清華大學 Tsinghua University

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

CXTRACK: IMPROVING 3D POINT CLOUD TRACKING WITH CONTEXTUAL INFORMATION

TIAN-XING XU[1], YUAN-CHEN GUO[1], YU-KUN LAI[2] AND SONG-HAI ZHANG[1]

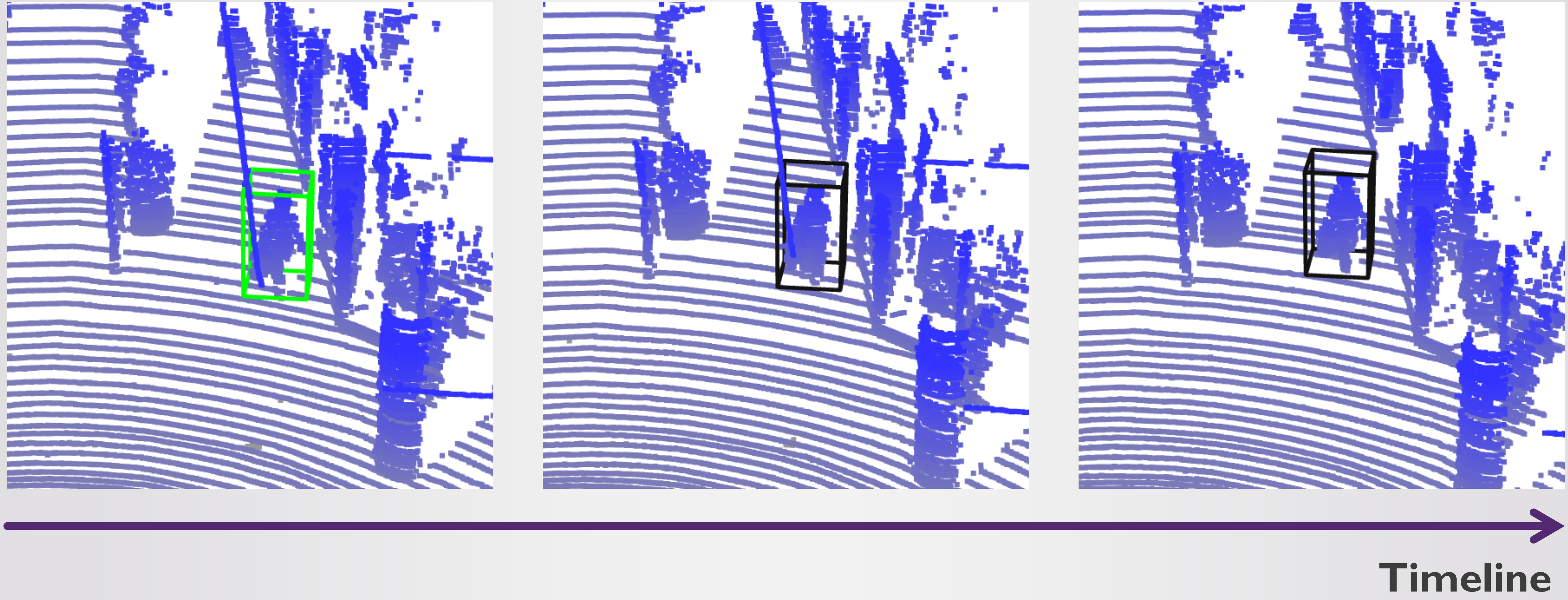
[1]TSINGHUA UNIVERSITY, [2]CARDIFF UNIVERSITY

PAPER ID: 3021

TAG: TUE-AM-103

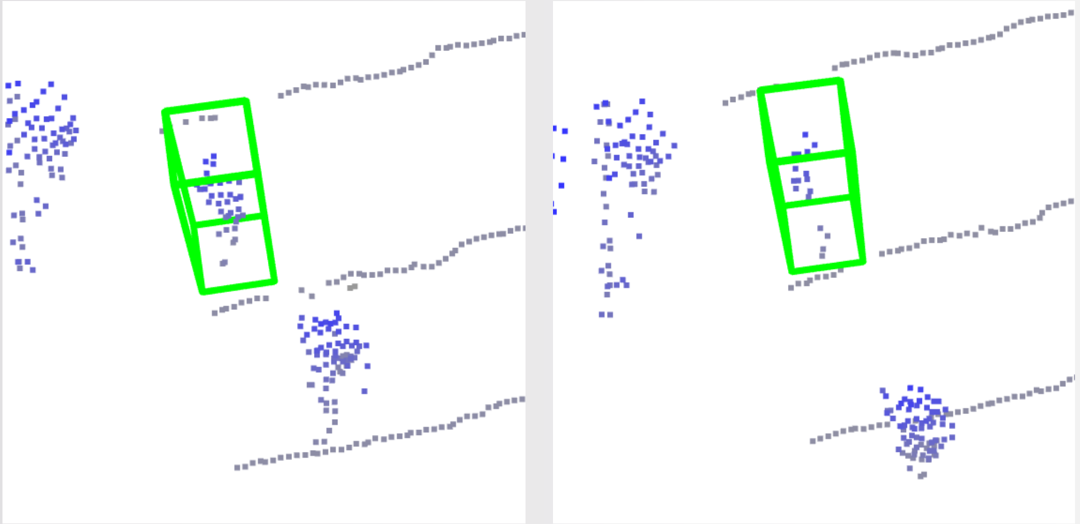
JUNE 20, 2023

3D SINGLE OBJECT TRACKING

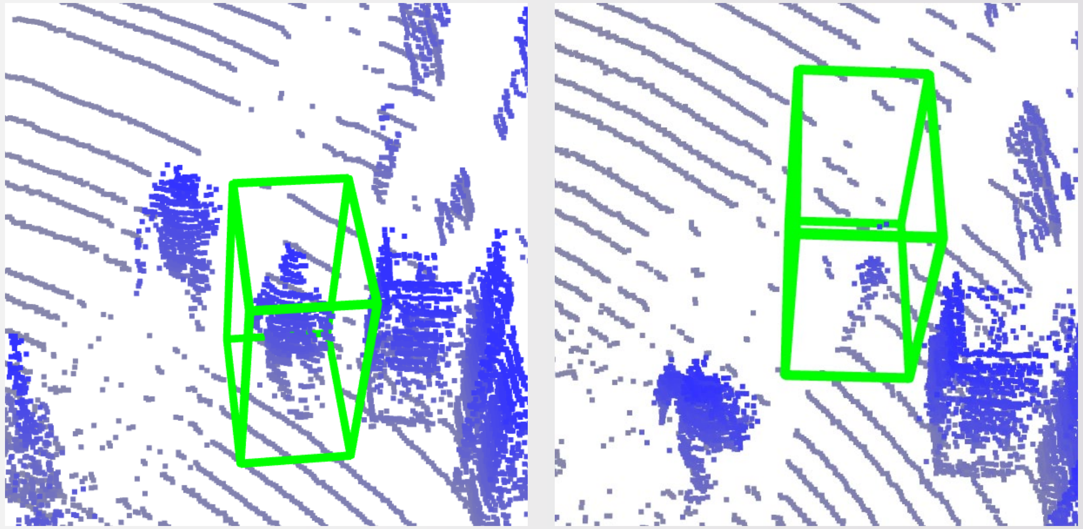


CHALLENGE

Appearance variation

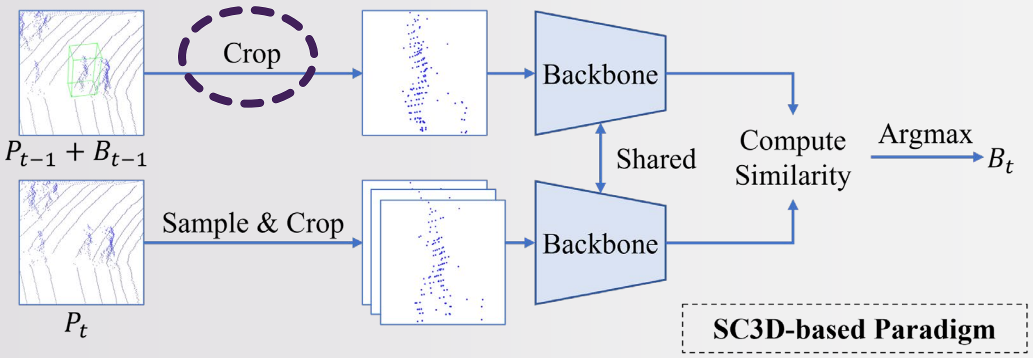


Distractor

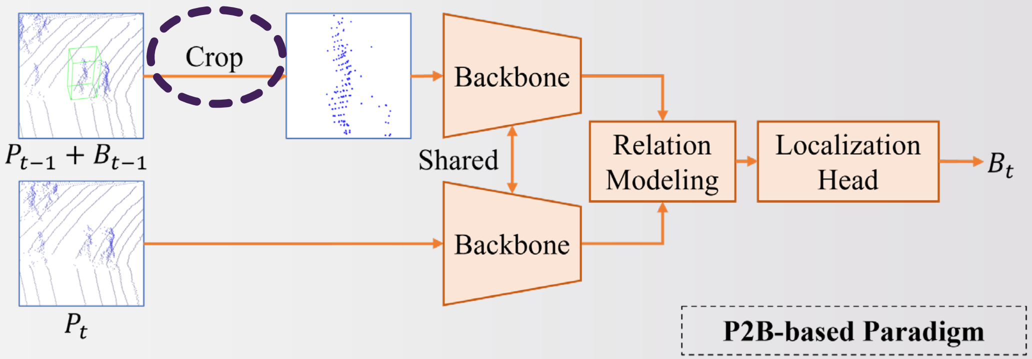


Contextual information across frames is crucial for single object tracking!

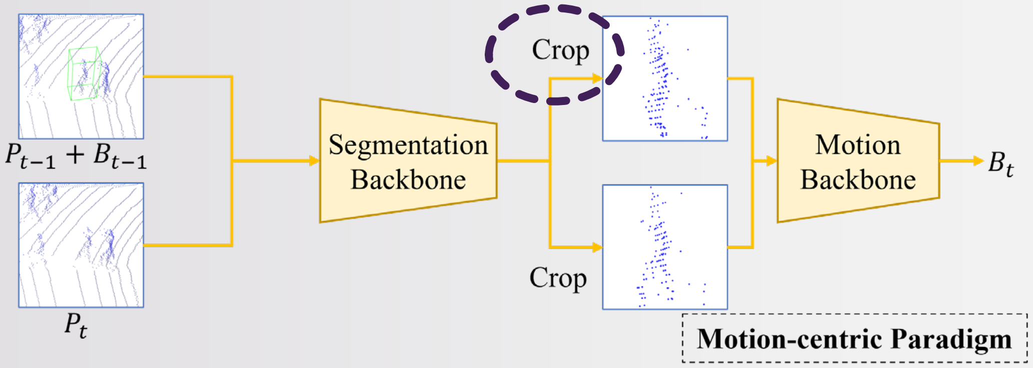
IMPROVING SOT WITH CONTEXTUAL INFORMATION



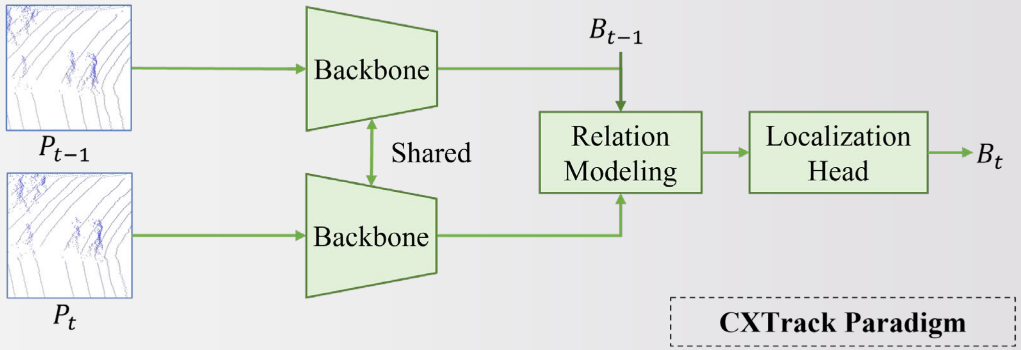
SC3D-based Paradigm



P2B-based Paradigm



Motion-centric Paradigm



CXTrack Paradigm

MORE ACCURATE LOCALIZATION

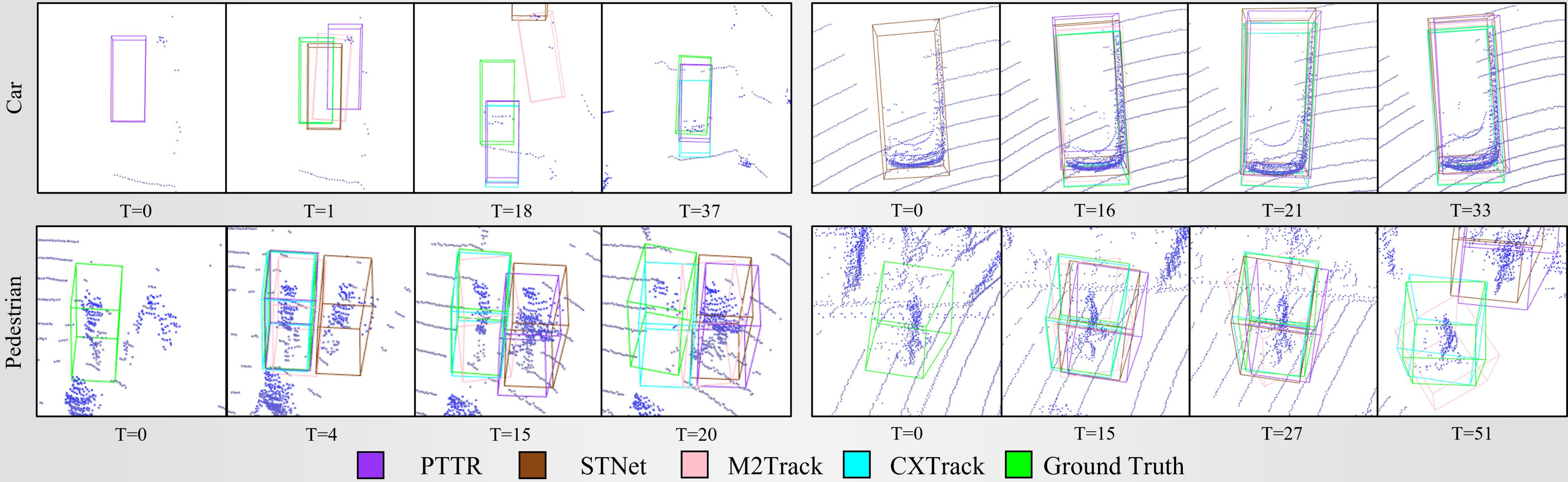
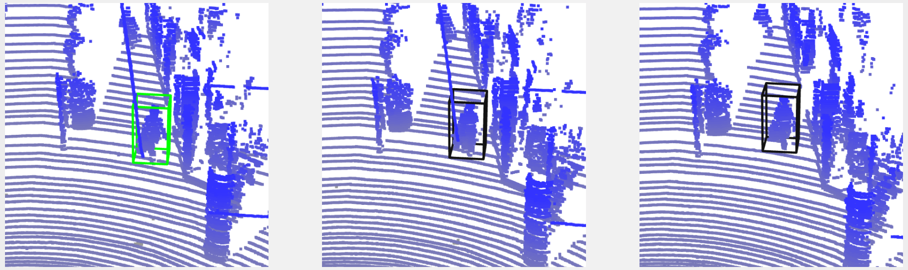


TABLE OF CONTENT

- Problem Definition
- CXTrack
- Experimental Results
- Limitations & Future work

PROBLEM DEFINITION



$$F(P_{t-1}, B_{t-1}, P_t) \rightarrow B_t$$

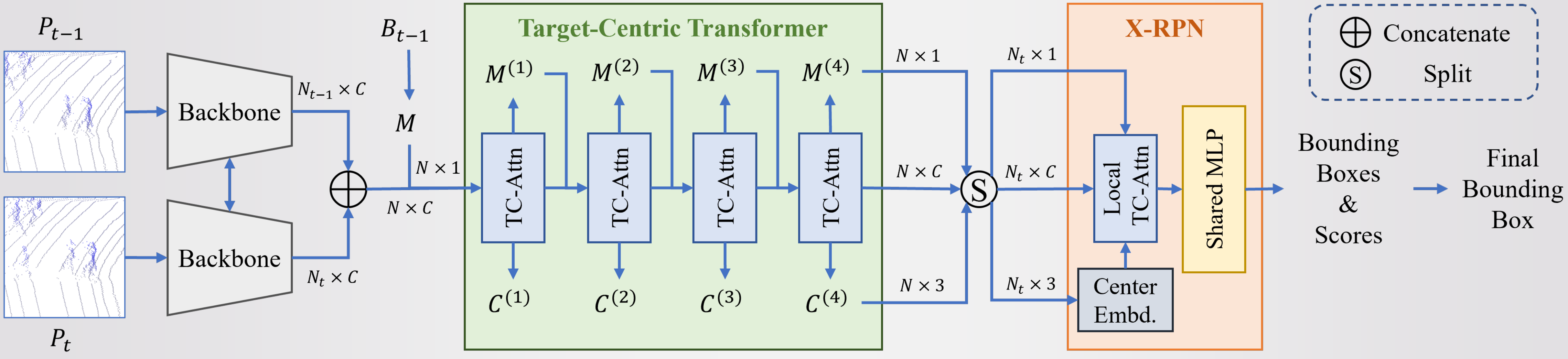
Previous Paradigm

- $F(T_{t-1}, P_t) \rightarrow (\Delta x, \Delta y, \Delta z, \Delta\theta)$
  - B_t and P_t are the bounding box and point cloud at time t, respectively
  - $\Delta x, \Delta y, \Delta z, \Delta\theta$ are the offset vectors between B_{t-1} and B_t
  - T_{t-1} is the template point cloud of the target cropped from P_{t-1} using B_{t-1}

Ours

- $F(P_{t-1}, M_{t-1}, P_t) \rightarrow (\Delta x, \Delta y, \Delta z, \Delta\theta)$
  - B_t and P_t are the bounding box and point cloud at time t, respectively
  - $\Delta x, \Delta y, \Delta z, \Delta\theta$ are the offset vectors between B_{t-1} and B_t
  - B_{t-1} is encoded into the point-wise mask M_{t-1} to indicate the tracking target

CXTRACK

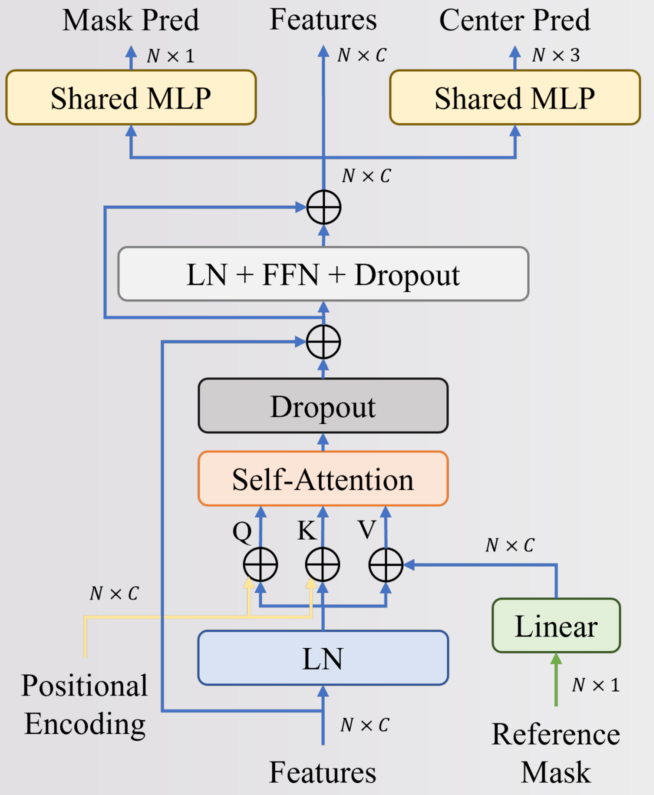


Embed local geometry into point features

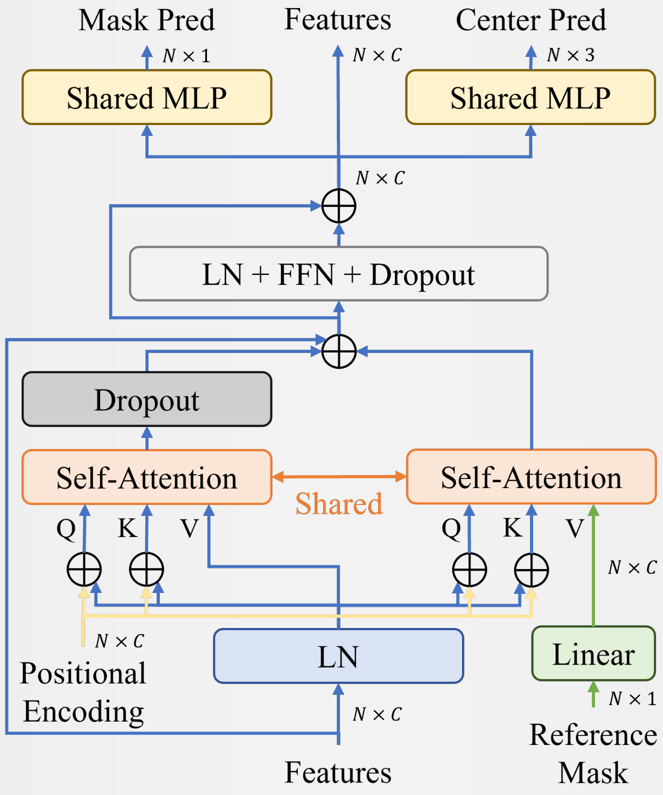
Enhance features with contextual information & Propagate target cues to the current frame

Localization

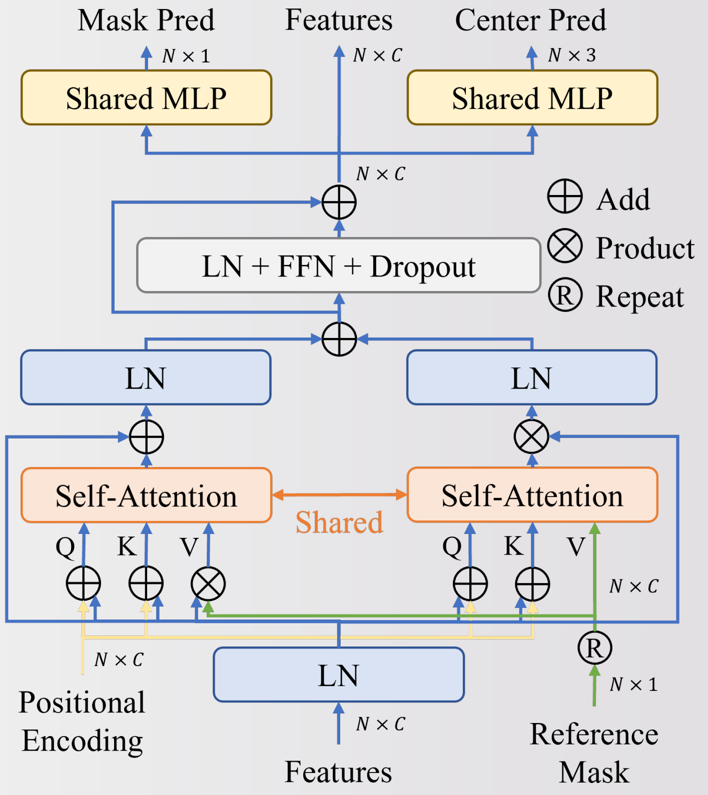
TARGET-CENTRIC TRANSFORMER



Vanilla

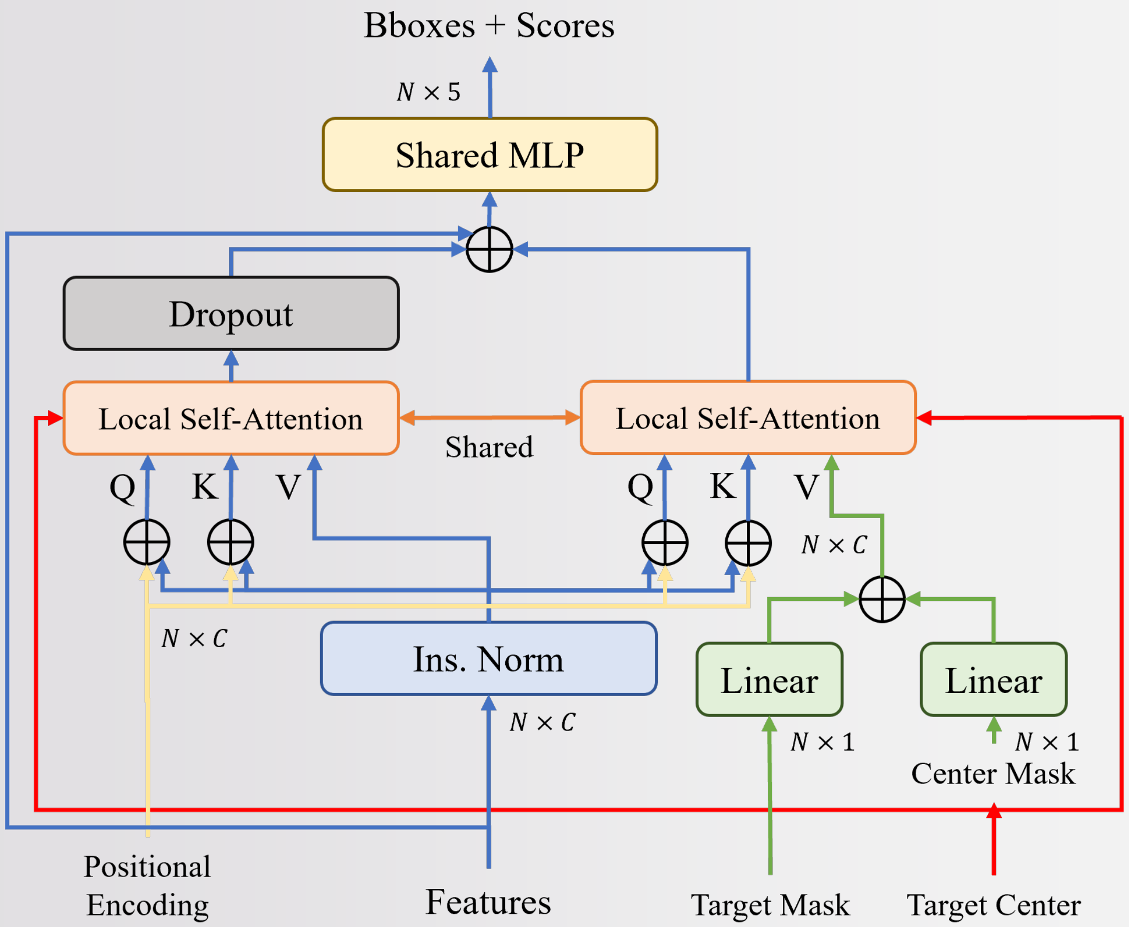


Semi-dropout 🙂



Gated

X-RPN



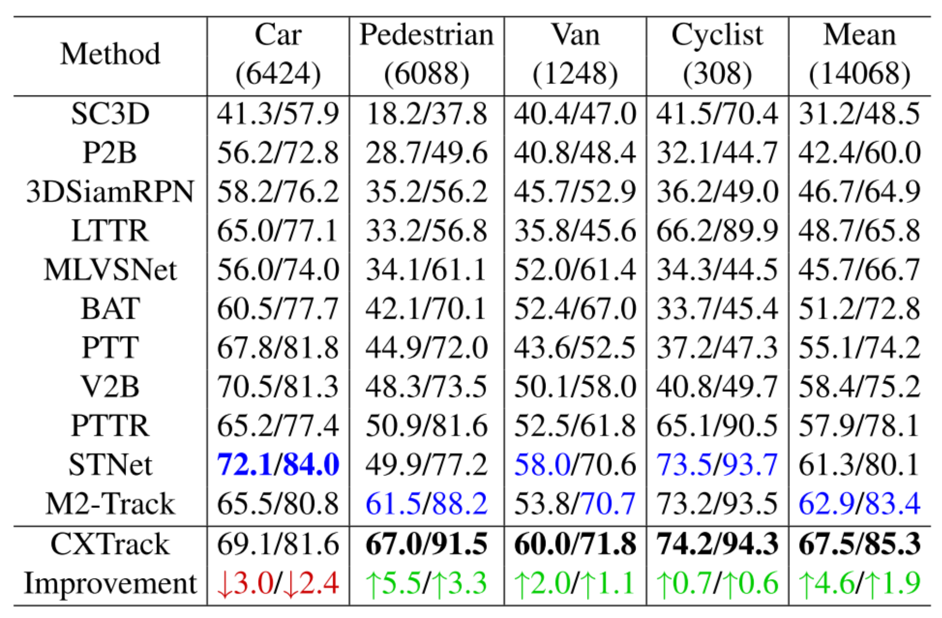
- Local Attention : each point should only interact with points belonging to the same object

$$N(p_i) = \{p_j | ||c_i - c_j||_2 < r\}$$

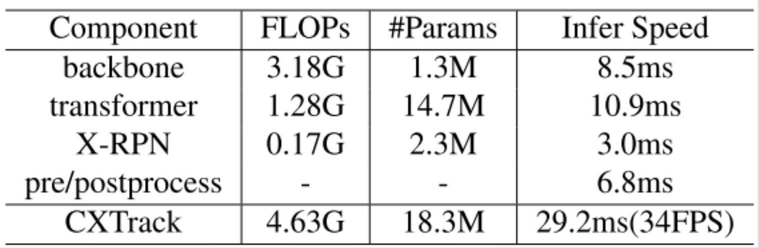
- Center Embedding : the tracked target is closer to its previous position than intra-class distractors (if the sample frequency is relatively high)

$$m_i^c = \exp(-\frac{||c_i - \bar{c}||_2^2}{2\sigma^2})$$

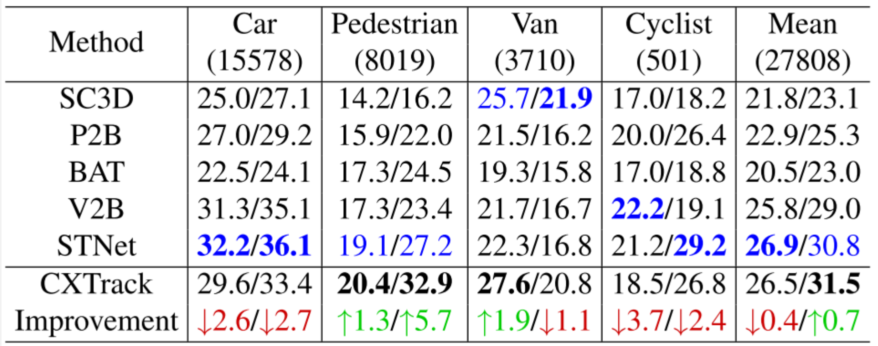
EXPERIMENTAL RESULTS



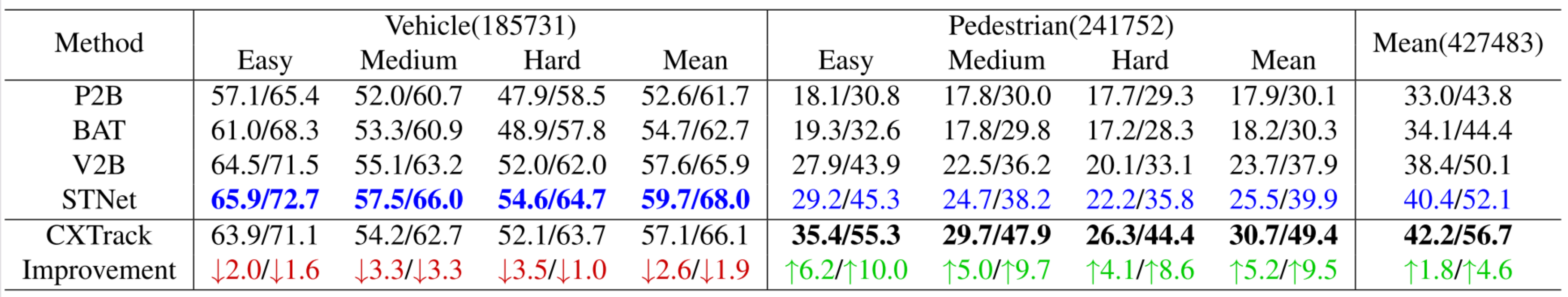
Method	Car (6424)	Pedestrian (6088)	Van (1248)	Cyclist (308)	Mean (14068)
SC3D	41.3/57.9	18.2/37.8	40.4/47.0	41.5/70.4	31.2/48.5
P2B	56.2/72.8	28.7/49.6	40.8/48.4	32.1/44.7	42.4/60.0
3DSiamRPN	58.2/76.2	35.2/56.2	45.7/52.9	36.2/49.0	46.7/64.9
LTTR	65.0/77.1	33.2/56.8	35.8/45.6	66.2/89.9	48.7/65.8
MLVSNet	56.0/74.0	34.1/61.1	52.0/61.4	34.3/44.5	45.7/66.7
BAT	60.5/77.7	42.1/70.1	52.4/67.0	33.7/45.4	51.2/72.8
PTT	67.8/81.8	44.9/72.0	43.6/52.5	37.2/47.3	55.1/74.2
V2B	70.5/81.3	48.3/73.5	50.1/58.0	40.8/49.7	58.4/75.2
PTTR	65.2/77.4	50.9/81.6	52.5/61.8	65.1/90.5	57.9/78.1
STNet	**72.1/84.0**	49.9/77.2	58.0/70.6	73.5/93.7	61.3/80.1
M2-Track	65.5/80.8	61.5/88.2	53.8/70.7	73.2/93.5	62.9/83.4
CXTrack	69.1/81.6	**67.0/91.5**	**60.0/71.8**	**74.2/94.3**	**67.5/85.3**
Improvement	↓3.0/↓2.4	↑5.5/↑3.3	↑2.0/↑1.1	↑0.7/↑0.6	↑4.6/↑1.9



Component	FLOPs	#Params	Infer Speed
backbone	3.18G	1.3M	8.5ms
transformer	1.28G	14.7M	10.9ms
X-RPN	0.17G	2.3M	3.0ms
pre/postprocess	-	-	6.8ms
CXTrack	4.63G	18.3M	29.2ms(34FPS)

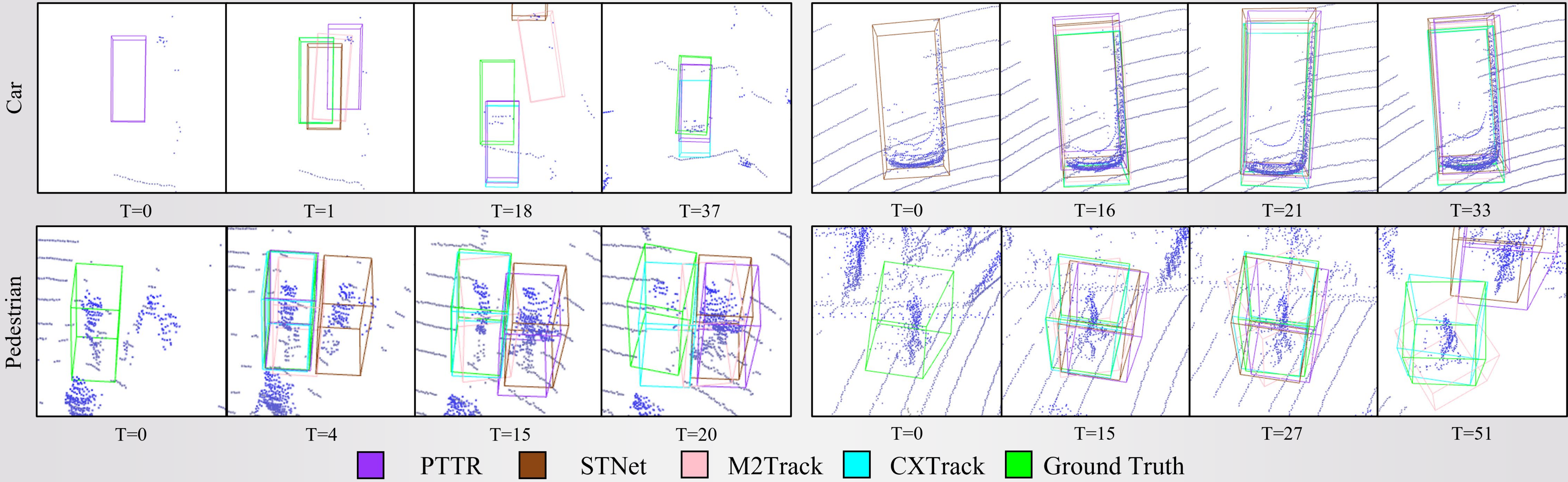


Method	Car (15578)	Pedestrian (8019)	Van (3710)	Cyclist (501)	Mean (27808)
SC3D	25.0/27.1	14.2/16.2	25.7/**21.9**	17.0/18.2	21.8/23.1
P2B	27.0/29.2	15.9/22.0	21.5/16.2	20.0/26.4	22.9/25.3
BAT	22.5/24.1	17.3/24.5	19.3/15.8	17.0/18.8	20.5/23.0
V2B	31.3/35.1	17.3/23.4	21.7/16.7	**22.2**/19.1	25.8/29.0
STNet	**32.2/36.1**	19.1/27.2	22.3/16.8	21.2/**29.2**	**26.9**/30.8
CXTrack	29.6/33.4	**20.4/32.9**	**27.6**/20.8	18.5/26.8	26.5/**31.5**
Improvement	↓2.6/↓2.7	↑1.3/↑5.7	↑1.9/↓1.1	↓3.7/↓2.4	↓0.4/↑0.7



Method	Vehicle(185731)				Pedestrian(241752)				Mean(427483)
	Easy	Medium	Hard	Mean	Easy	Medium	Hard	Mean	
P2B	57.1/65.4	52.0/60.7	47.9/58.5	52.6/61.7	18.1/30.8	17.8/30.0	17.7/29.3	17.9/30.1	33.0/43.8
BAT	61.0/68.3	53.3/60.9	48.9/57.8	54.7/62.7	19.3/32.6	17.8/29.8	17.2/28.3	18.2/30.3	34.1/44.4
V2B	64.5/71.5	55.1/63.2	52.0/62.0	57.6/65.9	27.9/43.9	22.5/36.2	20.1/33.1	23.7/37.9	38.4/50.1
STNet	**65.9/72.7**	**57.5/66.0**	**54.6/64.7**	**59.7/68.0**	29.2/45.3	24.7/38.2	22.2/35.8	25.5/39.9	40.4/52.1
CXTrack	63.9/71.1	54.2/62.7	52.1/63.7	57.1/66.1	**35.4/55.3**	**29.7/47.9**	**26.3/44.4**	**30.7/49.4**	**42.2/56.7**
Improvement	↓2.0/↓1.6	↓3.3/↓3.3	↓3.5/↓1.0	↓2.6/↓1.9	↑6.2/↑10.0	↑5.0/↑9.7	↑4.1/↑8.6	↑5.2/↑9.5	↑1.8/↑4.6

EXPERIMENTAL RESULTS



LIMITATIONS & FUTURE WORK

- Failure cases
  - The point clouds are too sparse to capture informative local geometry → Light-weight design
  - Large appearance variations occur(target missing) → Exploiting historical information
  - The scale of the displacement between training and testing data differs significantly

清華大學
Tsinghua University

CARDIFF
UNIVERSITY
PRIFYSGOL
CAERDYDD

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

THANK YOU!

Tian-Xing Xu
xutx21@mails.tsinghua.edu.cn