UNIVERSITY OF WASHINGTON
ELECTRICAL ENGINEERING

Single-camera and Inter-camera Vehicle Tracking and 3D Speed Estimation Based on Fusion of Visual and Semantic Features

Team 48

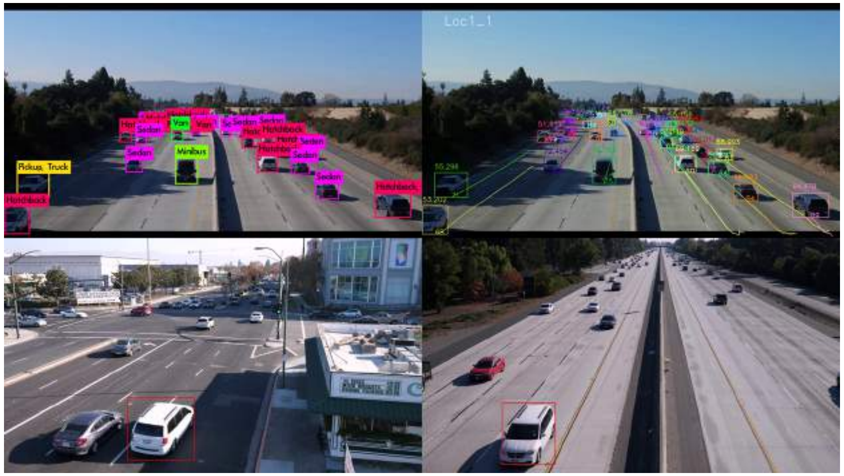
Zheng Tang, Gaoang Wang, Hao Xiao, Aotian Zheng, Jenq-Neng Hwang

Information Processing Lab, Department of Electrical Engineering

University of Washington

Introduction

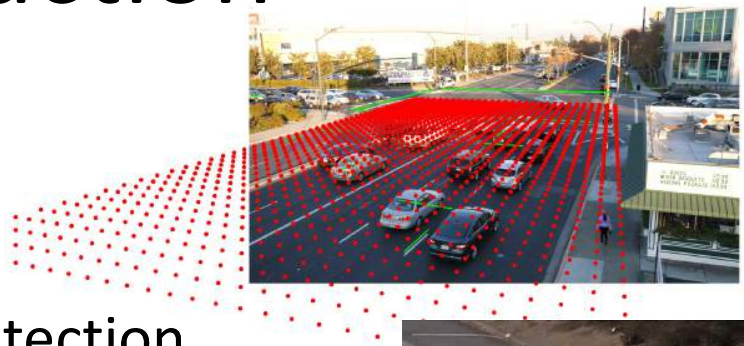
- Intelligent Transportation System (ITS)
  - Estimating traffic flow
  - Anomalies detection
  - Multi-camera tracking and re-identification



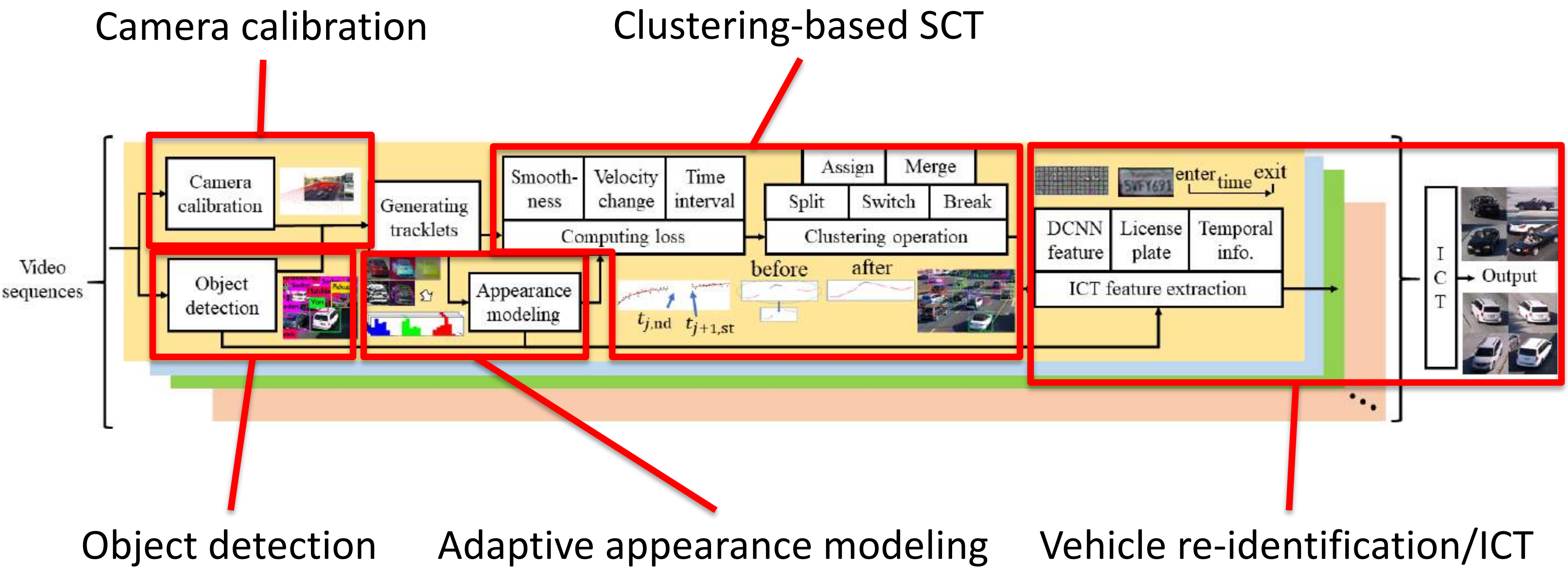
- Single-Camera Tracking (SCT)
  - Object detection/classification + data association
- Inter-Camera Tracking (ICT)
  - Re-identification of the same object(s) across multiple cameras

Introduction

- Challenges in SCT & ICT
  - Extraction of 3D information
  - Failure/confusion in object detection
  - High similarity among vehicle models
  - Frequent occlusion
  - Large variation in different viewing perspectives
  - Low video resolution (for license plate recognition)



Overview



Camera calibration
Clustering-based SCT
Video sequences
Camera calibration
Generating tracklets
Object detection
Appearance modeling
Smooth-ness
Velocity change
Time interval
Computing loss
Assign
Merge
Split
Switch
Break
Clustering operation
before
after
$t_{j,\text{nd}}$
$t_{j+1,\text{st}}$
enter
time
exit
DCNN feature
License plate
Temporal info.
ICT feature extraction
ICT
Output
Object detection
Adaptive appearance modeling
Vehicle re-identification/ICT

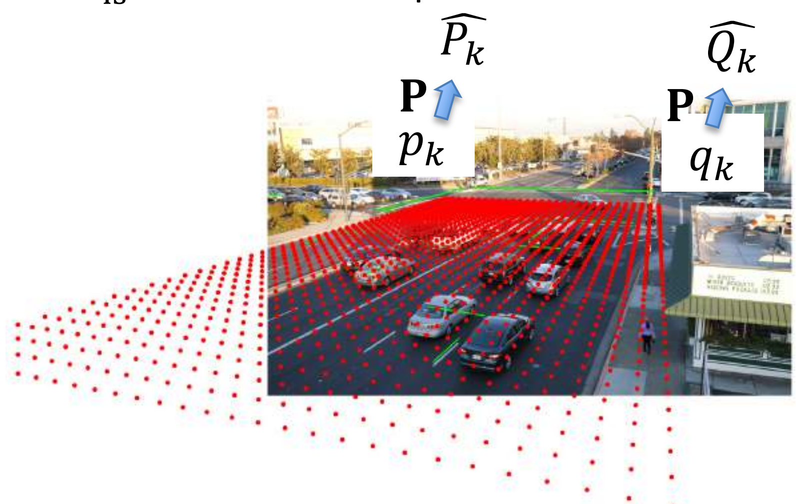
Camera Calibration

- Minimization of reprojection error solved by EDA

$$\min_{\mathbf{P}} \sum_{k=1}^{N_{\mathrm{ls}}} \left| \|P_k - Q_k\|_2 - \left\|\widehat{P_k} - \widehat{Q_k}\right\|_2 \right|$$

$$\mathrm{s.t.}\ \mathbf{P} \in \mathrm{Rng}_{\mathbf{P}}, p_k = \mathbf{P} \cdot \widehat{P_k}, q_k = \mathbf{P} \cdot \widehat{Q_k}$$

$\mathbf{P}$: Camera projection matrix
$\mathrm{Rng}_{\mathbf{P}}$: Range for optimization
P_k, Q_k: True endpoints of line segments
$\widehat{P_k}$, $\widehat{Q_k}$: Estimated endpoints of line segments
p_k, q_k: 2D endpoints of line segments
N_{ls}: Number of endpoints

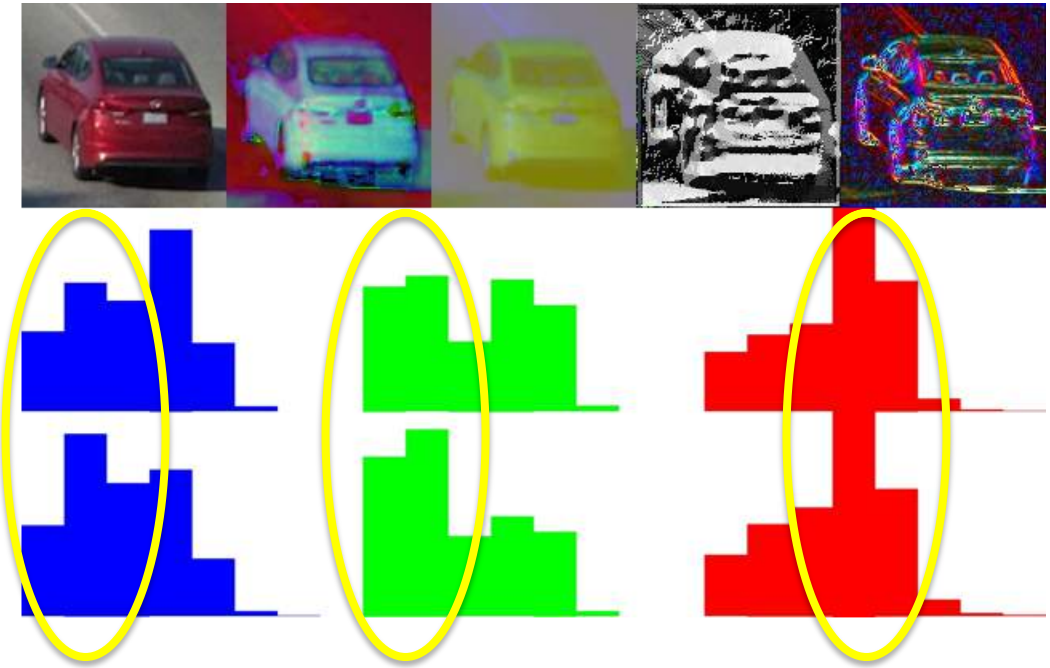


Object Detection

- YOLOv2 [Redmon et al., CVPR 2017]
  - Trained on ~4,500 manually labeled frames
  - 8 categories: Sedan, hatchback, bus, pickup, minibus, van, truck and motorcycle
  - Initialization: Provided pre-trained weights

Adaptive Appearance Modeling

- Histogram-based adaptive appearance model
  - A **history** of **spatially weighted (kernel)** histogram combinations will be kept for each vehicle



The first row respectively presents the RGB, HSV, Lab, LBP and gradient feature maps for an object instance in a tracklet, which are **used to build feature histograms**.
The second row shows the original RGB color histograms.
The third row demonstrates the Gaussian spatially weighted (kernel) histograms, where the contribution of background area is suppressed.

Clustering-based SCT

$$l = \sum_{i=1}^{n_{\mathrm{v}}} l_i$$

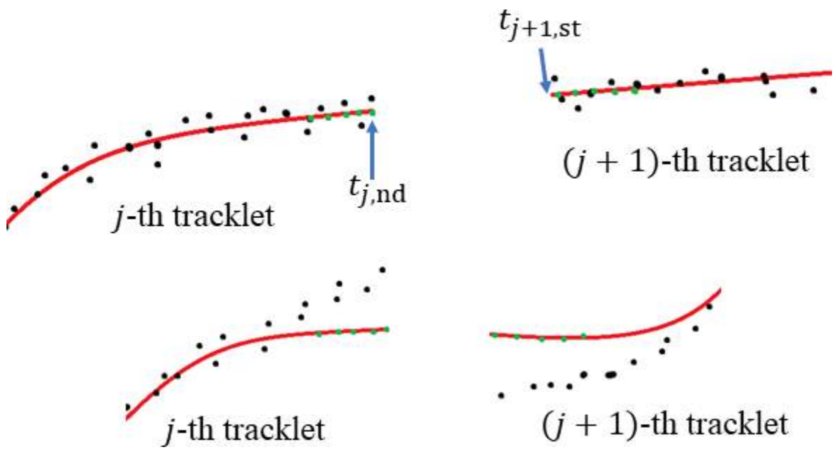
$$l_i = \lambda_{\mathrm{sm}} l_{i,\mathrm{sm}} + \lambda_{\mathrm{vc}} l_{i,\mathrm{vc}} + \lambda_{\mathrm{ti}} l_{i,\mathrm{ti}} + \lambda_{\mathrm{ac}} l_{i,\mathrm{ac}}$$

Smoothness Velocity Time interval Appearance

n_{v}: No. of vehicles in a single camera
l_i: Loss for the i-th vehicle
$l_{i,\mathrm{sm}}$: Smoothness loss
$l_{i,\mathrm{vc}}$: Velocity change loss
$l_{i,\mathrm{ti}}$: Time interval loss
$l_{i,\mathrm{ac}}$: Appearance change loss
λ's: Regularization parameters

Same trajectory

Different trajectory



Black dots show the detected locations at time *t.*
Red curves represent trajectories from Gaussian regression.
Green dots show n_{k} neighboring points on the red curves around the endpoints of the tracklets at $t_{j,\mathrm{nd}}$ and $t_{j+1,\mathrm{st}}$.

Clustering-based SCT

- Smoothness loss
  - The total distance between the regression trajectory and observed trajectory
- Velocity change loss
  - Maximum acceleration around each end point of the tracklets
- Time interval loss
  - Time interval between two adjacent tracklets
- Appearance change loss
  - (Average) Bhattacharyya distance between each pair of histograms in the adaptive appearance models

Clustering-based SCT

- Clustering operations

$$\Delta l_j{}^* = \arg\min_{\Delta l_j}\left(\Delta l_{j,\mathrm{as}}, \Delta l_{j,\mathrm{mg}}, \Delta l_{j,\mathrm{sp}}, \Delta l_{j,\mathrm{sw}}, \Delta l_{j,\mathrm{bk}}\right)$$

  - $\Delta l_{j,\mathrm{as}}$, $\Delta l_{j,\mathrm{mg}}$, $\Delta l_{j,\mathrm{sp}}$, $\Delta l_{j,\mathrm{sw}}$ and $\Delta l_{j,\mathrm{bk}}$ respectively stand for the changes of loss for *assign*, *merge*, *split*, *switch* and *break* operations.
  - The operation with minimum loss-change value is chosen.
  - If $\Delta l_j{}^* > 0$, no change is made for this tracklet.
  - Convergence is guaranteed.

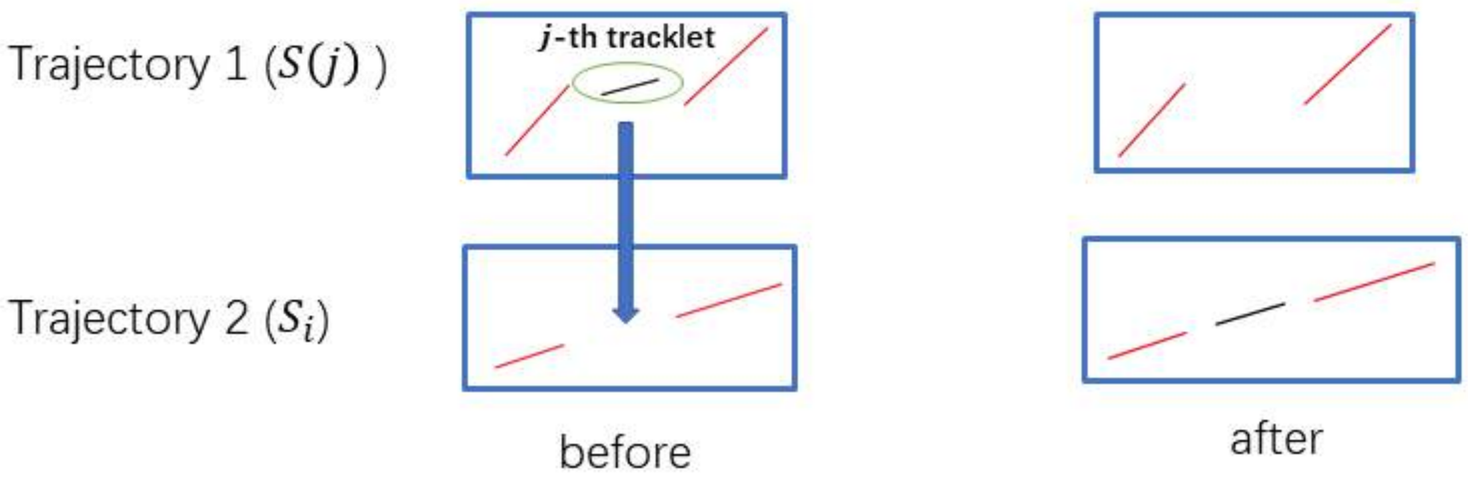
Clustering-based SCT

- Assign operation

$$\Delta l_{j,\text{as}} = \min_i \left(l\left(S(j) \backslash \tau_j\right) + l\left(S_i \cup \tau_j\right) \right) - \left(l\left(S(j)\right) + l\left(S_i\right) \right)$$

Loss after operation Loss before operation

  - τ_j : The tracklet of interest
  - $S(j)$: The trajectory set of τ_j, noted $S(j)$

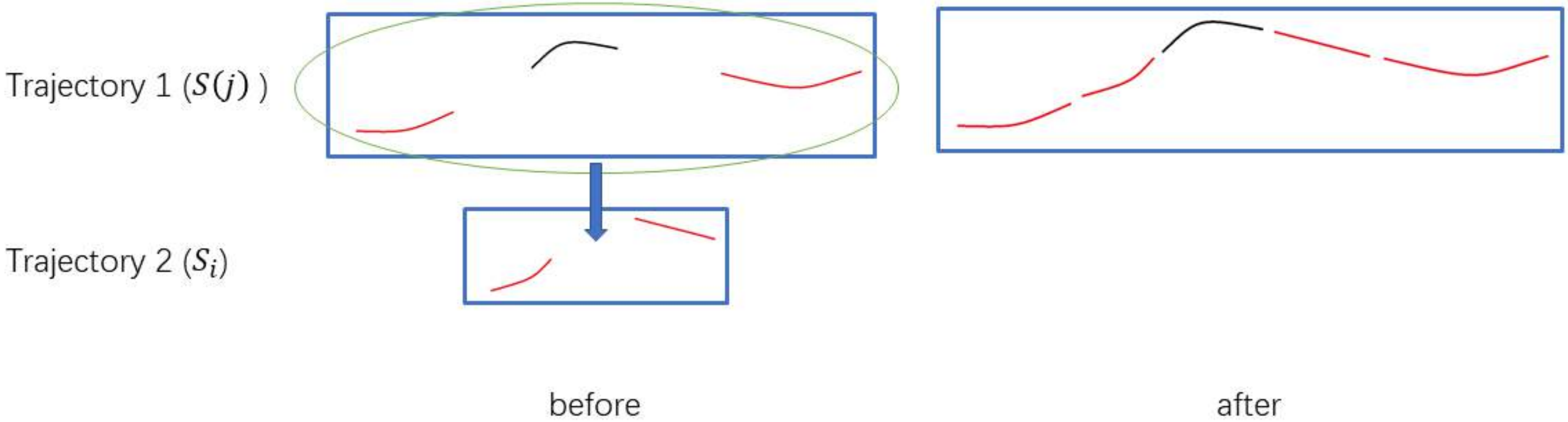


Clustering-based SCT

- Merge operation

$$\Delta l_{j,\mathrm{mg}} = \min_i \left(l(S(j) \cup S_i) \right) - \left(l(S(j)) + l(S_i) \right)$$

Loss after operation Loss before operation



Clustering-based SCT

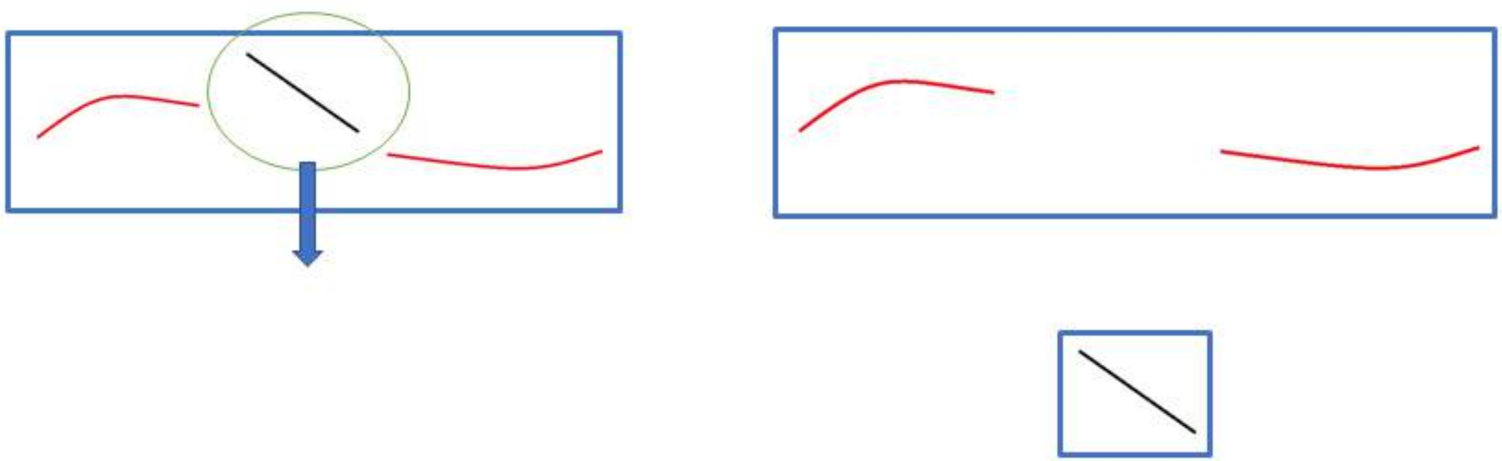
- Split operation

$$\Delta l_{j,\mathrm{sp}} = \left(l(\tau_j) + l(S(j)\backslash\tau_j)\right) - l(S(j))$$

Loss after operation Loss before operation

Trajectory 1 ($S(j)$)

Trajectory 2 (S_i)



before after

Clustering-based SCT

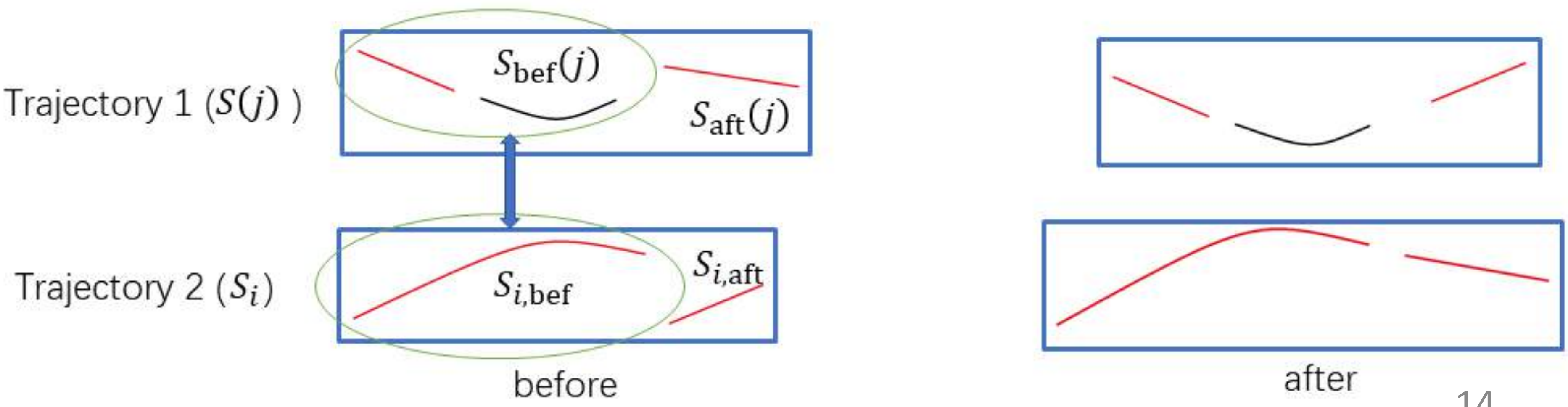
- Switch operation

$$\Delta l_{\mathrm{sw}} = \min_i \left(l\left(S_{\mathrm{bef}}(j) \cup S_{i,\mathrm{aft}}\right) + l\left(S_{\mathrm{aft}}(j) \cup S_{i,\mathrm{bef}}\right) \right) - \left(l\left(S(j)\right) + l(S_i) \right)$$

Loss after operation

Loss before operation

  - $S_{\mathrm{bef}}(j)$: Tracklets before τ_j in $S(j)$
  - $S_{\mathrm{aft}}(j)$: Tracklets after τ_j in $S(j)$

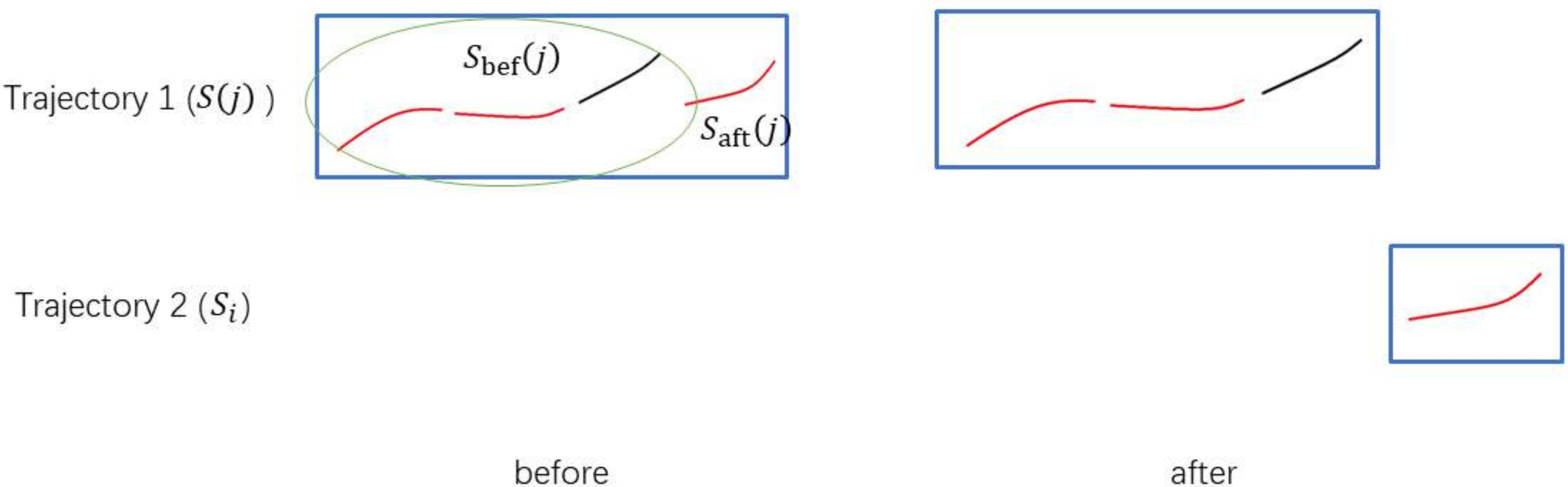


Clustering-based SCT

- Break operation

$$\Delta l_{\mathrm{bk}} = \Big(l\big(S_{\mathrm{bef}}(j)\big) + l\big(S_{\mathrm{aft}}(j)\big)\Big) - l\big(S(j)\big)$$

Loss after operation   Loss before operation



Vehicle Re-identification/ICT

$$L = \sum_{I=1}^{N_{\mathrm{v}}} L_I$$

$$L_I = L_{I,\mathrm{ac}} \times L_{I,\mathrm{nn}} \times L_{I,\mathrm{lp}} \times L_{I,\mathrm{ct}} \times L_{I,\mathrm{tt}}$$

Appearance DCNN License plate Car type Travel time

N_{v}: No. of vehicles appeared in all cameras
L_I: Loss for the I-th vehicle
$L_{I,\mathrm{ac}}$: Appearance change loss
$L_{I,\mathrm{nn}}$: Matching loss of DCNN features
$L_{I,\mathrm{lp}}$: License plate comparison loss
$L_{I,\mathrm{ct}}$: Mis-classified car type loss
$L_{I,\mathrm{tt}}$: Traveling time loss

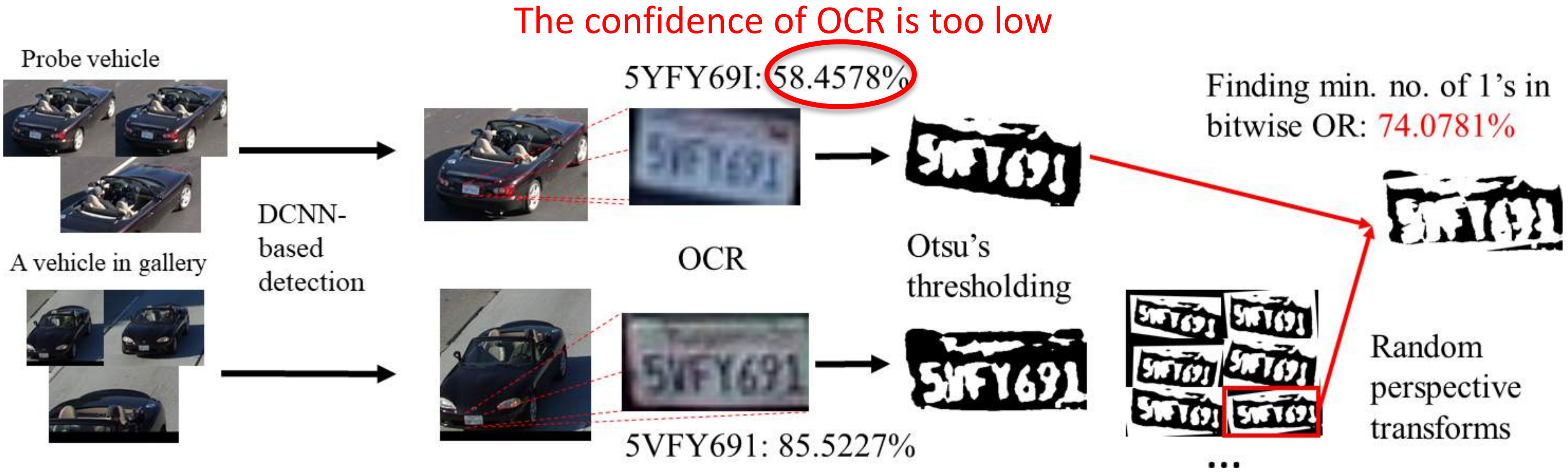
- Appearance change loss
  - (Average) Bhattacharyya distance between each pair of histograms in the adaptive appearance models
- Mis-classified car type loss
  - Different detected categories (majority vote) between vehicles will cause penalty.

Vehicle Re-identification/ICT

- Matching loss of DCNN features
  - Pre-trained model on the Comprehensive Cars (CompCars) dataset
  - 3 images are chosen for each vehicle for feature extraction
  - The dimension of each feature vector is 1024
  - Comparison given by Bhattacharyya distance

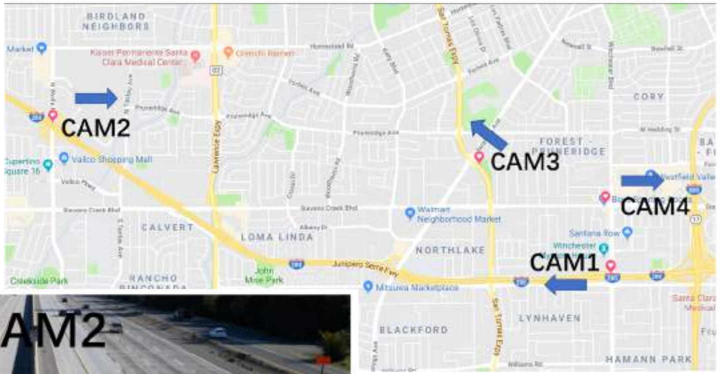
Vehicle Re-identification/ICT

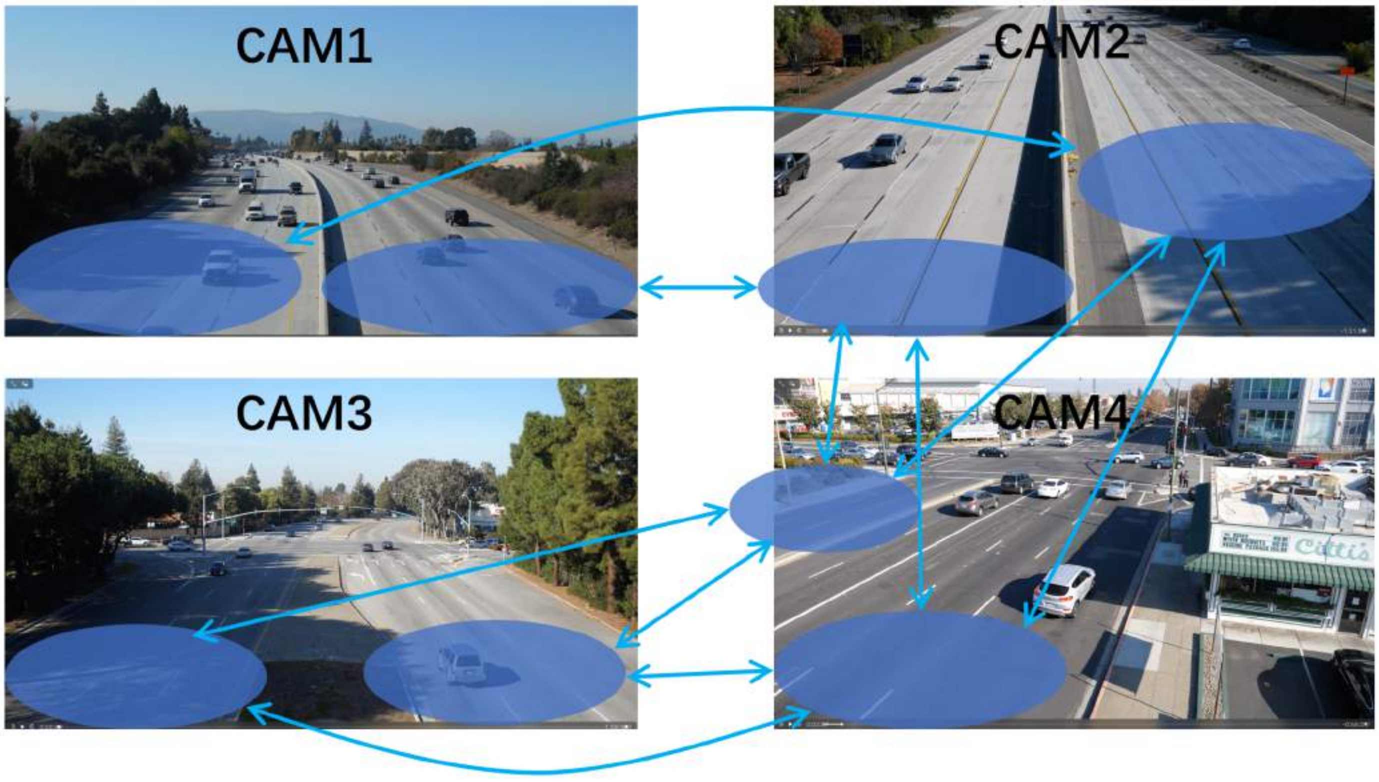
- License plate comparison loss



Vehicle Re-identification/ICT

- Traveling time loss





Experimental Results

- Track 1 - Traffic flow analysis
  - 27 videos, each 1 minute in length, recorded at 30 fps and 1080p resolution
  - Performance evaluation: $S1 = DR \times (1 - NRMSE)$
  - *DR* is the detection rate and *NRMSE* is the normalized Root Mean Square Error (RMSE) of speed
- Track 3 - Multi-camera vehicle detection and re-identification
  - 15 videos, each around 0.5-1.5 hours long, recorded at 30 fps and 1080p resolution
  - Performance evaluation: $S3 = 0.5 \times (TDR + PR)$
  - *TDR* is the trajectory detection rate and *PR* is the localization precision

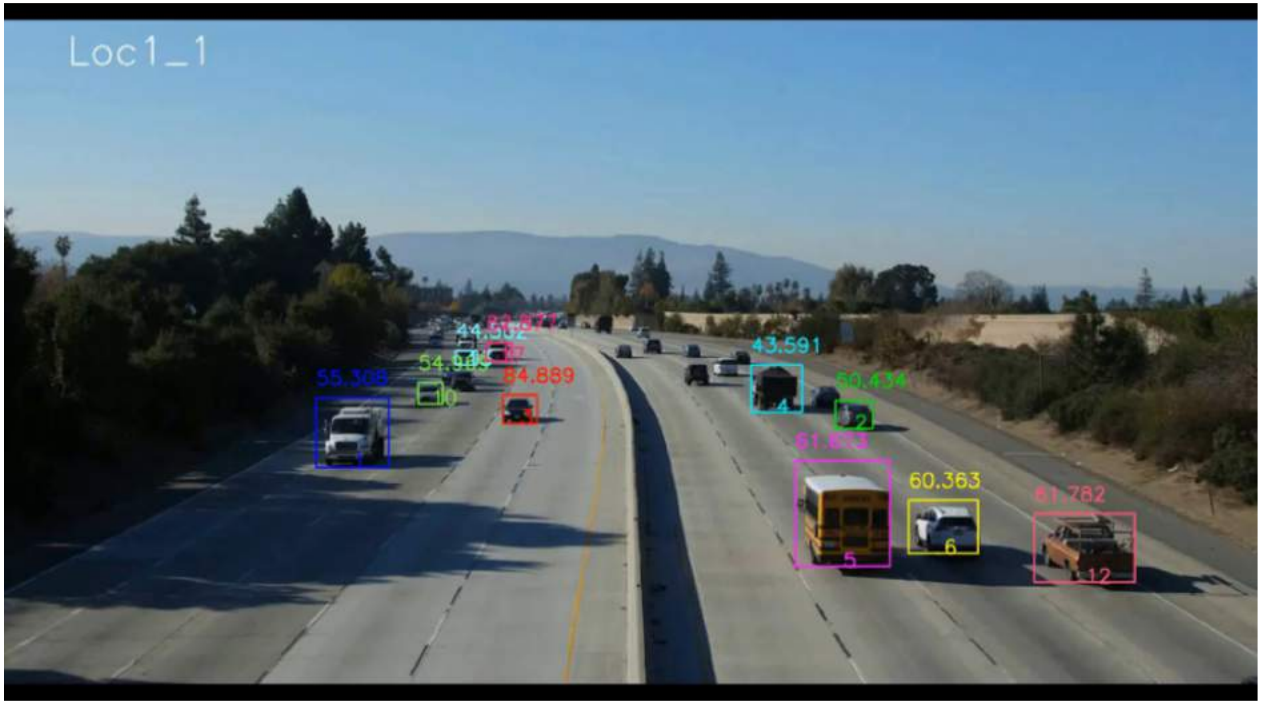
Track 1 Experimental Results

Table 1. Quantitative comparison of speed estimation on the *NVIDIA AI City Dataset* [9]

Rank	Team	S1 Score
1	**team48**	**1.0000**
2	team79	0.9162
3	team78	0.8892
4	team24	0.8813
5	team12	0.8331
6	team4	0.7924
7	team65	0.7654
8	team6	0.7174
9	team40	0.6564
10	team26	0.6547
11	team18	0.6226
12	team45	0.5953
13	team39	0.0000

DR: 1.0000 RMSE: 4.0963 mi/h

https://youtu.be/_i4numqiv7Y



Track 3 Experimental Results

Table 2. Quantitative comparison of multi-camera tracking on the *NVIDIA AI City Dataset* [9]

Rank	Team	S3 Score
1	**team48**	**0.7106**
2	team37	0.2861
3	team79	0.0785
4	team18	0.0074
5	team28	0.0026
6	team41	0.0024
7	team53	0.0002
8	team6	0.0001
9	team10	0.0000
10	team31	0.0000

TDR: 3/7 PR: 0.9925

https://youtu.be/Jlvh_KxHl40

Conclusion

- Fusion of visual and semantic features for SCT: motion, temporal and appearance attributes
- Fusion of visual and semantic features for ICT: appearance, license plate, vehicle type and temporal attributes
- Adaptive appearance model to robustly encode long-term appearance change
- Camera calibration based on EDA optimization for reliable 2D-to-3D backprojection
- Top performance in both Track 1 & Track 3 on the challenge dataset
- GitHub: https://github.com/zhengthomastang/2018AICity_TeamUW