复现 MOTRv2：基于预训练物体检测器的端到端多目标跟踪方法

摘要

本文基于论文《MOTRv2: Bootstrapping End-to-End Multi-Object Tracking by Pre-trained Object Detectors》进行了复现。文章的作者提出了 MOTRv2，这是一种简单高效的多目标跟踪管道，旨在通过预训练的目标检测器引导端到端的多目标跟踪。现有的端到端方法，如 MOTR [30] 和 TrackFormer [15]，由于检测性能不足，难以与基于检测的跟踪方法相媲美。而文章作者的目标是优雅地结合一个额外的目标检测器来提升 MOTR 的性能。他们首先采用查询的锚点形式，并引入额外的目标检测器生成提议，以提供检测先验，作为 MOTR 的锚点。这一简单的修改有效缓解了 MOTR 中联合学习检测与关联任务之间的冲突问题。MOTRv2 保留了查询传播的特性，并在大规模基准测试中表现出良好的可伸缩性。值得一提的是，MOTRv2 在第一届多人跟踪集体舞蹈挑战赛中获得第一名（在 DanceTrack 上达到 73.4% HOTA）。此外，MOTRv2 在 BDD100K 数据集上也达到了最先进的性能。作者希望这个简单而有效的管道能够为端到端多目标跟踪社区提供新的见解。

关键词：端到端多目标跟踪；MOTR；预训练目标检测器

1 引言

多目标跟踪（MOT）的目标是预测流媒体视频中所有物体的轨迹。它可以分为两个部分：检测和关联。长期以来，基于检测的跟踪方法 [4, 25, 31, 32] 凭借出色的检测性能在 MOT 领域占据了领先地位，以应对各种外观分布。这些跟踪器 [31] 首先使用物体检测器（例如，YOLOX [10]）在每帧中定位物体，然后通过 ReID 特征或 IoU 匹配来关联轨迹。这些方法的优越性能部分归因于数据集和评估指标对检测性能的偏向。然而，正如 DanceTrack 数据集 [19] 所揭示的，它们在复杂运动中的关联策略仍有待改进。

最近，MOTR [30] 被引入，作为一个完全端到端的框架来进行 MOT。关联过程通过更新跟踪查询来执行，而新出现的物体则通过检测查询来检测。它在 DanceTrack 上的关联性能令人印象深刻，但其检测结果却不如基于检测的跟踪方法，特别是在 MOT17 数据集上。作者将检测性能较差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 [6, 9, 31] 倾向于使用额外的物体检测器，因此一个自然的问题是如何将 MOTR 与额外的物体检测器结合，以获得更好的检测性能。一种直接的方法是在跟踪查询的预测与额外物体检测器之间执行 IoU 匹配（类似于 TransTrack [20]）。在原论文作者的实践中，这仅在物体检测中带来了边际的改善，而违背了 MOTR 的端到端特性。

受到将检测结果作为输入的基于检测的跟踪方法的启发，作者想知道是否可以将检测结果作为输入，并将 MOTR 的学习减少到关联上。最近，在 DETR 中出现了一些针对基于锚点建模的进展 [13, 23]。例如，DAB-DETR 使用锚框的中心点、高度和宽度初始化物体查询。类似地，作者修改了 MOTR 中检测查询和跟踪查询的初始化。他们将 MOTR 中检测查询的可学习位置嵌入（PE）替换为锚点的正弦余弦 PE [22]，从而生成一个基于锚点的 MOTR 跟踪器。通过这种基于锚点的建模，额外物体检测器生成的提议可以作为 MOTR 的锚点初始化，提供局部先验。变换器解码器用于预测相对于锚点的相对偏移，使得检测任务的优化变得更加容易。

与原始 MOTR 相比，所提出的 MOTRv2 带来了许多优势。它极大地受益于额外物体检测器带来的良好检测性能。检测任务隐式地与 MOTR 框架解耦，从而缓解了共享变换器解码器中检测与关联任务之间的冲突。MOTRv2 在获得来自额外检测器的检测结果的基础上，学习在帧之间跟踪实例。

2 相关工作

2.1 基于检测的跟踪

主流的方法 [6, 31] 主要遵循基于检测的跟踪流程：物体检测器首先预测每帧的物体边界框，然后使用单独的算法将相邻帧中的实例边界框进行关联。这些方法的性能在很大程度上依赖于物体检测的质量。

许多研究尝试使用匈牙利算法 [12] 进行关联：SORT [4] 为每个被跟踪的实例应用卡尔曼滤波器 [26]，并使用卡尔曼滤波器预测框和检测框之间的交并比（IoU）矩阵进行匹配。DeepSORT [27] 引入了一个单独的网络来提取实例的外观特征，并在 SORT 的基础上使用成对的余弦距离。JDE [25]、TrackRCNN [18]、FairMOT [32] 和 Unicorn [29] 进一步探索了物体检测和外观嵌入的联合训练。ByteTrack [31] 利用强大的基于 YOLOX 的检测器 [10]，并取得了最先进的性能。它引入了一种增强的 SORT 算法，除了关联高分检测框外，也关联低分检测框。BoT-SORT [1] 进一步设计了更好的卡尔曼滤波状态、相机运动补偿和 ReID 特征融合。TransMOT [9] 和 GTR [34] 在计算分配矩阵时，采用空间-时间变换器进行实例特征交互和历史信息聚合。OCSORT [6] 放宽了线性运动假设，并使用可学习的运动模型。虽然作者的方法也受益于强大的检测器，但他们不计算相似性矩阵，而是使用带锚点的跟踪查询来联合建模运动和外观。

2.2 基于查询传播的跟踪

另一种 MOT 方法扩展了基于查询的物体检测器 [7, 21, 35] 以进行跟踪。这些方法强制每个查询在不同帧中回忆相同的实例。查询与图像特征之间的交互可以并行或串行进行。

并行方法将短视频作为输入，使用一组查询与所有帧进行交互，以预测实例的轨迹。VisTR [24] 和后续工作 [8, 28] 将 DETR [7] 扩展到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。

串行方法逐帧与图像特征进行查询交互，并迭代更新与实例相关的跟踪查询。Tracktor++ [2] 利用 R-CNN [11] 回归头进行跨帧的实例重新定位。TrackFormer [15] 和 MOTR [30] 从可

变形 DETR [35] 扩展。它们预测物体边界框并更新跟踪查询，以在后续帧中检测相同的实例。MeMOT [5] 构建短期和长期实例特征记忆库，以生成跟踪查询。TransTrack [20] 传播跟踪查询一次，以找到下一帧中的物体位置。P3AFormer [33] 采用流引导的图像特征传播。与 MOTR 不同，TransTrack 和 P3AFormer 仍然在历史轨迹和当前检测中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。

原论文作者的方法继承了用于长期端到端跟踪的查询传播方法，同时利用强大的物体检测器提供物体位置先验。所提出的方法在复杂运动中的跟踪性能上大大超过了现有的匹配和基于查询的方法。

3 本文方法

3.1 回顾 MOTR

MOTR [30] 是一个完全端到端的多目标跟踪框架，基于可变形 DETR [35] 架构构建。它引入了跟踪查询和对象查询。对象查询负责检测新出现或遗漏的物体，而每个跟踪查询则负责随着时间跟踪一个独特的实例。为了初始化跟踪查询，MOTR 使用与新检测到的物体关联的对象查询输出。跟踪查询随着时间的推移通过其状态和当前图像特征进行更新，这使得它们能够以在线方式预测轨迹。

MOTR 中的轨迹感知标签分配将跟踪查询分配给它们之前跟踪的实例，同时通过二分匹配将对象查询分配给剩余的实例。MOTR 引入了一个时间聚合网络，以增强跟踪查询的特征，并采用集体平均损失来平衡跨帧的损失。

3.2 总体结构

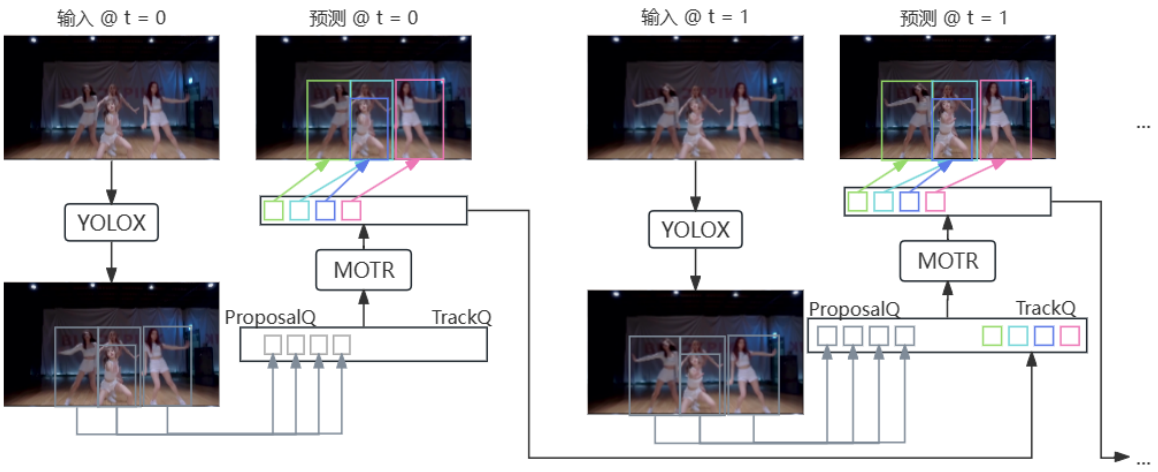


图 1. MOTRv2 的整体架构

如图 1 所示，MOTRv2 架构由两个主要组件组成：一个最先进的物体检测器和一个修改后的基于锚的 MOTR 跟踪器。物体检测器组件首先生成用于训练和推理的建议。对于每一帧，YOLOX 生成一组建议，包括中心坐标、宽度、高度和置信度值。修改后的基于锚的

MOTR 组件负责根据生成的建议学习轨迹关联。第 3.3 节描述了在原始 MOTR 框架中用提议查询替换检测查询的过程。修改后的 MOTR 现在将跟踪查询和提议查询的拼接作为输入。第 3.4 节描述了拼接查询与帧特征之间的交互，以更新被跟踪物体的边界框。

3.3 提议查询生成

在本节解释了提议查询生成模块如何为 MOTR 提供来自 YOLOX 的高质量提议。该模块的输入是 YOLOX 为视频中的每一帧生成的一组提议框。与 DETR [7] 和 MOTR 使用固定数量的可学习查询进行物体检测不同，MOTRv2 框架根据 YOLOX 生成的选定提议动态确定提议查询的数量。

具体来说，对于帧 t，YOLOX 生成 N_t 个提议，每个提议由一个边界框表示，包含中心坐标 (x_t, y_t)、高度 h_t、宽度 w_t 和置信度得分 s_t。如图 2 的橙色部分所示，引入了一个共享查询 q_s 来生成一组提议查询。共享查询是一个大小为 $1\times D$ 的可学习嵌入，首先被广播到 $N_t\times D$ 的大小。N_t 个提议框的预测得分 s_t 通过正弦余弦位置编码生成大小为 $N_t\times D$ 的得分嵌入。广播的查询与得分嵌入相加，以生成提议查询。YOLOX 的提议框作为提议查询的锚点。

在实践中，作者还使用 10 个可学习的锚点（类似于 DAB-DETR [13]），并将它们与 YOLOX 提议拼接，以回忆 YOLOX 检测器遗漏的物体。

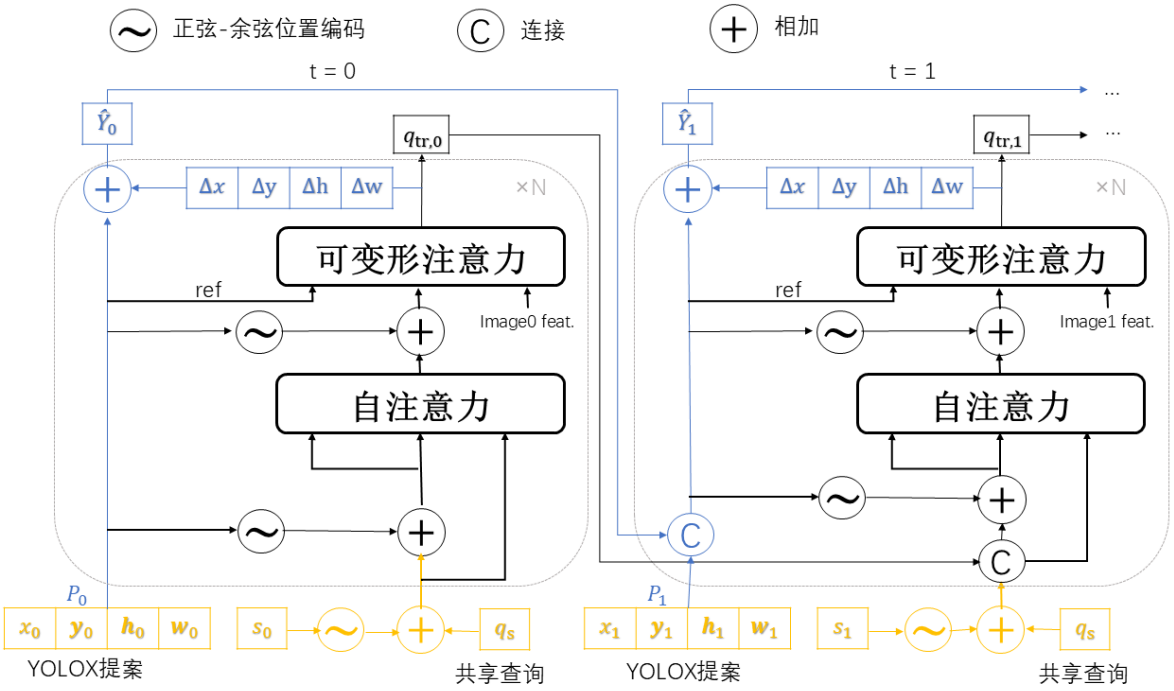


图 2. 提案查询生成和提案传播用于跟踪。橙色部分标记了提案查询的生成，蓝色部分标记了提案传播路径；虚线灰框表示 N 个变换器解码器。为简化起见，图中省略了 MOTR 中的查询交互模块。

3.4 提议传播

在 MOTR [30] 中，跟踪查询和检测查询被拼接并输入到变换器解码器中，以同时进行物体检测和轨迹关联。来自前一帧生成的跟踪查询表示被跟踪的物体，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，用于检测新出现的物体。与 MOTR 不同，MOTRv2 使用提议查询来检测新出现的物体，而跟踪查询的预测则基于前一帧的预测。

对于第一帧（$t = 0$），只有新出现的物体，这些物体由 YOLOX 进行检测。如上所述，提议查询是基于共享查询 q_s 和 YOLOX 提议的预测得分生成的。在 YOLOX 提议 P_0 进行位置编码后，提议查询进一步通过自注意力更新，并通过可变形注意力与图像特征交互，以生成跟踪查询 $q_{tr,0}$ 及相对于 YOLOX 提议 P_0 的相对偏移 $(\Delta \mathrm{x}, \Delta \mathrm{y}, \Delta \mathrm{w}, \Delta \mathrm{h})$。预测 $\hat{Y_0}$ 是提议 P_0 和预测偏移的总和。

对于其他帧（$\mathrm{t} > 0$），与 MOTR 相似，来自前一帧生成的跟踪查询 $q_{tr,t-1}$ 将与当前帧的提议查询 $q_{p,t}$ 拼接。前一帧的边界框预测 $\hat{Y}_{t-1}$ 也将与 YOLOX 提议 P_t 拼接在一起，作为当前帧的锚点。锚点的正弦余弦编码用作拼接查询的位置嵌入，然后输入到变换器解码器中，以生成预测和更新的跟踪查询。边界框预测由置信度得分和相对于锚点的相对偏移组成，更新后的跟踪查询 $q_{tr,t}$ 进一步传递到下一帧，以检测被跟踪的物体。

4 复现细节

4.1 与已有开源代码对比

本次复现主要使用 MOTRv2 的官方代码库，在 MOTRv2 的基础上，添加了一个新的可视化模块。这一模块可以实时展示跟踪过程中的目标状态，帮助用户更直观地理解跟踪效果。

4.2 数据集

本文使用 DanceTrack [19] 数据集来评估 MOTRv2。DanceTrack [19] 是一个大规模的数据集，用于舞蹈场景中的多人跟踪。该数据集具有统一的外观和多样的运动，这对跨帧关联实例提出了挑战。DanceTrack 包含 100 个视频：40 个用于训练，25 个用于验证，35 个用于测试。视频的平均长度为 52.9 秒。

4.3 评估指标

本文使用高阶指标来评估多目标跟踪方法（Higher Order Tracking Accuracy；HOTA）[14]，并分析其在检测准确率（DetA）和关联准确率（AssA）方面的贡献，同时列出了 MOTA [3] 和 IDF1 [16] 指标。

4.4 实现细节

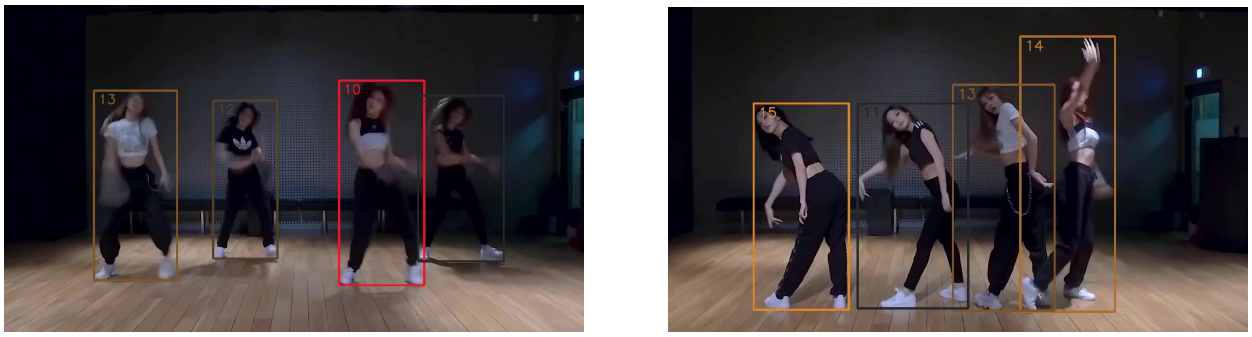
为了提高检测性能，作者还利用了大量静态 CrowdHuman 图像。对于 DanceTrack 数据集，类似于 MOTR 中 MOT17 与 CrowdHuman 的联合训练，为 CrowdHuman 生成伪视频剪辑，并与 DanceTrack 进行联合训练。伪视频剪辑的长度也设置为 5。一共使用了 DanceTrack [19] 数据集的 41,796 个训练样本和 CrowdHuman [17] 数据集的 19,370 个训练和验证样本进行联合训练。

4.5 实验环境

操作系统为 Ubuntu 18.04 的 Linux，Python 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 和 opencv-python 库。使用的显卡为 4 块 NVIDIA Tesla V100。

5 实验结果分析

本次实验结果基于 DanceTrack 测试集得出，输出结果包括当前视频帧的编号、被跟踪目标的唯一标识符，以及目标边界框左上角的 x 和 y 坐标、宽度和高度。每一帧的结果如图 3 所示，所有舞者均有唯一标识符进行标记和跟踪，目标框基本上包围了舞者。然而，测试结果显示同一舞者被分配了两个跟踪查询，例如图 3a 中的目标 10 和图 3b 中的目标 14 实际上是同一舞者，模型出现了重复跟踪的情况。



(a) 未重复跟踪情况　　(b) 重复跟踪情况

图 3. 实验结果可视化

表 1 展示了原论文在 DanceTrack 数据集上的实验结果与我的复现结果的对比。除了 DetA 和 MOTA 的值与原论文相近外，其他指标均低于原论文实验结果 2% 以上。这主要是因为复现时训练轮数不足，仅为 5 轮，而原论文中训练了 15 轮。这可能导致模型未能充分学习数据特征，因此未能达到理想的收敛效果。

表 1. 原论文实验结果与我复现结果对比

结果	HOTA	DetA	AssA	MOTA	IDF1
原论文	69.9	83.0	59.0	91.9	71.7
复现	67.9	82.6	55.9	91.1	69.3

6 总结与展望

在本文中，原论文作者提出的 MOTRv2，它优雅地结合了 MOTR 跟踪器和 YOLOX 检测器。YOLOX 生成高质量的物体提议，帮助 MOTR 更轻松地检测新物体。这降低了物体检测的复杂性，使 MOTR 能够专注于关联过程。MOTRv2 突破了普遍认为端到端框架不适合高性能多目标跟踪的观念，并解释了之前端到端多目标跟踪框架失败的原因。

尽管使用 YOLOX 提议大大简化了 MOTR 的优化问题，但该方法仍然对数据需求较高，并且在较小数据集上的表现不够理想。此外，我们观察到在某些情况下，比如两个个体相交时，会出现一些重复的跟踪查询。在这种情况下，一个跟踪查询可能会跟随错误的目标，导致对同一个体存在两个跟踪查询。

参考文献

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

[2] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[3] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008.

[4] 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, 2016.

[5] Jiarui Cai, Mingze Xu, Wei Li, Yuanjun Xiong, Wei Xia, Zhuowen Tu, and Stefano Soatto. Memot: Multi-object tracking with memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8090–8100, June 2022.

[6] Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, and Kris Kitani. Observation-centric sort: Rethinking sort for robust multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9686–9696, June 2023.

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 213–229, Cham, 2020. Springer International Publishing.

[8] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1290–1299, June 2022.

[9] Peng Chu, Jiang Wang, Quanzeng You, Haibin Ling, and Zicheng Liu. Transmot: Spatial-temporal graph transformer for multiple object tracking. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4870–4880, January 2023.

[10] Z Ge. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

[11] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[12] H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955.

[13] 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.

[14] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *International journal of computer vision*, 129:548–578, 2021.

[15] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixé, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8844–8854, June 2022.

[16] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In Gang Hua and Hervé Jégou, editors, *Computer Vision – ECCV 2016 Workshops*, pages 17–35, Cham, 2016. Springer International Publishing.

[17] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. Crowdhuman: A benchmark for detecting human in a crowd, 2018.

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

[19] Peize Sun, Jinkun Cao, Yi Jiang, Zehuan Yuan, Song Bai, Kris Kitani, and Ping Luo. Dancetrack: Multi-object tracking in uniform appearance and diverse motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20993–21002, June 2022.

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

[21] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14454–14463, June 2021.

[22] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[23] Yingming Wang, Xiangyu Zhang, Tong Yang, and Jian Sun. Anchor detr: Query design for transformer-based object detection. *arXiv preprint arXiv:2109.07107*, 3(6), 2021.

[24] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8741–8750, June 2021.

[25] Zhongdao Wang, Liang Zheng, Yixuan Liu, Yali Li, and Shengjin Wang. Towards real-time multi-object tracking. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 107–122, Cham, 2020. Springer International Publishing.

[26] G Welch. An introduction to the kalman filter. 1995.

[27] 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, 2017.

[28] Junfeng Wu, Yi Jiang, Song Bai, Wenqing Zhang, and Xiang Bai. Seqformer: Sequential transformer for video instance segmentation. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 553–569, Cham, 2022. Springer Nature Switzerland.

[29] Bin Yan, Yi Jiang, Peize Sun, Dong Wang, Zehuan Yuan, Ping Luo, and Huchuan Lu. Towards grand unification of object tracking. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 733–751, Cham, 2022. Springer Nature Switzerland.

[30] Fangao Zeng, Bin Dong, Yuang Zhang, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. Motr: End-to-end multiple-object tracking with transformer. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 659–675, Cham, 2022. Springer Nature Switzerland.

[31] 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 Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 1–21, Cham, 2022. Springer Nature Switzerland.

[32] 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.

[33] Zelin Zhao, Ze Wu, Yueqing Zhuang, Boxun Li, and Jiaya Jia. Tracking objects as pixel-wise distributions. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 76–94, Cham, 2022. Springer Nature Switzerland.

[34] Xingyi Zhou, Tianwei Yin, Vladlen Koltun, and Philipp Krähenbühl. Global tracking transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8771–8780, June 2022.

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