Multi-Camera based Vehicle Tracking Using Collaborative Deep Learning

2020. 2. 11

HyungWon Kim (MSIS Lab)

Chungbuk National University, South Korea

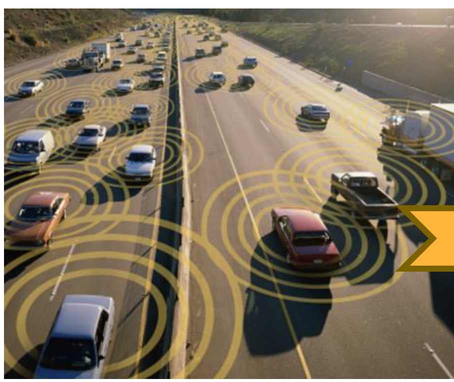
Agenda

- Introduction
- Multi-camera multi-object tracker
- Feature extraction
- Dataset for feature extraction
- Evaluation

V2X Based Multi-Camera Tracking System (Enabling Safer Autonomous Driving)

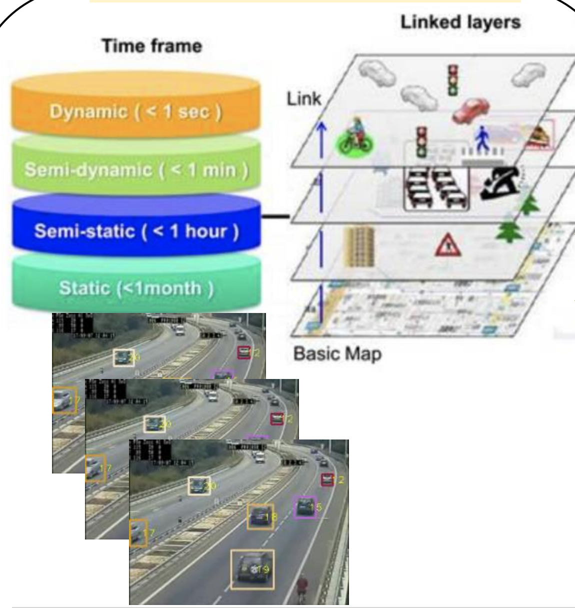
- **Smart V2X Network for Wide Coverage of Local Dynamic Map**
- ✓ Problem: Single Camera ADAS ➔ Serious limitation for Autonomous driving
- ✓ Solution: Multi-Camera Tracking with V2X ➔ Higher Accuracy & Wider LDM Coverage

Limitation of Existing V2X



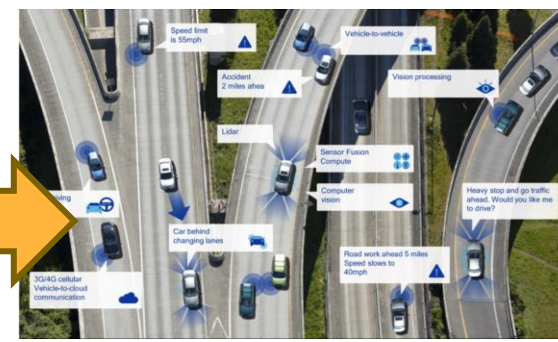
Dumb Short Range Position Data

Share Tracking Data Over V2X



LDM (Local Dynamic Map) + Sensing Fusion

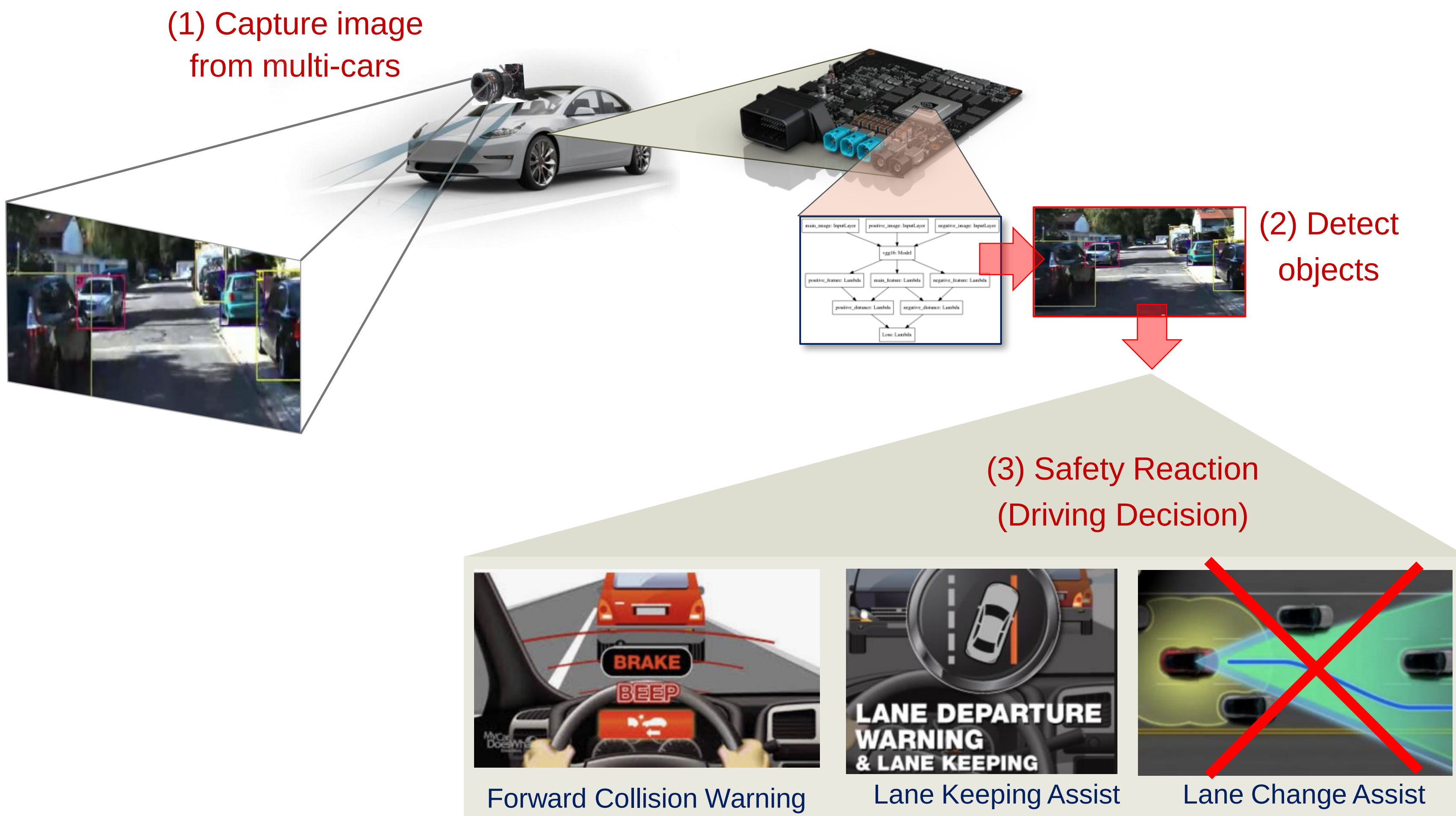
Higher Safety ADAS & Collaborative Self-Driving



Intelligent Wide Range Mobility Map

Problem of Current ADAS System
(Cannot Build LDM → Limited Sensing Range)

- Rely on ADAS sensors on a single vehicle with no V2X
- Object detection is limited only to visible vehicles
- Cannot allow full autonomous driving (High Speed Lane change assist is not possible)

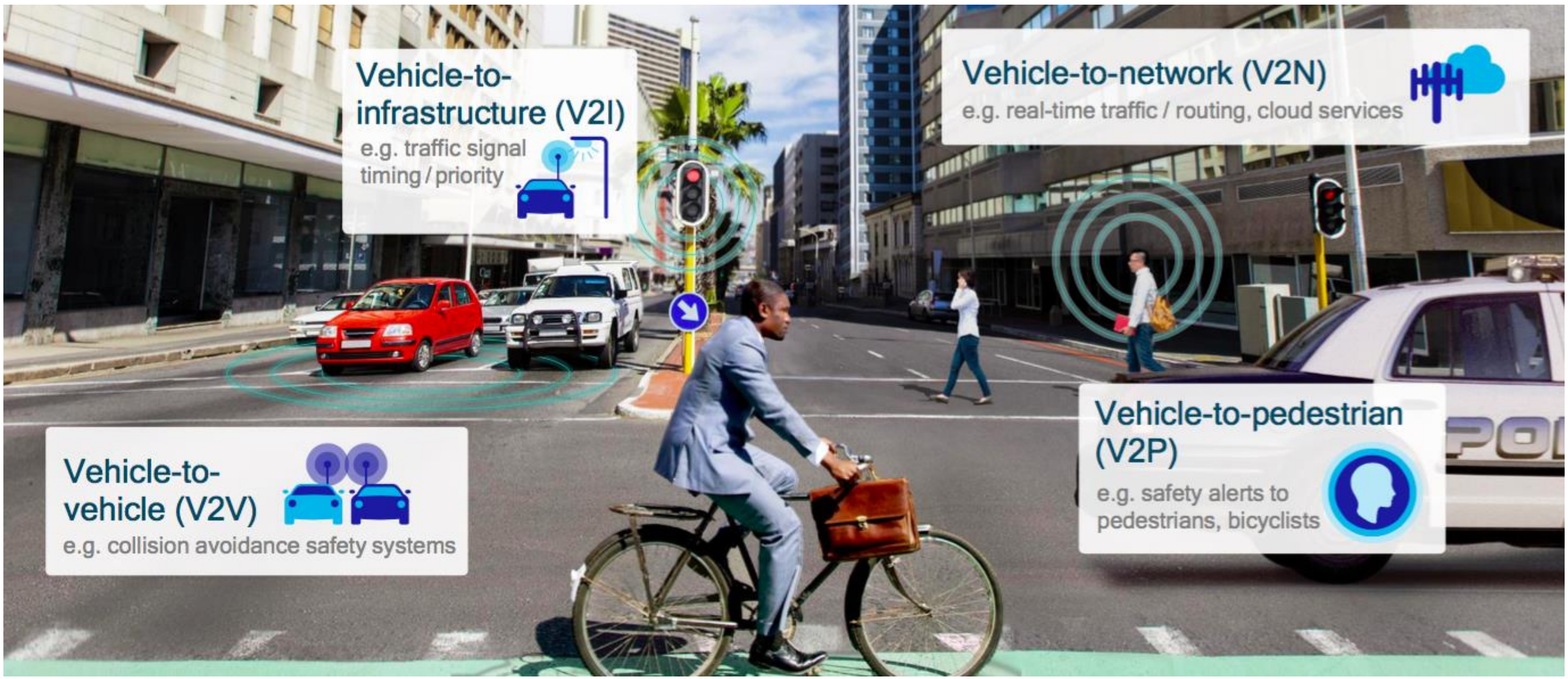


Current Limitation of Autonomous Driving

- ✓ Tesla in Autopilot drove underneath a white trailer
- ✓ Tesla in Autopilot drove towards the concrete barrier
- ✓ Uber’s Volvo SUV did not sense a walker with a bike at night
- ✓ A self-driving bus crashed in Vegas.
- ✓ A commercial van running a red light truck Google's autonomous Lexus SUVs

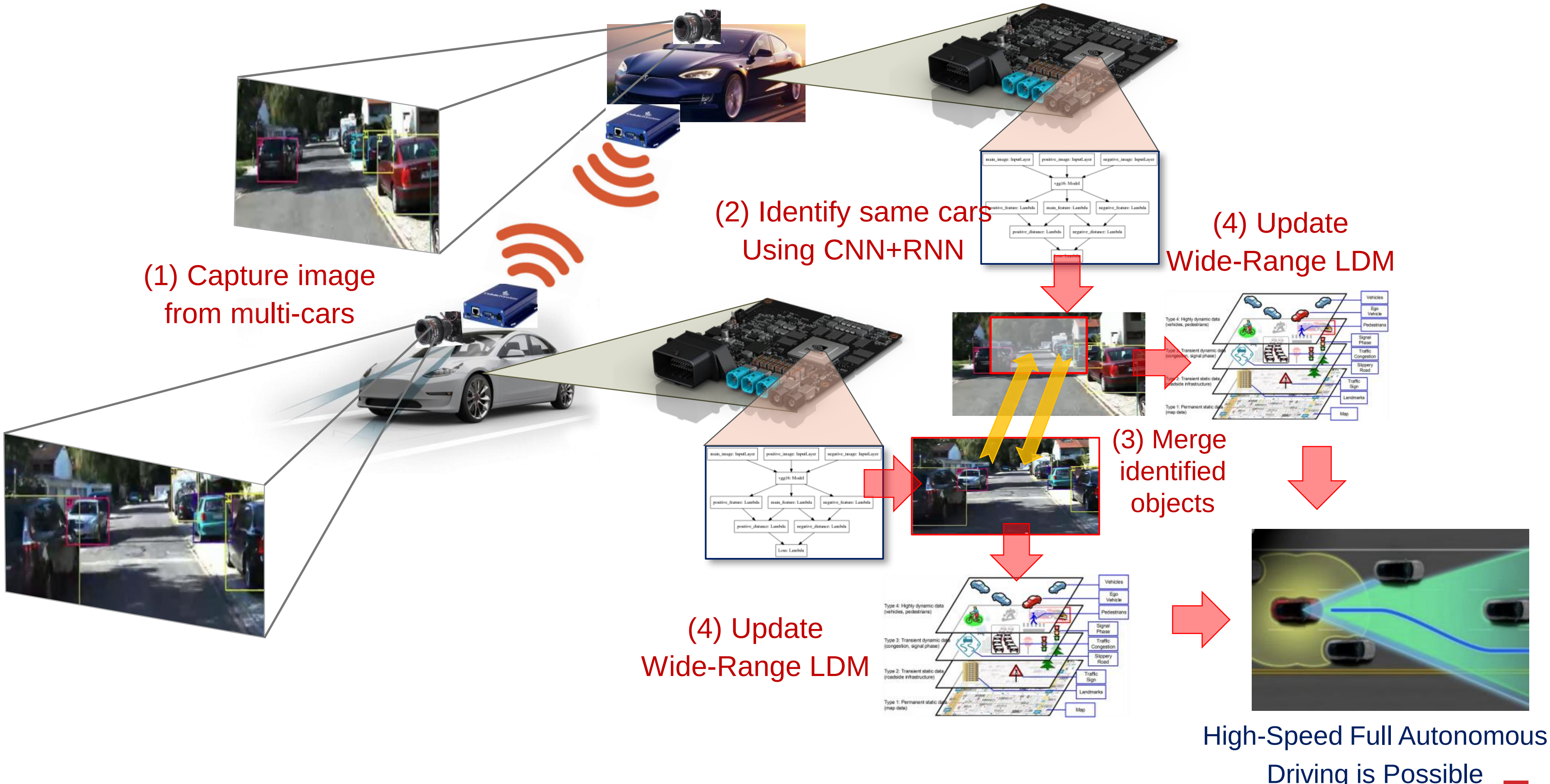
Objectives of V2X (Connected Car)

- Accident Avoidance: V2I & V2V (V2X)
- Cloud service: V2N
- Pedestrian safety: V2P



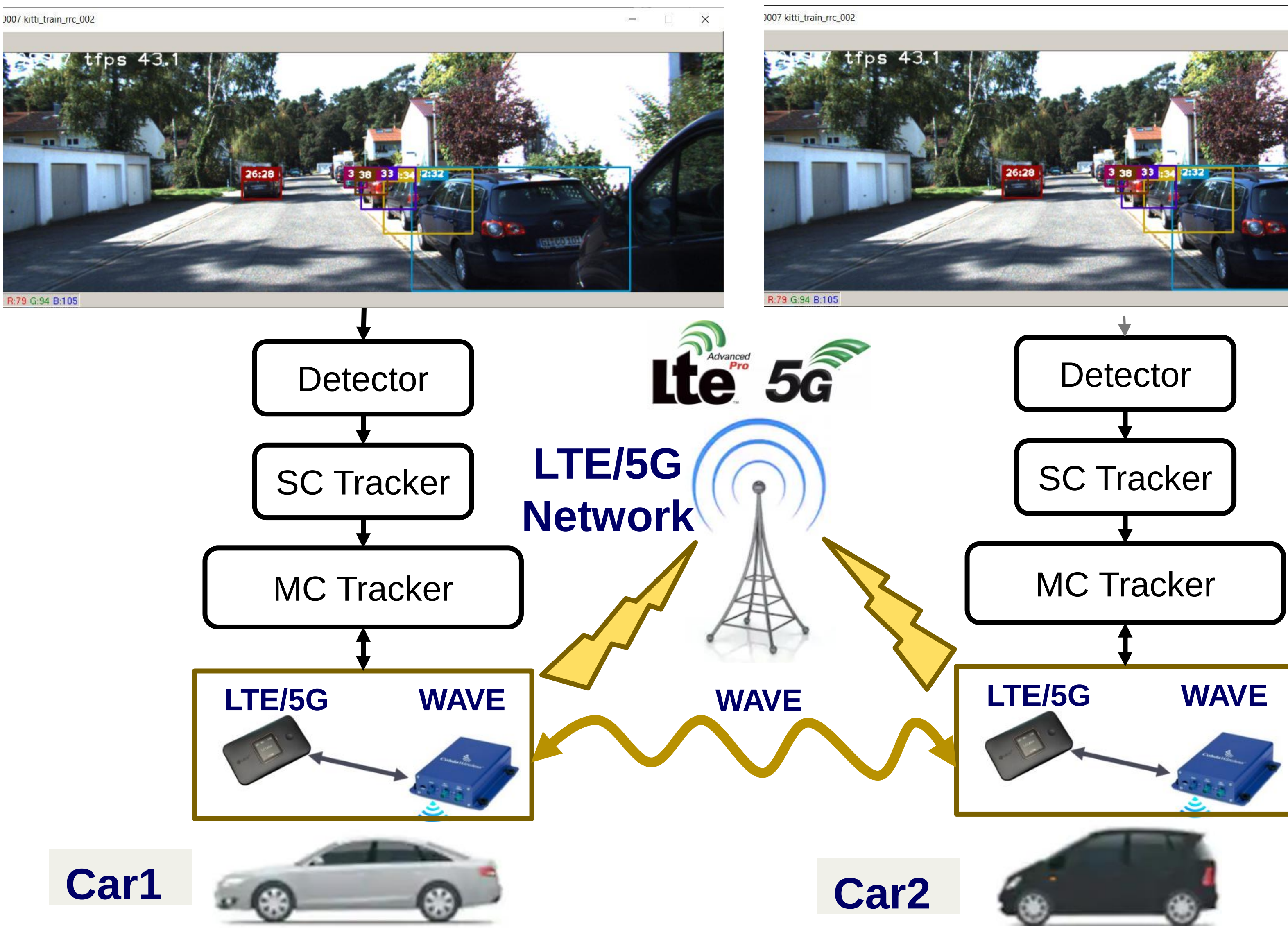
Multi-Camera Vehicle Tracking With V2X

- Can be used even before V2X is widely adopted (DSRC/WAVE, LTE/5G C-V2X)
- Share Multi-car sensing data with surrounding cars using V2X ➔ Wide Range LDM
- Allow full autonomous driving with high speed lane change and driving decisions

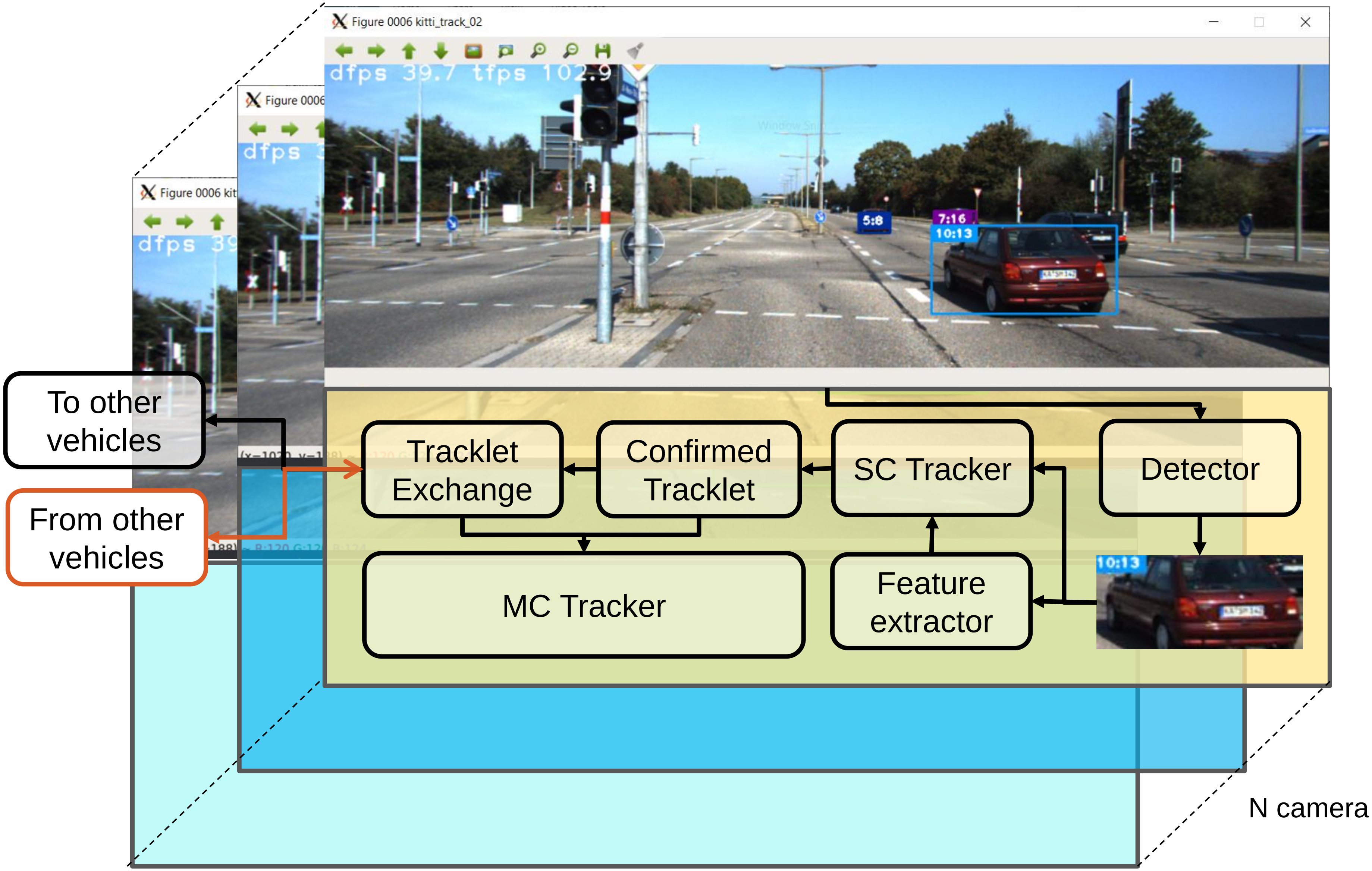


Multi-Camera Based Vehicle tracking with V2X

❖ System overview



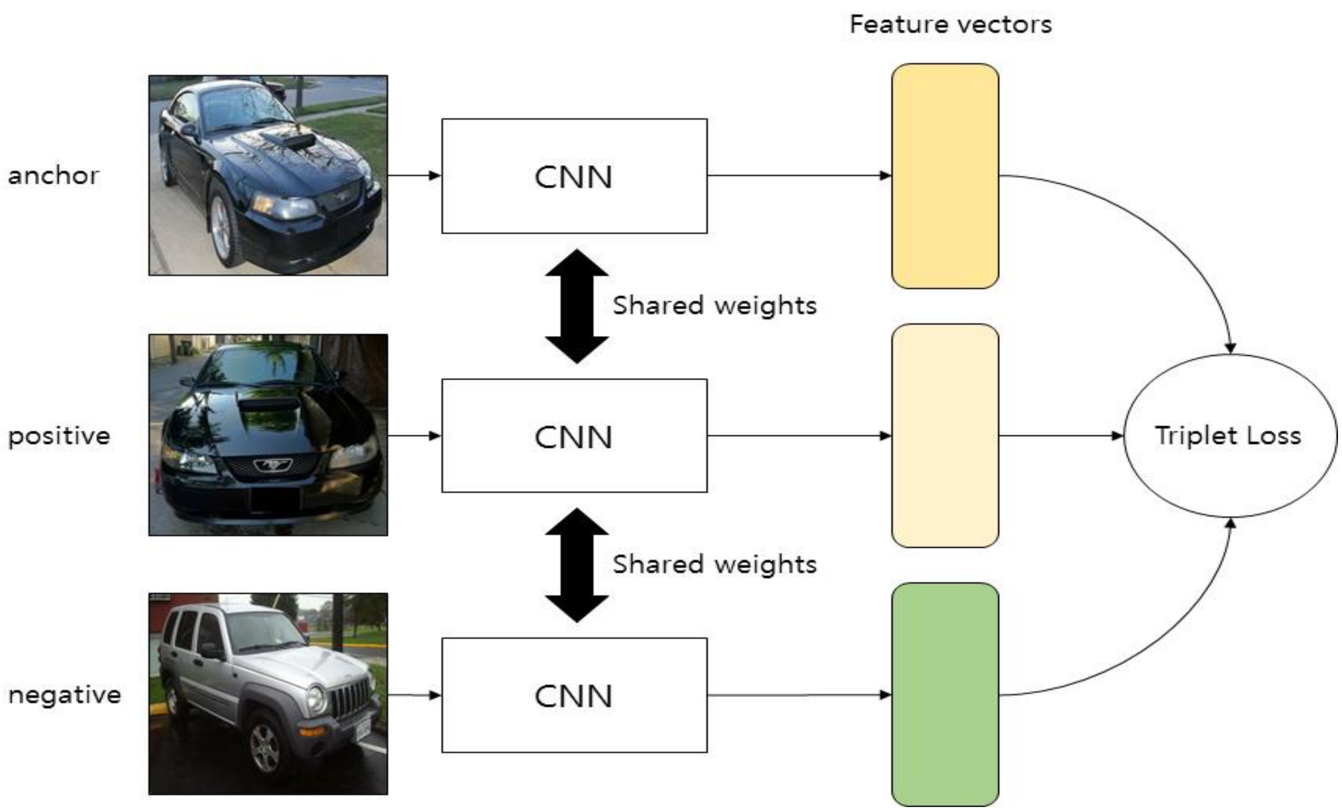
Overall Flow of Multi-Camera Tracker Algorithm



Appearance Feature Extraction

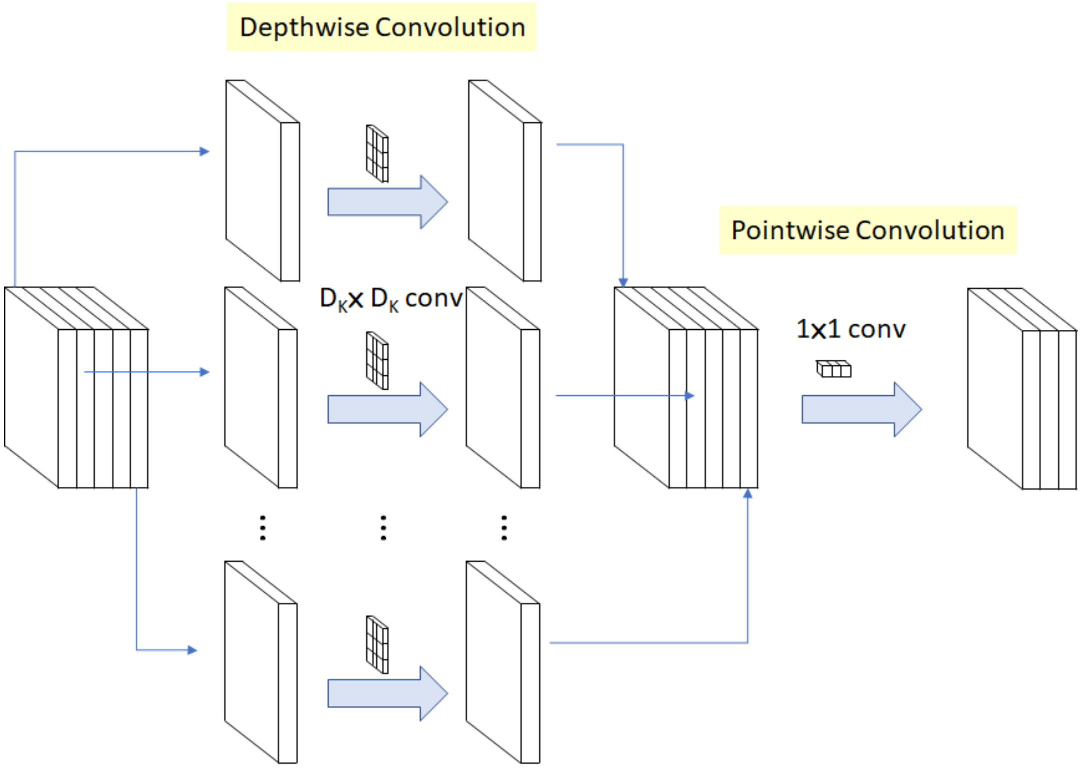
❖ Triplet loss for Training Feature Extraction CNN

- General classification: use fixed number of classes
- Multi vehicle tracking: there are many vehicles appeared, and some vehicles appearance may be similar in real-world.
- Compare multiple detected vehicles and determine if it is new or same as previous tracked vehicles



Neural Network Compression Decomposition of 3D Conv. Filters (Depthwise Separable Convolution)

- **Depthwise separable convolution**
  - ✓ Depthwise convolution followed by a pointwise convolution
- **Depthwise convolution**
  - ✓ Channel-wise DK×DK spatial convolution.
- **Pointwise convolution** :
  - ✓ 1×1 convolution to change the dimension



✓ Depthwise computation cost: $D_K \times D_F$

✓ Conventional convolution computation cost: $D_K \times D_K \times M \times N \times D_F \times D_F$

✓ Computation cost reduction: $\frac{D_K D_K M D_F D_F + M N D_F D_F}{D_K D_K M N D_F D_F} = \frac{1}{N} + \frac{1}{D_K D_K}$

➔ Significant Computation Reduction

M: # of input channels,

N: # of output channels,

D_K: Kernel size

D_F: Feature map size

Appearance feature extractor

❖ Euclidean Distance for Feature Quality Metric

- A squared difference distance between two feature vector in Euclidean space

$$D_{ED}(f_\theta(x_i), f_\theta(x_j)) = \|f_\theta(x_i) - f_\theta(x_j)\|_2^2$$

❖ Triplet Loss function for Training

- **Using P-K Batch with hardest samples**

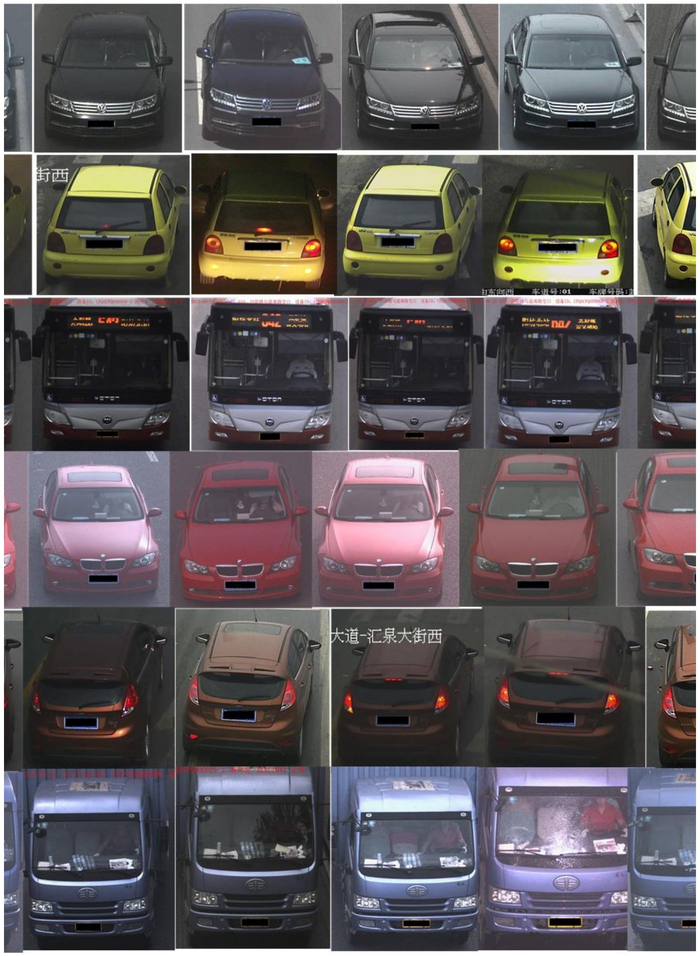
$$\mathcal{L}_{\text{BH}}(\theta; X) = \overbrace{\sum_{i=1}^{P} \sum_{a=1}^{K}}^{\text{all anchors}} \Big[m + \overbrace{\max_{p=1\ldots K} D\left(f_\theta(x_a^i), f_\theta(x_p^i)\right)}^{\text{hardest positive}} - \underbrace{\min_{\substack{j=1\ldots P \\ n=1\ldots K \\ j\neq i}} D\left(f_\theta(x_a^i), f_\theta(x_n^j)\right)}_{\text{hardest negative}} \Big]_+ , \quad (5)$$

✓ **P-K batch**

- randomly sampling *P* classes (vehicle IDs)
- Randomly sampling *K* images from each class (vehicle ID)

In Defense of the Triplet Loss for Person Re-Identification

Dataset for Training Feature Extractor CNN



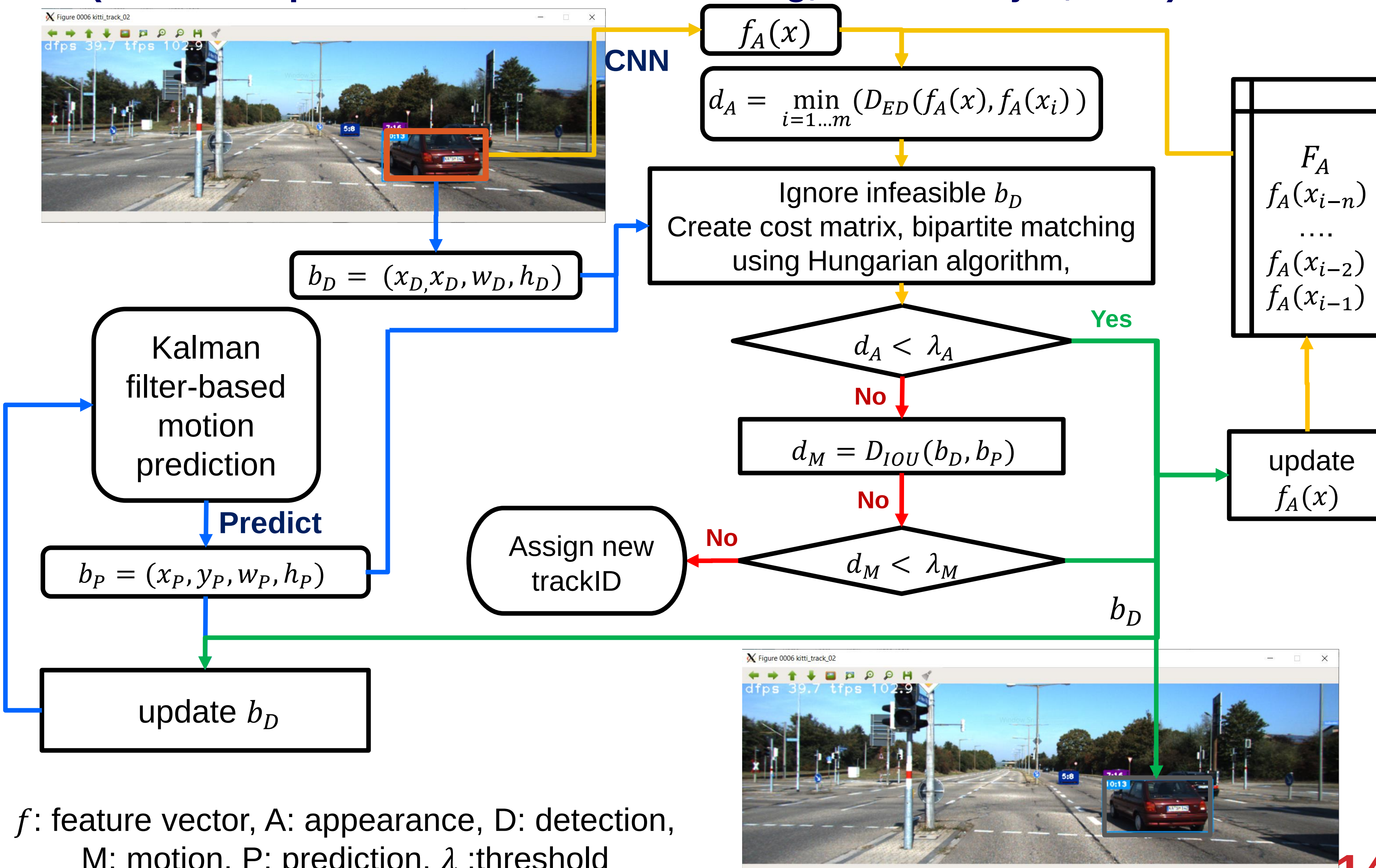
❖ Vehicle-1M Dataset

- Captured across day and night, from head or rear, by multiple surveillance cameras
- Total of 936,051 images from 55,527 vehicles and 400 vehicle models in the dataset.
- Real world vehicle model label indicating the maker, model and year of the vehicle (i.e. "Audi-A6-2013")

Dataset	# of ID	# of img
Training	50000	844571
query_test_1000	1000	16123
query_test_2000	2000	32539
query_test_3000	3000	49259
query_test_full	5527	91480

Single-Camera Vehicle Tracking (SC-VT)

- **Employing Deep SORT tracking with Enhancement (SORT: Simple Online Real Time Tracking, Nicolai Wojke, et al.)**



f: feature vector, A: appearance, D: detection, M: motion, P: prediction, λ :threshold

Multi-Camera Vehicle Tracking (MC-VT)

❖ Proposed MC-VT architecture

- Assuming communication has no errors

Algorithm 1 MC-VT - Multi-camera multi vehicle tracking

Input: consecutive images frames Frames $= frame_1, frame_2, ..., frame_t$
Output: Σ: list of multi-camera tracklet with pairs of egoId and remoteIds
for $frame$ in $frame_1, frame_2, ..., frame_t$ **do**
  $\tau_m^{ego} \leftarrow$ list of m confirmed tracklet from SC-VT of ego camera
  $\tau_n^{remote} \leftarrow$ list of n confirmed tracklet from SC-VT of remote camera
  M^{aff} : affinity score matrix
  for $i \in (1, m), j \in (1, n)$ **do**
    $f_A(x^{ego}) \leftarrow$ apperance feature history of τ_i^{ego}
    $f_A(x^{rem}) \leftarrow$ newest apperance feature of τ_j^{remote}
    $M_{i,j}^{aff} = min(D_{ED}(f, f_A(x^{rem})); \forall f \in f_A(x^{ego})$
    Discard $M_{i,j}^{aff}$ if $M_{i,j}^{aff} > \lambda_A$
  end for
  Compute M^{agn} from M^{aff} using Hungarian algorithm
  Assign τ_m^{ego} and τ_n^{remote} into Σ from M^{agn}
  Generate $(egoId : remoteId)$ from Σ
end for

Evaluation Metric for SC-VT

❖ MOTP : Multiple Object Tracking Precision

- $d_{i,t}$: the bounding box overlap between the ground-truth object and its corresponding estimated bounding box i for frame t
- c_t : the number of matches found for frame t

$$MOTP = \frac{\sum_{i,t} d_{i,t}}{\sum_t c_t}$$

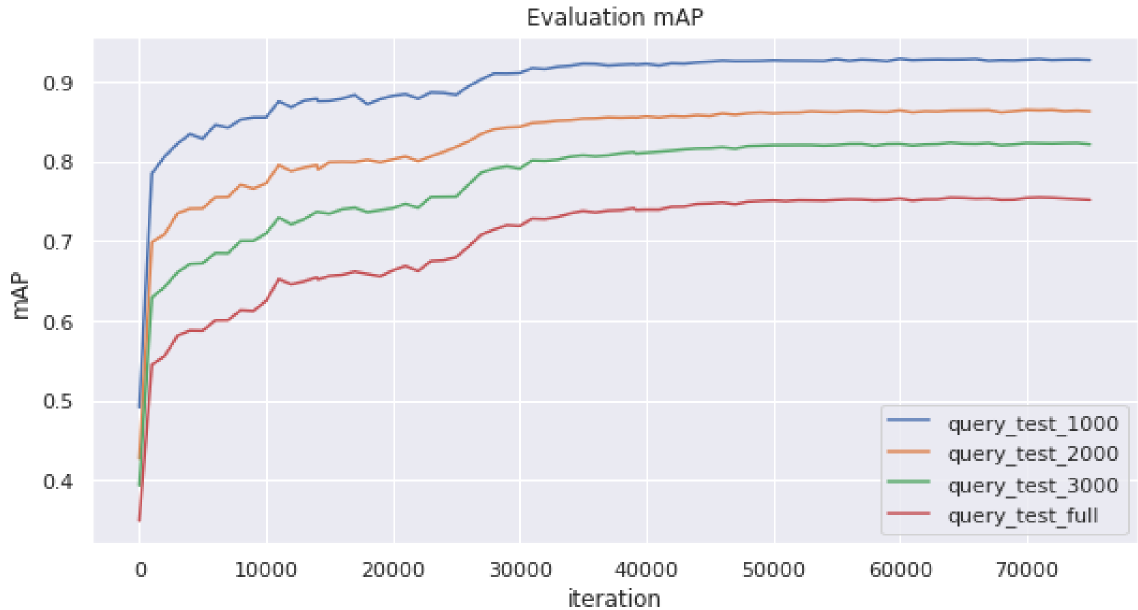
❖ MOTA : Multiple Object Tracking Accuracy

- m_t : the number of misses for frame t or false negative (FN)
- $f_{p,t}$: the number of false positives (FP)
- mme_t : the number of mismatches (correctly tracked but ID is changed) or IDS
- g_t : the number of ground-truth objects

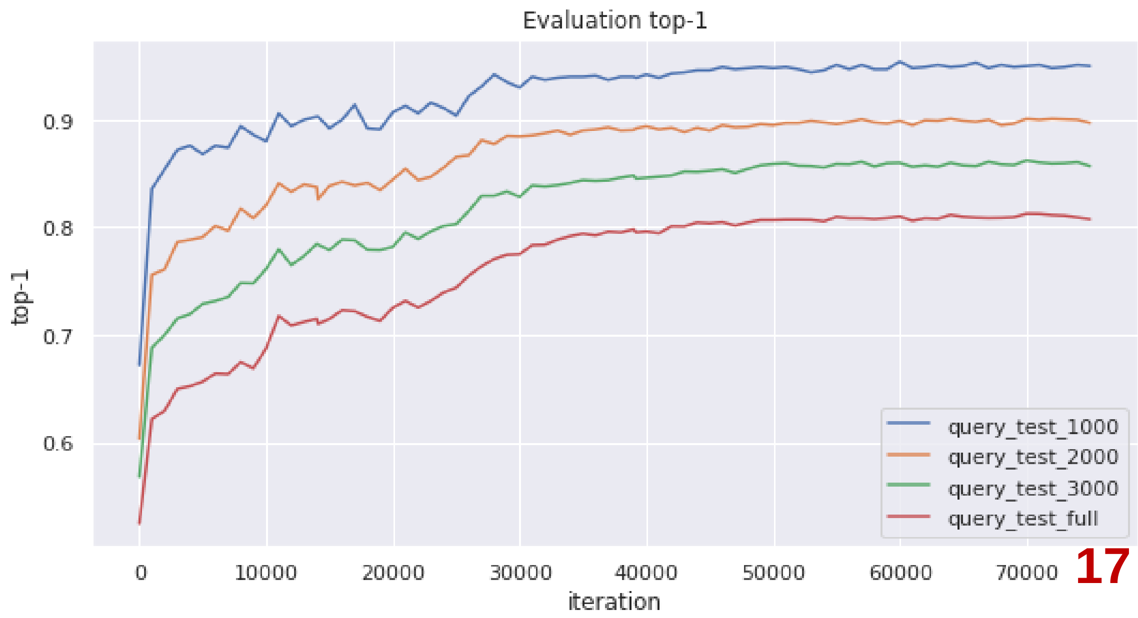
$$MOTA = 1 - \frac{\sum_t (m_t + f_{p,t} + mme_t)}{\sum_t g_t}$$

Feature Extractor Training Results

- **Mean Avg Precision (mAP) of feature extractor on Vehicle-1M dataset**



- **Top-1 accuracy of feature extractor on Vehicle-1M dataset**



Evaluation of MC-VT with KITTI Data set

❖ Ground Truth Generation for MC-VT

- Using KITTI stereo image data set (img_02 is Left image, img_03 is Right image), but ground truth is given for img_02 set only
- Stereo camera gap is 54 cm ➔ Two camera mostly cover the same number of vehicles in the same frame *t*
- Exploit this properties to calculate MCMOTA by matching remoteID (right image) with its corresponding egoID based on the bounding box of egoID's ground truth (GT)
- Right images have no GT, so we generated GT for each right image by running SC-VT on each dataset for comparison.
- For MC-MT, defined a new evaluation metric : MC-MOTA

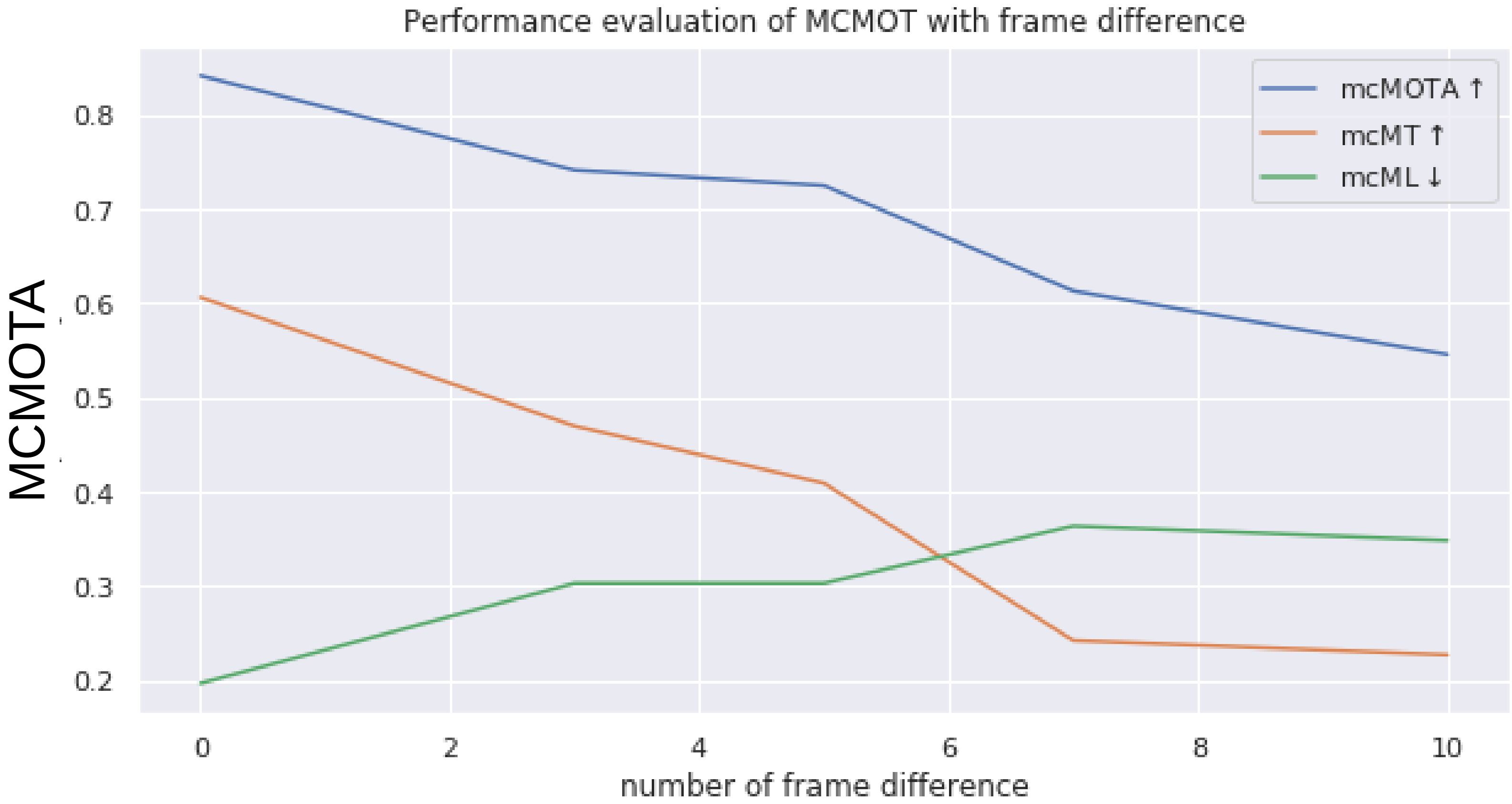
❖ MC-MOTA : Multi-Camera Multiple Object Tracking Accuracy

- m'_t : the number of missed remoteIDs w.r.t egoIDs (egoIDs without remoteID) for frame *t* (also called *mcFN)*
- $f'_{p,t}$: the number of false positives (remoteID without egoID) (also called mcFP)
- mme'_t : the number of remoteID switches w.r.t. egoID (also called mcIDS)
- g'_t : the number of ground-truth object pairs [egoID, remoteID]

$$MC{-}MOTA = 1 - \frac{\sum_t (m'_t + f'_{p,t} + mme'_t)}{\sum_t \ g_t}$$

MCMOTA Analysis with KITTI Seq7

- **Evaluation over various frame gap between two cameras**
  - **Frame gaps : 0 ~ 10 frames between Left image (img_02) and Right image (img_03)**



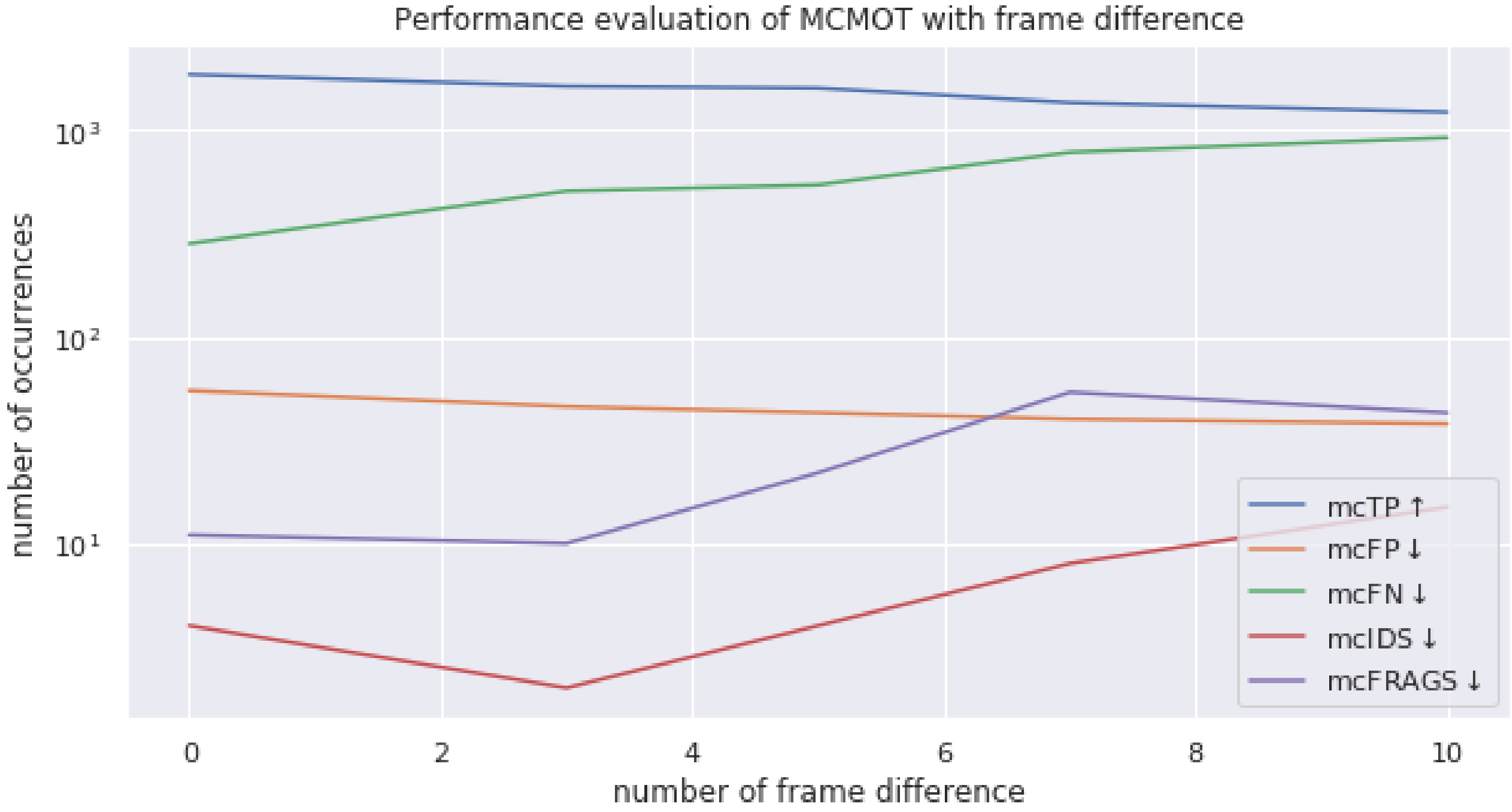
Up arrow: higher is better; down arrow: lower is better

MC-VT Failure Analysis with KITTI Seq7

❖ Multi-Camera TP, FP, FN, IDS, FRAGS

TP: True positive :correctly match
FP: False positive :wrong match
FN: False negative: miss match

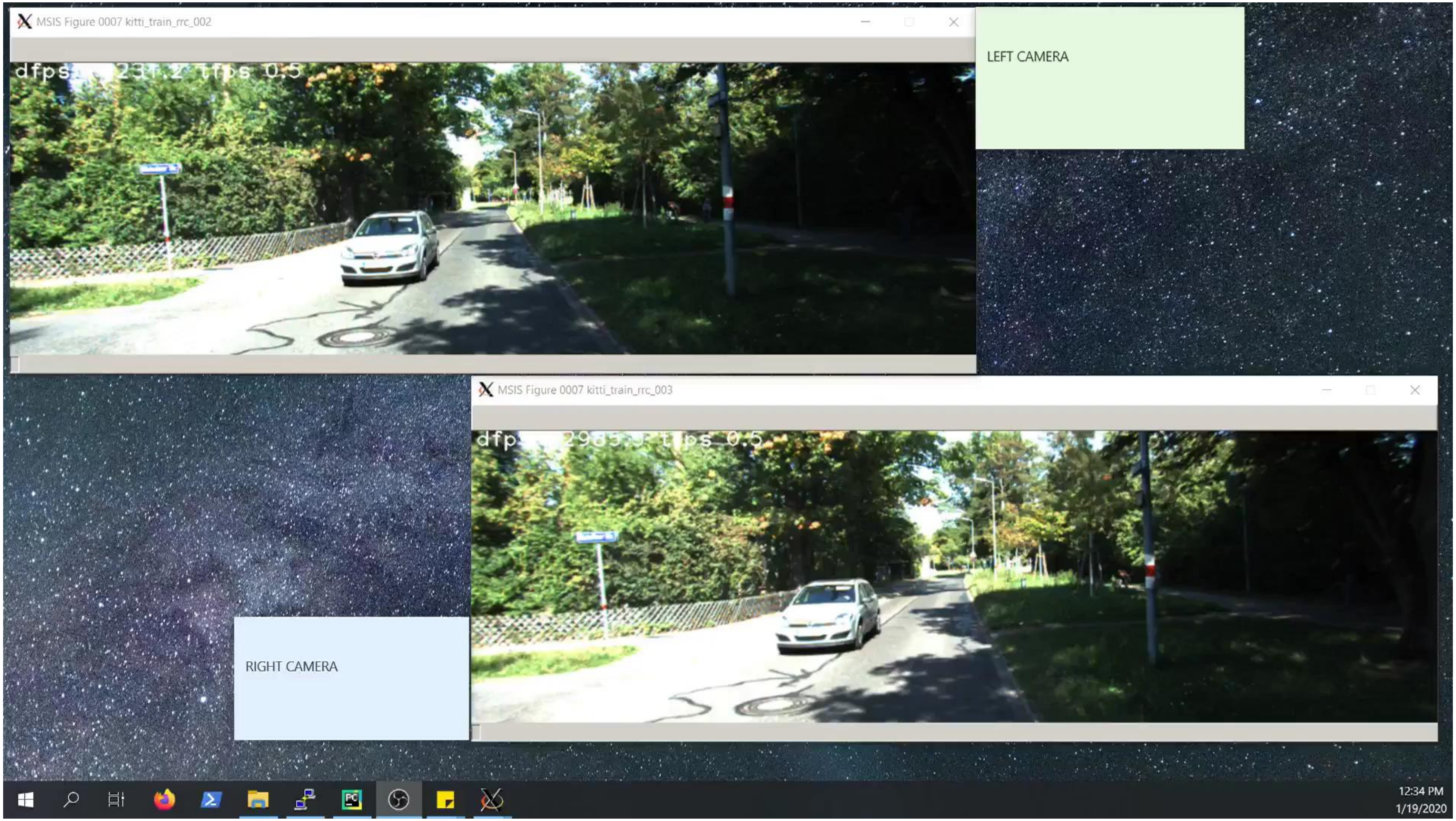
IDS: ID Switching
Frags: Fragmented Sequence



Up arrow: higher is better; down arrow: lower is better

Test Result with Sterio Camera Video

- Multi-Cam Tracking on KITTI seq7 (Stereo Cam with 0 frame gap)



Test Result with Two Vehicles

- **Multi-Cam Tracking on 2 vehicle's ADAS Cameras**