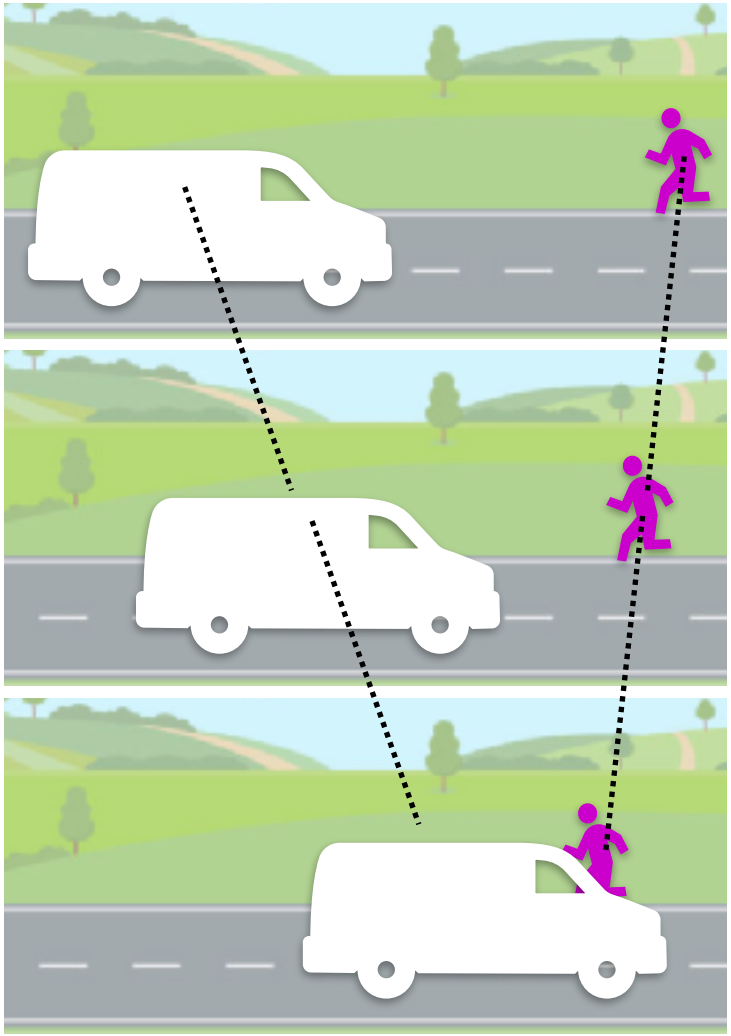
Tracking From Object Segmentation Models

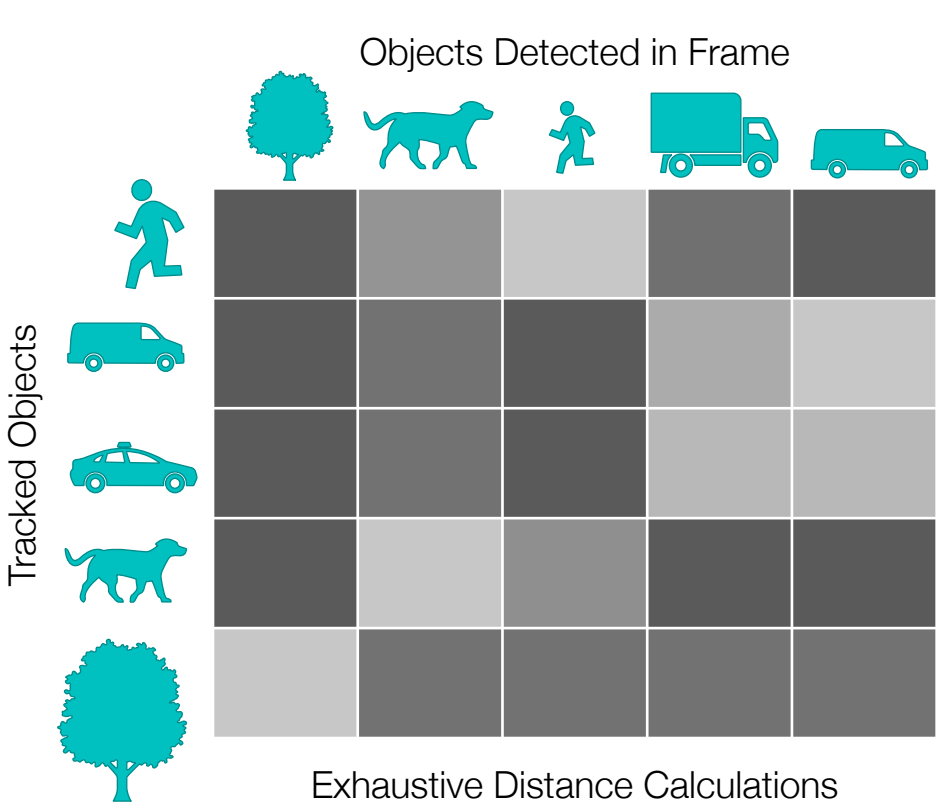
Tracking, In General

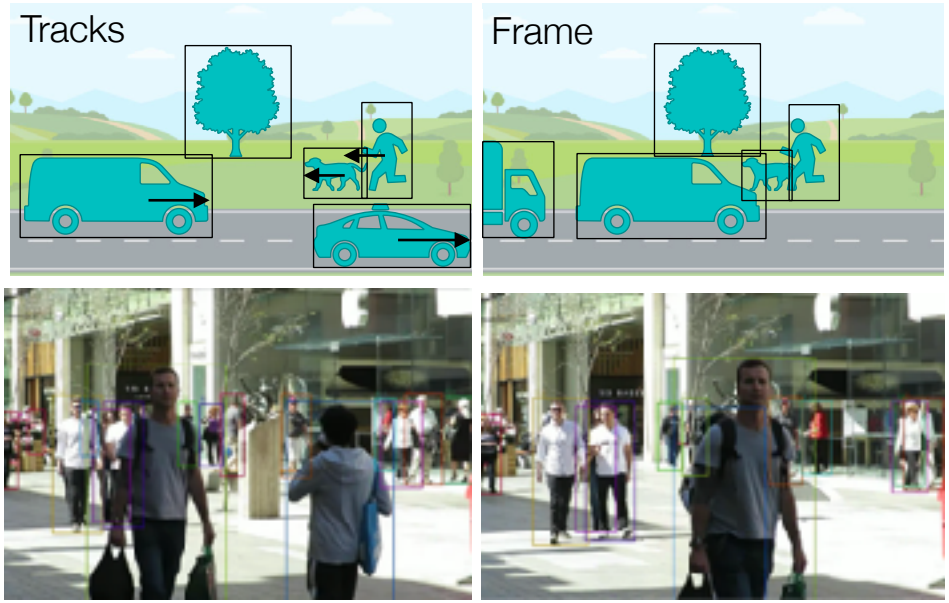
- Typically separated into two tasks
  - Identify in frame
  - Track across frame
- Hardest aspects:
  - Identifying partially visible objects
  - Tracking same object across occlusion
  - Recognizing objects that leave frame and return
- Traditional Approaches:
  - Classifier followed by Hungarian assignment and Kalman filtering (through occlusion)



Minimum Assignment

- Minimum Assignment problem: Match detections from to current frame that are “closest” to one another.



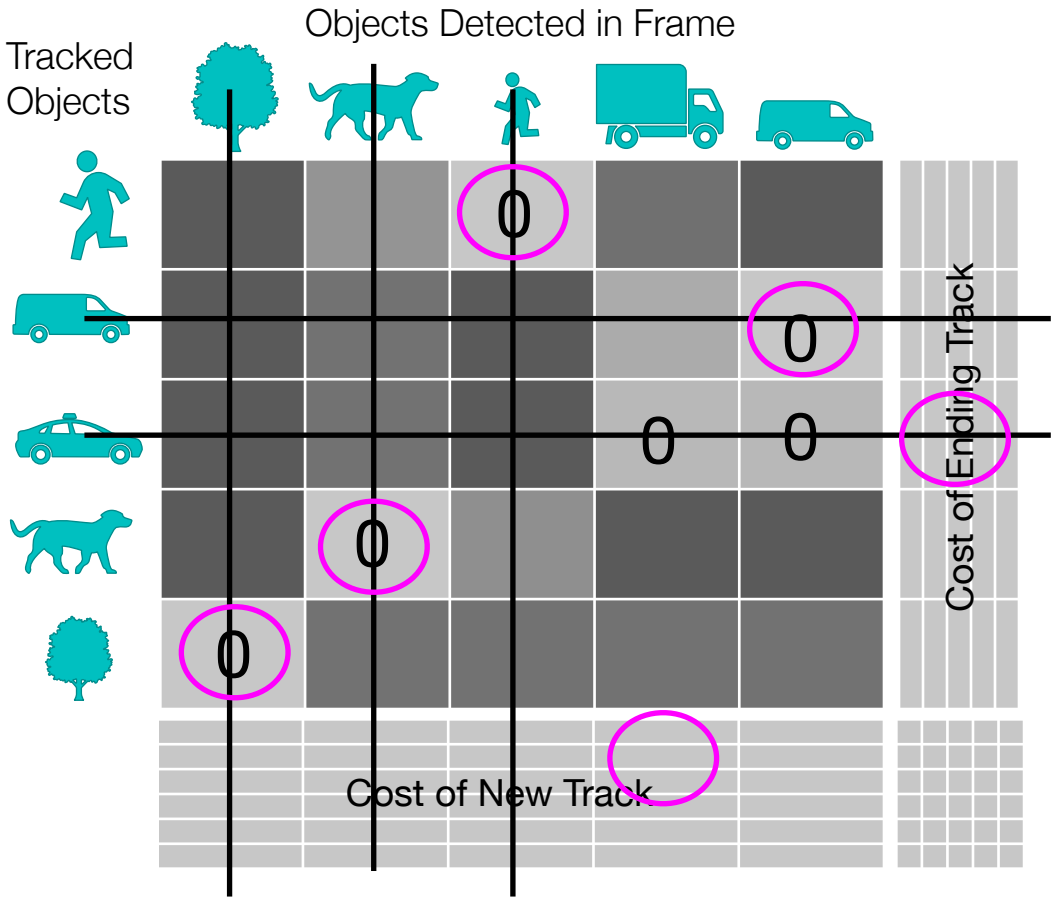


Common Distance Metrics:

- Difference in centers
- IoU of objects
- IoU of “Future Position”
- Appearance

Hungarian Algorithm

Brute force matching is $O(n!)$, looking at every possible assignment and selecting the minimum.

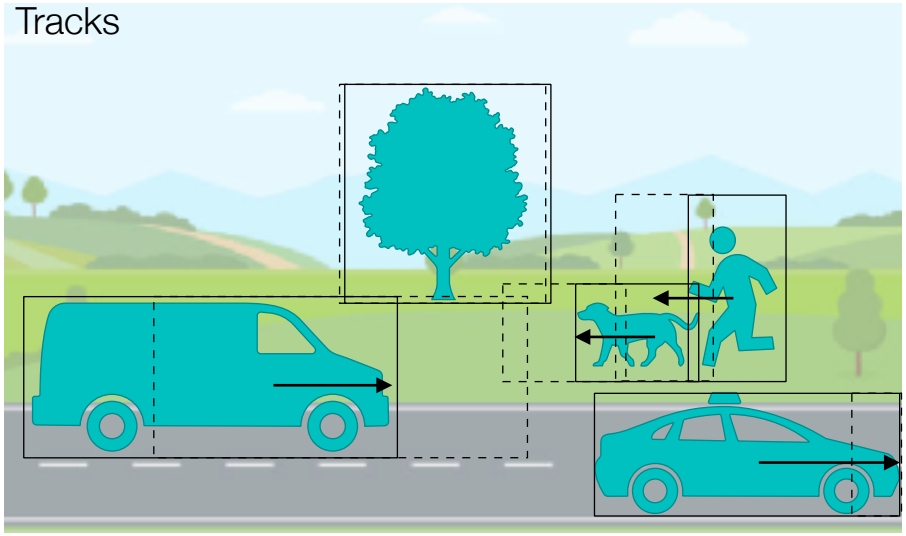


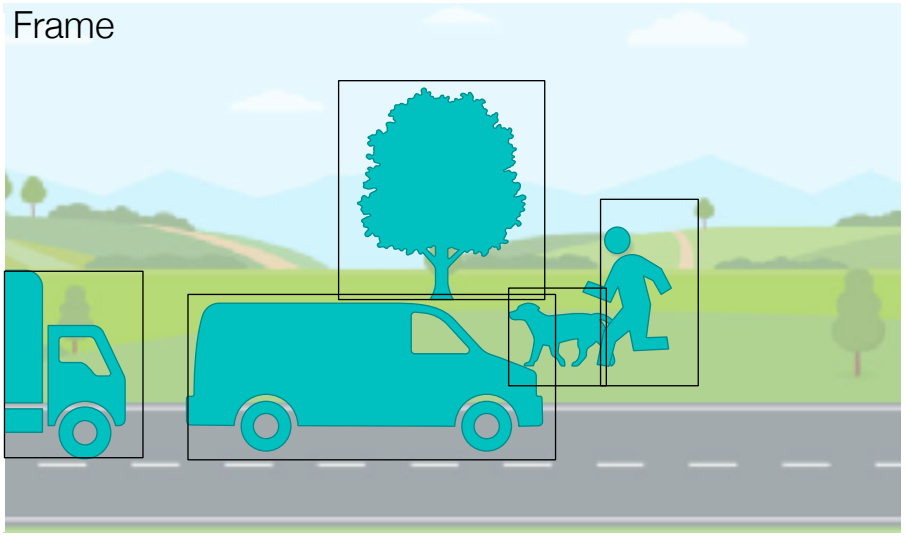
Steps in Hungarian Algorithm:

- Row reduction (subtract min)
- Column reduction (subtract min)
- Zero covering (repeated, as needed)
  - Find minimum lines to cover all zeros.
  - If number of lines does not equal number of tracks, perform additional reductions
- Assignment (one zero per line), including matching with New/End

Worst case becomes $O(n^3)$ and gives the optimal assignment guaranteed!

Kalman Filtering

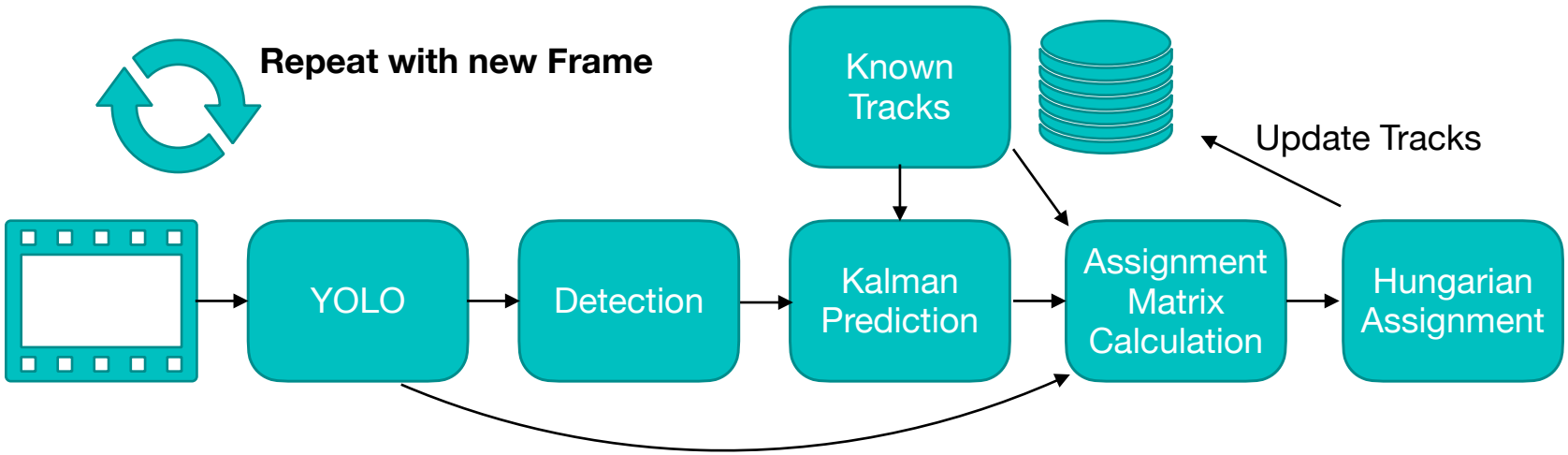


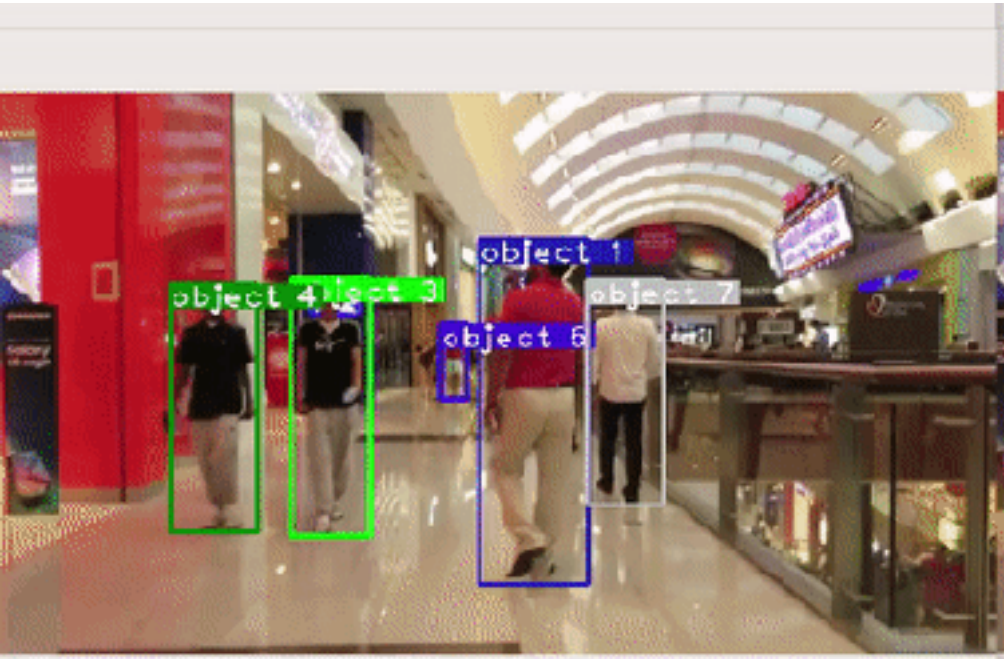


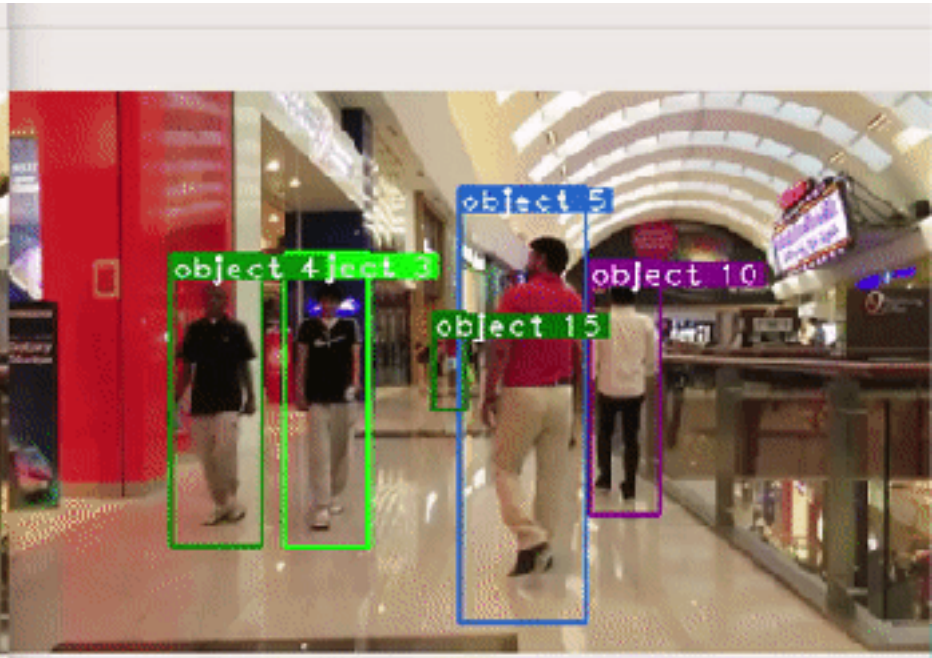
- For each object with multiple frames:
  - Estimate the trajectory of the bounding box center based on estimate of state space for each object
  - Kalman Filtering is an operation that computes the filter coefficients in the state space to minimize estimate assuming linear operations

Tracking with YOLO, Deep-SORT

Simple Online Realtime Tracking







Tracking with Transformers: Trackformer

Replace frame to frame assignment and tracking with Transformer. **Encoder** processes the CNN patches and **decoder** takes output and known tracks for assignment. Output: track IDs.

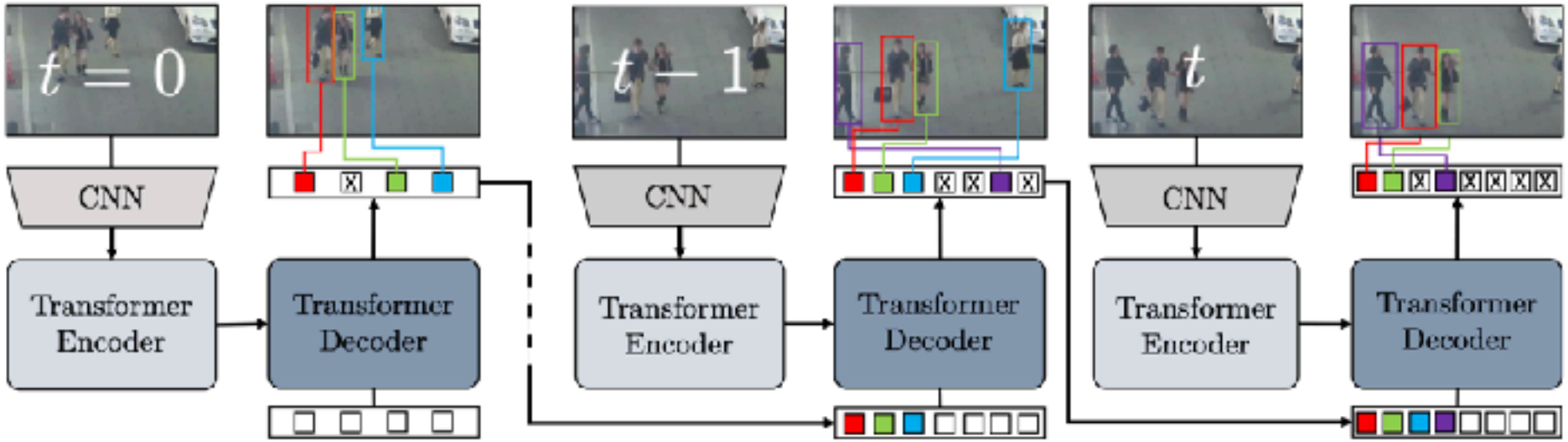
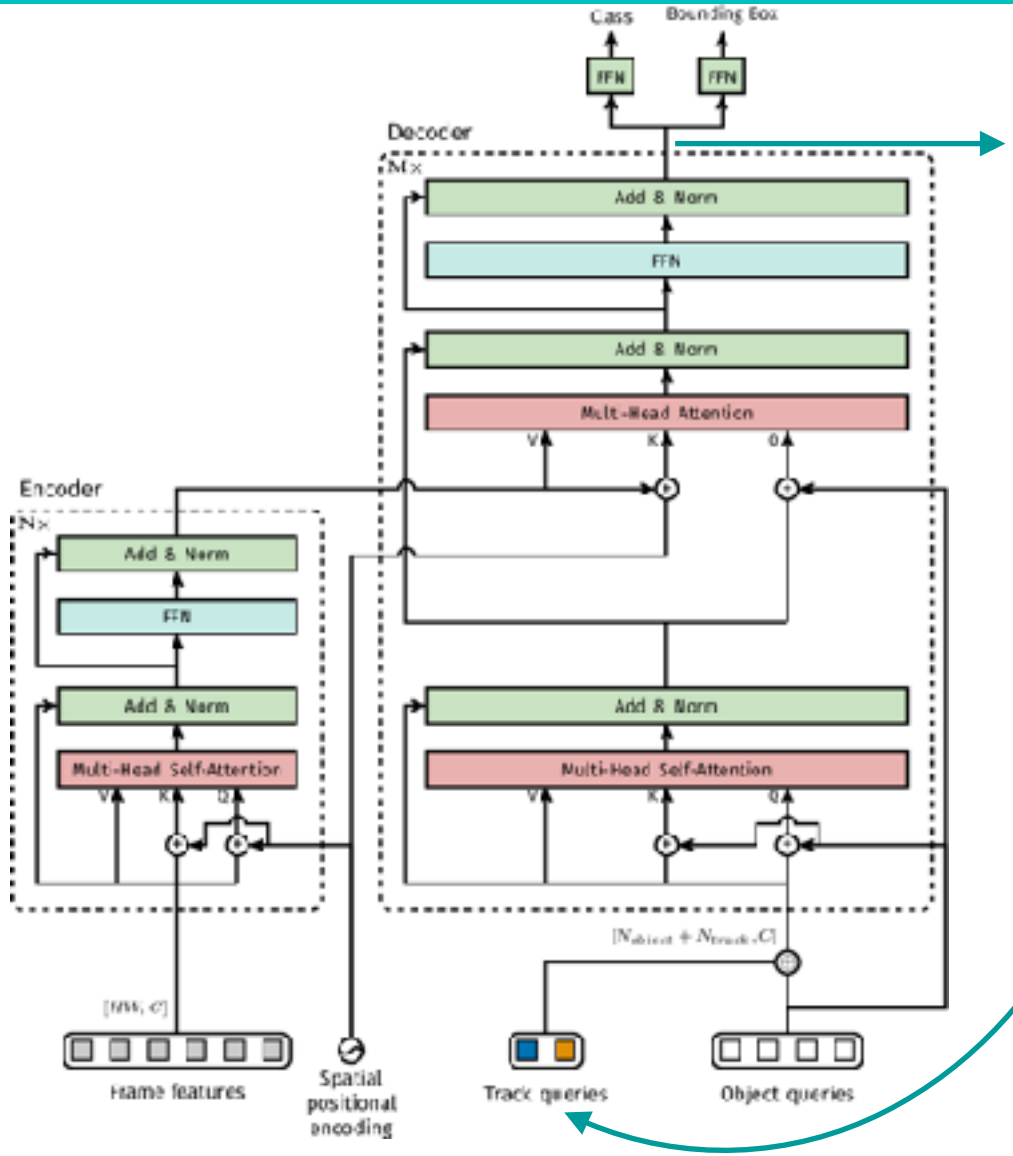


Figure 2. **TrackFormer** casts multi-object tracking as a set prediction problem performing joint detection and **tracking-by-attention**. The architecture consists of a CNN for image feature extraction, a Transformer [50] encoder for image feature encoding and a Transformer decoder which applies self- and encoder-decoder attention to produce output embeddings with bounding box and class information. At frame $t = 0$, the decoder transforms N_{object} object queries (white) to output embeddings either initializing new autoregressive **track queries** or predicting the background class (crossed). On subsequent frames, the decoder processes the joint set of $N_{\text{object}} + N_{\text{track}}$ queries to follow or remove (blue) existing tracks as well as initialize new tracks (purple).

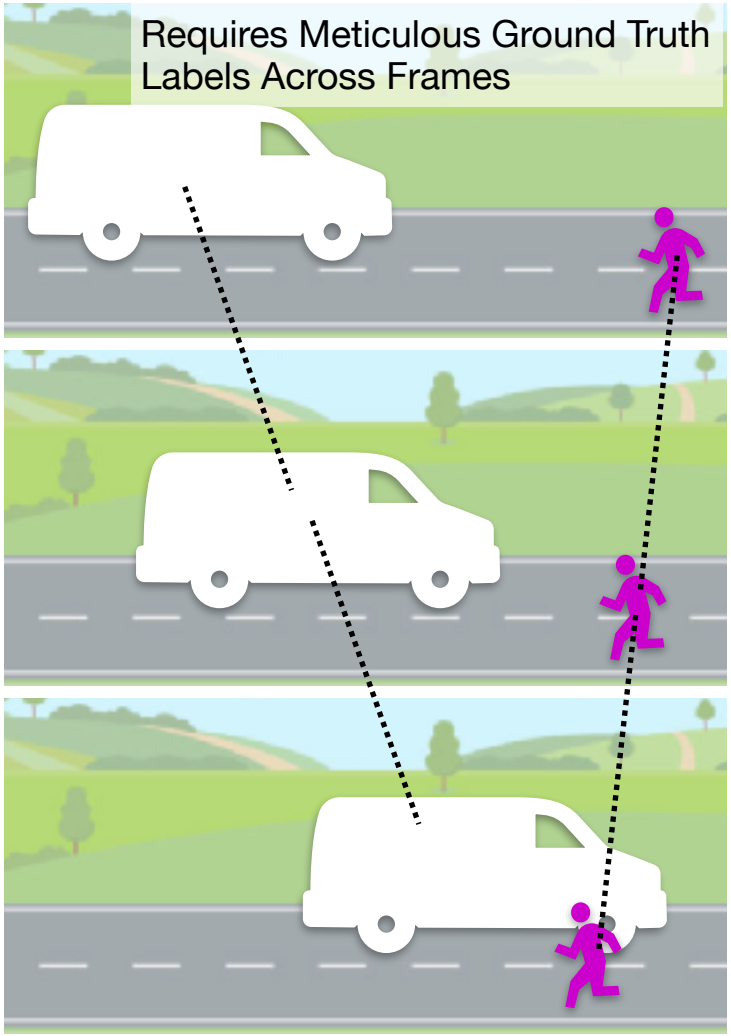
Meinhardt, Tim, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. "Trackformer: Multi-object tracking with transformers." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8844-8854. 2022.

Trackformer, more Details



Per object features
Saved, become features of track

Measures of Performance



Evaluation Measures

Lower is better. Higher is better.

Measure	Better	Perfect	Description
Avg Rank	lower	1	This is the rank of each tracker averaged over all present evaluation measures.
MOTA	higher	100 %	Multiple Object Tracking Accuracy [1]. This measure combines three error sources: false positives, missed targets and identity switches.
MOTP	higher	100 %	Multiple Object Tracking Precision [1]. The misalignment between the annotated and the predicted bounding boxes.
IDF1	higher	100 %	ID F1 Score [2]. The ratio of correctly identified detections over the average number of ground-truth and computed detections.
FAF	lower	0	The average number of false alarms per frame.
MT	higher	100 %	Mostly tracked targets. The ratio of ground-truth trajectories that are covered by a track hypothesis for at least 80% of their respective life span.
ML	lower	0 %	Mostly lost targets. The ratio of ground-truth trajectories that are covered by a track hypothesis for at most 20% of their respective life span.
FP	lower	0	The total number of false positives.
FN	lower	0	The total number of false negatives (missed targets).
ID Sw.	lower	0	The total number of identity switches. Please note that we follow the stricter definition of identity switches as described in [3].
Frag	lower	0	The total number of times a trajectory is fragmented (i.e. interrupted during tracking).
Hz	higher	Inf.	Processing speed (in frames per second excluding the detector) on the benchmark.

https://learnopencv.com/understanding-multiple-object-tracking-using-deepsort/

Evaluating Tracking with Transformer

Domain Mismatch

Source Movement

Distance

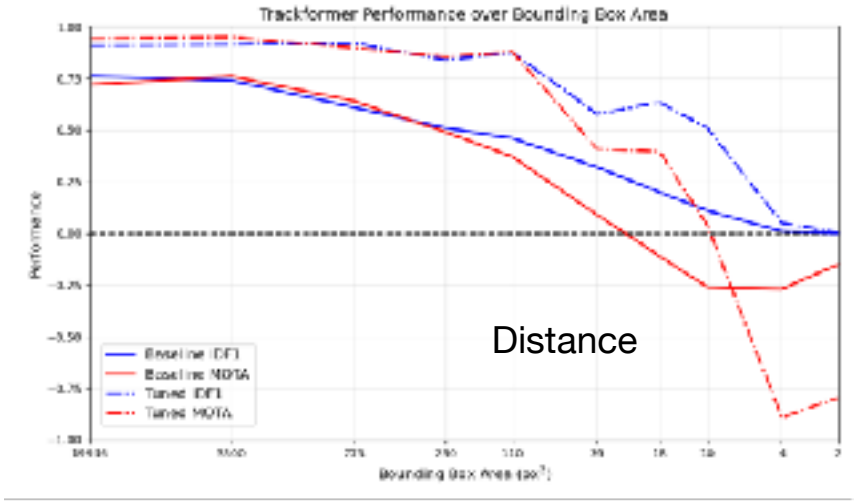
25 meters 300 meters 750 meters

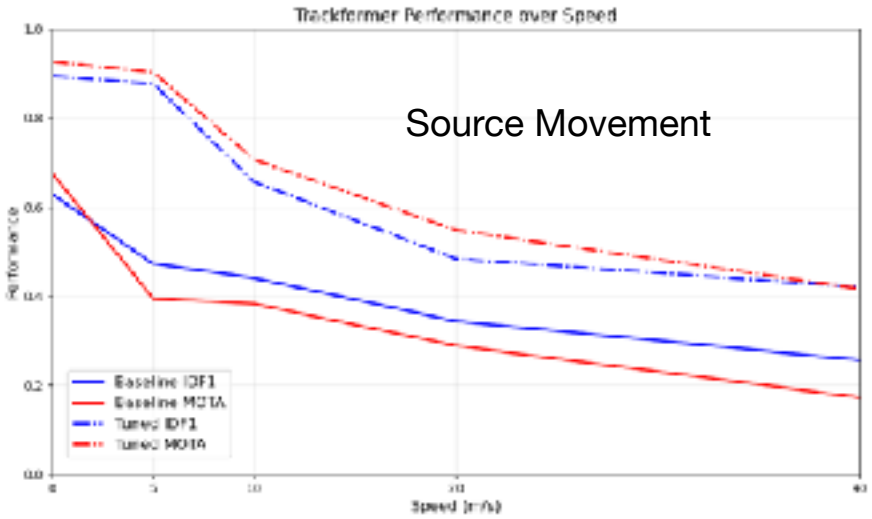
Lighting

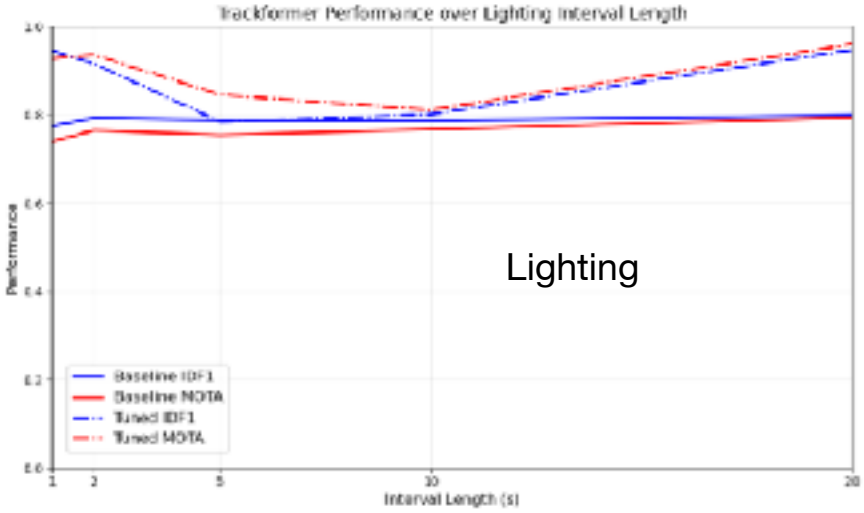
Impacts of Synthetically Generated Data on Trackformer-based Multi-Object Tracking

Matthew Lee, Clayton Harper, William Flinchbaugh, Eric C. Larson, and Mitchell A. Thornton

Evaluating Tracking with Transformer







Impacts of Synthetically Generated Data on Trackformer-based Multi-Object Tracking

Matthew Lee, Clayton Harper, William Flinchbaugh,
Eric C. Larson, and Mitchell A. Thornton

Lecture Notes for

Neural Networks and Machine Learning

FCN Learning: Detection

Next Time:
Instance Segmentation

Reading: None