TrackFormer

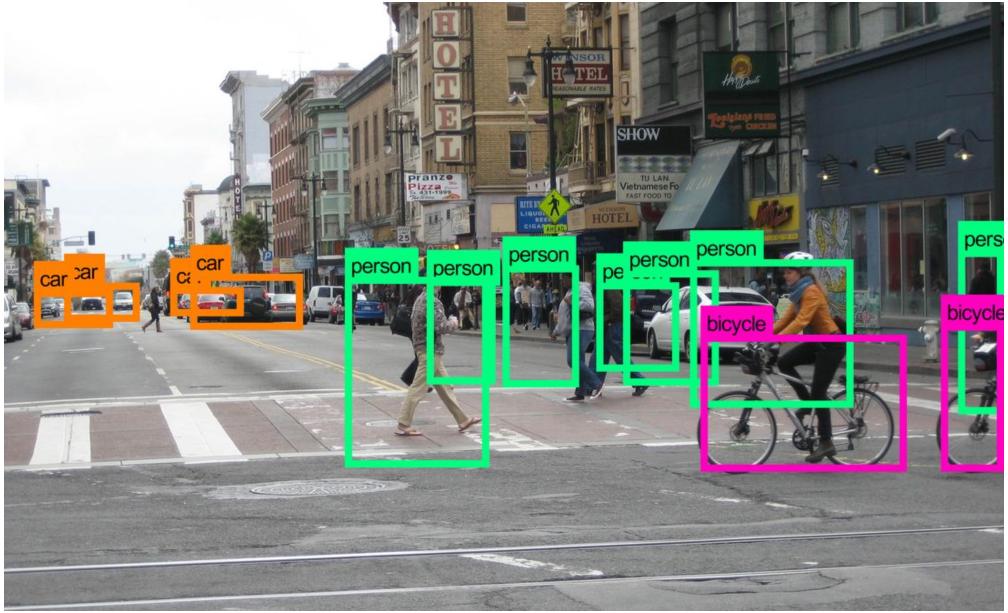
Multi-Object Tracking with Transformers (CVPR '22)

Presenters: Kaan Nyman & Sizhe Liu

Authors: T. Meinhardt, A. Kirillov, L. Leal-Taixe, & C. Feichtenhofer

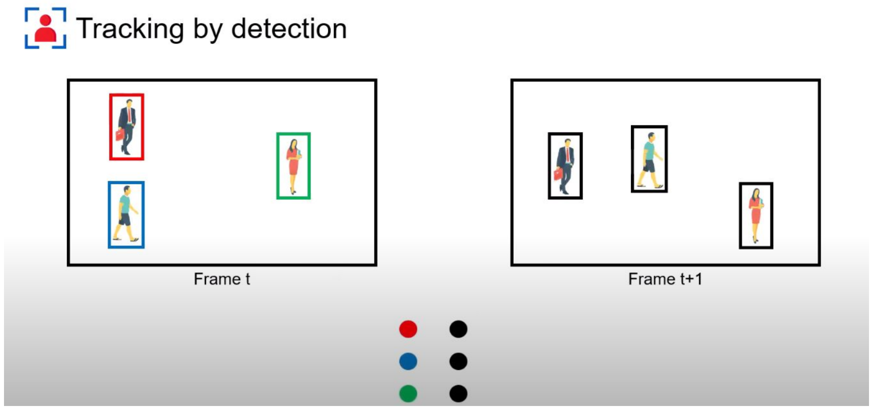
Motivation

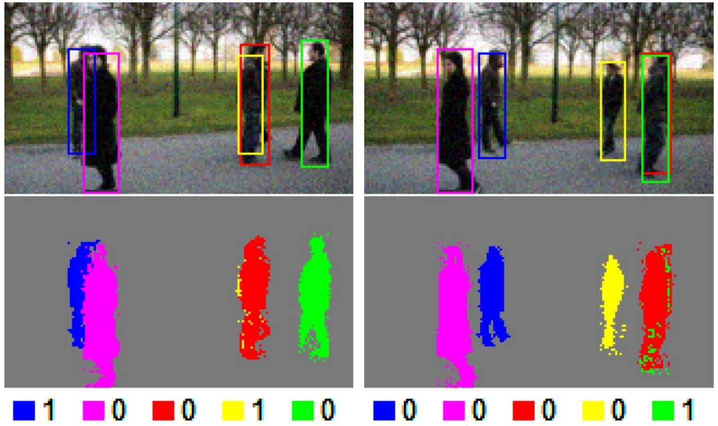
- Challenges in multi-object tracking (MOT)

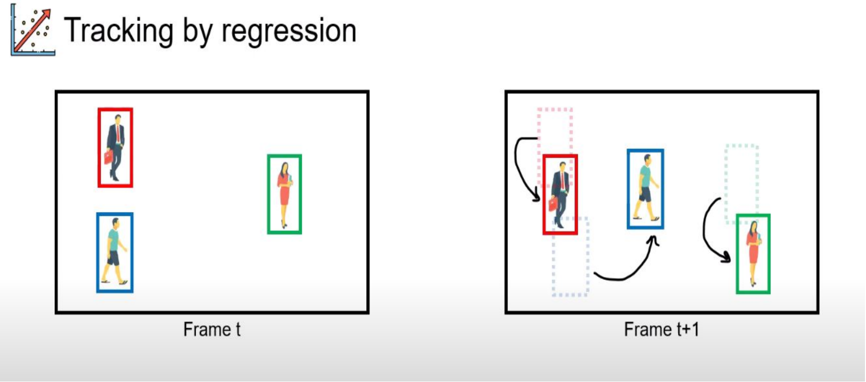


Related Work

- Tracking-by-detection
- Tracking-by-regression
- Tracking-by-segmentation

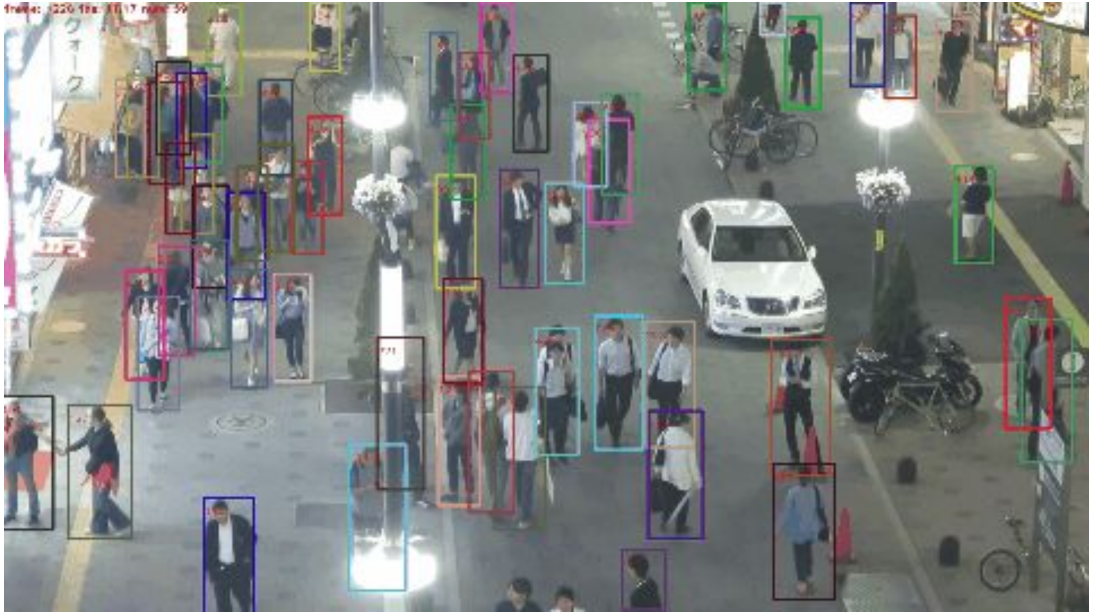






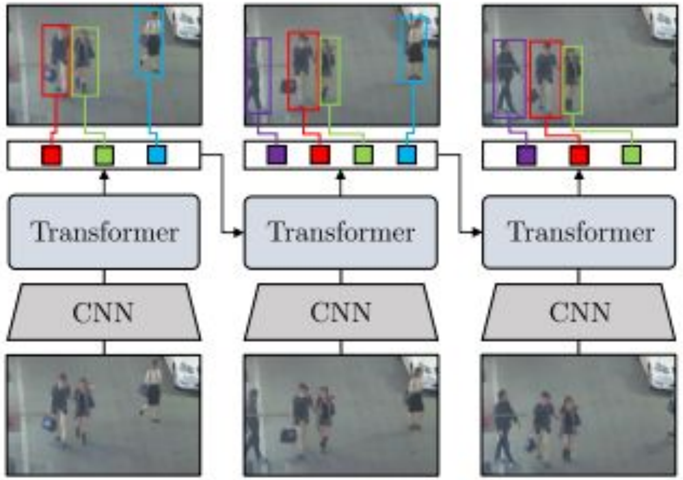
Introduction To TrackFormer

- A new approach using Transformers



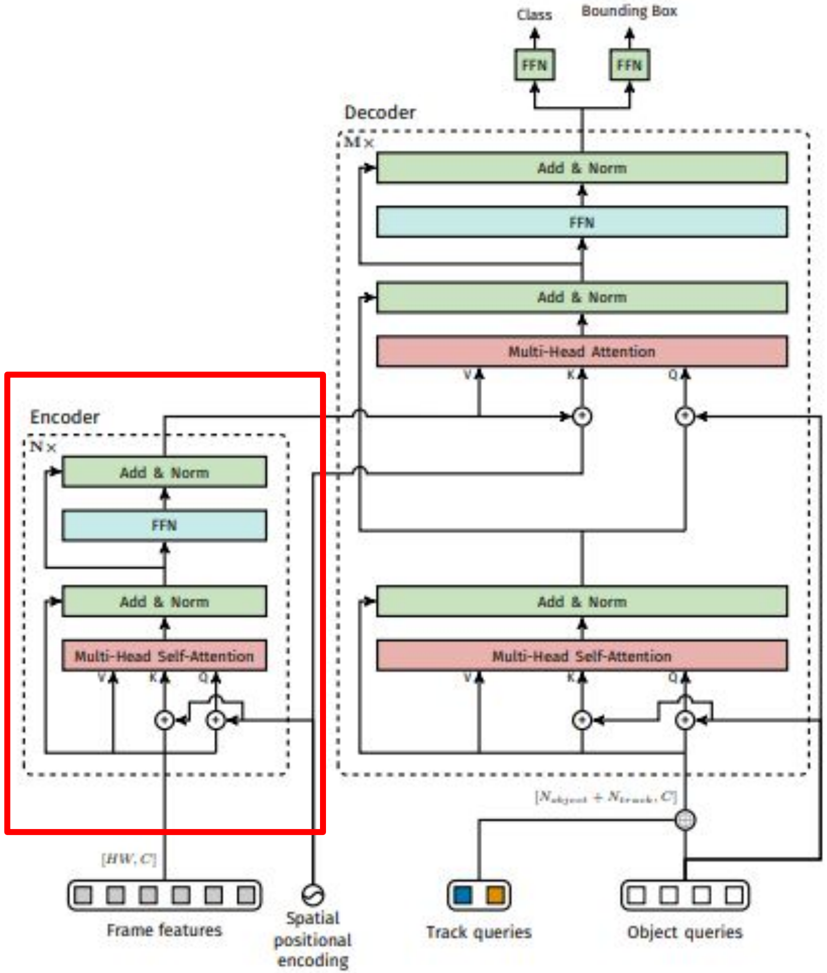
Track Queries and Object Queries

- **Combining CNN and Transformer**: Illustrates the fusion of Convolutional Neural Network (CNN) features with Transformer architecture for enhanced object detection and tracking.
- **Frame-by-Frame Processing**: Each frame is processed sequentially, with CNN extracting features and the Transformer managing object and track queries.
- **Track Queries**: Represent the continuity of an object's trajectory over time, updated by the Transformer with each new frame.
- **Object Queries**: Utilized for detecting and classifying objects within a single frame, providing real-time information for the track queries to associate.



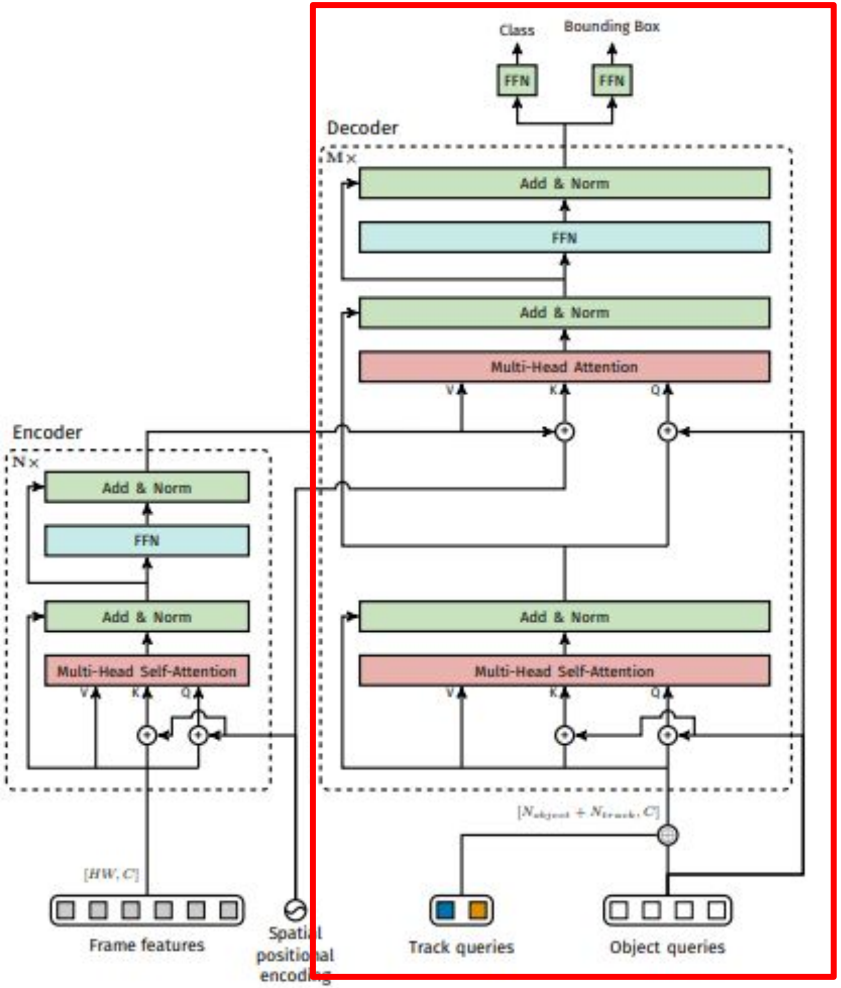
TrackFormer Architecture

- **Encoder**: Processes video frames to extract feature representations, capturing the essential details of each object and the scene context.



TrackFormer Architecture

- **Decoder**: Utilizes these features along with track queries to predict the trajectories of objects. Track queries, representing the identities of objects across frames, are updated autoregressively, ensuring accurate tracking over time.

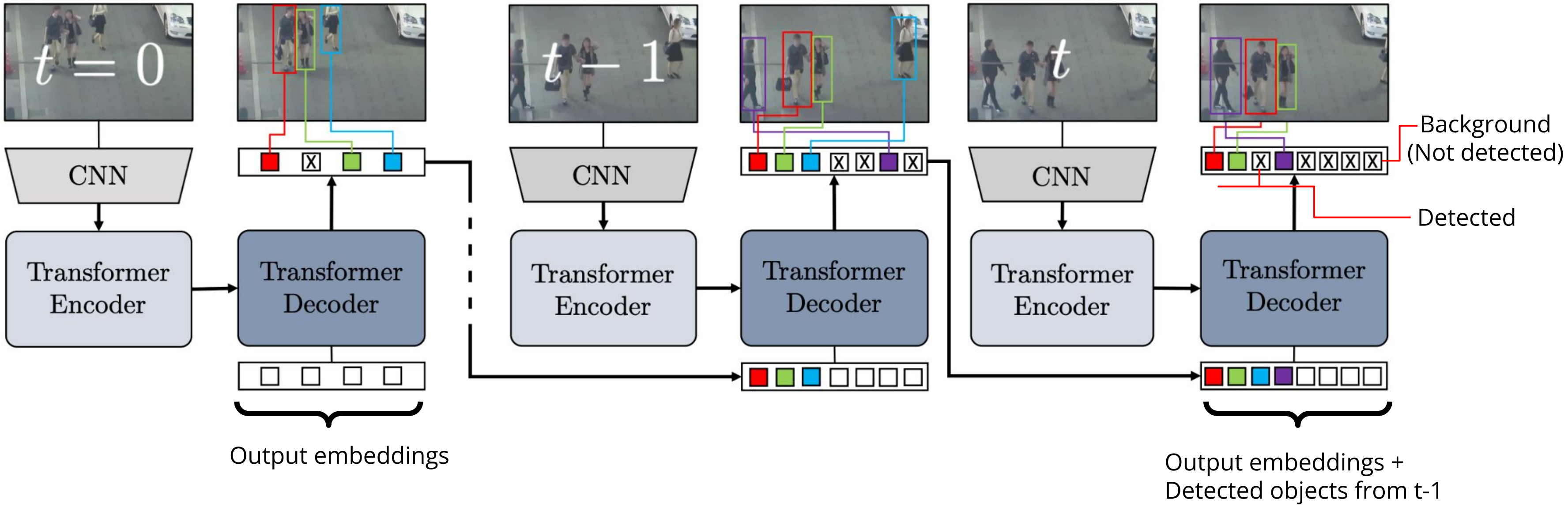


Training and Implementation

- The unified framework provided by TrackFormer simplifies the MOT pipeline, which traditionally consists of two distinct phases: object detection and data association for tracking.
- By learning these tasks jointly, TrackFormer eliminates the need for separate optimization stages and complex post-processing, which are common in conventional approaches.
- **End-to-End Trainability.**

Training and Implementation - Loss Calculation

TrackFormer train two adjacent frames together and optimize the MOT objective at once.

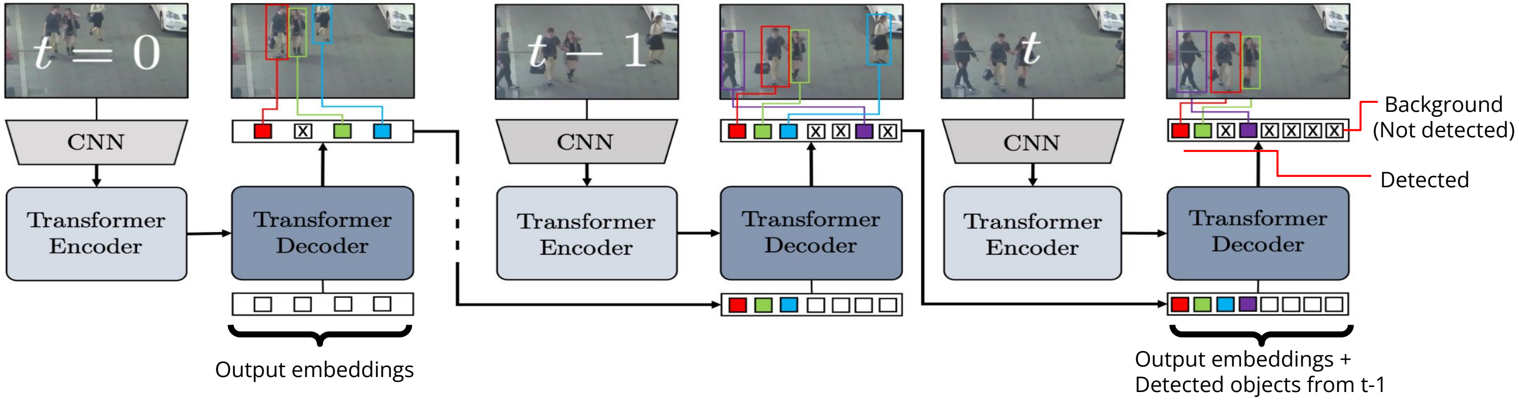


Training and Implementation - Loss Calculation

The tracking outcome of frame t is dependent on successfully detected objects from t-1

As such, the loss for frames is computed in two steps (detailed in paper) against ground truth:

- Object detection on frame t-1 of output embeddings
- Tracking of all query objects (output embeddings + detected objects from frame t-1)



Training and Implementation - Track Augmentation

To improve the joint learning pipeline, augmentations are used during training.

- Sample frame t-1 from a range of frames: simulating camera motion
- Add false negative (remove some track queries): keep false negative high as to detect new objects better.
- Add false positives: better handling of occlusion or removal.

Metrics for MOT(S)

- Multiple Object Tracking Accuracy (MOTA)
  - Measures object coverage (i.e. covers all object?)

$$MOTA = 1 - \sum^{N_{frame}} \frac{(FN + FP + Mismatch)}{n}$$

- Identity F1 Score (IDF1)
  - Measures identity preservation (i.e. same object?)

$$\text{F1 Score} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}}$$

- MOTSA
  - Similar to MOTA, but with IoU definition of tp

Performance and Benchmarks - Public Detection

Result: TrackFormer can achieve **competitive accuracy** in tracking. (As result was evaluated independent of the detection)

***Public detection**: using detection model architectures provided by MOT challenge

	Method	Data	FPS ↑	MOTA ↑	IDF1 ↑	MT ↑	ML ↓	FP ↓	FN ↓	ID Sw. ↓
	MOT17 [30] - Public									
Offline	jCC [22]	–	–	51.2	54.5	493	872	25937	247822	1802
Offline	FWT [19]	–	–	51.3	47.6	505	830	24101	247921	2648
Offline	eHAF [46]	–	–	51.8	54.7	551	893	33212	236772	1834
Offline	TT [65]	–	–	54.9	63.1	575	897	20236	233295	1088
Offline	MPNTrack [6]	M+C	–	58.8	61.7	679	788	17413	213594	1185
Offline	Lif_T [20]	M+C	–	60.5	65.6	637	791	14966	206619	1189
Online	FAMNet [10]	–	–	52.0	48.7	450	787	14138	253616	3072
Online	Tracktor++ [4]	M+C	1.3	56.3	55.1	498	831	**8866**	235449	1987
Online	GSM [29]	M+C	–	56.4	57.8	523	813	14379	230174	**1485**
Online	CenterTrack [69]	–	17.7	60.5	55.7	580	777	11599	208577	2540
Online	TMOH [47]	–	–	62.1	**62.8**	633	739	10951	201195	**1897**
Online	**TrackFormer**	–	7.4	**62.3**	57.6	**688**	**638**	16591	**192123**	4018

Performance and Benchmarks - Private Detection

Result: With private detection model, TrackFormer can achieve **higher accuracy** MOT compared against models trained only on CH (CrowdHuman).

It also achieves **competitive performance** compared against architectures with additional training dataset.

	Method	Data	FPS ↑	MOTA ↑	IDF1 ↑	MT ↑	ML ↓	FP ↓	FN ↓	ID Sw. ↓
	MOT17 [30] - Private									
Online	TubeTK [32]	JTA	–	63.0	58.6	735	468	27060	177483	4137
	GSDT [55]	6M	–	73.2	66.5	981	411	26397	120666	3891
	FairMOT [66]	CH+6M	–	73.7	72.3	1017	408	27507	117477	3303
	PermaTrack [50]	CH+PD	–	73.8	68.9	1032	405	28998	115104	3699
	GRTU [54]	CH+6M	–	75.5	76.9	1158	495	27813	108690	1572
	TLR [53]	CH+6M	–	76.5	73.6	1122	300	29808	99510	3369
	CTracker [36]	–	–	66.6	57.4	759	570	22284	160491	5529
	CenterTrack [69]	CH	17.7	67.8	64.7	816	579	**18498**	160332	3039
	QuasiDense [33]	–	–	68.7	66.3	957	516	26589	146643	3378
	TraDeS [57]	CH	–	69.1	63.9	858	507	20892	150060	3555
	TrackFormer	CH	7.4	**74.1**	**68.0**	**1113**	**246**	34602	**108777**	**2829**

Performance and Benchmarks - MOTS

Result: For tracking + segmentation, TrackFormer is also **on par** with SOTA approaches.

Method	TbD	sMOTSA ↑	IDF1 ↑	FP ↓	FN ↓	ID Sw. ↓
Train set (4-fold cross-validation)						
MHT_DAM [23]	×	48.0	–	–	–	–
FWT [19]	×	49.3	–	–	–	–
MOTDT [8]	×	47.8	–	–	–	–
jCC [22]	×	48.3	–	–	–	–
TrackRCNN [52]		52.7	–	–	–	–
MOTSNet [38]		56.8	–	–	–	–
PointTrack [58]		58.1	–	–	–	–
TrackFormer		**58.7**	–	–	–	–
Test set						
Track R-CNN [52]		40.6	42.4	**1261**	12641	567
TrackFormer		**54.9**	**63.6**	2233	**7195**	**278**

Performance and Benchmarks - MOTS

Result: Hand-picked comparison between TrackFormer and R-CNN. It shows TrackFormer has clear superiority over R-CNN.

Ablation Study

Pretraining significantly improved accuracy.

The track augmentation made a particularly huge difference.

Method	MOTA ↑	Δ	IDF1 ↑	Δ
TrackFormer	71.3		73.4	
w\o				
Pretraining on CrowdHuman	69.3	-2.0	71.8	-1.6
Track query re-identification	69.2	-0.1	70.4	-1.4
Track augmentations (FP)	68.4	-0.8	70.0	-0.4
Track augmentations (Range)	64.0	-4.4	59.2	-10.8
Track queries	61.0	-3.0	45.1	-14.1

Conclusion

- TrackFormer represents a novel approach to multi-object tracking
    - Joint detection+tracking
    - Transformer-based architecture
- Its success on challenging benchmarks
    - Track augmentations
- Questions?