MIFT: 《Refinements in Motion and Appearance for Online Multi-Object Tracking》

Abstract

现代的多目标跟踪（MOT）系统通常涉及分离的模块，例如用于位置的运动模型和用于数据关联的外观模型。但是，运动模型和外观模型中的兼容问题始终被忽略。在本文中，通过**无缝融合运动集成**，**三维（3D）集成图像**和**自适应外观特征融合**，提出了一种称为 MIF 的通用体系结构。由于不确定的行人和摄像机运动通常是分开处理的，因此**使用我们确定的摄像机运动强度来设计集成运动模型**。具体来说，**提出了一种基于 3D 积分图像的空间分块方法，以有效地切断具有空间约束的轨迹和候选对象之间的无用连接**。然后共同建立外观模型和可见性预测。考虑到比例，姿势和可见性，外观特征会进行自适应融合以克服特征未对准问题。我们的基于 MIF 的跟踪器（MIFT）在 MOT16 和 17 挑战中均达到了 60.1 MOTA 的最新精度。

1、Introduction

多目标跟踪（MOT）在场景分析视频分析任务中起着至关重要的作用。 它旨在估计对象的轨迹，并将其与给定的检测结果以在线或整批方式关联。 随着对象检测任务的最新进展，逐个检测跟踪策略成为解决跟踪多个对象问题的首选范例。 然而，尽管依赖于检测具有优点，但是由于检测的质量，它也成为复杂场景中的主要限制。

使用“按检测跟踪”范例时，跟踪任务通常分为几个单独的部分，例如运动模型，特征提取，数据关联。 在本文中，我们**探索了**在线多目标跟踪在**运动**和**外观方面的改进**。 运动模型用于处理行人运动和摄像机运动的估计，这对于数据关联很有用。 此外，当存在遗漏的检测时，通过运动模型预测的位置也可以视为轨迹。 但是，行人运动模型和摄像机运动模型总是分开使用(如 deepsort)或简单地相互组合使用，因此无法精确估计对象的运动状态。**为了建立具有高鲁棒性的运动集成模型，我们探索了非刚性运动与刚性运动之间的关系，并将它们集成在一起**。

此外，考虑到轨迹与候选对象之间的无用联系，**在数据关联阶段通过 3D 积分图像将空间约束应用于每个轨迹**。 因此，**所有检测都将以 one-hot 编码的方式转移到特征图中**（请参见第 3.2 节）。 因此，将**以恒定的时间复杂性整体获得每个跟踪位置的目标区域内的检测结果，这可以显着降低数据关联的时间成本**。

由于类别内的遮挡和不可靠的检测，提取的特征通常会受到前景对象的影响，前景对象不是框中的目标。 每个轨迹都包含不同比例，姿势和质量的各种历史特征，这会引起特征错位。 为了解决此问题，**设计了一种可识别遮挡的外观模型，以利用尺度不变性和可见性估计更好地提取对象的特征**。 考虑到特征的未对准性，轨迹的历史特征与每个即将到来的检测之间的差异是通过遮挡（可见性），比例，姿势和时间间隔来衡量的。 因此，历史特征被自适应地融合。

我们的**工作贡献**概括如下：

–我们提出了一种通用架构（MIF），该架构可以在所有 MOT 基准上都具有最新性能的情况下应用于多对象跟踪和检测任务。

–我们采用建议的运动强度度量标准，整合了行人和摄像机的运动，以克服它们之间的相互作用。

–我们**应用空间约束来减少 3D 积分图像的数据关联时间成本**。

–我们**设计了一种遮挡感知的外观模型和自适应外观特征融合机制，以处理轨迹和检测之间的不对准**。我们的代码将在论文被接受后发布。

2、Related Work

关于 MOT 的最新相关研究可以归纳如下：

Motion models for trajectory prediction. 视频序列中的运动可以概括为**非刚性运动（行人）**和**刚性运动（摄像机姿势）**。 非刚性运动通常由恒速模型描述[9]。 在[29]中，基于观测，通过高斯分布对轨迹进行平滑处理。 最近，使用提供的检测作为观测值，卡尔曼滤波器趋于被更广泛地接受[4,27]。 此外，由于在拥挤的场景中行人运动复杂，因此应用了社会力量模型[19]。 至于由相机姿态变化引起的刚性运动，研究人员已在两个方向进行了研究。一种是基于 3D 信息的方法，例如自我运动[26]和 SFM [9]。另一个是基于仿射变换[2]。此外，还提出了带有递归神经网络的条件概率模型[13]，以预测下一帧目标的位置和形状。此外，逐渐采用基于单对象跟踪（SOT）的方法[11,15,33]直接搜索目标。

Appearance feature extraction and selection. 目标对象的识别以及轨迹和候选对象之间的未对准是外观特征模型的关键方面。识别任务通常被认为是人的重新识别问题[27]。由于背景对象和遮挡的影响，提取的特征通常比较嘈杂。为了解决这个问题，一些基于空间注意力的方法被用来关注前景目标[7]。至于每个轨迹中的各种历史特征，还需要通过特征选择和融合任务来解决未对准问题。最直接的方法是将每个历史特征与每个即将来临的候选人进行比较[27]。但是，这种方法将花费大量时间，并且对未对准问题的影响很小。为了解决这个问题，[15]提出了一种外观质量评估模型，以选择每个时间窗口内最具代表性的特征。另外，[31]使用隐马尔可夫模型探索了时间动态来预测外观特征。最近，还研究了外观和运动特征以端对端的方式融合，例如将外观特征与单对象跟踪器（SOT）结合[11]以及外观和位置特征的联合学习[26]。

3、Proposed Method

在这项工作中，我们提出了一种用于多对象跟踪的通用体系结构 MIF。 它也可以扩展到检测任务。 我们的框架（参见图 1）包含考虑**非刚性运动（行人运动）**和**刚性运动（摄像机运动）**的运动集成，空间块使用 **3D 集成图像**以及**自适应外观特征融合**来对检测和轨迹之间的姿态对齐。 此外，空间块模块旨在将空间约束应用于每个被跟踪的框，这对于度量计算和图形构建都非常节省时间。

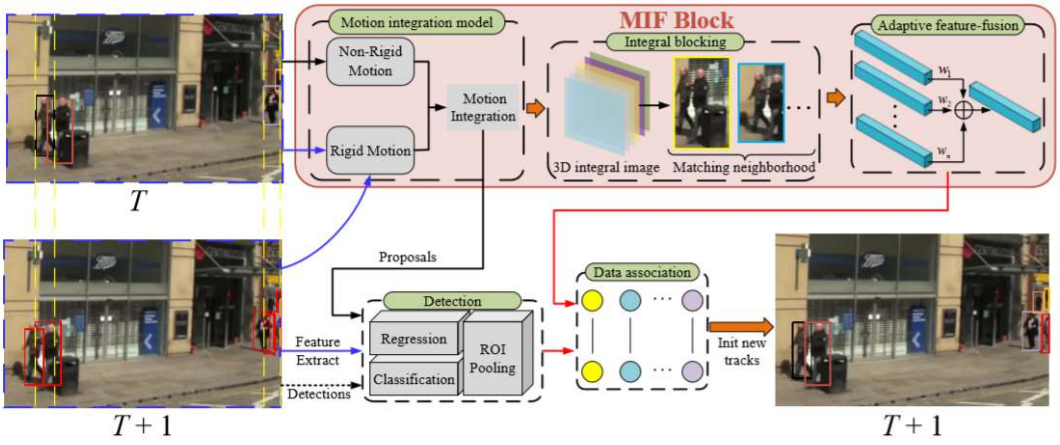

Fig. 1. Illustration of the MIF architecture with an regression based tracker. For a given frame T, the integrated motions are applied to predict each track's positions in frame T+1 considering camera pose variances. Second, each track object will be constrained to a local searching region using 3D integral image and the trajectories' historical features are adaptively weighted for different detections. After the regression and classification of tracked boxes, new detections will be associated with the trajectories.

我们使用 Tracktor [2]作为基线跟踪器，该跟踪器将跟踪预测和提供的检测视为自定义建议，以取代 RPN 网络。 然后，提案将被传递到 ROI Pooling 块进行回归和分类。

3.1 Motion Integration

在某些情况下，基于 IOU 的数据关联可以胜过许多常规方法，这是由于高质量的检测和高的帧速率。 但是，如果存在较大的行人运动，摄像机运动或低帧频，则必须将它们考虑在内。 对于相机运动，可以通过对极几何（Ego Motion）约束或精细变换来建立顺序帧之间的像素对应关系。 假设目标具有慢速运动和静态形状，可以将目标的状态表述为 Ego Motion 的优化问题[26]，其中 F 是基本矩阵，x 表示目标边界框的坐标。

$$f\left(x_{i,t+1}\right)=\sum_{i=1}^{4}\left\|x_{i,t+1}^{T}Fx_{i,t}\right\|^{2}+\left\|\left(x_{3,t+1}-x_{1,t+1}\right)-\left(x_{3,t}-x_{1,t}\right)\right\|_{2}^{2} \quad (1)$$

在这种方法中，基本矩阵需要在没有相机信息的情况下通过特征匹配来估计。 但是特征点通常依赖于包含大量渐变信息的区域，这些区域也会受到人体部分的严重干扰。结果，预测目标的状态将不可靠。 在这里，我们**研究了使用增强相关系数最大化（ECC）和卡尔曼滤波器将刚体运动和非刚体运动紧密结合在一起**。

考虑到空间一致性，需要在摄像机运动模型之前处理行人运动模型。 详细来说，每个目标的位置都需要首先通过卡尔曼滤波器进行预测，然后通过 ECC 模型（被称为卡尔曼+ ECC）对准。此外，由于相机运动和不均匀运动状态引起的不确定性，我们将衰落存储器应用于卡尔曼滤波器，以将更多注意力集中在最近的运动上。 因此，可以如下建立卡尔曼+ ECC 运动模型：

$$\begin{cases} s_{t+1}=warp\left(Fs_{t}\right) \\ P_{t+1}=\alpha FP_{t}F^{T}+Q \end{cases} \quad (2)$$

Where α denotes the fading memory coefficient, Q denotes the process covariance, s and P denote the predicted states and prior covariance of Kalman Filter, $warp$ denotes the ECC model.

但是，Kalman + ECC 解决方案的独立运动处理将引发兼容问题。 因此，我们通过使用仿射矩阵来调整整体运动模型，从而将摄像机运动和行人运动混合在一起。 这样，集成的运动模型无需预定义的参数即可适应各种运动场景。 首先，我们将相机运动的强度定义为等式 3。

$$Ic=1-\frac{\boldsymbol{W}\times\boldsymbol{R}}{\|\boldsymbol{W}\|_{2}\times\|\boldsymbol{R}\|_{2}},R=[I;O] \quad (3)$$

Where the Ic denotes the intension of camera motion. $\boldsymbol{W}$ denotes the vectorization of the affine matrix. The R means the the affine matrix of static frames. I is the identity matrix and O is the all-zero matrix.

通过上面定义的意图，我们可以通过更改状态转换矩阵来调整卡尔曼滤波器：

$$\begin{cases} s_{t+1}=warp\left(F_{c}s_{t}\right) \\ P_{t+1}=\alpha F_{c}P_{t}{F_{c}}^{T}+Q \end{cases},F_{c}=\begin{bmatrix} I & \left(dt+Ic\right)I \\ O & I \end{bmatrix} \quad (4)$$

其中F_c表示调整后的状态转换矩阵，而dt表示卡尔曼滤波器的原始时间步长。

3.2 Spatial Blocking via 3D Integral Image

通常，计算轨道边界框与候选边界之间的成本矩阵的时间复杂度为$O(n\text{^}2)$。 为了为每个轨道边界框分配附近的检测，我们将检测转化为基于掩码的 one-hot 编码描述符（见图 2）。 使用 3D 积分图像可以非常快速地计算此特征表示。

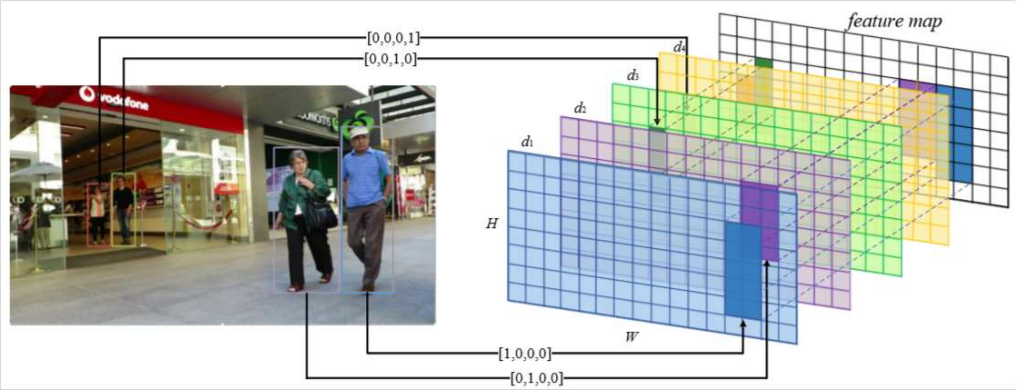

Fig. 2. Illustration of feature mapping. The input image of $W \times H$ is divided into $M \times N$ cells to reduce space complexity. Thus each cell has the size of $\frac{W}{M} \times \frac{H}{N}$. Then each detection bounding box d_i will be mapped into corresponding regions in the graph. Every cell contains a D-dim vector which uses one-hot encoding style, where the D denotes the number of detections in this frame. For example, [1 0 0 0] in cell(m, n) represents that this cell overlaps with the first detection box.

获取特征图的$M \times N$个 cell 后，每个 cell 都包含具有 D-dim 矢量的候选者的位置信息。位置m，n处的 3D 积分图像包含特征图中从（0,0）到（m，n）的候选总和，包括:

$$I(m,n) = \sum_{m' \le m, n' \le n} f(m', n') \tag{5}$$

Where $I(m,n)$ is the 3D integral image and $f(m',n')$ is the feature map. We can simplify the process by dynamic programming:

$$I(m,n) = I(m,n-1) + I(m-1,n) - I(m-1,n-1) + f(m,n) \tag{6}$$

对于每个即将到来的轨迹的边界框$[x_1, x_2, y_1, y_2]$，将分配一个包含多个 cell 的空间块区域。 使用 3D 积分图像，我们可以直接以恒定复杂度获得每个空间块区域的候选列表。

$$\begin{aligned} I(x_1 : x_2, y_1 : y_2) = {} & I(x_2, y_2) + I(x_1 - 1, y_1 - 1) \\ & - I(x_1 - 1, y_2) - I(x_2, y_1 - 1) \end{aligned} \tag{7}$$

尽管数据关联的时间复杂度仍为$O(mn)$。 m和n表示检测次数和轨迹数。大多数操作是赋值，加法和减法。 因此，实际上减少了该阶段的时间成本。 此外，它还需要空间来保存 3D 积分图像，因此空间复杂度从$O(1)$增加到$O(n)$。

3.3 Adaptive Appearance Feature Fusion

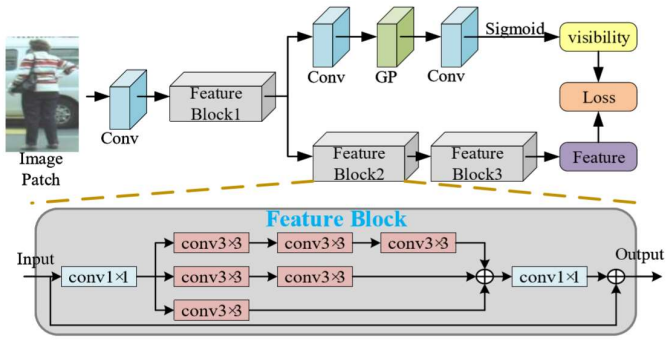

Fig. 3. Architecture of our proposed occlusion aware appearance model. The appearance branch contains three feature blocks, which are formed as multi-scale residual inception blocks. The visibility prediction branch is performed after the first feature block of appearance branch.

考虑到行人的遮挡，不同的姿势和规模，我们提出了一种姿势和遮挡感知的自适应外观特征融合模型。 该模型包括两个方面，包括如图 3 所示的遮挡感知外观模型和如图 4 所示的自适应特征融合。

有人提出了分辨率不变的表示方法[21]，以解决人员重新识别领域中的规模和分辨率失调问题。 因此，我们提出了一种轻量级的特征块，该特征块使用级联卷积形成为多尺度的残差起始块，以获得不同的接收域尺度。

$$L_a = \frac{1}{N}\sum_{i=1}^{N} y_i^* log\left(y_i\right) + \left(1 - y_i^*\right) log\left(1 - y_i\right) \tag{8}$$

外观模型的交叉熵损失定义为等式 8。相反，受位置敏感蒙版[7]的启发，我们训练了与外观模型结合的可见性预测分支。 但是，大多数物体是完全可见的，这带来了不平衡的问题。 通过添加系数ϕ来设计多任务损失，以利用对象不同可见性之间的不平衡。

$$Loss = L_a + \frac{\phi}{N}\sum_{i=1}^{N}\left(v_i - v_i^*\right)^2 \tag{9}$$

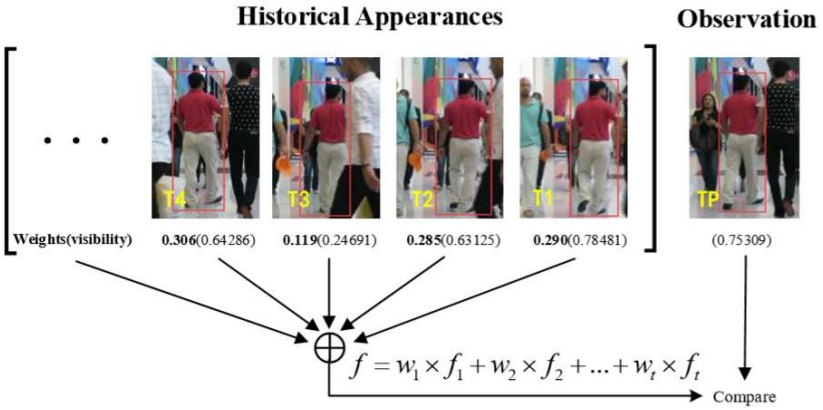

Fig. 4. Illustration of adaptive feature fusion algorithm. Each historical appearance in the trajectory will be automatic weighted as the bold texts show. Besides, the visibility of each appearance is also shown within brackets.

利用外观特征，除了那些汇总的端到端框架外，轨迹和候选者之间的相似性度量可以简单地作为特征选择或特征融合问题来执行。 但是，特征选择会带来很大的未对准风险。 因此，我们**提出了一种结合可见度，尺度，aspect 和时间信息的自适应特征融合模型**。 至于在这四个方面，候选人特征与一条轨迹的历史特征之间的差异，我们仅使用最小-最大规格化来统一尺寸。 同样，对这四个方面进行加权求和以获得总距离，如图 4 所示。

$$d = \lambda_1 d_{scale} + \lambda_2 d_{aspect} + \lambda_3 d_{visibility} + \lambda_4 d_{time} \tag{10}$$

The weight coefficients of each trajectory's historical features can be calculated as:

$$weights_i = \frac{\exp\left(-d_i\right)}{\sum_j \exp\left(-d_j\right)} \tag{11}$$

3.4 MIF based Tracker

结合上面提到的 MIF 方法，我们可以轻松地将其扩展为使用现有的跟踪器跟踪多个对象。 在这里，我们使用马氏距离来评估运动距离和卡尔曼滤波器的协方差。

$$d_m = \left(det - track\right)^T S^{-1}\left(det - track\right) \tag{12}$$

其中 S 表示卡尔曼滤波器的系统不确定性。然后通过归一化余弦度量来计算出现距离。

实际上，**运动距离更适合于短期关联，而外观距离更可能用于长期关联**。 因此，我们提出了一种平衡的方式来将它们两者整合在一起。

$$
\begin{aligned}
w &= miss\_rate^{time\_gap} \\
d &= wd_m + (1-w)\, d_a
\end{aligned}
\tag{13}
$$

其中，$time_gap$表示自轨迹消失以来的时间长度。 随着不活动轨迹的长度增加，保留的位置将变得不可靠。 因此，将采用外观特征来进行补偿，并且外观成本的权重将随着$time_gap$而增加。

Algorithm 1: MIF based tracking algorithm

Input: Video sequences $\mathcal{I} = \{\mathcal{I}_1, \mathcal{I}_2, \cdots, \mathcal{I}_T\}$ and provided detections $\mathcal{D} = \{D_1, D_2, \cdots, D_T\}$.
Output: Trajectories $\mathcal{T}$
$\mathcal{T} \leftarrow \phi$;
$\mathcal{L}_k$: Lost length of $\mathcal{T}_k$;
$\mathcal{F}_k$: Appearance Feature of $\mathcal{T}_k$;
for $t = 1, \cdots, T$ **do**
  Extract features of I_t;
  Apply Integrated motions to $\mathcal{T}_k$ referring to Eq. 4;
  $B_t, S_t \leftarrow Regress\_and\_Classify\{D_t, \mathcal{T}_k\}$;
  $\mathcal{B} \leftarrow NMS([B_t, D_t], thresh)$;
  Build 3D integral image with $\mathcal{B}$ as in Section 3.2;
  Extract appearance features of each $\mathcal{B}$;
  for $\mathcal{T}_k \in \mathcal{T}$ **do**
    $b \leftarrow SpatialBlock(B, \mathcal{T}_k)$;
    $F_k \leftarrow AdaptiveWeightedFeatures(b, \mathcal{T}_k)$;
    $Cost \leftarrow GetCost(b, F_k, \mathcal{T}_k)$ by Eq. 13;
  Associate the $\mathcal{B}$ with $\mathcal{T}$ using $Cost$;
  for $\mathcal{T}_k \in \mathcal{T}$ **do**
    if *Assigned with* B_i **then**
      $\mathcal{T}_k \leftarrow B_i$;
      $\mathcal{F}_k \leftarrow \mathcal{F}_k + F_i$;
      $\mathcal{L}_k = 0$;
    else
      if $\mathcal{L}_k >$ *time gap* **then**
        $\mathcal{T} = \mathcal{T} - \mathcal{T}_k$;
      else
        $\mathcal{L}_k ++$;
  $\mathcal{T} \leftarrow \mathcal{T} + \{\mathcal{B} - \mathcal{T}_k\}$;
Delete the inactive trajectories.

4、 Experiments

4.1 Experiment Setup

在广泛使用的多目标跟踪基准 MOTChallenge 上进行了实验。 该基准测试包括几个具有挑战性的行人跟踪和检测序列，这些序列具有频繁的遮挡和拥挤的场景，这些场景在摄像机视角，物体比例和帧速率方面有所不同。 我们的基于 MIF 的跟踪器已经针对三个独立的挑战进行了评估，分别称为 2D MOT2015 [18]，MOT16 和 MOT17 [22]。 另外，MOT16 和 MOT17 都包含相同的 7 个序列和 7 个测试序列。 区别在于 MOT17 基准测试提供了三种性能提高的不同公开检测（DPM [14]，Faster R-CNN [23]，SDP [30]），而 MOT16 基准测试仅提供了 DPM 一种。 2D MOT2015 基准还提供了 22 个序列的 ACF [12]检测。

Evaluation Metric. 根据广泛接受的 CLEAR MOT 指标[3]进行评估，包括多对象跟踪（MOTA）的准确性，ID 开关的数量（ID Sw。），假阳性（FP）和假阴性的总数 （FN）等。在这些指标中，MOTA 和 ID Sw。 可以量化两个主要方面，即对象覆盖率和识别率。

Implementation Details. 拟议的方法在 Pytorch 中实现，并在台式机上运行，该台式机具有 10 个核@ 2.2GHz 的 CPU 和两个 RTX2080Ti GPU。卡尔曼滤波器的衰落记忆和时间步长设置为 1.2 和 0.15。为了利用效率和准确性，将 3D 整体图像分为 16×8 块。对于外观模型，网络经过 150 个时期的训练，学习率为 3e-3，批处理大小为 64。外观模型的输入图像补丁大小为 64×256，特征尺寸为 512。对于可变数量的目标，特征提取阶段将花费大量时

间。实验表明，随着 batchsize 大小的增加，特征提取的速度最初呈线性增长。然后它趋于稳定。因此，无论特征数量是否可被 batchsize 整除，都将使用固定的 batchsize（26）提取所有目标的特征。另外，我们通过在 COCO 数据集上训练的预训练权重，将锚点方面更改为{1.0,2.0,3.0}，从而重新实现 Faster RCNN 检测器。请注意，对象检测器和 re-id 模型都采用多尺度策略从头开始进行训练。具体而言，**轨迹的重新连接机制仅适用于具有时间间隔可以达到 10 的具有摄像机运动的场景。每个轨迹最多可以包含 26 个历史特征。在后处理阶段，将删除长度小于 5 的轨迹**。

Table 1. Ablation study in terms of different motion models. The Ego denotes the Epipolar Geometry model and MI denotes the integrated motion model.

Method	MOTA↑	IDF1↑	ID Sw.↓
Ego	53.31	44.22	1829
ECC	59.2	59.44	**481**
Kalman	59.33	58.81	604
Kalman+ECC	59.48	59.86	569
MI	**60.23**	**59.87**	509

Table 2. Comparisons of different appearance models. The first and the latest selected historical features are compared with features fused in average way and our proposed adaptive way.

Method	MOTA↑	IDF1↑	ID Sw.↓
ReID(avg)	57.48	53.15	1486
ReID(latest)	57.76	53.38	1123
ReID(fusion)	57.92	53.78	1238
ReID+MI	**60.38**	**61.47**	**484**

4.2 Ablation Study

根据从 MOT17 训练集中提取的验证集对消融研究进行了评估。 由于我们已将训练集的一部分用于训练，因此在这里我们仅将验证集用于消融研究。

Motion Integration. 表 1 显示了不同的运动模型对 MOT 任务的影响。这里的 Ego 模型（极地几何）是指等式。 1 也添加到实验中。 MI 表示参考方程式的运动积分。 4.基线模型是我们重新实现的 Tractor ++ [2]。 从表 1 中可以看出，同时使用卡尔曼滤波器和 ECC 模型而不进行积分（卡尔曼+ ECC）（参考公式 1）。 2 可能比仅采用其中任何一种更好。 此外，参考方程式的综合运动模型。 图 4 显示了 MOTA 和 IDF1 相对于非集成 Kalman + ECC 模型的明显优势。 同样很明显，对极模型（Ego）无法很好地处理运动对齐。 请注意，ID 为 Sw。 单个 ECC 模型的特征值略小于集成运动模型。 事实是，集成运动模型跟踪的轨迹数量远多于 ECC 模型。

Fig. 6. Heatmaps' visualizations of the first feature block's outputs.

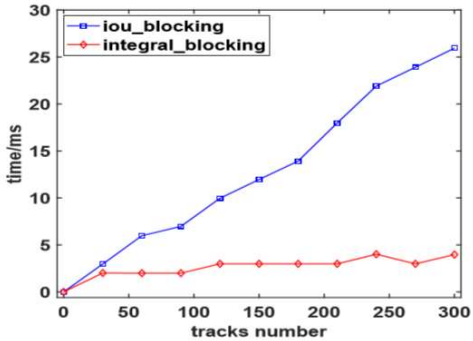

Fig. 7. Speed comparison using IOU based blocking and 3D integral image based blocking

Appearance Model. 为了评估外观模型和自适应特征融合方法，表 2 中评估了几种不同的特征选择和融合方法。ReID 表示我们提出的遮挡感知外观模型。 在表 2 中，我们通过

简单地使用轨迹中的最新特征，实验性地选择了具有代表性的历史特征，并通过平均特征和我们的自适应特征融合模型来融合这些特征。 相比之下，自适应外观特征融合模型可以更好地提高 MOTA 和 IDF1 分数。 与简单使用运动积分模型和自适应特征融合模型相结合，结合了短期和长期线索，MOTA 得分分别提高了 1.5%，2.46%和 IDF1 分别提高了 1.6%，7.69%。 此外，外观特征的可视化效果如图 6 所示。由于我们使用可见性预测分支训练了外观模型，因此该模型更多地关注前景目标。

3D Integral image. 为了演示使用 3D 积分图像（称为积分阻挡）提出的空间阻挡方法的速度，我们还提出了一种基于 IOU 的称为 iou 阻挡的区域阻挡方法进行比较。 如果检测结果与轨道边界框的扩展区域重叠，则检测结果将分配给轨道。 速度比较如图 7 所示，它显示了我们的整体速度对基于 IOU 的速度的整体阻塞的显着优势，特别是当每帧存在大量轨迹或检测时。

4.3 Evaluation on Benchmarks

我们已对所有 MOT 测试序列进行了基于 MIF 的跟踪器（MIFT）的性能评估 4。 表 3 列出了正式发布的结果。在线和批处理方法均在同一表中进行了说明。 如表中所示，我们的跟踪器（MIFT）在大多数指标上均优于所有现有的在线跟踪器，尤其是对于 MOTA，IDF1，MT，ML，FN。 此外，与大多数跟踪器相比，我们的在线跟踪器的计算成本要低得多。 在 2DMOT2015 挑战中，由于检测质量差，我们提出的跟踪器的性能要比唯一的批处理方法（MPNTracker)稍差。综上所述,我们提出的方法由于运动集成可以显着改善 ML 和 MOTA，从而保持了轨迹的连续性。使用 3D 积分图像,我们的跟踪器的速度比基线跟踪器(Tracktor)快得多。

Table 3. Comparison of our method with the methods on the MOT Challenge. **O** denotes the online methods.

Methods	MOTA↑	IDF1↑	MT↑	ML↓	FP↓	FN↓	ID Sw.↓	Frag↓	Hz↑
2D MOT 2015									
Ours(O)	46.7	51.6	29.4%	25.7%	11003	**20839**	878	1265	6.7
MPNTrack [5]	**48.3**	**56.5**	**32.2%**	**24.3%**	9640	21629	504	1074	9.3
Tracktor(**O**) [2]	44.1	46.7	18.0%	26.2%	6477	26577	1318	1790	0.9
KCF(**O**) [10]	38.9	44.5	16.6%	31.5%	7321	29501	720	1440	0.3
AP_HWDPL_p(**O**) [8]	38.5	47.1	8.7%	37.4%	**4005**	33203	586	1263	6.7
STRN(**O**) [28]	38.1	46.6	11.5%	33.4%	5451	31571	1033	2665	**13.8**
AMIR(**O**) [24]	37.6	46.0	15.8%	26.8%	7933	29397	1026	2024	1.9
JointMC [17]	35.6	45.1	23.2%	39.3%	10580	28508	**457**	**969**	0.6
MOT16									
Ours(O)	**60.1**	56.9	**26.1%**	**29.1%**	6964	**65044**	739	951	6.9
MPNTrack [5]	55.9	**59.9**	26.0%	35.6%	7086	72902	431	921	11.9
Tracktor(**O**) [2]	54.4	52.5	19.0%	36.9%	**3280**	79149	682	1480	1.5
NOTA [7]	49.8	55.3	17.9%	37.7%	7248	83614	614	1372	**19.2**
HCC [20]	49.3	50.7	17.8%	39.9%	5333	86795	**391**	**535**	0.8
LSSTO(**O**) [15]	49.2	56.5	13.4%	41.4%	7187	84875	606	2497	2.0
TNT [26]	49.2	56.1	17.3%	40.3%	8400	83702	606	882	0.7
AFN [25]	49.0	48.2	19.1%	35.7%	9508	82506	899	1383	0.6
MOT17									
Ours(O)	**60.1**	56.4	**28.5%**	**28.1%**	23168	**199483**	2556	3182	**7.2**
MPNTrack [5]	55.7	59.1	27.2%	34.4%	25013	223531	1433	3122	4.2
LSST [15]	54.7	**62.3**	20.4%	40.1%	26091	228434	**1243**	3726	1.5
Tracktor(**O**) [2]	53.5	52.3	19.5%	36.6%	**12201**	248047	2072	4611	1.5
LSSTO(**O**) [15]	52.7	57.9	17.9%	36.6%	22512	241936	2167	7443	1.8
JBNOT [16]	52.6	50.8	19.7%	35.8%	31572	232659	3050	3792	5.4
FAMNet [11]	52.0	48.7	19.1%	33.4%	14138	253616	3072	5318	0.0
TNT [26]	51.9	58.1	23.1%	35.5%	36164	232783	2288	**3071**	0.7

其他定性结果显示在图 5 中。MIFT 跟踪所有测试序列，并以 SDP 检测作为观察结果。显然，我们提出的跟踪器可以精确跟踪目标箱。 而且，MIFT 对于不规则的摄像机运动（例如 MOT17-06 和 MOT17-14），拥挤的场景（例如 MOT17-03），不同的摄像机视点（例如 MOT17-03 和 MOT17-07）具有鲁棒性。 尤其是在 MOT17-14 序列上，这些序列是由安装

在繁忙路口公交车上的快速移动摄像头捕获的。 我们提出的跟踪器仍然可以稳定，持久地跟踪目标。

4.4 Extension to Detection

在本节中，我们首先使用 FPN 重新实现 Faster RCNN 检测器。 然后，将使用新的检测器代替跟踪任务中的公共检测。 表 4 中显示的结果表明，基于 MIF 的检测器（MIFD）也可以在 MOT17 Det 挑战中获得有希望的结果。

Table 4. Comparison with the state-of-arts MOT Detection methods. Our MIFD detector is a MIF based detector, which is combined with a re-implemented Faster RCNN detector with FPN.

Method	AP↑	MODA↑	FAF↓	Precision↑	Recall↑
MSCNN [6]	**0.89**	**76.7**	2.8	86.2	91.3
POI [32]	0.89	67.1	4.8	78.7	92.1
ViPeD [1]	0.89	-14.4	20.8	46.4	**93.2**
FRCNN [23]	0.72	68.5	**1.7**	**89.8**	77.3
FRCNN+FPN	0.88	65.9	5.1	77.7	92.4
MIFD	**0.88**	**67.4**	**4.9**	**78.6**	**92.6**

此外，基于 MIF 的检测器几乎在每个指标上都优于基于非 MIF 的检测器（FRCNN + FPN）。

5、Conclusion

在本文中，我们探讨了运动和外观模型的改进。 因此，提出了一种称为 MIF（运动集成，3D 集成图像和自适应外观特征融合）的通用体系结构，该体系结构可以嵌入跟踪和检测任务中。在广泛使用的 MOT 挑战赛上进行了实验，展示了我们基于 MIF 的跟踪器（MIFT）和基于 MIF 的检测器（MIFD）的优势。 具体来说，由于运动和外观模型通常用于跟踪方法中，因此我们提出的方法可以帮助克服它们之间的相互作用和错位。 此外，可以通过我们提出的 3D 积分图像简化检测与轨迹之间的关联，如图 3 以及表 3 的最后一列所示，该 3D 积分图像非常有效。