Robust Multi-Modality Multi-Object Tracking

Wenwei Zhang[1], Hui Zhou[2], Shuyang Sun[3], Zhe Wang[2], Jianping Shi[2], Chen Change Loy[1]

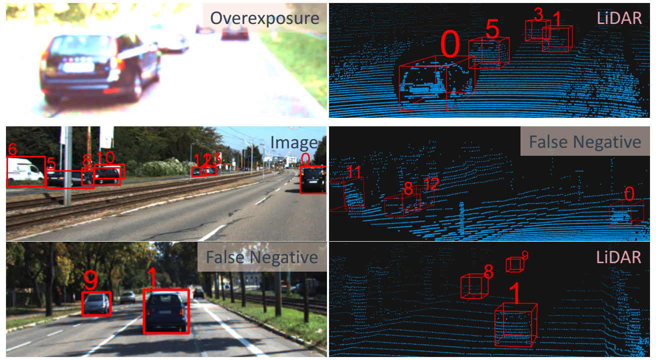
[1]Nanyang Technological University, [2]SenseTime Research, [3]University of Oxford

ICCV 2019 Seoul, Korea

Why Multi-Modality MOT in Autonomous Driving?

Sequential information of moving objects is helpful. **But:**

1. Relying on single sensor lacks reliability
2. Multi-sensor information could reinforce the perception ability



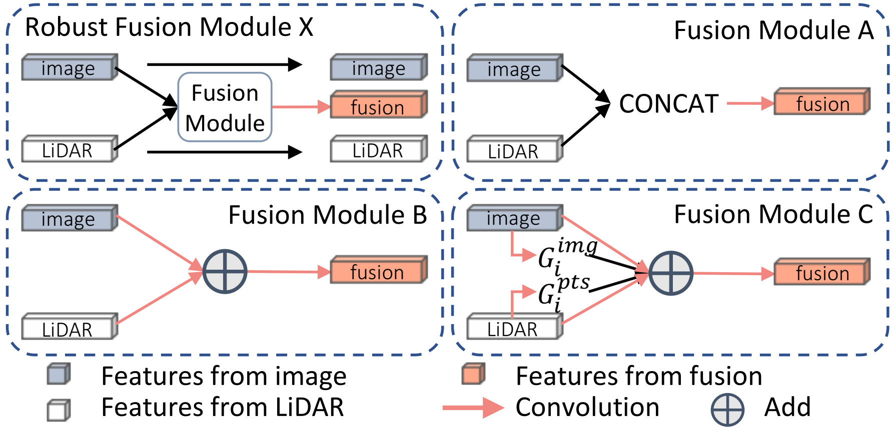
Contributions:

1. A multi-modality MOT framework with a robust fusion module that exploits multi-modality information to improve both reliability and accuracy.
2. A novel end-to-end training method that enables joint optimization of cross-modality inference.
3. The first attempt to apply deep features of point cloud for tracking and obtain competitive results.

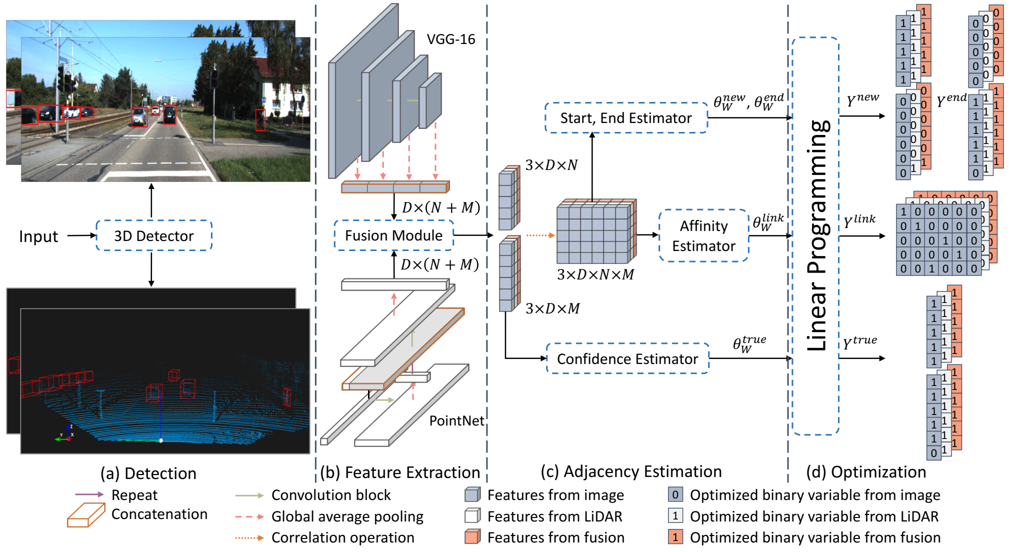
How to Exploit Multi-Modality and Keep Robust to Sensor Failure?

Fusion Module: Only provides fused modality. (Lacks reliability!!)

Robust Fusion Module: Also provides single modality.



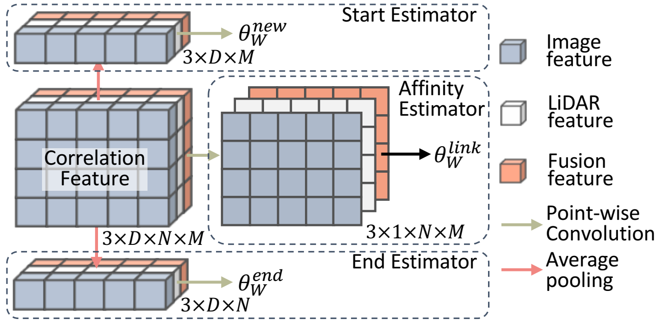
The Robust Multi-Modality Multi-Object Tracking Framework



1. The feature extractors extract features from image and LiDAR for each bounding box.
2. The robust fusion module fuses the multi-sensor features and outputs all the modalities.
3. The correlation operator produces the correlation features for each detection pair.
4. The adjacency estimator predicts the adjacency matrix based on correlation features.

How to Deal with Multi-Modality?

- Features of different modalities are concatenated in the batch dimension.
- Correlation operation is **batch-agnostic**.
- Convolution and pooling in the adjacency estimator are **batch-agnostic.**



Correlation Operation:

Element-wise Multiplication:

$$F_{jk} = F_j^i \odot F_k^{i+1}$$

Subtraction:

$$F_{jk} = F_j^i - F_k^{i+1}$$

Absolute subtraction:

$$F_{jk} = |F_j^i - F_k^{i+1}|$$

Advantages:

- **Accurate:** State-of-the-art on the KITTI benchmark.
- **Robust:** Still competitive under sensor failure.

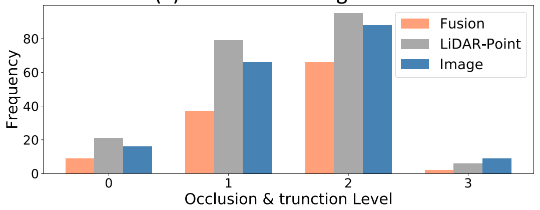
Comparison on the testing set of KITTI tracking benchmark

Method	MOTA	MOTP	Prec.	Recall	FP	FN	ID-s	Frag.	MT	ML
DSM [Frossard et al., ICRA2018]	76.15	83.42	98.09	80.23	578	7328	296	868	60.00	8.31
extraCK [Gunduz et al., IV2018]	79.99	82.46	98.04	84.51	642	5896	343	938	62.15	5.54
PMBM [Scheidegger et al., IV2018]	80.39	81.26	96.93	85.01	1007	5616	121	613	62.77	6.15
JCSTD [Tian et al., IEEE TITS2019]	80.57	81.81	98.72	83.37	405	6217	**61**	643	56.77	7.38
IMMDP [Xiang et al., ICCV2015]	83.04	82.74	**98.82**	86.11	**391**	5269	172	**365**	60.62	11.38
MOTBeyondPixels [Sharma, et al., ICRA2018]	84.24	**85.73**	97.95	88.80	705	4247	468	944	73.23	2.77
mmMOT with multi-modality	**84.77**	85.21	97.93	**88.81**	711	**4243**	284	753	**73.23**	**2.77**
mmMOT with point cloud only	84.53	85.21	97.93	88.81	711	4243	368	832	73.23	2.77
mmMOT with image only	84.59	85.21	97.93	88.81	711	4243	347	809	73.23	2.77

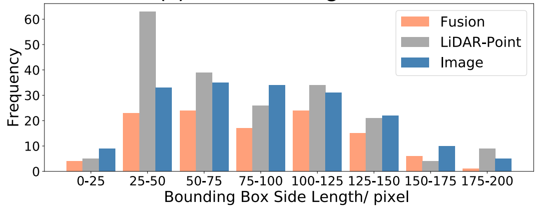
Failure Analysis:

- Occlusion, illumination and long distance are still challenging.
- Detector could cause early failures (False Negative).

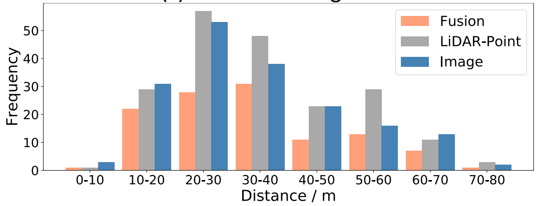
(a) Occlusion Histogram



(b) Box Size Histogram



(c) Distance Histogram



Interesting Phenomenon:

1. Most of ID switches come with occlusion. Occlusion causes more errors when only using point cloud than image.

2. More errors come with small bounding box size and long distance. Point cloud modality faces more errors in such cases.

Paper, code at:

https://github.com/ZwwWayne/mmMOT