1, S2, S3, S4, S5和S6。由于特征融合的结构在高层的语义特征融合效果并不好，因此本文根据实际情况选择前4层(S1, S2, S3和S4)采用TwFFM进行特征融合。其中S1和S2是一对特征层，经过TwFFM生成F1和F2, S3和S4是另一对特征层，并经过TwFFM生成F3和F4，而S5和S6保持不变，并改记为F5和F6。随后对F1, F2, F3, F4, F5和F6采用JAU生成P1, P2, P3, P4, P5和P6并构成ASSD算法的检测层。ASSD算法的检测层在尺度大小和特征通道数上与传统SSD算法保持一致。

(4)训练细节：ASSD算法输入图像的尺度分300×300和512×512两种。采用批量梯度下降算法进行优化。批量大小设置为32，最大迭代次数为120 k次，采用Warm-up策略来设置学习率，前1 k次迭代的学习率为10^{-4}，从1 k～80 k次，学习率则提升至10^{-3}，随后从80 k～100 k次，学习率则降低至10^{-4}，最后100 k～120 k次，学习率再降为10^{-5}。为保证实验的公平性，本文首先训练和评估了传统SSD算法，然后对ASSD算法进行对应的仿真实验。

3.2 评价指标

目标检测领域最常用的评价指标是多类别平均精度(mean Average Precision, mAP)和每秒帧速率(frame per second, fps)。每一类都可以根据召回率(Recall)和准确率(Precision)绘制Precision-Recall(P-R)曲线，AP是P-R曲线下区域的面积。图4显示了ASSD算法和SSD算法的两个输入尺度(300×300和512×512)在Pascal VOC2007 test下的P-R曲线对比图。从图中可以看到，ASSD相比于SSD算法在召回率R≤0.5时有着几乎相同的平均精度，而当召回率R>0.5时，ASSD算法明显比SSD算法更好。考虑到大多数目标检测算法的实际应用情况，在高召回率(>0.6)下测量的精度值比在低召回率下测量的精度值更为重要，通过P-R曲线对比图，可以清楚地看到ASSD算法比SSD算法更有效。

3.3 消融实验

如表1所示，为了验证和评价本文提出的ASSD算法以及双向特征融合模块TwFFM和联合注意力单元JAU的有效性，本文在Pascal VOC2007 test进行了一系列的消融实验。为保证实验的公平性，本文首先训练并评估了传统SSD算法，结果表明，对于300×300的输入，ASSD算法的mAP为79.1%，比传统SSD提高了1.6%，fps由于参数量的增加(Δ24.57 M)略有下降，达到了39.6。对于512×512的输入，ASSD算法的mAP为81.0%，比传统SSD提高了1.2%，参数量增加了24.57 M，因此fps也略下降至20.8。而进一步的消融实验结果显示，单独利用TwFFM来改进SSD算法取得的mAP为78.8%和fps为47.1，参数量则增加了14.04 M，单独利用JAU来改进SSD算法取得的mAP为78.6%和fps为48.1，对应增加的参数量为10.53 M，而单独利用SDPA单元和SEB网络来改进SSD算法其mAP则分别达到了78.2和78.4，fps则分别为53.1和51.5，参数量则分别增加了6.24 M和4.29 M。可以发现，本文提出的双向特征融合模块TwFFM和联合注意力单元JAU是十分有效的，且JAU明显比SDPA和SEB更有效，因为其结合了两者的优点，可以从空间和通道两个方向同时挖掘重点信息。而在参数量方面，TwFFM则相比于JAU更复杂，其增加的参数量相对也更多，JAU由于结合了SDPA和SEB，因此其增加的参数量为两者增加的参数量之和。整体而言，ASSD算法增加的全部参数量都来源于TwFFM和JAU，相比利用复杂的基础骨干网和复杂的特征融合方法，本文的ASSD算法在增加少量计算开销的同时取得了较大的性能提升，是十分有效的。

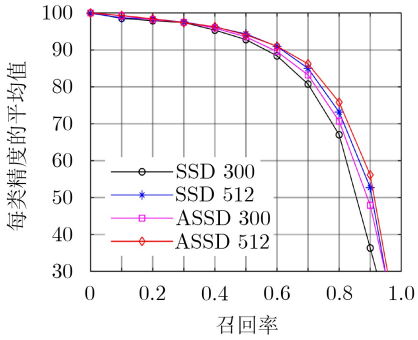


图 4 SSD和ASSD的P-R曲线对比图

表 1　Pascal VOC2007 test上的ASSD消融实验 (1 M=10^6)

方法	输入尺寸	mAP	#参数量	fps
SSD300	300×300	77.5	50.14 M	60.2
SSD512	512×512	79.8	51.98 M	25.2
SSD300+TwFFM	300×300	78.8	64.18 M	47.1
SSD300+SDPA[22]	300×300	78.2	56.38 M	53.1
SSD300+SEB[23]	300×300	78.4	54.43 M	51.5
SSD300+JAU	300×300	78.6	60.67 M	48.1
ASSD300	300×300	79.1	74.71 M	39.6
ASSD512	512×512	80.9	76.55 M	20.8

3.4 与其它先进目标检测算法的性能比较

为了进一步验证和评价本文的ASSD算法，本

文将其与其它一些先进的目标检测算法进行了比较分析，所涉及的算法训练集和测试集都与本文的ASSD算法保持一致。表2表明对于300×300的输入尺度，ASSD算法相比于YOLOv2[11], Faster R-CNN[7], SSD, DSSD[16], RSSD[20]和FSSD[21]精度更高，而对于512×512的输入尺度，只比DSSD精度略低，这是因为DSSD使用了更好的基础骨干网ResNet-101。在检测速度上，由于TwFFM和JAU带来网络复杂度的增加，ASSD算法相比于传统SSD算法略有下降，但与Faster R-CNN、DSSD和RSSD相比明显更快，这主要是因为本文提出的TwFFM复杂度低，JAU的网络参数也较小。DSSD速度较慢的主要原因是其基础骨干网ResNet的复杂度过高，而RSSD速度较慢主要是由于其特征融合方式相对更复杂。ASSD相比FSSD速度略低，主要是因为FSSD通过将传统SSD算法的多个不同尺度的浅层特征层进行特征融合来生成一个大尺度特征，随后再通过下采样的方式重新构建新检测层，虽然相比传统SSD算法在速度上和精度上取得了不错的提升，但算法可移植性较低。

表 2 各种算法在Pascal VOC2007测试集上的性能对比

方法	基础骨干网	输入尺寸	mAP	fps
YOLOv2[11]	Darkent-19	416×416	76.8	67
YOLOv2+[11]	Darkent-19	544×544	78.6	40
Faster RCNN[7]	ResNet-101	~1000×600	76.4	2.4
SSD300	VGG-16	300×300	77.5	60.2
SSD512	VGG-16	512×512	79.8	25.2
DSSD321[16]	ResNet-101	300×300	78.6	9.5
DSSD513[16]	ResNet-101	513×513	81.5	5.5
RSSD300[20]	VGG-16	300×300	78.5	35.0
RSSD512[20]	VGG-16	300×300	80.8	16.6
FSSD300[21]	VGG-16	300×300	78.8	65.8
FSSD512[21]	VGG-16	512×512	80.9	35.7
ASSD300	VGG-16	300×300	79.1	39.6
ASSD512	VGG-16	512×512	81.0	20.8

3.5 Pascal VOC2007测试集中20类性能对比

为了进一步验证ASSD算法能否有效提高传统SSD算法对于小目标检测的精度，表3对ASSD算法和传统SSD算法在Pascal VOC2007测试集中20个类别的精确度进行了比较。结果表明，与传统SSD算法相比，ASSD算法在aerp, bike, bird, boat, bottle, cow, dog, plant, sheep, sofa和tv等类别上精确度有明显提高，尤其是对于bird, bottle, dog, plant, sheep和tv这些小目标出现频次高的类别。

3.6 Pascal VOC2012测试集中样例图片对比

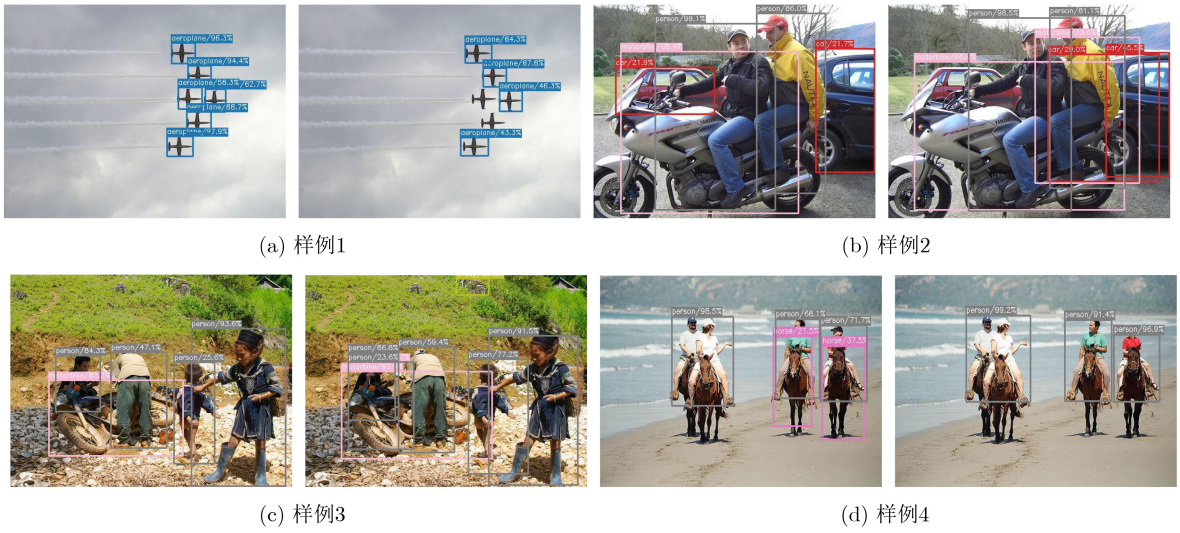
为更加直观地评价本文的ASSD算法。如图5所示，本文在Pascal VOC2012测试集随机选取样例图片对ASSD算法和传统SSD算法进行了仿真测试和对比分析。其中，图5中左侧是ASSD算法的效果图，右侧是传统SSD算法的结果，不同彩色的框代表不同的类别。结果表明，ASSD算法相比于传统SSD算法有明显的优势，能有效减少检测过程中目标的遗漏和误检的发生，并且对于目标的定位更加准确，具有更强的检测和识别小规模物体的能力。

4 结束语

本文提出了一种基于注意力机制的单发多框检测器ASSD算法。首先，提出一种双向特征融合模块TwFFM来改进传统SSD算法中检测层单一的利用方式，并对其进行高效的特征融合，融合后的特征层包含丰富的几何细节和强大的语义信息，可以有效提高检测器的性能。其次，在TwFFM的基础上还提出了一种联合注意力单元JAU，通过自适应学习的方式从空间和通道两个方向同时进行重点特征信息挖掘，从而增强有用信息和抑制无用信息进而指导模型优化，有利于提高目标检测器的检测精度。最后，本文在公共数据集Pascal VOC2007和

表 3 各种算法在Pascal VOC2007测试集中20个类别的性能对比

方法	mAP	aero	bike	bird	boat	bottle	bus	car	cat	chair	**cow**
SSD300	77.5	79.5	83.9	76.0	69.6	50.5	87.0	85.7	88.1	60.3	81.5
SSD512	79.8	85.8	85.6	79.5	74.1	59.9	86.8	88	89.1	63.8	86.3
ASSD300	79.1	85.4	84.1	78.7	71.8	54.0	86.2	85.3	89.5	60.4	87.4
ASSD512	81.0	86.8	85.2	84.1	75.2	60.5	88.3	88.4	89.3	63.5	87.6
方法	mAP	table	dog	horse	mbike	person	plant	sheep	sofa	train	**tv**
SSD300	77.5	77.0	86.1	87.5	84.0	79.4	52.3	77.9	79.5	87.6	76.8
SSD512	79.8	75.8	87.4	87.5	83.2	83.5	56.3	81.6	77.7	87	77.4
ASSD300	79.1	77.1	87.4	86.8	84.8	79.5	57.8	81.5	80.1	87.4	76.9
ASSD512	81.0	76.6	88.2	86.7	85.7	82.8	59.2	83.6	80.5	87.5	80.8



(a) 样例1　(b) 样例2

(c) 样例3　(d) 样例4

图 5 ASSD算法与SSD算法样例检测结果对比图

VOC2012上进行了相关的实验，并与其它一些先进目标检测算法进行了定量和定性的比较分析。结果表明，ASSD算法的性能优于传统SSD算法和其它一些先进的目标检测器算法。ASSD算法能够有效减少误检和漏检的情况，且定位更加准确，对小目标检测具有更强的辨识能力。最后，相关的消融实验表明本文提出的双向特征融合模块TwFFM和联合注意力单元JAU是十分有效的。

参 考 文 献

[1] 赵凤, 孙文静, 刘汉强, 等. 基于近邻搜索花授粉优化的直觉模糊聚类图像分割[J]. 电子与信息学报, 2020, 42(4): 1005–1012. doi: 10.11999/JEIT190428.

ZHAO Feng, SUN Wenjing, LIU Hanqiang, *et al*. Intuitionistic fuzzy clustering image segmentation based on flower pollination optimization with nearest neighbor searching[J]. *Journal of Electronics & Information Technology*, 2020, 42(4): 1005–1012. doi: 10.11999/JEIT190428.

[2] 孙彦景, 石韫开, 云霄, 等. 基于多层卷积特征的自适应决策融合目标跟踪算法[J]. 电子与信息学报, 2019, 41(10): 2464–2470. doi: 10.11999/JEIT180971.

SUN Yanjing, SHI Yunkai, YUN Xiao, *et al*. Adaptive strategy fusion target tracking based on multi-layer convolutional features[J]. *Journal of Electronics & Information Technology*, 2019, 41(10): 2464–2470. doi: 10.11999/JEIT180971.

[3] GIRSHICK R, DONAHUE J, DARRELL T, *et al*. Rich feature hierarchies for accurate object detection and semantic segmentation[C]. Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, USA, 2014: 580–587. doi: 10.1109/CVPR.2014.81.

[4] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, *et al*. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, 37(9): 1904–1916. doi: 10.1109/TPAMI.2015.2389824.

[5] GIRSHICK R. Fast R-CNN[C]. Proceedings of 2015 IEEE International Conference on Computer Vision, Santiago, Chile, 2015: 1440–1448. doi: 10.1109/ICCV.2015.169.

[6] SIMONYAN K and ZISSERMAN A. Very deep convolutional networks for large scale image recognition[C]. Proceedings of the 3rd International Conference on Learning Representations, San Diego, USA, 2015.

[7] 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. doi: 10.1109/TPAMI.2016.2577031.

[8] REDMON J, DIVVALA S, GIRSHICK R, *et al*. You only look once: Unified, real-time object detection[C]. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 779–788. doi: 10.1109/CVPR.2016.91.

[9] LIU Wei, ANGUELOV D, ERHAN D, *et al*. SSD: Single shot MultiBox detector[C]. Proceedings of the 14th European Conference on Computer Vision, Amsterdam, The Netherlands, 2016: 21–37. doi: 10.1007/978-3-319-46448-0_2.

[10] SZEGEDY C, LIU Wei, JIA Yangqing, *et al*. Going deeper with convolutions[C]. Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015: 1–9. doi: 10.1109/CVPR.2015.7298594.

[11] REDMON J and FARHADI A. YOLO9000: Better, faster, stronger[C]. Proceedings of 2017 IEEE Conference on

Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 7263–7271. doi: 10.1109/CVPR.2017.690.

[12] ZEILER M D and FERGUS R. Visualizing and understanding convolutional networks[C]. Proceedings of the 13th European Conference on Computer Vision, Zurich, Switzerland, 2014: 818–833. doi: 10.1007/978-3-319-10590-1_53.

[13] CHEN Chenyi, LIU Mingyu, TUZEL O, *et al*. R-CNN for small object detection[C]. Proceedings of the 13th Asian Conference on Computer Vision, Taipei, China, 2016: 214–230. doi: 10.1007/978-3-319-54193-8_14.

[14] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, *et al*. Deep residual learning for image recognition[C]. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 770–778. doi: 10.1109/CVPR.2016.90.

[15] HUANG Gao, LIU Zhuang, VAN DER MAATEN L, *et al*. Densely connected convolutional networks[C]. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 4700–4708. doi: 10.1109/CVPR.2017.243.

[16] FU Chengyang, LIU Wei, RANGA A, *et al*. DSSD: Deconvolutional single shot detector[J]. arXiv: 1701.06659, 2017.

[17] SHEN Zhiqiang, LIU Zhuang, LI Jianguo, *et al*. Dsod: Learning deeply supervised object detectors from scratch[C]. Proceedings of 2017 IEEE International Conference on Computer Vision, Venice, Italy, 2017: 1919–1927. doi: 10.1109/ICCV.2017.212.

[18] ZEILER M D and FERGUS R. Visualizing and understanding convolutional networks[C]. Proceedings of the 13th European Conference on Computer Vision, Zurich, Switzerland, 2014: 818–833.

[19] LIN T Y, DOLLÁR P, GIRSHICK R, *et al*. Feature pyramid networks for object detection[C]. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 2117–2125. doi: 10.1109/CVPR.2017.106.

[20] JEONG J, PARK H, and KWAK N. Enhancement of SSD by concatenating feature maps for object detection[C]. Proceedings of 2017 British Machine Vision Conference, London, UK, 2017.

[21] LI Zuoxin and ZHOU Fuqiang. FSSD: Feature fusion single shot Multibox detector[J]. arXiv: 1712.00960, 2017.

[22] MNIH V, HEESS N, GRAVES A, *et al*. Recurrent models of visual attention[C]. Proceedings of the 27th International Conference on Neural Information Processing Systems, Montreal, Canada, 2014: 2204–2212.

[23] BAHDANAU D, CHO K, BENGIO Y, *et al*. Neural machine translation by jointly learning to align and translate[C]. Proceedings of the 3rd International Conference on Learning Representations, San Diego, USA, 2015.

[24] 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, USA, 2017: 5998–6008.

[25] HU Jie, SHEN Li, and SUN Gang. Squeeze-and-excitation networks[C]. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 7132–7141. doi: 10.1109/CVPR.2018.00745.

[26] EVERINGHAM M, VAN GOOL L, WILLIAMS C K I, *et al*. The pascal visual object classes (VOC) challenge[J]. *International Journal of Computer Vision*, 2010, 88(2): 303–338. doi: 10.1007/s11263-009-0275-4.

赵 辉：女，1980年生，教授，博士生导师，研究方向为深空通信、信号理论与信息处理、图像信息处理.

李志伟：男，1996年生，硕士，研究方向为计算机视觉、深度学习、目标检测.

张天琪：男，1971年生，教授，博士生导师，研究方向为无线通信的智能信号处理、通信抗干扰和信息对抗.

责任编辑：陈 倩