智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

抗遮挡的行人多目标跟踪算法

张国印, 王传博, 高伟

引用本文:
张国印, 王传博, 高伟. 抗遮挡的行人多目标跟踪算法[J]. 智能系统学报, 2024, 19(5): 1248-1256.
ZHANG Guoyin, WANG Chuanbo, GAO Wei. Pedestrian multiobject tracking algorithm with anti-occlusion[J]. *CAAI Transactions on Intelligent Systems*, 2024, 19(5): 1248-1256.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202307002

DOI: 10.11992/tis.202307002
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20240828.1350.028

抗遮挡的行人多目标跟踪算法

张国印, 王传博, 高伟
(哈尔滨工程大学 计算机科学与技术学院, 黑龙江 哈尔滨 150001)

摘　要: 为了解决在复杂场景下行人相互遮挡导致跟踪系统精度降低的问题, 提出了基于 FairMOT 的抗遮挡多目标跟踪算法 (multiple obeject tracking algorithm with anti-occlusions, AOMOT)。首先通过轻量化平衡模块, 解耦不同层次的语义信息, 减少检测任务和重识别任务的语义冲突, 降低重识别任务的性能提升对检测任务的影响。其次应用自注意力结构提取行人的外观特征, 加强局部窗口下的类内特征的区分度, 增强行人身份信息的匹配一致性并减少身份标识的频繁切换。最后优化身份关联算法, 挖掘低置信度目标中的被遮挡对象, 将其重新纳入目标身份关联并更新其重识别特征。实验结果表明, AOMOT 相比原有 FairMOT 在 MOT17 数据集中高阶跟踪精度提升 1.5 百分点, 身份 F1 分数提升 3 百分点, 身份切换数量降低 32%。

关键词: 计算机视觉; 行人跟踪; 目标检测; 重识别; 关联算法; 抗遮挡; 自注意力; 特征提取
中图分类号: TP391　　**文献标志码:** A　　**文章编号:** 1673−4785(2024)05−1248−09

中文引用格式: 张国印, 王传博, 高伟. 抗遮挡的行人多目标跟踪算法 [J]. 智能系统学报, 2024, 19(5): 1248–1256.
英文引用格式: ZHANG Guoyin, WANG Chuanbo, GAO Wei. Pedestrian multiobject tracking algorithm with anti-occlusion[J]. CAAI transactions on intelligent systems, 2024, 19(5): 1248–1256.

Pedestrian multiobject tracking algorithm with anti-occlusion

ZHANG Guoyin, WANG Chuanbo, GAO Wei
(College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China)

Abstract: A multiobject tracking algorithm with anti-occlusion (referred to as AOMOT), which is based on the FairMOT framework, is proposed to improve the accuracy of tracking systems in crowded pedestrian scenes. First, the lightweight balance module decouples semantic information at different levels to minimize semantic conflicts between detection and recognition tasks and decrease the impact of performance improvement in re-identification tasks. Second, the self-attention structure is adopted to extract pedestrian appearance features and improve the discrimination of intra-class features under local windows. The matching consistency of pedestrian identity information is enhanced, and frequent switching of identity signs is reduced. Finally, the identity association algorithm is optimized to mine occluded objects in low-confidence targets, reincorporate them into the target identity association, and update their recognition features. Experimental results show that, compared with the original model in the MOT17 dataset, the improved model enhances the higher-order tracking accuracy by 1.5 percentage points, improves identity F1 score by 3 percentage points, and reduces identity switching by 32%.

Keywords: computer vision; pedestrian tracking; target detection; re-identification; association algorithm; anti-occlusion; self-attention; feature extraction

随着计算机视觉技术的快速发展, 多目标跟踪 (multi object tracking, MOT) 作为计算机视觉领域的一项关键技术, 是姿态估计[1]、行为识别[2]、行为分析[3] 等高级计算机任务的基础, 是近些年计算机视觉研究的热点方向之一。在实际应用方面, 多目标跟踪系统在自动驾驶、智能监控、人机

收稿日期: 2023−07−10.　**网络出版日期:** 2024−08−29.
通信作者: 高伟. E-mail: gaowei@hrbeu.edu.cn.

交互、车辆监控[4] 等领域也发挥着越来越重要的作用。然而阴影、遮挡、运动模糊等许多因素对算法的鲁棒性提出了挑战。特别是在行人密集的场景下, 行人在运动的过程中会发生频繁的遮挡。同时人体作为非刚性物体, 运动的过程中也会发生外观和形态上的不规则变化。因此减少这些因素造成的漏检和误检, 提升抗遮挡跟踪的精确度成为研究热点。

多目标跟踪系统[5] 在视频序列的第一帧中检测需要跟踪的多个目标, 并在后续视频帧中对同属一个身份的目标进行关联以此形成其运动轨迹。目前多目标跟踪算法主要分为 2 个部分, 即目标检测和目标关联。其中目标检测需要确定当前帧中目标的位置以获得相应的运动信息或外观信息, 而目标关联阶段是将当前帧和历史帧中同一目标分配相同的身份编号, 维持跟踪轨迹的一致性。Zhou 等[6] 提出的 GTR(global tracking transformer) 基于 Transformer 结构的查询机制, 以概率的方式将一个短时间视频内的对象分组为轨迹从而避免了成对关联。但是需要一个时间窗口数量的视频帧完成查询对象的初始化。Sun 等[7] 提出的 TransTrack 通过 2 个并行 Transformer 结构在检测目标的同时进行目标跟踪, 采用交并比匹配的方法将检测与跟踪关联起来。Meinhardt 等[8] 提出的 TrackFormer 是跟踪模型利用 Transformer 结构的查询机制隐式地在检测的同时在前后帧之间传递跟踪目标的轨迹标识, 端到端地到跟踪轨迹。这些模型将目标的检测和关联结合在一个模型中, 难以针对行人相互遮挡的情况进行有效改进, 并且模型的训练也需要较多计算资源。在将检测和数据关联分开处理的模型中, Wu 等[9] 提出的 TraDes(track to detect and segment) 利用 2 帧目标之间的匹配相似度推断跟踪偏移, 利用位置偏移进行 2 帧的短程关联, 利用重识别特征的余弦相似度进行长程关联, 其在低帧率视频中目标位移过大时效果不佳。在 Zhou 等[10] 提出的 CenterTrack 中, 对象被表示为中心点, 通过将过去的对象位置回归到当前帧中的新位置来完成跟踪, 通过基于距离的贪婪匹配算法进行关联。该方法只考虑了连续帧的对象关联, 没有重新初始化丢失的轨迹, 无法进行长程跟踪。Hyun 等[11] 提出的稀疏图跟踪器 (sparse graph tracker, SGT), 利用高阶关系特征进行跟踪, 通过聚合相邻检测特征及其关系, 生成具有更强的鉴别性的高阶关系特征, 并利用高级关系特征进行追踪。其在行人被非行人物体遮挡时, 聚合低维特征能力降低导致身份切换次数增加, 跟踪准确度下降。

FairMOT[12] 的核心思想是在单个网络中同时完成目标检测和身份特征提取, 通过共享大部分计算来减少推理时间。FairMOT 使用深度聚合 (deep layer aggregation, DLA) 网络作为骨干网络提取输入图像的特征表示, 将输入图像转化为高分辨率特征图。检测器使用无锚框的 CenterNet[13] 来估计高分辨率特征图上的目标中心和位置, 以减少使用锚框提取特征产生的歧义, 同时添加并行分支来抽取重识别特征用于预测目标的身份标识。目标身份关联则结合外观和位置信息计算相似度矩阵, 采用 DeepSORT[14] 算法进行级联匹配以保证可以进行长程关联。

综上, 为了解决多目标跟踪系统在行人密集、遮挡严重等场景下产生大量漏检和误检导致跟踪精度下降的问题, 提出以 FairMOT 为基础的抗遮挡多目标跟踪算法 (multiple obeject tracking algorithm with anti-occlusions, AOMOT)。

首先, 在主干网络提取高分辨率特征图后增加一个轻量化平衡模块, 平衡不同层次特征在目标检测和重识别中的语义特征, 降低由于检测和重识别冲突对模型造成的影响。

其次, 在重识别中引入了窗口注意力机制来加强局部窗口下不同目标之间的差异特征的提取, 从而生成更具辨别力的表观特征。减少跟踪目标过程中身份标识的频繁切换。

最后, 对关联算法进行了优化。特别是在低置信度检测中, 挖掘被遮挡的对象, 并将它们重新纳入目标身份关联, 同时更新它们的重识别特征, 促使目标轨迹得到正确关联匹配, 从而提升在严重遮挡情况下的目标跟踪效果。

1　模型结构

AOMOT 遵循与 FairMOT 相同的联合检测和追踪的范式, 图像输入骨干网络进行编码和解码以提取高分辨率特征图。然后通过轻量化平衡网络 (lightweight balanced network, LBN) 平衡主干网络的语义信息, 分别为目标检测任务和重识别特征提取任务生成特征图。目标检测分支使用卷积块注意力模块 (convolutional block attention module, CBAM) 增强生成中心点热力图及其他相关位置信息 (包括目标的长宽和中心点偏移量) 的鲁棒性。重识别特征提取分支使用窗口注意力网络

(window attention network, WAN) 的全局语义信息提升目标外观信息的差异。

本节将从平衡不同任务的语义信息、提取更加明确表观特征与设计抗遮挡的数据关联匹配算法 3 个方面进行对 FairMOT 的改进，AOMOT 的整体模型框架如图 1 所示。

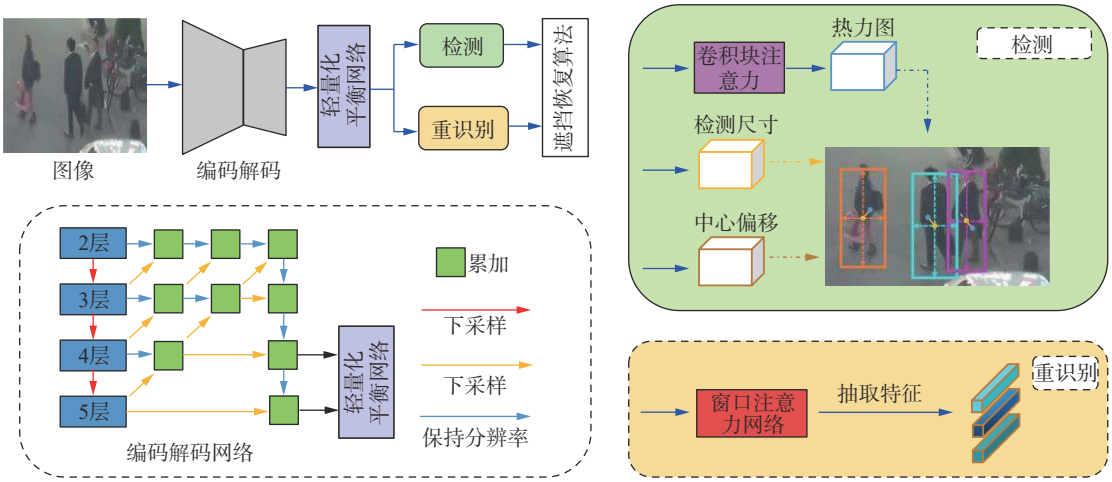


图 1　AOMOT 的模型框架

Fig. 1　AOMOT model frame

1.1　轻量化平衡网络

遵循联合检测和重识别的模型[15-17]通过单一网络进行目标检测和重识别特征提取。其中目标检测更加注重目标和背景之间的类间差异，而重识别特征提取任务更加注重类内差异，2 个任务目的上的冲突，导致主干网络提取的特征语义不明确，不利于后续 2 个任务分支的训练。特征语义的混淆，也导致针对特征提取任务的优化会影响目标检测的准确性。对于卷积神经网络而言，不同深度对应着不同层次的语义特征，相同的行人在相邻的 2 帧下中层语义属性很少变化，中层网络可以学到目标更多的细节特征，而高层网络的特征语义信息比较丰富，可以很好地从背景中分离出感兴趣的行人目标。轻量化平衡网络设计目的在于平衡目标检测任务和重识别任务中不同层次语义信息在特征的中比重，从而生成更适合各自任务的特征图。轻量化平衡网络 (lightweight balanced network, LBN) 的结构如图 2 所示。

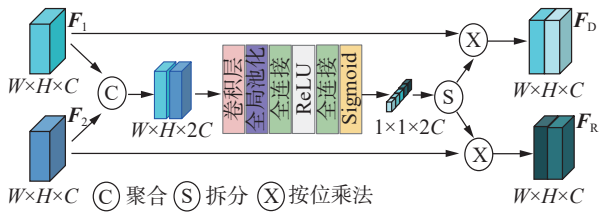


图 2　LBN 结构

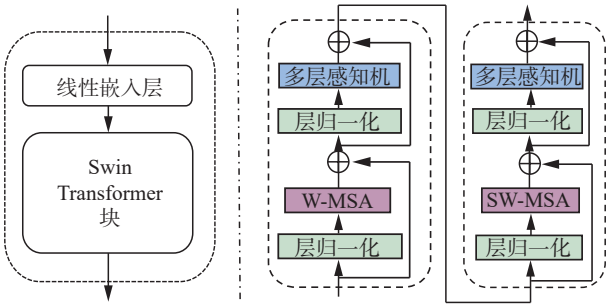
Fig. 2　LBN structure

首先抽取主干网络 DLA-34 的第 4 层和第 5 层的特征图分别记 $\boldsymbol{F}_1\in\mathbf{R}^{W\times H\times C}$ 和 $\boldsymbol{F}_2\in\mathbf{R}^{W\times H\times C}$。将 $\boldsymbol{F}_1$ 和 $\boldsymbol{F}_2$ 在通道维度进行叠加聚合，生成的特征图记作 $\boldsymbol{F}\in\mathbf{R}^{W\times H\times 2C}$，使用 3×3 的卷积将聚合后的特征映射为高维特征，经过全局平局池化将一个通道上的整体空间的信息都编码成一个特征值来代表特征图在不同通道上的语义特征，后经过多层感知机利用通道间整体信息的相关性来计算特征权重，用 Sigmoid 激活函数将权重限制到 [0,1] 的范围，最后将特征权重拆分并分别为初始特征加权，生成 2 个聚合不同层次语义信息的特征图 $\boldsymbol{F}_{\mathrm{D}}$ 和 $\boldsymbol{F}_{\mathrm{R}}$，加权后的特征图将分别用于目标检测和重识别特征提取。

此外，还可以将卷积块注意力模块[18]作为网络的子模块作用于目标检测任务中的热力图分支，进一步整合空间信息，提高目标检测中心点预测的准确度，减少特征提取对目标检测的影响。

1.2　窗口注意力网络

重识别特征提取任务主要是在视频序列提取出同一行人的相似性表观特征，判断在不同帧中的行人是否是同一人。在重识别特征提取的过程中 FairMOT 使用 2 层 3×3 的卷积作为特征提取器。这种结构只利用了检测目标的局部信息，忽略了目标和周围区域之间的全局语义关系。为了解决这一问题设计了基于 Swin-Transformer[19] 结构的窗口注意力网络，在窗口注意力网络中行人的重识别特征是由移动窗口计算的。移动窗口方案使得自注意力计算限制在非重叠的窗口上，并且允许跨窗口连接，这使得不同窗口之间的特征可以进行注意力关联。网络增强特征抽取能力的同时也降低了计算复杂度，窗口注意力网络 (window attention network, WAN) 的结构如图 3 所示。



(a) 特征提取结构　　(b) Swim-Transformer 块

图 3　WAN 结构

Fig. 3　WAN structure

图 3(a) 为 WAN 的结构, 其中特征提取结构由 1 个线性嵌入层和 1 个 Swin-Transformer 块组成。线性嵌入层使用 1×1 卷积提升特征维度至 128 维。Swin-Transformer 块由图 3(b) 所示的基于移位窗口的 MSA(multi-head self-attention) 和具有 GELU(gaussian error linear unit) 非线性的多层感知机组成。在每个 MSA 模块和每个多层感知机之前应用归一化层, 在每个模块之后接 1 个残差连接。W-MSA 和 SW-MSA 分别表示基于窗口的标准自注意力和移动窗口的多头自注意力。其中 W-MSA 模块表示局部窗口内计算自注意力, 这些窗口不重叠地均匀划分特征图。假设特征图为 $\boldsymbol{F}\in\mathbf{R}^{w\times h\times c}$, 每个窗口包含 $M\times M$ 特征像素, 基于窗口的计算复杂度为

$$\Omega_{\text{W-MSA}}=4hwc^2+2M^2hwc$$

当 M 固定时, 窗口注意力的计算复杂度随着特征图的分辨率大小线性增长, 这对于计算高分辨率特征图是有利的。但是窗口自注意力缺乏跨窗口关联上下文信息的能力, 所以窗口注意力网络在保持非重叠窗口计算效率的同时还通过移动窗口进行了跨窗口连接, 在 1 个 Swin-Transformer 块中的 2 种注意力交替进行, 其计算公式为

$$\hat{z}^{l}=\text{W-MSA}\left(\text{LN}\left(z^{l-1}\right)\right)+z^{l-1},$$

$$z^{l}=\text{MLP}\left(\text{LN}\left(\hat{z}^{l}\right)\right)+\hat{z}^{l}$$

$$\hat{z}^{l+1}=\text{SW-MSA}\left(\text{LN}\left(z^{l}\right)\right)+z^{l},$$

$$z^{l+1}=\text{MLP}\left(\text{LN}\left(\hat{z}^{l+1}\right)\right)+\hat{z}^{l+1}$$

式中 $\hat{z}^{l}$ 和 z^{l} 分别表示 (S)W-MSA 模块和多层感知机模块的输出特征。最后在生成的高分辨率特征图中根据目标检测分支的热力图中心点的位置提取重识别特征。

1.3　遮挡恢复算法

在行人密集的场景下行人和行人之间会发生频繁的遮挡, Zhang 等[20] 提出的 ByteTrack 指出在对行人进行目标检测时, 行人间的相互遮挡导致其目标检测的置信度下降从而造成漏检, 这使得漏检的目标无法进行轨迹关联。而单纯地升目标检测的置信度阈值又会造成大量的误检, 导致跟踪的准确度下降。遮挡发生的过程中, 被遮挡的行人在视频序列的前一帧中被检测到并且关联成为轨迹。虽然在当前帧中因为遮挡导致置信度降低, 但是其位置信息依旧可信的, 其外观信息也与上一帧被关联的轨迹相似, 因此提出遮挡恢复算法。

遮挡恢复算法 (occlusion restoration algorithm, OraSORT) 的主要目的是在低置信度目标检测中挖掘出被遮挡的物体, 更新其外观信息后重新纳入目标检测并进行与轨迹的关联匹配。算法 1 给出了遮挡恢复算法的伪代码。

算法 1　遮挡恢复算法

输入　当前帧检测 $\boldsymbol{D}$; 上一帧关联轨迹 $\boldsymbol{T}$; 检测阈值 τ_{high} 和 τ_{low}; 余弦距离阈值 e

输出　满足条件的检测 $\boldsymbol{D}_{\text{result}}$

1) 初始化: $\boldsymbol{D}_{\text{result}}\leftarrow\varnothing$
2) $\boldsymbol{D}_{\text{low}}\leftarrow\varnothing$
3) $\boldsymbol{D}_{\text{high}}\leftarrow\varnothing$
4) for d in $\boldsymbol{D}$ do
5) if $d_{\text{score}}>\tau_{\text{high}}$ then
6) $\boldsymbol{D}_{\text{high}}\leftarrow\boldsymbol{D}_{\text{high}}\cup\{d\}$
7) else if $d_{\text{score}}>\tau_{\text{low}}$ and $d_{\text{score}}<\tau_{\text{high}}$ then
8) $\boldsymbol{D}_{\text{low}}\leftarrow\boldsymbol{D}_{\text{low}}\cup\{d\}$
9) end if
10) end for
11) /*判断目标是否被遮挡*/
12) $\boldsymbol{D}_{\text{remain}}\leftarrow\boldsymbol{D}_{\text{low}}$ 与 $\boldsymbol{T}$ 交并比最高
13) $\boldsymbol{D}_{\text{remain}}\leftarrow\boldsymbol{D}_{\text{result}}$ 与 $\boldsymbol{D}_{\text{high}}$ 特征相似度>e
14) /*$\boldsymbol{D}_{\text{result}}$ 特征更新为关联轨迹 $\boldsymbol{T}$ 特征*/
15) $\boldsymbol{D}_{\text{result-feature}}\leftarrow\boldsymbol{T}_{\text{feature}}$
16) /*恢复被遮挡的目标*/
17) $\boldsymbol{D}_{\text{result}}\leftarrow\boldsymbol{D}_{\text{result}}\cup\boldsymbol{D}_{\text{high}}$
18) return $\boldsymbol{D}_{\text{result}}$

遮挡恢复算法的输入是当前帧的检测集合 $\boldsymbol{D}$ 和上一帧完成关联的轨迹集合 $\boldsymbol{T}$。此外还设置了 3 个阈值 τ_{high}、τ_{low} 和 e。其中 τ_{high} 和 τ_{low} 为检测置信度阈值, e 为目标相似度阈值。遮挡恢复算法的输出是满足阈值条件的检测集合。对于视频中的每一帧图片得到检测集合 $\boldsymbol{D}$, 其中包含目标的检测框位置、重识别特征和置信度。

首先, 根据置信度阈值 τ_{high} 和 τ_{low} 将检测集合 $\boldsymbol{D}$ 分为 2 类。置信度高于 τ_{high} 的检测目标放入高置信度检测集合 $\boldsymbol{D}_{\text{high}}$ 中。对于那些置信度范围从

τ_{low}到τ_{high}的检测，将它们放入低置信度的检测集合$\boldsymbol{D}_{high}$中。

其次，计算低置信度检测集合$\boldsymbol{D}_{low}$和完成关联的轨迹集合$\boldsymbol{T}$之间所有检测框的交并比，为每一个完成关联的轨迹匹配一个交并比最高的低置信度检测。如果轨迹和检测的检测框的交并比为 0，则说明检测不属于同一个目标进而不进行匹配。

将所有被匹配到的低置信度检测放入一个新的低置信度检测集合$\boldsymbol{D}_{remain}$。假设在$\boldsymbol{D}_{remain}$中的检测是当前帧中被遮挡的目标，通过计算交并比的最大值，强制$\boldsymbol{T}$中的轨迹关联一个$\boldsymbol{D}_{low}$中的检测。也就是在中$\boldsymbol{D}_{remain}$检测到的目标是在上一帧被关联为轨迹但是在当前帧因为被遮挡而置信度降低的目标。

然后，注意到当前帧中受遮挡影响的目标与已经被检测到的高置信度目标有着明显的外观区别。所以计算$\boldsymbol{D}_{remain}$和$\boldsymbol{D}_{high}$之间的特征相似度，在$\boldsymbol{D}_{result}$中保留特征相似度大于 0.2 的检测，降低误检数量。

最后，将$\boldsymbol{D}_{result}$中的特征对应更新为上一帧关联的轨迹$\boldsymbol{T}$的特征，将$\boldsymbol{D}_{result}$作为从遮挡中恢复的检测集合并入高置信度检测集合$\boldsymbol{D}_{high}$，并进行后续的关联计算。

由于模型的特征提取能力得到了加强并且为低置信度检测更新了可靠的外观特征。所以在 DeepSORT 级联匹配的第二级匹配中，匈牙利匹配的损失矩阵 $\boldsymbol{C}$ 的计算可以引入特征相似度矩阵$\boldsymbol{A}_a$修正部分遮挡目标的损失，计算其与交并比矩阵$\boldsymbol{A}_m$的加权和。权重因子λ设置为 0.6，其计算公式为

$$\boldsymbol{C} = \lambda \boldsymbol{A}_a + (1-\lambda)\boldsymbol{A}_m$$

2　实验与分析

2.1　数据集和指标

2.1.1　MOT Challenge 数据集

MOT Challenge 数据集中包括 MOT15[21]、MOT16[22]、MOT17 等数据集。视频来源于多个不同角度的街景镜头，主要跟踪目标均为行人，包含多种光照条件、遮挡程度不同的场景，能够充分地代表多目标跟踪的实际应用场景。其中 MOT15 数据集有 22 个视频序列，训练集 11 个视频共包含 5 500 张图片，测试集 11 个视频共包含 5 783 张图片。该数据集共含有 1 221 个行人身份的标注。MOT16 和 MOT17 数据集来源于相同的视频序列，数据集中一共有 14 个视频，训练集和测试集各 7 个视频；训练集有 5 316 帧图像，测试集有 5 919 帧图像，总共有 11 235 帧。该数据集共含有 1 342 个行人身份的标注，其中训练集有 512 个，测试集有 830 个。MOT16 和 MOT17 的区别在于使用了不同检测器。

2.1.2　评估指标

使用 MOT15 的验证集来进行消融实验并在 MOT17 测试集上对 AOMOT 进行评估，采用高阶跟踪精度 (higher order tracking accuracy, HOTA)[23] 作为主要指标，因为它在目标检测的准确性和目标关联之间保持了适当的平衡。除此之外也使用身份 F1 分数 (identity F1 score, IDF1)、多目标跟踪精度 (multiple object tracking accuracy, MOTA)、身份切换数量 (identity switches, IDSW) 作为多目标跟踪系统抗遮挡能力的评价指标，并且使用每秒帧数作为评价模型推理速度的指标，误检数 (false positive, FP) 和漏检数 (false negative, FN) 作为评价实验鲁棒性的指标。

2.2　实验细节

使用 DLA-34 的一个变体作为默认主干，将 CrowdHuman[24] 数据集上预先训练的模型参数用于初始化 AOMOT。使用 Adam 优化器对 AOMOT 进行了 26 个轮次的训练，初始学习速率为10^{-4}，学习速率在 20 个轮次后衰减到10^{-5}。批处理大小被设置为 4。使用标准的数据增强技术，包括旋转、缩放和颜色抖动。

在训练时除了使用 MOT17 数据集之外还引入 ETH[25]、CityPerson[26]、CalTech[27]、CUHK-SYSU[28]、PRW[29] 进行训练。输入图像大小调整为 1 088 像素×608 像素，特征图的分辨率为 272 像素×152 像素。在 NVIDIA GeForce RTX 3060 上进行训练，训练大约需要 150 h。

2.3　实验结果与分析

2.3.1　消融实验

AOMOT 使用 MOT17 验证集进行训练，在 MOT15 验证集测试各模块有效性。消融实验将分别验证对模型网络结构的改进和遮挡恢复算法对多目标跟踪的影响。首先验证窗口注意力网络的有效性，其实验结果如表 1 所示。

表 1　不同模块对跟踪效果的影响

Table 1　Impact of different modules on tracking performance

LBN	WAN	MOTA/%↑	IDF1/%↑	IDSW↓	速度/(f/s)↑
		62.5	70.3	235	**14.4**
	√	62.1	71.0	**217**	11.6
√	√	62.4	**71.5**	228	10.7

实验结果表明在基线网络中加入窗口注意力网络进行重识别特征提取时 IDF1 提升 0.7 百分点 IDSW 降低 18, MOTA 下降 0.4 百分点。这表明窗口注意力网络通过关联上下文信息提取了更具代表性的重识别特征。但是 MOTA 的下降表明目标检测任务和重识别特征提取任务的冲突依旧存在, 重识别特征提取任务的优化降低模型目标检测的性能。为了验证轻量化平衡网络对这个冲突的缓解作用。使用带有 CBAM 的轻量化平衡网络进行实验。实验结果显示联合使用上述模块后 MOTA 相比原模型下降 0.1 百分点, IDF1 上升 1.2 百分点, IDSW 下降 7。这说明轻量化平衡网络通过平衡不同任务分支之间的不同层次的语义信息缓解了任务冲突带来性能降低。值得注意的是, 比较不同模型的推理速度是相当困难的, 因为每种方法给出的速度取决于实现它们的设备, 并且检测所花费的时间通常排除在系统跟踪目标的全部时间之外。因此实验结果只对比 AOMOT 与基线模型 FairMOT 的推理速度。观察实验数据发现, 在使用窗口注意力网络后模型的推理速度下降 2.8 f/s, 在增加轻量化平衡网络后推理速度下降 0.9 f/s。AOMOT 相比于 FairMOT 增加的网络结构造成了推理时间的增加, 但是可以通过提升硬件性能来满足高视频速率的推理。

接下来验证遮挡恢复算法对多目标跟踪系统的影响, 实验使用基线模型预测目标位置和提取重识别特征。实验对比了基于 DeepSORT 改进的遮挡恢复算法 OraSORT 与 FairMOT 使用的 DeepSORT 算法对跟踪效果的影响, 实验结果如表 2 所示。通过实验对比可以发现算法在 MOT15 和 MOT16 数据集上得出的结果中 MOTA 分别提升 0.1 百分点和 0.2 百分点, IDF1 分别提升 0.7 百分点和 0.8 百分点, IDSW 分别下降 10% 和 6%。结果表明基于 DeepSORT 改进的 OraSORT 算法在不同的数据集上维持被跟踪目标身份时都有着稳定的表现, 并且恢复被遮挡目标并在级联匹配中融合重识别特征进行加权, 对原始 DeepSORT 的改进是有效的。

表 2　不同关联算法对跟踪效果的影响

Table 2　Impact of different association algorithms on tracking performance

数据集	方法	MOTA/%↑	IDF1/%↑	IDSW↓
MOT15	DeepSORT	66.2	73.2	146
	OraSORT	**66.3**	**73.9**	**131**
MOT16	DeepSORT	83.3	81.9	545
	OraSORT	**83.5**	**82.5**	**515**

2.3.2　MOT Challenge 上的结果及与其他方法的对比

为了全面评估 AOMOT 模型的抗遮挡能力, AOMOT 使用 MOT17 的测试集进行评估, 并与近几年的其他模型进行对比, 其结果如表 3 所示。观察对比结果可以发现 AOMOT 相比 FairMOT 在 HOTA、IDF1 上分别提升 1.5 百分点、3 百分点, 且 IDSW 下降 32%, 且相比其他模型在 HOTA、IDF1 和 IDSW 均达到了最好的效果。表明 AOMOT 在跟踪目标时, 对于目标发生因遮挡导致的轨迹交换或目标轨迹中断有着很好的鲁棒性。但是模型在 MOTA 指标上的结果低于 CSTrack、GTR 与 SGT。分析原因是 CSTrack 使用了更复杂的相关矩阵来解耦特征。GTR 的目标检测器在 DETR[30] 上进行微调, 从而继承了 DETR 的目标检测能力。SGT 针对目标检测聚合了高维特征。这 3 种方式虽然提升了模型在目标检测上的能力但同时也增加了计算量。

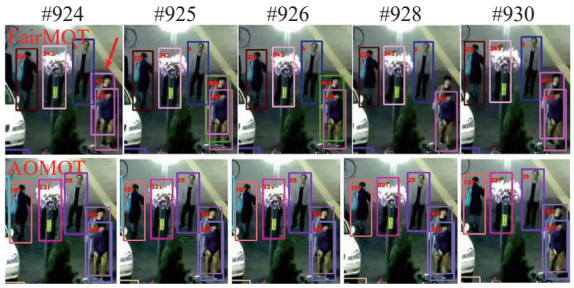
表 3　MOT17 测试集与其他模型的对比

Table 3　Comparison of MOT17 test set with other models

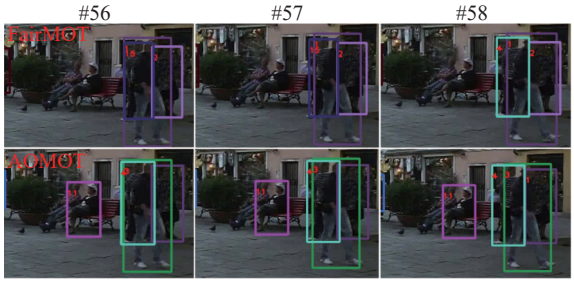
方法	HOTA/%↑	MOTA/%↑	IDF1/%↑	FP↓	FN↓	IDSW↓
CenterTrack[10]	52.2	67.8	64.7	**18 498**	160 332	3 039
TraDeS[9]	52.7	69.1	63.9	20 892	150 060	3 555
TransCenter[31]	54.4	73.2	62.2	23 112	123 738	4 614
MeMOT[32]	56.9	72.5	69.0	37 221	115 248	2 724
GTR[6]	59.1	75.3	71.5	26 793	109 854	2 859
CSTrack[33]	59.3	74.9	72.6	23 847	114 303	3 567
FairMOT[12]	59.3	73.7	72.3	27 507	117 477	3 303
SGT[11]	60.6	**76.3**	72.4	25 983	**102 984**	4 578
AOMOT	**60.8**	73.7	**75.3**	29 799	116 187	**2 257**

2.3.3　实验结果可视化

AOMOT 在行人密集场景 MOT17 验证集视频中可视化效果如图 4 所示。



(a) MOT17-03 可视化对比



(b) MOT17-01 可视化对比

(c) MOT17-06 可视化对比

(d) MOT17-12 可视化对比

图 4　MOT17 跟踪结果可视化

Fig. 4　MOT17 visualization of tracking results

通过与 FairMOT 在图 4(a) 中 MOT17-03 第 924、925、926、928、930 帧的跟踪效果对比可以看出，使用 FairMOT 跟踪的 924 帧中红色箭头指向的 102 号行人，在 925 帧中因为误检而被标注了 102、253 两个身份信息，在 926 帧中产生了身份切换，由 102 切换成了 253，接下来在 928 帧中又由于遮挡导致漏检丢失轨迹，在 930 帧中被重新检测成功后又产生了一次身份切换。而 AOMOT 在 924 ~ 930 帧期间一直稳定地行人标注相同的 192 号身份。在图 4(b) 视频 MOT17-01 中 FairMOT 在 56 ~ 58 帧中未能准确检测在椅子上的老人，并且在 58 帧 15 号目标发生身份切换。AOMOT 检测到座椅上的老人并标注且稳定跟踪 4 号目标。在图 4(c) 视频 MOT17-06 中 FairMOT 在 219 ~ 221 帧中 113 号目标在 220 帧丢失，在 221 帧发生 ID 切换，而 AOMOT 稳定跟踪相同的目标。在图 4(d) 视频 MOT17-12 中 599 ~ 605 帧，AOMOT 可以稳定标记被遮挡的 94 号目标，而 FairMOT 到 606 帧才开始标记相同位置的目标。可视化结果表明了 AOMOT 在有遮挡发生的情况下，也能稳健和精确地跟踪目标。

3　结束语

针对行人运动过程中相互遮挡时多目标跟踪系统的跟踪身份频繁切换的问题，AOMOT 基于 FairMOT 进行了改进。提出了轻量化平衡网络，平衡主干网络不同层次间的语义信息，缓解特征图的语义冲突，为后续的分支任务提供更有利的语义信力，增大重识别特征的类内距离，促使模型提取更加具有区分度的重识别特征。最后针对遮挡造成的漏检提出遮挡恢复算法，通过判断目标之间的交并比和外观相似度恢复低置信度的检测目标，提升模型的身份跟踪能力。实验结果表明 AOMOT 所作的改进有效地提升了模型的抗遮挡能力，相较于原模型在相同应用场景下有着更好的表现。不过 AOMOT 在目标检测任务上效果仍有不足。未来将进一步优化检测任务的网络结构，并将视频时序信息产生的运动特征用于目标关联，进一步提升关联的准确度。

参考文献：

[1] 魏旋旋, 黄子健, 曹乐, 等. 采用轻量级姿态估计网络的脊柱侧弯筛查方法 [J]. 智能系统学报, 2023, 18(5): 1039−1046.
WEI Xuanxuan, HUANG Zijian, CAO Le, et al. Scoliosis screening method using lightweight pose estimation network[J]. CAAI transactions on intelligent systems, 2023, 18(5): 1039−1046.

[2] DAVID R, SÖFFKER D. A review on machine learning-based models for lane-changing behavior prediction and recognition[J]. Frontiers in future transportation, 2023, 4: 950429.

[3] 高尚兵, 黄子赫, 耿璇, 等. 视觉协同的违规驾驶行为分析方法 [J]. 智能系统学报, 2021, 16(6): 1158−1165.
GAO Shangbing, HUANG Zihe, GENG Xuan, et al. A visual collaborative analysis method for detecting illegal driving behavior[J]. CAAI transactions on intelligent sys-

tems, 2021, 16(6): 1158−1165.

[4] 胡硕, 王洁, 孙妍等. 无人机视角下的多车辆跟踪算法研究 [J]. 智能系统学报, 2022, 17(4): 798−805.
HU Shuo, WANG Jie, SUN Yan, et al. Research on multi-vehicle tracking algorithm from the perspective of UAV[J]. CAAI transactions on intelligent systems, 2022, 17(4): 798−805.

[5] BASHAR M, ISLAM S, HUSSAIN K K, et al. Multiple object tracking in recent times: a literature review[EB/OL]. (2022−09−11)[2023−07−10]. https://arxiv.org/abs/2209.04796.

[6] ZHOU Xingyi, YIN Tianwei, KOLTUN V, et al. Global tracking transformers[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 8761−8770.

[7] SUN Peize, CAO Jinkun, JIANG Yi, et al. Transtrack: multiple object tracking with transformer[EB/OL]. (2020−12−01)[2023−07−10]. https://arxiv.org/abs/2012.15460.

[8] 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: IEEE, 2022: 8834−8844.

[9] WU Jialian, CAO Jiale, SONG Liangchen, et al. Track to detect and segment: an online multi-object tracker[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 12347−12356.

[10] ZHOU Xingyi, KOLTUN V, KRÄHENBÜHL P. Tracking objects as points[C]//European Conference on Computer Vision. Cham: Springer, 2020: 474−490.

[11] HYUN J, KANG M, WEE D, et al. Detection recovery in online multi-object tracking with sparse graph tracker[C]//2023 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa: IEEE, 2023: 4839−4848.

[12] ZHANG Yifu, WANG Chunyu, WANG Xinggang, 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] DUAN Kaiwen, BAI Song, XIE Lingxi, et al. CenterNet: keypoint triplets for object detection[C]//2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 6568−6577.

[14] 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: IEEE, 2017: 3645−3649.

[15] 王凯, 戴芳, 郭文艳, 等. 融合目标相似性和作用力的多目标跟踪 [J]. 中国图象图形学报, 2024, 29(7): 1984−1997.
WANG Kai, DAI Fang, GUO Wenyan, et al. Integrating similarity and interaction force between objects for multiple object tracking[J]. Journal of image and graphics, 2024, 29(7): 1984−1997.

[16] YU En, LI Zhuoling, HAN Shoudong, et al. RelationTrack: relation-aware multiple object tracking with decoupled representation[J]. IEEE transactions on multimedia, 2023, 25: 2686−2697.

[17] 张英俊, 白小辉, 谢斌红. CNN-Transformer 特征融合多目标跟踪算法 [J]. 计算机工程与应用, 2024, 60(2): 180−190.
ZHANG Yingjun, BAI Xiaohui, XIE Binhong. Multi-object tracking algorithm based on CNN-transformer feature fusion[J]. Computer engineering and applications, 2024, 60(2): 180−190.

[18] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module[C]//European conference on computer vision. Cham: Springer, 2018: 3−19.

[19] LIU Ze, LIN Yutong, CAO Yue, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]//2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021: 9992−10002.

[20] ZHANG Yifu, SUN Peize, JIANG Yi, et al. ByteTrack: multi-object tracking by associating every detection box[C]//Lecture Notes in Computer Science. Cham: Springer, 2022: 1−21.

[21] LAURA LEAL-TAIXÉ, MILAN A , REID I , et al. MOTChallenge 2015: towards a benchmark for multi-target tracking[EB/OL]. (2015−04−08) [2023−07−10]. https://arxiv.org/abs/1504.01942.

[22] MILAN A , LEAL-TAIXE L , REID I , et al. MOT16: a benchmark for multi-object tracking[EB/OL]. (2016−03−02)[2023−07−10]. https://arxiv.org/pdf/1603.00831.

[23] LUITEN J, OŠEP 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.

[24] SHAO Shuai, ZHAO Zijian, LI Boxun, et al. Crowdhuman: a benchmark for detecting human in a crowd[EB/OL]. (2018−04−30)[2023−07−10]. https://arxiv.org/abs/1805.00123.

[25] 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: IEEE, 2008: 1−8.

[26] ZHANG Shanshan, BENENSON R, SCHIELE B. CityPersons: a diverse dataset for pedestrian detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 4457−4465.

[27] DOLLAR P, WOJEK C, SCHIELE B, et al. Pedestrian detection: a benchmark[C]//2009 IEEE Conference on

Computer Vision and Pattern Recognition. Miami: IEEE, 2009: 304−311.

[28] XIAO Tong, LI Shuang, WANG Bochao, et al. Joint detection and identification feature learning for person search[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 3376−3385.

[29] ZHENG Liang, ZHANG Hengheng, SUN Shaoyan, et al. Person re-identification in the wild[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 3346−3355.

[30] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers[C]//European conference on computer vision. Cham: Springer, 2020: 213−229.

[31] XU Yihong, BAN Yutong, 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.

[32] CAI Jiarui, XU Mingze, LI Wei, et al. MeMOT: multi-object tracking with memory[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 8080−8090.

[33] LIANG Chao, ZHANG Zhipeng, ZHOU Xue, et al. Rethinking the competition between detection and ReID in multiobject tracking[J]. IEEE transactions on image processing, 2022, 31: 3182−3196.

作者简介：

张国印，教授，博士生导师，博士，主要研究方向为智能感知与决策。主持国家自然科学基金等各类科研项目 20 余项，发表学术论文 100 余篇。E-mail：zhangguoyin@hrbeu.edu.cn。

高伟，副教授，博士，中国计算机学会会员、中国计算机学会黑龙江省计算机网络专委会委员。主要研究方向为计算机网络、数据库和信息安全，发表学术论文 30 余篇。E-mail：gaowei@hrbeu.edu.cn。

王传博，硕士研究生，主要研究方向为计算机视觉。E-mail：694882809@qq.com。