中国科技期刊卓越行动计划项目入选期刊

控制与决策

CONTROL AND DECISION

带有深度特征重平衡网络的多目标跟踪方法

郭文, 全五洲

引用本文:
郭文,全五洲. 带有深度特征重平衡网络的多目标跟踪方法[J]. 控制与决策, 2024, 39(8): 2521−2529.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2023.0213

带有深度特征重平衡网络的多目标跟踪方法

郭 文†, 全五洲

(山东工商学院 信息与电子工程学院，山东 烟台 264005)

摘 要: 针对基于联合检测嵌入范式的多目标跟踪方法中,检测与re-ID任务间冲突导致系统性能劣化的问题,首先设计一种用于富集多层语义信息并能针对不同分支倾向重构特征图的网络,有效缓解检测与re-ID分支在优化中对特征信息需求的恶性竞争;其次采用一种强化的关联策略,该策略将检测信息更深入地引入到关联流程中,旨在为更多检测结果提供关联机会,同时抑制环境干扰在关联中带来的长期损害,有效降低关联过程中误关联和漏关联的发生. 实验结果表明,所提出的方法相对于当前的先进方法展现了强大的潜力,在MOT17测试集上取得了75.7 % MOTA、73.4 % IDF1及60.0 % HOTA的性能.

关键词: 多目标跟踪；联合检测嵌入；一步式在线跟踪；数据关联

中图分类号: TP391 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2023.0213

引用格式: 郭文,全五洲. 带有深度特征重平衡网络的多目标跟踪方法[J]. 控制与决策, 2024, 39(8): 2521-2529.

Multiple object tracking method with deep feature rebalancing network

GUO Wen†, QUAN Wu-zhou

(School of Information and Electronic Engineering，Shandong Technology and Business University，Yantai 264005，China)

Abstract: Aiming at the decreasing of system performance caused by conflicts between detection and re-ID branches in the multiple object tracking method based on the joint detection and embedding paradigm, we design a network for concentrating multi-level semantic information and constructing differentiated feature maps for different branch. This network effectively relieves the vicious competition between detection and re-ID branches for feature information demands. Secondly, a modified association strategy is adopted, which introduces further information from detection branch into the association process. This strategy provides more opportunities for detections to associate while suppressing the long-term damage caused by environmental noise, effectively reducing the occurrence of misassociation and missed association. The experiments show that the method in this paper has great potential compared with the current advanced methods, and achieves the performance of 75.7 % MOTA, 73.4 % IDF1 and 60.0 % HOTA on the MOT17 test set.

Keywords: multiple object tracking；joint detection and embedding；one-shot online tracking；data association

0 引 言

多目标跟踪是计算机视觉领域举足轻重的研究方向之一. 多目标跟踪的任务是对存在某种时序关系的图像序列中特定对象组的轨迹进行描述,这有助于更多高阶任务的发展,诸如姿态与动作识别[1-2]、视频分析[3]、自动驾驶[4]等. 随着硬件水平的发展以及更多、更复杂的实际使用场景的出现,对于多目标跟踪方法性能的需求愈发迫切.

在此之前,多目标跟踪领域中已经提出了许多行之有效的方法[5-6]. 其中最具主导地位的方法被称为基于检测跟踪(track by detection, TBD),这类方法往往以一个检测器作为系统的开端,而后依据检测结果进行进一步关联.

随着深度学习的日趋流行,使用深度神经网络实现的方法逐步替换了TBD范式中各个模块,但是,相关的深度学习方法在检测与表观特征提取模块中存在重复计算[7]. 为了解决这一问题,以Wang等[8]的工作为代表的联合检测跟踪方法应运而生. 这种范式被称为联合检测嵌入(joint detection and embedding, JDE),只需进行单次特征提取就能输出检测结果和

收稿日期: 2023-02-24；**录用日期:** 2023-08-02.

基金项目: 国家自然科学基金项目(62072286, 61572296)；山东省研究生教育创新计划项目(SDYAL21211).

责任编委: 张文安.

†通讯作者. E-mail: wguo@sdtbu.edu.cn.

*本文附带电子附录文件,可登录本刊官网该文“资源附件”区自行下载阅览.

表观特征,大大降低了重复运算带来的效能损失.

然而,许多遵循该范式的方法[9-10]将“联合”一词简单地理解为平行分支. 这有利于快速地将优秀的检测架构修改成为多目标跟踪架构,但正是这种简单性为JDE方法带来了隐患. 联合跟踪器(CSTrack)直接指出:在文献[8]的工作中,检测与身份再识别(re-ID)分支间存在着不可调和的特征需求冲突[11];公平多目标跟踪器(FairMOT)则指出基于锚框的检测方法会为表观特征信息提取带来不可预测的噪声[12];关系跟踪器(RelationTrack)再次强调,即便是基于中心点的方法[12],也无法避免因为语义层级需求不同导致的分支间竞争[13].

这一问题主要源于检测与re-ID分支对于特征信息有着截然相反的需求:检测分支强调共性,它期望的特征信息是深层语义信息,能够最大化地将实体分类;而re-ID分支强调差异性,它期望的特征信息是浅层语义信息,更多的细节信息有利于其分离不同实体的嵌入向量.

虽然文献[11-13]各自提出了不同缓解特征需求冲突的方法,但是FairMOT使用的方法过于简单,它侧重于将CenterNet[14]框架引入联合检测嵌入领域;而CSTrack和RelationTrack只是在文中声称为不同的分支提供特异化的特征信息,实际上,它们提出的系统更多关注于在后端为re-ID分支提供更深的隐参数空间. 这体现在两者高度复杂的re-ID分支后处理模块,其中RelationTrack甚至单独为re-ID分支引入了Transformer[15]结构,导致与其基线方法FairMOT相比速度减损超过70 %.

在获得检测与嵌入信息后,往往会采用如交并比、运动估计及表观特征嵌入距离最优匹配等方式来完成轨迹与检测间的关联[8,16]. 然而,由于关联过程的设计理论往往基于对前向信息的绝对信任,前向错误的推理会在系统中缓慢累计,造成显著的身份跳变、漏关联或误关联.

包括但不限于上述提及的方法虽然已经在多目标跟踪任务上展现出令人赞许的效果,然而,这些方法仍然存在以下缺陷:

1) 联合检测嵌入方法普遍忽视了各个分支间对于特征信息需求的不同一性,即使部分方法意识到该问题存在,使用的改进方法也存在解耦不足或复杂度过高的问题.

2) 关联过程中无条件盲信网络输出的信息,使系统在短期和长期上都受到影响:当前帧关联前便丢弃大量检测结果,导致一些受到短时影响难检的目标被忽略,同时将所有目标一同关联,导致某些特殊情况的误检被关联;更新长期特征时,上述短期信息不加筛选地被用于更新长期特征,导致这些错误长时间地影响着关联流程.

为了缓解上述问题,本文基于FairMOT提出一种新颖的改进型方法,该方法主要贡献如下:

1) 提出一种轻量级的深度特征重平衡网络. 该网络包含一个基于空间与通道注意力的强化特征融合模块和一个特征重耦合模块. 前者在融合分级特征的同时,能够保留更丰富的语义特征;后者首先将融合特征图解耦为多个子特征图,而后重新加权耦合这些特征图以使输出特征更适应多目标优化.

2) 提出一种围绕检测得分设计的强化关联策略. 该策略的核心思想在于将检测分支可信度信息深层次地引入关联流程. 将可信度用于分区关联,更高可信度的检测结果优先被关联;在更新长期表观特征时,可信度被引入更新机制中,用以降低误关联带来的性能损失.

1 方 法

本文方法流程如图1所示.

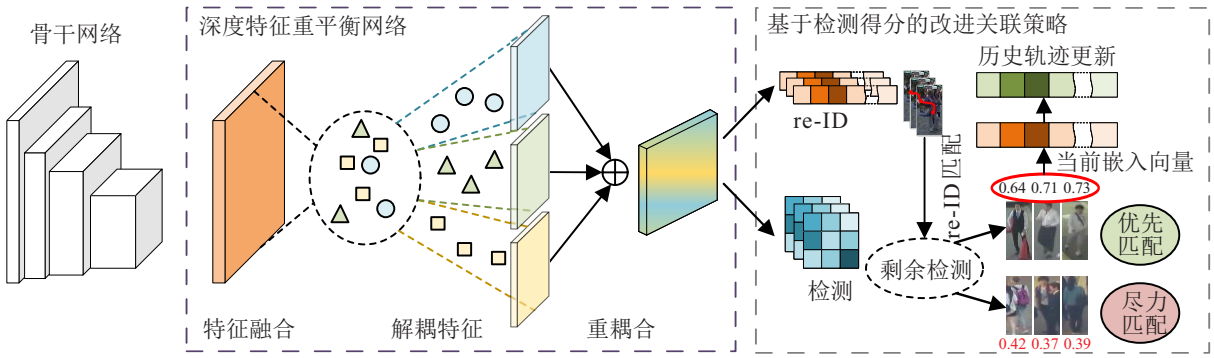


图 1 本文方法总体流程

1.1 深度特征重平衡网络

深度特征重平衡网络(DFRN)前端由骨干网络输出的金字塔特征图[17]作为输入,经由特征融合和特征重耦合两个主要模块完成图像的特征提取,而后使用数个多层感知机进行各个分支的信息推理,总体结构如图2所示.

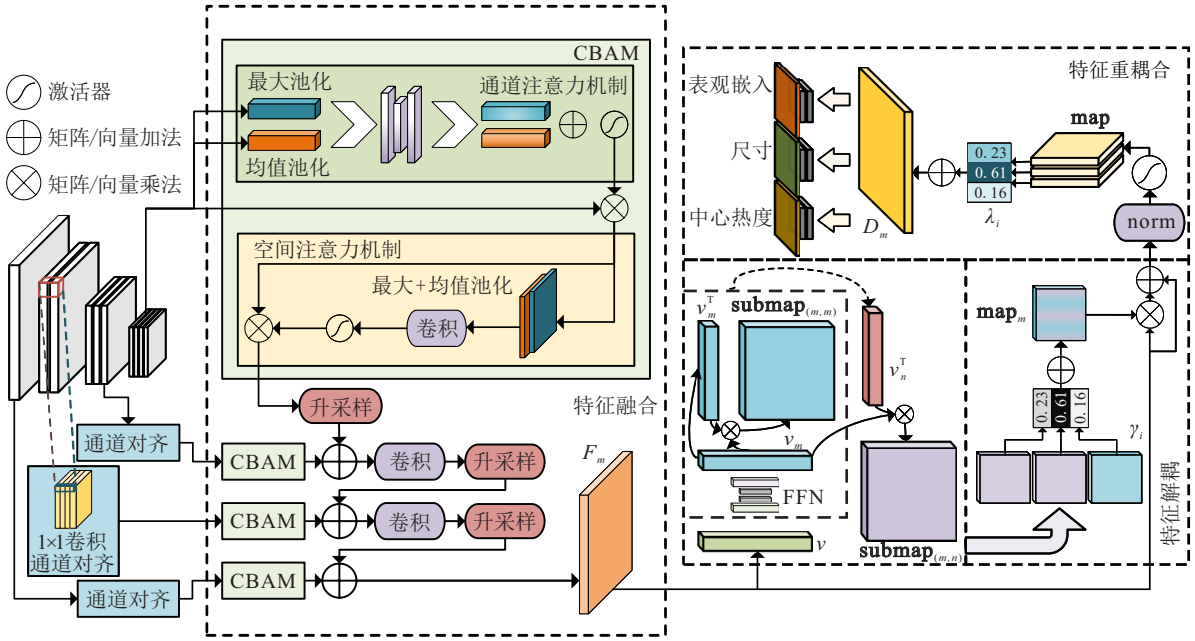


图 2　深度特征重平衡网络结构

1.1.1　特征融合

该子模块基于CenterNet[14]中使用的迭代特征融合(IDA)[18]方法,这种方法使用反卷积层进行升采样后融合,但是由此导致了潜在的语义分级特征丢失. 为了让网络能够重新关注多级性的特征,本文对主干网络输出的多尺度特征图使用一种改良的特征融合方法. 主干网络输出的金字塔特征图首先经过通道对齐,它们被对齐到相同的通道数后经过串行的通道与空间注意力模块(convolutional block attention module, CBAM)[19],然后逐个与上层升采样后的特征图累加. 升采样使用双线性插值方法,因其无参特性,需要使用一个可变卷积层[20]来重新富集融合特征图中的信息.

对于输入图像 $\boldsymbol{I} \in \mathbb{R}^{3\times 608\times 1\,088}$ 经过骨干网络提取得到的多层特征 $\{\boldsymbol{F}_k\}$,其第 k 级特征图表示为 $\boldsymbol{F}_k \in \mathbb{R}^{C_k\times H_k\times W_k}$,共计 K 级特征,本文中 $K=4$. 对于最高层级 $k=K$ 的特征图,通过一个CBAM后进行升采样得到中间特征图 $\boldsymbol{F}'_K$;对于 $1<k<K$ 的每个中间层级特征,首先需要使用 1×1 卷积层将通道数对齐至最低层级通道数 C_1,通过CBAM再与上级中间特征图 $\boldsymbol{F}'_{k+1}$ 相加,经过一个用于重新富集特征的可变卷积层[20]后再次升采样至下级中间特征图尺寸,可得第 k 层对应的中间特征 $\boldsymbol{F}'_k \in \mathbb{R}^{C_1\times H_k\times W_k}$;对于 $k=1$ 层,经过通道提升与CBAM后只需要与 $\boldsymbol{F}'_2$ 相加即获得最终的融合特征图 $\boldsymbol{F}' \in \mathbb{R}^{C_1\times H'\times W'}$. 在经上述卷积层后均采用批归一化(batch normalization)和ReLU激活器,升采样使用了双线性插值方法.

1.1.2　深度特征重耦合

由于联合检测嵌入方法的网络结构普遍由检测框架修改而来,其设计时所有输出分支对于特征信息有着相同或近似的需求. 具体而言,常见的基于中心点方法往往输出中心(类别)热度图、尺寸信息以及偏移修正. 这为需要提供表观特征嵌入向量的多目标跟踪系统带来了困难,在优化过程中,自由度过大的网络结构导致难以针对这种完全相悖的多任务进行优化. 为了解决这一问题,本文受自编码器设计结构的启发,提出深度特征重耦合网络.

对于由特征融合模块得到的融合特征图 $\boldsymbol{F}'$,首先经过全局平均池化层获得总初始向量 $\boldsymbol{v} \in \mathbb{R}^{C_1\times 1}$,再由平行的 M 个全连接网络映射获得每个自相关块的初始向量 $\boldsymbol{v}_m \in \mathbb{R}^{C_1\times 1}$. 对于每个自相关块的初始向量 $\boldsymbol{v}_m$,与另一个自相关块的初始向量 $\boldsymbol{v}_n$ 相乘,获得子强化图 $\mathbf{submap}_{(m,n)} \in \mathbb{R}^{C'\times C'}$,将所有子强化图加权平均即可获得强化图 $\mathbf{map}_m = \sum_{i=1}^{M} \gamma_i \times \mathbf{submap}_i$,将每个强化图进行归一化 $\mathrm{norm}(\,)$ 与激活 $\mathrm{activate}(\,)$ 后,与融合特征图相乘并相加,获得解耦特征图 $\boldsymbol{D}_m = \mathrm{activate}(\mathrm{norm}(\mathbf{map}_m)) \times \boldsymbol{F}_m + \boldsymbol{F}_m$. 而后,对解耦特征图再次进行加权和,即可获得重耦合特征图 $\boldsymbol{C} = \sum_{i=1}^{M} \lambda_i \boldsymbol{D}_i$.

深度特征重耦合网络是一个近似即插即用的模块,在逻辑上,它输出与输入特征完全相同的张量. 与此同时,深度特征重耦合网络也是一个轻量级的网络,在本文方法下,该网络待优化参数数量仅为

96.51 K.

1.2 基于检测得分的改进关联策略(SAS)

实际场景中,目标组及背景的状态非常复杂,利用二维平面图像进行的特征判别很难像人类神经系统一样提供准确无误的信息,检测器提供的检测结果往往存在大量的缺陷;对于数据关联而言,这些脏污的数据会极大影响关联效果. 因此,需要使用更复杂的策略以降低前端检测器不足之处带来的负面影响.

经过网络推理获得的目标,首先按照表观特征嵌入向量进行匹配,表观特征嵌入向量为一个维度为128的向量,计算该向量与各个轨迹的历史表观特征向量之间的马氏距离,以此作为代价进行二分图匹配,未关联的检测对象和轨迹被送入下一步关联.

在第2步中,检测结果会被按照检测得分(中心点热度)阈值划分为两类,高于该阈值的检测结果首先进行关联,使用卡尔曼滤波器计算剩余轨迹的当前帧预计位置,而后计算该位置与高分检测结果的交并比,以此作为代价进行二分图匹配,最后,未被匹配的检测结果与低分检测结果一同再次重复上述步骤.

在基于运动估计的关联结束后,仍可能存在少数未被关联的检测与轨迹,这时可继续按照阈值分割剩余检测结果,优先对高分检测结果组使用当前检测信息与轨迹最后一帧信息直接进行交并比计算,再继续在低分组上重复上述步骤.

经过3部分的关联后,仍未成功关联的跟踪会被视为新轨迹的开始,而未成功关联的轨迹会被挂起数帧,等待重新关联,以避免短暂的遮挡造成的目标丢失,这些挂起的轨迹不参与第3步中直接交并比关联.

对于成功关联的轨迹而言,需要对其历史表观特征向量进行更新. 许多方法中使用时间滑动均值方法[21-22]记录表观特征的缓慢变化. 然而,该方法可能会导致错误的累积. 即便每一次更新,新特征的权重始终很低,对于那些检测分支得分很低的目标而言,其可信度相对较低,因此错误的匹配可能持续地将错误的信息引入历史特征,最后引起不可逆转的错误.

为了避免这种情况的出现,基于检测得分的表观特征更新方法应该被应用至全局的特征更新中. 将检测可信度纳入更新考虑范围,对于那些高可信度的检测结果所提取的表观特征,给予更高的更新权重.

该方法具体表述为,对于第k帧中的第i个实体的检测得分为$s_{(i,k)}$,当前帧表观特征嵌入向量为$\mathrm{emb}_{(i,k)}$,其成功关联于第t个轨迹,则该轨迹在第k帧的历史表观特征向量$\boldsymbol{\mathrm{EMB}}_{(t,k)}$为

$$\boldsymbol{\mathrm{EMB}}_{(t,k)}=\frac{\mathrm{e}^{s_{(i,k)}}-1}{\mathrm{e}-1}\times\boldsymbol{\mathrm{EMB}}_{(t,k-1)},\tag{1}$$

其中e表示自然常数.

1.3 推理与训练细节

本文网络输入分辨率为1 088 × 608的RGB三通道图像,经过推理网络后输出尺寸为272 × 152的多个结果图,它们分别是中心热度图、尺寸信息、偏移信息及re-ID分支需要的表观特征嵌入向量. 中心热度图输出通道数量与目标分类数量相同,在MOT17的研究中恒定为1;尺寸信息输出通道与偏移信息输出通道数量均恒定为2;表观特征嵌入向量输出通道数量恒定为128.

检测分支包括中心热度图、尺寸信息及偏移信息. 中心热度图表示各个类别目标的几何中心,尺寸信息描述该目标对应边界框的大小,偏移信息用于修正中心点在重新映射回原图尺寸时产生的位置偏移. 检测分支的损失L_d由中心热度图损失L_{hm}和偏移与尺寸信息损失L_{os}两部分组成,可表示为

$$L_d=L_{hm}+L_{os}.\tag{2}$$

对于获得的中心热度图,需要使用非极大值抑制提取峰值,本文参照CenterNet[14]方法,选用了一个简单的3 × 3最大池化操作进行非极大值抑制. 在非极大值抑制后,使用一个较低的阈值筛选出响应较强的中心点,而后依据尺寸信息和偏移信息重建边界框. 此外,中心热度信息还会作为下一节re-ID分支表观特征嵌入向量提取的依据使用.

对于某一帧图像第i个目标而言,其边界框客观真实的空间信息可以使用向量$\boldsymbol{g}_i=(x_i^{lt},y_i^{lt},x_i^{rb},y_i^{rb})$表示. 其中:$(x_i^{lt},y_i^{lt})$表示边界框左上角在图像中的位置,$(x_i^{rb},y_i^{rb})$表示边界框右下角在图像中的位置. 由此可得目标边界框中心$(c_x'^i,c_y'^i)=\left(\dfrac{x_i^{lt}+x_i^{rb}}{2},\dfrac{y_i^{lt}+y_i^{rb}}{2}\right)$. 由于融合并重耦合后的特征图$\boldsymbol{F}'$在分辨率的两个维度上均为原图的0.25倍,在训练过程中,将目标中心点的客观真实位置规定为$(c_x^i,c_y^i)=\left(\left\lfloor\dfrac{c_x'^i}{4}\right\rfloor,\left\lfloor\dfrac{c_y'^i}{4}\right\rfloor\right)$. 该点的中心热度响应值表示为

$$M_{(x,y)}=\sum_{N}^{i=1}\mathrm{e}^{-\frac{(x-c_x^i)^2+(y-c_y^i)^2}{2\sigma_c^2}}.\tag{3}$$

其中:N表示当前帧中目标总数,σ表示标准差.

中心热度图损失L_{hm}的计算即对逐像素使用焦点损失(focal loss[23])进行逻辑回归,具体表示为

$$L_{hm}=-\frac{1}{N}\sum_{xy}\begin{cases}D^{\alpha}\log D,\ M_{xy}=1;\\ D^{\beta}\hat{M}_{xy}^{\alpha}\log D,\ M_{xy}\neq 1.\end{cases}\tag{4}$$

其中: $\hat{M}_{xy}$ 表示对于 (x,y) 点相应的估计值,即由网络输出获得的响应值; $D=1-\hat{M}_{xy}$; α 和 β 表示焦点损失中使用的超参数.

如上文所述,推理中心点热度图时,由于图像与输入的特征图在长宽两个维度上各自有着4倍的差距,导致中心点的输出包含对应网络输入图像多达 4×4 个像素的位置. 为了解决这一问题,需要使用一个输出头估计推理出的中心点与原图像中实际位置的位移,以获得更精确的位置信息. 而尺寸信息头则负责输出边界框的长宽信息.

对于某一帧图像中第 i 个目标而言,它的尺寸 $\boldsymbol{s}_i$ 可以表示为 $\boldsymbol{s}_i=(x_i^{rb}-x_i^{lt},y_i^{rb}-y_i^{lt})$,而偏移可以表示为 $\boldsymbol{o}_i=(c_x'^i-c_x^i,c_y'^i-c_y^i)$. 对于由网络推理获得的尺寸 $\hat{\boldsymbol{s}}_i$ 和偏移 $\hat{\boldsymbol{o}}_i$,使用 l_1 损失计算其与客观真值间的差距,损失 L_{os} 为

$$L_{os}=\sum_{i=1}^{N}\|\boldsymbol{o}_i-\hat{\boldsymbol{o}}_i\|_1+\lambda\|\boldsymbol{s}_i-\hat{\boldsymbol{s}}_i\|_1. \tag{5}$$

其中: N 表示当前帧中目标总数; λ 表示一个超参数权重,按照CenterNet[14]的实验,其被设置为0.1.

re-ID分支网络部分输出一个特征向量图 $\boldsymbol{E}\in\boldsymbol{R}^{128\times608\times1\,088}$,而后,可根据检测分支输出的中心点 (x,y) 提取表观特征嵌入向量 $\boldsymbol{E}_{(x,y)}\in\mathbb{R}^{128}$. 该特征向量需要具有能够区分不同目标身份的作用,理想情况下,同一个目标的检测对应的特征嵌入向量要尽可能地相似,反之则应该尽可能地差异化. 因此,re-ID分支的工作逻辑上可以被视为一个分类任务: 每一个目标对应的所有检测被视为一个分类,使类间差距尽可能大,类内差距尽可能小.

对于位于估计中心点 $(\hat{x},\hat{y})$ 上的当前帧中的第 i 个目标而言,其表观特征嵌入向量 $\boldsymbol{E}_{(\hat{x},\hat{y})}$ 经由一个全连接网络与Softmax层被映射至为一个分布向量 $\boldsymbol{C}_i$, $\boldsymbol{C}_i$ 中共包含 K 个元素,其中第 k 个元素 $c(k)$ 逻辑意义上表示该向量归属于第 k 个身份的可能性. 身份的客观真实值由一个独热编码矩阵 $\boldsymbol{G}$ 表示,其维度为 $K\times N$,定义 $\boldsymbol{G}(n,k)$ 为当前帧数第 n 个目标是否为第 k 个身份的标记,若是则为1,否则为0. 则re-ID分支的损失 L_r 可以表示为

$$L_r=-\sum_{i=1}^{N}\sum_{t=1}^{K}\boldsymbol{G}(i,k)\log(c(t)). \tag{6}$$

为了能够联合优化检测与re-ID任务,并且确保在训练过程中更有效地平衡两种任务,参照文献[24]中多任务学习中不确定性损失的思路,最终将本文的系统总损失 L 定义为

$$L=\frac{1}{2}\Big(\frac{L_d}{\mathrm{e}^{\mu_d}}+\frac{L_r}{\mathrm{e}^{\mu_r}}+\mu_d+\mu_r\Big). \tag{7}$$

其中: μ_d 和 μ_r 表示可学习参数; L_d 表示检测分支总体损失,它评估了检测分支输出的中心热度、尺寸及偏移信息与客观真值的差距; L_r 表示re-ID分支表观特征嵌入向量输出的损失函数.

2　实验结果与分析

2.1　数据集与评价标准

本文主要使用的评估基准数据集为MOT Challenge中的MOT17[25]和MOT20[26]. 此外,在训练过程中,本文引入了额外的6个数据集.

补充数据集主要被用于额外训练以增强模型的泛化性,避免过拟合,其中包括6个数据集,共分为两类: 其一由无关的单张图片组成,所有目标均被视为一个独立的身份进行训练; 其二包括边界框和身份信息的标注. CrowdHuman[27]、ETH[28]和City Persons[29]隶属于第1类, Caltech[30]、 Pedestrian[31]、CUHK-SYSU[9]、PRW[32]隶属于第2类. 由于ETH中部分数据与MOT17验证集重复,重复部分在实验训练中已经被剔除以确保测试集结果的公平性.

本文使用CLEAR MOT[33]、IDF1以及HOTA[34]作为系统综合性能评价指标. MOTA是一种评价错检、漏检和身份跳变缺陷的指标; IDF1则更侧重于关注身份准确性,它综合评估了系统身份分配结果的准确性及全面性; 需要指出的是, HOTA是一种更全面的评价指标,与MOTA和IDF1相比, HOTA进一步平衡了检测性能与关联性能、查全性与查准性之间的偏颇.

2.2　实验细节

本文采用DLA-34[14]作为骨干网络原型,并应用FairMOT[12]中提供的在CrowdHuman[27]上预训练的权重作为初始权重. 系统使用AdamW作为优化器共训练40轮,初始学习率设置为2e-4,并且在第20轮时降低为2e-5,训练批次大小被设定为8. MOT数据集中图像的默认分辨率为1 920×1 080,考虑到效率和内存问题,所有输入的分辨率都被缩放并填充为1 088×608. 此外,训练过程中对训练集启用了一些常见的数据增强技术,包括随机旋转、随机尺寸和颜色随机变换,以减少过拟合. 实验内容不包括最优参数的网格搜索,这意味着可以进行更多的参数微调以达到更好的效果.

本文实验训练过程使用Intel Xeon Silver 4210 CPU和单块Nvidia Tesla V100 GPU,整个训练过程

需要约30 h及21 G字节的显卡内存空间. 由于显卡性能对于推理速度(IS)的影响十分显著,为了该指标比较的公平性,在推理阶段,GPU被更换为Nvidia Geforce RTX Titan,它与Nvidia Geforce RTX 2080Ti性能相当,表1和表2列出的大多数实验均使用该级别的显卡评估IS指标. 其中: TrackFormer为注意力跟踪器,GSDT为联合检测与跟踪的图跟踪器,TransCenter为中心注意力跟踪器,CTracker为链式跟踪器,DeepSORT为深度简单实时跟踪器,MLT为多重标记图跟踪器.

表 1 本文方法与其他先进方法在MOT17测试集上的性能对比

方法	HOTA↑	MOTA↑	IDF1↑	MT↑	ML↓	FP↓	FN↓	ID Sw.↑	IS/Hz↑
RelationTrack[13]	**61.0**	74.7	**73.8**	41.70	23.20	27 999	118 623	1 374	6.6
CSTrack[11]	59.3	74.9	72.6	41.53	17.45	23 847	114 303	3 567	15.8
FairMOT[12]	59.3	73.7	72.3	43.18	17.32	27 507	117 477	3 303	**25.9**
TrackFormer[35]	57.3	74.1	68.0	**47.26**	**10.45**	34 602	**108 777**	2829	5.7
GSDT[36]	55.2	73.2	66.5	41.66	17.45	26 397	120 666	**1 297**	4.9
TransCenter[37]	54.5	73.2	62.2	40.76	18.47	23 112	123 738	1 598	1
CenterTrack[38]	52.2	67.8	64.7	34.65	24.59	**18 498**	160 332	3 039	3.8
CTracker[39]	49.0	66.6	57.4	32.19	24.20	22 284	160 491	5 529	6.8
JDE [8]	–	63.0	59.5	35.70	17.30	39 888	162 927	–	18.8
DeepSORT [7]	–	61.4	–	32.80	18.20	38 556	170 004	–	40
文献[40]	–	33.4	–	18.90	–	–	–	–	7.1
文献[41]	–	23.5	–	9.50	52.20	43 188	208 785	–	0.6
Ours	60.0	**75.7**	73.4	41.15	14.39	28 683	111 210	2 375	17.9

表 2 本文方法与其他先进方法在MOT20测试集上的性能对比

方法	HOTA↑	MOTA↑	IDF1↑	MT↑	ML↓	FP↓	FN↓	ID Sw.↑	IS/Hz↑
RelationTrack[13]	**55.1**	61.8	67.9	62.2	8.9	112 927	**85 062**	4 243	2.7
FairMOT[12]	54.6	61.8	67.3	**68.8**	**7.6**	103 440	88 901	5 243	**13.2**
CSTrack[11]	54.0	66.6	68.6	50.4	15.5	**25 404**	144 358	3 196	4.5
TransCenter[37]	43.5	58.5	49.6	48.6	14.9	64 217	14 6019	4 695	8.8
MLT[42]	43.2	48.9	54.6	30.9	22.1	45 660	216 803	**2 187**	3.7
Ours	54.9	**67.4**	**68.7**	52.71	16.92	31 572	136 935	4 392	9.6

2.3 对比实验

在本章节中,将对本文方法与其他先进方法进行性能对比分析. 表1展示了本文方法在MOT17测试集上与其他使用私有检测器的先进方法的关键性能指标对比,表2展示了本文方法在MOT20测试集上与其他使用私有检测器的先进方法的关键性能指标对比. 在表1和表2中,各指标最优项已使用粗体标注.

本文方法在MOTA、IDF1和HOTA上均展现了优秀的效果. 在MOT17上: 与本文基线方法FairMOT相比取得了+2.0 MOTA、+1.1 IDF1和+0.7 HOTA的提升; 与基于相似理论、但基于锚框的方法CSTrack相比,本文取得+0.8 MOTA、+0.8 IDF1和+0.7 HOTA的提升; 与同样基于FairMOT、并且与本文研究方向相仿的RelationTrack方法相比,本文仍然有着+1.0的MOTA提升. 与RelationTrack相比,本文方法的性能差距主要集中在身份分配上,其主要原因在于RelationTrack中为re-ID分支单独引入的Transformer的结构复杂、参数量大. 这一点在其与CSTrack对比时也尤为明显: 虽然RelationTrack与CSTrack相比有着+1.7 HOTA和+1.2 IDF1的优势,但是在MOTA上却有着−0.2的劣势. 在MOT20上,本文方法与基线方法FairMOT相比取得了+0.3 HOTA、+5.6 MOTA以及+1.0 IDF1的优势.

在MOT17上,本文方法的MT和FP指标相较于基线方法FairMOT略有下降,这是因为本文使用的SAS策略相比FairMOT使用的关联策略引入了更多的检测结果; 这一点在ML和ID Sw.指标中得以验证,其中ML指标在对比中排行第2,显著高于基线方法FairMOT, ID Sw.也获得了显著的提升,这验证了引入更多的低分检测结果并设计更合理的更新策略有助于更多地为每一个目标分配到正确的身份. 此外,由于更多低可信度的检测结果的引入,FN指标相较于基线方法FairMOT也有少许提升. 在MOT20上,MT和ML指标相较于FairMOT都较低,但是结合其FN较低同时FP极高的特点,可以推测FairMOT在MOT20上的实验使用了一个极低的检测结果可信度阈值,导致大量结果引入,这也间接导致了ID Sw.的劣化. 本文方法的FP指标相对较低,这是因为本文方法虽然在设计时引入了更多的检测结果,但是通过

SAS 策略, 本文方法能够有效地过滤 FP 结果.

此外, 本文方法在推理速度上也具有优势, 相较于基线方法 FairMOT, 本文提出的方法推理速度仅降低了 30.89 %. RelationTrack 相较本文方法整体速度降低约 63.12 %; CSTrack 相较本文方法整体速度降低约 11.73 %. 虽然本文方法仍未达到实时性的标准 (>24 fps), 但是本文方法在速度和精度上取得了较好的平衡, 且相较于其他多目标跟踪方法, 本文方法在速度上有着明显的优势.

2.4　消融实验

本节将深入研究各个模块的实际效用. 由于 MOT Challenge 每种方法在每个数据集仅允许提交 4 次跟踪结果, 为了能够进行更多的消融实验, 本节模型基于 MOT17-half 训练集, 并且在 MOT17-validation 上进行测试. 这是一种在多目标研究领域被广泛采用的分割方法: 将 MOT17 训练集分为等量的两部分, 即每段帧序列分割为等帧数的两个集合, 前一半被用于训练, 记为 MOT17-half, 后一半被用于验证, 记为 MOT17-validation.

表 3 展示了各个模块单独运用在系统中产生的效果, 其中首行可等价视为 FairMOT 方法.

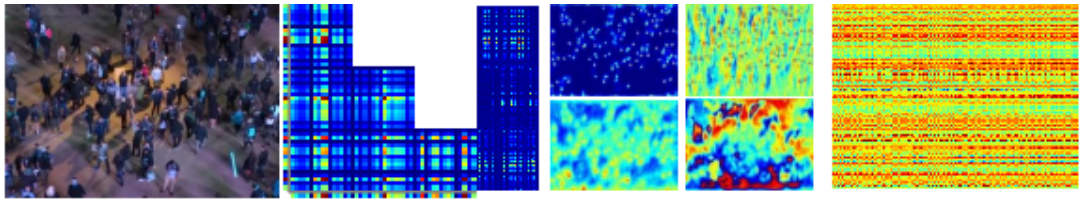
表 3　在 MOT17 验证集上的消融实验性能对比

DFRN	SAS	MOTA	IDF1	HOTA
–	–	69.4	72.4	56.3
–		69.8	74.5	57.1
	–	71.2	73.0	57.3
		71.3	74.3	59.8

由实验数据可见, DFRN 能够提升 1.8 % MOTA、0.6 % IDF1 和 1.0 % HOTA, SAS 能够提升 0.4 % MOTA、2.1 % IDF1 和 3.5 % HOTA. 这表明, DFRN 和 SAS 均能够有效提升系统的性能, 其中 DFRN 更注重 MOTA 侧的提升, 它有效抑制了错误的检测结果的过度采信导致的身份跳变和框偏移, 而 SAS 明显在身份关联中提供了有益的帮助.

2.5　可视化实验

图 3 展示了在高度稠密场景下的各关键环节上可视化的展现, 该段帧序列来自于 MOT20[26] 数据集测试集; 模型在上文基础上在 MOT20 训练集上进行了 20 轮的精调.



(a) 原图像　(b) 解耦特征　(c) 重耦合并推理　(d) 表观特征嵌入向量

(e) 关联结果

图 3　在高度稠密场景下的各关键环节上可视化的展现

图3(a)展示了原始图像;图3(b)展示了特征解耦过程中3个自相关块的初始向量$\boldsymbol{v}_m$(为了更好的视觉展示,初始向量使用尺度步幅均为4的均值池化算子处理,子强化图根据池化后向量计算,强化图展示了网络的实际输出效果)、 共计6幅子强化图$\mathbf{submap}_{(m,n)}$以及3幅强化图$\mathbf{map}_m$;图3(c)展示了中心热度图(左上)、偏移(右上)、边界框尺寸(左下)以及re-ID(右下)输出信息的热力图;图3(d)展示了当前帧表观特征向量可视化展示,其尺寸为132×128,横向表示目标序号,纵向表示嵌入向量维度;图3(e)展示了关联结果可视化.

3 结 论

本文提出了一种能够缓解当前多目标跟踪部分缺陷的应用深度特征重平衡网络及改进关联策略的基于联合检测嵌入的在线多目标跟踪算法. 实验表明,本文方法展现了令人满意的效果. 具体而言,本文的系统关注联合检测嵌入方法中各个分支间存在难以调和特征需求冲突的问题,并针对这一问题提出了深度特征重平衡网络,使用一个带有特征融合-解耦-重耦合的网络降低了这种冲突,提升了系统性能;此外,本文为关联过程引入了更多的检测信息,设计了围绕检测得分进行的新颖策略,在关联和更新两大重要步骤中改善了关联过程固有的缺陷,提升了关联成功率,这归功于关联过程中有效地抑制了误关联和漏关联发生次数. 本文提出的方法仍有改进空间,如更深入探讨分支间潜在的复杂关系等,或许能够成为未来研究的重要一环.

参考文献(References)

[1] Pfister T, Charles J, Zisserman A. Flowing ConvNets for human pose estimation in videos[C]. 2015 IEEE International Conference on Computer Vision. Santiago, 2016: 1913-1921.

[2] Choi W, Savarese S. A unified framework for multi-target tracking and collective activity recognition[C]. European Conference on Computer Vision. Berlin, 2012: 215-230.

[3] Takahashi N, Gygli M, Van Gool L. AENet: Learning deep audio features for video analysis[J]. IEEE Transactions on Multimedia, 2018, 20(3): 513-524.

[4] Luo W J, Yang B, Urtasun R. Fast and furious: Real time end-to-end 3D detection, tracking and motion forecasting with a single convolutional net[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 3569-3577.

[5] Luo W H, Xing J L, Milan A, et al. Multiple object tracking: A literature review[J]. Artificial Intelligence, 2021, 293: 103448.

[6] 朱姝姝, 王欢, 严慧. 基于帧内关系建模和自注意力融合的多目标跟踪方法[J]. 控制与决策, 2023, 38(2): 335-344.
(Zhu S S, Wang H, Yan H. Multi-object tracking based on intra-frame relationship modeling and self-attention fusion mechanism[J]. Control and Decision, 2023, 38(2): 335-344.)

[7] Wojke N, Bewley A, Paulus D. Simple online and realtime tracking with a deep association metric[C]. 2017 IEEE International Conference on Image Processing. Beijing, 2018: 3645-3649.

[8] Wang Z D, Zheng L, Liu Y X, et al. Towards real-time multi-object tracking[C]. European Conference on Computer Vision. Cham, 2020: 107-122.

[9] Xiao T, Li S, Wang B C, et al. Joint detection and identification feature learning for person search[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, 2017: 3376-3385.

[10] Lu Z C, Rathod V, Votel R, et al. RetinaTrack: Online single stage joint detection and tracking[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, 2020: 14656-14666.

[11] Liang C, Zhang Z P, Zhou X, et al. Rethinking the competition between detection and ReID in multiobject tracking[J]. IEEE Transactions on Image Processing, 2022, 31: 3182-3196.

[12] Zhang Y F, Wang C Y, Wang X G, et al. FairMOT: On the fairness of detection and re-identification in multiple object tracking[J]. International Journal of Computer Vision, 2021, 129(11): 3069-3087.

[13] Yu E, Li Z L, Han S D, et al. RelationTrack: Relation-aware multiple object tracking with decoupled representation[J]. IEEE Transactions on Multimedia, 2023, 25: 2686-2697.

[14] Zhou X Y, Wang D Q, Krähenbühl P. Objects as points[J/OL]. 2019, arXiv: 1904.07850.

[15] Ashish V, Noam S, Niki P, et al. Attention is all you need[C]. Proceeding of the Advances in Neural Information Processing Systems. LongBeach, 2017: 5998-6008.

[16] Chen L, Ai H Z, Zhuang Z J, et al. Real-time multiple people tracking with deeply learned candidate selection and person re-identification[C]. 2018 IEEE International Conference on Multimedia and Expo. San Diego, 2018: 1-6.

[17] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, 2017: 936-944.

[18] Yu F, Wang D Q, Shelhamer E, et al. Deep layer aggregation[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City,

2018: 2403-2412.

[19] Woo S, Park J, Lee J Y, et al. CBAM: Convolutional block attention module[C]. European Conference on Computer Vision. Cham, 2018: 3-19.

[20] Dai J F, Qi H Z, Xiong Y W, et al. Deformable convolutional networks[C]. 2017 IEEE International Conference on Computer Vision. Venice, 2017: 764-773.

[21] Bolme D S, Beveridge J R, Draper B A, et al. Visual object tracking using adaptive correlation filters[C]. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco, 2010: 2544-2550.

[22] Henriques J F, Caseiro R, Martins P, et al. High-speed tracking with kernelized correlation filters[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(3): 583-596.

[23] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]. 2017 IEEE International Conference on Computer Vision. Venice, 2017: 2999-3007.

[24] Cipolla R, Gal Y, Kendall A. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 7482-7491.

[25] Milan A, Leal-Taixe L, Reid I, et al. MOT16: A benchmark for multi-object tracking[J/OL]. 2016, arXiv: 1603.00831.

[26] Dendorfer P, Rezatofighi H, Milan A, et al. MOT20: A benchmark for multi object tracking in crowded scenes[J/OL]. 2020, arXiv: 2003.09003.

[27] Shao S, Zhao Z J, Li B X, et al. CrowdHuman: A benchmark for detecting human in a crowd[J/OL]. 2018, arXiv: 1805.00123.

[28] Ess A, Leibe B, Schindler K, et al. A mobile vision system for robust multi-person tracking[C]. 2008 IEEE Conference on Computer Vision and Pattern Recognition. Anchorage, 2008: 1-8.

[29] Zhang S S, Benenson R, Schiele B. CityPersons: A diverse dataset for pedestrian detection[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, 2017: 4457-4465.

[30] Gregory G, Alex H, Pietro P. Caltech-256 object category dataset[EB/OL]. (2016-05-03)[2019-06-20]. https://arxiv.org/abs/1603.00831.

[31] Dollar P, Wojek C, Schiele B, et al. Pedestrian detection: A benchmark[C]. 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, 2009: 304-311.

[32] Zheng L, Zhang H H, Sun S Y, et al. Person re-identification in the wild[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, 2017: 3346-3355.

[33] Bernardin K, Stiefelhagen R. Evaluating multiple object tracking performance: The CLEAR MOT metrics[J]. Journal on Image and Video Processing, 2008, 2008: 1-10.

[34] Luiten J, Osep A, Dendorfer P, et al. HOTA: A higher order metric for evaluating multi-object tracking[J]. International Journal of Computer Vision, 2021, 129(2): 548-578.

[35] Meinhardt T, Kirillov A, Leal-Taixé L, et al. TrackFormer: Multi-object tracking with transformers[C]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, 2022: 8834-8844.

[36] Wang Y X, Kitani K, Weng X S. Joint object detection and multi-object tracking with graph neural networks[C]. 2021 IEEE International Conference on Robotics and Automation. Xi'an, 2021: 13708-13715.

[37] Xu Y H, Ban Y T, Delorme G, et al. TransCenter: Transformers with dense representations for multiple-object tracking[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(6): 7820-7835.

[38] Zhou X Y, Koltun V, Krähenbühl P. Tracking objects as points[C]. European Conference on Computer Vision. Cham, 2020: 474-490.

[39] Peng J L, Wang C G, Wan F B, et al. Chained-tracker: Chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking[C]. European Conference on Computer Vision. Cham, 2020: 145-161.

[40] 刘洪彬, 常发亮, 刘春生, 等. 基于时空渐进特征模型的抗遮挡多目标跟踪[J]. 控制与决策, 2019, 34(10): 2171-2177.
(Liu H B, Chang F L, Liu C S, et al. Anti-occlusion multi-target tracking with progressive spatio-temporal feature model[J]. Control and Decision, 2019, 34(10): 2171-2177.)

[41] 薛俊韬, 马若寒, 胡超芳. 基于MobileNet的多目标跟踪深度学习算法[J]. 控制与决策, 2021, 36(8): 1991-1996.
(Xue J T, Ma R H, Hu C F. Deep learning algorithm based on MobileNet for multi-target tracking[J]. Control and Decision, 2021, 36(8): 1991-1996.)

[42] Zhang Y, Sheng H, Wu Y B, et al. Multiplex labeling graph for near-online tracking in crowded scenes[J]. IEEE Internet of Things Journal, 2020, 7(9): 7892-7902.

作者简介

郭文(1978–), 男, 教授, 博士, 从事计算机视觉、多媒体计算等研究, E-mail: wguo@sdtbu.edu.cn;

全五洲(1996–), 男, 硕士生, 从事计算机视觉的研究, E-mail: 2020420008@sdtbu.edu.cn.