

中图法分类号: TP391 文献标识码: A 文章编号: 1006-8961(2021)03-0542-14

论文引用格式: Yu W Q , Yu J , Bai M Y and Xiao C B. 2021. Video object detection using fusion of SSD and spatiotemporal features. Journal of Image and Graphics 26(03):0542-0555( 尉婉青 禹晶 柏曼晏 肖创柏. 2021. SSD 与时空特征融合的视频目标检测. 中国图象图形学报 26(03):0542-0555) [DOI: 10.11834/jig.200020]

# SSD 与时空特征融合的视频目标检测

尉婉青 禹晶 柏曼晏 肖创柏

北京工业大学信息学部 北京 100124

**摘要:** 目的 视频目标检测旨在序列图像中定位运动目标, 并为各个目标分配指定的类别标签。视频目标检测存在目标模糊和多目标遮挡等问题, 现有的大部分视频目标检测方法是在静态图像目标检测的基础上, 通过考虑时空一致性来提高运动目标检测的准确率, 但由于运动目标存在遮挡、模糊等现象, 目前视频目标检测的鲁棒性不高。为此, 本文提出了一种单阶段多框检测( single shot multibox detector, SSD) 与时空特征融合的视频目标检测模型。方法 在单阶段目标检测的 SSD 模型框架下, 利用光流网络估计当前帧与近邻帧之间的光流场, 结合多个近邻帧的特征对当前帧的特征进行运动补偿, 并利用特征金字塔网络提取多尺度特征用于检测不同尺寸的目标, 最后通过高低层特征融合增强低层特征的语义信息。结果 实验结果表明, 本文模型在 ImageNet VID( Imagelevel for video object detection) 数据集上的 mAP( mean average precision) 为 72.0%, 相对于 TCN( temporal convolutional networks) 模型、TPN + LSTM( tubelet proposal network and long short term memory network) 模型和 SSD + 李生网络模型, 分别提高了 24.5%、3.6% 和 2.5%, 在不同结构网络模型上的分离实验进一步验证了本文模型的有效性。  
**结论** 本文模型利用视频特有的时间相关性和空间相关性, 通过时空特征融合提高了视频目标检测的准确率, 较好地解决了视频目标检测中目标漏检和误检的问题。

**关键词:** 目标检测; 单阶段多框检测; 特征融合; 光流; 特征金字塔网络

## Video object detection using fusion of SSD and spatiotemporal features

Yu Wanqing , Yu Jing , Bai Manyan , Xiao Chuangbai

Faculty of Information Technology , Beijing University of Technology , Beijing 100124 , China

**Abstract:** **Objective** Object detection is a fundamental task in computer vision applications, which provides support for subsequent object tracking, semantic segmentation, and behavior recognition. Recent years have witnessed substantial progress in still image object detection based on deep convolutional neural network ( DCNN) . The task of still image object detection is to determine the category and position of each object in an image. Video object detection aims to locate a moving object in sequential images and assign a specific category label to each object. The accuracy of video object detection suffers from degenerated object appearances in videos, such as motion blur, multiobject occlusion, and rare poses. The methods of still image object detection achieve excellent results, but directly applying them to video object detection is challenging. According to the temporal and spatial information in videos, most existing video object detection methods improve the accuracy of moving object detection by considering spatiotemporal consistency based on still image object detection.

**Method** In this paper, we propose a video object detection method using fusion of single shot multibox detector ( SSD) and

收稿日期: 2020-02-11; 修回日期: 2020-06-23; 预印本日期: 2020-06-30

基金项目: 北京市教育委员会科技发展计划项目( KM201910005029) ; 北京市自然科学基金项目( 4212014)

**Supported by:** Scientific Research Common Program of Beijing Municipal Commission of Education ( KM201910005029) ; Beijing Municipal Natural Science Foundation ( 4212014)

spatiotemporal features. Under the framework of SSD, temporal and spatial information of the video are applied to video object detection through the optical flow network and the feature pyramid network. On the one hand, the network combining residual network (ResNet) 101 with four extra convolutional layers is used for feature extraction to produce the feature map in each frame of the video. An optical flow network estimates the optical flow fields between the current frame and multiple adjacent frames to enhance the feature of the current frame. The feature maps from adjacent frames are compensated to the current frame according to the optical flow fields. The multiple compensated feature maps as well as the feature map of the current frame are aggregated according to adaptive weights. The adaptive weights indicate the importance of all compensated feature maps to the current frame. Here, the cosine similarity metric is utilized to measure the similarity between the compensated feature map and the feature map extracted from the current frame. If the compensated feature map is close to the feature map of the current frame, then the compensated feature map is assigned a larger weight; otherwise, it is assigned a smaller weight. Moreover, an embedding network that consists of three convolutional layers is applied on the compensated feature maps and the current feature map to produce the embedding feature maps, and the embedding feature maps are used to compute the adaptive weights. On the other hand, the feature pyramid network is used to extract multiscale feature maps that are used to detect the object of different sizes. The low-and high-level feature maps are used to detect smaller and larger objects, respectively. For the problem of small object detection in the original SSD network, the low-level feature map is combined with the high-level feature map to enhance the semantic information of the low-level feature map via upsampling operation and a  $1 \times 1$  convolutional layer. The upsampling operation is used to extend the high-level feature map to the same resolution as the low-level feature map, and the  $1 \times 1$  convolution layer is used to reduce the channel dimensions of the low-level feature map to be consistent with those of the high-level feature map. Then, multiscale feature maps are input into the detection network to predict bounding boxes, and nonmaximum suppression is carried out to filter the redundant bounding boxes and obtain the final bounding boxes. **Result** Experimental results show that the mean average precision (mAP) score of the proposed method on the ImageNet VID (ImageNet for video object detection) dataset can reach 72.0%, which is 24.5%, 3.6%, and 2.5% higher than those of the temporal convolutional network, the method combining tubelet proposal network with long short memory network, and the method combining SSD and siamese network, respectively. In addition, an ablation experiment is conducted with five network structures, namely, 16-layer visual geometry group (VGG16) network, ResNet101 network, the network combining ResNet101 with feature pyramid network, and the network combining ResNet101 with spatiotemporal fusion. The network structure combining ResNet101 with spatiotemporal fusion improves the mAP score by 11.8%, 7.0%, and 1.2% compared with the first four network structures. For further analysis, the mAP scores of the slow, medium, and fast objects are reported in addition to the standard mAP score. Our method combined with optical flow improves the mAP score of slow, medium, and fast objects by 0.6%, 1.9%, and 2.3%, respectively, compared with the network structure combining ResNet101 with feature pyramid network. Experimental results show that the proposed method can improve the accuracy of video object detection, especially the performance of fast object detection.

**Conclusion** Temporal and spatial correlation of the video by spatiotemporal fusion are used to improve the accuracy of video object detection in the proposed method. Using the optical flow network in video object detection can compensate the feature map of the current frame according to the feature maps of multiple adjacent frames. False negatives and false positives can be reduced through temporal feature fusion in video object detection. In addition, multiscale feature maps produced by the feature pyramid network can detect the object of different sizes, and the multiscale feature map fusion can enhance the semantic information of the low-level feature map, which improves the detection ability of the low-level feature map for small objects.

**Key words:** object detection; single shot multibox detector (SSD); feature fusion; optical flow; feature pyramid network

## 0 引言

静态图像目标检测旨在判断一幅图像中目标的类别及其位置。视频目标检测的任务是在序列图像

中定位运动目标，并为各个目标分配指定类别标签。与静态图像目标检测不同，目标模糊、多目标遮挡等因素会影响视频目标检测的性能。在目标检测领域，静态图像目标检测从传统目标检测方法到基于深度学习的目标检测方法，取得了较大进展。传统

目标检测首先使用不同尺寸的滑动窗口在图像中遍历生成候选区域,然后利用特征提取方法对各个候选区域进行特征提取,典型的特征提取方法有Haar、方向梯度直方图(histogram of gradient,HoG)和稀疏编码直方图(histogram of sparse code,HSC)等,最后将特征向量输入到分类器,如支持向量机(support vector machine,SVM)、迭代器(AdaBoost)等实现分类识别。这类方法通过滑动窗口策略生成的候选区域存在大量冗余,会降低检测效率。此外,使用的手工设计特征并不能很好地描述目标,目标检测的鲁棒性不高。

深度学习广泛应用于目标检测领域,基于深度学习的目标检测方法主要分为两阶段检测和单阶段检测。两阶段检测在生成候选区域后,对候选区域进行分类实现目标检测。Girshick等人(2014)设计了基于区域的卷积神经网络(regions with convolutional neural network,R-CNN),首先使用选择性搜索(selective search)(Uijlings等,2013)生成大量的候选区域,然后利用卷积神经网络(convolutional neural network,CNN)(Krizhevsky等,2012)提取各个候选区域的特征,最后应用SVM分类器和目标框回归方法分别对候选区域进行分类和位置的细化,但是该模型需要固定尺寸的输入图像。在R-CNN模型的基础上,He等人(2015)构造了空间金字塔池化网络模型(spatial pyramid pooling network,SPP-Net),将每一个特征图划分为 $4 \times 4$ 、 $2 \times 2$ 和 $1 \times 1$ 的块,通过最大池化操作对各个特征图提取固定长度的特征向量。Fast R-CNN(Girshick,2015)将分类任务和定位任务合并为多任务损失函数并进行模型训练,然而该模型使用的选择性搜索方法会导致候选区域的冗余。Ren等人(2017)提出了Faster R-CNN利用候选区域生成网络(region proposal network,RPN)提取少量且准确的候选区域,并实现了模型的端到端训练。Lin等人(2017)在Faster R-CNN的框架下构造了特征金字塔网络(feature pyramid networks,FPN)模型,通过上采样和横向连接融合高层特征和低层特征形成一组多尺度特征。这类方法分为候选区域生成和目标检测两个阶段,因此检测的准确率较高,但是会增加时间开销。单阶段检测根据位置、尺寸和宽高比在图像中进行密集采样,直接在图像的各个位置上回归出目标边框以及目标所属的类别。Redmon等人(2016)提出了

YOLO(you only look once)模型,利用CNN提取特征并直接输出目标的类别置信度和坐标位置,此模型存在定位不准确和查全率低的问题。Liu等人(2016)提出了单阶段多框检测(single shot multibox detector,SSD)模型,借鉴Faster R-CNN中的锚(anchor)机制构建多尺度检测模型。Zhang等人(2018a)在SSD模型的基础上筛除负锚框来缩小分类器的搜索空间。Zhang等人(2018b)提出了DES(detection with enriched semantics)模型,通过目标检测和基于目标框的弱监督语义分割的多任务学习增强低层特征的语义信息。

视频目标检测通常在静态图像目标检测的基础上,通过考虑时空一致性来提高检测的准确率。Kang等人(2016)在静态图像目标检测的基础上,结合目标跟踪模型提取候选区域,并利用时域卷积网络(temporal convolutional networks,TCN)重新评估候选区域的置信度,但是该模型步骤烦琐,计算量很大。Kang等人(2017)提出了管道候选网络(tubelet proposal network,TPN),利用静态图像目标检测获取数百个不同的管道,一个管道是由视频中每帧图像的同一目标的候选区域串联组成,然后结合长短时记忆网络(long short term memory network,LSTM)进行分类用于视频目标检测。Zhu等人(2017)提出了光流指导的特征融合(flow-guided feature aggregation,FGFA)模型,结合光流网络将多帧图像特征进行融合以提高检测的准确率。Xiao和Lee(2018)提出了时空记忆网络模型(spatial-temporal memory network,STMN),利用时空记忆模块保存视频中的时间信息,从而结合前后帧的特征补偿当前帧的特征。Zhao等人(2018)在SSD目标检测框架下,结合相邻帧之间的相关损失训练模型。

本文提出了一种SSD与时空特征融合的视频目标检测模型,在单阶段目标检测的SSD框架下,利用视频特有的时间相关性和空间相关性,通过时空特征融合提高视频目标检测的准确率。在本文模型中,根据光流场利用多个近邻帧的特征对当前帧的特征进行运动补偿,通过时间特征融合减弱了视频中的目标模糊、多目标遮挡等因素对视频目标检测的影响;利用特征金字塔网络提取多尺度特征用于检测不同尺寸的目标,并通过高低层特征融合以增强低层特征的语义信息,解决小尺寸目标检测问题。

# 1 光流网络与特征金字塔网络

## 1.1 光流网络

光流场表示运动目标在观察成像平面上像素运动的瞬时速度, 利用图像序列中像素在时域上的变化找到近邻帧与当前帧之间的对应关系, 具体为

$$I(x, y, t) = I(x + u, y + v, t + \tau) \quad (1)$$

式中,  $I(x, y, t)$  表示  $t$  时刻图像中  $(x, y)$  处的像素的灰度值,  $\tau$  表示近邻帧与当前帧的距离,  $(u, v)$  表示像素的运动向量。根据光流场以及近邻帧对当前帧进行运动补偿, 从而增强当前帧的目标信息。

Dosovitskiy 等人(2015) 提出了一种基于 CNN 的光流网络 FlowNetS, 利用卷积神经网络提取图像特征并估计光流场, 该模型包括收缩( contracting) 阶

段和扩张( expanding) 阶段。其中, 收缩阶段由卷积层构成, 用于提取图像特征; 扩张阶段由转置卷积层构成, 用于恢复特征的空间分辨率。

图 1 为 FlowNetS 的网络模型, 其输入是沿通道方向堆叠的两幅图像, 输出是  $x$  方向和  $y$  方向上的光流场。该网络模型的收缩阶段包括 6 组卷积层, 第 1 组和第 2 组分别由一个卷积核为  $7 \times 7$  和  $5 \times 5$  的卷积层构成, 步长都为 2, 其余 4 组由两个卷积层构成, 其中第 3 组的第一个卷积层的卷积核为  $5 \times 5$ , 余下卷积层的卷积核都为  $3 \times 3$ , 并且每组的第 2 个卷积层的步长为 2, 因此每经过一组卷积层, 特征的尺寸缩小 1 倍; 扩张阶段由 4 个转置卷积层组成, 为了提供精确的位置信息, 除了上一层的输出, 转置卷积层的输入还包括收缩阶段中同样尺寸的特征以及根据上一层特征估计的光流场。

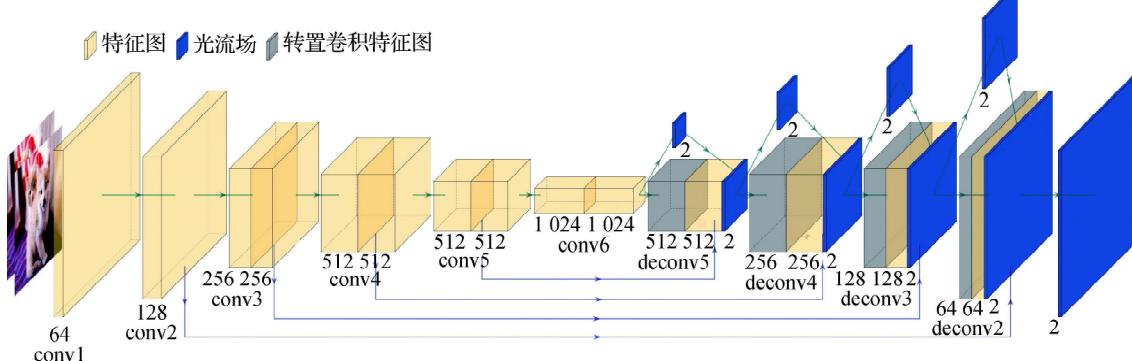


图 1 光流网络 FlowNetS

Fig. 1 Optical flow network FlowNetS

利用卷积神经网络构建的光流网络需要大规模的数据集进行训练, FlowNetS 使用 Middlebury、KITTI ( karlsruhe institute of technology and Toyota technological institute)、Sintel 和 Flying Chairs 数据集进行训练, 其中 Flying Chairs 数据集是合成数据集, 包括 22 872 对标注的图像。

根据真实光流场和估计的光流场, FlowNetS 采用端点误差( endpoint error, EPE) 作为损失函数, 定义为

$$e_{EPE} = \sum_j \sqrt{(\mathbf{u}_j^g - \mathbf{u}_j^e)^2 + (\mathbf{v}_j^g - \mathbf{v}_j^e)^2} \quad (2)$$

式中,  $(\mathbf{u}_j^g, \mathbf{v}_j^g)$  表示像素  $j$  的真实运动向量,  $(\mathbf{u}_j^e, \mathbf{v}_j^e)$  表示在像素  $j$  上估计的运动向量。由式(2) 可知,  $e_{EPE}$  采用欧氏距离度量网络输出误差, 真实运动向量与估计的运动向量的距离越小, 输出误差越小,

否则, 输出误差越大。

## 1.2 特征金字塔网络

为了提升对不同尺寸目标的检测能力, SSD 模型提出了利用卷积神经网络提取的多层特征进行检测。Lin 等人(2017) 构建了特征金字塔网络模型, 充分利用更多层的特征进行检测, 尤其是低层特征, 同时为了增强低层特征的语义信息, 该模型逐层将高层特征传递给低层特征, 使低层特征能够更好地检测小尺寸目标。

特征金字塔网络模型分为特征提取和特征融合两部分, 如图 2 所示。在特征提取过程中, 利用卷积神经网络对一幅图像提取多层不同尺寸的特征, 并按序排列构成特征金字塔。特征融合部分的输入包括两部分, 一部分为高层特征经过特征融合部分的输出, 另一部分为特征金字塔中的低层特征。由于

高层特征的尺寸小于低层特征,通过上采样操作将其扩展至与低层特征相同的尺寸,而低层特征在通道数上与高层特征不一致,采用 $1 \times 1$ 卷积操作统一特征的通道数,最终输入到检测网络中的特征的通道数均为256。

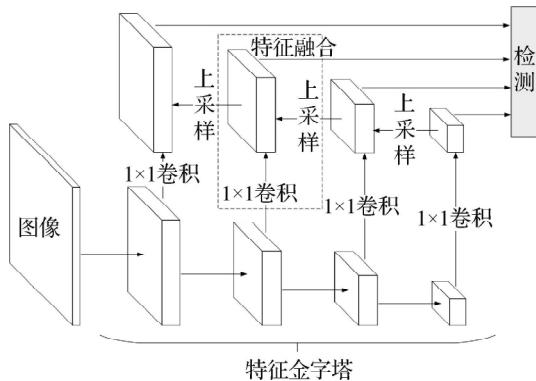


图2 特征金字塔网络

Fig. 2 Feature pyramid network

## 2 SSD 与时空特征融合的视频目标检测模型

根据视频特有的时间相关性,结合光流网络将多帧图像特征相融合,利用特征金字塔网络实现高低层特征融合以提高低层特征的检测能力。

### 2.1 基于 ResNet101 的 SSD 模型

两阶段检测中的候选区域生成过程会增加检测

的时间开销,因此本文采用单阶段检测的SSD模型作为基本检测框架。原SSD模型使用VGG16(16-layer visual geometry group)提取图像特征,VGG16由13个 $3 \times 3$ 卷积核的卷积层和3个全连接层组成,去除全连接层后作为特征提取网络。由于浅层网络模型无法拟合大规模的数据集,本文在SSD框架中采用深层残差网络模型ResNet(residual network)101(He等,2016)作为特征提取网络。ResNet101利用残差结构构建深层网络模型,具有100个卷积层和1个全连接层,避免了深层网络模型中卷积层的增加导致的梯度消失问题。本文模型在ResNet101的基础上,额外增加4组卷积层,为检测网络提供更多层不同尺寸的特征。

图3为基于ResNet101的SSD模型,ResNet101的参数如下:第1组由1个 $7 \times 7$ 卷积核的卷积层组成,滑动步长为2,第2、3、4和5组分别由3、4、23和3个残差结构构成。其中,残差结构包括3个卷积层,卷积核大小分别为 $1 \times 1$ 、 $3 \times 3$ 和 $1 \times 1$ 。此外,在ResNet101的后面增加4组卷积层,每组卷积层包括1个卷积核为 $1 \times 1$ 、滑动步长为1的卷积层和1个卷积核为 $3 \times 3$ 、滑动步长为2的卷积层。最后将提取的7层特征输入检测网络,输出目标框,并通过非极大值抑制形成最终的检测结果。这7层特征的通道数分别为512、1 024、512、512、256、256、128,尺寸分别为 $64 \times 64$ 、 $32 \times 32$ 、 $16 \times 16$ 、 $8 \times 8$ 、 $4 \times 4$ 、 $2 \times 2$ 、 $1 \times 1$ 。

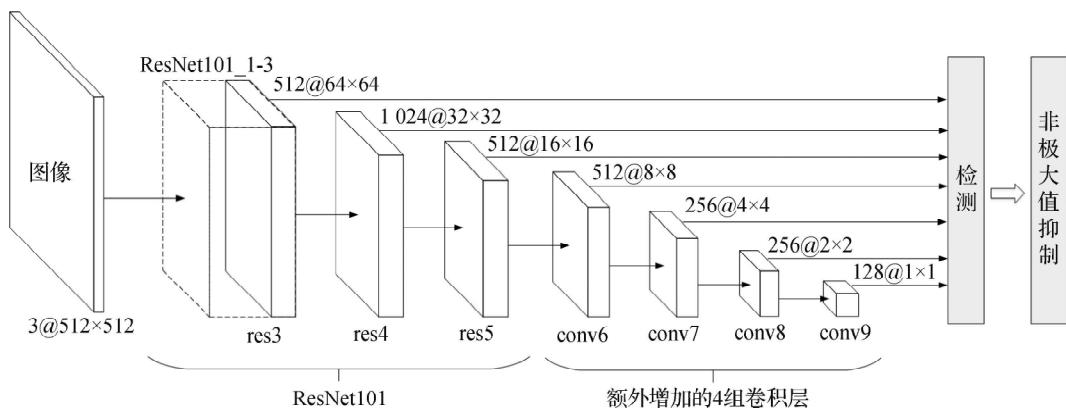


图3 基于 ResNet101 的 SSD 模型

Fig. 3 SSD model based on ResNet101

### 2.2 空间特征融合

在原SSD模型中,检测网络使用多尺度特征进行检测,每个特征中生成不同尺寸和不同宽高比的

先验框。其中,低层特征对应小尺寸先验框,高层特征对应大尺寸先验框,先验框的引入使模型能够检测出各种不同比例和尺寸的目标框。但是SSD模

型对小尺寸目标的检测效果仍然不理想,主要是因为低层特征的感受域范围较小,且包含的语义信息较弱。本文利用特征金字塔网络将高层特征的语义信息传递给低层特征,通过空间特征融合增强低层特征的语义信息。

图 4 为基于 ResNet101 和特征金字塔的 SSD 模型。将图像输入到该模型中提取 7 个不同尺寸的特征,表示为 res3、res4、res5、conv6、conv7、conv8、conv9,由这些特征构成特征金字塔,然后逐层将高

层特征融合到低层特征中。对高层特征,利用最近邻插值进行上采样操作,将其扩展至与低层特征相同的尺寸;对低层特征,通过卷积核为  $1 \times 1$  的卷积操作,使低层特征在通道维度上与高层特征一致。然后将高层特征与低层特征相加生成空间特征融合后的特征,表示为 P0、P1、P2、P3、P4、P5、P6,其对应的尺寸分别为  $64 \times 64$ 、 $32 \times 32$ 、 $16 \times 16$ 、 $8 \times 8$ 、 $4 \times 4$ 、 $2 \times 2$ 、 $1 \times 1$ ,最后将特征输入检测网络中形成最终的检测结果。

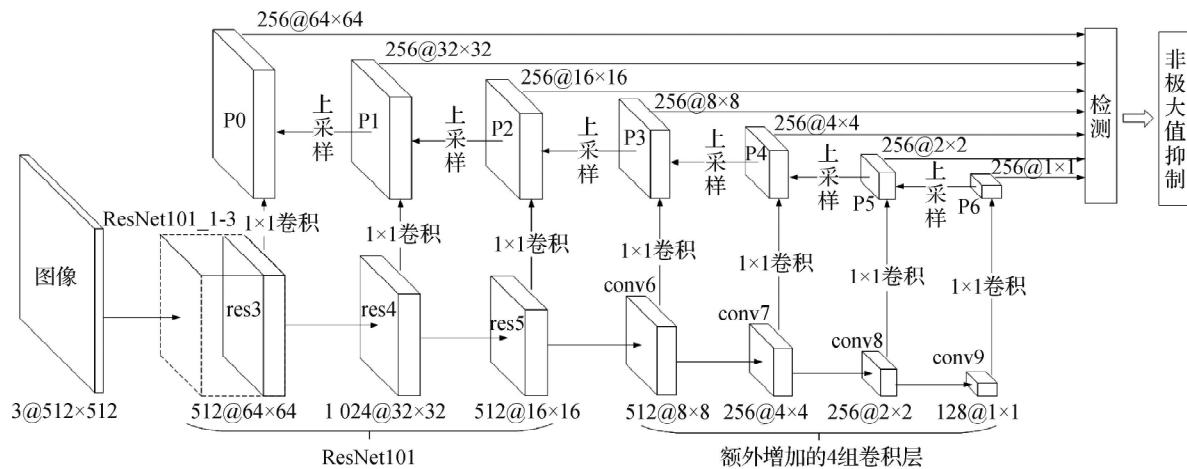


图 4 基于 ResNet101 与特征金字塔的 SSD 模型

Fig. 4 SSD model based on ResNet101 and feature pyramid

检测网络的输出包括目标框的类别置信度和位置偏移量,依照 Liu 等人(2016)的研究,损失函数定义为

$$L(\{p_i\}, \{l_i\}) = \frac{1}{N} (L_{\text{conf}}(\{p_i\}) + \lambda L_{\text{loc}}(\{l_i\})) \quad i = 1, 2, \dots, N \quad (3)$$

式中,  $L_{\text{conf}}(\cdot)$  和  $L_{\text{loc}}(\cdot)$  分别表示分类损失函数和定位损失函数,  $i$  表示目标框的索引,  $p_i$  表示第  $i$  个目标框 softmax 归一化的类别置信度,  $l_i$  表示第  $i$  个目标框相对于先验框的位置偏移量,  $N$  表示正样本的数量,  $\lambda$  为权衡分类损失和定位损失的超参数。

### 2.3 时间特征融合

视频的相邻帧之间通常具有一定的相关性,利用视频的时间相关性,可以将多帧图像特征进行融合以补偿当前帧的特征,从而减弱目标模糊和多目标遮挡等因素对视频目标检测的影响。本文利用 FlowNetS 估计当前帧与近邻帧之间的光流场,并计算当前帧特征与近邻帧特征之间的余弦相似度作为权重进行多帧图像特征融合。

设当前帧和近邻帧的特征分别表示为  $F^t \in$

$\mathbf{R}^{m \times n \times c}$  和  $F^{t+\tau} \in \mathbf{R}^{m \times n \times c}$ , 其中,  $m$ 、 $n$  和  $c$  分别为当前帧特征和近邻帧特征的高、宽和通道数。根据光流场,使用近邻帧的特征  $F^{t+\tau}$  补偿当前帧的特征,表示为  $F^{t+\tau \rightarrow t}$ , 可定义为

$$F^{t+\tau \rightarrow t} = \text{warp}(F^{t+\tau} M^{t+\tau \rightarrow t}) \quad (4)$$

式中,  $M^{t+\tau \rightarrow t} \in \mathbf{R}^{m \times n \times 2}$  表示当前帧与近邻帧之间的光流场,包括在  $x$  和  $y$  两个方向上的光流场,  $\text{warp}(\cdot)$  表示变换函数。

由于视频特有的时间相关性,根据多个近邻帧特征补偿的当前帧特征对当前帧特征进行线性预测,并计算补偿的当前帧特征与当前帧特征的余弦相似度作为线性预测的权重。参照 Zhu 等人(2017)的研究,本文将近邻帧特征补偿的当前帧特征  $F^{t+\tau \rightarrow t}$  和当前帧特征  $F^t$  输入到由 3 个  $1 \times 1$  卷积核的卷积层构成的嵌入卷积神经网络中, 提取其嵌入特征  $E^{t+\tau \rightarrow t} = \psi(F^{t+\tau \rightarrow t})$  和  $E^t = \psi(F^t)$ , 用于权重的计算, 这里  $\psi(\cdot)$  表示嵌入卷积神经网络。为便于描述, 将嵌入特征  $E^{t+\tau \rightarrow t}$  和  $E^t$  表示为 2 维矩阵的形式, 即  $E^{t+\tau \rightarrow t} = [e_1^{t+\tau \rightarrow t}, \dots, e_m^{t+\tau \rightarrow t}] \in \mathbf{R}^{c \times mn}$  和

$E^t = [e_1^t; \dots; e_{mn}^t] \in \mathbf{R}^{c \times mn}$  利用嵌入特征的余弦相似度计算权重  $w_i^{t+\tau-t}$  具体为

$$w_i^{t+\tau-t} = \frac{\langle e_i^{t+\tau-t}, e_i^t \rangle}{\|e_i^{t+\tau-t}\| \|e_i^t\|} \quad (5)$$

式中,  $e_i^{t+\tau-t} \in \mathbf{R}^c$  和  $e_i^t \in \mathbf{R}^c$  分别为  $E^{t+\tau-t}$  和  $E^t$  的第  $i$  个列向量, 表示空间位置  $i$  沿通道方向的特征,  $i = 1, \dots, mn$ 。由式(5)可知, 特征的不同空间位置对应的权重不同。权重表示近邻帧特征补偿的当前帧特征与当前帧特征的相似性, 余弦相似度越大, 权重越大。即在空间位置  $i$  处, 如果  $e_i^{t+\tau-t}$  接近  $e_i^t$ , 那么给予近邻帧特征补偿的当前帧特征较大的权重, 否则, 给予其较小的权重。对式(5)中的权重进行归一化处理, 具体为

$$\tilde{w}_i^{t+\tau-t} = \frac{\exp(w_i^{t+\tau-t})}{\sum_{\tau=-T}^T \exp(w_i^{t+\tau-t})} \quad (6)$$

式中,  $T$  表示待融合的近邻帧与当前帧的最大间距。

将近邻帧特征补偿的当前帧特征  $F^{t+\tau-t}$  也表示为 2 维矩阵形式, 即  $F^{t+\tau-t} = [f_1^{t+\tau-t}; \dots; f_{mn}^{t+\tau-t}] \in \mathbf{R}^{c \times mn}$ , 利用上述计算的归一化权重  $\tilde{w}_i^{t+\tau-t}$  对多个近邻帧补偿的当前帧特征进行加权平均, 可表示为

$$\hat{f}_i^t = \sum_{\tau=-T}^T \tilde{w}_i^{t+\tau-t} f_i^{t+\tau-t} \quad (7)$$

式中,  $\hat{F}^t = [\hat{f}_1^t; \dots; \hat{f}_{mn}^t] \in \mathbf{R}^{c \times mn}$  为时间特征融合的当前帧特征。

图 5 给出了基于光流的时间特征融合模块, 在 res4 特征上将多帧图像特征进行融合。首先使用 ResNet101 提取当前帧和近邻帧的 res4 特征, 并利用光流网络估计当前帧与近邻帧之间的光流场, 根

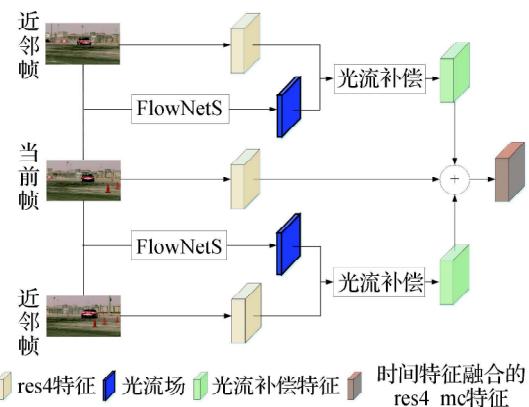


图 5 基于光流的时间特征融合模块

Fig. 5 Temporal feature fusion module based on optical flow

据光流场利用多个近邻帧的 res4 特征补偿当前帧的 res4 特征, 然后利用多个光流补偿特征对当前帧的 res4 特征进行线性预测, 形成时间特征融合的 res4\_mc 特征。

## 2.4 整体模型

图 6 为本文结合前述内容构建的视频目标检测整体模型。该模型首先使用 ResNet101 提取当前帧和近邻帧的 res4 特征, 根据当前帧与近邻帧之间的光流场对 res4 特征进行时间特征融合形成 res4\_mc 特征。然后在 res4\_mc 特征的基础上进行卷积操作, 提取光流补偿后的高层特征, 即 res5\_mc、conv6\_mc、conv7\_mc、conv8\_mc、conv9\_mc。随后利用这些特征以及当前帧的 res3 和 res4\_mc 特征构成特征金字塔, 并对其进行空间特征融合, 形成 P0、P1、P2、P3、P4、P5、P6 特征。最后利用检测网络进行检测, 并通过非极大值抑制得到检测结果。

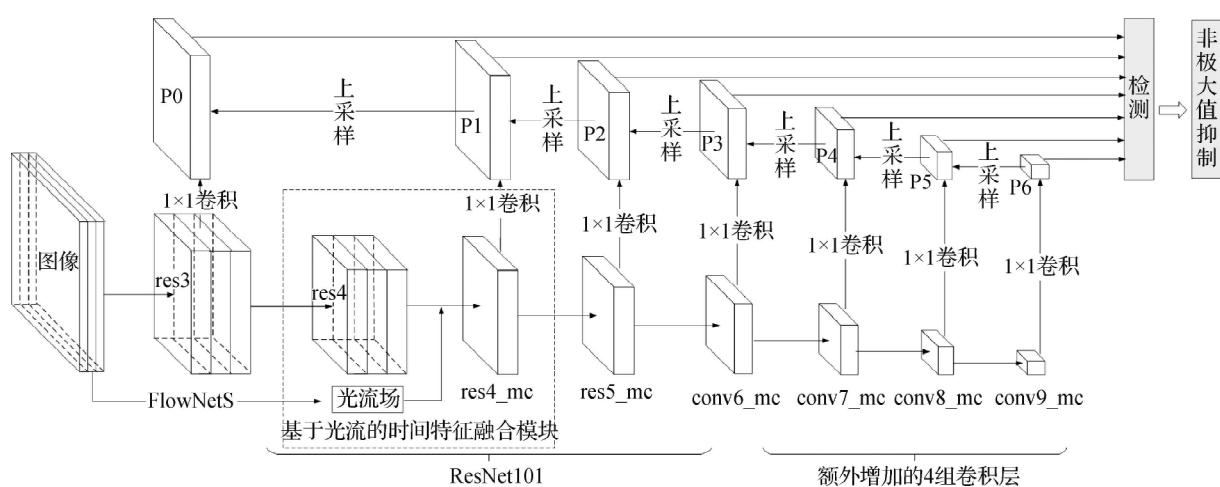


图 6 本文模型

Fig. 6 Proposed model

在模型训练阶段, 从间距  $\tau \leq T$  的帧中随机选取两帧图像作为近邻帧。在输出检测结果后, 根据损失函数利用反向传播方法更新参数, 在指定迭代次数的训练后保存模型参数。在模型测试阶段, 计算当前帧的  $2T + 1$  个近邻帧与当前帧之间的光流场, 从而利用多帧图像进行时间特征融合。

### 3 实验结果与分析

为了验证本文模型的性能, 使用 ImageNet VID (ImageNet for video object detection) 和 ImageNet DET (ImageNet for object detection) 数据集进行两阶段的模型训练。在 ImageNet VID 数据集的验证集上进行测试。本文模型使用 Python 语言开发, 基于深度学习框架 MXNet 实现。实验均在单频 2.10 GHz 的 ES-2620 v4CPU、64 GB 内存、NVIDIA Geforce RTX 2080Ti 显卡的服务器上进行。

#### 3.1 ImageNet VID 数据集的实验

本文利用 ImageNet VID 和 ImageNet DET 数据集的训练集共同训练模型。ImageNet VID 数据集是视频目标检测数据集, 训练集有 3 862 个视频片段, 验证集有 555 个视频片段, 每个视频片段的帧率是 25 帧/s 或 30 帧/s。视频中的每一帧图像都有标注, 整个数据集标注了 30 个目标类别。ImageNet DET 数据集是图像目标检测数据集, 其训练集包含 456 567 幅图像和 200 个类别。而 ImageNet VID 数据集中的类别是 ImageNet DET 数据集中的类别的子集, 因此使用 ImageNet DET 数据集中与 ImageNet VID 数据集的类别相对应的图像进行训练。

在模型训练前, 对训练集中的图像进行预处理, 通过双线性插值方法将图像扩展为  $512 \times 512$  像素, 使其符合模型的输入。模型的训练均采用动量梯度下降法, 动量系数设置为 0.9。整个训练过程分为两个阶段。第 1 阶段使用 ImageNet DET 和 ImageNet VID 数据集中两个训练集共同训练基于 ResNet101 和特征金字塔的 SSD 模型, 设定每批的数据量为 16 幅图像, 共迭代 5 个 epoch, 即使用这两个训练数据集中的全部训练样本对模型训练 5 次。其中, 初始学习率为  $1.0 \times 10^{-3}$ , 迭代 2 个 epoch 后学习率降为原来的  $1/10$ , 在后面 3 个 epoch 中, 每迭代 1 个 epoch 学习率降为原来的  $1/10$ , 训练终止后保存模型参数; 第 2 阶段使用 ImageNet VID 数据集的训练

集训练整体模型, 并使用第 1 阶段训练的模型参数作为初始权重, 设定每批的数据量为 8 幅图像。对于每幅图像, 从间距  $\tau \leq T$  的帧中随机选取两帧图像作为近邻帧用于训练, 共迭代 2 个 epoch, 其中, 初始学习率为  $7.5 \times 10^{-6}$ , 当模型迭代 1.333 个 epoch 后, 学习率降为  $7.5 \times 10^{-7}$ , 直至训练终止, 保存模型参数。在测试过程中, 使用 ImageNet VID 数据集的验证集进行测试, 对每幅图像融合  $2T + 1$  个近邻帧的特征来增强当前帧的特征。Zhu 等人(2017)的实验表明,  $T$  越大, 融合的近邻帧越多, 检测准确率越高, 但是随着  $T$  的增大, 检测准确率的提升不再明显, 并且时间开销也会增加, 因此本文设置待融合的近邻帧与当前帧的最大间距  $T = 10$ 。

本文采用平均准确率的均值 (mean average precision, mAP) 作为评价指标来分析本文提出的模型。平均准确率 (average precision, AP) 的计算以查全率和查准率为基础, 当目标框的置信度大于阈值时, 检测为正样本, 否则为负样本。查全率  $r$  为正确检测的正样本数占实际正样本总数的比例, 定义为

$$r = \frac{TP}{TP + FN} \quad (8)$$

式中,  $TP$  表示正确检测的正样本数 (true positive, TP),  $FN$  表示实际为正样本但检测为负样本的数量 (false negative, FN),  $TP$  与  $FN$  之和表示实际正样本的总数。

查准率  $p$  为正确检测的正样本数占检测结果中正样本总数的比例, 定义为

$$p = \frac{TP}{TP + FP} \quad (9)$$

式中,  $FP$  表示实际为负样本但检测为正样本的数量 (false positive, FP),  $TP$  与  $FP$  之和表示检测为正样本的总数。

设从样本总数  $n$  中检测  $k$  个样本, 对应地, 查全率表示为  $r_k$ ,  $p_k$  表示查全率大于  $r_k$  对应的最大查准率。平均准确率定义为

$$AP = \sum_{k=1}^n p_k(r_{k+1} - r_k) \quad (10)$$

mAP 为所有类别的平均准确率的均值, 定义为

$$mAP = \frac{1}{Q} \sum_{q=1}^Q AP_q \quad (11)$$

式中,  $Q$  表示类别总数,  $AP_q$  表示类别  $q$  的平均准确率。

在实验中, 将本文模型与 TCN (Kang 等 2016)、TPN + LSTM (Kang 等 2017) 和 Zhao 等人 (2018)

的模型进行比较,比较模型的结果均由模型作者提供。表1列出了在ImageNet VID验证集上3个比较

表1 不同模型在ImageNet VID数据集上对应的各类别AP值和mAP值比较

Table 1 Comparison of AP and mAP values among different models on ImageNet VID dataset

目标	模型			
	TCN	TPN + LSTM	SSD + 李生网络	本文
airplane	72.7	84.6	81.2	<b>85.7</b>
antelope	75.5	78.1	73.5	81.9
bear	42.2	72.0	70.2	75.6
bicycle	39.5	67.2	<b>72.3</b>	65.2
bird	25.0	68.0	71.5	73.6
bus	64.1	<b>80.1</b>	78.6	67.6
car	36.3	54.7	50.1	<b>64</b>
cattle	51.1	61.2	<b>65.3</b>	57.5
dog	24.4	61.6	60.4	<b>63.9</b>
domestic cat	48.6	<b>78.9</b>	70.1	78.8
elephant	65.6	71.6	<b>82.6</b>	69.4
fox	73.9	83.2	85.9	<b>88.6</b>
giant panda	61.7	78.1	<b>81.2</b>	79.8
hamster	82.4	<b>91.5</b>	87.5	91.4
horse	30.8	66.8	75.2	<b>77.9</b>
lion	34.4	21.6	47.8	<b>54.3</b>
lizard	54.2	74.4	71.5	<b>78.5</b>
monkey	1.6	36.6	<b>50.3</b>	47.1
motorbike	61	76.3	72.5	<b>81.3</b>
rabbit	36.6	51.4	<b>61.8</b>	45.4
red panda	19.7	70.6	71.9	<b>76.5</b>
sheep	55.0	<b>64.2</b>	40.0	56.7
snake	38.9	61.2	62.3	<b>77.3</b>
squirrel	2.6	42.3	<b>85.1</b>	49.4
tiger	42.8	84.8	78.1	<b>90.5</b>
train	54.6	78.1	71.2	<b>85.7</b>
turtle	66.1	<b>77.2</b>	74.5	76.5
watercraft	<b>69.2</b>	61.5	65.6	63.8
whale	26.5	66.9	51.8	<b>68.1</b>
zebra	68.6	<b>88.5</b>	75.2	86.5
mAP	47.5	68.4	69.5	<b>72.0</b>

注:加粗字体为每行最优结果。

模型与本文模型在各类目标上的AP和mAP值,目标包括飞机、羚羊、熊等30个类别。TCN和TPN+LSTM模型均在两阶段目标检测的基础上,分别结合目标跟踪和LSTM构建视频目标检测模型,步骤烦琐,时间开销大。而单阶段目标检测方法检测速度快,本文模型和SSD+李生网络(Zhao等人,2018)的模型均建立在单阶段目标检测中SSD框架的基础上,SSD+李生网络模型的单帧图像平均检测时间为31 ms,本文模型的单帧图像平均检测时间为95 ms,但是在检测准确率方面,本文模型比SSD+李生网络的模型提高了2.5%。

### 3.2 分离实验

在本文模型中,ResNet101利用残差结构加深网络模型,使用深层网络模型描述数据。而特征金字塔网络和光流网络分别通过空间特征融合和时间特征融合来增强特征。为了验证这3个模块对检测性能的影响,在SSD目标检测框架下,在ImageNet VID数据集上分别对基于VGG16、ResNet101、ResNet101与特征金字塔(ResNet101+FPN)、ResNet101与光流网络(ResNet101+FlowNetS)以及ResNet101与时空融合(ResNet101+FPN+FlowNetS)等5种结构的网络模型进行实验,实验结果如表2所示。表2列出了5种不同结构的网络模型在ImageNet VID验证集上对应的类别的AP以及mAP值。可以看出,VGG16模型的mAP明显小于ResNet101模型的mAP,表明深层网络模型的使用可以提高检测准确率。与ResNet101模型相比,ResNet101+FPN、ResNet101+FlowNetS、ResNet101+FPN+FlowNetS模型的mAP从65.0%分别提升到70.8%、65.9%和72.0%,表明时空特征融合通过结合近邻帧的特征以及多尺度的特征能够增强当前帧的特征,提升检测准确率。

为了更好地分析模型的性能,根据Zhu等人(2017)的研究,在目标的运动速度和尺寸两方面分别对不同结构的网络模型进行评测,将目标分为慢速、中速和快速目标,以及小型、中型和大型尺寸目标。运动速度依据当前目标与其近邻帧( $\pm 10$ 帧)中对应目标的交并比(intersection-over-union,IoU)平均值划分,IoU平均值越小,目标运动越快,反之亦然。IoU平均值大于0.9的目标属于慢速目标,小于0.7的目标属于快速目标,在0.7~0.9之间的目标属于中速目标。目标的尺寸依据目标的像素

划分, 小于  $50 \times 50$  像素的目标属于小型尺寸目标, 大于  $150 \times 150$  像素的目标属于大型尺寸目标, 在  $50 \times 50 \sim 150 \times 150$  像素之间的目标属于中型尺寸目标。

表 2 不同结构网络模型在 ImageNet VID 数据集上对应的各类别 AP 和 mAP 值比较

Table 2 Comparison of AP and mAP values among different network structures on ImageNet VID dataset

目标	模型					/%
	VGG16	ResNet101	ResNet101 + FPN	ResNet101 + FlowNetS	ResNet101 + FPN + FlowNetS	
airplane	84.4	81.9	<b>87.8</b>	81.4	85.7	
antelope	77.3	78.2	<b>82.2</b>	77.1	81.9	
bear	65.1	70.6	72.2	73.5	<b>75.6</b>	
bicycle	62.3	54.7	64.7	55.8	<b>65.2</b>	
bird	63.4	68.6	72.4	69.5	<b>73.6</b>	
bus	66.9	69.2	66.1	<b>71.3</b>	67.6	
car	63.2	58.5	<b>65.5</b>	56.4	64	
cattle	52.2	46.3	53.8	48.9	<b>57.5</b>	
dog	48.7	55.3	61.7	55.6	<b>63.9</b>	
domestic cat	47.6	73.4	75.7	78.7	<b>78.8</b>	
elephant	<b>71.3</b>	66.9	70.4	65.6	69.4	
fox	72.2	81.7	86.9	83.4	<b>88.6</b>	
giant panda	80.3	82.0	<b>84.1</b>	79.1	79.8	
hamster	65.7	89.3	90.1	<b>91.5</b>	91.4	
horse	69.7	63.2	76.1	63.2	<b>77.9</b>	
lion	7.5	24.4	47.0	27.5	<b>54.3</b>	
lizard	60.1	76.4	75.8	<b>80.2</b>	78.5	
monkey	26.5	39.2	<b>48.1</b>	39.2	47.1	
motorbike	79.9	74.9	79.3	75.8	<b>81.3</b>	
rabbit	39.1	<b>47.8</b>	46.3	45.6	45.4	
red panda	45.3	56.5	69.6	62.2	<b>76.5</b>	
sheep	52.7	51.7	<b>58.0</b>	49.5	56.7	
snake	41.8	67.7	72.5	68.7	<b>77.3</b>	
squirrel	29.2	45.7	48.6	46.5	<b>49.4</b>	
tiger	81.3	88.6	90.4	87.9	<b>90.5</b>	
train	77.7	78.2	84.4	79.6	<b>85.7</b>	
turtle	70.2	77.3	76.4	<b>77.9</b>	76.5	
watercraft	62.0	51.9	63.5	51.8	<b>63.8</b>	
whale	59.3	51.0	66.9	52.0	<b>68.1</b>	
zebra	84.6	79.9	86.1	80.3	<b>86.5</b>	
mAP	60.2	65.0	70.8	65.9	<b>72.0</b>	

注: 加粗字体为每行最优结果。

表3列出了不同结构网络模型在不同运动速度目标上的mAP。其中,慢速mAP、中速mAP、快速mAP分别表示在慢速、中速、快速目标上的mAP。从表3可以看出,在结合光流网络之后,与ResNet101模型相比,ResNet101+FlowNetS模型的慢速、中速、快速mAP分别提高了0.8%、1.6%、1.2%;与ResNet101+FPN模型相比,本文整体模型的慢速、中速、快速mAP分别提高了0.6%、1.9%、2.3%表明光流补偿有效提高了视频目标检测的准确率,尤其能够更好地提升中速和快速目标的检测能力。

表3 不同结构网络模型在不同运动速度目标上的mAP比较

Table 3 Comparison of mAP values among different network structures for objects of different speed

模型	目标运动速度 /%		
	慢速	中速	快速
VGG16	70.9	58.0	36.3
ResNet101	75.7	64.5	40.7
ResNet101 + FPN	80.1	69.0	47.3
ResNet101 + FlowNetS	76.5	66.1	41.9
ResNet101 + FPN + FlowNetS	<b>80.7</b>	<b>70.9</b>	<b>49.6</b>

注:加粗字体为每列最优结果。

表4列出了不同结构网络模型在小、中、大型尺寸目标上的mAP。其中,小型mAP、中型mAP、大型mAP分别表示在小尺寸、中尺寸、大尺寸目标上的mAP。

从表4可以看出:1)结合特征金字塔后,与ResNet101和ResNet101+FlowNetS模型相比,ResNet101+FPN和ResNet101+FPN+FlowNetS模型的mAP在小尺寸目标上分别提高了9.0%和8.0%,在中尺寸目标上分别提高了9.8%和10.6%,在大尺寸目标上分别提高了2.2%和2.4%表明特征金字塔可以较好地提高模型对中小尺寸目标的检测能力;2)结合光流网络后,与ResNet101和ResNet101+FPN模型相比,ResNet101+FlowNetS和ResNet101+FPN+FlowNetS模型的mAP在小尺寸目标上分别提高了0.4%和降低了0.4%,在中尺寸目标上分别提高了2.1%和2.9%,在大尺寸目标上分别提高了0.7%和1.2%表明光

流补偿能更有效地提高大、中尺寸目标的检测准确率。

表4 不同结构网络模型在不同尺寸目标上的mAP比较

Table 4 Comparison of mAP values among different network structures for objects of different size

模型	目标尺寸 /%		
	小型	中型	大型
VGG16	15.4	36.4	72.5
ResNet101	11.6	36.2	80.1
ResNet101 + FPN	<b>20.6</b>	46.0	82.3
ResNet101 + FlowNetS	12.2	38.3	80.8
ResNet101 + FPN + FlowNetS	20.2	<b>48.9</b>	<b>83.2</b>

注:加粗字体为每列最优结果。

本文模型在SSD目标检测框架下,通过时空特征融合提高了视频目标检测的准确率。为了从视觉效果上观察特征金字塔和光流网络的有效性,图7—图10分别显示了在ImageNet VID验证集中编号为7016、19000、9000和10000的视频上的检测结果。

图7和图8给出了使用特征金字塔前后的检测结果比较,图7(a)和图8(a)为ResNet101+FlowNetS模型的检测结果,图7(b)和图8(b)为ResNet101+FPN+FlowNetS模型的检测结果。

从图7和图8可以看出,对于图7和图8右上角的小尺寸目标以及图8中目标运动造成的尺寸变化,ResNet101+FlowNetS模型出现了漏检问题,而ResNet101+FPN+FlowNetS模型准确检测出目标。由此可见,结合特征金字塔的整体模型能够提高小尺寸目标的检测能力,同时较好地解决目标尺寸变化问题。

图9和图10给出了使用光流网络前后的检测结果比较。图9(a)和图10(a)为ResNet101+FPN模型的检测结果,图9(b)和图10(b)为ResNet101+FPN+FlowNetS模型的检测结果。从图9可以看出,图像中的运动模糊导致检测出现目标漏检以及误检问题,结合光流补偿可以准确检测出正确目标并抑制错误目标;从图10可以看出,场景中存在目标遮挡情况,利用光流场通过近邻帧的特征对当前帧的特征进行补偿,从而检测出遮挡目标。

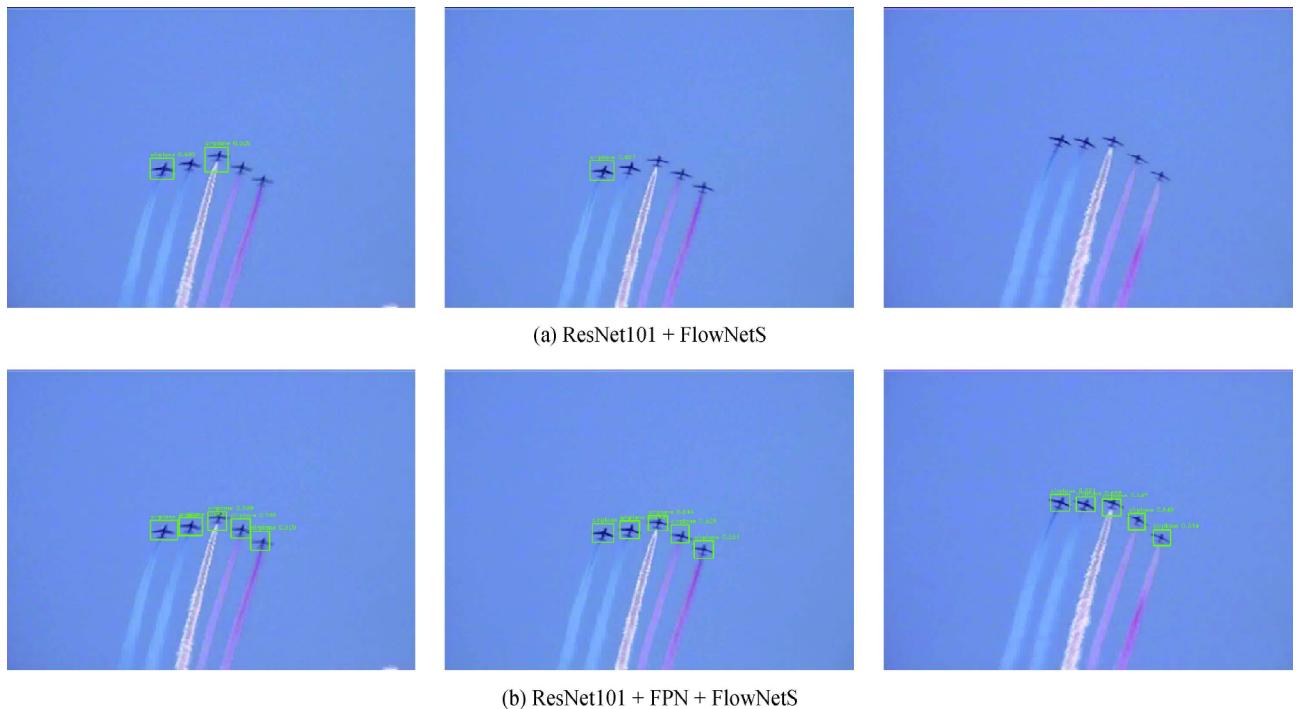


图7 不同结构网络模型在ImageNet VID验证集中编号为7016视频上的结果比较

Fig. 7 Comparison results of different network structures on the video #7016 from ImageNet VID validation set

((a) ResNet101 + FlowNetS; (b) ResNet101 + FPN + FlowNetS)

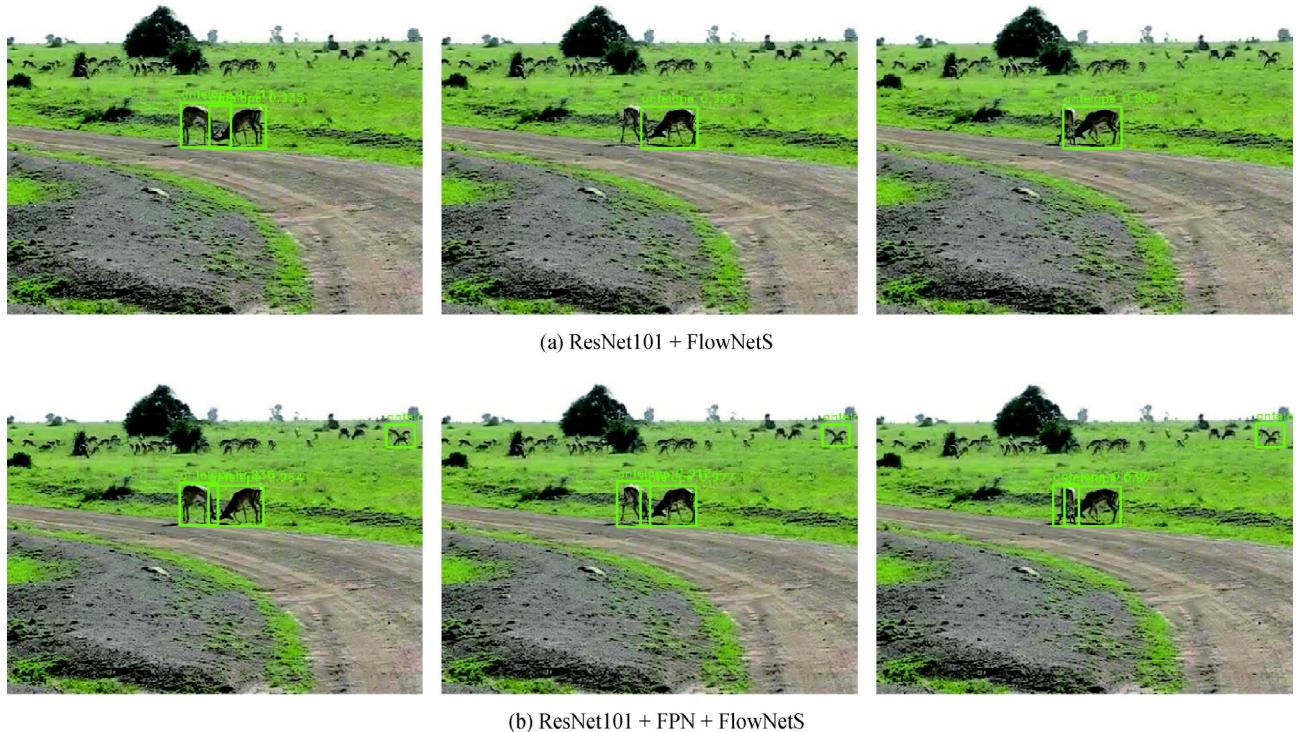


图8 不同结构网络模型在ImageNet VID验证集中编号为19000视频上的结果比较

Fig. 8 Comparison results of different network structures on the video #19000 from ImageNet VID validation set

((a) ResNet101 + FlowNetS; (b) ResNet101 + FPN + FlowNetS)

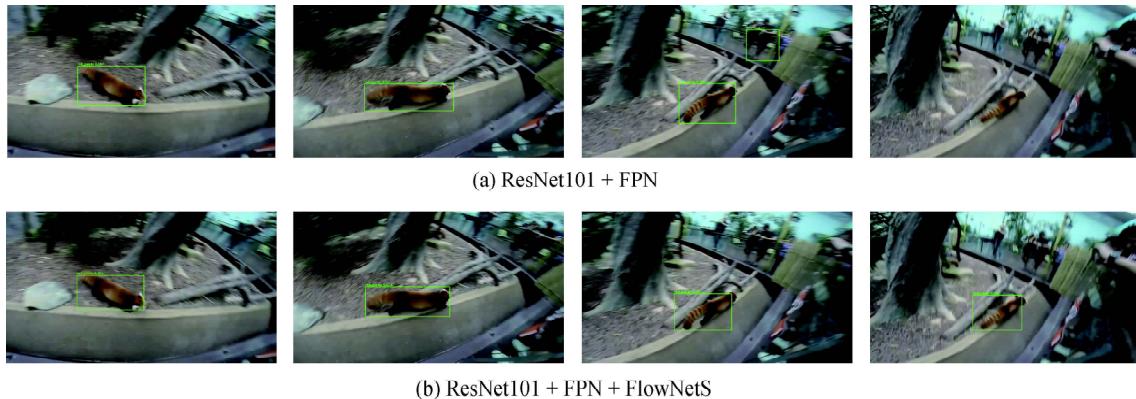


图9 不同结构网络模型在ImageNet VID验证集中编号为9000视频上的结果比较

Fig. 9 Comparison results of different network structures on the video #9000 from ImageNet VID validation set

( ( a) ResNet101 + FPN; ( b) ResNet101 + FPN + FlowNetS)

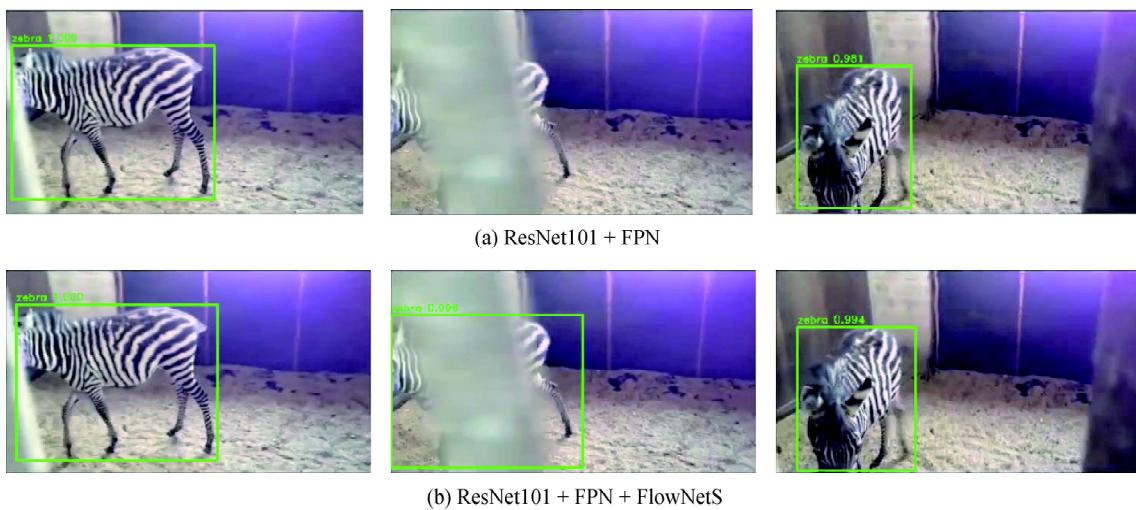


图10 不同结构网络模型在ImageNet VID验证集中编号为10000视频上的结果比较

Fig. 10 Comparison results of different network structures on the video #10000 from ImageNet VID validation set

( ( a) ResNet101 + FPN; ( b) ResNet101 + FPN + FlowNetS)

视频目标检测的影响。

本文方法在进行多帧图像特征融合时,依赖于光流网络估计的光流场,光流场的准确性决定了融合时间特征的准确性,后续研究工作将进一步专注于运动补偿和特征融合的方法。

### 参考文献( References)

- Dosovitskiy A ,Fischer P ,Ilg E ,Häusser P ,Hazirbas C ,Golkov V ,van der Smagt P ,Cremers D and Brox T. 2015. FlowNet: learning optical flow with convolutional networks//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago , Chile: IEEE: 2758-2766 [DOI: 10.1109/ICCV.2015.316]
- Girshick R ,Donahue J ,Darrell T and Malik J. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation//

## 4 结 论

本文提出了一种 SSD 与时空特征融合的视频目标检测模型。在单阶段目标检测中 SSD 的框架下,结合光流网络计算当前帧与近邻帧之间的光流场并融合多帧图像特征以补偿当前帧的特征,最后将高低层特征相融合,利用融合后的特征进行多尺度特征检测,从而增强低层特征对小尺寸目标的检测能力。在 ImageNet VID 数据集上的实验结果表明,本文提出的视频目标检测模型的 mAP 达到 72.0%,优于其他模型,通过分离实验进一步验证了本文模型的有效性,本文模型提高了小尺寸目标的检测能力,减弱了目标模糊和多目标遮挡等因素对

- Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE: 580-587 [DOI: 10.1109/CVPR.2014.81]
- Girshick R. 2015. Fast R-CNN // Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 1440-1448 [DOI: 10.1109/ICCV.2015.169]
- He K M, Zhang X Y, Ren S Q and Sun J. 2015. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9): 1904-1916 [DOI: 10.1109/TPAMI.2015.2389824]
- He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition // Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]
- Kang K, Li H S, Xiao T, Ouyang W L, Yan J J, Liu X G and Wang X G. 2017. Object detection in videos with tubelet proposal networks // Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 889-897 [DOI: 10.1109/CVPR.2017.101]
- Kang K, Ouyang W L, Li H S and Wang X G. 2016. Object detection from video tubelets with convolutional neural networks // Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 817-825 [DOI: 10.1109/CVPR.2016.95]
- Krizhevsky A, Sutskever I and Hinton G E. 2012. ImageNet classification with deep convolutional neural networks // Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe, USA: NIPS: 1097-1105
- Lin T Y, Dollár P, Girshick R, He K M, Hariharan B and Belongie S. 2017. Feature pyramid networks for object detection // Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 936-944 [DOI: 10.1109/CVPR.2017.106]
- Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C Y and Berg A C. 2016. SSD: single shot multibox detector // Proceedings of the 14th European Conference on Computer Vision. Amsterdam, The Netherlands: Springer: 21-37 [DOI: 10.1007/978-3-319-46448-0\_2]
- Redmon J, Divvala S, Girshick R and Farhadi A. 2016. You only look once: unified, real-time object detection // Proceedings of 2016 IEEE Conference on Computer Vision and Pattern. Las Vegas, USA: IEEE: 779-788 [DOI: 10.1109/CVPR.2016.91]
- Ren S Q, He K M, Girshick R and Sun J. 2017. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6): 1137-1149 [DOI: 10.1109/TPAMI.2016.2577031]
- Uijlings J R R, van de Sande K E A, Gevers T and Smeulders A W M. 2013. Selective search for object recognition. *International Journal of Computer Vision*, 104(2): 154-171 [DOI: 10.1007/s11263-013-0620-5]
- Xiao F Y and Lee Y J. 2018. Video object detection with an aligned spatial-temporal memory // Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 494-510 [DOI: 10.1007/978-3-030-01237-3\_30]
- Zhang S F, Wen L Y, Bian X, Lei Z and Li S Z. 2018a. Single-shot refinement neural network for object detection // Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 4203-4212 [DOI: 10.1109/CVPR.2018.00442]
- Zhang Z S, Qiao S Y, Xie C H, Shen W, Wang B and Yuille A L. 2018b. Single-shot object detection with enriched semantics // Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 5813-5821 [DOI: 10.1109/CVPR.2018.00609]
- Zhao B J, Zhao B Y, Tang L B, Han Y Q and Wang W Z. 2018. Deep spatial-temporal joint feature representation for video object detection. *Sensors*, 18(3): #774 [DOI: 10.3390/s18030774]
- Zhu X Z, Wang Y J, Dai J F, Yuan L and Wei X C. 2017. Flow-guided feature aggregation for video object detection // Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 408-417 [DOI: 10.1109/ICCV.2017.52]

## 作者简介



尉婉青,1995年生,女,硕士研究生,主要研究方向为计算机视觉和深度学习。

E-mail: 1454072136@qq.com



肖创柏 通信作者 男 教授 博士生导师,主要研究方向为数字信号处理、音视频信号处理和网络通信。

E-mail: cbxiao@bjut.edu.cn

禹晶,女 副教授 硕士生导师,主要研究方向为图像逆处理、稀疏表示和深度学习。

E-mail: jing.yu@bjut.edu.cn

柏曼晏,女 硕士研究生,主要研究方向为计算机视觉和深度学习。E-mail: baimanyan@163.com