DOI:10.19651/j.cnki.emt.2212176

融合 SimAM 注意力机制的实时多目标跟踪算法*

程之星 杨 帆
(河北工业大学电子信息工程学院 天津 300401)

摘 要：多目标跟踪中的 JDE 算法首次将目标检测与重识别进行联合学习，极大提升了跟踪速度，但由于复杂背景干扰和遮挡导致跟踪准确度下降。为了解决跟踪速度与准确度的平衡问题，本文提出了 SAM-JDE，该模型融合了 SimAM 注意力机制、多尺度融合等思想，通过增强特征提取能力提高目标跟踪的准确性。使用 CIoU_Loss 作为回归损失函数，通过准确地构建目标框和预测框之间的位置关系来提升定位精度。关联匹配部分使用卡尔曼滤波预测运动信息，匈牙利匹配算法完成时序维度上的目标关联。在 MOT16-test 数据集上进行测试，MOTA 达到 66.4%，跟踪速度为 20.6 FPS，在保证实时性的基础上跟踪准确度较 JDE 算法提升 2.3%，较好地优化了准确度与速度的平衡问题。

关键词：机器视觉；多目标跟踪；注意力机制；实时跟踪
中图分类号：TP391.4 **文献标识码**：A **国家标准学科分类代码**：510.4

Real-Time multi-object tracking algorithm based on SimAM attention mechanism

Cheng Zhixing Yang Fan
(School of Electronics Information Engineering, Hebei University of Technology, Tianjin 300401, China)

Abstract: JDE algorithm in multi-object tracking jointly learns target detection and re-identification for the first time, which greatly improves the tracking speed. However, the tracking accuracy is reduced due to the poor tracking effect caused by complex background interference and occlusion processing. In order to balance the tracking speed and accuracy, SAM-JDE is proposed in this paper. This model integrates SimAM attention mechanism, multi-scale fusion and other ideas to improve the accuracy of target tracking by enhancing the ability of feature extraction. CIoU_Loss is used as the regression loss function to improve the positioning accuracy by accurately building the position relationship between the target box and the prediction box. In the association matching part, Kalman filtering is used to predict the motion information, and the Hungarian matching algorithm completes the target association in the time series dimension. Testing on MOT16-test dataset shows that MOTA reaches 66.4% and tracking speed is 20.6 FPS. On the basis of ensuring real-time performance, tracking accuracy is 2.3% higher than JDE algorithm, which better optimizes the balance between accuracy and speed.

Keywords: machine vision; multi-object tracking; attention mechanism; real time tracking

0 引 言

多目标跟踪(multiple object tracking,MOT)旨在预测视频连续帧中多个目标的位置和身份信息，广泛应用于无人驾驶、智能监控等领域。基于深度学习实现多目标跟踪的方法分为两类：单阶段和双阶段[1]。双阶段方法遵循检测跟踪范式，即将多目标跟踪任务分离为目标检测和重识别(re-identification,ReID)两个单独的任务。这类方法首先使用比较成熟的目标检测网络，例如：YOLOV3[2]或 Faster RCNN[3]等算法，基于输入的视频序列对多个目标进行分类和定位，提取目标的运动特征。而后根据目标框在序列图像上的预测位置进行裁剪，输入到重识别网络提取嵌入向量，用于确认目标的外观特征。根据运动特征和外观特征计算代价矩阵，利用匹配算法进行时序关联。DeepSort[4]是双阶段的经典算法。但此类方法中的检测和重识别需要两个独立的网络串联计算完成推理，整个算法

收稿日期：2022-11-23
* 基金项目：国家重点研发计划智能机器人专项(2019YFB1312102)、河北省自然科学基金(F2019202364)项目资助

的推理时间大致是二者之和,导致跟踪速度较慢。

由于速度是多目标跟踪技术落地的关键问题,Wang 等[5]提出了联合检测与嵌入模型(jointly learns the Detector and Embedding model,JDE)方法,即单阶段。该方法将目标检测和重识别用一个共享网络进行联合学习,在推理阶段较双阶段方法减少了一个网络的计算,两个任务共享了大量的底层计算,极大减少了 MOT 系统运行的时间。但在背景复杂、频繁遮挡等干扰下,JDE 算法的跟踪效果一般,容易出现因目标定位不准导致的漏跟、误跟等现象。为了改善跟踪效果,Zhan 等[1]提出了基于无瞄框目标检测算法的 FairMOT 算法,采用降采样的方式优化目标外观特征的提取,提升了跟踪准确度;Zhang 等[6]改进了关联匹配策略,提出减少因检测得分较低导致漏检的 ByteTrack 算法。近年来,有研究人员在 JDE 范式下探索新框架进行创新,利用 Query-Key 机制来跟踪当前帧中已经存在的目标并检测新目标,例如 TransTrack[7]、MeMOT[8]、Trackformer[9]等,但该框架引入 transformer 后跟踪速度有明显下降。

为了确保算法实时性的同时提升跟踪准确性,本文针对 JDE 算法进行优化,提出单阶段范式的多目标跟踪算法 SAM-JDE。在此之前,通过引入通道和空间注意力模块改进 JDE 模型在目标发生重叠时的跟踪效果[10],从而获得更高的跟踪精度,但两个注意力模块都存在一定的参数,增加了网络的参数量且影响了跟踪速度。从实际应用价值出发,算法的参数量和速度尤为重要,SAM-JDE 则从以下两点进行优化:1)将 3D 无参注意力机制 SimAM[11]首次引入速度较快的 JDE 范式多目标跟踪,在不增加参数的基础上进行特征增强,强化网络对感兴趣目标的定位与提取能力,改善目标跟踪中的误报漏报,提升目标跟踪的准确度;2)由于损失函数只作用于网络训练过程,不会影响网络的参数量和跟踪过程中网络的推理速度,SAM-JDE 将回归任务损失函数中的 Smooth L1 Loss[12]改进为 CIoU_Loss[13],同时考虑重叠面积、中心点距离、长宽比三个因素来引导预测框的回归,提升检测框回归的速度和准确度,获得较好的检测和跟踪效果。以上改进促使 SAM-JDE 算法在保证实时性的同时还提高了跟踪准确性,对于实际部署使用较为友好。

1 SAM-JDE 算法原理及改进

本文提出的算法整体采用单阶段范式,完成单镜头下的多目标跟踪。算法流程如图 1 所示。大致流程为将视频的连续帧输入到本文设计的 SAM-JDE 模型完成目标检测和 ReID 两个任务的联合学习,输出当前帧的目标边界框和目标外观嵌入向量,即检测信息和外观特征。然后将两部分特征通过卡尔曼滤波算法[14]和匈牙利算法完成视频帧中轨迹的更新与目标的匹配,完成匹配关联步骤,最后输出整个视频中的跟踪轨迹。

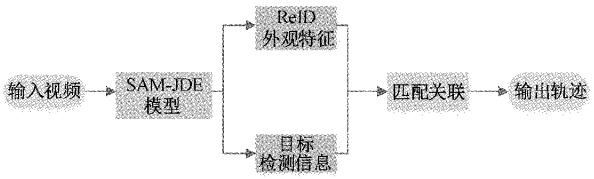


图 1 算法流程

1.1 SAM-JDE 模型结构

SAM-JDE 模型由骨干网络 Darknet-53、特征金字塔 FPN[2]与预测头 3 个部分组成,如图 2 所示。其中,Darknet-53 借鉴残差连接思想加深网络的层数,从而提高网络的特征表达能力。视频中每帧图像尺寸 resize(1 088×608)大小,通过 Darknet-53 进行特征提取操作,首先经过一个卷积层(卷积核为 3×3,通道数为 32),再通过 5 个残差块进行 5 次下采样,图像的宽高不断压缩,通道数不断扩张。将降采样率为 8、16、32 不同分辨率下的特征图通过 3 个 SimAM 注意力模块,根据不同尺寸特征图下各神经元的重要性分别赋予相应权重,使网络对一帧图像中大小不一的物体都更好地提取到重要特征,弱化非重点特征,提升网络对目标的定位能力,缓解物体背景复杂,增强检测器的性能。经过强化后的特征再通过 FPN 网络,对 3 个尺度的特征进行融合,具体为将最小尺寸(34×19)的特征图上采样与第二小的特征图(68×38)跳跃连接进行融合,其余尺度类似,融合后获得语义信息更加完善的特征,提升对小尺度目标的跟踪能力。输出(34×19×75)、(68×38×75)和(136×76×75)3 个尺度的融合特征图,送入预测头中进行分类、回归和嵌入外观特征提取多任务联合学习。

1.2 3D 无参注意力机制 SimAM

在多目标跟踪场景中,存在遮挡、背景干扰等现象,对目标跟踪造成了很大的干扰。受目标检测中为了提高模型表达能力加入非 3D 带参注意力机制的 Attention-YOLO[15]的启发,针对上述问题,本文在骨干网络后引入了 SimAM 注意力模块,提高网络提取关键目标的能力并通过直接有效的方式对跟踪目标的特征进行优化,减弱上述干扰。

根据视觉神经科学研究发现,信息最丰富的神经元与周围神经元相比有独特的放电方式,且会抑制周围神经元的活性,这种现象称为空间抑制[16]。具有明显空间抑制效应的神经元应当是更重要的,SimAM 通过衡量某个神经元与其他神经元的线性可分性来找到更为重要的神经元,并赋予更大的权重,在多目标跟踪任务中这些神经元往往负责提取出目标的关键特征并对其进增强。如图 2 所示,本文在骨干网络后加入 SimAM 模块,优化骨干网络提取出来的特征,通过赋予重要神经元更高的权重,从而获得更加完整和精细的目标特征,减少复杂背景的干扰及遮挡导致提取的目标特征较弱或漏检。基于神经科学的发现,定义了神经元的能量函数,如式(1)所示。

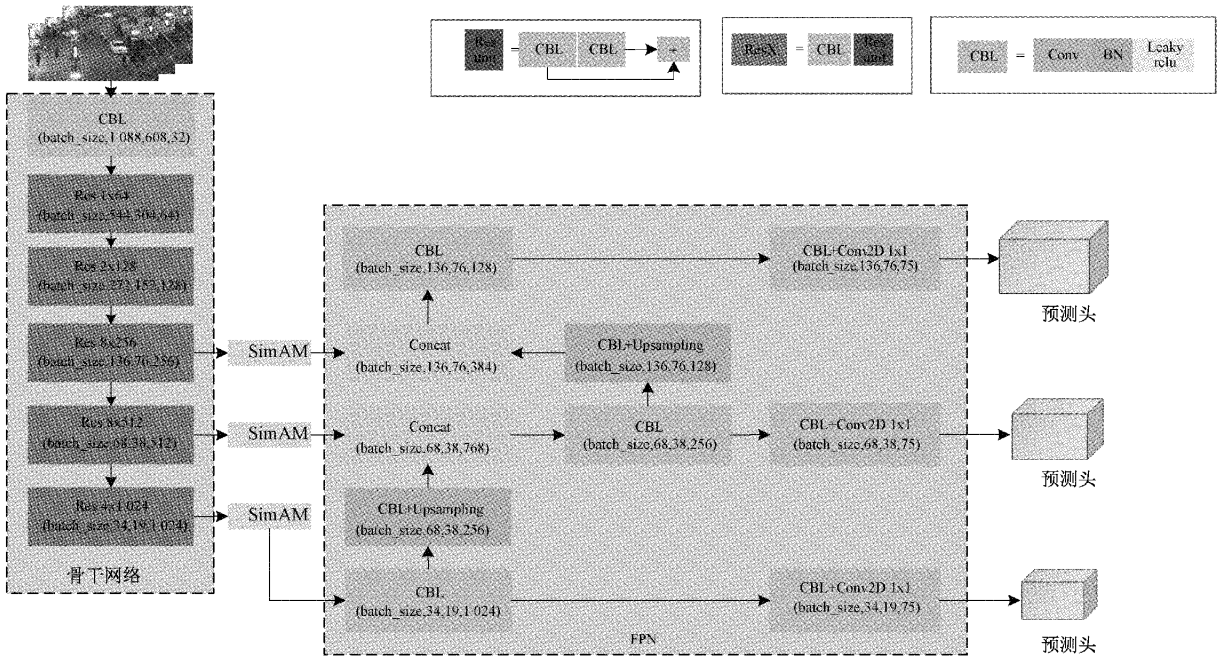


图 2　SAM-JDE 模型结构图

$$e_t^* = \frac{4(\hat{\sigma}^2+\lambda)}{(t-\hat{\mu})^2+2\hat{\sigma}^2+2\lambda} \tag{1}$$

其中，$\hat{\mu}=\frac{1}{M}\sum_{i=1}^{M}x_i$，$\hat{\sigma}^2=\frac{1}{M}\sum_{i=1}^{M}(x_i-\hat{\mu})^2$。

式(1)表明，神经元 t 的能量越低，其重要性越强，所以每个神经元的重要性由能量函数的倒数得到。由于注意力模块通常在人类大脑表现为每个神经元反应的增强效应，例如标量增强。所以，可以通过标量运算而不是添加其余的模块，通过 $1/e_t^*$ 对神经元根据重要性进行加权，从而提高网络对跟踪目标的特征表达能力，如式(2)所示。

$$\tilde{X}=\mathrm{sigmoid}\left(\frac{1}{E}\right)\odot X \tag{2}$$

SimAM 的权重赋予方式如图 3 所示，直接在通道和空间上同时给每个神经元赋予不同的权重，可以更加全面的评估各个神经元的重要性，其权重的计算都是根据神经科学研究理论获得，相比于 SENet[17]、CBAM[18] 等通过手工设计注意力模块不同，其更具解释性且无需引入可学习参数。

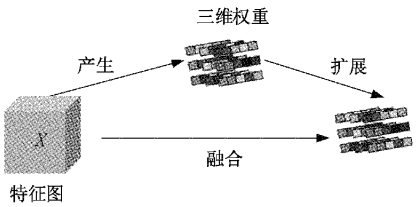


图 3　SimAM 注意力

在 SAM-JDE 模型中 SimAM 模块通过上述能量函数的封闭解计算出了 Darknet-53 提取出的特征重要性分布，对具有更多有效信息的神经元输出进行强化，并有效抑制其余无关特征的干扰，使得网络具有更强的特征表达能力，提高了算法面对复杂背景和遮挡的抗干扰能力和定位能力。

1.3　预测输出

图 2 中预测头由几个堆叠的卷积层构成且输出密集预测图，预测图被分为检测分支和外观嵌入分支。检测分支输出检测框的分类结果、检测框的回归结果，外观嵌入分支输出嵌入向量。如图 4 所示。

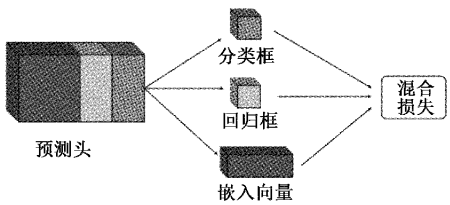


图 4　预测头

检测分支采用与 YOLOV3 相同的方式完成目标的分类及回归，使用交叉熵损失函数及 CIoU_Loss，在网络最后 3 个尺度的特征图上利用预先设置的 anchor 完成检测。嵌入分支将 ReID 任务视为分类任务来完成，使用交叉熵损失函数将检测到的目标通过卷积操作提取出 512 维的特征向量，用于后续关联匹配。最后将 3 个分支的损失函数聚合成混合损失，通过误差反向传播算法完成模型的训练。

在 JDE 算法中，每个预测头的回归损失函数是 Smooth L1 Loss[12]，直接回归边界框中的四个顶点，将 4 个点的 Loss 求和作为边界框归回损失，忽略掉了 4 个点

之间的相关性且实际评价指标是交并比(intersection over union,IoU),两者不匹配,因为存在多个检测框可能有相同的 Smooth _ L1 _ Loss,但 IoU 差异很大。之后 Unitbox[19]提出 IoU-Loss,将 4 个点构成目标边界框看成一个整体进行回归,计算真实边界框与预测边界框的交并比如式(3),并利用式(4)作为回归任务的损失函数,提高预测框回归的准确性。

$$R_{IoU}=\frac{|B\cap B_{gt}|}{|B\cup B_{gt}|} \tag{3}$$

其中,B 为预测框,B_{gt} 为真实框。

$$L_{IoU}=1-R_{IoU} \tag{4}$$

但 IoU_Loss 存在 3 点不足:1)当预测框和真实框不相交的时候,$IoU=0$,此时无法反应两个框的位置关系,且反向传播对损失函数不可导,无法优化这种情况。2)当真实框包含预测框时,存在预测框位置不同但 IoU 值相同,此时无法反应两者的位置关系。3)严重依赖 IoU 的值,导致收敛速度降低。如图 5 所示,3 种不同情况下的 IoU_Loss 值都相同,但显然图 5(c)回归的更加准确。

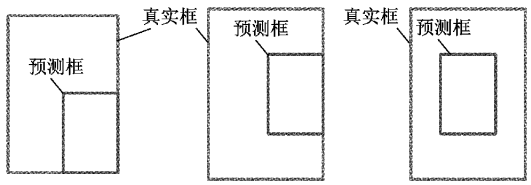


图 5 CIOU_Loss 与 IoU_Loss 对比图

为了加速收敛且提升定位目标的准确性,采用 CIoU_Loss[13]作为回归任务的损失函数。如式(5)所示。其同时考虑了两个框之间的重叠面积、中心点距离和长宽比 3 个因素,通过比较两个中心点的距离可以解决两个框包含情况下的问题,由图 5 观察到,图 5(c)中间位置的 CIoU_Loss 是更低的,对应定位更加精准。通过增加长宽,增加了网络对目标的定位速度和准确性,减少了由于背景或者遮挡导致的漏检、跟踪轨迹断裂等问题,使得网络的连续性更好。

$$L_{CIoU}=1-IoU+\frac{\rho^2(b,b_{gt})}{c^2}+\alpha\nu \tag{5}$$

其中,ρ 是两个框中心点坐标的欧氏距离,c 则是保住它们的最小方框的对角线距离,a 是用于做 trade-off 的参数,$\alpha=\frac{\nu}{(1-IoU)+\nu}$,$\nu$ 是用来衡量长宽比一致性的参数,$\nu=\frac{4}{\pi^2}\left(\arctan\frac{w_{gt}}{h_{gt}}-\arctan\frac{w}{h}\right)^2$。CIoU_Loss 作为回归损失函数,使预测框回归的准确度和速度更高一些。

1.4 匹配关联

利用预测头输出的检测信息和嵌入向量完成关联匹配。具体为首先对视频第 1 帧中检测出的目标初始化轨迹,形成一个轨迹池。然后从第 2 帧开始计算当前帧观测信息的嵌入向量与现有轨迹池的嵌入向量之间的距离和框之间的交并比计算出代价矩阵,送入匈牙利算法将当前帧的观测目标分配给现有轨迹,即与轨迹进行关联匹配。卡尔曼滤波器用于平滑轨迹,并预测先前的轨迹在当前帧中的位置,如果当前帧被分配的观测结果在空间上与预测位置距离太远的话,则将拒绝分配。如果某条轨迹没有被分配到观测目标,则将该轨迹标记为丢失态,若丢失态持续时间超过设定的阈值,则将该轨迹从轨迹池中删除。如果观测结果没有被分配轨迹,视为新出现的目标,则初始化一个新的轨迹。然后,对轨迹的外观嵌入向量进行如下更新:

$$\boldsymbol{f}_t=\gamma\boldsymbol{f}_{t-1}+(1-\gamma)\tilde{\boldsymbol{f}} \tag{6}$$

其中,$\tilde{\boldsymbol{f}}$ 表示指定观测值的嵌入向量;$\boldsymbol{f}_t$ 表示 t 时刻轨迹的嵌入向量;γ 是平滑的动量项;匹配更新完成后,最后输出跟踪结果,每帧图像中框出行人位置且分配唯一的身份标识(identity document,ID)。

2 实验及实验数据分析

2.1 数据集与评价指标

数据集分为训练集和测试集。由于小数据集得出的模型迁移性较差,故本文训练集由 6 个行人检测与跟踪的公开数据集 MOT17、caltech、citypersons、cuhksysu、prw、eth 中的子训练集相加构成。测试集是 MOT16 公开数据集,并将 MOT16 与 ETH 中重复的视频删除以保证评估的准确性。训练集中有图片 54 000 张,已标注真实框 270 000 个以及 ID 标注 8 700 个。

为了评估整个 MOT 系统的性能,使用了与人类感知最一致的 MOTA 指标。具体指标详情如下:

1)MT:正确跟踪的帧数在总帧数中占比高于 80%的 GT(ground truth)轨迹数量。

2)ML:正确跟踪的帧数在总帧数中占比低于 20%的 GT 轨迹数量。

3)IDs:每个跟踪的对象都有唯一标识的真实 ID,但在跟踪过程中由于跟踪算法性能不足够强,导致对象的 ID 发生切换,ID 发生切换的总次数就是 IDS。

4)FPS:帧率。

5)IDF1:代表被检测和跟踪的目标中获取正确 ID 的检测目标的比例,考察跟踪的连续性和重识别的准确性。

6)FP(false positive):指模型将负样本预测为了正,也称作误报。MOTA 中指总的误报数量,即整个视频序列中每一帧 FP 数量和。

7)FN(false negative):指模型将正样本预测为负,也称作漏报。MOTA 中指总的漏报数量,即整个视频序列中

每一帧 FN 数量和。

8)MOTA:多目标跟准确度,公式如下:

$$MOTA = 1 - \frac{FN + FP + IDSW}{GT} \tag{7}$$

2.2　实验环境

本实验的硬件条件为 Intel Xeon Silver 4114 CPU,64 G 内存,2.2 GHz 主频,4 块 NVIDIA GeForce GTX 1080 Ti GPU,显卡内存为 12 GB。软件环境以 pytorch1.6 为开发框架,在 Ubuntu20.04 的系统及 Python3.6 的环境中运行。

2.3　训练策略

为了减少过拟合,采取数据增强技术,例如:随机翻转,缩放,色彩抖动等,将增强后的图像调整到固定的分辨率,如果没有指定,输入分辨率默认为 1 088×608。采用随机梯度下降方法训练了 30 个 epoch,初始学习率为 10^{-2} 并在第 15 和 23 个 epoch 的时候降低 0.1。Batch size 设置为 8,动量为 0.9。

2.4　实验结果分析

SAM-JDE 模型在上述训练策略下的损失函数曲线如图 6 所示,横轴表示训练的步数,纵轴表示损失值。可以发现无论是混合损失 loss 还是预测头中的分类损失 loss_conf、回归损失 loss_box、外观嵌入损失 loss_ide 都随着优化算法的迭代不断变小,且随着网络的训练,各个损失函数的值趋势都是先快速下降然后缓慢下降并趋于稳定,说明模型的训练过程征程并且最后收敛。

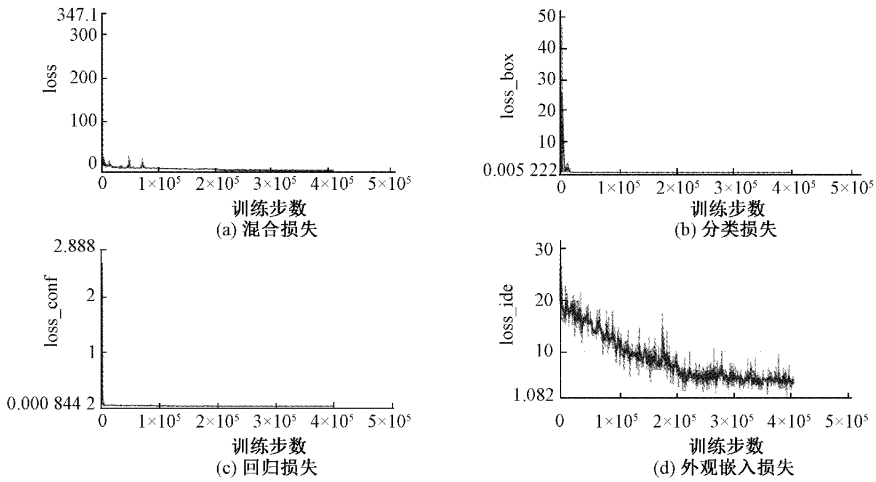


图 6　训练过程中损失函数值变化

1)将改进后的算法在 MOT16[20] 测试集上进行了评估,并与 DeepSort[4]、JDE[5] 等模型进行了对比,评估结果如表 1 所示。实验结果表明,SAM-JDE 算法在 MOTA 指标上可以达到 66.4%,同时跟踪速度可以达到 20.6 FPS,取得了较好的性能。与原 JDE 算法相比,跟踪速度基本不变的前提下,MOTA 提升了 2.3%。Trancstrack 在 MOTA 和 IDF1 指标中最高,但其跟踪速度较低,算法的实时性无法得到保障,而 SAM-JDE 在跟踪准确性较优的同时能够满足实时跟踪。与 Attention-JDE 算法相比,SAM-JDE 在跟踪速度和准确度上都更有优势,因为本算法优化特征的注意力机制是三维无参,而 Attention-JDE 中特征增强模块中的注意力机制引入了一定参数量和计算量,使得跟踪速度受到影响。且 SAM-JDE 对不影响速度的损失函数进行改进,使得回归框更加准确,检测器被优化进而提高了跟踪性能。

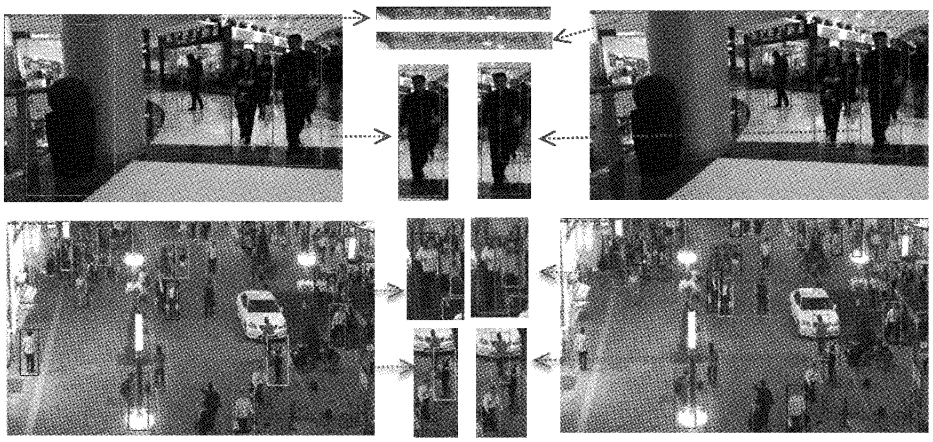
改进前后的跟踪结果对比如图 7 所示。图 7(a)为优化前的跟踪结果图,图 7(b)为优化后的 SAM-JDE 算法跟踪结果图,由高亮部分可以看出本文提出的 SAM-JDE 算法,首先在跟踪速度上可以达到 20 FPS 左右,保证了很好的实时性。其次,改进后的算法提升了跟踪的准确性,在复杂场景下减少了漏检了现象,同一帧图像准确跟踪的数量上升。同时在多目标之间存在遮挡时,跟踪效果也得到改善。经观察可知本文算法较好地平衡了多目标跟踪速度和准确度,追踪效果较好。

表 1　不同方法在 MOT16 测试集上的结果对比

方法	MOTA	IDF1	MT	ML	IDs	FPS
RAR16wVGG	63.0	63.8	39.9	22.1	482	<1.5
CNNMTT[21]	65.2	62.2	32.4	21.3	946	<6.4
Tube_TK	64.0	61.7	25.4	18.9	1 117	<2.1
DeepSort[3]	61.4	62.2	32.8	18.2	781	<8.1
TAP	64.8	62.5	32.1	17.3	794	<8.2
JDE[5]	**64.1**	60.9	32.4	20.7	1 231	**21.3**
TransTrack[7]	68.8	62.6	36.5	15.4	1 434	9.1
Attention-JDE[10]	65.7	61.8	32.1	20.3	1 191	18.2
Ours-1088	**66.4**	61.3	35.8	20.1	1 173	**20.6**

2)为了验证 SimAM 在跟踪速度和准确度上相比于其他主流注意力机制更加有效。本文将 SE[17]、SAM[21]、CBAM[18] 及 SimAm 注意力机制分别嵌入到 JDE 模型的相同位置,模型其余部及训练策略不变,训练完成后在 MOT17 测试集上进行了评估,实验结果如表 2 所示。

通过观察可以发现 4 种注意力机制在 MOTA 上均有



(a) 改进前跟踪结果　　(b) 改进后跟踪结果

图 7　跟踪结果对比图

表 2　添加不同注意力机制后的性能对比

注意力模块	MOTA	FPS
+SE	69.81	19.7
+SAM	69.25	19.4
+CBAM	70.39	19.1
+SimAM	**70.84**	**20.8**

提升,但 SimAM 效果最好。因为 SE 通道注意力机制如图 8(a)所示是对特征的通道赋予不同权重,但平均处理了每个空间位置的神经元,忽视了空间位置的信息。空间注意力机制如图 8(b)所示对特征的空间位置赋予了不同权重,但平均处理了每个通道的神 经元,忽略了通道位置的信息。CBAM[18]是先进行通道注意力再进行空间权重,将一维权重和二维权重进行合并,但这种方法并不能直接生成真正的三维权值。且两步方式需要太多的计算时间。

而 SimAM 直接生成三维权重,不同于通道注意力和空间注意力,消除了通道和空间注意力分别在空间维度和通道维度上的同一性,更加符合人类注意力的方法。从表 2 中可以看出,SimAM 的 FPS 最高,跟踪速度最快,且参数增加量为 0,因为 SimAM 的推理构成不需要进行卷积、池化等操作。

不同注意力机制输出结果的热力图可视化结果如图 9 所示。颜色越深代表模型最当前区域的关注度越高,对结果的影响越大。通过观察热力图可知,图 9(a)未引入注意力机制的热力图最为弥散,包含了过多的背景信息。从左到右图像中对 3 个目标的注意力越来越集中,模型对目标的跟踪定位更加精准,针对图像中目标物体的提取能力更加强大,使得跟踪准确度不断提升,图 9(e)的 SimAM 注意力机制对小目标和遮挡目标的提取能力都得到了改善,提高了模型整体的跟踪准确度。

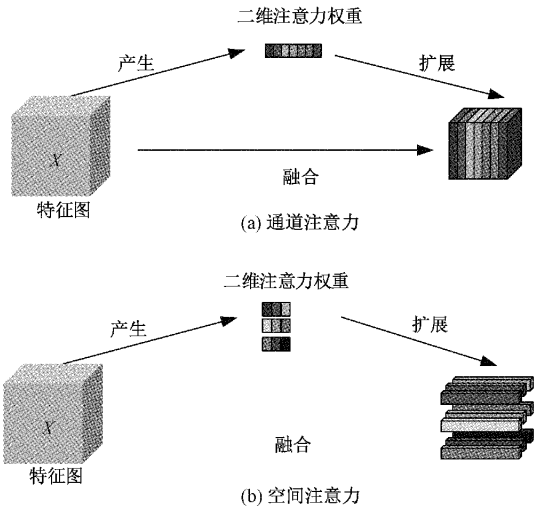
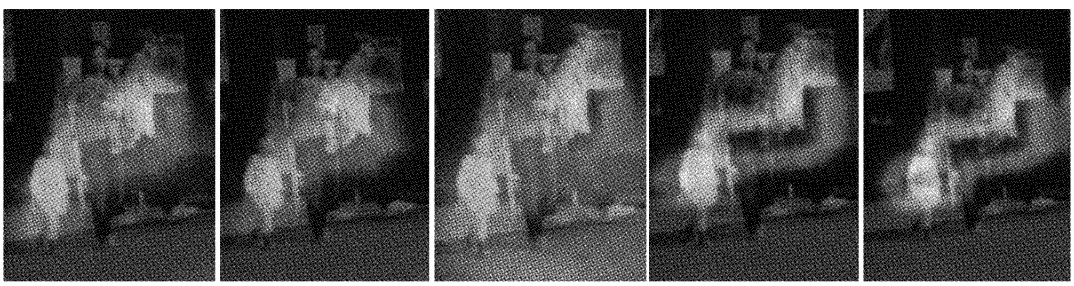


图 8　注意力对比示意图



(a) JDE　(b) +SAM　(c) +SE　(d) +CBAM　(e) +SimAM

图 9　不同注意力机制的热力图可视化

3)为了验证 SimAM 和 CIOU_Loss 的有效性,本文在 MOTA16 训练集上进行了消融实验,结果如表 3 所示。

表 3　各模块消融实验

SimAM	CIoU_LOSS	MOTA	IDs	FPS
		68.3	1 173	**21.8**
√		71.3	1 082	21.5
	√	70.1	1 214	20.9
√	√	**72.0**	1 121	20.5

JDE 算法加入 SimAM 模块后,MOTA 指标提升了 3.0 并降低了 IDs;将回归损失函数改进为 CIoU_LOSS 后,MOTA 指标提升了 1.8,结果表明两者对原 JDE 算法在 MOTA 上均有提高。由于 SimAM 无参且计算简单,对多目标跟踪速度的影响很小,而损失函数的改进也不会影响测试推理的速度,使得优化之后跟踪速度仍然保持在 20 PFS 左右,保持了较好的实时性,较好地平衡了跟踪速度和准确度。

3　结　　论

本文提出了 SAM-JDE 多目标跟踪算法,在 JDE 算法的基础上进行了两点优化:1)在骨干网络后面加入 SimAM 注意力模块,结合通道和空间维度进行了三维权重的分配,使得神经网络更加关注与目标类别物体,提升卷积网络的表达能力和定位能力,对关键区域的提取更加精准,在不增加参数量和影响实时性的前提下提升了跟踪准确性。2)在网络预测头中的回归任务分支中,将归回损失函数 Smooth_L1 改进为 CIoU_LOSS,收敛速度更快,检测框的回归准确度更高,使得目标跟踪准确度提升,且两处改进基本不会影响跟速度。通过实验结果表明,SAM-JDE 算法在实时推理的同时,跟踪准确性有了明显提升,相比于原 JDE 算法及相关改进算法在准确度和速度的平衡性之间做的更好。但嵌入分支提取的特征不够鲁棒,导致 ID 切换频繁、IDF1 指标不高的问题,下一步工作将对此部分进行探索和优化。

参考文献

[1] ZHANG Y, WANG C, WANG X, et al. Fairmot: On the fairness of detection and re-identification in multiple object tracking[J]. International Journal of Computer Vision, 2021: 1-19.

[2] REDMON J, FARHADI A. Yolov3: An incremental improvement [J]. ArXiv Preprint, 2018, ArXiv:1804.02767.

[3] 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.

[4] WOJKE N, BEWLEY A, PAULUS D. Simple online and realtime tracking with a deep association metric[C]. 2017 IEEE International Conference on Image Processing (ICIP), 2017: 3645-3649.

[5] WANG Z, ZHENG L, LIU Y, et al. Towards real-time multi-object tracking[C]. European Conference on Computer Vision, 2020: 107-122.

[6] ZHANG Y, SUN P, JIANG Y, et al. Bytetrack: Multi-object tracking by associating every detection box[C]. European Conference on Computer Vision, 2022: 1-21.

[7] SUN P, CAO J, JIANG Y, et al. Transtrack: Multiple object tracking with transformer[J]. ArXiv Preprint, 2020, ArXiv:2012.15460.

[8] CAI J, XU M, LI W, et al. MeMOT: Multi-object tracking with memory[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 8090-8100.

[9] MEINHARDT T, KIRILLOV A, LEAL-TAIXE L, et al. Trackformer: Multi-object tracking with transformers [C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 8844-8854.

[10] 晏康, 曾凤彩, 何宁, 等. 引入注意力机制的 JDE 多目标跟踪方法[J]. 计算机工程与应用, 2022, 58(21): 189-196.

[11] YANG L, ZHANG R Y, LI L, et al. Simam: A simple, parameter-free attention module for convolutional neural networks [C]. International Conference on Machine Learning, 2021: 11863-11874.

[12] GIRSHICK R. Fast r-cnn [C]. Proceedings of the IEEE International Conference on Computer Vision, 2015: 1440-1448.

[13] ZHENG Z, WANG P, LIU W, et al. Distance-IoU loss: Faster and better learning for bounding box regression[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020: 12993-13000.

[14] 刘振宇, 赵彬, 邹风山, 等. Mean-Shift 和 Kalman 算法在工件分拣技术中的应用[J]. 仪器仪表学报, 2012, 33(12): 2796-2802.

[15] 徐诚极, 王晓峰, 杨亚东. Attention-YOLO: 引入注意力机制的 YOLO 检测算法[J]. 计算机工程与应用, 2019, 55(6): 13-23.

[16] WEBB B S, DHRUV N T, SOLOMON S G, et al. Early and late mechanisms of surround suppression in striate cortex of macaque[J]. Journal of Neuroscience, 2005, 25(50): 11666-11675.

[17] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]. Proceedings of the IEEE Conference on

Computer Vision and Pattern Recognition,2018: 7132-7141.

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

[19] YU J, JIANG Y, WANG Z, et al. Unitbox: An advanced object detection network[C]. Proceedings of the 24th ACM International Conference on Multimedia,2016: 516-520.

[20] MILAN A, LEAL T L, REID I, et al. MOT16: A benchmark for multi-object tracking [J]. ArXiv Preprint,2016, ArXiv:1603.00831.

[21] ZHU X, CHENG D, ZHANG Z, et al. An empirical study of spatial attention mechanisms in deep networks [C]. Proceedings of the IEEE/CVF International Conference on Computer Vision,2019: 6688-6697.

作者简介

程之星,硕士研究生,主要从事计算机视觉,视频多目标跟踪方面的研究。

E-mail:chengzhixing_x@163.com

杨帆,博士研究生,教授,博士生导师,主要从事电子电路与计算机视觉方面的研究。

E-mail:yangfan@hebut.edu.cn