# ByteTrack 的复现与改进

**摘要**

多目标跟踪是计算机视觉领域的热点任务，旨在视频场景下的目标检测与目标身份的重新识别。大部分的 MOT 方法会选择一个阈值，只保留高于这个阈值的检测结果来做关联得到跟踪结果，低于这个阈值的检测结果直接丢弃。但是这种做法会带来不可逆转的错误，例如漏检和轨迹碎片化。为了解决之前方法直接丢弃低分检测框带来的后果，本文提出了一种新的数据关联方法 BYTE，将匹配过程分为两个阶段，优先匹配高分检测结果，其次在低分结果中挖掘出真实对象，并过滤掉背景。通过实验，将 BYTE 应用到 9 个不同的 SOTA 跟踪器上，IDF1 指标提升了 1 到 10 个点不等。基于 BYTE 方法，本文提出了 ByteTrack，在 MOT17 等数据集上取得了 SOTA 性能。

**关键词：**计算机视觉；多目标跟踪

## 1 引言

Tracking-by-detection 是当前多目标跟踪 (MOT) 领域最有效的范式。由于视频场景的复杂性，检测器往往难以做出完美的预测。许多 SOTA 的 MOT 方法 [1, 10, 13] 为了权衡 true positive 和 false positive，会选择直接丢弃所有低置信度的检测框，只保留高于某个阈值的检测结果来进行轨迹关联。本文 [12] 认为这种做法是不合理的，低置信度检测框有时表示物体的存在，例如被遮挡的物体。直接过滤掉这些结果会导致 MOT 出现不可逆的误差，带来不可忽略的缺失检测和轨迹碎片化。

如图1，在 $t_1$ 时刻 (b) 初始化了 3 条轨迹，这些轨迹的检测框置信度均大于 0.5。然而，在 $t_2$ 和 $t_3$ 时刻，人物出现了遮挡，红色轨迹对应的检测框置信度从 0.8 降到了 0.4 和 0.1，这些检测框被阈值机制消除，红色轨迹也随之消失。如果我们把每一个检测框都考虑到，就会很容易出现 false positive。很少有 MOT 方法能够处理低分检测框而不引入 false positive。

在本文中，我们发现在低分数检测框与轨迹的相似性为区分目标和背景提供了强有力的线索。如图1(c) 所示，通过运动模型预测，2 个低分数检测框与轨迹匹配，从而正确地恢复了目标。背景框由于没有匹配的轨迹也被移除。

为了在匹配过程中充分利用从高分检测框和低分检测框，本文提出了一种简单有效的关联方法 BYTE。首先根据运动相似度或外观相似度将高分检测框与轨迹匹配。类似于 [1]，我们使用卡尔曼滤波 [5] 来预测轨迹在下一帧的位置。1(b) 可以看作是第一阶段的匹配结果。然后，我们使用运动相似度在未匹配的轨迹和低分框间执行第二次匹配。1(c) 展示了第二次匹配的结果，低分检测框被正确地匹配到轨迹上，右侧的背景也被移除。

MOT 的理想解决方案永远不是一个检测器和随后的关联，设计好检测器和关联的连接处也很重要。BYTE 的创新之点在于检测和关联的交界处，它没有改变检测器，也没有改变关联方法，而是利用了低分检测框。得益于这一创新点，将 BYTE 应用到 9 个 SOTA 的跟踪器上时，在 MOTA,IDF1 和 IDS 等指标上都取得了显著提升。

为了进一步提升 MOT 性能，本文提出了一个简单而强大的跟踪器 ByteTrack。通过将高性能的目标检测器 YOLOX [4] 和本文提出的关联方法 BYTE 结合起来从而得到 ByteTrack。在 MOT17 和 MOT20 数据集上，ByteTrack 都取得了第一名的成绩。

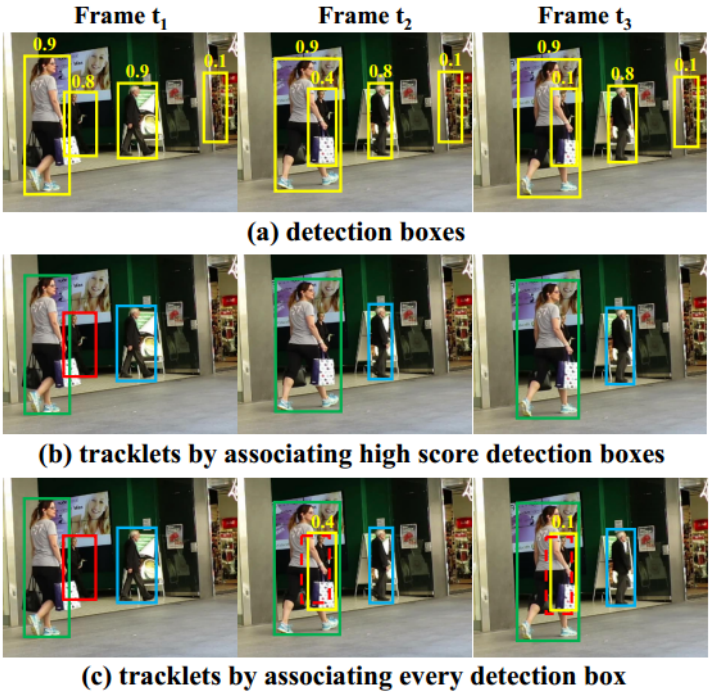


图 1. (a) 连续帧上的所有检测框及其得分。(b) 通过之前的方法得到的轨迹图，这些轨迹图仅关联得分高于阈值 (即 0.5) 的检测框。相同的框色代表相同的身份。(c) 本文方法得到的轨迹图，低分检测框与之前的轨迹正确地关联。

## 2 相关工作

### 2.1 目标检测

目标检测是计算机视觉中最活跃的研究课题之一，是实现多目标跟踪的基础。MOT17 数据集提供了常用检测器获得的检测结果，如 DPM [3], Faster R-CNN [8] 和 SDP [11]。大量的 MOT 方法都关注于在给定的检测结果下改进跟踪性能。

随着目标检测技术的快速发展越来越多的方法开始利用更强大的检测器来获得更高的跟踪性能。one-stage 的 RetinaNet 开始被应用，CenterNet 由于其简单高效成为最流行的目标检测器。YOLO 系列检测器因其在精度和速度上的优异平衡，也被大量方法采用。这些方法大多直接使用单幅图像上的检测框进行跟踪。然而，当视频序列中发生遮挡或运动模糊时，漏检和低分检测结果的数量开始增加。为了提高视频检测性能，通常会结合使用之前帧的信息。

目标跟踪也可以用来获取更准确的检测框。有些方法会利用单目标跟踪或者卡尔曼滤波来预测轨迹，将预测结果和检测框结合可以获得更好的检测结果。也有一些方法会利用之前帧的跟踪结果来增强后续帧的特征表示。本文方法同样利用与轨迹的相似度来提高检测框的

可靠性。

大多数 MOT 方法在获得各种检测器的检测框后，只保留得分较高的检测框，并将其作为数据关联的输入。这是因为低分数检测框中包含了很多背景，影响了跟踪性能。然而，本文观察到许多被遮挡的物体可以被正确检测到，但得分很低。为了减少漏检并保持轨迹的连续性，本文保留了所有的检测框，并在每个检测框之间进行关联。

### 2.2 数据关联

数据关联是多目标跟踪的核心，它首先计算轨迹与检测框之间的相似度，然后根据相似度利用不同的策略对检测框和轨迹进行匹配。

在进行多目标跟踪时，位置（Location）、运动（motion）和外观（appearance）都是关联目标的有用线索。经典方法 SORT [1] 使用了位置和运动信息，它先用卡尔曼滤波 [5] 预测轨迹在下一帧的位置，然后计算检测框和预测框的 IoU 作为匹配相似度。也有一些方法设计网络来学习物体运动，以在相机运动的场景下获得更鲁棒的结果。通常情况下，位置和运动相似度在短期追踪的效果不错，而外观相似度在长期追踪的表现更好。一个物体在经过长时间的遮挡后，可以利用外观相似度重新追踪到。外观相似度可以通过 Re-ID 特征的余弦相似度来计算，DeepSORT [12] 就使用使用独立的 Re-ID 网络从检测框中提取外观特征。

在计算完相似度之后，由匹配策略为对象分配身份。匹配一般是由匈牙利算法 [6] 或者贪婪分配 [14] 来实现。SORT [1] 一次性地将检测框和轨迹进行匹配，DeepSORT [10] 提出了级联匹配策略，该策略首先将检测框与最近的轨迹进行匹配，然后再与丢失的轨迹进行匹配。MOTDT [2] 先利用外观相似度进行匹配，然后利用 IoU 相似度对未匹配的轨迹进行匹配。QDTrack [7] 通过双向 softmax 运算将外观相似度转化为概率，并采用最近邻搜索完成匹配。注意力机制 [9] 可以直接在帧之间传播检测框，并隐式地进行关联而不使用匈牙利算法。

上述的方法都关注于如何设计更好的关联方法。然而，本文认为检测框的好坏决定了数据关联的上界，并聚焦于如何在匹配过程中充分利用高分检测框和低分检测框。

## 3 BYTE

本文提出了一种简单、有效、通用的数据关联方法 BYTE。与之前只保留高分检测框的方法不同，BYTE 保留了几乎所有的检测框，并将这些检测框分为高分检测框和低分检测框。BYTE 首先将高分检测框与轨迹关联起来，一些轨迹不匹配是因为它们没有匹配到合适的高分检测框，这通常发生在遮挡、运动模糊或大小变化发生时。然后，我们将低分检测框与这些不匹配的轨迹关联起来，恢复低分检测框中的物体，同时过滤掉背景。伪代码如2所示。

BYTE 的输入是是视频序列 V 和目标检测器 Det，输出是视频轨迹，包括每一帧中的检测框和目标的身份。对于视频中的每一帧，先使用目标检测器 Det 来得到检测框，通过一定的阈值将检测框分为高分检测框和低分检测框。然后采用卡尔曼滤波来预测每条轨迹在当前帧中的新位置。

第一次关联会将所有轨迹（包括丢失的轨迹）和高分检测框进行匹配，相似度度量采用高分检测框和轨迹预测框的 IoU 或者 Re-ID 特征。然后采用匈牙利算法根据相似度完成匹配过程，未匹配的轨迹和高分检测框会被保留。

第二次关联会将低分检测框和未匹配的轨迹进行匹配，匹配不上的低分检测框会被当做

背景直接删除，仍然匹配不上的轨迹会被保留 30 帧。对于第一次匹配剩余的高分检测框，对其新建一个轨迹。

为了进一步提升 MOT 性能，本文还设计了一个简单而强大的跟踪器 ByteTrack。ByteTrack 由高性能的目标检测器 YOLOX [4] 和本文提出的关联方法 BYTE 结合得到。

**Algorithm 1:** Pseudo-code of BYTE.

**Input:** A video sequence $V$; object detector Det; detection score threshold $\tau$
**Output:** Tracks $\mathcal{T}$ of the video

Initialization: $\mathcal{T} \leftarrow \emptyset$
**for** *frame* $f_k$ *in* $V$ **do**
  /* predict detection boxes & scores */
  $\mathcal{D}_k \leftarrow \text{Det}(f_k)$
  $\mathcal{D}_{high} \leftarrow \emptyset$
  $\mathcal{D}_{low} \leftarrow \emptyset$
  **for** $d$ *in* $\mathcal{D}_k$ **do**
    **if** $d.score > \tau$ **then**
      $\mathcal{D}_{high} \leftarrow \mathcal{D}_{high} \cup \{d\}$
    **end**
    **else**
      $\mathcal{D}_{low} \leftarrow \mathcal{D}_{low} \cup \{d\}$
    **end**
  **end**
  /* predict new locations of tracks */
  **for** $t$ *in* $\mathcal{T}$ **do**
    $t \leftarrow \text{KalmanFilter}(t)$
  **end**
  /* first association */
  Associate $\mathcal{T}$ and $\mathcal{D}_{high}$ using Similarity#1
  $\mathcal{D}_{remain} \leftarrow$ remaining object boxes from $\mathcal{D}_{high}$
  $\mathcal{T}_{remain} \leftarrow$ remaining tracks from $\mathcal{T}$
  /* second association */
  Associate $\mathcal{T}_{remain}$ and $\mathcal{D}_{low}$ using similarity#2
  $\mathcal{T}_{re-remain} \leftarrow$ remaining tracks from $\mathcal{T}_{remain}$
  /* delete unmatched tracks */
  $\mathcal{T} \leftarrow \mathcal{T} \setminus \mathcal{T}_{re-remain}$
  /* initialize new tracks */
  **for** $d$ *in* $\mathcal{D}_{remain}$ **do**
    $\mathcal{T} \leftarrow \mathcal{T} \cup \{d\}$
  **end**
**end**
Return: $\mathcal{T}$

图 2. 方法伪代码

# 4 复现工作

## 4.1 与已有开源代码对比

原文作者已经给出了开源代码，https://github.com/ifzhang/ByteTrack，本工作主要在原文的基础上思考检测性能对多目标跟踪的影响，衡量不同检测器在同个数据集的效果，再结

合本文提出的 BYTE 方法，通过实验验证检测性能对 MOT 的影响。

具体实验设置为，使用检测性能相对较差的 YOLOv3，PP-YOLOE-l 与 BYTE 结合，以及检测性能高于原文使用的 YOLOX 的 YOLOv8 与 BYTE 结合，比较这些方法在 MOT-17 half Val 数据集上的结果。

## 5 实验结果分析

通过比较不同性能的检测器与 BYTE 结合得到的检测效果对比，我们可以发现在检测性能比较低的时候，随着检测性能的提升，MOTA 和 IDF1 指标均出现上升的趋势，然而，当检测性能到了一定的高度，虽然 MOTA 上升了，但是 IDF1 反而却下降了。

通过分析 MOTA(1)和 IDF1(2)的公式，我们可以发现，MOTA 这个指标中，FN 和 FP 的权重占比比较大，相对于 IDF1 来说，检测性能较好的方法的 MOTA 往往会得到更好的结果。总的来说，MOTA 更偏向于衡量检测结果的质量，而 IDF1 更倾向于轨迹的稳定。更好的检测性能也会带来更多的低分检测结果，从而容易降低跟踪性能，具体体现在 IDF1 的降低。

| MOT17 half val | | | |
|---|---|---|---|
| | MOTA | IDF1 | mAP |
| YOLOv3 | 49.5 | 54.8 | 42.7 |
| PP-YOLOE-l | 50.4 | 59.7 | 52.9 |
| byteTrack | 76.6 | **79.3** | 62.4 |
| YOLOv8 | **81.5** | 76.2 | **64.2** |

表 1. MOT17 half val 数据集上，BYTE 在不同性能的检测器上的性能对比

$$MOTA = 1 - \frac{FN + FP + IDSW}{GT} \tag{1}$$

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

## 6 总结与展望

原文提出了一种简单有效的关联方法 BYTE，并与 YOLOX 结合得到了性能良好的 ByteTrack，同时也强调了检测结果的质量对 MOT 性能的重要性。受到启发，本工作进一步探讨了检测性能对 MOT 的影响，通过简单的实验发现 MOTA 和 IDF1 有一定的制衡关系，特别是在检测性能较高的时候。但本工作的实验数量较少，得到的结果仍然有进一步验证的空间，未来将进一步实验探讨 MOT 指标的关系，力求为 MOT 领域做出一份贡献。

# 参考文献


[1] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pages 3464–3468. IEEE, 2016.

[2] Long Chen, Haizhou Ai, Zijie Zhuang, and Chong Shang. Real-time multiple people tracking with deeply learned candidate selection and person re-identification. In *2018 IEEE international conference on multimedia and expo (ICME)*, pages 1–6. IEEE, 2018.

[3] Pedro Felzenszwalb, David McAllester, and Deva Ramanan. A discriminatively trained, multiscale, deformable part model. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. Ieee, 2008.

[4] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

[5] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. 1960.

[6] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[7] Jiangmiao Pang, Linlu Qiu, Xia Li, Haofeng Chen, Qi Li, Trevor Darrell, and Fisher Yu. Quasi-dense similarity learning for multiple object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 164–173, 2021.

[8] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[10] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE, 2017.

[11] Fan Yang, Wongun Choi, and Yuanqing Lin. Exploit all the layers: Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2129–2137, 2016.

[12] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *European Conference on Computer Vision*, pages 1–21. Springer, 2022.

[13] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision*, 129:3069–3087, 2021.

[14] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *European conference on computer vision*, pages 474–490. Springer, 2020.