ISSN 1003-501X CN 51-1346/O4

光电工程

Opto-Electronic Engineering

中文核心期刊 中国科技核心期刊
Scopus CSCD

时空特征对齐的多目标跟踪算法

程稳，陈忠碧，李庆庆，李美惠，张建林，魏宇星

引用本文:

程稳，陈忠碧，李庆庆，等. 时空特征对齐的多目标跟踪算法[J]. 光电工程，2023， **50**(6): 230009.

Cheng W, Chen Z B, Li Q Q, et al. Multiple object tracking with aligned spatial-temporal feature[J]. *Opto-Electron Eng*, 2023, **50**(6): 230009.

https://doi.org/10.12086/oee.2023.230009

收稿日期: 2023-01-12; 修改日期: 2023-04-02; 录用日期: 2023-04-03

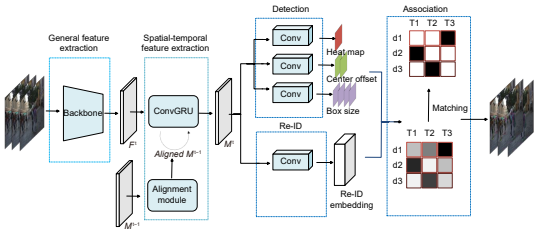
相关论文

Opto-Electronic Engineering
光 电 工 程

Article
2023 年, 第 50 卷, 第 6 期

DOI: 10.12086/oee.2023.230009

时空特征对齐的多目标跟踪算法



程　稳[1,2,3]，陈忠碧[2*]，李庆庆[2]，
李美惠[2]，张建林[2]，魏宇星[2]

[1]中国科学院光场调控科学技术全国重点实验室，四川 成都 610209；
[2]中国科学院光电技术研究所，四川 成都 610209；
[3]中国科学院大学电子电气与通信工程学院，北京 100049

摘要： 多目标跟踪 (Multi-object tracking, MOT) 是计算机视觉领域的一项重要任务，现有研究大多针对目标检测和数据关联进行改进，通常忽视了不同帧之间的相关性，未能充分利用视频时序信息，导致算法在运动模糊，遮挡和小目标场景中的性能显著下降。为解决上述问题，本文提出了一种时空特征对齐的多目标跟踪方法。首先，引入卷积门控递归单元 (convolutional gated recurrent unit, ConvGRU)，对视频中目标的时空信息进行编码；该结构通过考虑整个历史帧序列，有效提取时序信息，以增强特征表示。然后，设计特征对齐模块，保证历史帧信息和当前帧信息的时间一致性，以降低误检率。最后，本文在 MOT17 和 MOT20 数据集上进行了测试，所提算法的 MOTA (multiple object tracking accurary) 值分别为 74.2 和 67.4，相比基准方法 FairMOT 提升了 0.5 和 5.6；IDF1 (identification F1 score) 值分别为 73.9 和 70.6，相比基准方法 FairMOT 提升了 1.6 和 3.3。此外，定性和定量实验结果表明，本文方法的综合跟踪性能优于目前大多数先进方法。

关键词： 多目标跟踪；时空特征；ConvGRU；时间一致性；特征对齐

中图分类号： TP391.41　　**文献标志码：** A

程稳，陈忠碧，李庆庆，等. 时空特征对齐的多目标跟踪算法 [J]. 光电工程，2023，**50**(6): 230009
Cheng W, Chen Z B, Li Q Q, et al. Multiple object tracking with aligned spatial-temporal feature[J]. *Opto-Electron Eng*, 2023, **50**(6): 230009

Multiple object tracking with aligned spatial-temporal feature

Cheng Wen[1,2,3], Chen Zhongbi[2*], Li Qingqing[2], Li Meihui[2], Zhang Jianlin[2], Wei Yuxing[2]

[1]National Key Laboratory of Optical Field Manipulation Science and Technology, Chinese Academy of Sciences, Chengdu, Sichuan 610209 China;
[2]Institute of Optics and Electronics, Chinese Academy of Science, Chengdu, Sichuan 610209 China;
[3]University of Chinese Academy of Science School of Electronic, Electrical, Communication Engineering, Beijing 100049 China

Abstract: Multiple object tracking (MOT) is an important task in computer vision. Most of the MOT methods

收稿日期：2023-01-12；修回日期：2023-04-02；录用日期：2023-04-03
基金项目：国家自然科学基金青年科学基金资助项目 (62101529)
*通信作者：陈忠碧，chenzb@ioe.ac.cn。

improve object detection and data association, usually ignoring the correlation between different frames. They don't make good use of the temporal information in the video, which makes the tracking performance significantly degraded in motion blur, occlusion, and small target scenes. In order to solve these problems, this paper proposes a multiple object tracking method with the aligned spatial-temporal feature. First, the convolutional gated recurrent unit (ConvGRU) is introduced to encode the spatial-temporal information of the object in the video; By considering the whole history frame sequence, this structure effectively extracts the spatial-temporal information to enhance the feature representation. Then, the feature alignment module is designed to ensure the time consistency between the historical frame information and the current frame information to reduce the false detection rate. Finally, this paper tests on MOT17 and MOT20 datasets, and multiple object tracking accuracy (MOTA) values are 74.2 and 67.4, respectively, which is increased by 0.5 and 5.6 compared with the baseline FairMOT method. Our identification F1 score (IDF1) values are 73.9 and 70.6, respectively, which are increased by 1.6 and 3.3 compared with the baseline FairMOT method. In addition, the qualitative and quantitative experimental results show that the overall tracking performance of this method is better than that of most of the current advanced methods.

Keywords: multiple object tracking; spatial-temporal feature; ConvGRU; time consistency; feature alignment

1 引　言

多目标跟踪是计算机视觉的重要组成部分，广泛应用在监控视频分析和自动驾驶等领域，其目的是定位多个感兴趣的目标，并维持每个目标唯一的身份编号 (identification, ID)，记录连续运动轨迹[1]。多目标跟踪面临诸多挑战，首先在背景复杂的场景中，目标的运动具有不确定性和随意性，而且目标与目标之间存在相互遮挡以及被物体遮挡现象，导致目标特征发生改变；其次在低分辨率场景中，目标与背景差异较小，分辨出感兴趣目标就十分困难；并且在多目标跟踪过程中目标数目具有不确定性，容易带来误检、漏检以及 ID 切换等现象。针对上述问题，研究者提出了一系列多目标跟踪方法。最早的多目标跟踪算法主要关注优化检测以及数据关联。随着目标检测和行人重识别的迅速发展，MOT 也有了相当大的突破[2-10]。但是这些方法的检测步骤是完全独立于先前的历史帧，一旦目标变得部分或完全被遮挡，检测器就失效了，从而造成轨迹丢失。Zhou 等人 [11-12] 将成对的帧作为输入，直接输出检测和成对的关联，虽然这些方法提高了跟踪的健壮性，但是它们输入的是成对的帧，不能提取多帧的相关性，只能处理单帧遮挡。最近，随着端到端的目标检测器 DETR[13] 的提出，开始有学者提出了基于注意力机制的多目标跟踪算法[14-16]。虽然这些算法是端到端的 (联合检测和跟踪)，但是其中的检测部分也只是将基于卷积的检测器换成了基于 Transformer[17] 的检测器，仍是独立地对每一帧进行特征提取，没有对目标的时序信息进行直接地建模。

目前主流的多目标跟踪方法大多是单独地提取每一帧的信息，忽略了不同帧之间的关联，虽然近几年也有一些方法开始尝试构建不同帧之间相关性，但是它们都仅停留在相邻帧，没有对视频中存在的时序信息进行显式建模。而在视频目标检测和视频行人重识别领域中，视频时序信息已被证实在处理运动模糊，遮挡和小目标等问题上很有效。受此启发，本文提出了一种时空特征对齐的多目标跟踪方法。本文主要贡献如下：

1) 提出时空特征对齐的多目标跟踪方法，充分利用时空特征以及保证时间一致性，提升多目标跟踪性能；

2) 引入卷积门控递归单元，对视频时空信息进行建模，该结构可以学习整个历史帧序列信息，输入任意长度的视频，构建任意长度视频帧之间的相关性；

3) 设计特征对齐模块，利用前后帧目标的位置对应关系，将历史帧信息与当前帧信息对齐，保证时间一致性，降低误检率；

4) 将设计的方法在公开数据集 MOT17 和 MOT20 上进行实验验证，结果表明所提方法较基准方法提升明显且优于目前同类先进方法。特别是在 MOT20 上，MOTA 值达到了 67.4，IDF1 值达到了 70.6。

2 相关工作

本文方法从视频理解相关领域出发，探究多目标跟踪中视频时序信息的有效性，下面为这些领域中与本文方法相关的工作以及本文方法的不同之处。

2.1 多目标跟踪

多目标跟踪方法大致可以分为三类，分别为基于检测[2-3,18-19]，联合检测与重识别[9-10,20-22]以及联合检测与跟踪[11-12,14-16,23]。基于检测的算法将多目标跟踪任务分为四步，分别为目标检测、特征提取、相似度计算和数据关联。由于目标检测和行人重识别的迅速发展，大多数学者的目光聚焦在前两步，而后两步采用传统方法。SORT[2]是最早利用卷积神经网络检测行人的多目标跟踪算法之一，该算法依靠卡尔曼滤波[24]和匈牙利算法[25]来解决目标关联，但是相似度计算只利用了运动信息——检测框和跟踪框的交并比(intersection over union, IOU)，对于遮挡问题效果不佳，DeepSORT[3]在SORT的基础上引入行人重识别(re-identification, Re-ID)网络来提取目标的深度表观特征，使得数据关联更准确，还有一些方法利用了更复杂的特征，如Xu等人[8]使用了时空图卷积来提取轨迹深度特征表示。不过复杂特征的提取大大增加了计算量，算法实时性较差。为了提高实时性，JDE[9]提出联合检测与重识别这一跟踪范式，用一个网络来实现目标检测和Re-ID特征提取，平衡了跟踪精度和跟踪速度，而针对JDE方法在单一网络中检测和Re-ID特征存在不公平等问题，一系列算法如FairMOT[10]，CSTrack[20]，RelationTrack[21]，CorrTrack[22]相继提出。也有研究者尝试为MOT构建端到端的解决方案，也就是联合检测与跟踪，该范式旨在同时输出检测和跟踪结果，Tracktor[11]直接利用检测器的回归模块预测目标下一帧的位置来完成多目标跟踪任务，CenterTrack[12]通过在成对的图像上执行检测，并结合先前帧的目标检测结果来预测当前帧的目标位置偏移，从而将前后帧中相同目标建立起联系，实现多目标跟踪。ChainedTrack[26]使用相邻帧作为输入，并生成代表相同目标的框对，将跨帧关联问题转化成目标检测问题。简单有效的端到端目标检测器DETR[13]的出现给目标检测领域带来革新的同时，也给多目标跟踪带来了新思路，有学者开始构建基于Transformer[17]的端到端的多目标跟踪器，如TransTrack[14]，TrackFormer[15]，MOTR[16]，这些算法主要在DETR解码器的查询输入这块进行一定的改进以适应MOT任务。可以看出，MOT的发展与目标检测和行人重识别的发展是一致的，不过本文从视频目标检测和视频行人重识别出发来研究视频时序信息对MOT的重要性。

2.2 视频行人重识别

在视频行人重识别方面，视频比静止图像包含更丰富的空间和时间信息，基于视频的行人重识别最直接的方法是先把视频拆成一帧一帧的图像，利用深度学习提取每帧图像的帧级别特征，然后通过不同操作如平均池化或最大池化[27]，递归循环网络(recurrent neural networks, RNN)[28-29]和时间注意力[30]来聚合多帧特征得到视频级别特征。另一种策略是通过3D卷积同时捕获空间和时间信息[31]，不同于基于2D卷积的模型需要诸如循环网络来提取时间信息，3D卷积自然处理输入视频以输出时空特征。尽管性能良好，但3D卷积通常需要更多的计算和内存资源，因此本文方法没有采用3D卷积模型，而是采用先提取图像单帧级别特征，再聚合多帧特征这一策略。

2.3 视频目标检测

在视频目标检测方面，相比于图像目标检测，视频具有高冗余度的特性，其中包含了大量的时空信息[32]。充分利用好时序上下文关系，可以解决视频中连续帧之间的大量冗余的情况，提高检测速度[33]；还可以解决视频相对于图像存在的运动模糊、视频失焦、部分遮挡和奇异姿势等问题。对于高冗余度特性，学者们希望利用运动信息来进行检测，其中最常用的运动信息是光流。DFF[34]只对关键帧进行特征提取，而对于关键帧附近的非关键帧，通过计算光流来聚集关键帧特征，大大减少了计算量。对于时空信息的提取，T-CNN[35]用检测算法学习图像中目标的空间信息，用跟踪算法学习图像中目标的时序信息，D&T[36]利用孪生网络来提取不同帧的相关性也就是时序信息，STMN[37]在单帧检测器上加入时空存储模块来提取时空信息，使其能够处理任意长度的视频。与STMN结构类似，本文在通用特征提取模块后引入了ConvGRU[18]来提取时空信息，并用特征对齐模块来保证了时空一致性。

3 本文方法

本节对本文方法进行详细描述。首先在3.1节对本文方法整体架构进行概述，然后在3.2节、3.3节和3.4节详细介绍各个模块，分别为时空特征提取模块、检测头与Re-ID头和数据关联模块。

3.1 方法整体架构

针对目前主流多目标跟踪算法未能有效地提取时序信息这一问题，本文提出时空特征对齐的多目标跟

踪方法，结构如图 1 所示，是联合检测与重识别这一范式下的多目标跟踪方法。算法模型由通用特征提取、时空特征提取、检测头、Re-ID 头和数据关联 5 部分组成。给定连续视频帧序列$\{I^1, I^2, \cdots, I^n\}$, 本文方法将每个帧单独地通过骨干网络得到单帧级别特征图$\{F^1, F^2, \cdots, F^n\}$，本文方法的骨干网络与 FairMOT[10] 相同，采用 DLA-34[38] 网络来提取图像单帧级别特征，该网络包含很多高维特征与低维特征的连接，能更好地聚合空间信息和语义信息，提取目标位置与表观信息；对于任意时间步，先将存放历史帧序列信息的M^{t-1}通过特征对齐模块得到 *Aligned* M^{t-1}，然后F^t和 *Aligned* M^{t-1}一同输入到 ConvGRU[18] 得到M^t；经过通用特征提取模块和时空特征提取模块后，直接将特征图送入检测头和 Re-ID 头分别输出位置信息和 Re-ID 特征；然后利用位置信息和 Re-ID 特征计算当前帧检测目标$\{d^1, d^2, \cdots, d^N\}$与轨迹$\{T^1, T^2, \cdots, T^K\}$的相似度矩阵，最后结合分配算法实现数据关联完成多目标跟踪。

3.2 时空特征提取模块

本文方法使用 ConvGRU[18] 来学习目标的时空信息，ConvGRU 是如图 2 所示的门控循环单元 (gated recurrent unit, GRU) 的改进版本。GRU 常用于自然语言处理中捕捉序列数据的长时间依赖关系，不过自然语言处理领域处理的是一维信息，而图像是二维的，为了同时捕捉时间和空间信息，ConvGRU 将一维状态向量替换成二维状态特征图，将全连接层替换成卷积层。ConvGRU 的计算公式如下：

$$z^t = \sigma(W^z * F^t + U^z * M^{t-1}), \tag{1}$$

$$r^t = \sigma(W^r * F^t + U^r * M^{t-1}), \tag{2}$$

$$\widetilde{M}^t = \tanh\left(W * F^t + r^t * \left(U \cdot M^{t-1}\right)\right), \tag{3}$$

$$M^t = (1 - z^t) \cdot \widetilde{M}^t + z^t \cdot M^{t-1}, \tag{4}$$

其中$*$表示卷积，$\cdot$表示点乘，W^z，W^r，W 和U^z，U^r，U 都是 2D 卷积核，F^t表示当前帧特征图，M^{t-1}表示过去状态特征图，代表了整个历史帧信息。

由于视频中目标是运动的，目标在当前帧的空间位置与前一帧的空间位置不同，那么代表历史帧特征的M^{t-1}可能没有和当前帧特征F^t在空间位置上进行对齐，这可能会导致 ConvGRU 难以忘记历史帧目标的空间位置，从而叠加了未对齐的特征，造成拖尾现象——特征图上历史帧目标所在的空间位置存在高响应，从而使得检测器认为目标还处在前一时刻的空间位置，造成大量误检。为了解决这一问题，本文引入特征对齐模块如图 3，充分利用相邻帧之间的一致性信息。具体来说，就是根据当前帧特征图$\boldsymbol{F}^t$和前一帧特征图F^{t-1}的位置对应关系来修正过去状态特征图M^{t-1}，使其与$\boldsymbol{F}^t$对齐。首先计算F^t中位置(x,y)的特征向量$\boldsymbol{F}^t(x,y) \in R^D$与$\boldsymbol{F}^{t-1}$中位置$(x,y)$附近区域的特征向量$\boldsymbol{F}^{t-1}(x,y) \in R^D$的余弦相似度，然后对$M^{t-1}$进行加权使其对齐到当前帧特征图$\boldsymbol{F}^t$。具体计算如式 (5)

$$C_{x,y}(i,j) = \frac{\boldsymbol{F}^t(x,y) \cdot \boldsymbol{F}^{t-1}(x+i, y+j)}{\sum_{i,j\in\{-d,\cdots,d\}} \boldsymbol{F}^t(x,y) \cdot \boldsymbol{F}^{t-1}(x+i, y+j)}, \tag{5}$$

$$Aligned\ M^{t-1}(x,y) = \sum_{i,j\in\{-d,\cdots,d\}} C_{x,y}(i,j) \cdot M^{t-1}(x+i, y+j), \tag{6}$$

其中i, j限制在范围$[-d, d]$，d 是个超参数，本文设置 $d = 2$, 基于的假设是相邻帧不会有过大的位移，当然这样也可以减少计算量。

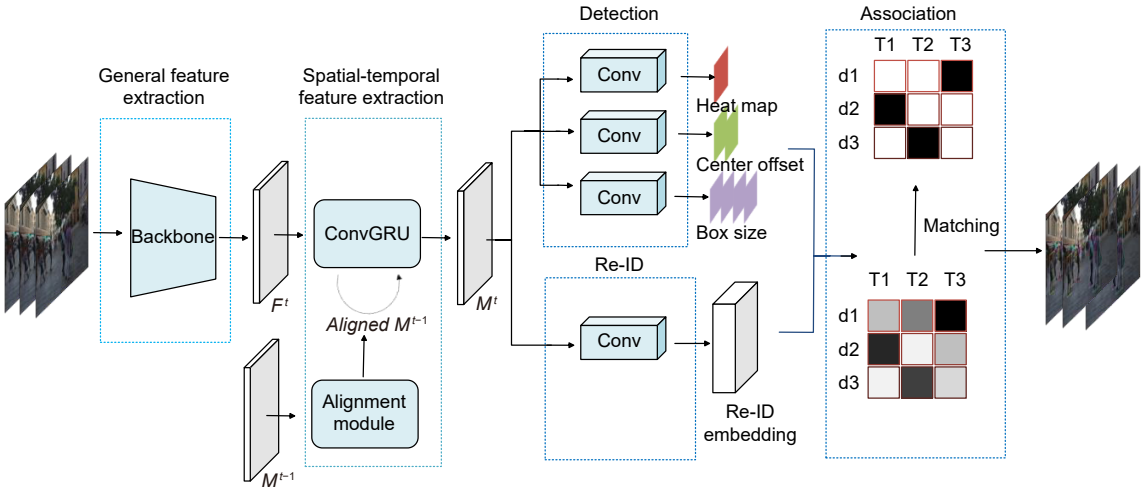


图 1 算法整体框架

Fig. 1 Overall framework of the algorithm

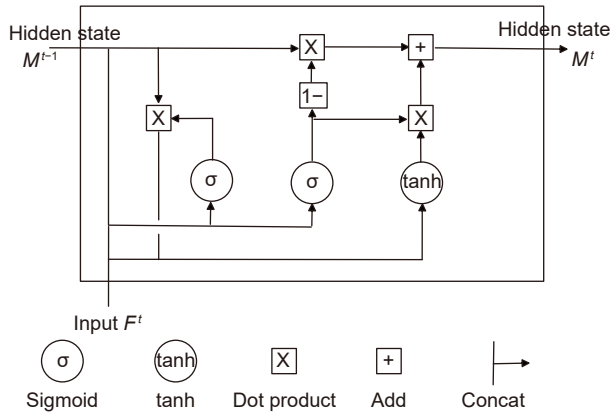


图 2 门控循环单元结构图

Fig. 2 Gated recurrent unit

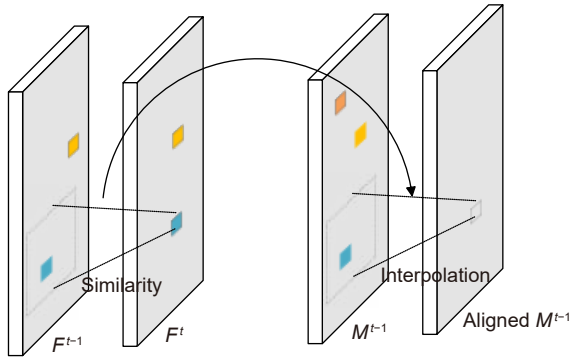


图 3 特征对齐

Fig. 3 Feature alignment

3.3 检测头和 Re-ID 头

检测模块主要由三个并行的卷积模块 (卷积核大小为 3x3，输出通道数为 256 的卷积+卷积核大小为 1x1 的卷积) 组成，分别输出目标中心点热力图，目标中心点偏移和检测框宽高。热力图分支负责预测目标中心点位置，训练时需要将标签转化为热力图形式来计算损失，假设目标真实框为$b=(x_1,y_1,x_2,y_2)$，则中心点为$(c_x,c_y)=\left(\frac{x_1+x_2}{2},\frac{y_1+y_2}{2}\right)$，经下采样得低分辨率坐标$c_d=\left(\frac{c_x}{4},\frac{c_y}{4}\right)$，则该目标的中心点分散至热力图上$H_{xy}=\exp\left(-\frac{(x-c_{dx})^2+(y-c_{dy})^2}{2\sigma^2}\right)$，其中$\sigma$为标准差。热力图分支的损失函数为：

$$L_{\text{heatmap}}=-\frac{1}{N}\cdot\sum_{xy}\begin{cases}\left(1-\widehat{H}_{xy}\right)^{\alpha}\log\widehat{H}_{xy}, & H_{xy}=1\\ \left(1-H_{xy}\right)^{\beta}\left(\widehat{H}_{xy}\right)^{\alpha}\log\left(1-\widehat{H}_{xy}\right) & \text{otherwise}\end{cases}, \tag{7}$$

其中，参数α用于控制易分类样本权重，参数β用于减少负样本权重占比，N是图像中心点个数，$\widehat{H}_{xy}$是热力图估计。中心点偏移分支用于估计目标中心点的偏移补偿，检测框宽高分支用于估计目标中心点到检测框四条边的距离，损失函数为：

$$L_{\text{box}}=\sum_{i=1}^{N}\left\|o^i-\hat{o}^i\right\|_1+\lambda_s\left\|s^i-\hat{s}^i\right\|_1, \tag{8}$$

其中，o^i和s^i分别为中心点位置和检测框宽高的真实值，$\hat{o}^i$和$\hat{s}^i$为对应的网络估计值。

Re-ID 模块旨在提取同一类别中可以区分不同身份的细粒度表观特征，主要在时空特征提取模块后应用卷积提取 128 维的特征图。训练时将 Re-ID 作为分类任务，将真值处目标特征向量经过一个线性分类层，得到每个 ID 分类的概率值$p(k),k\in[1,K]$，其中 K 为类别数目。损失函数为：

$$L_{\text{id}}=-\sum_{i=1}^{N}\sum_{k=i}^{K}Y^i(k)\ln p(k), \tag{9}$$

其中，$Y^i(k)$表示第 i 个目标的真实 ID 概率分布。本文同时训练检测任务和 Re-ID 任务，使用不确定性损失来自动平衡两个任务，计算如下：

$$L_{\text{det}}=L_{\text{heatmap}}+L_{\text{box}}, \tag{10}$$

$$L=\frac{1}{2}\left(\frac{1}{\mathrm{e}^{w_1}}L_{\text{det}}+\frac{1}{\mathrm{e}^{w_2}}L_{\text{id}}+w_1+w_2\right), \tag{11}$$

其中，w_1和w_2为可学习参数，用于平衡检测和重识别任务

3.4 数据关联

数据关联策略与 FairMOT[10] 保持一致。首先基于第一帧中检测到的框初始化轨迹片段。然后在后续的帧中，使用两阶段匹配策略实现检测框与轨迹片段的连接。在第一阶段，通过网络得到输入图像的目标位置信息和 Re-ID 特征，首先利用卡尔曼滤波和马氏距离排除相距较远的匹配，然后将 Re-ID 特征余弦距离D_{r}和马氏距离D_{m}融合在一起计算相似度$D=0.98D_{\text{r}}+0.02D_{\text{m}}$，利用匈牙利算法完成目标和轨迹的第一次匹配；在第二阶段，对未匹配的轨迹片段和未匹配的目标计算交并比 (Intersection over union, IoU)，然后利用匈牙利算法完成目标和轨迹的第二次匹配；最后更新轨迹，将未匹配的目标初始化为新轨迹，对未匹配的轨迹做记录，当轨迹连续 30 帧都没匹配到新目标，则丢失该轨迹。

4 实验结果与分析

4.1 数据集与模型评价

实验主要在多目标跟踪数据集 MOT17 和

MOT20 上进行，并与现有方法进行对比分析。MOT17 数据集主要标注目标为移动的行人，包含了不同天气状况、相机静止或运动、多个拍摄角度和光照变化的视频，涵盖了多目标跟踪过程中可能遇到各种挑战的场景。MOT17 数据集共 14 个视频序列，分为 7 个训练集和 7 个测试集，视频序列长度平均为 800 帧，其中训练集包含 112297 个检测框标注和 548 个 ID 标注且提供 3 种检测器 SDP、DPM 和 Faster R-CNN 的检测结果。为了进行公平的对比分析，实验在训练时还使用了与 FairMOT[10] 相同的额外数据集 ETH、CityPerson、CalTech、CUHK-SYSU、PRW、CrowdHuman。CityPerson 是行人检测数据集，数据是由车载摄像机在城市中收集，总计 25000 张图片，350000 个标注框；ETH 包含 5598 张图片。不过这两个数据集只提供了目标真实检测框，所以训练时忽略了这些数据集中的 Re-ID 损失。CalTech、CUHK-SYSU、PRW、MOT17 提供了目标真实检测框和 ID，可以用来同时训练检测分支和 Re-ID 分支。对于消融实验，本文使用上述 6 个额外数据集和 MOT17 的前半序列作为训练集，MOT17 的后半序列作为验证集。MOT 任务中的评价指标主要包括如下：

多目标跟踪准确度 (Multiple object tracking accurary, MOTA)：同时考虑误检、漏检和 ID 切换，能够直接衡量算法检测和跟踪的性能。计算公式如式 (12) 所示，其中，t 表示时间帧的索引，FN 表示漏检数，FP 表示误检数，IDSW 表示 ID 切换次数，GT 表示真实检测框数。

$$\frac{\sum_t \mathrm{FN}_t + \mathrm{FP}_t + \mathrm{IDSW}_t}{\sum_t \mathrm{GT}_t}. \tag{12}$$

识别 F1 值 (Identification F1 Score，IDF1)：用来衡量 ID 识别准确率与召回率之间的平衡性，评估跟踪器的 ID 识别性能。计算公式如式 (13) 所示，其中，IDTP 表示真阳性 ID，IDFP 表示假阳性 ID，IDFN 表示假阴性 ID，与检测指标的 TP、FP、FN 相对应。

$$\mathrm{IDF1} = \frac{2\mathrm{IDTP}}{2\mathrm{IDTP} + \mathrm{IDFP} + \mathrm{IDFN}}. \tag{13}$$

高阶跟踪精度 (Higher order tracking accuracy, HOTA)：同时考虑了检测精度、关联和定位精度。

误检数 (False positive, FP)：被预测为正样本的负样本数。

漏检数 (False negatives, FN)：被预测为负样本的正样本数。

命中轨迹比 (Mostly tracked targets, MT)：跟踪轨迹占真实轨迹 80% 以上的轨迹数与轨迹总数之比。

丢失轨迹比 (Mostly lost targets, ML)：跟踪轨迹占真实轨迹 20% 以下的轨迹数与轨迹总数之比。

ID 切换 (Identity switches, IDs)：目标 ID 切换的总数。

4.2 实验环境与训练细节

实验硬件环境为搭载 Inter Xeon(R) Platinum 8163 CPU 2.50GHz 处理器和 4 张 NVIDIA GeForce RTX 3090(24G 显存) 的深度学习服务器。软件环境为 Ubuntu 20.04 操作系统下的 Pytorch1.7 深度学习框架。实验按照 FairMOT[10] 的设置，采用了如随机翻转和随机裁剪等数据增强方法。为了解决不同帧率的问题，本文对视频序列进行了有间隔的随机采样。ConvGRU 的卷积核大小设为 5×5，特征对齐模块的局部区域大小设为 5。训练时输入图片大小为 1088×608，首先使用在 COCO 数据集预训练得到的模型参数来初始化骨干网络模型，然后采用 Adam 优化器训练 30 个轮次，batch size 设置为 12，初始化学习率为 1e-4，在第 20 个轮次更改学习率为 1e-5。

4.3 定量分析

为了验证本文提出的时空特征对齐的多目标跟踪方法的效果，在 MOT Challenge 上与当前一些先进 MOT 算法进行了指标对比。表 1、表 2 分别为在 MOT17、MOT20 测试集对比结果。从表 1 可以看出，本文方法在 IDF1 指标上超过大部分现有方法且具有较高的 MOTA 值。对比基准方法 FairMOT，IDF1 值由原 72.3 提升至 73.9，提升了 1.6，MOTA 值由原 73.7 提升至 74.2，提升了 0.5，MT 和 IDS 指标也有所提升。不过特征对齐模块需要计算前后帧点对点的相似度，计算量较大，导致帧率有所下降。值得注意的是，尽管 CSTrack 方法的 MOTA 值较本文方法高，但是 IDF1 值较本文方法低，这也可以从 FP、FN 指标和 MT、ML、IDS 指标中可以看出，CSTrack 方法的检测效果比本文方法好，但跟踪效果比本文方法差。MOT20 数据集的目标更加稠密，目标遮挡现象更严重，因此更具有挑战性。实验结果如表 2 所示，可以看出，本文方法在 MOTA 指标上超过大部分现有方法且具有较高的 IDF1 值，并且带来的性能提升比在 MOT17 数据集上更加明显，说明本文方法的时空特征提取模块在遮挡和小目标等困难场景中更能发挥作用。对比基准方法 FairMOT，IDF1 值由原 67.3 提升至 70.6，提升了 3.3，MOTA 值由原 61.8 提升至 67.4，提升了 5.6，IDS 指标也有所提升。值得注意的是，

表 1 本文方法与其他先进方法在 MOT17 数据集上的对比结果

Table 1 The tracking performance comparision between our method and other advanced methods on MOT17 data set

Method	Year	MOTA↑	IDF1↑	HOTA↑	FP↓	FN↓	MT↑	ML↓	IDS↓	FPS↑
TubeTK[39]	CVPR2020	63.0	58.6	48.0	27060	177483	31.2	19.9	5529	3.0
CTracker[26]	ECCV2020	66.6	57.4	49.0	22284	160491	32.2	24.2	5529	6.8
CenterTrack[12]	ECCV2020	67.8	64.7	52.2	**18489**	160332	34.6	24.6	3309	**22.0**
TraDes[40]	CVPR2021	69.1	63.9	52.7	20892	150060	36.4	21.5	3555	3.4
FairMOT[10]	IJCV2021	73.7	72.3	59.3	27507	117477	43.2	**17.3**	3303	18.9
TrackFormer[15]	CVPR2022	65.0	63.9	-	70443	123552	-	-	3528	-
MOTR[16]	ECCV2022	67.4	67.0	-	32355	149400	34.6	24.5	**1992**	-
CSTrack[20]	TIP2022	**74.9**	72.3	-	23847	**114303**	41.5	17.5	3567	16.4
Ours		74.2	**73.9**	**60.1**	27129	116337	**43.8**	19.1	2367	10.9

表 2 本文方法与其他先进方法在 MOT20 数据集上的对比结果

Table 2 The tracking performance comparision between our method and other advanced methods on MOT20 data set

Method	Year	MOTA↑	IDF1↑	HOTA↑	FP↓	FN↓	MT↑	ML↓	IDS↓	FPS↑
FairMOT[10]	IJCV2021	61.8	67.3	54.6	103440	**88901**	**68.8**	**7.6**	5243	**8.9**
TransTrack[14]	arXiv2021	64.5	59.2	-	28566	151377	49.1	13.6	3565	-
CorrTracker[22]	CVPR2021	65.2	**73.6**	-	29808	99510	47.6	12.7	3369	-
CSTrack[20]	TIP2022	66.6	68.6	54.0	**25404**	144358	50.4	15.5	3196	4.5
Ours		**67.4**	70.6	**55.6**	49358	117370	59.6	12.3	**2066**	4.8

尽管 CorrTracker 方法的 IDF1 指标较本文方法高，但 MT、ML 以及 IDS 这些评价跟踪器的指标都较本文方法低，说明本文方法的跟踪效果不比 CorrTracker 差。

4.4 消融实验

本文的消融实验使用上述 6 个额外数据集和 MOT17 的前半序列作为训练集，MOT17 的后半序列作为验证集。本文探究了 ConvGRU 和特征对齐模块对整体跟踪性能的影响。从表 3 可以看出，使用 ConvGRU 和特征对齐模块均能有效提升多目标跟踪性能，其中最重要的指标 MOTA 由原 69.1 提升至 70.0，IDF1 由原 72.8 提升至 74.8，但是误检率有所升高，不过加入特征对齐模块后有所缓解。值得注意的是，本文方法的 IDs 较基准方法也有所增加，但 ML 较高，ML 较低，导致 IDs 占据总匹配数较小，从 IDF1 指标也能看出整体跟踪性能更好。

本文还设计了消融实验探究视频序列输入长度对跟踪性能的影响，如表 4 所示。当视频序列输入长度从 2 增加到 8 时，MOTA 和 IDF1 指标分别提高了 1.1 和 1.3，说明视频序列输入长度的增加可以提高跟踪性能，模型能够很好地学习目标长时间的依赖关系。尽管随着视频序列输入长度的增加，IDs 也随之增加，但 MT 随之增加，ML 随之减小，导致 IDs 占据总匹配数的比例越来越小，因此匹配错误越来越低，这也从 IDF1 指标中可以看出。

表 3 不同模块对跟踪性能的影响

Table 3 The impact of different components on the overall tracking performance

Method	MOTA↑	IDF1↑	FP↓	FN↓	MT↑	ML↓	IDS↓
Baseline	69.1	72.8	**1976**	14443	143	53	**299**
Baseline+ConvGRU	69.6	73.4	2434	13729	150	**50**	321
Baseline+ConvGRU+Alignment Module	**70.0**	**74.8**	2201	**13715**	**153**	51	320

表 4 视频序列输入长度对跟踪性能的影响

Table 4 The impact of video sequence input length on the overall tracking performance

Input length	MOTA↑	IDF1↑	FP↓	FN↓	MT↑	ML↓	IDS↓
2	68.9	73.5	2412	14092	143	52	311
3	69.6	74.1	**2108**	13990	144	51	319
4	69.6	73.9	2156	13949	152	52	**293**
5	69.5	74.1	2221	13947	151	52	313
8	**70.0**	**74.8**	2201	**13715**	**153**	**51**	320

4.5 定性分析

除了在基准数据集上进行量化指标的测试，本节也对本文方法与基准方法进行了对比定性分析，通过可视化的结果来分析本文方法在面对多目标跟踪中遮挡、目标形变等问题时表现出的效果。与消融实验一样，将 MOT17 的前半序列作为训练集，MOT17 的后半序列作为验证集，本文在验证集上进行定性分析。图 4 表示本文方法和基准方法的多目标跟踪结果对比图，由于原数据集的图片过大，含有的目标比较多，不方便对比分析，所以截取了中间比较有代表性的一

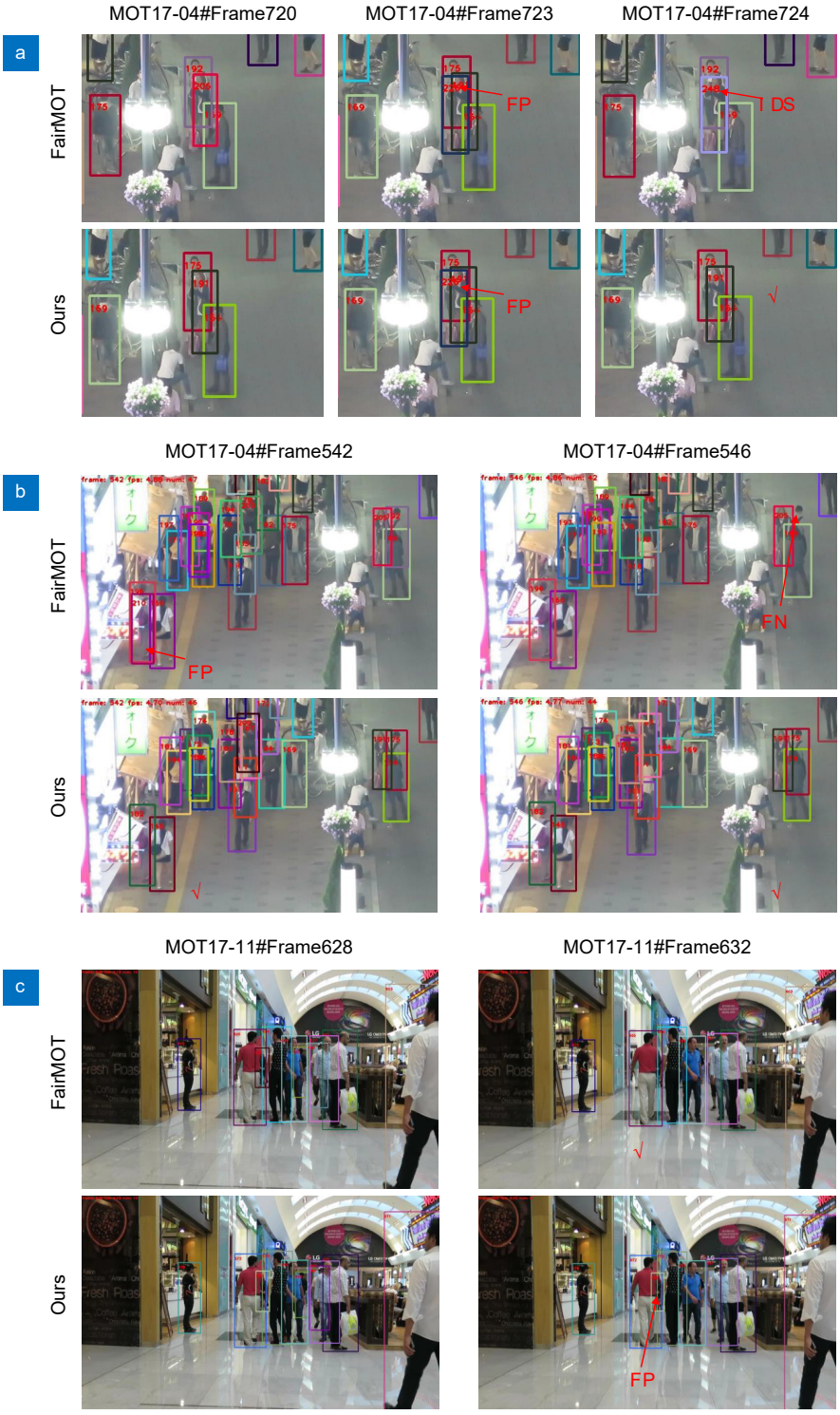


图 4 本文方法与基准方法在验证集上的可视化结果对比。(a) ID 切换；(b) 误检和漏检；(c) 特定的误检

Fig. 4 The visualization results comparison between baseline and our method on validation set. (a) ID switch; (b) FP and FN; (c) special FP

些跟踪结果来比较效果。

图 4 中 MOT17-04 视频序列，第 723 帧时基准方法和本文方法都出现了误检，但在第 724 帧时基准方法将目标关联到了第 723 帧的误检目标上，出现 ID 切换现象，而本文方法能排除误检干扰，保持目标 ID。这是因为原模型的 Re-ID 特征只利用了图像单帧信息，一旦出现误检且提取的 Re-ID 特征相似，很容易发生 ID 切换；而本文方法的时空特征提取模块可以充分利用前几帧的信息，使得提取的 Re-ID 特征更鲁棒，更能避免因为一时误检造成的 ID 切换。

图 4 中 MOT17-04 视频序列，第 542 帧中左下角的背着红色书包的行人被店铺招牌遮挡，第 546 帧中右上角穿着白色上衣背着书包的行人被周围的行人严重遮挡。在这样的情况下，基准方法分别出现了错误识别目标和丢失了部分目标的现象，而本文方法仍然能够正确框出目标，这充分说明了时空特征提取模块的作用。在当前帧严重遮挡而缺失信息的情况下，时空特征提取模块可以利用前几帧的信息进行补齐，使得本文方法在面对因为遮挡而出现误检和漏检等问题上更有鲁棒性。

但时空特征提取模块也存在一定的缺陷，会出现基于单帧的方法中不存在的误检现象。比如图 4 中 MOT17-11 视频序列，第 628 帧中最中间的穿着黑色衣服的行人随着时间的推移在第 632 帧时几乎被完全遮挡而消失在视野中，但本文方法仍将其框出，出现了误检现象。这是因为本文方法利用的时序模型具有视频外推的能力，即可以将历史帧信息传递到当前帧，使得目标虽然在当前帧已经消失了但模型仍然保留了目标过去信息，从而造成误检现象。

5 结　论

现有多目标跟踪方法大多是单独提取每一帧的信息，没有对视频中存在的时序信息进行显式建模，这使得方法在运动模糊、遮挡和小目标等场景中的性能显著下降。针对这一问题，本文提出了时空特征对齐的多目标跟踪方法，主要通过 ConvGRU 提取视频中的时空信息，不过由于前后帧目标的空间位置不同，且实验结果表明时序模型难以忘记过去帧目标所处位置，使得误检增多，因此进一步提出特征对齐模块将前后帧目标信息对齐。实验结果表明，本文方法可以有效提取时序信息，提升多目标跟踪性能，这也体现了多目标跟踪中时序信息的有效性。不过，特征对齐模块中相似度计算的运算量较大以及时序信息的引入会造成一些单帧检测器中不会出现的误检现象，影响检测器的性能，因此，下一步研究工作的重点在于更高效的特征对齐模块和检测模块上，通过改进检测模块来避免因为时序信息引入而出现的特定误检现象。

6 附　录

6.1 KITTI 数据集上的实验结果与分析

自动驾驶数据集 KITTI 共 50 个视频序列，分为 21 个训练集和 29 个测试集，包含市区、乡村和高速公路等场景采集的真实图像数据，可用于车辆多目标跟踪。值得注意的是，相比于 MOT 系列数据集，该数据集帧率较低，只有 10 fps。KITTI 车辆类测试集的定量结果如表 5。许多方法会使用额外的合成数据集或者 KITTI 点云数据集进行训练，为了进行公平对比，对比的算法需未使用额外数据集。而且，大多数使用了 Re-ID 模块的算法未在 KITTI 上进行实验。因此，可对比的算法较少。对比算法中 CenterTrack 是只利用了运动信息，而 QDTrack 是采用对比学习提取了鲁棒的 Re-ID 特征。

KITTI 上的定量结果显示了本文方法的一个缺陷，即在本文方法的特征对齐模块中，计算前后帧相似度是在局部区域，假设目标在相邻帧不会有过大的位移，这在目标速度过快或帧率过低时有些不成立。KITTI 数据集帧率低且车辆速度快，尽管如此，对于 KITTI 的大多数场景，本文方法仍有效。从可视化结果可以看出，在相机运动目标静止以及相机静止目标运动这些场景，本文方法跟踪性能良好。对于图 5 中 0008 视频序列的跟车和转向等场景，相机一直运动，本文

表 5　本文方法与其他先进方法在 KITTI 车辆类测试集上的对比结果

Table 5　The tracking performance comparision between our method and other advanced methods on KITTI vehicle class test set

Method	Year	HOTA↑	MOTA↑	FP↓	FN↓	MT↑	ML↓	IDS↓
CenterTrack[12]	ECCV2020	**73.0**	**88.8**	**2703**	886	**82.2**	15.4	**254**
QDTrack[41]	CVPR2021	68.5	84.9	4320	549	69.5	**3.8**	313
Ours		69.6	82.2	5403	**433**	58.6	8.3	274

图 5 本文方法在 KITTI 测试集上的可视化结果。图片左侧为视频号。图片左上角为帧号

Fig. 5 Visualization results of this method on the KITTI test set. The video number is in the left side of the figure. The frame number is in the upper left of the figure

方法稳定跟踪目标，对于图 5 中 0010 视频序列的红绿灯场景，相机静止但车辆的视角和形状变化较大，本文方法稳定跟踪目标。不过对于图 5 中 0002 视频序列的会车场景，本文方法会出现跟踪框漂移现象。这是因为会车时车辆和相机是相对运动的，目标速度最快，此时的相邻帧位移特别大。会车在自动驾驶场景频繁出现，导致本文方法的整体跟踪性能不佳。很自然的改进点是特征对齐模块中前后帧相似度计算的区域采取一定方式自适应调整。

尽管如此，对比同样使用了 Re-ID 模块的 QDTrack，本文方法的 HOTA、IDS 指标较好，但在同时衡量检测和跟踪的 MOTA 指标上较低，表明本文方法跟踪器较好，提取的 Re-ID 特征较好，而检测器较差，这也可以从 FP 这个指标可以看出，因为本文方法在会车场景时检测框会漂移，漂移的检测框与真实检测框的 IOU 可能小于 0.5，就被认为是误检。不过使用 Re-ID 模块的方法对比未使用 Re-ID 模块的 CenterTrack，跟踪性能都较差，表明提取的 Re-ID 特征不够鲁棒，这可以使用额外的车辆重识别数据弥补。

利益冲突: 所有作者声明无利益冲突

参考文献

[1] Ciaparrone G, Sánchez F L, Tabik S, et al. Deep learning in video multi-object tracking: a survey[J]. *Neurocomputing*, 2020, **381**: 61−88.

[2] Bewley A, Ge Z Y, Ott L, et al. Simple online and realtime tracking[C]//2016 *IEEE International Conference on Image Processing* (*ICIP*), 2016: 3464–3468. https://doi.org/10.1109/ICIP.2016.7533003.

[3] Wojke N, Bewley A, Paulus D. Simple online and realtime tracking with a deep association metric[C]//*2017 IEEE International Conference on Image Processing*, 2018: 3645–3649. https://doi.org/10.1109/ICIP.2017.8296962.

[4] E G, Wang Y X. Multi-candidate association online multi-target tracking based on R-FCN framework[J]. *Opto-Electron Eng*, 2020, **47**(1): 190136.
鄂贵, 王永雄. 基于R-FCN框架的多候选关联在线多目标跟踪[J]. 光电工程, 2020, **47**(1): 190136.

[5] Berclaz J, Fleuret F, Fua P. Robust people tracking with global trajectory optimization[C]//*2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (*CVPR'06*), 2006: 744–750. https://doi.org/10.1109/CVPR.2006.258.

[6] Pirsiavash H, Ramanan D, Fowlkes C C. Globally-optimal greedy algorithms for tracking a variable number of objects[C]//*CVPR 2011*, 2011: 1201–1208. https://doi.org/10.1109/CVPR.2011.5995604.

[7] Brasó G, Leal-Taixé L. Learning a neural solver for multiple object tracking[C]//*Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 6246–6256. https://doi.org/10.1109/CVPR42600.2020.00628.

[8] Xu J R, Cao Y, Zhang Z, et al. Spatial-temporal relation networks for multi-object tracking[C]//*Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision*, 2019: 3987–3997. https://doi.org/10.1109/ICCV.2019.00409.

[9] Wang Z D, Zheng L, Liu Y X, et al. Towards real-time multi-object tracking[C]//*Proceedings of the 16th European Conference on Computer Vision*, 2020: 107–122. https://doi.org/10.1007/978-3-030-58621-8_7.

[10] 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]. *Int J Comput Vision*, 2021, **129**(11): 3069−3087.

[11] Bergmann P, Meinhardt T, Leal-Taixé L. Tracking without bells

and whistles[C]//*Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision*, 2019: 941–951. https://doi.org/10.1109/ICCV.2019.00103.

[12] Zhou X Y, Koltun V, Krähenbühl P. Tracking objects as points[C]//*Proceedings of the 16th European Conference on Computer Vision*, 2020: 474–490. https://doi.org/10.1007/978-3-030-58548-8_28.

[13] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers[C]//*Proceedings of the 16th European Conference on Computer Vision*, 2020: 213–229. https://doi.org/10.1007/978-3-030-58452-8_13.

[14] Sun P Z, Cao J K, Jiang Y, et al. Transtrack: Multiple object tracking with transformer[Z]. arXiv: 2012.15460, 2020. https://arxiv.org/abs/2012.15460.

[15] Meinhardt T, Kirillov A, Leal-Taixé L, et al. TrackFormer: Multi-object tracking with transformers[C]//*Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022: 8834–8844. https://doi.org/10.1109/CVPR52688.2022.00864.

[16] Zeng F G, Dong B, Zhang Y A, et al. MOTR: End-to-end multiple-object tracking with transformer[C]//*Proceedings of the 17th European Conference on Computer Vision*, 2022: 659–675. https://doi.org/10.1007/978-3-031-19812-0_38.

[17] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//*Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017: 6000–6010.

[18] Ballas N, Yao L, Pal C, et al. Delving deeper into convolutional networks for learning video representations[C]//*Proceedings of the 4th International Conference on Learning Representations*, 2015.

[19] Yu F W, Li W B, Li Q Q, et al. POI: Multiple object tracking with high performance detection and appearance feature[C]//*Proceedings of the European Conference on Computer Vision*, 2016: 36–42. https://doi.org/10.1007/978-3-319-48881-3_3.

[20] Liang C, Zhang Z P, Zhou X, et al. Rethinking the competition between detection and ReID in multiobject tracking[J]. *IEEE Trans Image Process*, 2022, **31**: 3182−3196.

[21] Yu E, Li Z L, Han S D, et al. RelationTrack: Relation-aware multiple object tracking with decoupled representation[J]. *IEEE Trans Multimedia*, 2022. https://doi.org/10.1109/TMM.2022.3150169.

[22] Wang Q, Zheng Y, Pan P, et al. Multiple object tracking with correlation learning[C]//*Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021: 3875–3885. https://doi.org/10.1109/CVPR46437.2021.00387.

[23] Tokmakov P, Li J, Burgard W, et al. Learning to track with object permanence[C]//*Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision*, 2021: 10840–10849. https://doi.org/10.1109/ICCV48922.2021.01068.

[24] Welch G, Bishop G. *An Introduction to the Kalman Filter*[M]. Chapel Hill: University of North Carolina at Chapel Hill, 1995.

[25] Kuhn H W. The Hungarian method for the assignment problem[J]. *Naval Res Logist Q*, 1955, **2**(1–2): 83–97.https://doi.org/10.1002/nav.3800020109.

[26] Peng J L, Wang C A, Wan F B, et al. Chained-tracker: Chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking[C]//*Proceedings of the 16th European Conference on Computer Vision*, 2020: 145–161. https://doi.org/10.1007/978-3-030-58548-8_9.

[27] Zheng L, Bie Z, Sun Y F, et al. MARS: A video benchmark for large-scale person re-identification[C]//*Proceedings of the 14th European Conference on Computer Vision*, 2016: 868–884. https://doi.org/10.1007/978-3-319-46466-4_52.

[28] McLaughlin N, Del Rincon J M, Miller P. Recurrent convolutional network for video-based person re-identification[C]//*Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 1325–1334. https://doi.org/10.1109/CVPR.2016.148.

[29] Zhou Z, Huang Y, Wang W, et al. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification[C]//*Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 6776–6785. https://doi.org/10.1109/CVPR.2017.717.

[30] Fu Y, Wang X Y, Wei Y C, et al. STA: Spatial-temporal attention for large-scale video-based person re-identification[C]//*Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, 2019: 8287–8294. https://doi.org/10.1609/aaai.v33i01.33018287.

[31] Li J N, Zhang S L, Huang T J. Multi-scale 3D convolution network for video based person re-identification[C]//*Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, 2019: 8618–8625. https://doi.org/10.1609/aaai.v33i01.33018618.

[32] Wang D C, Bai C S, Wu K J. Survey of video object detection based on deep learning[J]. *J Front Comput Sci Technol*, 2021, **15**(9): 1563−1577.
王迪聪, 白晨帅, 邬开俊. 基于深度学习的视频目标检测综述[J]. 计算机科学与探索, 2021, **15**(9): 1563−1577.

[33] Lu K L, Xue J, Tao C B. Multi target tracking based on spatial mask prediction and point cloud projection[J]. *Opto-Electron Eng*, 2022, **49**(9): 220024.
陆康亮, 薛俊, 陶重犇. 融合空间掩膜预测与点云投影的多目标跟踪[J]. 光电工程, 2022, **49**(9): 220024.

[34] Zhu X Z, Xiong Y W, Dai J F, et al. Deep feature flow for video recognition[C]//*Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 4141–4150. https://doi.org/10.1109/CVPR.2017.441.

[35] Kang K, Ouyang W L, Li H S, et al. Object detection from video tubelets with convolutional neural networks[C]//*Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 817–825. https://doi.org/10.1109/CVPR.2016.95.

[36] Feichtenhofer C, Pinz A, Zisserman A. Detect to track and track to detect[C]//*Proceedings of the 2017 IEEE International Conference on Computer Vision*, 2017: 3057–3065. https://doi.org/10.1109/ICCV.2017.330.

[37] Xiao F Y, Lee Y J. Video object detection with an aligned spatial-temporal memory[C]//*Proceedings of the 15th European Conference on Computer Vision*, 2018: 494–510. https://doi.org/10.1007/978-3-030-01237-3_30.

[38] Yu F, Wang D Q, Shelhamer E, et al. Deep layer aggregation[C]//*Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 2403–2412. https://doi.org/10.1109/CVPR.2018.00255.

[39] Pang B, Li Y Z, Zhang Y F, et al. TubeTK: adopting tubes to track multi-object in a one-step training model[C]//*Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and*

Pattern Recognition, 2020: 6307–6317. https://doi.org/10.1109/CVPR42600.2020.00634.

[40] Wu J J, Cao J L, Song L C, et al. Track to detect and segment: an online multi-object tracker[C]//*Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021: 12347–12356. https://doi.org/10.1109/CVPR46437.2021.01217.

[41] Pang J M, Qiu L L, Li X, et al. Quasi-dense similarity learning for multiple object tracking[C]//*Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021: 164–173. https://doi.org/10.1109/CVPR46437.2021.00023.

作者简介

程稳 (1998-)，男，硕士研究生，主要研究多目标跟踪。

E-mail: chengwen20@mails.ucas.ac.cn

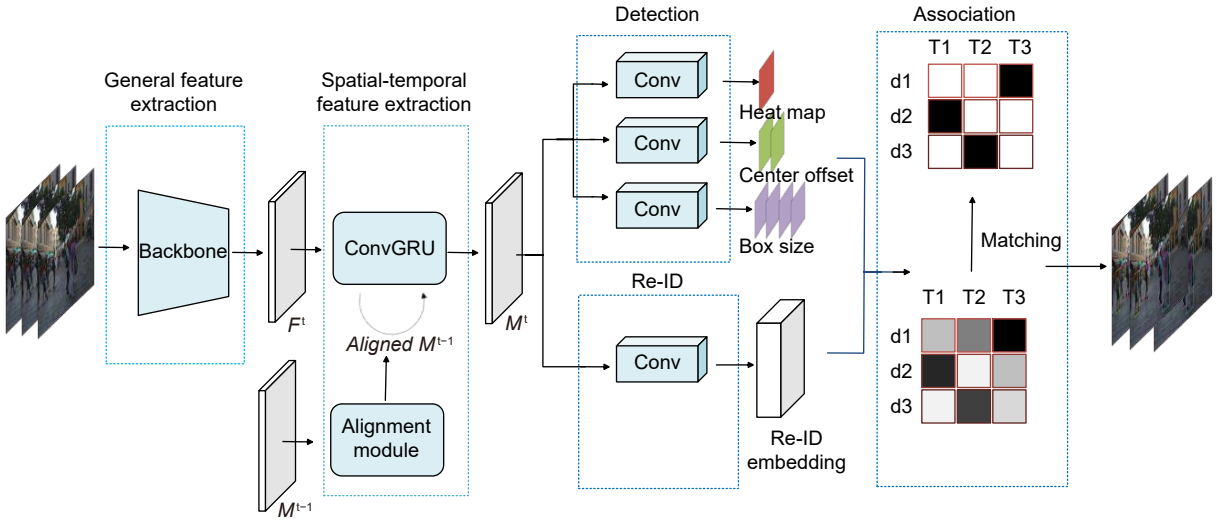
【通信作者】陈忠碧，女，副研究员，硕士生导师，主要从事目标检测识别与跟踪技术研究。

E-mail: chenzb@ioe.ac.cn

扫描二维码，获取PDF全文

Multiple object tracking with aligned spatial-temporal feature

Cheng Wen[1,2,3], Chen Zhongbi[2*], Li Qingqing[2], Li Meihui[2], Zhang Jianlin[2], Wei Yuxing[2]



Overall framework of the algorithm

Overview: Multiple object tracking (MOT) is an important task in computer vision. It is widely used in the fields of surveillance video analysis and automatic driving. MOT is to locate multiple objects of interest, maintain the unique identification number (ID) of each object, and record continuous tracks. The difficulty of multi-target tracking is false positives (FP), false negatives (FN), ID switches (IDs), and the uncertainty of the target number. Most of the MOT methods improve object detection and data association, usually ignoring the correlation between different frames. Although some methods have tried to construct the correlation between different frames in recent years, they only stay in the adjacent frames and do not explicitly model the temporal information in the video. They don't make good use of the temporal information in the video, which makes the tracking performance significantly degraded in motion blur, occlusion, and small target scenes. In order to solve these problems, this paper proposes a multiple object tracking method with the aligned spatial-temporal feature. First, the convolutional gated recurrent unit (ConvGRU) is introduced to encode the spatial-temporal information of the object in the video; By considering the whole history frame sequence, this structure effectively extracts the spatial-temporal information to enhance the feature representation. However, the target in the video is moving, and the spatial position of the target in the current frame is different from that in the previous frame, and ConvGRU is difficult to forget the spatial position of the target in the historical frame, thus overlaying the misaligned features, resulting in the spatial position of the target in the historical frame on the feature map has a high response, which makes the detector think that the target is still in the spatial position of the previous frame. Then, the feature alignment module is designed to ensure the time consistency between the historical frame information and the current frame information to reduce the false detection rate. Finally, this paper tests MOT17 and MOT20 datasets, and the multiple object tracking accuracy (MOTA) values are 74.2 and 67.4, respectively, which are increased by 0.5 and 5.6 compared with the baseline FairMOT method. Our identification F1 score (IDF1) value is 73.9 and 70.6, respectively, which is increased by 1.6 and 3.3 compared with the baseline FairMOT method. In addition, the qualitative and quantitative experimental results show that the overall tracking performance of this method is better than that of most of the current advanced methods.

Cheng W, Chen Z B, Li Q Q, et al. Multiple object tracking with aligned spatial-temporal feature[J]. *Opto-Electron Eng*, 2023, **50**(6): 230009; DOI: 10.12086/oee.2023.230009

Foundation item: National Natural Science Foundation of China (62101529)

[1]National Key Laboratory of Optical Field Manipulation Science and Technology, Chinese Academy of Sciences, Chengdu, Sichuan 610209 China; [2]Institute of Optics and Electronics, Chinese Academy of Science, Chengdu, Sichuan 610209 China; [3]University of Chinese Academy of Science School of Electronic, Electrical, Communication Engineering, Beijing 100049 China

* E-mail: chenzb@ioe.ac.cn