DOI: 10.13382/j.jemi.B2104235

基于时空一致性约束视频目标车辆的检测与跟踪算法研究*

洪 锋 鲁昌华 蒋薇薇 王 涛 方恒阳
(合肥工业大学计算机与信息学院 合肥 230011)

摘 要:复杂场景中的目标感知是深度学习在计算机视觉中最重要的研究领域之一,而复杂交通场景中的车辆检测与跟踪是当今众多学者研究的热点问题。在视频目标检测过程中由于运动物体的时间维度特征信息利用不充分,导致在长序列之间的时间特征极其容易被忽略,本文提出一种时空一致性的视频车辆的检测跟踪算法。该算法由双分支网络结构组成:分支一是由基于空间相关性的 Transformer 网络模块组成,该分支网络主要用于判断前后帧的相关性、感知相邻帧之间的一致性,预测目标车辆时空一致性的关联度;另一网络分支是由基于交叉特征金字塔融合的网络模块组成,该模块主要是提取检测对象的局部信息结合浅层的空间边缘信息和深层的语义特征信息,提取对象空间位置的特征信息。该网络结构将 Transformer 机制和交叉特征金字塔模块相结合,利用 Transformer 对长序列之间时间关联性敏感和特征金字塔网络模块对边缘信息敏感的特性,对视频帧对象进行检测和跟踪,确保相邻帧的长程相关性以及边缘和深层的特征信息深度融合。实验结果表明,本文设计的双分支网络结构在视频目标跟踪和检测中取得更好精度和更快的收敛速度;同时在显著性视频目标检测中,实验表明算法的有效性和泛化性。

关键词: 时空一致性;车辆跟踪;Transformer;交叉特征金字塔网络。
中图分类号: TP391.4 **文献标识码**: A **国家标准学科分类代码**: 520.2050

Research on vehicle detection& tracking algorithm based on spatio-temporal consistent dual-stream network video target

Hong Feng Lu Changhua Jiang Weiwei Wang Tao Fang Hengyang
(School of Computer and Information, Hefei University of Technology, Hefei 230011, China)

Abstract: Target perception in complex scenes is one of the most important research fields of deep learning in computer vision, and vehicle detection in complex traffic scenes is the object of research by many scholars today. In the process of video target detection, due to the insufficient utilization of the time dimension feature information of moving objects, time features between long sequences are extremely easy to be ignored. This paper proposes a spatio-temporal consistent video vehicle detection and tracking algorithm. The algorithm is composed of a two-branch network structure: one of branch is composed of transformer network modules based on spatial correlation. The branch network is mainly used to determine the correlation between the previous and subsequent frames, perceive the consistency between adjacent frames, and predict the temporal and spatial consistency of the target vehicle relevance; another network branch is composed of network modules based on cross-feature pyramid fusion. This module mainly extracts the local information of the detected object combined with shallow spatial edge information and high-level semantic feature information. This branch extracts the spatial position of the object characteristic information. The network structure combines the Transformer mechanism and the cross-feature pyramid module, and uses the advantages of Transformer's sensitivity to the time correlation between long sequences and the feature pyramid network module's sensitivity to edge information to detect and track video frame objects to ensure neighboring the long-range correlation of the frame is deeply integrated with the feature information of the edge and the deep layer. The experimental results show that the dual-branch network structure designed in this paper achieves better accuracy and faster convergence speed in video target tracking and detection. At the same time, experiments in saliency video target detection show the effectiveness and generalization of the algorithm.

Keywords: spatio-temporal consistency; vehicle tracking; transformer; cross-feature pyramid network

收稿日期: 2021-04-29 Received Date: 2021-04-29
*基金项目:国家重大科技攻关项目(JZ2015KJZZ0254)、中科院 STS 重大项目(KFJ-STS-ZDTP-079)、安徽省优秀拔尖人才培育计划(gxyq2018110,gxyq2019111)、国家基金培育项目(CZ2021GP08)项目资助

0 引　言

在复杂场景中如自动驾驶和智能交通领域,对视频目标的感知是其开展研究的方向之一[1-4]。在视频目标对象的检测过程中,由于前后帧的遮挡、变形、体态姿势变化等因素的影响,较静止的目标检测挑战性更大。视频目标对象的感知和检测的方法主要是利用相邻帧的特征增强技术实现目标的匹配[5]。传统视频目标的感知、检测以及跟踪算法主要是利用卡尔曼滤波器[6]等进行实时检测+跟踪的方式进行目标定位检测和跟踪,传统方法没有考虑时间特征之间的关系,而且实时性较差。随着深度学习的兴起,利用卷积神经网络进行目标对象的感知和检测跟踪,在实时性和准确性上面得到很大的提升。Danelljan 等[7]提出在连续卷积的滤波器中插入多分辨率的深度特征图,将各层特征图进行融合,利用传统 HOG 的特征和深度特征融合,实现目标的检测跟踪;Wang 等[8]将卷积神经网络看成是一个集成学习的过程,通过一个集成学习器,得到热图,利用设计的损失函数,最终训练出来的网络具有很好的泛化能力,但是 CNN 训练的过程中必须离线训练,而且目标检测和跟踪都是利用损失函数来度量确定,存在很大的局限性;Bertinetto 等[9]提出的全卷积网络将图像通过两个卷积网络分别得到特征的降维映射,利用卷积计算目标兑现位置的相关性,通过插值原图实现目标的定位,利用孪生网络分别确定检测跟踪的目标,但是由于 CNN 无法实现端到端的实时性训练;为了实现快速地跟踪,He 等[10]提出的 SA-Siam 双分支网络,考虑到图像分类中的语义特征和相似性匹配中的学习外观特征的相关性,设计一个语义分支网络注入注意力机制,根据目标位置附近的通道激活情况计算通道权重,设计一个外观特征网络,两个通道在训练过程中保持相互的异质性,通过双重设计和注意力机制提高了目标跟踪的实时性和效果,但相对考虑浅层的优化过程中采用 SGD 对网络进行微调,没有充分利用 CNN 网络的端到端的优势,没有保证网络的实时性,随着卷积网络的加深和下采样等操作,空间特征的语义信息可能会有所丢失,导致目标跟踪的实效性低;Li 等[11]提出的多任务引导策略的网络结构,提高视频目标中的显著性目标的感知和检测,设计两个子网络分别对静止目标和运动目标的光流信息进行融合引导,在视频目标的检测中优于现有的方法。

视频目标检测主要是利用各帧之间的时空关系对帧中的目标进行检测。目标检测过程中难点是充分提取目标对象的特征,有效的特征融合是进行目标检测的关键因素之一。为将目标中的特征信息进行多尺度的融合,大量的研究者提取各种网络结构的模型,特征融合的方式有设计多分支的网络结构进行多尺度的特征融合[12],有进行特征金字塔式的特征融合策略[13]和加入注意力机制等策略提高检测和定位精度。Hu 等[14]提出一种时空分割的多维度特征融合模型,关注相邻像素之间的局部交互,提高增强的语义信息;Song 等[15]提出的多尺度特征融合的模块构建高级的语义特征,利用图像素级标签的全局语义信息的引导策略,快速的定位不同比例的对象,并通过注意力机制减少模型推理时间的基础上提高了检测的精度。Cao 等[16]提出了一种基于注意力引导的上下文特征金字塔网络,结合注意力机制和上下文引导模块,通过注入注意力机制捕获目标对象区域和进行定位,同时特征金字塔的模块融合目标对象的多尺度特征信息。

视频目标跟踪和检测的过程中主要方法有利用多分支的网络结构融合时空特征进行检测跟踪,或者利用对象的深度语义信息进行检测和跟踪,本质是在视频帧之间寻求类间的全局优化,对视频中各帧的目标对象进行检测和感知。鉴于以上描述,在视频目标的跟踪和检测中主要考虑两个问题:要求网络实时性高,在目标对象的特征一致性感知过程中,计算量不能过大,推理时间不能太长;其次,判断跟踪对象的位置和提取特征的过程,对象的空间位置特征和时间位置信息一致性的相关度被综合考虑。因此,本文主要解决的问题包含:1)采用 Transformer 网络块,定位跟踪车辆的空间位置信息,判断前后视频帧之间的相关性;2)利用交叉特征金字塔网络提取相邻视频帧之间的特征信息,提取对象的局部边缘的特征信息,融合浅层边缘空间信息和深层高语义信息。通过双分支的网络结构,将视频帧中的时间相关性和空间相关性进行融合,实现相邻帧之间的跟踪车辆的定位和特征提取,对目标对象进行精准感知和检测。

1 网络结构的设计

本文采用多分支网络的结构形式,分支一获取检测、跟踪车辆的时空信息,考虑确定车辆位置以及时间变化的过程中车辆的空间信息,该分支采用循环网络结构和 Transformer 机制。视频监控对象的本质是利用前一帧中对象的特征信息、位置信息、空间信息预测下一帧物体的位置信息,利用循环网络捕获前后帧之间车辆位置和空间信息相关性,通过注入 Transformer 实现自我学习上下文信息进行非全局交互建模,捕获对象空间和时间的全局上下文信息;另一个分支网络主要获取跟踪车辆的特征信息,采用交叉特征金字塔的结构,获取浅层的空间信息和深层的语义特征,能够捕获对象具有强语义和精确位置的强依赖关系的同时获得目标对象的特征信息。图 1 是利用本文提出的时间-空间一致性的双流网络进

行目标跟踪和感知的真实检测和跟踪结果示意图。

图 1　车辆跟踪检测结果示意图

Fig. 1　Schematic diagram of vehicle tracking detection results

1.1　基于 Transformer 的编码-解码器网络结构

循环神经网络是一种具有短时记忆能力的网络结构，主要应用在视频和自然语言等顺序的网络结构中。Yang 等[17]提出将 Transformer 结合循环网络，在视频中判断跟踪车辆的空间和时间信息的相关性效果显著。早期的跟踪检测主要是将跟踪和检测结合在一起，该分支网络主要是对视频对象进行跟踪，结合时间-空间信息的一致性，注入非局部的上下文信息，利用注意力机制确定目标。注意力机制是根据人眼在观察目标的工作原理，自动注意到目标对象的某些关键信息上，不用进行逐点扫描，但是注意力机制仍然存在信息丢失、实时性差以及不能并行训练和预测等问题。

为提升注意力机制在目标检测中的性能，Vaswani 等[18]提出的网络结构中完全由注意力机制的网络组成，形成编码-解码器的结构，可以进行并行的网络训练，提出的 Transformer 网络结构随后在各种顺序数列中得到广泛的应用；Bilkhu 等[19]提出的网络结构将 Transformer 引入到捕获视频目标的信息中，但是没有考虑前后帧之间检测目标的特征融合。

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{1}$$

式中：$\boldsymbol{Q}$ 是注意力函数映射的查询矩阵，$\boldsymbol{K}$，$\boldsymbol{V}$ 分别表示关键字和向量值的矩阵，d_k 是输入的查询和关键字的维度，最后输出的之表示加入的注意力机制之后的输出，式(1)本质是将注意力机制转换为概率之后输出，命名为转换器 Transformer[20]。Transformer 的内部工作原理由图 2、3 可以获悉[21]；图 2 是 Transformer 编码阶段其各层之间的工作原理；图 3 是 Transformer 解码阶段各层之间的工作原理。

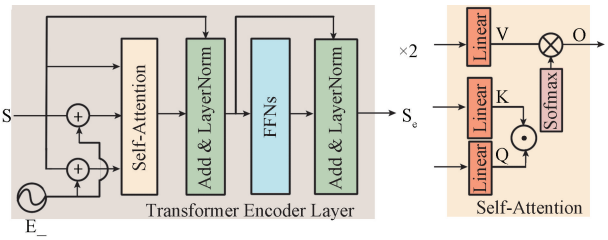


图 2　Transformer 的编码器结构图

Fig. 2　Encoder structure diagram of Transformer

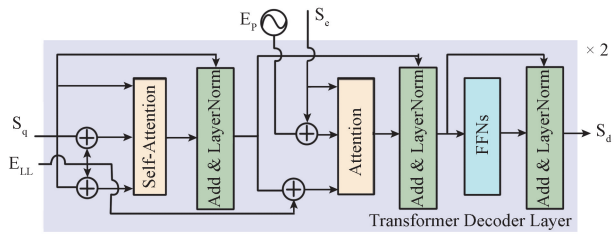


图 3　Transformer 的解码器结构图

Fig. 3　Decoder structure diagram of Transformer

本文在 Transformer 编码阶段通过嵌入空间信息和令牌信息的循环结构对视频帧进行长程之间的关联信息融合，然后对其维度进行扩展之后拆分为 3 个矩阵信息分别为($\boldsymbol{Q}$,$\boldsymbol{K}$,$\boldsymbol{V}$)，利用矩阵的互乘和激活函数得到相关帧之间长程联系，将矩阵($\boldsymbol{Q}$,$\boldsymbol{K}^{\mathrm{T}}$)互乘结果经过 Softmax 函数之后得到概率值；然后将概率值和矩阵 $\boldsymbol{V}$ 做点积操作，结果为多头注意力机制的输出值。

Transformer 解码阶段的主要工作原理与编码阶段的工作原理相似，结合循环网络和残差网络的思想。在预处理之前参考循环网络的设计思想，对编码处理后端利用残差网络的思想，对多头注意力机制之后进行线性归一化之后结合开始的令牌和原始序列信息。

本文提出的基于时空一致性约束的网络结构，该网络主要负责对视频帧中车辆进行跟踪匹配，其中对车辆信息的检测由交叉的特征金字塔网络分支。视频目标在检测和跟踪过程中，由于检测和推理过程中占据网络大量的资源，引入 Transformer 的机制，利用 Vision transformers[22]将视频检测中的时间信息进行分解，对提取对象的空间特征信息进行引导，节约了网络本身的推理和检测时间。

1.2　基于交叉金字塔的网络结构

金字塔的网络结构[23]在特征提取和融合方面表现出巨大的优势。设置不同尺度的金字塔提取特征，在多尺度的目标检测过程中效果显著。本文在金字塔的网络结构基础上提出交叉金字塔的网络结构，充分利用浅层金字塔能提取丰富的空间特征和深层金字塔提取的语义

特征相结合,避免不同视频帧之间由于目标对象的尺度发生改变而发生漏检。在特征提取的骨干网络中利用[P_3,P_4,P_5]层,图4是本文提出的交叉特征金字塔的网络结构。每一层分别对应预测目标位置和前景、背景的确定,设定阈值;之后经过一个提炼阶段,对预测的目标进行排序提炼;P_3层的预测结果引导P_4层进行预测目标的位置和目标边界确定的阈值设定;最终深层的提炼的结果反馈到浅层的骨干网络,如P_5引导→P_4预测目标的位置以及前景和背景的确定;最终,提炼之后的各层特征进行融合,通过该模块可以提高多尺度目标检测精度和准确性,针对视频中的目标取得了良好的检测效果。

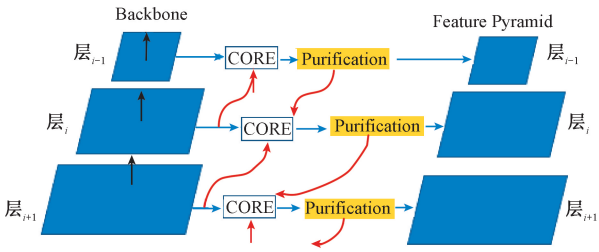


图4　交叉金字塔的网络结构

Fig. 4　Network structure of the cross pyramid

定检测对象的概率值。为了混合不同Channel不同位置的空间特征以及不同空间位置同一Channel的特征,文中利用Mix-MLP操作,既可以混合给定空间位置的特征,也可以混合不同空间位置的特征。

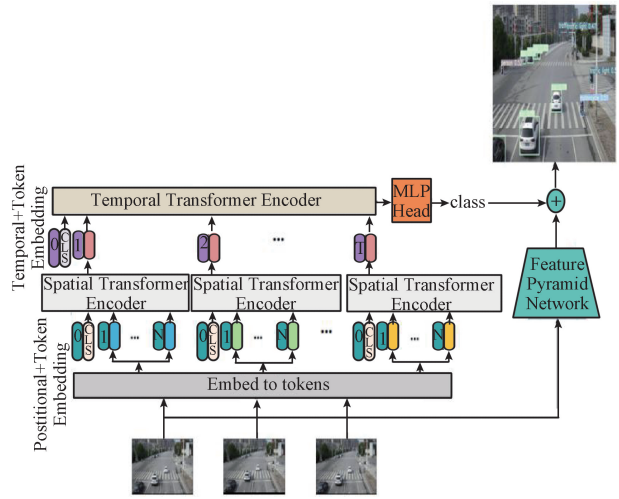


图5　基于时空-一致性约束的多分支网络结构

Fig. 5　Multi-branch network structure based on time-space-consistency constraints

1.3　网络结构及其损失函数

1)网络结构

本文的网络结构如图5所示,利用双分支的网络结构,分别利用视频帧之间的长程相关性和每帧之间的特征信息相结合的网络结构,进行视频车辆的跟踪和检测。基于Transformer的网络结构主要检测检测车辆的位置和空间信息,虽然其对长程相关性建模中主要通过多头注意力机制建模,它的输入是序列信号,在局部区域内缺少对图像局部信息提取的能力,而局部机制对于图像的检测来说必不可少,因为它与线条、边缘以及物体的结构等有关;交叉特征金字塔的网络结构主要负责前后帧中的特征信息的提取,通过金字塔的特征提取模块,提取相邻帧中的目标对象的特征,保证提取对象的局部信息的完整性;提取特征之后将时空一致性Transformer模块的长程相关性结合,对视频目标中的对象进行感知和检测,最后将空间的位置信息和空间的语义信息融合之后进行目标的跟踪检测。

本文提出的网络结构中,如图5中所示空间Transformer编码模块(spatial transformer encoder,STE)主要对各帧的空间位置信息进行编码;时间Transformer编码模块(temporal transformer encoder,TTE)主要是编码关联帧之间的信息;多层感知器(multilayer perceptron,MLP)模块主要功能是对TTE模块中被检测对象进行检测,在MLP中主要是点乘、全连接以及经过Softmax函数操作,检测对象进行全连接层之后经过Softmax函数,确

2)网络优化训练

本文提出双分支的网络结构在训练度量特征向量的损失函数过程中分别独立完成。各分支中设定,视频中的目标为G,每一帧中的候选区域为c,搜索区域为C,每一个分支的输入为:(G,C);每一个分支的特征提取的过程可以用函数$F_1(\cdot)$表示,在Transformer模块的网络结构隐射过程中,用式(2)表示如下:

$$\Gamma_1(G,C)=\mathrm{Coff}(F_1(G),F_1(C)) \tag{2}$$

式(2)表示在Transformer网络结构中的特征提取的训练过程,Coff是该分支中的相关系数的函数,在实际的网络训练的过程中大量的(G_i,C_i)数值会产生大量的预测值,真实的区域和预测区域的映射函数H_i,该分支的优化过程可以用式(3)表达:

$$\underset{\theta_1}{\operatorname{argmin}}\frac{1}{N}\sum_{i=1}^{N}\{L(\Gamma_1(G_i,C_i);\theta_1);H_1\} \tag{3}$$

交叉金字塔的网络结构在训练的过程中定义输入为(G^n,C)一个多特征的融合函数$F_2(\cdot)$,特征向量融合之后可以用函数$\mu(F_2(C))$表示,分支的映射过程函数表达式如式(4)所示:

$$\Gamma_1(G,C)=\mathrm{Coff}(\mu(F_2(G)),\mu(F_2(C))) \tag{4}$$

在经过该分支处理之后的优化方向如式(5)所示:

$$\underset{\theta_2}{\operatorname{argmin}}\frac{1}{N}\sum_{i=1}^{N}\{L(\Gamma_2(G_i^n,C_i);\theta_2);H_2\} \tag{5}$$

经过两个分支融合之后的函数表达式如式(6)所示:

$$\Gamma(G^n,C)=\lambda\Gamma_1((G,C))+(1-\lambda)\Gamma_2((G^n,C))+\beta\Delta\omega \tag{6}$$

其中，$\beta\Delta\omega$ 是惩罚项，根据实际的数据库中的数据特征设定该值。

在网络的训练的过程中损失函数包含 3 个部分；基于目标区域的 Box-loss[24]、时空一致性的目标运动的位置 Conf-loss[25] 以及基于注意力机制的 ID-loss[26]。

$$\mathcal{L}_{Total}=\mathcal{L}_{bl}+\mathcal{L}_{cl}+\mathcal{L}_{IDl} \tag{7}$$

训练过程中为了精准定位跟踪对象的范围本文中利用 Box-loss 损失；检测对象之间的目标位置之间运动的相关性损失利用 Conf-loss 损失；针对多目标的分类问题利用 ID-loss 损失确定。网络在训练的过程中采用随机梯度下降的方式，在前 100 个 epoch 中学习率设置为 0.001，以后每 100 个 epoch 学习率降低 10 倍，直到学习率调整为 0.000 01 为止，停止训练。

图 6 中是在每一个 batch 中训练的是该数据集在 Box-loss、Conf -loss、ID-loss 中以及总 Total-loss 的损失函数的网络收敛情况。

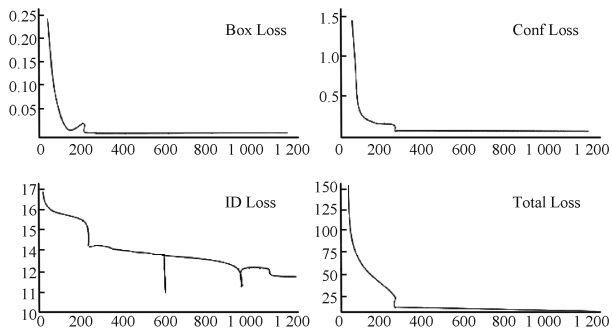


图 6　KITTI 数据集训练信息的基本情况

Fig. 6　Basic situation of training information in KITTI data set

2　实验以及结果分析

2.1　数据库以及实验分析

本文提出的网络结构在 KITTI 数据集上进行训练，该数据集是德国卡尔斯鲁厄理工学院和丰田美国技术研究院联合创办，数据集是复杂背景下的真实场景数据，数据集中车辆最多达到 15 辆。本文分别在数据集 D²-City[27]、UA-DETRAC[28] 上进行测试，对视频中车辆进行跟踪和检测的结果展开说明。D²-City 数据集采集在中国 5 个城市的滴滴运营车辆，所提供的原始数据均存储视频的帧的频率为 25 fps/s、时长 30 s 的短视频。UA-DETRAC 是一个现实世界多目标检测和多目标跟踪基准的中型数据集，数据集由 Cannon EOS 550 D 摄像头在中国北京和天津的 24 个不同地点拍摄的 10 h 的视频组成。视频以 25 帧/s 的速度录制，录制的分辨率为 960×540 pixel。在 UA-DETRAC 数据集中，有超过 140 000 帧和 8 250 辆车被人工标注，总共标记了 1 210 000 物体的边界盒，该数据被广泛的应用在多目标和单目标的跟踪和检测中作为基准数据，图 7 是 UA-DETRAC 数据集各类数据的真实基本情况展示。

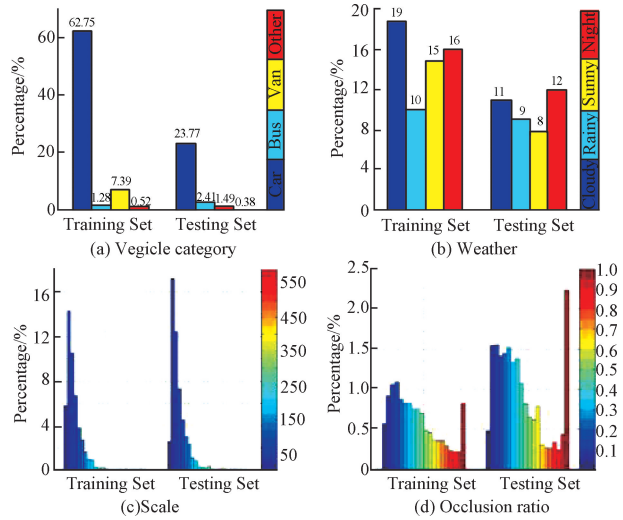


图 7　UA-DETRAC 数据集属性信息的基本情况[28]

Fig. 7　Basic information of attribute information of UA-DETRAC data set

2.2　实验分析

本文提出的网络结构利用 Keras 框架完成，网络的训练过程是在两块 P100-PCIE-16GB GPU 上进行。本文提出的网络结构在 KITTI 数据集上训练之后，在两种基准数据集上的检测结果分析如表 1 所示。

表 1　各种方法在 D²-City 和 UA-DETRAC 集上的对比

Table 1　Comparison of various methods on D²-City and UA-DETRAC sets

方法	策略	D²-City 数据集		UA-DETRAC 数据集	
		多目标跟踪	单目标跟踪	多目标跟踪	单目标跟踪
Faster R-CNN	数据增强	13.7	33.7	14.2	32.5
	蒸馏	12.9	22.4	13.9	34.5
	数据+蒸馏	11.5	31.3	13.4	30.2
CompACT	数据增强	14.2	36.6	14.0	34.5
	蒸馏	13.1	34.2	13.9	32.4
	数据+蒸馏	14.5	30.2	13.8	31.4
R-CNN	数据增强	13.8	32.8	14.2	34.1
	蒸馏	11.2	34.5	12.4	35.2
	数据+蒸馏	13.2	35.3	14.2	35.1
ACF	数据增强	12.8	37.1	14.2	35.1
	蒸馏	13.5	32.1	13.5	32.1
	数据+蒸馏	13.9	34.5	12.4	33.1
DPM	数据增强	10.4	32.1	11.4	32.9
	蒸馏	11.8	33.4	14.2	32.9
	数据+蒸馏	12.3	32.4	14.4	32.9
Ours method	数据增强	**14.7**	**35.2**	**14.6**	**35.8**
	蒸馏	**15.2**	**35.8**	**14.9**	**37.7**
	数据+蒸馏	**14.9**	**34.6**	**14.7**	**37.6**

根据表1中的数据可知,网络结构在两个数据集上对单目标检测和跟踪的效果和多目标跟踪的效果比之前的网络结果有所提高;同时利用本文提出的网络结构在两种数据集上测试的结果表明,提出的网络结构在多目标和单目标的跟踪和检测过程中时间和精度都得到提高。本文算法在加入了数据增强策略之后精度提高1%~4.3%;加入蒸馏技术之后,明显的精度提高了2.1%~3.4%;但是将蒸馏技术和数据增强都加入检测的精度反而有所下降,主要原因是在对视频目标的跟踪过程中,由于本身数据量有限,从而导致加入蒸馏和数据增强之后,网络对运动对象的特征提取过程中效果不佳。

如图8、9所示,同时和其他的网络方法进行比较,显示本文提出的网络结构的性能在精度和时间的实时性得到了很好的折中,主要比较各类算法在召回率和精度之间的关系。在测试集上面的各种指标比较,最终在真实的数据集对车辆进行跟踪,如表2所示,本文的方法的时间上提高4%,精度提高2.3%。

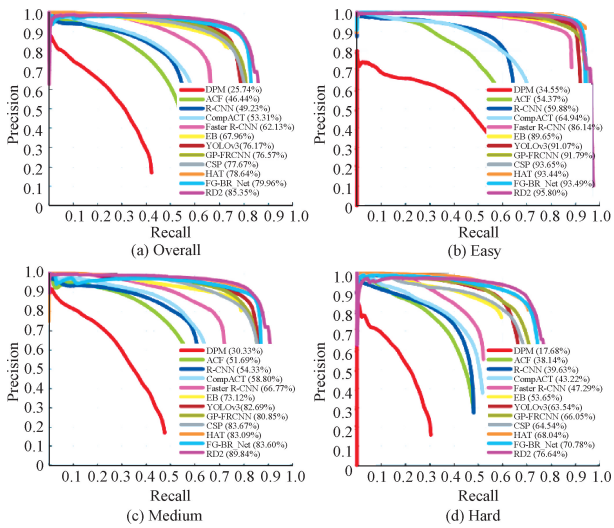


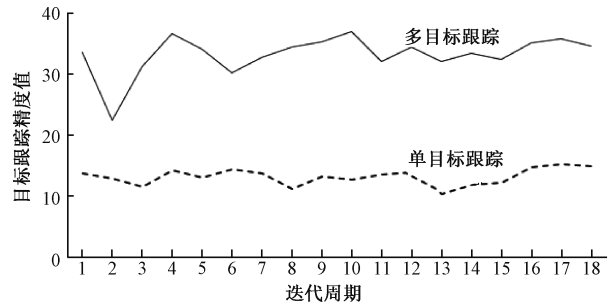
图8　各种网络结构在数据集上的召回率的结果

Fig. 8　The result of the recall rate of various network structures on the data set

表2　本文提出的算法在增加数据处理之后的指标结果

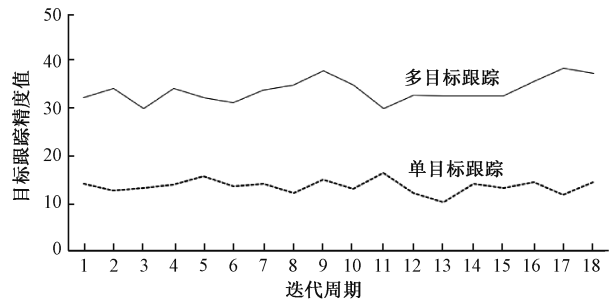
Table 2　The index result of the algorithm proposed in this paper after adding data processing

方法		数据增强	知识蒸馏	光流特征
本文算法			√	√
		√		√
		√	√	√
结果	MAE	0.018	0.014	0.015
	$maxF_\beta$	0.964	0.954	0.951
	S	0.965	0.967	0.971
检测时间		↓	↑	↑



(a) 各种网络结构在数据集D^2-City检测结果

(a) Comparison of detection results of various network structures on the data set of D^2-City



(b) 各种网络结构在数据集UA-DETRAC检测结果

(b) Comparison of detection results of various network structures on the data set of UA-DETRAC

图9　各种网络结构在数据集上的检测结果

Fig. 9　Comparison of detection results of various network structures on the data set

2.3　消融实验

消融实验主要分为两个部分,一是利用本文提出的网络结构增加知识蒸馏和数据增强技术之后,对网络的检测的精度和实时性进行比较,说明本文提出的算法的泛化性和检测的效率;结果显示,本文提出的网络结构检测结果的有效性,由表2可知;另一方面,通过在其他数据集上利用本文提出的基于时空一致性的网络结构的P-R值,如图10所示,说明本文提出的网络结构的泛化能力和检测效率。

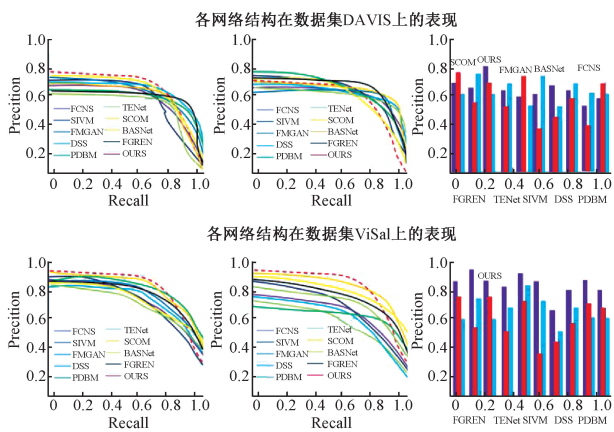


图10　本文提出的算法在其他数据集上的性能表现

Fig. 10　The performance of the algorithm proposed in this article on other data sets

在测试数据集上的 3 个衡量显著性视频目标的指标 MAE、$maxF_\beta$、S 以及检测的时间 4 个指标说明本文提出的算法的有效性和泛化性。

为了验证本文提出的算法在泛化能力上的表现,利用本文提出的算法在数据集 ViSal 和数据集 DAVIS 上的对显著性视频目标的检测结果如图 10 所示。

图 10 显示该算法在两个数据集上准确率和召回率之间的关系,本文提出的算法较其他的网络结构在显著性目标检测方面,通过 P-R 值表现出强大的性能,明显优于其他网络算法。

3　结　论

本文利用孪生网络 Siamese-Net 的结构思想,设计基于时空一致性的双分支网络结构,利用 Transformers 结构对复杂场景下对视频车辆的空间和时间信息进行关联度分析,确定其精准的位置和对象信息;另一方面利用交叉特征金字塔对对象的特性进行多尺度的特征融合。通过在基准数据库上的实验以及实际的预测结果显示,本文提出的网络结构在精度和速度上面较之前的网络都有所改进;同时利用本文提出的网络在视频显著性目标检测中进行消融实验,在精度和速度上分别提高 4% 和 2.3%,说明本算法较其他算法具有更强的泛化性和有效性。下一步将重点研究如何降低网络模型的参数量和实时性以及防止网络出现过拟合,尝试对网络进行知识蒸馏和迁移学习等措施,使模型在轻量级上表现优越,提高检测精度的同时改善模型的泛化能力。

参考文献

[1] 董永昌, 单玉刚, 袁杰. 基于改进 SSD 算法的行人检测方法 [J]. 计算机工程与设计, 2020, 41 (10): 2921-2926.
DONG Y CH, SHAN Y G, YUAN J. Pedestrian detection method based on improved SSD algorithm [J]. Computer Engineering and Design, 2020, 41 (10): 2921-2926.

[2] 窦鑫泽, 盛浩, 吕凯, 等. 基于高置信局部特征的车辆重识别优化算法 [J]. 北京航空航天大学学报, 2020, 46 (9): 1650-1659.
DOU X Z, SHENG H, LYU K, et al. An optimization algorithm for vehicle re-identification based on high-confidence local features [J]. Journal of Beijing University of Aeronautics and Astronautics, 2020, 46(9): 1650-1659.

[3] 洪伟, 王吉通, 刘宇, 等. 基于 DBSCAN 的复杂环境下车道线鲁棒检测与跟踪 [J]. 吉林大学学报(工学版), 2020, 50(6): 2122-2130.
HONG W, WANG J T, LIU Y, et al. Robust detection and tracking of lane lines in complex environments based on DBSCAN [J]. Journal of Jilin University (Engineering and Technology Edition), 2020, 50(6): 2122-2130.

[4] 李熙莹, 周智豪, 邱铭凯. 基于部件融合特征的车辆重识别算法[J]. 计算机工程, 2019, 45 (6): 12-20.
LI X Y, ZHOU ZH H, QIU M K. Vehicle re-identification algorithm based on component fusion features [J]. Computer Engineering, 2019, 45 (6): 12-20.

[5] 邱铭凯, 李熙莹. 用于车辆重识别的基于细节感知的判别特征学习模型 [J]. 中山大学学报(自然科学版), 2021, 60(4): 111-120.
QIU M K, LI X Y. Discriminant feature learning model based on detail perception for vehicle re-identification [J]. Journal of Sun Yat-sen University (Natural Science Edition), 2021, 60(4): 111-120.

[6] 盛涛, 夏海宝, 肖冰松. 基于 AIMM-SRCKF 的机动目标跟踪算法 [J]. 电子测量与仪器学报, 2021, 35 (1): 159-164.
SHENG T, XIA H B, XIAO B S. Maneuvering target tracking algorithm based on AIMM-SRCKF [J]. Journal of Electronic Measurement and Instrumentation, 2021, 35(1): 159-164.

[7] DANELLJAN M, ROBINSON A, SHAHBAZ KHAN F, et al. Beyond correlation filters: Learning continuous convolution operators for visual tracking [C]. European Conference on Computer Vision. Springer, Cham, 2016: 472-488.

[8] WANG L, OUYANG W, WANG X, et al. Stct: Sequentially training convolutional networks for visual tracking [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 1373-1381.

[9] BERTINETTO L, VALMADRE J, HENRIQUES J F, et al. Fully-convolutional siamese networks for object tracking [C]. European Conference on Computer Vision. Springer, Cham, 2016: 850-865.

[10] HE A, LUO C, TIAN X, et al. A twofold siamese network for real-time object tracking [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 4834-4843.

[11] LI H, CHEN G, LI G, et al. Motion guided attention for video salient object detection [C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 7274-7283.

[12] 卢笑, 曹意宏, 周炫余, 等. 基于深度强化学习的两

阶段显著性目标检测[J]. 电子测量与仪器学报, 2021, 35(6): 34-42.
LU X, CAO Y H, ZHOU X Y, et al. Two-stage salient target detection based on deep reinforcement learning [J]. Journal of Electronic Measurement and Instrumentation, 2021, 35(6): 34-42.

[13] 郝腾龙, 李熙莹. 提升预测框定位稳定性的视频目标检测[J]. 中国图象图形学报, 2021, 26(1): 113-122.
HAO T L, LI X Y. Video target detection to improve the stability of prediction frame positioning [J]. Chinese Journal of Image and Graphics, 2021, 26(1): 113-122.

[14] HU B, GAO B, WOO W L, et al. A lightweight spatial and temporal multi-feature fusion network for defect detection[J]. IEEE Transactions on Image Processing, 2021, 30: 472-486.

[15] SONG K Y, YANG H, YIN Z P. Multi-scale attention deep neural network for fast accurate object detection[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 29(10): 2972-2985.

[16] CAO J, CHEN Q, GUO J, et al. Attention-guided context feature pyramid network for object detection[J]. arXiv Preprint, 2020, arXiv: 2005.11475.

[17] YANG W, ZHANG X, LEI Q, et al. Lane position detection based on long short-term memory (LSTM)[J]. Sensors, 2020, 20(11): 3115.

[18] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [J]. Advances in Neural Information Processing Systems, 2017, 30.

[19] BILKHU M, WANG S, DOBHAL T. Attention is all you need for videos: Self-attention based video summarization using universal transformers[J]. arXiv Preprint, 2019, arXiv: 1906.02792.

[20] TAN H, LIU X, TIAN S, et al. Mhsa-net: Multi-head self-attention network for occluded person re-identification[J]. arXiv Preprint, 2020, arXiv: 2008.04015.

[21] LIU R, YUAN Z, LIU T, et al. End-to-end lane shape prediction with transformers [C]. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021: 3694-3702.

[22] ARNAB A, DEHGHANI M, HEIGOLD G, et al. Vivit: A video vision transformer[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 6836-6846.

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

[24] WU S, YANG J, WANG X, et al. Iou-balanced loss functions for single-stage object detection [J]. Pattern Recognition Letters, 2019, arXiv: 1908.05641.

[25] KERVADEC H, BOUCHTIBA J, DESROSIERS C, et al. Boundary loss for highly unbalanced segmentation [C]. International Conference on Medical Imaging with Deep Learning. PMLR, 2019: 285-296.

[26] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017(99): 2999-3007.

[27] CHE Z, LI G, LI T, et al. D^2-City: A large-scale dashcam video dataset of diverse traffic scenarios [J]. 2019, 10.48550/arXiv.1904.01975.

[28] WEN L, DU D, CAI Z, et al. UA-DETRAC: A new benchmark and protocol for multi-object detection and tracking[J]. Computer Vision and Image Understanding, 2020, 193: 102907.

作者简介

洪锋, 2011年于桂林理工大学获得硕士学位, 2017年进入合肥工业大学计算机与信息学院信号与信息处理专业攻读博士学位, 主要研究方向为复杂场景的目标检测、智能信息处理。
E-mail: hongfeng@czu.edu.cn

Hong Feng received his M. Sc. degree in 2011 from Guilin University of Technology. Since 2017, he has been a Ph. D. candidate in the School of Computer and Information, Hefei University of Technology, majoring in signal and information processing. His main research interest includes target detection in complex scenes; intelligent information processing and deep learning.

鲁昌华(通信作者), 1983年于合肥工业大学获学士学位, 1988年于哈尔滨工程大学获硕士学位, 2001年于中国科学院获博士学位, 现为合肥工业大学教授, 主要研究方向为智能信息处理等。
E-mail: lch6208@163.com

Lu Changhua (Corresponding author) received his B. Sc. degree from Hefei University of Technology in 1983, M. Sc. degree from Harbin Engineering University in 1988 and Ph. D. degree from Chinese Academy of Sciences in 2001. Now he is a professor in Hefei University of Technology. His main research interests include intelligent information processing and so on.