Computer Science and Application 计算机科学与应用, 2022, 12(8), 1952-1963
Published Online August 2022 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2022.128196

Hans汉斯

基于改进FairMOT特征解耦的多目标跟踪算法

刘文强，李　阳，王家宝，王彩玲，苗　壮，裘杭萍*

陆军工程大学指挥控制工程学院，江苏　南京

收稿日期：2022年7月12日；录用日期：2022年8月12日；发布日期：2022年8月18日

摘　要

联合检测和重识别跟踪模型(Joint-Detection-and-Embedding Models, JDE)的两个子任务所需要的特征存在矛盾，通过目标中心点提取重识别特征的方式难以得到遮挡目标的有效特征，这导致在复杂环境下模型提取的目标重识别特征可靠性下降，造成数据关联错误。针对目标检测和重识别任务间的矛盾问题，文中基于FairMOT跟踪算法提出了一种特征解耦模块。该模块使用协调注意力(Coordinate Attention, CA)将骨干网输出的多尺度特征图进行初步解耦，然后以自底向上的方式融合不同分辨率的重识别特征图。为了提取遮挡目标的有效信息，文中提出一种根据目标可视度调整高斯核方差的策略，用于构建目标中心点监督热图，加大训练时对遮挡目标及其周围区域的关注。最后在MOT17数据集上对所提算法进行了测试，实验结果验证了各模块的有效性，表明了算法能够有效应对遮挡，实现稳定跟踪。

关键词

多目标跟踪，目标重识别，特征解耦，注意力机制

Multi-Object Tracking Algorithm Based on Improved FairMOT Feature Decoupling

Wenqiang Liu, Yang Li, Jiabao Wang, Cailing Wang, Zhuang Miao, Hangping Qiu*

Command & Control Engineering College, Army Engineering University of PLA, Nanjing Jiangsu

Received: Jul. 12th, 2022; accepted: Aug. 12th, 2022; published: Aug. 18th, 2022

Abstract

The features required for the two sub-tasks of joint object detection and re-identification multi-object tracking algorithm (Joint-Detection-and-Embedding Models, JDE) are contradictory. It is difficult to extract the effective features of occluded objects by extracting object re-identification

*通讯作者。

文章引用: 刘文强, 李阳, 王家宝, 王彩玲, 苗壮, 裘杭萍. 基于改进 FairMOT 特征解耦的多目标跟踪算法[J]. 计算机科学与应用, 2022, 12(8): 1952-1963. DOI: 10.12677/csa.2022.128196

features through the object center point. This leads to unreliable object re-identification features extracted by the model in complex environments, resulting in data association errors. A feature decoupling module is proposed based on the FairMOT tracking algorithm, aiming at the contradiction between object detection and re-identification tasks. This module uses coordinate attention to initially decouple the multi-scale feature maps output by the backbone network, and then fuses re-identification feature maps of different resolutions in a bottom-up manner. In order to extract the effective information of the occluded object, a strategy of adjusting the variance of the Gaussian kernel according to the visibility of the object is proposed, which is used to construct a supervised heat map of the object center point, and pay more attention to the occluded object and its surrounding areas during training. The proposed algorithm is tested on the MOT17 dataset, and the experimental results verify the effectiveness of each module, indicating that the algorithm can effectively deal with occlusion and achieve stable tracking.

Keywords

Multi-Object Tracking, Object Re-Identification, Feature Decoupling, Attention Mechanism

1. 引言

多目标跟踪要求定位并识别视频中不同的目标[1] [2]，作为计算机视觉的一个重要研究课题，它为多种实际应用提供了关键的核心技术支持，包括视频分析[3]、自动驾驶[4]、机器人[5]等。目前，多目标跟踪主要依靠目标检测技术进行定位，然后根据运动预测和重识别特征关联不同帧中的目标形成轨迹。相比于只依靠运动预测进行关联，使用重识别特征能够获得更强的抗遮挡能力，因为它可以识别出被遮挡后重新出现的目标。因此，为被跟踪的目标提取良好的重识别特征是近年来的一个热点研究方向。基于检测的跟踪模型(Detection-based-Tracking Models, DBT) [6] [7]使用一个离线训练的目标重识别网络提取图像帧中检测切片的重识别特征，但这类方法会导致模型的复杂度显著升高，并且跟踪速度随着目标的数量增多而明显下降。

为了兼顾跟踪算法的关联性能和实时性，Wang 等人[8]首先构建了联合检测和重识别的跟踪模型(Joint-Detection-and-Embedding Models, JDE)，在检测算法上加入一个分支用来学习重识别特征，极大地降低了提取重识别特征的复杂度。随后 Zhang 等人[9]针对锚不适合重识别特征提取的问题进一步提出了基于目标中心点的联合检测和重识别的跟踪模型，实现了良好的跟踪准确度和实时性。

联合检测和重识别的跟踪模型虽然简单且有效，但是它们的性能通常不如基于检测的跟踪模型，性能的差异来源于联合检测和重识别模型面临的两个挑战：

1) 目标检测和重识别任务之间存在矛盾。检测任务需要区分不同的类，例如背景中的行人和车辆，因此需要为相同类的对象提取更相似的特征；而重识别任务要求区分的是个体，它需要为同一类中的不同对象提取具有差异的特征。在同一个骨干网中提取这两个相矛盾的特征导致了模型优化的困难。

2) 被遮挡的目标难以提取有效的重识别特征。不同于基于检测的跟踪算法[6] [7]可以使用大量额外的重识别数据集训练目标重识别模型，JDE 模型[8] [9] [10] [11] [12]的重识别分支依赖检测模型聚合的特征进行分类学习。基于目标中心点的模型通过可变形卷积将目标的特征向其中心位置聚合。但在实际的

聚合过程中，中心点特征通常重点关注到目标中心位置附近的信息，这导致当目标中心位置被遮挡时，网络很难找到目标正确的特征区域。

针对任务间的矛盾问题，近两年来，研究者使用特征解耦模块将主干网提取的原始特征图映射到子任务分支中[11] [12]。这些解耦模块采用不同的注意力机制建模子特征图的相关关系，从而生成丰富且具有差异的检测特征图和目标重识别特征图。但是，这些模型没有从目标重识别本身的特性进行解耦。由于要区分所有目标，重识别特征更加关注深层特征图中的语义信息。为此，我们设计了一种关注深层特征的融合解耦方法，该方法使用不同的方式融合骨干网输出的多分辨率特征图，分别生成适合检测和重识别任务的特征图。

此外，为了聚合遮挡目标的有效特征，研究者通常在目标重识别分支后加入一个复杂的注意力网络学习不同尺度目标以及目标的相关性[11] [12]。但是这导致了计算量的增加，在训练时需要更多的计算资源。为了缓解这一问题，本文提出了一种依据目标可视度构建监督热图的方法。该方法采用数据集中标注的可视度信息调整构建监督热图时的高斯核参数，使被遮挡的目标得到更大的关注，从而令目标区域的信息以更大概率聚合到目标中心，引导模型关注遮挡目标与其周围物体的联系。

综上所述，本文的主要贡献如下：

1) 改进了一种带特征解耦的无锚框的 JDE 跟踪算法，使用 CA 注意力机制和自底向上的融合模块将原始特征图解耦为检测特征图和目标重识别特征图。

2) 针对联合检测和重识别的跟踪算法难以提取被遮挡目标有效特征的问题，将目标的可视度融入到监督信息中，根据可视度调整目标中心热图的高斯核参数提高训练时对遮挡样本的关注。

3) 在 MOT17 验证集中对本文方法进行了测试，实验结果表明本文算法能够明显超过基准方法，并且改进的解耦方案优于其它同类型的解耦方法。

2. 相关工作

2.1. 联合检测与重识别模型

联合检测与重识别的跟踪模型在单个神经网络中实现目标检测和重识别特征提取，以简化多阶段的网络设计，减小模型并提升跟踪速度。这些模型通常建立在先进的检测算法上，例如 JDE [8]使用一个单独的 Re-ID 头在 YOLOv3 框架[13]原有的检测分支之外学习表观特征；FairMOT [9]基于 CenterNet 网络[14]设计了目标中心点的特征提取分支；RetinaTrack [15]修改 RetinaNet [16]，使用独立的分支对每个检测锚框进行分类、回归和表观特征提取，以防止为重叠目标提取相同的表观特征。为了在表观特征中融入时间和空间信息，GSDT [17]利用图网络实现帧间特征融合，提升遮挡和变形目标的特征提取能力；CorrTracker [18]采用自监督的方式同时学习目标的时空特性和尺度信息，提取轨迹的表观特征，明显提升了跟踪准确度。但以上算法均未解决目标检测和重识别的优化矛盾问题。本文在基于目标中心点的跟踪算法上，针目标重识别自身特点，将骨干网输出的原始特征解耦，使用不同的融合策略生成关注不同特征的检测特征图和重识别特征图。

2.2. 注意力机制与特征解耦

注意力机制是用来自动计算模型输出对不同特征关注度的一种特殊结构，通过注意力机制对中间特征进行加权从而过滤非典型信息。例如，SE [19]、CBAM [20]以及 CA (Coordinate Attention) [21]中的注意力模块可以广泛应用于多种网络架构中对特征的通道及空间进行加权注意，提升模型的效果。在多目标跟踪算法中，研究者也基于注意力机制处理目标检测和重识别任务，生成不同的注意力图融合适合各

自任务的典型特征，以提取更丰富且准确的重识别特征。例如，CSTrack [11]设计了一种通道注意力解耦模块，称为互相关网络(Cross-correlation Network, CCN)，通过目标检测与重识别特征的自相关和互相关操作生成通道注意力图，然后对各自特征图进行通道注意力加权；RelationTrack [12]提出了一种全局的通道注意力机制，融合全局信息解耦骨干网提取的原始特征图。近期，SimpleTrack [10]提出了一种自底向上的特征融合模块用于目标重识别分支，从而根据目标检测和重识别的任务特性进行了解耦。虽然SimpleTrack 使用与检测分支不同的结构进行重识别融合的解耦方式有效，但是重识别特征融合模块更加关注浅层特征，这可能导致融合后的特征缺失深层语义信息。为了在重识别分支中充分聚合深层的语义特征，本文在深层特征上采样的过程中使用空间和通道注意力机制保留重要信息，并且逐层相加融合各分辨率的特征图，以减少信息的丢失。

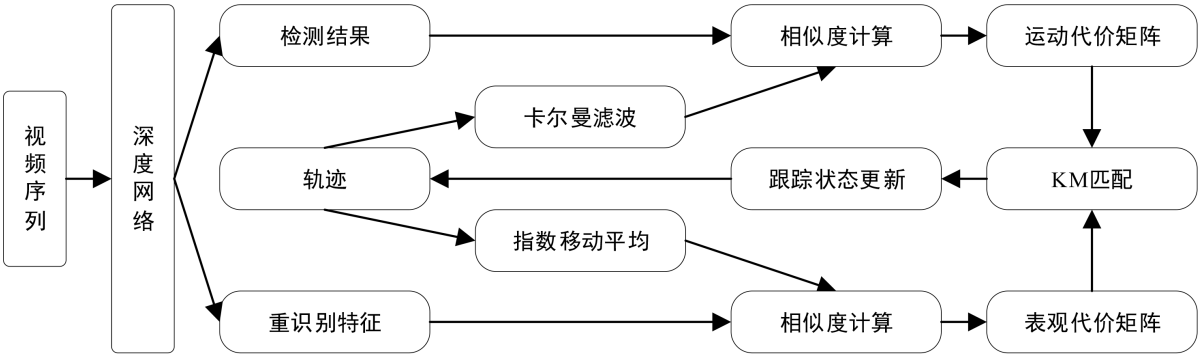

Figure 1. Flow chart of JDE algorithm
图 1. 联合检测与重识别算法跟踪流程

3. 算法设计

3.1. 算法框架

本文采用的联合检测和重识别在线跟踪框架基于目标中心点进行跟踪，如图 1 所示。该框架对视频序列进行单帧处理，其中的深度网络用来实现目标检测和重识别特征提取。输入的序列图片经骨干网提取深度特征图，再通过解耦器分解为用于目标检测的特征和用于目标重识别的特征，然后分别在头部网络中检测目标并提取对应目标的重识别特征向量。在跟踪阶段，数据关联模块根据由运动特征和重识别特征(表观特征)分别构建的代价矩阵确定检测与轨迹间的匹配关系，并迭代更行轨迹状态。其中轨迹的运动特征和表观特征分别通过卡尔曼滤波和指数移动平均(EMA) [8]逐帧进行更新。

3.2. 深度网络

本文采用 Stacked Hourglass Networks [22] [23]作为骨干网，该网络在输入图片 4 倍下采样的高分辨率特征图上进行密集的预测，可以有效区分空间位置相近的目标，防止目标在特征图上的中心点映射存在大尺度的偏移而导致对应位置的重识别特征存在偏差，从而适应拥挤的多目标跟踪场景。为了通过该网络获得良好的目标检测特征图和重识别特征图，本文对其中的多分辨率特征融合模块进行了改进。如图 2 所示，首先使用两个结构相同参数不同的 CA 注意力将同分辨率的特征图进行初步的分解，突出各任务关注的特征通道和位置；然后对检测分支使用 Stacked Hourglass Networks 原有的自顶向下的特征融合结构聚合不同分辨率的特征；对重识别分支使用一种自底向上的特征融合结构进行特征聚合，在融合过程中使用空间和通道注意力操作保持深层特征中的语义信息。

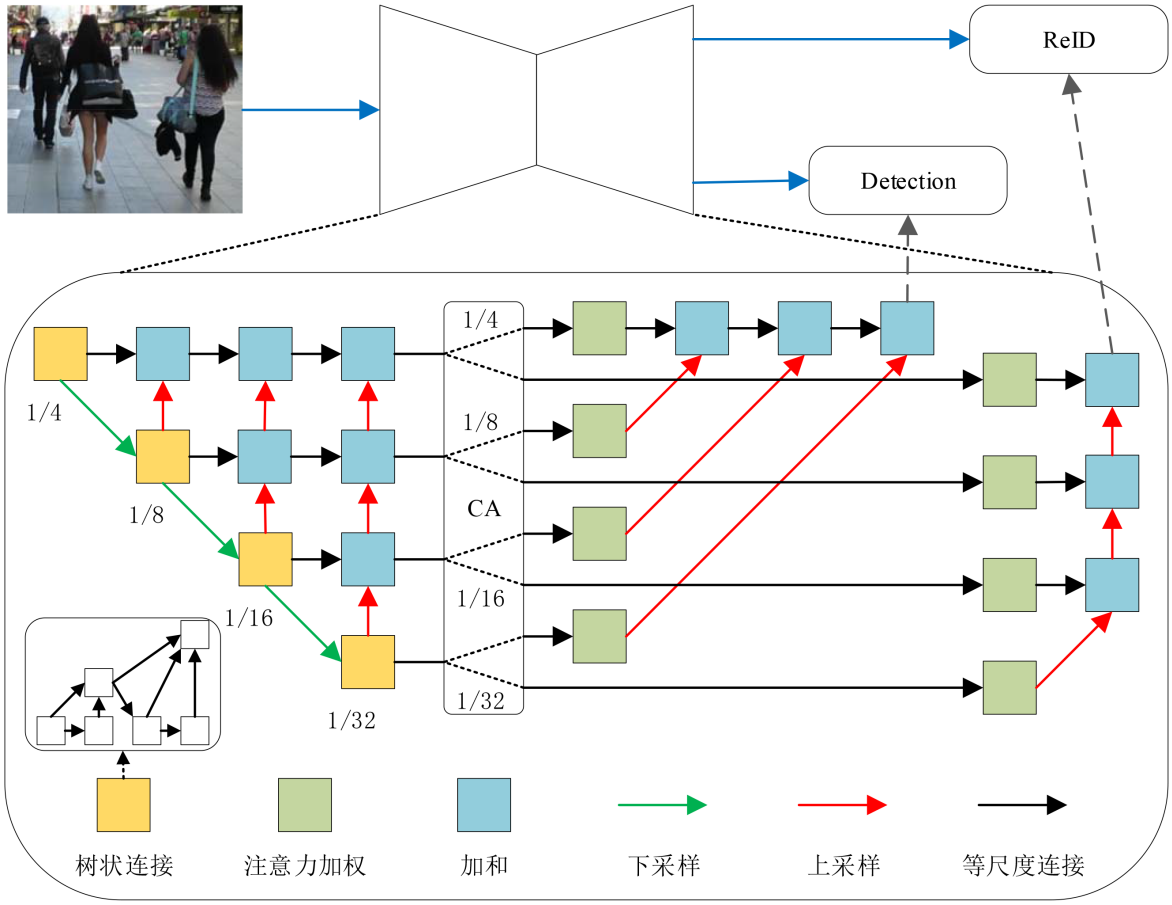

Figure 2. The overall framework of the decoupling algorithm
图 2. 解耦算法整体框架

3.2.1. CA 注意力

本文利用 CA 注意力[21]将骨干网提取的多分辨率特征图进行初步的解耦，假设 $\{F_i\}_{i=1}^{N}$ 表示骨干网输出的多分辨率特征图，N 为特征层数，通过公式(1)将所有分辨率的特征图分解成两组：

$$\left\{\hat{F}_i^d\right\}_{i=1}^{N}=\left\{CA_i^d\left(F_i\right)\right\}_{i=1}^{N},\quad \left\{\hat{F}_i^e\right\}_{i=1}^{N}=\left\{CA_i^e\left(F_i\right)\right\}_{i=1}^{N} \tag{1}$$

其中 $\left\{\hat{F}_i^d\right\}_{i=1}^{N}$ 和 $\left\{\hat{F}_i^e\right\}_{i=1}^{N}$ 分别表示解耦后的检测和重识别多尺度特征图，CA_i^d 和 CA_i^e 代表结构相同参数不同的 CA 注意力模块。其中 CA 注意力是一种兼顾特征图位置关系的通道注意力模块，它在输入特征图上分别进行宽和高两个方向的通道注意力操作，以此捕获目标的位置信息和通道关系，可以有效增强特征的表达能力。

3.2.2. 自底向上融合模块

低分辨率的深层特征包含了丰富的语义信息，可以帮助模型进行更细粒度的对象级别的识别。在解耦的重识别分支中，采用一种自底向上的方式向浅层特征中传递深层语义信息。具体来说，将低分辨率的深层特征转化为高分辨率特征一共包含四个步骤：一是特征图上采样，将下层特征图尺寸调整到与上层特征保持一致；二是空间注意力，利用上层特征准确的空间信息生成空间注意力图，对上采样后的特征图进行空间信息编码；三是通道注意力，利用下层特征图生成的通道注意力权重对上采样后的特征图

进行通道上的处理；四是将得到的特征图与上层特征图进行相加融合，从而保留原始信息。由此可以得到融合公式为：

$$\left\{\hat{F}_i\right\}_{i=N}^{1}=\begin{cases}F_i, & \text{if } i=N\\ UpSample\left(\hat{F}_{i+1}\right)\cdot cam\left(\hat{F}_{i+1}\right)\cdot\sigma\left(Conv_{1\times1}\left(F_i\right)\right)+F_i, & \text{otherwise}\end{cases} \tag{2}$$

其中 $UpSample(\bullet)$ 表示由可变形卷积和反卷积组成的上采样操作，$Conv_{1\times1}$ 表示 1 × 1 的卷积层，$\sigma(\bullet)$ 表示 Sigmoid 激活层，$cam(\bullet)$ 表示层间的通道注意力操作。$cam(\bullet)$ 的结构如图 3 所示，首先使用全局平局池化和全局最大池化聚合全局信息，然后使用一个标准的两层卷积模块将下层特征的通道维度压缩到与上层一致，最后与输入特征进行乘法调整各通道的权重。

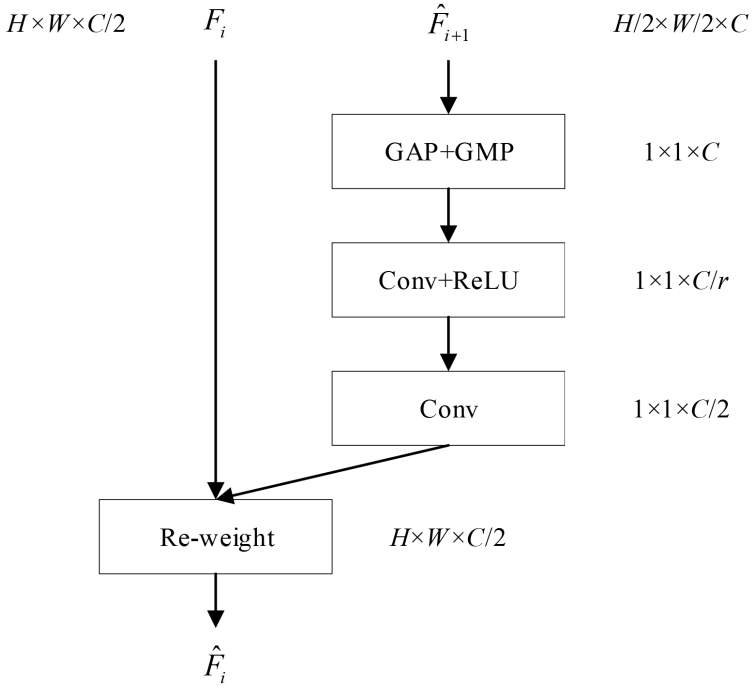

Figure 3. CAM interlayer channel attention
图 3. CAM 层间通道注意力

3.2.3. 遮挡目标高斯核自适应调整策略

中心点检测和跟踪算法在输出特征图上进行一个密集的预测，判定像素点是否为目标中心点。为此本文采用基于热图的表示方法将标签信息转化为热力图的形式，在单类多目标跟踪场景中(例如行人多目标跟踪)热力图的维度为 $1\times H\times W$，H 和 W 分别代表模型输出的特征图的高和宽，1 表示类别数。热图中与标注目标中心点重合的位置的预计响应值为 1，视为正样本；不在目标标注框区域的预计响应值为 0，视为负样本；标注目标中心到目标边界的响应值呈指数衰减，通过目标尺寸及可视度控制衰减速度。

具体来说，对于图中每个目标的真实边界框 $\left(X_i,Y_i,W_i,H_i\right)$，它在 4 倍下采样后的特征图中的映射边界框为 $\left(x_i,y_i,w_i,h_i\right)$，使用高斯核函数平缓地将目标的中心点分布到热图中：

$$Y_{xy}=\exp\left(-\frac{\left(x-x_i\right)^2+\left(y-y_i\right)^2}{2\sigma^2}\right) \tag{3}$$

其中 σ 为目标尺度自适应标准差，通过以下操作进行确定：

$$\sigma = 2 \cdot Guassian\ radius\left(\left(h_i, w_i\right)+1\right)\Big/\left(5+\max\left(0.5, S\right)\right) \tag{4}$$

其中 *Guassian radius* 操作根据输入边界框尺寸自适应确定高斯核半径，它随着目标尺寸增大和目标可见度减小而增加，S 表示目标的可见程度。

4. 实验

4.1. 实验细节

本文实验在两台 NVIDIA GeForce RTX 2080Ti GPU 上进行，操作系统为 Ubuntu 20.04.1，采用 PyTorch 实现深度学习模型的训练和测试。本文模型首先使用在 COCO 数据集[24]上预训练的参数进行初始化，然后使用 CrowdHuman [25]验证子集和 MOT17 [26]训练集的前半部分进行微调，最后在 MOT17 训练集的后半部分进行测试。其中 CrowdHuman 数据集只提供检测级别的标注，因此对该数据集使用自监督的方式[9]训练重识别分支。实验采用 Adam [27]优化器更新模型参数，输入图像的大小设置为 1088 × 608，Batch size 设为 12，总共训练 30 个 epoch，初始学习率设置为 10^{-4}，在最后 10 个 epoch 减为 10^{-5}。

4.2. 实验结果

本文在 MOT17 验证集上对所提方法的跟踪性能进行测试。该数据集共有 7 段标注视频序列，其中第 2、4、9 段序列为固定镜头的街道行人场景，其余为镜头发生抖动或移动的行人场景，跟踪结果如表 1 所示。表中使用的跟踪评价指标来自于 MOTChallenge 基准，它们的具体含义在表 2 中进行了总结。为了说明不同类型的样本对跟踪性能的影响，本文在图 4 中给出了不同序列上 HOTA、IDF1 和 MOTA 三个重要指标以及对应序列中标注可视度为 0 和小尺寸目标(定义为面积小于 900 平方像素)数量占比的折线图。从表 1 结果可以看到，整体上三个指标随着 0 可视度和小样本比例的增加而减小，其中，IDF1 指标受可视度为 0 的样本占比影响更大，而 HOTA 和 MOTA 受到小目标和可视度为 0 样本的双重影响。例如在序列 13 中没有 0 可视度样本但存在较多的小目标，此时 IDF1 依然取得较高得分，而 HOTA 和 MOTA 得分与序列 10 及序列 11 的分数基本持平，这说明目标遮挡是造成关联错误的主要原因。为了进一步说明本文算法的有效性，我们在表 3 中对比了同类型的跟踪算法。其中粗体代表最优结果，*代表使用了 SimpleTrack [10]提出的关联策略，没有*的表示采用 FairMOT 的关联策略。从表 3 结果可以看到，在使用 FairMOT 的关联策略时，本文方法比最先进的基于解耦的算法在 HOTA 和 IDF1 两个主要指标上分别提升了 0.8 和 1.1。这两个指标的提升说明本文所提算法能够提取更可区分的重识别特征。在使用了 SimpleTrack 所提关联策略后，本文方法比最先进的基于解耦的算法在 HOTA、IDF1 和 MOTA 三个主要指标上分别提升了 0.8、0.7 和 0.8，并且 MT 和 FN 都能够达到最优。这表明在使用特征解耦的跟踪算法中，本文所提的特征解耦模块能够有效将骨干网输出的原始特征图进行合理的解耦，从而提升跟踪效果。

Table 1. Tracking results of different video sequences
表 1. 不同视频序列的跟踪结果

序列	HOTA↑	IDF1↑	MOTA↑	MT↑	ML↓	FP↓	FN↓	IDS↓
2	39.5	50.4	42.2	10	17	658	4916	160
4	68.2	84.8	83.9	52	2	823	2970	96
5	53.0	72.9	67.1	22	12	31	1039	37
9	60.7	76.7	77.6	14	1	23	606	20

Continued

10	54.7	75.7	64.3	12	2	295	1775	55
11	55.9	65.5	67.0	17	12	275	1179	38
13	56.5	76.6	66.3	25	6	329	684	58
All	59.3	75.0	70.3	152	52	2434	13169	464

Table 2. Multi-object tracking evaluation index and its meaning
表 2. 多目标跟踪评价指标及含义

评价指标	指标含义
HOTA↑	多目标跟踪高阶指标
MOTA↑	多目标跟踪准确度
IDF1↑	多目标跟踪身份标识精度
MT↑	至少 80%被正确跟踪的轨迹数量
ML↓	低于 20%被正确跟踪的轨迹数量
FP↓	将背景误检测为目标的数量
FN↓	将目标误检测为背景的数量
IDS↓	轨迹身份标识改变次数

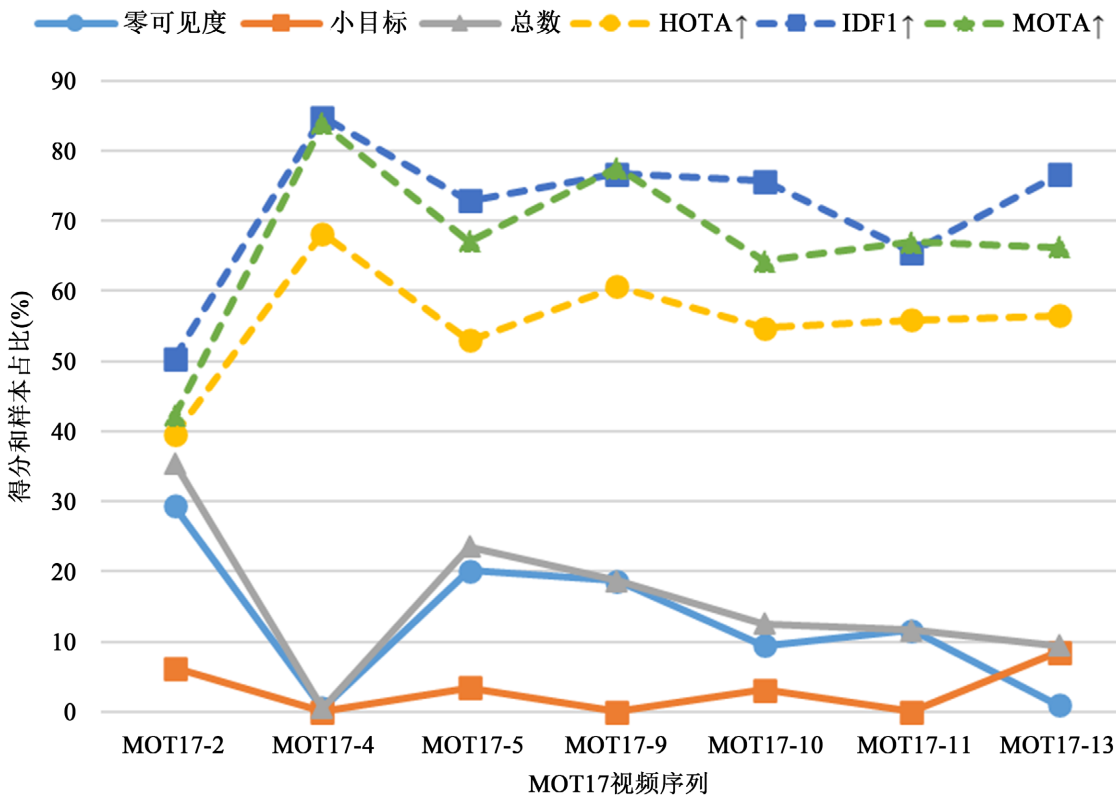

Figure 4. Line chart of tracking results
图 4. 跟踪结果折线图

Table 3. Tracking results of different tracking algorithms on MOT17 dataset
表 3. 不同跟踪算法在 MOT17 数据集上的跟踪结果

算法	HOTA↑	IDF1↑	MOTA↑	MT↑	ML↓	FP↓	FN↓	IDS↓
FairMOT	57.8	72.3	70.7	159	55	2350	13,117	370
SimpleTrack	58.5	73.9	70.8	158	49	2695	12,663	431
OUR	59.3	75.0	70.7	152	52	2425	13,169	464
SimpleTrack*	60.2	76.3	71.6	159	61	2119	13,036	191
OUR*	61.0	77.0	72.4	161	63	2150	12,582	203

4.3. 消融实验

本文的主要创新点为自底向上融合模块、CA 注意力解耦和遮挡目标高斯核自适应策略。下面分别对这三个结构进行消融实验，消融实验在 MOT17 验证集上进行。其中表 4 为在 FairMOT 算法上添加各个模块的消融结果，BU_D 代表自底向上融合模块，OA 表示使用遮挡目标高斯核自适应策略，CA 表示使用 CA 注意力解耦。从表 4 结果可以看到，在 FairMOT 算法上单独加入自底向上融合模块之后，HOTA 和 IDF1 分别提升了 0.9 和 2.2，MOTA 得分略微下降，单独加入使用遮挡目标高斯核自适应策略也能够分别提升 1.0 的 HOTA 得分和 1.8 的 IDF1 得分，并且 MOTA 得分也只有微小的下降。在同时使用这两个策略之后，模型的跟踪性能在 HOTA、IDF1 和 MOTA 三个指标上都有了进一步的提升。最后加入 CA 注意力进行初步解耦之后，模型在 HOTA、IDF1 和 MT 三个指标上取得了最佳表现。

最后，我们比较了 SE、CBAM 和 CA 三种不同注意力模块在初步解耦时的有效性，实验在 MOT17 训练集的前半部分进行训练，在 MOT17 训练集后半部分测试，结果如表 5 所示。从表 5 结果可以看到，如果使用 SE 注意力进行初步解耦，模型的整体跟踪效果将会下降，这可能是由于单独的通道注意力无法

Table 4. Ablation study of different modules on the MOT17 dataset
表 4. 在 MOT17 数据集上对不同模块的消融实验

算法	HOTA↑	IDF1↑	MOTA↑	MT↑	ML↓	FP↓	FN↓	IDS↓
FairMOT	57.8	72.3	70.7	159	55	2350	13,117	370
SimpleTrack	58.5	73.9	70.8	158	49	2695	12,663	431
BU_D	58.7 + 0.9	74.5 + 2.2	70.4−0.3	158	49	2766	12,803	402
OA	58.8 + 1.0	74.1 + 1.8	70.3−0.4	161	49	2542	13,068	440
BU_D+GA	59.1 + 1.3	74.6 + 2.3	71.1+0.4	158	51	2743	12,399	471
BU_D+OA+CA	59.3 + 1.5	75.0 + 2.7	70.7+0.0	152	52	2425	13,169	464

Table 5. Effects of different attention mechanisms in initial decoupling
表 5. 不同注意力机制在初步解耦时的效果

注意力	HOTA↑	IDF1↑	MOTA↑	MT↑	ML↓	FP↓	FN↓	IDS↓
SE	56.3	70.7	67.5	144	61	3024	14,113	437
CBAM	56.3	72.8	68.3	141	55	2541	14,175	438
CA	57.2	73.5	68.3	139	53	2583	14,079	447

有效建模各分辨率特征图中的有效信息。CBAM 和 CA 注意力取得了相同的 MOTA 得分，但是 CA 注意力在 HOTA、IDF1、MT、ML 和 FN 五项指标中都取得了最佳表现。

4.4. 可视化分析

为了从直观上验证本文算法的有效性，本文在一个困难场景中对所提算法进行了可视化分析，如图 5 所示。其中图 5(a)分别为本文方法和 FairMOT 算法在 MOT17-6 的序列的跟踪结果。在这幅图像中，一个人穿过了遮挡物。FairMOT 未能提取出具有代表性的 Re-ID 特征，导致目标没有与正确的轨迹关联。

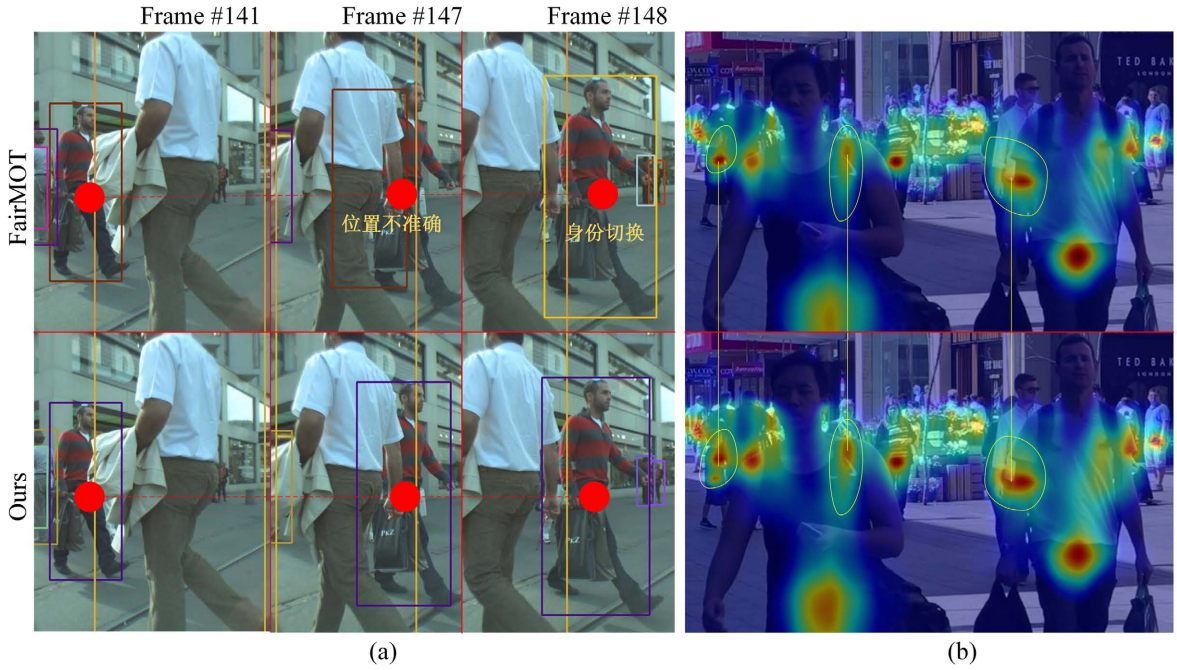

Figure 5. Visual analysis
图 5. 可视化分析

具体来说，FairMOT 在#147 帧时得到的目标位置不准确，然后在帧#148 目标的身份发生了切换。但是本文方法在行人穿过其它遮挡时依然能够稳定地跟踪。

此外，我们进一步使用 CAM 可视化方法[28]显示了模型对跟踪目标的关注区域，如图 5(b)所示，其中图像取自 MOT17-8 序列。从标出的三个区域中可以看到，使用了遮挡目标高斯核自适应策略后的模型对被遮挡目标的可见区域施加了更大的关注，这从侧面说明了该策略的有效性。

5. 结论

本文在基于中心点的联合检测和重识别跟踪模型中增加使用自底向上融合解耦模块，获得更加丰富且具有差异性的检测特征和重识别特征，较好地解决了联合检测和重识别跟踪模型中检测和目标重识别任务之间的优化矛盾问题。针对遮挡问题，提出遮挡目标高斯核自适应策略，提高了模型对遮挡目标的关注，使模型可以提取遮挡目标可见区域的有效特征。实验结果表明，本文算法所提部件均能提升基准方法的跟踪效果，在 MOT17 验证集上取得了良好的表现。但是，使用两个分离的融合模块处理检测和重识别分支不仅增加了模型参数，而且使两个分支无法共享重要信息。此外，联合跟踪模型的跟踪精度依旧逊色于基于检测的跟踪模型，因此更加合理的解耦方式依旧是后续值得进一步研究的内容。

参考文献

[1] Ciaparrone, G., Sánchez, F.L., Tabik, S., Troiano, L., Tagliaferri, R. and Herrera, F. (2020) Deep Learning in Video Multi-Object Tracking: A Survey. *Neurocomputing*, **381**, 61-88. https://doi.org/10.1016/j.neucom.2019.11.023

[2] Sun, Z., Chen, J., Chao, L., Ruan, W. and Mukherjee, M. (2021) A Survey of Multiple Pedestrian Tracking Based on Tracking-by-Detection Framework. *IEEE Transactions on Circuits and Systems for Video Technology*, **31**, 1819-1833. https://doi.org/10.1109/TCSVT.2020.3009717

[3] Takahashi, N., Gygli, M. and Van Gool, L. (2018) AENet: Learning Deep Audio Features for Video Analysis. *IEEE Transactions on Multimedia*, **20**, 513-524. https://doi.org/10.1109/TMM.2017.2751969

[4] Luo, W., Yang, B. and Urtasun, R. (2018) Fast and Furious: Real Time End-to-End 3D Detection, Tracking and Motion Forecasting with a Single Convolutional Net. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 3569-3577. https://doi.org/10.1109/CVPR.2018.00376

[5] Manglik, A., Weng, X., Ohn-Bar, E. and Kitanil, K.M. (2019) Forecasting Time-to-Collision from Monocular Video: Feasibility, Dataset, and Challenges. *IEEE/RSJ International Conference on Intelligent Robots and Systems* (*IROS*), Macau (China), 3-8 November 2019, 8081-8088. https://doi.org/10.1109/IROS40897.2019.8967730

[6] Wojke, N., Bewley, A. and Paulus D. (2017) Simple Online and Realtime Tracking with a Deep Association Metric. 2017 *IEEE International Conference on Image Processing* (*ICIP*), Beijing, 17-20 September 2017, 3645-3649. https://doi.org/10.1109/ICIP.2017.8296962

[7] Du, Y., Song, Y., Yang, B. and Zhao, Y. (2022) StrongSORT: Make DeepSORT Great Again. ArXiv, abs/2202.13514.

[8] Wang, Z., Zheng, L., Liu, Y., Li, Y. and Wang, S. (2020) Towards Real-Time Multi-Object Tracking. *European Conference on Computer Vision* (*ECCV*) *Workshops*, Glasgow, 23-28 August 2020, 107-122. https://doi.org/10.1007/978-3-030-58621-8_7

[9] Zhang, Y., Wang, C., Wang, X., Zeng, W. and Liu, W. (2021) FairMOT: On the Fairness of Detection and Re-identification in Multiple Object Tracking. *International Journal of Computer Vision*, **129**, 3069-3087. https://doi.org/10.1007/s11263-021-01513-4

[10] Li, J., Ding, Y., Wei, H.-L., Zhang, Y. and Lin, W. (2022) SimpleTrack: Rethinking and Improving the JDE Approach for Multi-Object Tracking. *Sensors*, **22**, Article No. 5863. https://doi.org/10.3390/s22155863

[11] Liang, C., Zhang, Z., Zhou, X., Li, B., Zhu, S. and Hu, W. (2022) Rethinking the Competition between Detection and ReID in Multi-Object Tracking. *IEEE Transactions on Image Processing*, **31**, 3182-3196. https://doi.org/10.1109/TIP.2022.3165376

[12] Yu, E., Li, Z., Han, S. and Wang, H. (2022) RelationTrack: Relation-Aware Multiple Object Tracking with Decoupled Representation. *IEEE Transactions on Multimedia*. https://doi.org/10.1109/TMM.2022.3150169

[13] Redmon, J. and Farhadi, A. (2018) YOLOv3: An Incremental Improvement. ArXiv, abs/1804.02767.

[14] Zhou, X., Wang, D. and Krähenbühl, P. (2019) Objects as Points. ArXiv, abs/1904.07850.

[15] Lu, Z., Rathod, V., Votel, R. and Huang, J. (2020) RetinaTrack: Online Single Stage Joint Detection and Tracking. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 14656-14666. https://doi.org/10.1109/CVPR42600.2020.01468

[16] Lin T, Y., Goyal, P., Girshick, R., He, K. and Dollár, P. (2017) Focal Loss for Dense Object Detection. 2017 *IEEE International Conference on Computer Vision* (*ICCV*), Venice, 22-29 October 2017, 2999-3007. https://doi.org/10.1109/ICCV.2017.324

[17] Wang, Y., Kitani, K. and Weng X. (2021) Joint Object Detection and Multi-Object Tracking with Graph Neural Networks. *IEEE International Conference on Robotics and Automation* (*ICRA*), Xi'an, 30 May-5 June 2021, 13708-13715. https://doi.org/10.1109/ICRA48506.2021.9561110

[18] Wang, Q., Zheng, Y., Pan, P. and Xu, Y. (2021) Multiple Object Tracking with Correlation Learning. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 20-25 June 2021, 3875-3885. https://doi.org/10.1109/CVPR46437.2021.00387

[19] Hu, J., Shen, L., Albanie, S., Sun, G. and Wu, E. (2020) Squeeze-and-Excitation Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **42**, 2011-2023. https://doi.org/10.1109/TPAMI.2019.2913372

[20] Woo, S., Park, J., Lee, J.Y. and Kweon, I.S. (2018) CBAM: Convolutional Block Attention Module. *European Conference on Computer Vision* (*ECCV*) *Workshops*, Munich, 8-14 September 2018, 3-19. https://doi.org/10.1007/978-3-030-01234-2_1

[21] Hou, Q., Zhou, D. and Feng, J. (2021) Coordinate Attention for Efficient Mobile Network Design. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 20-25 June 2021, 13708-13717. https://doi.org/10.1109/CVPR46437.2021.01350

[22] Newell, A., Yang, K. and Deng, J. (2016) Stacked Hourglass Networks for Human Pose Estimation. *European Conference on Computer Vision* (*ECCV*) *Workshops*, Amsterdam, 11-14 October 2016, 483-499. https://doi.org/10.1007/978-3-319-46484-8_29

[23] Yu, F., Wang, D., Shelhamer, E. and Darrell, T. (2018) Deep Layer Aggregation. 2018 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 2403-2412. https://doi.org/10.1109/CVPR.2018.00255

[24] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., *et al*. (2014) Microsoft COCO: Common Objects in Context. *European Conference on Computer Vision* (*ECCV*) *Workshops*, Zurich, 6-12 September 2014, 740-755. https://doi.org/10.1007/978-3-319-10602-1_48

[25] Shao, S., Zhao, Z., Li, B., Xiao, T., Yu, G., Zhang, X., *et al*. (2018) CrowdHuman: A Benchmark for Detecting Human in a Crowd. ArXiv, abs/1805.00123. http://arxiv.org/abs/1805.00123

[26] Milan, A., Leal-Taixé, L., Reid I, D., Roth, S. and Schindler, K. (2016) MOT16: A Benchmark for Multi-Object Tracking. ArXiv, abs/1603.00831. http://arxiv.org/abs/1603.00831

[27] Kingma, D.P. and Ba, J. (2015) Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations* (*ICLR*), San Diego, 7-9 May 2015, 13. https://hdl.handle.net/11245/1.505367

[28] Zhou, B., Khosla, A., Lapedriza, À., Oliva, A. and Torralba, A. (2016) Learning Deep Features for Discriminative Localization. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 2921-2929. https://doi.org/10.1109/CVPR.2016.319