计算机科学
COMPUTER SCIENCE

基于互相关注意力的链式帧处理多目标跟踪算法

陈云芳，陆洋洋，周鑫，张伟

引用本文

陈云芳，陆洋洋，周鑫，张伟.基于互相关注意力的链式帧处理多目标跟踪算法[J]．计算机科学, 2023, 50(1): 131-137.

CHEN Yunfang, LU Yangyang, ZHOU Xin, ZHANG Wei. Multi-object Tracking Based on Cross-correlation Attention and Chained Frames [J]. Computer Science, 2023, 50(1): 131-137.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.211100097

基于互相关注意力的链式帧处理多目标跟踪算法

陈云芳　陆洋洋　周　鑫　张　伟
南京邮电大学计算机学院　南京 210023
(chenyf@njupt.edu.cn)

摘　要　多目标跟踪的一阶段方法因其在推理速度方面的优势逐渐成为主流。然而,与两阶段方法相比,其跟踪精度较差。一方面是因为采用单幅图像输入,目标间的关联性不强,容易导致目标丢失,另一方面忽视了检测和跟踪两个任务之间的差异性。为了减轻上述限制,提出了一种基于互相关注意力的链式帧处理多目标跟踪算法(MOT-CCC)。MOT-CCC将连续的两帧图片作为输入,将目标关联问题转化为两帧检测框对回归的问题,增强了目标间的关联性;采用互相关注意力模块将检测任务和身份识别任务解耦,以平衡并减少这两个任务之间的竞争。此外,所提算法将目标检测、特征提取和数据关联3个模块融合到一个网络中,实现了端到端的优化,提高了跟踪准确性,减少了跟踪耗时。在MOT16和MOT17基准测试中,MOT-CCC比原有的基准CTracker算法的MOTA提高了1.3%,FP减少了13%。

关键词:多目标跟踪;链式跟踪器;互相关注意力;一阶段方法;端到端
中图法分类号　TP391

Multi-object Tracking Based on Cross-correlation Attention and Chained Frames

CHEN Yunfang, LU Yangyang, ZHOU Xin and ZHANG Wei
School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, China

Abstract　The one-stage method of multi-object tracking(MOT) has gradually become the mainstream of MOT due to its advantages in reasoning speed. However, compared with the two-stage method, its tracking accuracy is poor. One reason is that the target is easy to be lost due to the use of single frame input that cause the correlation between the targets is not strong, the other is that the difference between the two tasks of detection and tracking is ignored. In order to alleviate the limitations, a multi-object tracking algorithm based on cross-correlation attention and chained frames(MOT-CCC) is proposed. MOT-CCC takes two consecutive frames as input, and converts the target association problem into a two-frame detection frame pair regression problem, which enhances the correlation between targets. The cross-correlation attention module decouples the detection task and the identification task to balance and reduce the competition between the two tasks. In addition, the proposed algorithm integrates the three modules of target detection, feature extraction and data association into one whole network to achieve end-to-end optimization, which improves tracking accuracy and reduces tracking time. In the MOT16 and MOT17 benchmark tests, compared with the benchmark CTracker algorithm, the MOTA of MOT-CCC increases by 1.3% and the FP decreases by 13%.

Keywords　Multi-object tracking, Chained tracker, Cross-correlation attention, One-shot, End-to-End

1　引言

视觉是人类获取外界信息的重要途径,近80%的环境信息如颜色、亮度、形状、运动、深度等来自视觉。计算机视觉技术(Computer Vision,CV)让计算机能够像人类一样“看世界”。它利用摄像头来模拟人眼的功能,实现对物体的提取、识别和跟踪功能。目标跟踪是计算机视觉中最具挑战性的问题之一,可以分为单目标跟踪和多目标跟踪。单目标跟踪(Single Object Tracking,SOT)的目标是根据第一帧给定的矩形框,在接下来的连续帧中成功定位到该目标,主要方法是通过目标的表观建模或者运动特征建模,来处理光照、形变、遮挡等问题;多目标跟踪(Multiple Object Tracking,MOT)的目标是给定一个图像序列,找到图像序列中运动的多个物体,并给定不同帧中的运动物体一个特定的编号,然后得到不同物体的运动轨迹。除了单目标跟踪会遇到的问题外,多目标跟踪还需要解决目标的频繁遮挡、目标轨迹的开始和终止时刻无法预知、目标太小、表观相似、目标间交互等问题,处理方式更为复杂。相比单目标跟踪,多目标跟踪具有更广泛的应用场景,如自动驾驶、运动姿态分析和交通监控,近年来受到了越来越多的关注。

到稿日期:2021-11-08　返修日期:2022-06-30
基金项目:国家重点研发计划(2019YFB2101700)
This work was supported by the National Key R&D Program of China(2019YFB2101700).
通信作者:张伟(zhangw@njupt.edu.cn)

目前，MOT 可以概括为两类方法：两阶段方法和一阶段方法。两阶段方法遵循跟踪-检测[1-3]范式，它将 MOT 分为两个独立的任务：1)通过目标检测来获取目标边界框；2)通过目标特征提取和关联来寻找目标在整个时间过程中的轨迹。在检测阶段，卷积神经网络(Convolutional Neural Network，CNN)检测器(如 Yolov3[4]和 Faster-RCNN[5])被用于通过多个边界框定位图像中所有感兴趣的对象。在特征提取和关联阶段，基于重新识别进行跨帧数据关联[6]。在两阶段方法中，早期的跟踪算法使用运动、形状、外观等特征描述目标检测与轨迹的相关性，建立相关矩阵。匈牙利算法[7]、JPDA[8]、MHT[9]等算法以相关矩阵为输入完成数据关联。后来，大多数算法利用外观特征，特别是来自深度神经网络的复杂输出特征。例如，DeepSORT[2]使用预先提供的检测结果，并用离线训练的深度重新识别模型和卡尔曼滤波器运动模型将它们关联起来。Siamese CNN[10]使用 Siamese 网络直接学习一对检测目标之间的相似性。Deep Affinity Network[11]采用以两个视频帧为输入的连体网络，提取多尺度外观嵌入并输出所有检测目标对之间的相似性分数。文献[12]使用提取的视觉特征以及动态和位置特征进行对象关联。文献[13]将 CNN 的输出与形状和运动模型相结合。尽管在性能上有提升，但是计算耗时，模型需要从每个单独的边界框中提取 ReID 特征。此外，两阶段方法的准确性很大程度上取决于检测器的性能。

随着深度学习中多任务学习的成熟以及其在推理速度上的优势，一阶段方法在 MOT 研究界引起了更多的关注。一阶段模型将检测和表示任务集成到一个统一的系统中，缩短了推理时间，提升了系统的实时性。具体来说，一阶段框架由 3 部分组成，即特征提取器、检测分支和重新识别分支。例如，文献[14]建立在 Faster-RCNN[5]之上，并使其边界框回归头具有多种用途，不仅可以细化检测到的边界框，而且还可以作为回归对象的单个对象跟踪器从前一帧到当前帧的位置。联合检测嵌入(JDE)[15]方法将训练过程建模为具有锚分类、框回归和嵌入学习的多任务学习问题。JDE 架构选择 Feature Pyramid Network(FPN)[16]作为其基础架构，以提取多尺度特征。在检测分支中，它利用基于锚的检测范式，在特征图中的每个位置设置一定数量的不同大小的锚。同时，在重识别分支中，与一般的重识别一致，将重识别作为网络设计的分类任务。但是，目标可能会出现在两个锚点的中间位置，导致检测任务和身份识别任务之间存在歧义。因此，基于锚的方法会导致边界框和嵌入特征无法对齐。一些研究[17-18]表明，无锚方法更适合一阶段框架的检测分支，它们直接从特征图中对物体和背景进行分类，并回归边界框。然而，一阶段方法的跟踪精度不如两阶段方法，一阶段方法结合了检测和身份识别，没有考虑学习中不同任务的特异性和共性，它能够同时提供来自共享特征图的目标位置及其相关的嵌入特征。然而，使用相同的图像输入同时训练检测和重新识别两个任务，忽略了这两个任务之间的本质区别，过度竞争导致性能退化。当前的一阶段方法都采用单帧输入，目标间的关联性不强，特别是在有遮挡时容易导致目标丢失。

为了缓解一阶段方法的上述问题，本文提出了一种基于互相关注意力的链式帧处理多目标跟踪算法(Multi-object Tracking Based on Cross-correlation Attention and Chained Frames，MOT-CCC)，用于提升一阶段跟踪算法的性能。MOT-CCC 的贡献主要体现在：

(1)链式输入为连续两帧图片，将目标关联问题转化为两帧检测框对回归的问题，增强了目标间的关联性。

(2)互相关注意力模块将检测任务和身份识别任务解耦，以平衡并减少这两个任务之间的竞争。

(3)分析目标检测和身份识别任务的特征的共性和特异性。对于特异性学习，通过反映不同特征通道之间相关性的自相关来增强每个任务的特征表示。对于共性学习，通过交叉关系机制来学习两个任务之间的共享信息。

2 基于互相关注意力的链式跟踪网络

2.1 概述

基于互相关注意力的链式跟踪网络 MOT-CCC 是一种融合目标检测、特征提取和数据关联 3 个模块的完全端到端跟踪算法，如图 1 所示。

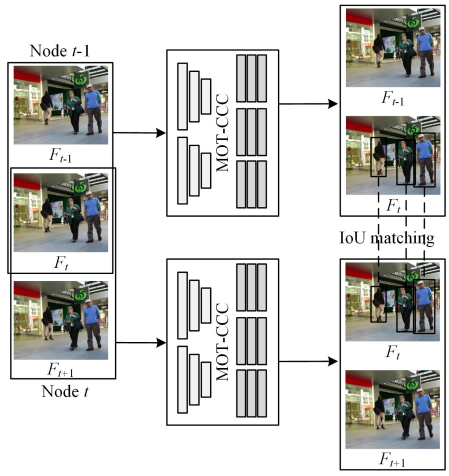

图 1 链式帧处理多目标跟踪算法概述

Fig. 1 Overview of multi-object tracking algorithm and chained frames

网络的输入是连续的两帧图片，在经过 MOT-CCC 处理后，相邻两个节点的目标关联采用公共帧的检测框对回归。MOT-CCC 采用了一种互相关注意力模块，使单独的分支学习依赖于任务的表征，将两帧的组合特征输入互相关注意力模块中，得到两个特征分支，并分别将其用于两个不同的任务，一个分支用于分类和回归，另一个分支用于身份识别。另一方面，为使得检测框对回归分支更加关注图像中的有效信息，MOT-CCC 保留了联合注意力模块，将目标分类分支与身份识别分支集中在回归分支上，根据分类置信度对边界框对进行过滤，生成的框对可以根据它们在公共帧中的框使用 IOU(联合相交)匹配关联，实现将所有相邻的节点即链节点依次链接，从而实现跟踪过程。

2.2 链式帧处理多目标跟踪算法的网络结构

MOT-CCC 把两个相邻帧图片输入到骨干网 ResNet-50 中，然后通过特征金字塔网络(FPN)生成多尺度特征表示，将来自两帧之间的特征图连接在一起，经过互相关注意力网络生成两个特征图，一个馈入分类和回归分支，另一个馈入身份验证分支，分类、身份识别输入到边界框回归分支，由于两个

分支是互补的，充分利用了分类分支与身份验证分支的信息，此外，不同节点之间因为公共帧的存在而仅有微小的差异，因此使用简单的 IOU 匹配即可以完成相邻节点之间的关联，MOT-CCC 的骨干网络结构如图 2 所示。

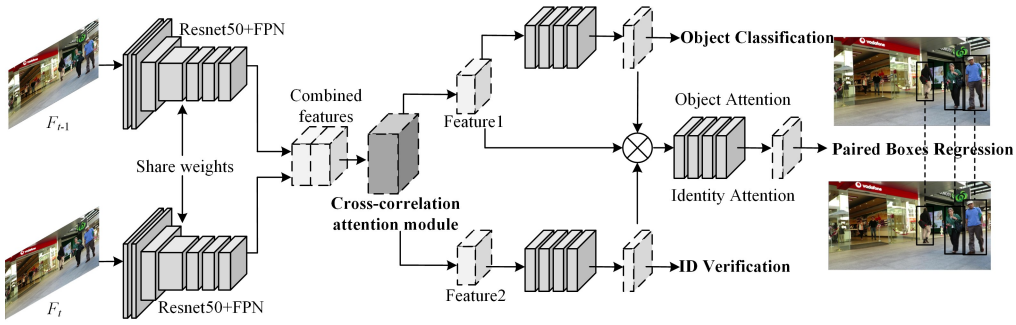

图 2 MOT-CCC 的骨干网络结构

Fig. 2 Backbone network structure of MOT-CCC

给定包含 N 帧的图像序列 $\{F_t\}_{t=1}^{N}$，多目标跟踪任务旨在针对出现的所有帧中的所有感兴趣对象，输出所有边界框 $\{g_t\}_{t=1}^{N}$ 和身份标签 $\{y_t^{GT}\}_{t=1}^{N}$。F_t 表示第 t 帧，g_t 表示第 t 帧中目标 K_t 数的真实边界框，y_t^{GT} 表示它们的身份。与其他仅需要一个帧作为输入的 MOT 模型不同，MOT-CCC 需要两个相邻的帧作为输入，本文称之为链节点。第一个链节点是 (F_1,F_2)，最后一个链节点（即第 N 个）是 (F_N,F_{N+1})。由于 F_N 是最后一帧，因此我们仅将 F_N 的副本作为 F_{N+1}。将 (F_{t-1},F_t) 作为输入，网络可以在两个帧中的相同目标生成边界框对 $\{(D_{t-1}^i,\hat{D}_t^i)\}_{i=1}^{n_{t-1}}$，其中 n_{t-1} 表示同一目标对的数量，D_{t-1}^i 和 $\hat{D}_t^i$ 分别表示节点内 F_{t-1} 与 F_t 中相同目标的两个边界框。类似地，也可以在下一个节点 (F_t,F_{t+1}) 中获得成对的边界框 $\{(D_t^j,\hat{D}_{t+1}^j)\}_{j=1}^{n_t}$，$D_t^j$ 和 $\hat{D}_{t+1}^j$ 表示位于相邻节点公共帧中的同一目标检测到的边界框，则这两个边界框之间应该只有微小的差异。因此可以使用简单的匹配策略将两个边界框链接起来，而不是像标准 MOT 方法那样使用复杂的外观特征。通过计算 $\hat{D}_t^i$ 和 D_t^j 中的框之间的 IOU（交并比），来获得 IOU 亲和力，通过应用 Kuhn-Munkres（KM）算法[19]来匹配 $\hat{D}_t^i$ 和 D_t^j 中检测到的框。对于每个匹配的框对 $\hat{D}_t^i$ 和 D_t^j，通过附加 D_t^j 来更新 $\hat{D}_t^i$ 所属的小轨道。任何不匹配的框 D_t^k 都将初始化为具有新标识的新轨迹。链接在所有相邻节点上顺序进行，并为单个目标建立较长的轨迹。

网络使用两个相邻的帧作为输入，并回归同一目标的边界框对。模型中采用 ResNet-50[20]作为骨干网来提取高级语义特征。然后，通过特征金字塔网络（FPN）生成多尺度特征表示，用于后续预测。为了关联相邻帧中的目标，首先将来自各帧的尺度级特征图连接在一起。目标检测头和身份识别头将来自检测器的相同特征作为输入，然而检测任务需要加强属于同一类别的目标的表观相似性身份识别趋向于最大化各种目标之间的特征差异。因此，本文算法将组合特征经过互相关注意力网络后生成两个特征图，一个特征图馈入分类和回归分支，另一个特征图馈入身份验证分支。从图 2 中可以看出，成对框回归分支为每个目标生成一个框对，对象分类分支为每对预测一个分数，表示作为前景的置信度。身份验证分支得到置信度分数，表明检测对中的两个框是否属于同一个目标，然后同时使用身份验证分支和对象分类分支的预测置信度图作为注意力图，将注意力图与组合特征相乘后再输入到边界框回归分支，两个分支是互补的，利用预测网络结构中 3 个分支的特性构造联合注意力模块，充分利用了分类分支与身份验证分支的信息。

在数据关联部分，连续两个链接节点之间因为公共帧的存在而仅有微小的差异，因此使用简单的 IOU 匹配即可以完成相邻节点之间的关联，使用基础的 Kuhn-Munkres（KM）算法就可以完成检测框之间的最优匹配，由于数据关联算法的简化，使算法提高了跟踪的速度，满足实时性的要求。另一方面，由于网络的输入包含两个相邻的帧，因此在跟踪过程中必须使用两个相邻节点的公共帧。为了避免推理中计算和内存的近乎双倍的成本，我们设计了一种内存共享机制，来临时保存当前帧的提取特征并重用它们直到处理下一个节点。为了对最后一个节点进行推理，复制了第 N 帧作为假设的第 $N+1$ 帧。为了进一步避免对第 $N+l$ 帧的重复计算，还对第 N 帧应用了特征恢复的技巧，将第 N 帧的特征复制为假设第 $N+1$ 帧的特征。综上，网络利用链式特性减少了误检的情况，同时降低了关联的复杂度，实现了端到端的实时跟踪。

2.3 互相关注意力模块

多任务学习往往存在任务不兼容的问题，导致性能下降甚至任务失败。我们提出了一个互相关注意力模块来学习目标检测和身份识别任务的特征的共性和特异性。对于特异性学习，学习反映不同特征通道之间相关性的自相关，以增强每个任务的特征表示。对于共性学习，可以通过精心设计的交叉关系机制来学习两个任务之间的共享信息。互相关注意力模块的结构如图 3 所示。

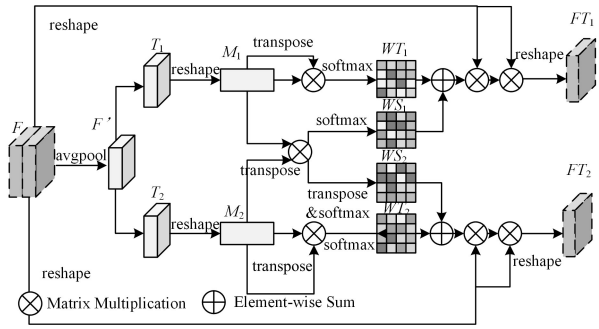

图 3 互相关注意力模块

Fig. 3 Cross-correlation attention module

在输入特征图上使用两个具有相同结构的卷积来聚合局部特征。然后，对特征图进行整形，使其在空间维度上展平，从而方便计算矩阵乘法以获得自注意力权重和相关权重。自注意力机制用于消除不同任务需求带来的歧义，而关联机制用于从两个任务中获取公共部分。最后，对原始特征图和获得的权重进行操作和融合，以输出适合不同任务的特征。形式上，假设检测器最后输出的特征为 $F\in R^{C\times H\times W}$。首先，通过平局池化操作(Avg-pooling)获得尺度更小的特征 $F'\in R^{C\times H'\times W'}$。其次，将其输入卷积层，以获得针对不同任务的特征映射 T_1 和 T_2。然后，将它们重塑成形式 $\{M_1,M_2\}\in R^{C\times N'}$，其中 $N'=H'\times W'$。最后，将不同的任务单独与它自身的转置相乘，并应用行向 softmax 计算每个任务独立的通道注意力图 $\{M_{T_1},M_{T_2}\}\in R^{C\times C}$。具体计算式如下：

$$\omega_{T_k}^{ij}=\frac{\exp(M_k^i\cdot M_k^j)}{\sum_{j=1}^{C}\exp(M_k^i,M_k^j)},k\in\{1,2\} \tag{1}$$

其中，$\omega_{T_k}^{ij}$ 表示任务通道注意力图中第 i 通道和第 j 通道之间的关系。

同样，在 M_1 和 M_2 的转置之间执行矩阵乘法，以学习不同任务之间的共性，然后遵循 Softmax 层生成互相关权重映射 $\{W_{S_1},W_{S_2}\}\in R^{C\times C}$：

$$\omega_S^{ij}=\frac{\exp(M_{1/2}^i\cdot M_{2/1}^i)}{\sum_{j=1}^{C}\exp(M_{1/2}^i\cdot M_{1/2}^j)} \tag{2}$$

其中，ω_S^{ij} 代表任务 1/2 的第 i 个通道对任务 2/1 的第 j 个通道的影响。通过可训练参数 λ，最终融合自相关和互相关权重，得到 $\{W_1,W_2\}\in R^{C\times C}$：

$$W_{1/2}=\lambda\times W_{T_1/T_2}+(1-\lambda)\times W_{S_1/S_2} \tag{3}$$

将原始特征 F 映射重新排列为形状 $R^{C\times N}$，其中 $N=H\times W$。然后，在重塑特征和学习的权重映射之间执行矩阵乘法，以获得每个任务的增强表示，并将增强表示与原始 F 通过残差计算进行融合，防止信息丢失。

2.4 标签分配和损失设计

对于任意链节点 (F_t,F_{t+1})，令 $A_t^i=(x_a^{t,i},y_a^{t,i},w_a^{t,i},h_a^{t,i})$，表示其第 i 个链锚(其中，$x_a^{t,i}$ 和 $y_a^{t,i}$ 是盒子的中心坐标；$w_a^{t,i}$ 和 $h_a^{t,i}$ 分别是宽度和高度)，采用与 SSD[24] 相似的真实边界匹配策略，使用矩阵 $\boldsymbol{M}$ 表示这种匹配的结果。如果 G_t^j 是 F_t 中对应于 A_t^i 的真实边界框，由 IOU 比判断，则 $M_{ij}=1$；如果 IOU 比低于另一个较小的阈值，则 $M_{ij}=0$。基于 $\boldsymbol{M}$，可以将真实标签 c_{cls}^i 分配给本文跟踪框架的 A_t^i 分类分支为：

$$c_{cls}^i=\begin{cases}1, & \text{if } \sum_{j=1}^{K_t}M_{ij}=1\\ 0, & \text{if } \sum_{j=1}^{K_t}M_{ij}=0\end{cases} \tag{4}$$

其中，K_t 是帧 F_t 的真实边界框总数。

使用 A_t^i，假设预测的一对边界框为 D_t^i，$\hat{D}_{t+1}^i$，相应的真实边界框为 $G_t^i\in G_{t+1}^k$，本文的 ID 验证分支应获得其真实标签：

$$c_{id}^i=\begin{cases}1, & \text{if } cls=1 \text{ and } L[G_t^j]=L[G_{t+1}^k]\\ 0, & \text{otherwise}\end{cases} \tag{5}$$

其中，$L[\cdot]$ 表示边界框中目标的身份。

遵循 Faster R-CNN[8] 回归 D_t^i，$\hat{D}_{t+1}^i$ 的偏移量，其中 $D_t^i=(x_d^{t,i},y_d^{t,i},w_d^{t,i},h_d^{t,i})$。令 $(\Delta_d^{t,i},\Delta_{\hat{d}}^{t+1,i})$ 表示这些偏移量，而 $(\Delta_g^{t,j},\Delta_g^{t+1,k})$ 表示真实的偏移量，列出了 $\Delta_d^{t,i}=(\Delta_{d,x}^{t,i},\Delta_{d,y}^{t,i},\Delta_{d,\omega}^{t,i},\Delta_{d,h}^{t,i})$ 的详细信息作为示例：

$$\begin{aligned}&\Delta_{d,x}^{t,i}=(x_d^{t,i}-x_a^{t,i})/\omega_d^{t,i},\Delta_{d,y}^{t,i}=(y_d^{t,i}-y_a^{t,i})/h_d^{t,i}\\&\Delta_{a,\omega}^{t,i}=\log(\omega_d^{t,i}/\omega_a^{t,i}),\Delta_{d,h}^{t,i}=\log(h_d^{t,i}/h_a^{t,i})\end{aligned} \tag{6}$$

配对框回归分支的损失定义如下：

$$L_{\text{reg}}(\Delta_d^{t,i},\Delta_{\hat{d}}^{t+1,i},\Delta_g^{t,j},\Delta_g^{t+1,k})=\frac{\sum_l[smooth_{L_1}(\Delta_{d,l}^{t,i}-\Delta_{g,l}^{t,j})+smooth_{L_1}(\Delta_{\hat{d},l}^{t+1,i}-\Delta_{g,l}^{t+1,k})]}{8} \tag{7}$$

其中，$smooth_{L_1}$ 是平滑 L_1 损失，$l\in\{x,y,\omega,h\}$。

跟踪算法的总损失为：

$$L_{\text{all}}=\sum_{t,i}[L_{\text{reg}}+\alpha F(p_{cls}^i,c_{cls}^i)+\beta F(p_{id}^i,c_{id}^i)] \tag{8}$$

其中，$F(p_{cls}^i,c_{cls}^i)$ 和 $F(p_{id}^i,c_{id}^i)$ 分别是 p_{cls}^i 和 p_{id}^i 的分类分支和 ID 验证分支(用于缓解样本不平衡问题)的焦点损失，id 表示它们的预测(置信度分数)，α 和 β 是加权因子。

3 实验结果与分析

3.1 数据集和评价指标

实验使用 MOT16 和 MOT17 数据集。MOT16 是 MOT 中普遍采用的基准，它由 14 个序列组成，涵盖各种场景、视点、相机姿势和天气条件。MOT16 中的 7 个序列用于训练，其他用于验证。MOT17 是通过重构 MOT16 建立的。与 MOT16 相比，MOT17 提供了更准确的 Ground Truth 和更多的由各种检测器产生的检测边界框，包括 DPM[22]，SDP，Faster-RCNN[5]，其余与 MOT16 相同。由于多目标跟踪算法的性能受检测器的影响很大，因此为了更加公平地评价多目标跟踪算法的性能，MOT Challenge 上设置了两种评估方法，一种是使用官方提供的公共检测器的结果，即 Public 方法，另一种是允许使用自己的检测器，即 Private 方法。本文的网络结构中检测部分利用了 Resnet-50[20] 的结构，属于 Private 方法。

实验的评估是基于 CLEAR-MOT 的评价指标[23]，其可以针对模型的整体性能进行评价，是一个普遍采用的度量集，包括 Multiple-Object Tracking Accuracy(MOTA)，Multiple-Object Tracking Precision(MOTP)，False Negatives(FN)，False Positives(FP)，Identity Switches(IDSW)，Mostly Tracked Trajectories(MT)，Mostly Lost Trajectories(ML)，IDF1 Score(IDF1)，也用于衡量轨迹识别准确度。在这些指标中，MOTA 是衡量整体检测和跟踪性能的主要指标。

$$MOTA=1-\frac{\sum_t(FP_t+FN_t+IDSW_t)}{\sum_t GT_t} \tag{9}$$

其中，下标 t 是帧索引，GT 是真实 bbox 的数量。

3.2 参数设置

实验平台采用 Pytorch，在训练过程中，选择可见分数高于 0.1 的真实框进行训练。为了避免过度拟合，我们使用了几种数据增强策略，例如光度失真、随机翻转和随机裁剪，增强图像对被调整大小或填充到原始图像较短边的一半。我们在时间维度上还添加了一种新的数据增强策略以形成链节点：并非总是选择两个相邻的帧，以随机的时间间隔对彼此靠近的两个帧进行采样。

作为速度与准确度的权衡，所有实验都使用 Resnet-50[20]网络作为主干。除了 Resnet50 中的 BN 参数之外，所有可训练的权重都是使用 Adam 优化器进行端到端训练。使用文献[24]中的 Kaiming 初始化方法为所有新添加的卷积层初始化参数，并将初始学习率设置为 $5\times e^{-5}$。模型训练过程需要 100 个 epoch，batch 大小为 8(4 个训练对)。权重因子 α 和损失函数中的 β 都设置为 1。在锚匹配阶段，使用 0.5 为正阈值，0.4 为负阈值。对于成对框后处理，对 soft-nms 使用 0.7 的阈值，然后进一步过滤剩余的成对，置信度阈值为 0.4。在链接阶段，IOU 匹配阈值为 0.5，σ 的保留阈值为 10。

3.3 基准评估

在 MOT16 和 MOT17 测试数据集上将本文的方法与 DeepSORT[2]，CNNMTT[12]，Tracktor + CTdet[19]，POI[25]，CTracker[26]进行比较。为了方便比较，分别使用 MOT16 训练数据、MOT17 训练数据来训练模型。

(1)算法指标测试

表 1、表 2 分别列出了 MOC-CCC 和对比方法在 MOT16 和 MOT17 测试数据集上的指标测试结果。

表 1 MOT16 测试数据集上的跟踪结果比较

Table 1 Comparisons of tracking results on MOT16 test dataset

Public Detection									
Process	Method	*MOTA* ↑	*IDF1* ↑	*MOTP* ↑	*MT* ↑/%	*ML* ↓/%	*FP* ↓	*FN* ↓	*IDS* ↓
Offline	LMP[17]	**48.8**	**51.3**	**79.0**	**18.2**	40.1	**6654**	**86245**	**481**
	MHT-bLSTM[27]	42.1	47.8	75.9	14.9	44.4	11637	93172	753
	EDMT[28]	45.3	47.9	75.9	17.0	**39.9**	11122	87890	639
Online	Tracktor[15]	**54.4**	52.5	**78.2**	**19.0**	**36.9**	**3280**	**79149**	682
	MOTDT[17]	47.6	50.9	74.8	5.2	38.3	9253	85431	792
	DMAN[29]	46.1	**54.8**	73.8	17.4	42.7	7909	89874	532
Private Detection									
Process	Method	*MOTA* ↑	*IDF1* ↑	*MOTP* ↑	*MT* ↑/%	*ML* ↓/%	*FP* ↓	*FN* ↓	*IDS* ↓
Offline	MCMOT-HDM[1]	62.4	51.6	78.3	31.5	24.2	9855	57257	1394
	KDNT[25]	**68.2**	60.0	79.4	**41.0**	**19.0**	11479	**45605**	933
	NOMT[30]	62.2	**62.6**	**79.6**	32.5	31.1	**5119**	63352	**406**
Online	DeepSORT[2]	61.4	62.2	79.1	32.8	**18.2**	12852	56668	**781**
	CNNMTT[12]	65.2	62.2	78.4	32.4	21.3	6578	55896	946
	JDE[15]	64.4	55.8	—	35.4	20.0	—	—	1544
	POI[25]	66.1	**65.1**	**79.5**	**34.0**	20.8	**5061**	55914	805
	CTracker[26]	67.6	57.2	78.4	32.9	23.1	8934	48305	1897
	Ours	**68.9**	56.1	78.2	33.3	22.4	7780	**46883**	1933

表 2 MOT17 测试数据集上的跟踪结果比较

Table 2 Comparisons of tracking results on MOT17 test dataset

Public Detection									
Process	Method	*MOTA* ↑	*IDF1* ↑	*MOTP* ↑	*MT* ↑/%	*ML* ↓/%	*FP* ↓	*FN* ↓	*IDS* ↓
Offline	MHT-bLSTM[27]	47.5	51.9	**77.5**	18.2	41.7	25981	268042	2069
	EDMT[28]	50.0	51.3	77.3	**21.6**	**36.3**	32279	**247297**	2264
	JCC[31]	**51.2**	**54.5**	75.9	20.9	37.0	**25937**	247822	**1082**
Online	Tracktor[15]	**53.5**	52.3	**78.0**	**19.5**	36.6	**12201**	**248047**	**2072**
	MOTDT[17]	50.9	52.7	76.6	17.5	**35.7**	24069	250768	2474
	DMAN[29]	48.2	**55.7**	75.9	19.3	38.3	26218	263608	2194
Private Detection									
Process	Method	*MOTA* ↑	*IDF1* ↑	*MOTP* ↑	*MT* ↑/%	*ML* ↓/%	*FP* ↓	*FN* ↓	*IDS* ↓
Online	DeepSORT[2]	60.3	**61.2**	**79.1**	31.5	**20.3**	36111	185301	**2442**
	Tracktor+CTdet[19]	54.4	56.1	78.1	25.7	29.8	44109	210774	2574
	CTracker[26]	66.6	57.4	78.2	**32.2**	24.2	22284	**160491**	5529
	Ours	**66.7**	55.6	78.5	31.5	25.1	**20154**	162105	5523

从表 1、表 2 中可以发现，在 MOT16 和 MOT17 的 Private 方法中，MOC-CCC 在 MOTA 方面明显优于现有的在线 MOT 方法，也优于基准网络 CTracker。在 MOT16 中，MOC-CCC 的 MOTA 比最佳离线方法 KDNT[25]高 0.7，比其在线版本 POI[25]高 2.8。此外，KDNT 和 POI 使用了许多额外的训练数据，包括 ETHZ 行人数据集、Caltech 行人数据集和他们自己收集的监视数据集。而 MOT-CCC 只使用了 MOT16 的训练数据。MOTA 是反映整体检测和跟踪性能的主要指标，这证明了 MOC-CCC 的有效性。

在 MOT16 的 Private 方法中，MOT-CCC 保持了可比较甚至更好的 MT 和 ML，这实际上证明了本文方法显著提高了跟踪关联。如表 2 所列，CTracker[26]和 MOT-CCC 网络结构都使用多帧线索来预测检测，MOT-CCC 的网络结构的 FP 大大减少了 13%，这进一步证明了本文网络结构卓越的跟踪性能。

(2)算法推理速度测试

两阶段方法由于将检测和关联分开处理，导致系统至少需要两个计算密集型组件，即一个检测器和一个嵌入模型，总推理时间是这两个分量的总和。而一阶段方法将检测和嵌入模型集成到一个统一的系统中，缩短了推理时间。本文提出的 MOT-CCC 属于一阶段方法，其采用单个网络同时输出检测结果和检测到的目标框的相应外观嵌入，实现上相对 CTracker 方法添加了一个特征分割模块，因此只会带来边际推理成本开销。

实验平台为 Intel Xeon 4210@2.20 GHz 支持下的搭载 4 块 2080ti 显卡的服务器，对比算法为二阶段方法 DMAN 和一阶段方法 Tracktor，JDE，分别在 MOT16 和 MOT17 数据

集上进行了推理速度的对比，结果如表 3 所列。

表 3 不同方法的速度

Table 3 Speed of different methods

Dataset	Method	Type	FPS
MOT16	JDE[18]	One-shot	18.5
	CTracker[29]	One-shot	17.6
	DMAN[32]	Two-stage	0.3
	MOT-CCC	One-shot	**16.4**
MOT17	JDE[18]	One-shot	18.5
	CTracker[29]	One-shot	17.6
	DMAN[32]	Two-stage	0.3
	MOT-CCC	One-shot	**16.4**

注：以 FPS 为单位

由表 3 所列，MOT-CCC 在 MOT16 和 MOT17 测试集上获得了 16.4FPS 的运行速度。相比之下，两阶段算法 DMAN 在两个测试集上仅以 0.3FPS 的速度运行。同为一阶段算法，MOT-CCC 虽然在处理速度上略低于 JDE，但相差约两帧每秒的处理速度对于系统实时性要求的影响有限，同时多目标跟踪的主要评价指标 MOTA 提高了 4.5，IDF1 也有所上升，可见 MOT-CCC 的综合性能更好。MOT-CCC 因为在 CTracker 上添加了一个特征分割模块，使得处理速度略有下降，但使 FP 大大减少了 13%，可见算法在对误报的处理上效果明显。

(3)算法可视化测试

在真实场景序列中，MOC-CCC 比 CTracker 方法[29]具有更好的跟踪能力。如图 4 所示，图 4(a)序列是在 CTracker 上的跟踪结果，图 4(b)序列是在 MOC-CCC 上的跟踪结果。对比实验结果可知，由于发生类内遮挡，CTracker 跟踪算法发生了 ID 切换，而 MOC-CCC 算法并没有发生 ID 切换。对比图 4(a)中的第三张图和图 4(b)中的第三张图，如图中的红色箭头所示，CTracker 跟踪算法错误地为同一个行人分配了两个不同的 ID。对比图 4 中的中间两张图，如图 4(d)中的红色箭头所示，CTracker 跟踪算法跟踪的 18 号 ID 发生了 ID 切换。实验结果证明了 MOC-CCC 算法的跟踪性能优于 CTracker 方法。

图 4 MOT16 和 MOT17 中选定序列的结果可视化(电子版为彩图)

Fig. 4 Visualization of results of selected sequences in MOT16 and MOT17

结束语 本文基于链式结构设计了一种名为 MOT-CCC 的新型网络，以连续的两帧图片为链节点输入，将目标关联问题转化为两帧检测框对的回归问题，增加了目标间的关联性，是一种融合目标检测、特征提取和数据关联 3 个模块的完全端到端跟踪算法。以往的方法使用单独的检测和身份识别任务，当在统一框架中优化网络参数时会产生歧义。MOT-CCC 使用互相关注意力模块来同时获得更适合不同任务的特征，该策略减少了训练过程中多任务联合学习引起的竞争冲突。在 MOT 挑战中，本文方法在 MOT16 和 MOT17 数据集上取得了有竞争力的表现。本文在数据关联阶段使用 IOU 进行匹配，虽然简单的关联算法使跟踪速度变快，但是影响了关联的准确性，我们未来的工作会关注优化数据关联，在保证速度的同时设计更合理的关联方法。

参 考 文 献

[1] LEE B, ERDENEE E, JIN S, et al. Multi-class multi-object tracking using changing point detection[C]// European Conference on Computer Vision. Cham: Springer, 2016: 68-83.

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

[3] FANG K, XIANG Y, LI X, et al. Recurrent autoregressive networks for online multi-object tracking[C]// 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2018: 466-475.

[4] FARHADI A, REDMON J. Yolov3: An incremental improve-

ment[C]// Computer Vision and Pattern Recognition. Berlin/Heidelberg, Germany: Springer, 2018: 1804-02.

[5] REN S, HE K, GIRSHICK R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in Neural Information Processing Systems, 2015, 28: 91-99.

[6] GONG X, LE Z C, WNAG H, et al. Survey of Data Association Technology in Multi-target Tracking[J]. Computer Science, 2020, 47(10): 136-144.

[7] SADEGHIAN A, ALAHI A, SAVARESE S. Tracking the untrackable: Learning to track multiple cues with long-term dependencies[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017: 300-311.

[8] REZATOFIGHI S H, MILAN A, ZHANG Z, et al. Joint probabilistic data association revisited[C]// Proceedings of the IEEE International Conference on Computer Vision. 2015: 3047-3055.

[9] KIM C, LI F, CIPTADI A, et al. Multiple hypothesis tracking revisited[C]// Proceedings of the IEEE International Conference on Computer Vision. 2015: 4696-4704.

[10] LEAL-TAIXÉ L, CANTON-FERRER C, SCHINDLER K. Learning by tracking: Siamese CNN for robust target association [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2016: 33-40.

[11] SUN S J, AKHTAR N, SONG H S, et al. Deep affinity network for multiple object tracking[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 43(1): 104-119.

[12] MAHMOUDI N, AHADI S M, RAHMATI M. Multi-target tracking using CNN-based features: CNNMTT[J]. Multimedia Tools and Applications, 2019, 78(6): 7077-7096.

[13] BAE S H, YOON K J. Confidence-based data association and discriminative deep appearance learning for robust online multi-object tracking[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(3): 595-610.

[14] BERGMANN P, MEINHARDT T, LEAL-TAIXE L. Tracking without bells and whistles[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 941-951.

[15] WANG Z, ZHENG L, LIU Y, et al. Towards real-time multi-object tracking[C]// Computer Vision-ECCV 2020: 16th European Conference, Glasgow, UK (Part XI 16). Springer International Publishing, 2020: 107-122.

[16] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2117-2125.

[17] ZHANG Y, WANG C, WANG X, et al. A simple baseline for multi-object tracking[J]. arXiv: 2004. 01888, 2020.

[18] CHEN L, AI H, ZHUANG Z, 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(ICME). IEEE, 2018: 1-6.

[19] KUHN H W. The Hungarian method for the assignment problem[J]. Naval Research Logistics Quarterly, 1955, 2(1/2): 83-97.

[20] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.

[21] LIU W, ANGUELOV D, ERHAN D, et al. Ssd: Single shot multibox detector[C]// European Conference on Computer vision. Cham: Springer, 2016: 21-37.

[22] FELZENSZWALB P F, GIRSHICK R B, MCALLESTER D, et al. Object detection with discriminatively trained part-based models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 32(9): 1627-1645.

[23] BERNARDIN K, STIEFELHAGEN R. Evaluating multiple object tracking performance: the clear mot metrics[J]. EURASIP Journal on Image and Video Processing, 2008, 2008: 1-10.

[24] HE K, ZHANG X, REN S, et al. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification [C]// Proceedings of the IEEE International Conference on Computer Vision. 2015: 1026-1034.

[25] YU F, LI W, LI Q, et al. Poi: Multiple object tracking with high performance detection and appearance feature[C]// European Conference on Computer Vision. Cham: Springer, 2016: 36-42.

[26] PENG J, WANG C, WAN F, 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: Springer, 2020: 145-161.

[27] KIM C, LI F, REHG J M. Multi-object tracking with neural gating using bilinear lstm[C]// Proceedings of the European Conference on Computer Vision(ECCV). 2018: 200-215.

[28] CHEN J, SHENG H, ZHANG Y, et al. Enhancing detection model for multiple hypothesis tracking[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2017: 18-27.

[29] ZHU J, YANG H, LIU N, et al. Online multi-object tracking with dual matching attention networks[C]// Proceedings of the European Conference on Computer Vision(ECCV). 2018: 366-382.

[30] CHOI W. Near-online multi-target tracking with aggregated local flow descriptor[C]// Proceedings of the IEEE International Conference on Computer Vision. 2015: 3029-3037.

[31] KEUPER M, TANG S, ANDRES B, et al. Motion segmentation & multiple object tracking by correlation co-clustering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 42(1): 140-153.

CHEN Yunfang, born in 1976, Ph. D, master supervisor. His main research interests include artificial intelligence algorithms, functional analysis of specific application areas, application development using intelligent systems.

ZHANG Wei, born in 1973, Ph.D, Ph.D supervisor. His main research interests include intelligent perception and cognition under UAV platform, privacy protection and artificial intelligence security.

（责任编辑：喻藜）