City-Scale Multi-Camera Vehicle Tracking by Semantic Attribute Parsing and Cross-Camera Tracklet Matching

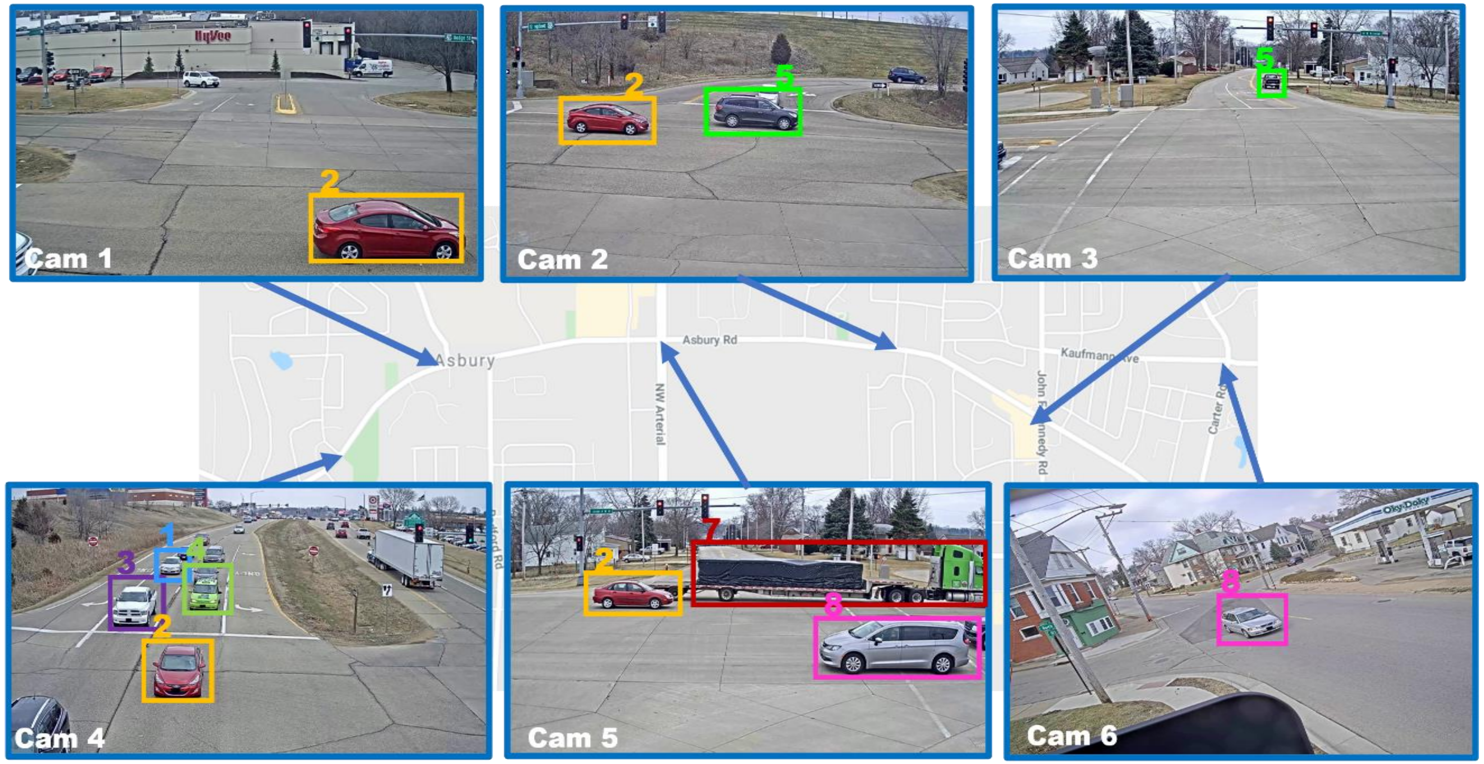
AI City Challenge 2020,
CVPR Workshop, Seattle

Yuhang He, Jie Han, Wentao Yu, Xiaopeng Hong*, Xing Wei, and Yihong Gong
Xi'an Jiaotong University (XTJU), P.R.C.

1. Introduction

1.1 Background

Goal : Track **multiple vehicles** in a city-scale **multi-camera** network.



Applications:

- Traffic flow management;
- Vehicle behavior analysis;
- Traffic anomaly detection;
- Auto-driving assistant;
- …

1. Introduction

1.2 Main Challenge

- Generate local tracklet in each camera (Single-camera multi-object tracking)
  - Object occlusion;
  - Background clutter;
  - Target interaction;
  - Targets enter and exit
  - …
- Cross-camera tracklet matching:
  - Visual appearance variations caused by different viewpoints;
  - The unknown target status caused by blind areas
  - The Occurrences of targets are different and unknown.
  - …

3. Methodology

3.1 Algorithm Overview

Algorithm 1 Tracking algorithm of the proposed method

Input: Image sequences collected from M cameras $\mathcal{I} = \{\mathcal{I}_1, \mathcal{I}_2, ..., \mathcal{I}_M\}$.

Output: Global trajectory set $\mathcal{G} = \{\mathcal{G}_1, \mathcal{G}_2, ..., \mathcal{G}_M\}$.

for camera $i = 1 : M$ **do**
    Generate local tracklet set $\mathcal{T}_i$ using single camera multi-object tracking technique.
end for
Generate robust representation $\mathbf{f}_i^j$ of each tracklet $\mathcal{T}_i^j$ using Eq. (9) and prune infeasible matching candidates by traffic topology reasoning.
Construct tracklet similarity matrix $\mathbf{S}$ using Eq. (10).
Compute tracklet-to-target assignment matrix $\mathbf{A}^*$ by optimizing Eq. (11).
Generate global trajectory set $\mathcal{G}$ according to $\mathbf{A}^*$.

Local tracklet generation

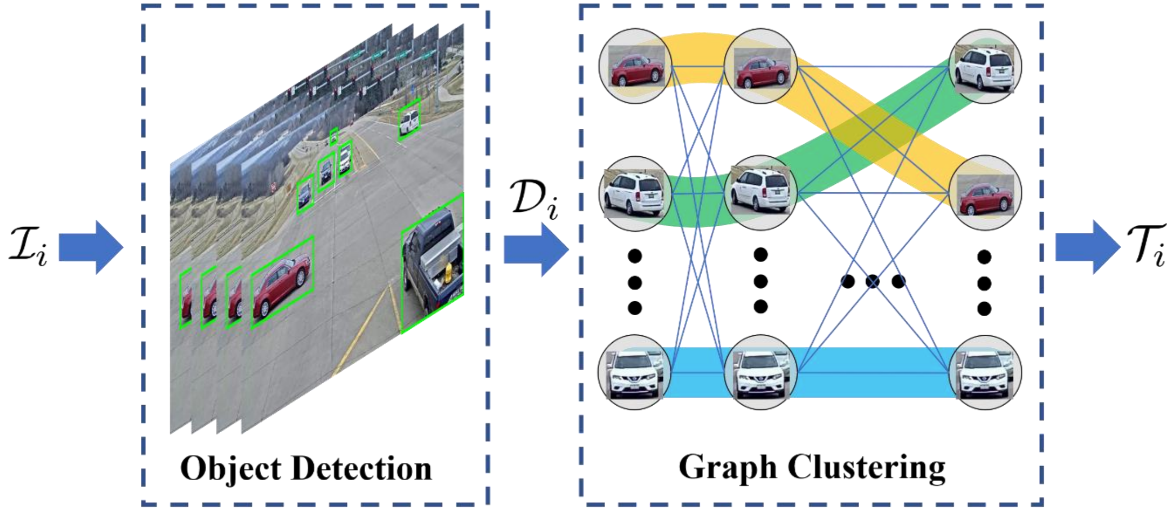
Semantic attribute parsing

Cross-camera tracklet matching

Tracklet-to-Target assignment

3. Methodology

3.2 Local Tracklet Generation



1. Detect target with an object detector

- Getting the detection collection $\mathcal{D}_i$ of the whole image sequence $\mathcal{I}_i$ of camera i.

2. Link detections into local tracklets by graph clustering

- Construct a graph model G = (V, E) based on $\mathcal{D}_i$.

$$w_{x,y} = \psi(v_x, v_y)$$

- Get the local tracklet collection $\mathcal{T}_i$ of camera i by graph clustering

3. Methodology

3.3 Semantic Attribute Parsing

1. Robust Tracklet Representation — Feature Extractor $\varphi(\cdot)$

- **Backbone:** ResNet50
- **Cross-entropy loss:** identity classification

$$y_i'(u) = \begin{cases} 1 - \frac{H-1}{H}\eta & \text{if } y_i(u) = 1, \\ \frac{\eta}{H} & \text{otherwise,} \end{cases} \qquad L'_{xent}(\mathbf{I}_i) = -\sum_{u=1}^{H} log(p_i(u)) \cdot y_i'(u).$$

- **Triplet loss:** metric learning

$$L_{trip}(\mathcal{I}_i) = \left[\|\varphi(\mathbf{I}_i^a) - \varphi(\mathbf{I}_i^p)\|_2 - \|\varphi(\mathbf{I}_i^a) - \varphi(\mathbf{I}_i^n)\|_2 + m\right]_+$$

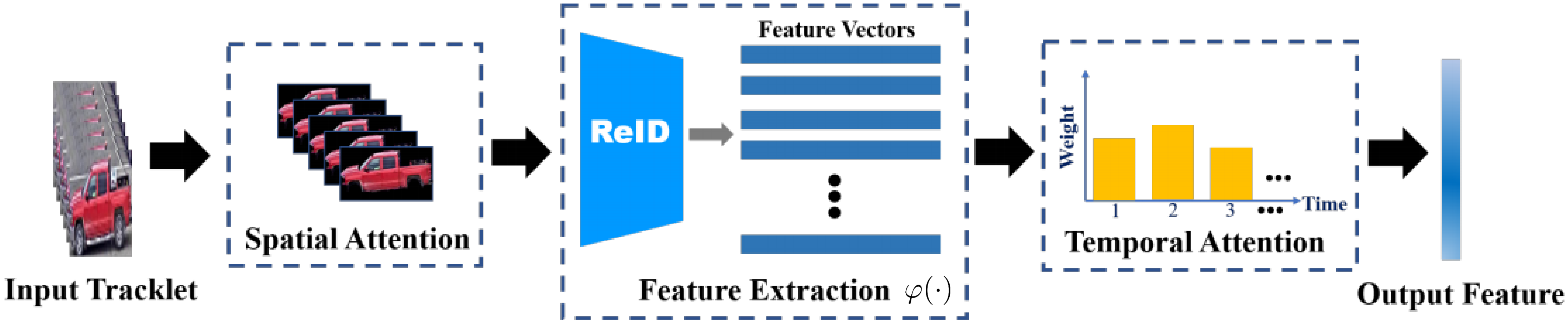
- **Overall objective function:**

$$L = \sum_{i=1}^{O} L'_{xent}(\mathbf{I}_i) + \lambda L_{trip}(\mathcal{I}_i)$$

3. Methodology

3.3 Semantic Attribute Parsing

1. Robust Tracklet Representation — Spatial-Temporal Attention



- **Spatial attention mechanism:** $$\mathbf{C}_{i,t}^{j*} = \mathbf{C}_{i,t}^{j} \odot \mathbf{M}_{i,t}^{j}$$
- **Temporal attention mechanism:** $$w_{i,t}^{j} = \frac{||\mathbf{M}_{i,t}^{j}||_2}{\sum_{t \in \pi_i^j} ||\mathbf{M}_{i,t}^{j}||_2}$$
- **Robust feature representation:** $$\mathbf{f}_i^j = \sum_{t \in \pi_i^j} \varphi(\mathbf{C}_{i,t}^{j*}) \cdot w_{i,t}^{j}$$

3. Methodology

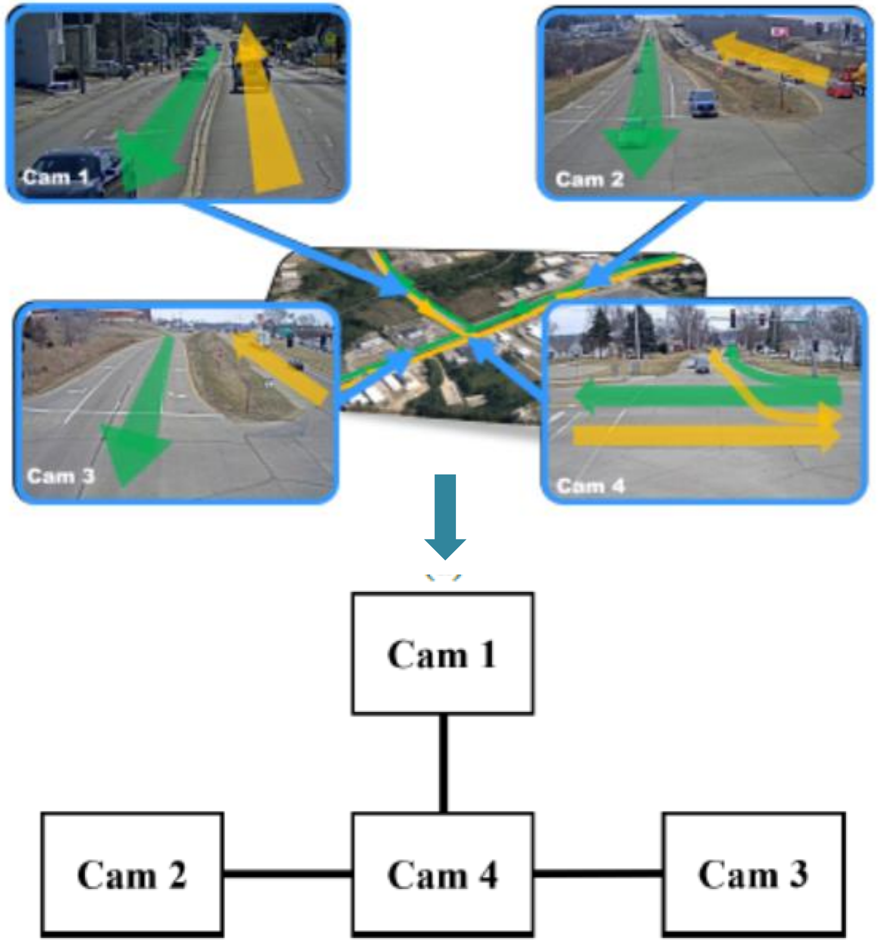
3.3 Semantic Attribute Parsing

2. Traffic Topology Reasoning

Objective:

- Prune infeasible matching candidates.

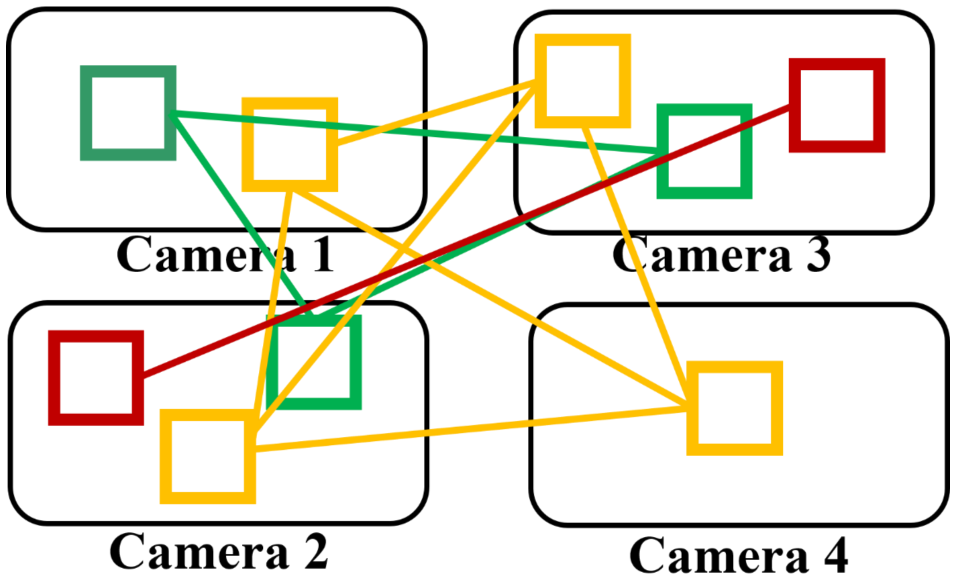
$$\delta(\mathcal{T}_i^u, \mathcal{T}_j^v) = \begin{cases} 1 & \text{if } \mathcal{T}_i^u, \mathcal{T}_j^v \text{ traffic connected,} \\ 0 & \text{otherwise,} \end{cases}$$



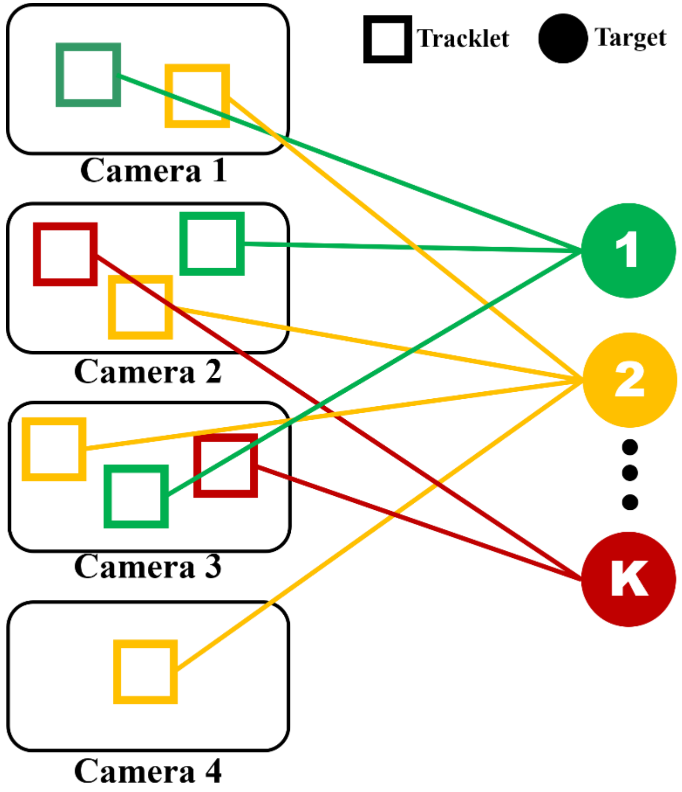
Traffic Topology

3.4 Tracklet-to-Target Assignment [2]

1. Illustration of tracklet-to-target assignment



Tracklet-to-Tracklet Matching



Tracklet-to-Target Assignment

He, Y., Wei, X., Hong, X., Shi, W., & Gong, Y. (2020). Multi-Target Multi-Camera Tracking by Tracklet-to-Target Assignment. *IEEE Transactions on Image Processing*, *29*, 5191-5205.

3. Methodology

3.5 Tracklet-to-Target Assignment

2. Advantages

- Smaller solution space. $N \times N \rightarrow N \times K$

- Determined assignment relationship:

  Each tracklet should be assigned to a target

- Matching consistency naturally satisfied.

3.4 Tracklet-to-Target Assignment

3. Method formulation

Tracklet-Tracklet Similarity Matrix:

$$\mathbf{S} \in [0, 1]^{N \times N}$$

Tracklet-Target Assignment Matrix:

$$\mathbf{A} \in \{0, 1\}^{N \times K}$$

3. Methodology

3.4 Tracklet-to-Target Assignment

3. Method formulation

Intuition:

$$\begin{cases} \mathbf{S}(u,v) \to 1 \Rightarrow \mathbf{A}(u,:)\mathbf{A}(v,:)^T = 1 \\ \mathbf{S}(u,v) \to 0 \Rightarrow \mathbf{A}(u,:)\mathbf{A}(v,:)^T = 0 \end{cases}$$

$$\Rightarrow \mathbf{A}\mathbf{A}^T \to \mathbf{S}$$

Objective Function:

$$\mathbf{A}^* = \underset{\mathbf{A}}{\operatorname{argmin}} \|\mathbf{S} - \mathbf{A}\mathbf{A}^T\|_2,$$
$$\text{s.t } \mathbf{A}\mathbf{I}_1 = \mathbf{I}_2,$$

3. Methodology

3.4 Tracklet-to-Target Assignment

4. Restricted non-negative matrix factorization

Relax constraint:

$$\mathbf{A}'^{*} = \operatorname*{argmin}_{\mathbf{A}' \geq 0} ||\mathbf{S} - \mathbf{A}'\mathbf{A}'^{T}||^2 + \alpha||\mathbf{A}'\mathbf{1}_1 - \mathbf{1}_2||^2,$$

Updating Rule:

$$\mathbf{A}' \leftarrow \mathbf{A}' \odot \text{sqrt}([4\mathbf{S}\mathbf{A}' + 2\alpha\mathbf{1}_2\mathbf{1}_1^T] \oslash [4\mathbf{A}'\mathbf{A}'^{T}\mathbf{A}' + 2\alpha\mathbf{A}'\mathbf{1}_1\mathbf{1}_1^T]),$$

4. Experimental Results

Main results

Rank	Team ID	IDF1 (%)
1	92	45.85
2	**11**	**44.00**
3	63	34.83
4	111	34.11
5	72	12.48
6	75	6.20
7	30	4.52
8	31	3.87

The proposed method achieves the second-best result and significantly outperforms most of the competitive methods by a large margin

4. Experimental Results

Ablation study

Method	IDF1 (%)	IDP (%)	IDR (%)
baseline	31.28	23.29	35.12
baseline+ST	34.51	29.54	41.50
baseline+ST+TT	38.61	47.19	32.80
baseline+ST+TT+TRACTA	44.00	53.63	37.31

- Different components are all effective for multi-camera tracking results
- Using all components achieves up to 12.72% on IDF1 than baseline

Conclusion

Main Contributions

- An efficient two-step MTMCT method for city-scale multi-camera vehicle tracking
- The semantic attribute parsing for tracklet affinity measurement
- A spatial-temporal attention mechanism to generate a robust representation for each tracklet.