MOTRv2: Bootstrapping End-to-End Multi-Object Tracking by Pretrained Object Detectors

摘要

在本文中，我们提出了 MOTRv2，这是一种简单而有效的管道，用于使用预训练的目标检测器引导端到端多目标跟踪。现有的端到端方法，如 MOTR[43] 和 TrackFormer[20]，主要由于其较差的检测性能而不如检测再跟踪的目标。我们的目标是通过加入一个额外的目标检测器来提高 MOTR。我们首先采用查询的锚点公式，然后使用额外的目标检测器生成提案作为锚点，在 MOTR 之前提供检测。简单的修改大大缓解了 MOTR 中联合学习检测和关联任务之间的冲突。MOTRv2 保持了查询传播功能，并在大规模基准测试上扩展良好。MOTRv2 在第一届团体舞多人追踪挑战赛中排名第一（DanceTrack 上有 73.4% 的 HOTA）。此外，MOTRv2 在 BDD100K 数据集上达到了最先进的性能。

关键词：目标检测；图像识别；轨迹跟踪

1 引言

多目标跟踪（MOT）旨在预测流媒体视频中所有目标的轨迹。它可以分为两个部分：检测和关联。长期以来，MOT 上最先进的性能一直由检测再跟踪方法 [3, 20, 25, 26] 所主导，这些方法具有良好的检测性能，可以应对各种外观分布。这些跟踪器首先采用目标检测器（例如 YOLOX [9]）来定位每帧中的目标，并通过 ReID 特征或 IoU 匹配来关联轨迹。这些方法的优越性能部分源于数据集和偏向检测性能的指标。然而，正如 DanceTrack 数据集所揭示的那样，它们的关联策略在复杂运动中仍有待改进。

最近，MOTR [24] 为 MOT 引入了一个完全端到端的框架。通过更新轨迹查询来执行关联过程，同时通过检测查询来检测新生目标。它在 DanceTrack 上的关联性能令人印象深刻，而检测结果不如检测再跟踪方法的结果，尤其是在 MOT17 数据集上。我们将较差的检测性能归因于联合检测和关联过程之间的冲突。由于最先进的跟踪器 [3, 20, 25] 倾向于使用额外的目标检测器，一个本质的问题是如何将 MOTR 与额外的目标检测器结合起来，以获得更好的检测性能。一种直接的方法是在轨迹查询的预测和额外的目标检测器之间执行 IoU 匹配（类似于 TransTrack [15]）。在作者的实践中，它只在目标检测方面带来了边际改进，而不符合 MOTR 的端到端特性。

受以检测结果为输入的检测再跟踪方法的启发，我们想知道是否有可能将检测结果作为输入，并减少对关联的 MOTR 学习。最近，DETR 中基于锚点的建模取得了一些进展 [12, 18]。

例如，DAB-DETR 使用定位框的中心点、高度和宽度初始化目标查询。与它们类似，我们修改了 MOTR 中检测和轨迹查询的初始化。我们将 MOTR 中检测查询的可学习位置嵌入（PE [17]）替换为锚点的正余弦 PE，产生了一个基于锚点的 MOTR 跟踪器。通过这种基于锚点的建模，由额外的目标检测器生成的提案可以作为 MOTR 的锚点初始化，提供局部先验。transformer 解码器用于预测锚的相对偏移，从而使检测任务的优化更加容易。

与最初的 MOTR 相比，所提出的 MOTRv2 带来了许多优点。它极大地受益于额外的目标检测器引入的良好检测性能。检测任务与 MOTR 框架隐式解耦，缓解了共享 transformer 解码器中检测任务和关联任务之间的冲突。MOTRv2 学习在给定来自额外检测器的检测结果的情况下跨帧跟踪实例。

2 相关工作

2.1 先检测再跟踪

主要方法 [5, 25] 遵循先检测再跟踪原则，目标检测器首先预测每个帧的目标边界框，然后使用单独的算法来关联相邻帧之间的实例边界框。这种方法的性能在很大程度上取决于目标检测的质量。

使用匈牙利算法 [11] 进行关联有多种尝试：SORT [3] 对每个跟踪的实例应用卡尔曼滤波器 [21]，并使用卡尔曼滤波器的预测框和检测框之间的交并比（IoU）矩阵进行匹配。Deep [?] 引入了一个单独的网络来提取实例的外观特征，并使用 SORT 之上的成对余弦距离。JDE [20]、Track-RCNN [14]、FairMOT [26] 和 Unicorn [23] 进一步探索了目标检测和外观嵌入的联合训练。ByteTrack [25] 利用了强大的基于 YOLOX [9] 的检测器，实现了最先进的性能。它引入了一种增强的 SORT 算法来关联低分数检测框，而不是只关联高分数检测框。BoT-SORT [2] 进一步设计了更好的卡尔曼滤波器状态、相机运动补偿和 ReID 特征融合。TransMOT [8] 和 GTR [28] 在计算分配矩阵时使用时空 transformers，例如特征交互和历史信息聚合。OC-SORT [5] 放松了线性运动假设，并使用了可学习的运动模型。

2.2 按查询传播进行跟踪

MOT 的另一个范例将基于查询的目标检测器 [6, 16, 29] 扩展到跟踪。这些方法强制每个查询在不同的框架中调用同一个实例。查询和图像特征之间的交互可以在时间上并行或串行执行。

并行方法以短视频作为输入，并使用一组查询与所有帧进行交互，以预测实例的轨迹。VisTR [19] 和随后的工作 [7, 22] 扩展了 DETR [6] 以检测短视频剪辑中的轨迹。并行方法需要将整个视频作为输入，因此它们消耗内存，并且仅限于几十帧的短视频剪辑。

串行方法执行与图像特征的逐帧查询交互，并迭代地更新与实例相关联的轨迹查询。Trackor++ [1] 利用 R-CNN [10] 回归头进行跨帧的迭代实例重新定位。TrackFormer [13] 和 MOTR [24] 从可变形 DETR [29] 延伸而来。它们预测目标边界框并更新轨迹查询，以便在后续帧中检测相同的实例。MeMOT [4] 构建短期和长期实例特征内存库，以生成轨迹查询。TransTrack [15] 传播轨迹查询一次，以在下一帧中找到目标位置。P3AFormer [27] 采用流引导图像特征传播。与 MOTR 不同，TransTrack 和 P3AFormer 在历史轨迹和当前检测中仍然使用基于位置的匈

牙利匹配，而不是在整个视频中传播查询。

作者的方法继承了用于长期端到端跟踪的查询传播方法，同时还利用强大的目标检测器来提供目标位置先验。在复杂运动的跟踪性能方面，该方法大大优于现有的基于匹配和查询的方法。

3 本文方法

3.1 本文方法概述

首先利用强大的目标检测器（YOLOX）来检测目标并提供目标先验，同时利用上一帧传来的轨迹查询预测（MOTR）被追踪的边界框，如图 1所示：

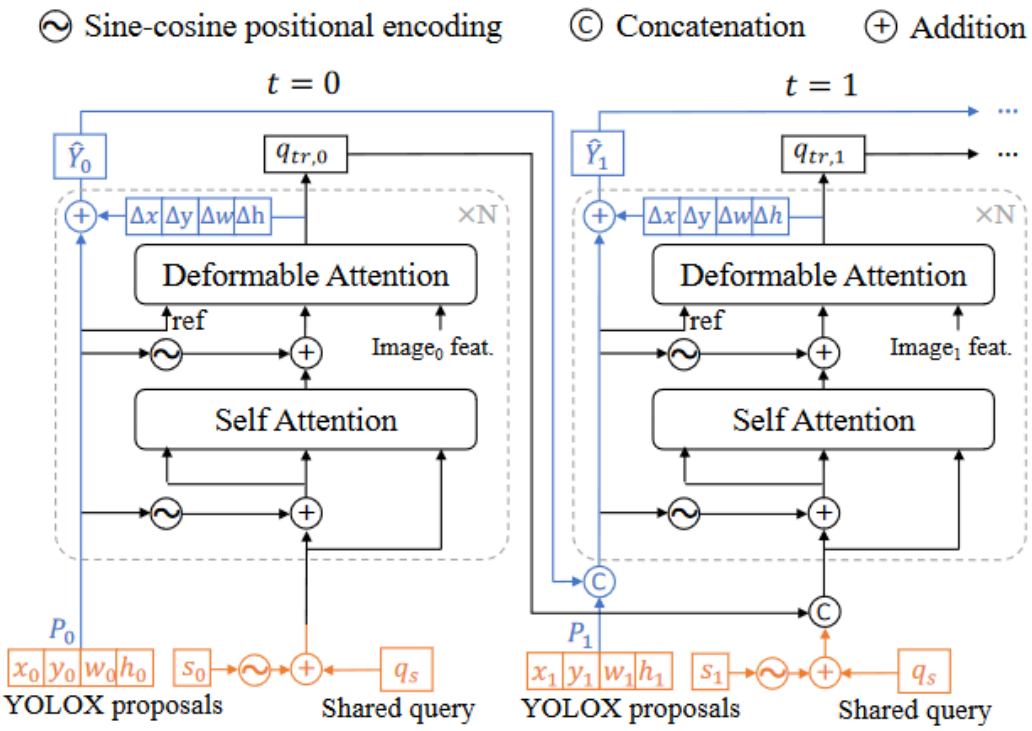

图 1. 用于跟踪的提案查询生成和提案传播。橙色标记提案查询生成，而蓝色标记提案传播路径；灰色虚线框代表 N 个 transformer 解码器。为了简单起见，省略了 MOTR 中的查询交互模块。

3.2 修改 MOTR

MOTR [24] 是一种基于可变形 DETR [29] 架构的完全端到端的多目标跟踪框架。它引入了轨迹查询和目标查询的概念。目标查询负责检测新出现或丢失的目标，并通过输出目标查询的结果来完成这一任务。每个轨迹查询负责在时间上跟踪一个独特的目标实例。为了初始化轨迹查询，MOTR 使用与新检测到的目标相关联的目标查询的输出。这样，轨迹查询可以根据其状态和当前图像特征随时间进行更新，从而能够在线预测跟踪结果。通过这种方式，MOTR 能够实现高效的多目标跟踪。MOTR 中的 tracklet 感知标签分配将轨迹查询分配给

先前跟踪的实例，同时通过二分匹配将目标查询分配给其余实例。MOTR 引入了一个时间聚合网络来增强轨迹查询的功能，并引入了一种集体平均损失来平衡跨帧的损失。

3.3 工作动机

端到端多目标跟踪框架的一个主要局限性是，与依赖独立目标检测器的检测再跟踪方法[5, 25] 相比，它们的检测性能较差。为了解决这一限制，我们建议结合 YOLOX [9] 目标检测器来生成作为目标锚的提案，在 MOTR 之前提供检测。它极大地缓解了 MOTR 中联合学习检测和关联任务之间的冲突，提高了检测性能。

3.4 总体架构

如图 1所示，所提出的 MOTRv2 体系结构由两个主要组件组成：最先进的目标检测器和改进的基于锚点的 MOTR 跟踪器。

目标检测器组件首先生成用于训练和推理的提案。对于每个帧，YOLOX 生成一组先验，其中包括中心坐标、宽度、高度和置信度值。修改后的基于锚点的 MOTR 组件负责基于生成的先验来学习轨迹关联。第 3.5 节描述了用提案查询替换原始 MOTR 框架中的检测查询。修改后的 MOTR 现在将轨迹查询和提案查询的连接作为输入。第 3.6 节描述了连接查询和框架特征之间的交互，以更新被跟踪目标的边界框。

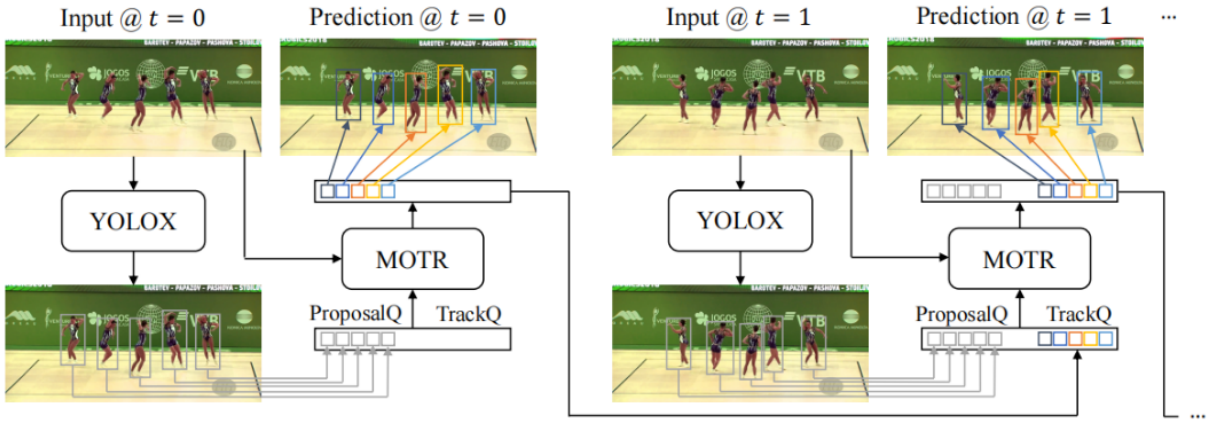

图 2. MOTRV2 的总体框架

3.5 提案查询生成

在本节中，我们将解释先验查询生成模块如何为 MOTR 提供来自 YOLOX 的高质量先验。该模块的输入是 YOLOX 为视频中的每一帧生成的一组提案框。与 DETR [29] 和 MOTR 使用固定数量的可学习查询进行目标检测不同，我们的框架基于 YOLOX 生成的所选先验来动态确定提案查询的数量。

具体来说，对于帧 t，YOLOX 生成 N_t 个提案，每个提案由一个具有中心坐标 (x_t, y_t)、高度 h_t、宽度 w_t 和置信度得分 s_t 的边界框表示。如图 1 所示的橙色部分，作者引入了一个共享查询 q_s 来生成一组提案查询。共享查询时大小为 1×D 的可学习嵌入，首先被广播到大小为 N_t×D。N_t 个提案框的置信度 s_t 通过正余弦位置编码产生大小为 N_t×D 的分数嵌入。然后将广播的查询与分数嵌入相加以后成提案查询。YOLOX 提案查询框充当提案查询的锚

点。在实践中，作者还使用了 10 个可学习锚，并将它们与 YOLOX 提案连接起来，以召回 YOLOX 检测器遗漏的目标。

3.6 提案传播

在 MOTR [24] 中，轨迹查询和检测查询被连接并输入到 transformer 解码器，用于同时进行目标检测和轨迹关联。从上一帧生成的轨迹查询表示被跟踪的目标，这些目标用于预测当前帧的边界框。检测查询是一组固定的可学习嵌入，用于检测新生目标。与 MOTR 不同，我们的方法使用提案查询来检测新生目标，并且轨迹查询的预测是基于先前帧预测的。

对于第一帧 (t=0)，只有新出现的目标，会被 YOLOX 检测到。如图 2 所示，给定 YOLOX 提案的共享查询 q_s 和置信度的情况下生成提案查询。在 YOLOX 提案 P_0 进行位置编码后，提案通过自注意力机制进一步更新，并通过可变形注意力机制与图像特征进行交互，产生轨迹查询 $q_{tr,0}$ 和 YOLOX 提案 P_0 的相对偏移量 $(\Delta x, \Delta y, \Delta w, \Delta h)$。预测 $\hat{Y_0}$ 是提案 P_0 和预测偏移的总和。

对于其他帧 (t>0)，类似于 MOTR，从上一帧生成的轨迹查询 $q_{tr,t-1}$ 将与当前帧的提案查询 $q_{p,t}$ 连接。前一帧的框预测 $\hat{Y_{t-1}}$ 也将与 YOLOX 提案 P_t 连接在一起，用作当前帧的锚。锚的正弦余弦编码被用作连接查询的位置嵌入，然后连接查询进入 transformer 解码器以产生预测和更新的轨迹查询。边界框预测由置信度得分和锚的相对偏移组成，并且更新的轨迹查询 $q_{tr,t}$ 被进一步转移到下一帧，用于检测被跟踪的目标。

分析在上述设计中，提案查询被限制为仅检测新生或丢失的目标，而轨迹查询负责重新定位被跟踪的目标。提案查询需要聚合来自轨迹查询的信息，以避免重复检测被跟踪的目标，并且轨迹查询可以利用 YOLOX 提案来改进目标定位。这是通过 transformer 解码器中的自注意力层来实现的。

4 复现细节

4.1 与已有开源代码对比

无太明显的改进。将测试出的结果进行可视化输出为视频。

4.2 实验环境搭建

操作系统使用的是 linux 操作系统，ubantu 18.04。Python 版本使用的版本为 3.7，cuda 版本为 10.2，pytorch 版本为 1.8.1。同时 python 中安装了 tqdm、scipy、opencv-python 库。使用的显卡型号为 NVIDIA Tesla V100，数量为 4。

5 实验结果分析

实验结果通过在 DanceTrack 数据集上进行训练，并使用 DanceTrack 的测试集进行测试。最终的结果是生成的坐标，及在一帧帧的图片中预测出目标在图片中 x 和 y 坐标的偏移量，同时对目标进行跟踪的结果，即对目标赋予一个 ID。该 ID 便是对目标跟踪的关键，如果只是检测那么只用 yolo 就可以得出结果，因此 ID 即为跟踪的一种凭证。实验结果数据中，大

约每秒钟有 20 帧，每帧有大概有 10 几个目标。如图 5所示，图中有很多舞者，围绕舞者的方框即对目标检测的结果所绘制的，左上角 ID 即为对该舞者跟踪的结果。

图 3. 实验结果示意

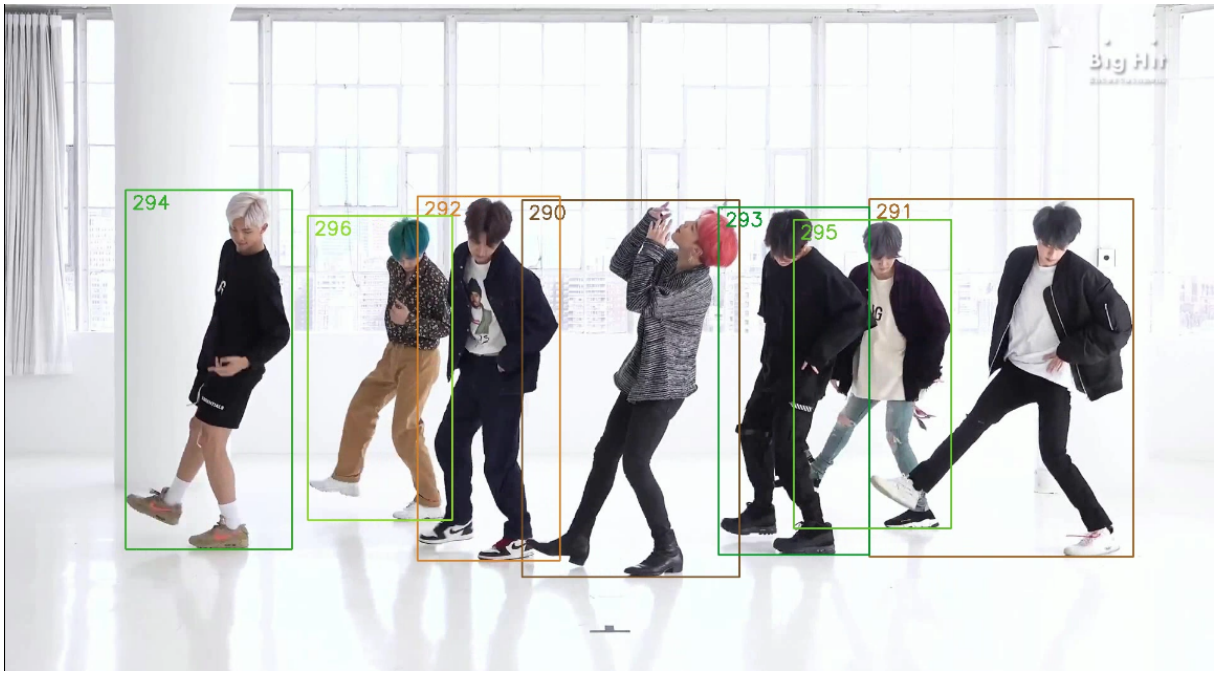

图 4. 实验结果示意

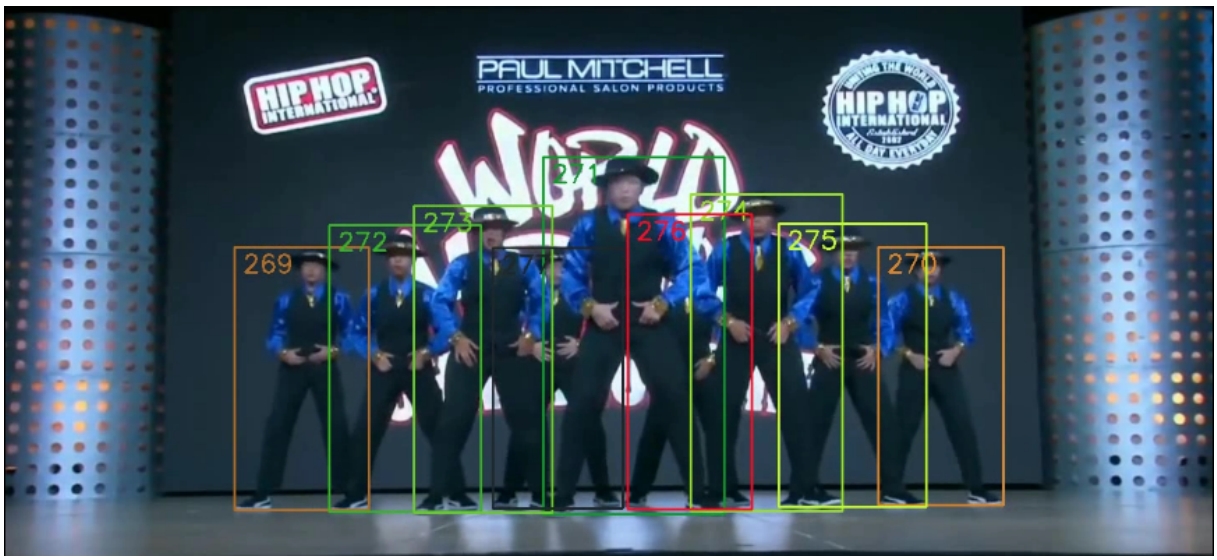

图 5. 实验结果示意

6 总结与展望

论文中提出的是一种由 MOTR [24] 跟踪模型和 YOLOX [9] 检测模型的结合。YOLOX 作为一种高质量的目标检测模型，能够生成高质量的目标提案，能使 MOTR 更加容易检测到新目标。由于 MOTR 模型，原本使将检测和跟踪放在一个模型中实现，而分离为两个模型后降低了目标检测的复杂，让 MOTR 能够更加专注于关联的过程。同时 MOTRv2 突破了以往端到端框架不适用于高性能 MOT 的普遍看法，并且还给出了为什么以往端到端的多目标检测框架的原因。

局限性方面，虽然 YOLOX 方案在解决 MOTR 的优化问题方面有了显著的改善，但该方法仍然需要大量的数据，并且在较小的数据集上表现不佳。实验结果中，还发现了一些重复的轨迹查询现象，例如当两个人相互交叉时，可能会出现一个轨迹查询跟随着错误的目标，导致对同一个人进行两个轨迹查询。这个观察结果为未来的潜在改进提供了宝贵的线索。另一个限制是效率问题。主要瓶颈出现在 MOTR 部分，整体速度不佳。

参考文献

[1] Abhinav Agarwalla, Xuhua Huang, Jason Ziglar, Francesco Ferroni, Laura Leal-Taixé, James Hays, Aljoša Ošep, and Deva Ramanan. Lidar panoptic segmentation and tracking without bells and whistles, 2023.

[2] Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky. Bot-sort: Robust associations multi-pedestrian tracking, 2022.

[3] 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)*. IEEE, September 2016.

[4] Jiarui Cai, Mingze Xu, Wei Li, Yuanjun Xiong, Wei Xia, Zhuowen Tu, and Stefano Soatto. Memot: Multi-object tracking with memory, 2022.

[5] Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, and Kris Kitani. Observation-centric sort: Rethinking sort for robust multi-object tracking, 2023.

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers, 2020.

[7] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation, 2022.

[8] Peng Chu, Jiang Wang, Quanzeng You, Haibin Ling, and Zicheng Liu. Transmot: Spatial-temporal graph transformer for multiple object tracking, 2021.

[9] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021, 2021.

[10] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation, 2014.

[11] Harold W. Kuhn. *The Hungarian Method for the Assignment Problem*, pages 29–47. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[12] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. Dab-detr: Dynamic anchor boxes are better queries for detr, 2022.

[13] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers, 2022.

[14] Bing Shuai, Andrew G. Berneshawi, Davide Modolo, and Joseph Tighe. Multi-object tracking with siamese track-rcnn, 2020.

[15] Peize Sun, Jinkun Cao, Yi Jiang, Rufeng Zhang, Enze Xie, Zehuan Yuan, Changhu Wang, and Ping Luo. Transtrack: Multiple object tracking with transformer, 2021.

[16] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, and Ping Luo. Sparse r-cnn: End-to-end object detection with learnable proposals, 2021.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[18] Yingming Wang, Xiangyu Zhang, Tong Yang, and Jian Sun. Anchor detr: Query design for transformer-based object detection, 2022.

[19] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers, 2021.

[20] Zhongdao Wang, Liang Zheng, Yixuan Liu, Yali Li, and Shengjin Wang. Towards real-time multi-object tracking, 2020.

[21] Greg Welch and Gary Bishop. *An Introduction to the Kalman Filter*. An Introduction to the Kalman Filter, 1995.

[22] Junfeng Wu, Yi Jiang, Song Bai, Wenqing Zhang, and Xiang Bai. Seqformer: Sequential transformer for video instance segmentation, 2022.

[23] Bin Yan, Yi Jiang, Peize Sun, Dong Wang, Zehuan Yuan, Ping Luo, and Huchuan Lu. Towards grand unification of object tracking, 2022.

[24] Fangao Zeng, Bin Dong, Yuang Zhang, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. Motr: End-to-end multiple-object tracking with transformer, 2022.

[25] 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, 2022.

[26] 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(11):3069–3087, September 2021.

[27] Zelin Zhao, Ze Wu, Yueqing Zhuang, Boxun Li, and Jiaya Jia. Tracking objects as pixel-wise distributions, 2022.

[28] Xingyi Zhou, Tianwei Yin, Vladlen Koltun, and Philipp Krähenbühl. Global tracking transformers, 2022.

[29] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection, 2021.