MOTR: トランスフォーマーによるエンドツーエンドの複数オブジェクト追跡

Fangao Zeng[1⋆], Bin Dong[1⋆], Yuang Zhang[2⋆], Tiancai Wang[1⋆⋆], Xiangyu Zhang[1], and Yichen Wei[1]

[1] MEGVII Technology
[2] Shanghai Jiao Tong University

概要 MOT(Multipleobject Tracking)において、物体の時間的モデリングは重要な課題である。既存の手法は、モーションベースとアピアランスベースの類似性ヒューリスティックによって、検出を関連付けることで追跡する。アソシエーションの後処理という性質は、ビデオシーケンスの時間的変化をエンドツーエンドで利用することを妨げる。
本論文では、DETR[6]を拡張し、ビデオ全体で追跡されたインスタンスをモデル化するために「追跡クエリ」を導入したMOTRを提案する。トラッククエリはフレームごとに転送され、更新され、時間の経過とともに反復予測が実行される。トラッククエリと新生オブジェクトクエリを学習するために、トラックレットを考慮したラベル割り当てを提案する。さらに、時間的関係モデリングを強化するために、時間的集約ネットワークと集団平均損失を提案する。DanceTrackでの実験結果から、MOTRはHOTA指標において、最先端手法であるByteTrack [42]を6.5%大幅に上回ることが示された。MOT17において、MOTRは我々の同時並行研究であるTrackFormer [18]とTransTrack [29]を連想性能で上回った。MOTRは、時間モデリングとTransformerベースのトラッカーに関する今後の研究のための、より強力なベースラインとして機能する。コードはhttps://github.com/megvii-research/MOTRにある。

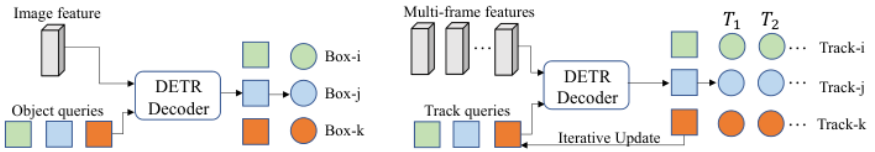
キーワード Multiple-Object Tracking, トランスフォーマー, エンドツーエンド

1 Introduction

複数物体追跡(MOT)は、連続画像シーケンスにおけるインスタンスの軌跡を予測する[39,2]。既存の手法の多くは、MOTの時間的関連性を外観と動きに分離している。外観の分散は通常、ペアワイズRe-ID類似度[37,43]で測定され、動きはIoU[4]またはカルマンフィルタリング[3]ヒューリスティックでモデル化される。これらの方法は、後処理のために類似性に基づくマッチングを必要とし、これがフレーム間の時間的情報フローのボトルネックとなる。本論文では、モーションと外観のジョイントモデリングを特徴とする、完全なエンドツーエンドのMOTフレームワークを紹介することを目的とする。

最近、エンドツーエンドの物体検出のためにDETR [6,45]が提案された。物体検出を集合予測問題として定式化する。

⋆ Equal contribution.
⋆⋆ Corresponding author. Email: wangtiancai@megvii.com



(a) One-shot prediction in DETR. (b) MOTRにおけるボックス列の反復予測。

図1：(a) DETRはオブジェクトクエリと画像特徴を相互作用させることでエンドツーエンドの検出を実現し、更新されたクエリとオブジェクトの間で1対1の割り当てを実行する。(b) MOTRはトラッククエリを更新することで、シーケンス予測セットを実行する。各トラッククエリはトラックを表す。カラーで見るのがベスト

図1(a)に示すように、オブジェクトの非結合表現として機能するオブジェクトクエリは、Transformerデコーダに供給され、画像特徴と相互作用してその表現を更新する。さらに、オブジェクトクエリとグランドトゥルース間の一対一の割り当てを達成するために、NMSのような後処理を排除した二分割マッチングが採用される。物体検出とは異なり、MOTはシーケンス予測問題とみなすことができる。エンドツーエンドのDETRシステムで配列予測を行う方法は未解決の問題である。

反復予測は機械翻訳でよく使われる[30,31]。出力文脈は隠れ状態によって表現され、文の特徴はデコーダの隠れ状態と繰り返し相互作用して翻訳単語を予測する。機械翻訳におけるこれらの進歩に触発され、我々はMOTを直感的にシーケンス予測の集合の問題とみなす。MOTはオブジェクトシーケンスの集合を必要とするからである。各シーケンスはオブジェクトの軌跡に対応する。技術的には、DETRのオブジェクトクエリを拡張し、オブジェクトシーケンスを予測するためのクエリを追跡する。トラッククエリはオブジェクトトラックの隠れ状態として提供される。トラッククエリの表現はTransformerデコーダで更新され、図1(b)に示すように、オブジェクトの軌跡を繰り返し予測するために使用される。具体的には、トラッククエリは自己アテンションとフレーム特徴によるクロスアテンションによって更新される。更新されたトラッククエリは、さらにバウンディングボックスを予測するために使用される。1つのオブジェクトのトラックは、異なるフレームにおける1つのトラッククエリのすべての予測から得ることができる。

上記の目標を達成するためには、2つの問題を解決する必要がある:1)1つのトラッククエリによって1つのオブジェクトを追跡する、2)新生オブジェクトと終了オブジェクトを扱う。最初の問題を解決するために、トラックレットを考慮したラベル割り当て(TALA)を導入する。これは、1つのトラッククエリの予測が、同じIDを持つバウンディングボックスシーケンスによって監督されることを意味する。第二の問題を解決するために、可変長のトラッククエリセットを保持する。新生オブジェクトのクエリはこのセットにマージされ、終了オブジェクトのクエリは削除される。このプロセスを入口と出口のメカニズムと呼ぶ。このように、MOTRは推論時に明示的なトラック関連付けを必要としない。さらに、トラッククエリの反復更新により、外観と動きの両方に関する時間的モデリングが可能になる。

時間的モデリングを強化するために、我々はさらに集団平均損失(CAL)と時間的集約ネットワーク(TAN)を提案する。CALでは、MOTRは

学習時の入力としてのビデオクリップMOTRのパラメータは、ビデオクリップ全体に対して計算された総合損失に基づいて更新される。TANはトラッククエリのショートカットを導入し、Transformerのキークエリメカニズムを介して、以前の状態からの履歴情報を集約する。
MOTRはシンプルなオンライントラッカーである。ラベル割り当てを少し修正したDETRに基づく開発は容易である。これは真にエンドツーエンドのMOTフレームワークであり、我々の同時並行研究であるTransTrack [29]やTrackFormer [18]で採用されているトラックNMSやIoUマッチングのような後処理を必要としない。MOT17とDanceTrackデータセットでの実験結果から、MOTRは有望な性能を達成していることがわかる。DanceTrack[28]において、MOTRは最先端のByteTrack[42]をHOTA指標で6.5%、AssAで8.1%上回った。

To summarize, our contributions are listed as below:

- MOTRと名付けられた完全なエンドツーエンドのMOTフレームワークを提示する。MOTRは、外観と位置の分散を暗黙のうちに共同で学習することができる。MOTをシーケンス予測の集合の問題として定式化する。反復更新と予測のために、以前の隠れ状態からトラッククエリを生成する。トラッククエリとオブジェクト間の一対一の割り当てのために、トラックレットを考慮したラベル割り当てを提案する。新生児と終了したトラックに対処するために、入口と出口のメカニズムが導入される。さらに、時間的モデリングを強化するために、CALとTANを提案する。

2 関連研究

トランスフォーマーベースのアーキテクチャTransformer[31]は、機械翻訳のために、入力シーケンス全体から情報を集約するために最初に導入された。主に自己注意と相互注意のメカニズムが関与している。それ以来、音声処理[13,7]やコンピュータビジョン[34,5]など、多くの分野に徐々に導入されている。最近、DETR [6]は畳み込みニューラルネットワーク(CNN)、Transformer、二分割マッチングを組み合わせて、エンドツーエンドの物体検出を行う。高速な収束を達成するために、変形可能なDETR [45]は変形可能な注意モジュールをTransformerエンコーダとTransformerデコーダに導入した。ViT[9]は、画像分類のための純粋なTransformerアーキテクチャを構築した。さらに、Swin Transformer [16]は、ローカルウィンドウ内で自己注意を行うために、シフトウィンドウ方式を提案し、より高い効率をもたらした。VisTR[36]は、ビデオインスタンスのセグメンテーションを実行するために、直接エンドツーエンドの並列シーケンス予測フレームワークを採用した。
複数物体追跡。支配的なMOT手法は、主にトラッキング・バイ・ディテクションのパラダイムに従った[3,12,22,24,39]。これらのアプローチは通常、まずオブジェクト検出器を用いて各フレーム内のオブジェクトを特定し、次に隣接するフレーム間のトラック関連付けを行い、トラッキング結果を生成する。SORT [3]は、カルマンフィルタ[38]とハンガリーアルゴリズム[11]を組み合わせてトラック関連付けを行った。DeepSORT[39]とTracktor[2]は余弦距離を追加導入し、トラック関連付けのための外観類似度を計算する。Track-RCNN[26]、JDE[37]、FairMOT[43]はさらに、物体検出とRe-ID特徴学習を組み込んだ共同学習フレームワークにおいて、物体検出器の上にRe-ID分岐を追加した。

TransMOT [8]は、関連付けのための空間-時間グラフ変換器を構築する。我々の同時進行の研究であるTransTrack [29]とTrackFormer [18]も、MOTのためのTransformerベースのフレームワークを開発している。と直接比較する場合
3.7節を参照してください。反復シーケンス予測エンコーダ・デコーダアーキテクチャを用いたsequence-to-sequence (seq2seq)による配列予測は、機械翻訳[30,31]やテキスト認識[25]でよく利用されている。seq2seqフレームワークでは、エンコーダネットワークは入力を中間表現にエンコードする。次に、タスク固有のコンテキスト情報を持つ隠れ状態を導入し、中間表現と繰り返し相互作用して、デコーダネットワークを介してターゲットシーケンスを生成する。反復デコード処理にはいくつかの反復が含まれる。各反復において、隠れ状態はターゲット配列の1つの要素を復号する。

3 Method

3.1 物体検出におけるクエリ

DETR [6]は、オブジェクトを検出するために、固定長のオブジェクトクエリのセットを導入した。オブジェクトクエリはTransformerデコーダに供給され、Transformerエンコーダから抽出された画像特徴と相互作用して、その表現を更新する。さらに、更新されたオブジェクトクエリとグランドトゥルース間の一対一の割り当てを達成するために、二分割マッチングが採用される。ここでは、物体検出に用いるクエリを指定するために、単純に物体クエリを「検出クエリ」と書く。

3.2 クエリの検出とクエリの追跡

物体検出からMOTにDETRを適応する場合、2つの主な問題が生じる:1)1つの追跡クエリによって1つの物体をどのように追跡するか、2)新生物体や終了物体をどのように扱うか。本論文では、検出クエリを拡張してクエリを追跡する。トラッククエリセットは動的に更新され、長さは可変である。図2に示すように、トラッククエリセットは空になるように初期化され、DETRの検出クエリは新生オブジェクト(T_2のオブジェクト3)を検出するために使用される。検出されたオブジェクトの隠された状態は、次のフレームのトラッククエリを生成する。終了したオブジェクトに割り当てられたトラッククエリは、トラッククエリセット(T_4のオブジェクト2)から削除される。

3.3 トラックレットを考慮したラベル割り当て

DETRでは、ラベルの割り当ては、すべての検出クエリとグランドトゥルースとの間で二分割マッチングを実行することによって決定されるため、1つの検出(オブジェクト)クエリが画像内の任意のオブジェクトに割り当てられる可能性がある。MOTRでは、検出クエリは新生オブジェクトの検出にのみ使用され、追跡クエリは追跡されたオブジェクトをすべて予測する。ここでは、この問題を解決するために、トラックレットを考慮したラベル割り当て(TALA)を導入する。
一般に、TALAは2つの戦略から構成される。検出クエリについては、DETRの割り当て戦略を新生児のみとして修正し、

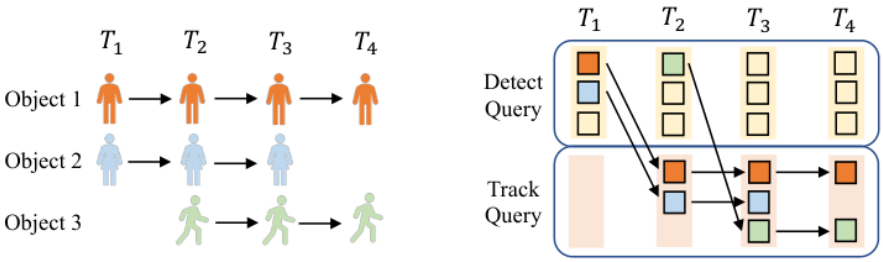


図2:いくつかの典型的なMOTケースにおける(オブジェクト)クエリの検出とクエリの追跡の更新プロセス。トラッククエリセットは動的に更新され、長さは可変である。トラッククエリセットは空になるように初期化され、検出クエリは新生オブジェクトの検出に使用される。検出されたすべてのオブジェクトの隠された状態は、次のフレームのためのトラッククエリを生成するために連結される。終了したオブジェクトに割り当てられたトラッククエリは、トラッククエリセットから削除される。

検出クエリと新生児オブジェクトのグランドトゥルースとの間で二分割マッチングを行う。トラッククエリに対して、ターゲットと矛盾しない割り当て戦略を設計する。トラッククエリは前のフレームと同じ割り当てに従うので、前述の二分割マッチングから除外される。
形式的には、トラッククエリの予測値をY_{btr}、検出クエリの予測値をY_{bdet}とする。Y_{new} は新生オブジェクトの基底真理である。トラッククエリと検出クエリのラベル割り当て結果は、ω_{tr}とω_{det}と書くことができる。フレームiについて、検出クエリのラベル割り当ては、検出クエリと新生オブジェクト間の二分割マッチングから得られる、

$$\omega_{det}^{i} = \underset{\omega_{det}^{i} \in \Omega_i}{\arg\min}\ \mathcal{L}(\widehat{Y}_{det}^{i}|_{\omega_{det}^{i}}, Y_{new}^{i}), \tag{1}$$

ここで、LはDETRで定義されたペアワイズマッチングコストであり、Ω_iは検出クエリと新生オブジェクト間の全ての二分割マッチングの空間である。トラッククエリの割り当てについては、新生オブジェクトと最後のフレームから追跡されたオブジェクトの割り当てをマージする、すなわち、i > 1の場合:

$$\omega_{tr}^{i} = \omega_{tr}^{i-1} \cup \omega_{det}^{i-1}. \tag{2}$$

最初のフレーム(i = 1)では、トラッククエリ割り当てω_{tr}^{1}は空集合ϕである。連続するフレーム(i > 1)に対して、トラッククエリ割り当てω_{tr}^{i}は、前のトラッククエリ割り当てω_{tr}^{i-1}と新生オブジェクト割り当てω_{det}^{i-1}の連結である。

実際には、TALA戦略はTransformerの強力な注意メカニズムのおかげで、シンプルで効果的である。各フレームについて、検出クエリとトラッククエリは連結され、それらの表現を更新するためにTransformerデコーダに供給される。Transformerデコーダの自己注意によるクエリの相互作用は、追跡されたオブジェクトを検出するクエリの検出を抑制するため、クエリの検出は新生オブジェクトのみを検出する。このメカニズムは、DETRにおける重複排除と同様に、重複ボックスは低スコアで抑制される。

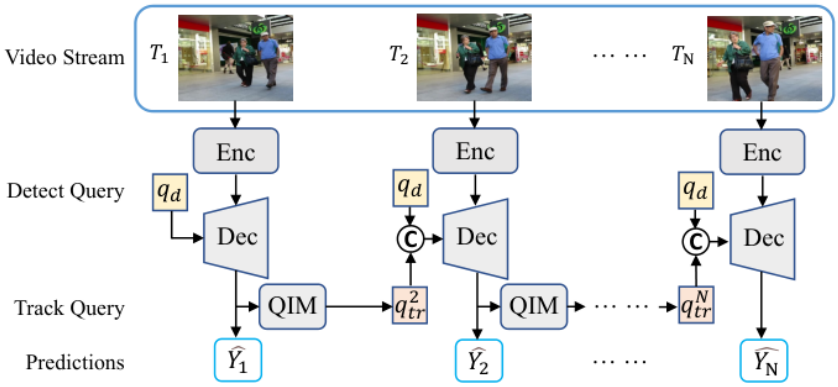


図3:MOTRの全体アーキテクチャ。”Enc ”は畳み込みニューラルネットワークのバックボーンを表し、各フレームの画像特徴を抽出するTransformerエンコーダを表す。検出クエリq_dと追跡クエリq_{tr}の連結は、Deformable DETRデコーダ(Dec)に供給され、隠れ状態を生成する。隠れ状態は、新生オブジェクトと追跡オブジェクトの予測値Y_bを生成するために使用される。クエリ相互作用モジュール(QIM)は、隠れ状態を入力とし、次のフレームのトラッククエリを生成する。

3.4 MOTRアーキテクチャ

MOTRの全体アーキテクチャを図3に示す。ビデオシーケンスは、フレーム特徴を抽出するために、畳み込みニューラルネットワーク(CNN)(例えばResNet-50 [10])と変形可能なDETR [45]エンコーダに供給される。
最初のフレームでは、トラッククエリは存在せず、固定長の学習可能な検出クエリ(図3のq_d)のみを変形可能なDETR[45]デコーダに入力する。連続するフレームに対して、前のフレームからのトラッククエリと学習可能な検出クエリの連結をデコーダに入力する。これらのクエリはデコーダの画像特徴と相互作用し、バウンディングボックス予測のための隠れ状態を生成する。また、隠れ状態は次のフレームのトラッククエリを生成するために、クエリ相互作用モジュール(QIM)に供給される。
学習段階では、各フレームのラベル割り当てを第3.3節で説明する。ビデオクリップの全ての予測は予測バンク{ Y_{b1}, Y_{b2},.. Y_{bN} }に集められ、監督のために3.6節で説明した提案された集団平均損失(CAL)を用いる。推論時間中、ビデオストリームはオンラインで処理され、各フレームの予測を生成することができる。

3.5 クエリ対話モジュール

本節では、クエリインタラクションモジュール(QIM)について説明する。QIMには、オブジェクトの出入り機構と時間集約ネットワーク(TAN)が含まれる。オブジェクトの出現と終了。上述したように、ビデオシーケンス中のいくつかのオブジェクトは、中間フレームに現れたり消えたりすることがある。

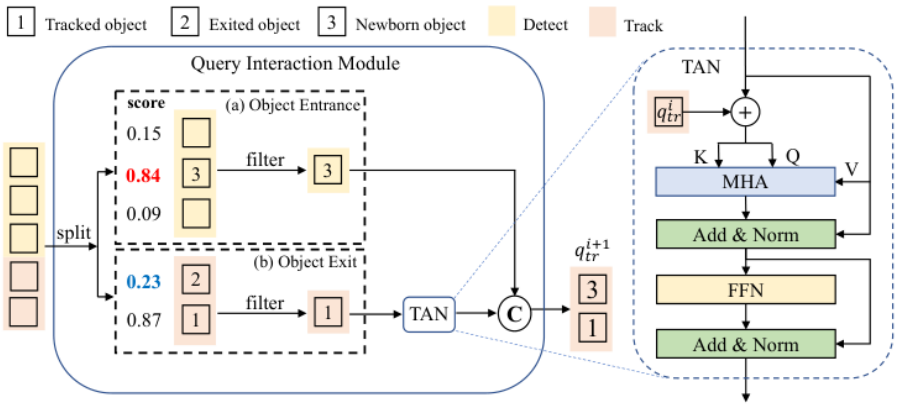


図4:クエリインタラクションモジュール(QIM)の構造。QIMの入力は、Transformerデコーダによって生成された隠れ状態と、それに対応する予測スコアである。推論段階では、新生オブジェクトを保持し、信頼度スコアに基づいて退出したオブジェクトを削除する。時間集約ネットワーク(TAN)は、長距離の時間モデリングを強化する。

ここでは、本手法における新生オブジェクトと終了オブジェクトの扱い方を紹介する。任意のフレームに対して、トラッククエリは検出クエリと連結され、Transformerデコーダに入力され、隠れ状態を生成する(図4の左側を参照)。
学習中、一致したオブジェクトがグランドトゥルースで消失した場合、または予測されたバウンディングボックスとターゲットの間のIoU(intersection-over-union)が閾値0.5以下であれば、終了オブジェクトの隠れ状態は削除される。これは、これらのオブジェクトが現在のフレームで消滅し、残りの隠れ状態が予約されている場合、対応する隠れ状態がフィルタリングされることを意味する。新生オブジェクトについては、式1で定義される新生オブジェクトω_{det}^{i}の割り当てに基づいて、対応する隠れ状態が保持される。

推論には、図4に示すように、予測された分類スコアを用いて、新生オブジェクトの出現と追跡オブジェクトの消失を決定する。オブジェクトクエリでは、分類スコアが入口閾値τ_{en}より高い予測は保持され、他の隠れ状態は除去される。トラッククエリでは、連続するMフレームに対して分類スコアが終了閾値τ_{ex}より低い予測は削除され、他の隠れ状態は保持される。
時間集約ネットワーク。ここでは、時間関係モデリングを強化し、追跡されたオブジェクトの文脈的事前分布を提供するために、QIMに時間集約ネットワーク(TAN)を導入する。

図4に示すように、TANの入力は追跡対象(物体 "1")のフィルタリングされた隠れ状態である。また、時間的集約のために、最後のフレームからトラッククエリq_{tr}^{i}を収集する。TANは修正Transformerデコーダ層である。最後のフレームからのトラッククエリとフィルタリングされた隠れ状態は、マルチヘッド自己アテンション(MHA)のキーとクエリのコンポーネントとして合計される。

隠れ状態のみがMHAの値成分である。MHAの後、フィードフォワードネットワーク(FFN)を適用し、その結果を新生オブジェクト(オブジェクト ”3”)の隠れ状態と連結して、次のフレームのトラッククエリセットq_{tr}^{i+1}を生成する。

3.6 集合平均損失

MOTRはカルマンフィルタリングのような手作業によるヒューリスティックではなく、データから時間的分散を学習するため、トラックの時間的モデリングには学習サンプルが重要である。2フレーム以内のトレーニングのような一般的なトレーニング戦略は、長距離のオブジェクトの動きのトレーニングサンプルを生成できない。MOTRは、それらとは異なり、ビデオクリップを入力とする。このようにして、時間学習のために、長距離物体運動の学習サンプルを生成することができる。
フレーム毎に損失を計算する代わりに、我々の集団平均損失(CAL)は複数の予測値$Y_b = \{ Y_{bi} \}_{i=1}^{N}$を収集する。次に、ビデオシーケンス全体の損失は、グランドトゥルース$Y = \{Y_i\}_{i=1}^{N}$とマッチング結果$\omega = \{\omega_i\}_{i=1}^{N}$によって計算される。CALは、ビデオシーケンス全体の損失であり、オブジェクトの数で正規化される:

$$\mathcal{L}_o(\widehat{Y}|_{\omega}, Y) = \frac{\sum_{n=1}^{N} (\mathcal{L}(\widehat{Y}_{tr}^{i}|_{\omega_{tr}^{i}}, Y_{tr}^{i}) + \mathcal{L}(\widehat{Y}_{det}^{i}|_{\omega_{det}^{i}}, Y_{det}^{i}))}{\sum_{n=1}^{N} (V_i)} \tag{3}$$

ここで、$V_i = V_{tr}^{i} + V_{det}^{i}$はフレームiにおける地上真理オブジェクトの総数を表す。V_{tr}^{i}とV_{det}^{i}はそれぞれフレームiにおける追跡対象物と新生対象物の数である。Lは単一フレームの損失であり、DETRの検出損失と同様である。シングルフレーム損失Lは次のように定式化できる:

$$\mathcal{L}(\widehat{Y}_i|_{\omega_i}, Y_i) = \lambda_{cls}\mathcal{L}_{cls} + \lambda_{l_1}\mathcal{L}_{l_1} + \lambda_{giou}\mathcal{L}_{giou} \tag{4}$$

ここで、L_{cls}は焦点損失である[14]。L_{l1}はL1損失、L_{giou}は一般化IoU損失[21]を表す。λ_{cls}, λ_{l1}, λ_{giou} は対応する重み係数である。

3.7 考察

DETRに基づき、我々の同時進行の研究であるTransTrack [29]とTrackFormer [18]もMOTのためのTransformerベースのフレームワークを開発している。しかし、我々の方法はそれらと比較して大きな違いを示す:
TransTrackは、フルトラックを複数の独立した短いトラックレットの組み合わせとしてモデル化する。トラックバイ検出パラダイムと同様に、TransTrackはMOTを2つのサブタスクとして分離する: 1) 隣接する2つのフレーム内の短いトラックレットとしてオブジェクトペアを検出する。2) IoUマッチングによって短いトラックレットを完全なトラックとして関連付ける。一方、MOTRでは、トラッククエリの反復更新により、エンドツーエンドでフルトラックをモデル化し、IoUマッチングを必要としない。

表1:Transformerに基づく他のMOT手法との比較。

Method	IoU match	NMS	ReID
TransTrack [29]	✓		
TrackFormer [18]		✓	✓
MOTR (ours)			

表2:評価のために選択したデータセットの統計量。

Datasets	Class	Frame	Video	ID
DanceTrack [28]	1	106k	100	990
MOT17 [19]	1	11k	14	1342
BDD100K [41]	8	318k	1400	131k

TrackFormerはトラッククエリのアイデアを共有している。しかし、TrackFormerは隣接する2つのフレーム内で学習する。3.6節で述べたように、短距離での学習は比較的弱い時間的学習となる。したがって、TrackFormerは重複トラックをフィルタリングするために、Track NMSやRe-ID特徴などのヒューリスティックを採用している。TrackFormerとは異なり、MOTRはCALとTANでより強い時間運動を学習するため、これらのヒューリスティックの必要性がなくなる。TransTrackやTrackFormerとの直接の比較は、表1を参照してください。

ここで、TrackFormerとTransTrackがarXivに登場するずっと前から、この研究を独自に開始していたことを明確にする。これらは正式には発表されていないことに加え、我々の研究が基礎としている過去の研究ではなく、並行的で独立した研究として扱う。

4 Experiments

4.1 データセットと指標

データセット。包括的な評価のために、3つのデータセットで実験を行った: DanceTrack [28]、MOT17 [19]、BDD100k [41]である。MOT17 [19]には7つの学習シーケンスと7つのテストシーケンスが含まれる。DanceTrack [28]は、均一な外観と多様な動きを特徴とする最近の多オブジェクト追跡データセットである。学習と評価のためのビデオをより多く含むため、トラッキング性能を検証するためのより良い選択となる。BDD100k [41]は、複数のオブジェクトクラスを特徴とするMOTトラックを持つ自律走行データセットである。詳細は、表2に示すデータセットの統計情報をご参照ください。評価指標本手法を評価するために、標準的な評価プロトコルに従う。一般的なメトリクスには、多オブジェクト追跡[17]を評価するための高次メトリクス(HOTA、AssA、DetA)、複数オブジェクト追跡
Accuracy (MOTA), Identity Switches (IDS) and Identity F1 Score (IDF1).

4.2 実装の詳細

CenterTrack[44]の設定に従い、MOTRはランダムフリップやランダムクロップなど、いくつかのデータ補強方法を採用している。入力画像の短辺は800にリサイズされ、最大サイズは1536に制限される。この解像度でのTesla V100での推論速度は約7.5FPSである。フレームレートが変化する問題を解決するために、キーフレームをランダムな間隔でサンプリングする。

また、追跡されたクエリを確率p_{drop}で消去して新生オブジェクトのサンプルをより多く生成し、確率p_{insert}で偽陽性の追跡クエリを挿入して終了オブジェクトをシミュレートする。すべての実験は、8台のNVIDIA Tesla V100 GPUを搭載したPyTorchで実施した。また、NVIDIA 2080 Ti GPUで学習可能なメモリ最適化バージョンも提供する。
ResNet50[10]を用いてDeformable-DETR[45]をベースにMOTRを構築し、高速に収束させた。バッチサイズは1に設定され、各バッチには5フレームのビデオクリップが含まれる。AdamWオプティマイザを用い、初期学習率2.0-10^{-4}でモデルを学習する。すべてのデータセットについて、COCO [15]データセットで事前に訓練された公式の変形可能なDETR [45]重みでMOTRを初期化する。MOT17では、MOTRを200エポック学習させ、100thエポックで学習率を10倍減衰させる。最先端の比較のために、ジョイントデータセット(MOT17トレーニングセットとCrowdHuman [23] valセット)でトレーニングを行う。CrowdHuman val setの約5k枚の静止画像に対して、[44]と同様にランダムシフトを適用し、擬似トラックを持つビデオクリップを生成する。ビデオクリップの初期長さは2であり、50th,90th,150thエポックでそれぞれ3,4,5まで徐々に増加させる。ビデオクリップの長さを漸進的に増加させることで、学習効率と安定性が向上する。アブレーション研究では、CrowdHumanデータセットを使用せずにMOT17トレーニングセットでMOTRをトレーニングし、2DMOT15トレーニングセットで検証する。DanceTrackでは、訓練セットで20エポック学習し、10thエポックで学習率が減衰する。5th,9th,15th エポックでクリップ長を2から3,4,5まで徐々に増加させる。BDD100kでは、訓練セットで20エポック学習し、16thエポックで学習率が減衰する。6thエポックと12thエポックでクリップ長を2から3、4まで徐々に増加させる。

4.3 MOT17における最先端技術の比較

表3は、MOT17テストセットにおいて、我々のアプローチと最先端手法を比較したものである。我々は主にMOTRをTransformerに基づく我々の同時並行的な研究と比較する: TrackFormer[18]とTransTrack[29]である。我々の手法はTransTrackとTrackFormerを4.5%上回り、より高いIDF1スコアを得た。HOTA メトリックにおける MOTR の性能は、TransTrack よりも 3.1%高い。MOTA指標では、我々の手法はTrackFormerよりもはるかに優れた性能を達成した(71.9%対65.0%)。興味深いことに、MOTAではTransTrackの性能が我々のMOTRよりも優れていることがわかった。TransTrackにおける検出枝と追跡枝のデカップリングは、確かに物体検出性能を向上させると考えられる。MOTRでは、検出と追跡のクエリは共有のTransformerデコーダを通して学習される。追跡された物体の検出において、検出クエリが抑制され、新生物体の検出性能が制限される。
ByteTrack[42]のような他の最先端手法と性能を比較すると、MOTRはMOT17データセットにおいて、それらに劣ることがわかる。通常、MOT17データセットにおける最先端の性能は、様々な外観分布に対応するために、検出性能の良いトラッカーが支配的である。また、トラッカーによって物体検出の検出器が異なる傾向がある。様々なトラッカーのモーションパフォーマンスを公平に検証することは、かなり難しい。したがって、MOT17データセットだけでは、MOTRのトラッキング性能を十分に評価することはできないと主張する。

表3:MOT17データセットにおける、プライベート検出プロトコルの下でのMOTRと既存手法の性能比較。数値はTransformerベースの手法の中で最も優れている場合、太字で表示されている。

Methods	HOTA↑	AssA↑	DetA↑	IDF1↑	MOTA↑	IDS↓
CNN-based:						
Tracktor++[2]	44.8	45.1	44.9	52.3	53.5	2072
CenterTrack[44]	52.2	51.0	53.8	64.7	67.8	3039
TraDeS [40]	52.7	50.8	55.2	63.9	69.1	3555
QDTrack [20]	53.9	52.7	55.6	66.3	68.7	3378
GSDT [35]	55.5	54.8	56.4	68.7	66.2	3318
FairMOT[43]	59.3	58.0	60.9	72.3	73.7	3303
CorrTracker [32]	60.7	58.9	62.9	73.6	76.5	3369
GRTU [33]	62.0	62.1	62.1	75.0	74.9	1812
MAATrack [27]	62.0	60.2	64.2	75.9	79.4	1452
ByteTrack [42]	63.1	62.0	64.5	77.3	80.3	2196
Transformer-based:						
TrackFormer [18]	/	/	/	63.9	65.0	3528
TransTrack[29]	54.1	47.9	**61.6**	63.9	**74.5**	3663
MOTR (ours)	**57.8**	**55.7**	60.3	**68.6**	73.4	**2439**

さらに、次に述べるように、一様な外観と多様な動きを持つDanceTrack [28]データセットでトラッキング性能を評価する。

4.4 DanceTrackにおける最先端技術の比較

最近、DanceTrack [28]が紹介された(表2参照)。DanceTrackは、外観が均一で動きが多様なデータセットである。評価用の動画が多く含まれており、トラッキング性能を検証するためのより良い選択肢となる。さらにDanceTrackデータセットで実験を行い、最先端手法との性能比較をTab.4で行う。4. MOTRはDanceTrackデータセットでより優れた性能を達成していることがわかる。我々の手法は、ByteTrackを6.5%上回り、HOTAスコアが大幅に向上した。AssA指標においても、我々の手法はByteTrackよりもはるかに優れた性能を達成した(40.2%対32.1%)。一方、DetA指標では、MOTRはいくつかの最先端手法に劣る。これは、MOTRが時間的運動学習において良い性能を発揮する一方で、検出性能がそれほど良くないことを意味する。HOTAの大きな改善は、主に時間的集約ネットワークと集団平均損失によるものである。

4.5 マルチクラスシーンでの汎化

FairMOT[43]のようなRe-IDベースの手法は、追跡された各オブジェクト(例えば人物)をクラスとみなし、検出結果を特徴の類似性によって関連付ける傾向がある。しかし、追跡対象数が非常に多い場合、関連付けは困難である。これらとは異なり、MOTRでは各オブジェクトを1つのトラッククエリと表記し、トラッククエリセットは動的長である。MOTRは簡単に対処できる

表4:DanceTrack[28]データセットにおけるMOTRと既存手法の性能比較。既存の手法の結果はDanceTrack [28]による。

Methods	HOTA	AssA	DetA	MOTA	IDF1
CenterTrack [44]	41.8	22.6	**78.1**	86.8	35.7
FairMOT [43]	39.7	23.8	66.7	82.2	40.8
QDTrack [20]	45.7	29.2	72.1	83.0	44.8
TransTrack [29]	45.5	27.5	75.9	88.4	45.2
TraDes [40]	43.3	25.4	74.5	86.2	41.2
ByteTrack [42]	47.7	32.1	71.0	**89.6**	**53.9**
MOTR (ours)	**54.2**	**40.2**	73.5	79.7	51.5

表5:BDD100k[41]検証セットにおけるMOTRと既存手法の性能比較。

Methods	mMOTA	mIDF1	IDSw
Yu *et al.* [41]	25.9	**44.5**	8315
DeepBlueAI [1]	26.9	/	13366
MOTR (ours)	**32.0**	43.5	**3493**

は、分類枝のクラス番号を変更するだけで、多クラス予測問題の性能を向上させることができる。多クラスシーンにおけるMOTRの性能を検証するために、さらにBDD100kデータセットで実験を行った(表5参照)。bdd100k検証セットでの結果は、MOTRがマルチクラスシーンで良好な性能を発揮し、より少ないIDスイッチで有望な性能を達成することを示している。

4.6 アブレーション研究

MOTRコンポーネント。表6aは、異なるコンポーネントを統合した場合の影響を示している。ベースラインにコンポーネントを統合することで、全体的なパフォーマンスを徐々に向上させることができる。ほとんどのオブジェクトは入口オブジェクトとして扱われるため、オリジナルとしてオブジェクトのクエリのみを使用すると、多数のIDが発生する。トラッククエリを導入することで、ベースラインはトラッキングアソシエーションを処理することができ、IDF1を1.2から49.8に改善することができる。さらに、ベースラインにTANを追加することで、MOTAは7.8%、IDF1は13.6%改善される。学習時にCALを使用した場合、MOTAとIDF1がそれぞれ8.3%と7.1%向上する。TANとCALを組み合わせることで、時間運動の学習を強化できることを示す。
集合平均損失。ここでは、ビデオシーケンスの長さがCALのトラッキング性能に与える影響を調べた。表6bに示すように、ビデオクリップの長さが徐々に2から5に増加すると、MOTAとIDF1のメトリクスはそれぞれ8.3%と7.1%向上する。このように、マルチフレームCALはトラッキング性能を大幅に向上させることができる。複数のフレームCALは、オクルージョンシーンのようないくつかの困難なケースをネットワークが処理するのに役立つことを説明した。オクルージョンのあるシーンでは、重複ボックス、IDスイッチ、オブジェクトの欠落が大幅に減少することが確認された。これを検証するために、図5にいくつかの可視化を示す。

表6：提案するMOTRのアブレーション研究。全ての実験では、ResNet50のシングルレベルC5特徴量を用いている。

(a) 我々の貢献の効果。TrackQ：トラッククエリ。TAN：時間集約ネットワーク。CAL：集団平均損失。

トラックQ	タンカル	運動↑	IDF1↑	IDS↓
		-	1.2	33198
✓		37.1	49.8	562
✓	✓	44.9	63.4	257
✓	✓	47.5	56.1	417
✓	✓ ✓	**53.2**	**70.5**	**155**

(b) 学習中のCollective Average Lossにおけるビデオクリップの長さの増加がトラッキング性能に与える影響。

Length	MOTA↑	IDF1↑	IDS↓
2	44.9	63.4	257
3	51.6	59.4	424
4	50.6	64.0	314
5	**53.2**	**70.5**	**155**

(c) 学習中のランダムトラッククエリ消去確率p_{drop}の解析。

p_{drop}	MOTA↑	IDF1↑	IDS↓
5e-2	49.0	60.4	411
0.1	**53.2**	**70.5**	**155**
0.3	51.1	69.0	180
0.5	48.5	62.0	302

(d) 学習中のランダムな偽陽性挿入確率p_{insert}の影響。

p_{insert}	MOTA↑	IDF1↑	IDS↓
0.1	51.2	**71.7**	**148**
0.3	**53.2**	70.5	155
0.5	52.1	62.0	345
0.7	50.7	57.7	444

(e) The exploration of different τ_{ex} and τ_{en} in QIM network.

τ_{ex}	0.6	0.6	0.6	0.5	0.6	0.7
τ_{en}	0.7	0.8	0.9	0.8	0.8	0.8
MOTA↑	52.7	**53.2**	53.1	**53.5**	53.2	52.8
IDF1↑	69.8	**70.5**	70.1	**70.5**	70.5	68.3
IDS↓	181	155	**142**	**153**	155	181

(f)ランダムサンプリングターバルがトラッキング性能 in- に与える影響。

Intervals	MOTA↑	IDF1↑	IDS↓
3	53.2	64.8	218
5	50.8	62.8	324
10	**53.2**	**70.5**	**155**
12	53.1	69	158

トラッククエリの消去と挿入MOTデータセットでは、ビデオシーケンスの入口オブジェクトと出口オブジェクトの2つのケースで、学習サンプルが少ない。そこで、トラッククエリ消去と挿入を採用し、それぞれ確率p_{drop}とp_{insert}でこの2つのケースをシミュレートする。表6cは、学習時にp_{drop}の値を変えても性能が向上することを報告している。MOTRはp_{drop}を0.1に設定した場合に最高の性能を達成する。入口オブジェクトと同様に、予測値が偽陽性である前のフレームから転送されたトラッククエリが、オブジェクトの出口の場合をシミュレートするために現在のフレームに挿入される。表6dでは、異なるp_{insert}がトラッキング性能に与える影響を調べている。p_{insert}を0.1から0.7まで徐々に増加させた場合、IDF1スコアが減少している間、p_{insert}を0.3に設定すると、MOTAで最高スコアを達成した。

物体の入口と出口のしきい値。表6eは、QIMにおけるオブジェクトの入口閾値τ_{en}と出口閾値τ_{ex}の組み合わせの違いによる影響を調べたものである。物体の入口閾値τ_{en}を変化させると、τ_{en}にあまり敏感ではなく(MOTAで0.5%以内)、入口閾値0.8を使用すると比較的良い性能が得られることがわかる。

(a) (b)

図5:(a)重複ボックスと(b)IDスイッチ問題を解く際のCALの効果。上段と下段はそれぞれCALなしとCALありのトラッキング結果である。

さらに、オブジェクトの出口閾値τ_{ex}を変化させて実験を行う。閾値を0.5とした場合、0.6とした場合よりも若干性能が良いことがわかる。我々の実践では、0.6のτ_{en}はMOT17テストセットでより良い性能を示す。

サンプリング間隔。表6fでは、ランダムサンプリング間隔がトレーニング中のトラッキング性能に与える影響を評価している。サンプリング間隔が2から10に増加すると、IDSは209から155に大幅に減少する。学習中、フレームが小さな間隔でサンプリングされると、ネットワークは局所最適解に陥りやすい。サンプリング間隔を適切に増やすことで、実際のシーンをシミュレートすることができる。ランダムサンプリング間隔が10より大きい場合、トラッキングフレームワークはこのような長距離ダイナミクスを捉えることができず、トラッキング性能が相対的に悪化する。

5 Limitations

オンライントラッカーであるMOTRは、エンドツーエンドの複数オブジェクト追跡を実現する。DETRアーキテクチャとトラックレットを考慮したラベル割り当てにより、外観と位置の分散を共同で暗黙的に学習する。しかし、いくつかの欠点もある。第一に、新生物体の検出性能は満足できるものには程遠い(MOTA指標の結果は十分ではない)。上記で分析したように、追跡された物体の検出では検出クエリが抑制されるため、物体クエリの性質に反して、新生物体の検出性能が制限される可能性がある。第二に、MOTRにおけるクエリのパッシングはフレームごとに行われるため、学習時のモデル学習の効率が制限される。我々の実践では、VisTR [36]の並列復号はMOTの複雑なシナリオに対応できない。上記の2つの問題を解決することは、TransformerベースのMOTフレームワークにとって重要な研究テーマとなる。

謝辞を述べる：本研究は、中国国家重点研究開発プログラム(No.2017YFA0700800)および北京人工知能研究院(BAAI)の支援を受けた。

References

1. CodaLab Competition - CVPR 2020 BDD100K Multiple Object Tracking Challenge (Jul 2022), https://competitions.codalab.org/competitions/24910, [Online; accessed 19. Jul. 2022] 12
2. Bergmann, P., Meinhardt, T., Leal-Taixe, L.: Tracking without bells and whistles. In: ICCV (2019) 1, 3, 11
3. Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B.: Simple online and realtime tracking. In: ICIP (2016) 1, 3
4. Bochinski, E., Eiselein, V., Sikora, T.: High-speed tracking-by-detection without using image information. In: AVSS (2017) 1
5. Camgoz, N.C., Koller, O., Hadfield, S., Bowden, R.: Sign language transformers: Joint end-to-end sign language recognition and translation. In: CVPR (2020) 3
6. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: ECCV (2020) 1, 3, 4
7. Chang, X., Zhang, W., Qian, Y., Le Roux, J., Watanabe, S.: End-to-end multi-speaker speech recognition with transformer. In: ICASSP (2020) 3
8. Chu, P., Wang, J., You, Q., Ling, H., Liu, Z.: Transmot: Spatial-temporal graph transformer for multiple object tracking. arXiv preprint arXiv:2104.00194 (2021) 4
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021) 3
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016) 6, 10
11. Kuhn, H.W.: The hungarian method for the assignment problem. Naval research logistics quarterly **2**(1-2), 83–97 (1955) 3
12. Leal-Taixé, L., Canton-Ferrer, C., Schindler, K.: Learning by tracking: Siamese cnn for robust target association. In: CVPRW (2016) 3
13. Li, N., Liu, S., Liu, Y., Zhao, S., Liu, M.: Neural speech synthesis with transformer network. In: AAAI (2019) 3
14. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: ICCV (2017) 8
15. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014) 10
16. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. arXiv preprint arXiv:2103.14030 (2021) 3
17. Luiten, J., Osep, A., Dendorfer, P., Torr, P., Geiger, A., Leal-Taixé, L., Leibe, B.: Hota: A higher order metric for evaluating multi-object tracking. IJCV **129**(2), 548–578 (2021) 9
18. Meinhardt, T., Kirillov, A., Leal-Taixe, L., Feichtenhofer, C.: Trackformer: Multi-object tracking with transformers. arXiv preprint arXiv:2101.02702 (2021) 1, 3, 4, 8, 9, 10, 11

19. Milan, A., Leal-Taixé, L., Reid, I., Roth, S., Schindler, K.: Mot16: A benchmark for multi-object tracking. arXiv preprint arXiv:1603.00831 (2016) 9
20. Pang, J., Qiu, L., Li, X., Chen, H., Li, Q., Darrell, T., Yu, F.: Quasi-dense similarity learning for multiple object tracking. In: CVPR (2021) 11, 12
21. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression. In: CVPR (2019) 8
22. Schulter, S., Vernaza, P., Choi, W., Chandraker, M.: Deep network flow for multi-object tracking. In: CVPR (2017) 3
23. Shao, S., Zhao, Z., Li, B., Xiao, T., Yu, G., Zhang, X., Sun, J.: Crowdhuman: A benchmark for detecting human in a crowd. arXiv preprint arXiv:1805.00123 (2018) 10
24. Sharma, S., Ansari, J.A., Murthy, J.K., Krishna, K.M.: Beyond pixels: Leveraging geometry and shape cues for online multi-object tracking. In: ICRA (2018) 3
25. Shi, B., Bai, X., Yao, C.: An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. TPAMI **39**(11), 2298–2304 (2016) 4
26. Shuai, B., Berneshawi, A.G., Modolo, D., Tighe, J.: Multi-object tracking with siamese track-rcnn. arXiv preprint arXiv:2004.07786 (2020) 3
27. Stadler, D., Beyerer, J.: Modelling ambiguous assignments for multi-person tracking in crowds. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 133–142 (2022) 11
28. Sun, P., Cao, J., Jiang, Y., Yuan, Z., Bai, S., Kitani, K., Luo, P.: Dancetrack: Multi-object tracking in uniform appearance and diverse motion. arXiv preprint arXiv:2111.14690 (2021) 3, 9, 11, 12
29. Sun, P., Jiang, Y., Zhang, R., Xie, E., Cao, J., Hu, X., Kong, T., Yuan, Z., Wang, C., Luo, P.: Transtrack: Multiple-object tracking with transformer. arXiv preprint arXiv: 2012.15460 (2020) 1, 3, 4, 8, 9, 10, 11, 12
30. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: NeurIPS (2014) 2, 4
31. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017) 2, 3, 4
32. Wang, Q., Zheng, Y., Pan, P., Xu, Y.: Multiple object tracking with correlation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3876–3886 (2021) 11
33. Wang, S., Sheng, H., Zhang, Y., Wu, Y., Xiong, Z.: A general recurrent tracking framework without real data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13219–13228 (2021) 11
34. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: CVPR (2018) 3
35. Wang, Y., Kitani, K., Weng, X.: Joint object detection and multi-object tracking with graph neural networks. In: 2021 IEEE International Conference on Robotics and Automation (ICRA). pp. 13708–13715. IEEE (2021) 11
36. Wang, Y., Xu, Z., Wang, X., Shen, C., Cheng, B., Shen, H., Xia, H.: End-to-end video instance segmentation with transformers. In: CVPR (2021) 3, 14
37. Wang, Z., Zheng, L., Liu, Y., Li, Y., Wang, S.: Towards real-time multi-object tracking. In: ECCV (2020) 1, 3
38. Welch, G., Bishop, G., et al.: An introduction to the kalman filter (1995) 3
39. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. In: ICIP (2017) 1, 3

40. Wu, J., Cao, J., Song, L., Wang, Y., Yang, M., Yuan, J.: Track to detect and segment: An online multi-object tracker. In: CVPR (2021) 11, 12
41. Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T.: Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020) 9, 12
42. Zhang, Y., Sun, P., Jiang, Y., Yu, D., Yuan, Z., Luo, P., Liu, W., Wang, X.: Bytetrack: Multi-object tracking by associating every detection box. arXiv preprint arXiv:2110.06864 (2021) 1, 3, 10, 11, 12
43. Zhang, Y., Wang, C., Wang, X., Zeng, W., Liu, W.: Fairmot: On the fairness of detection and re-identification in multiple object tracking. IJCV pp. 1–19 (2021) 1, 3, 11, 12
44. Zhou, X., Koltun, V., Krähenbühl, P.: Tracking objects as points. In: ECCV (2020) 9, 10, 11, 12
45. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. In: ICLR (2020) 1, 3, 6, 10