Visual Semantic Search: Retrieving Videos via Complex Textual Queries
[Lin et al]

CSC2523 Winter 2015: Paper Presentation

Micha Livne

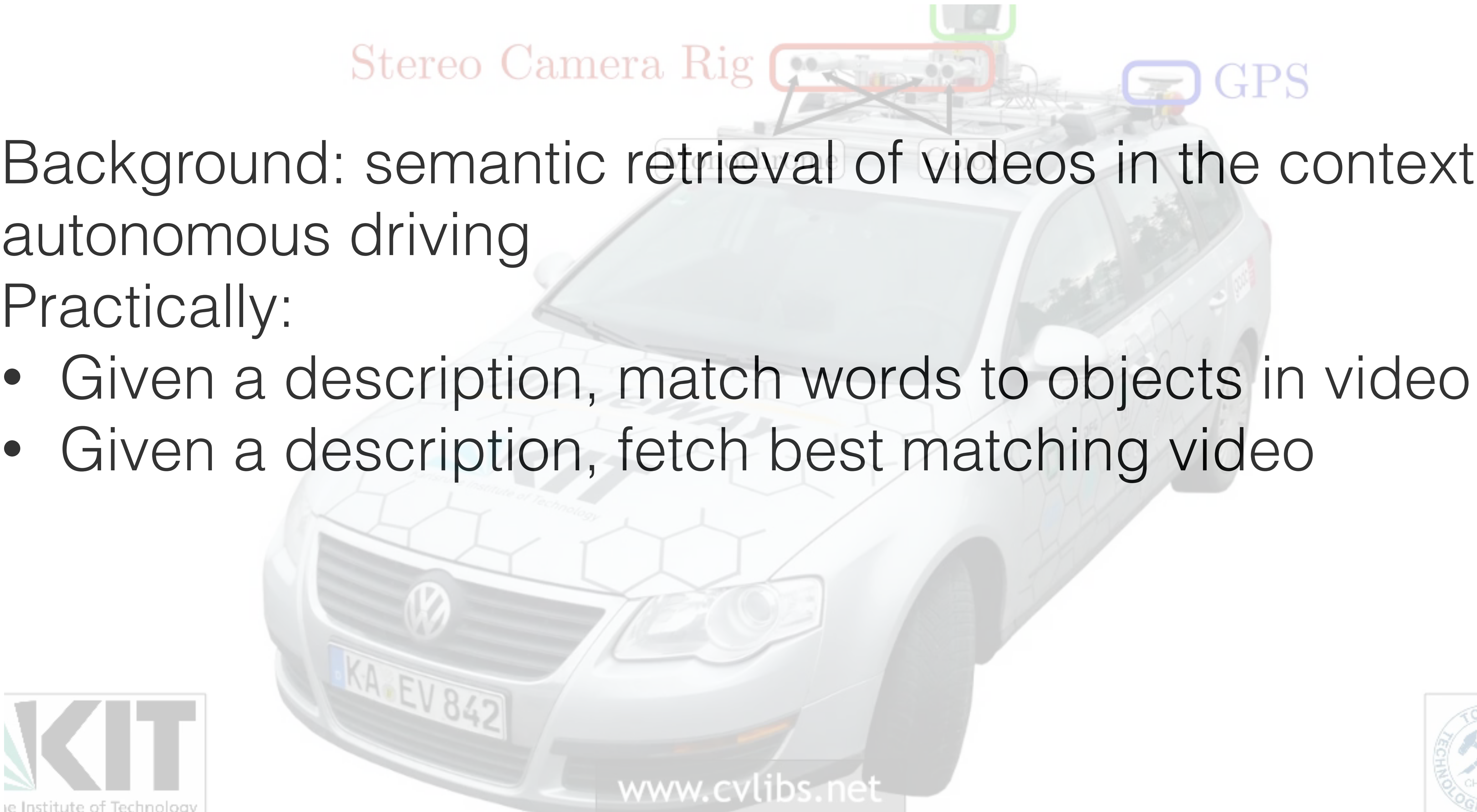
Goals

Goals

- Background: semantic retrieval of videos in the context of autonomous driving

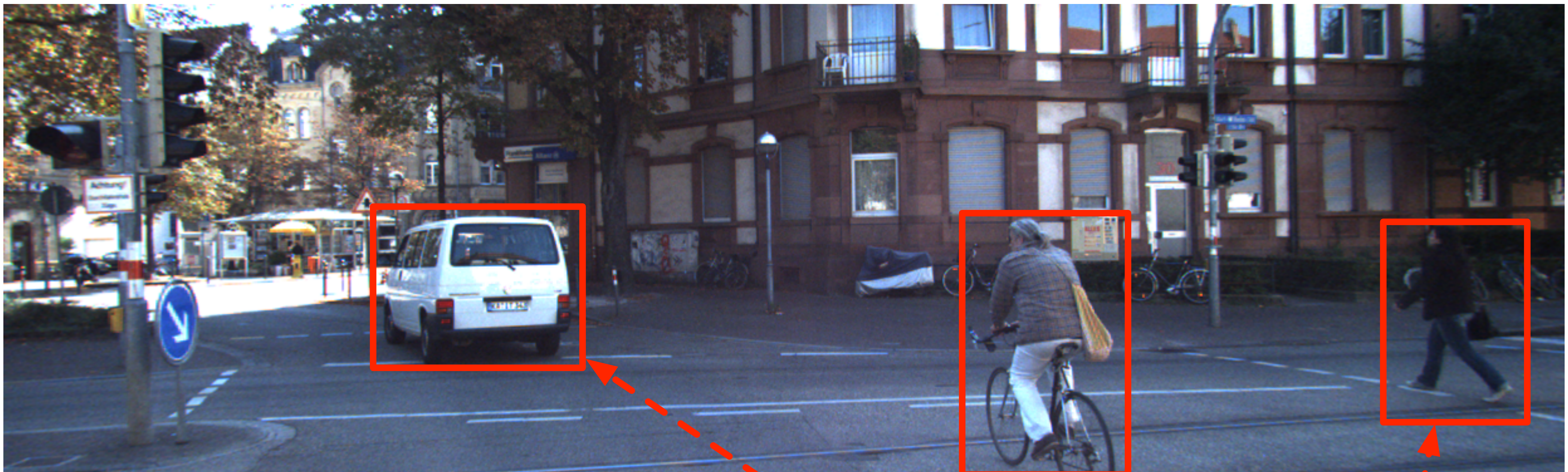
Goals

- Background: semantic retrieval of videos in the context of autonomous driving
- Practically:
  - Given a description, match words to objects in video
  - Given a description, fetch best matching video

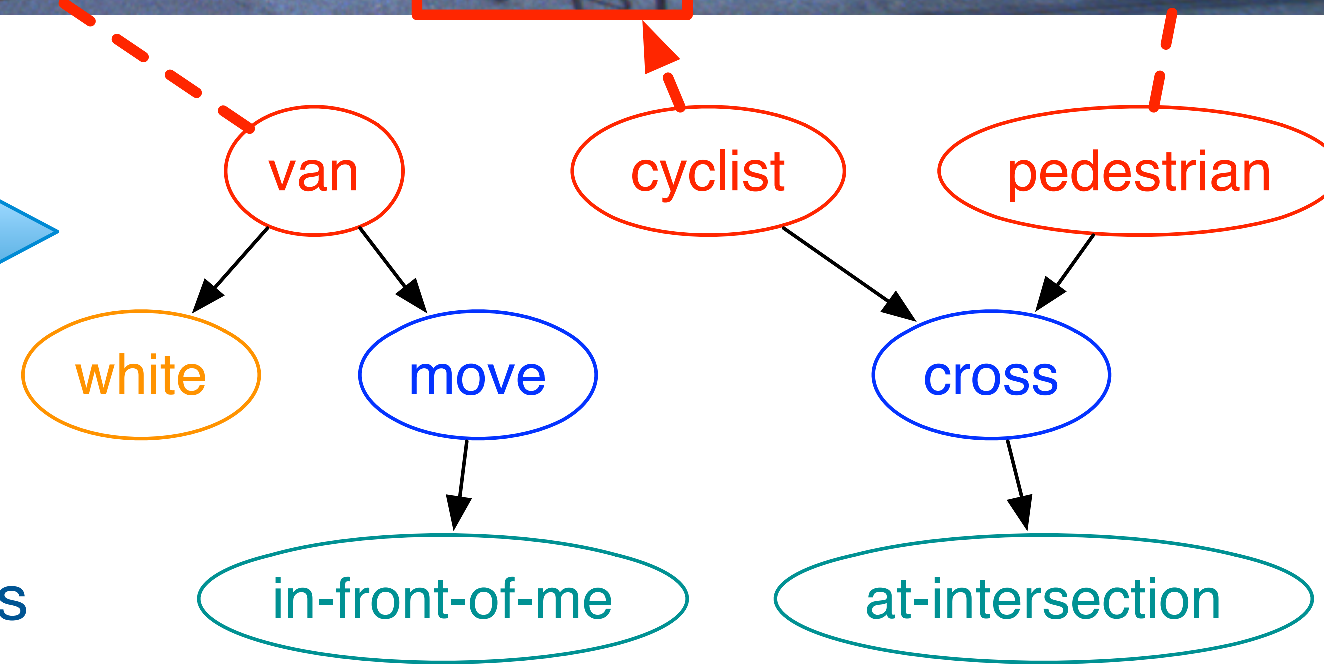


Goals

Goals

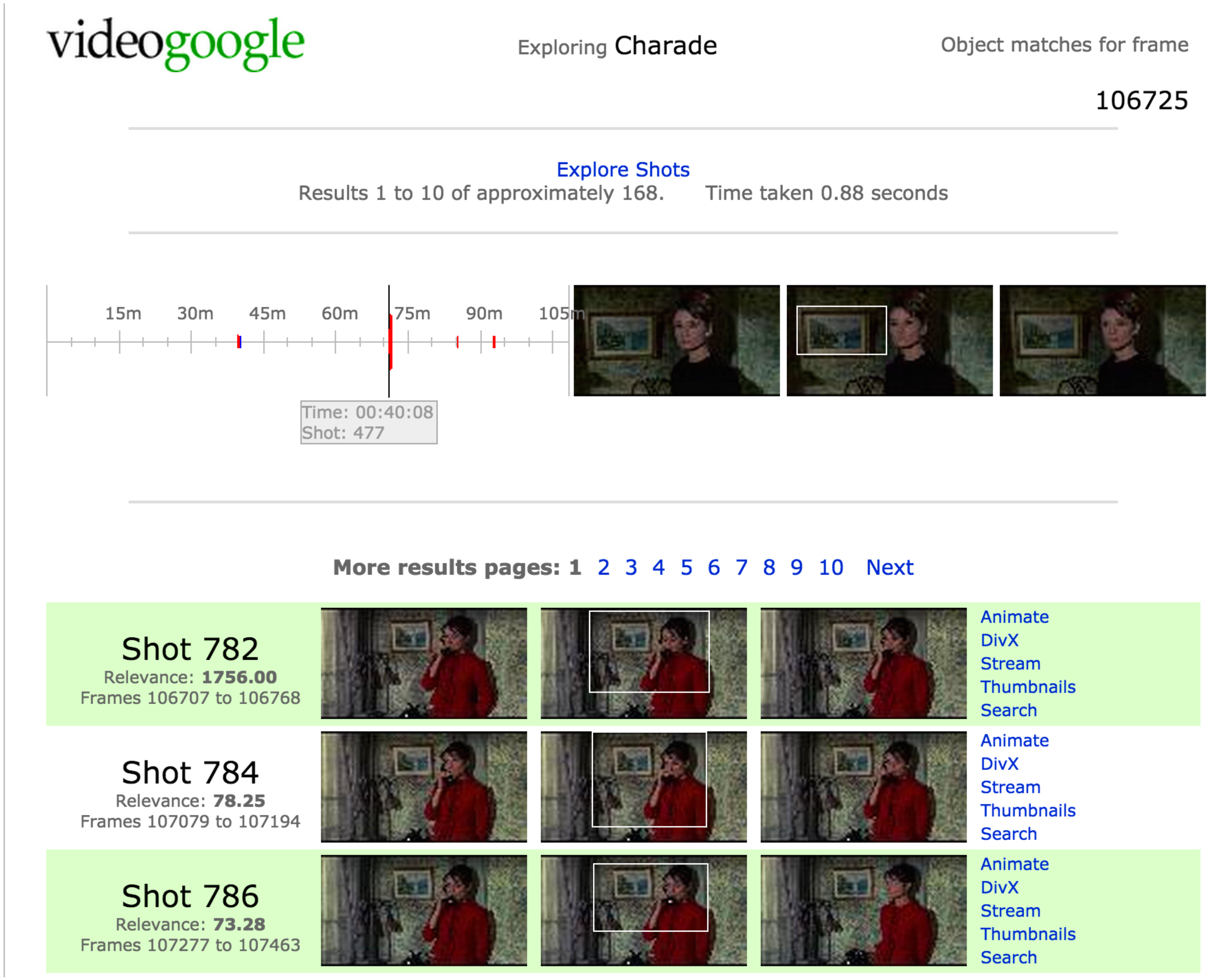


A white van is moving in front of me, while a cyclist and a pedestrian is crossing the intersection.



semantic graphs

Related Work



[Sivic and Zisserman, ’03]

Dataset

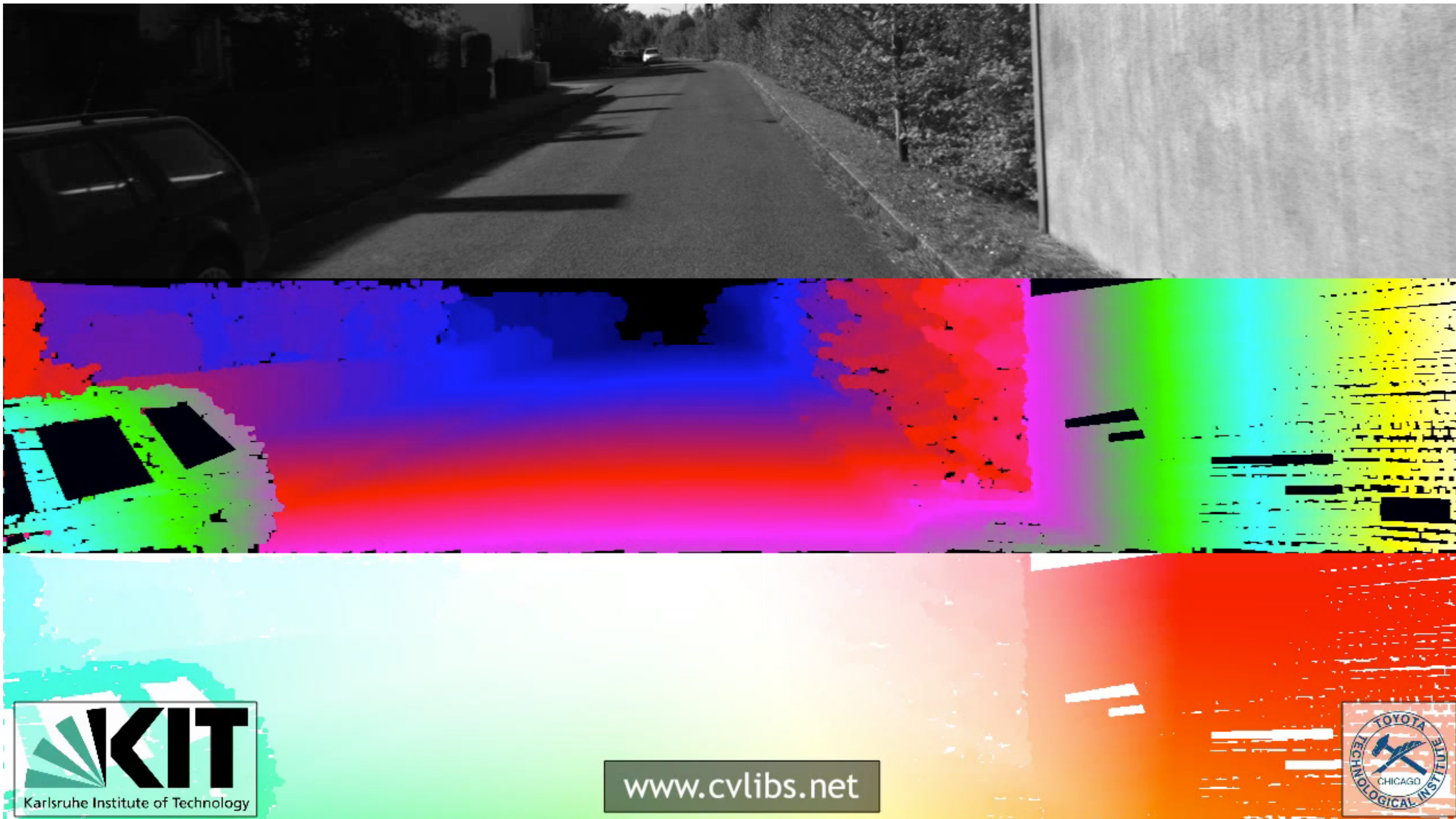
KITTI dataset [Geiger et al ‘12]

Dataset

➡ This paper adds text descriptions to parts of KITTI videos

KITTI dataset [Geiger et al '12]

Dataset

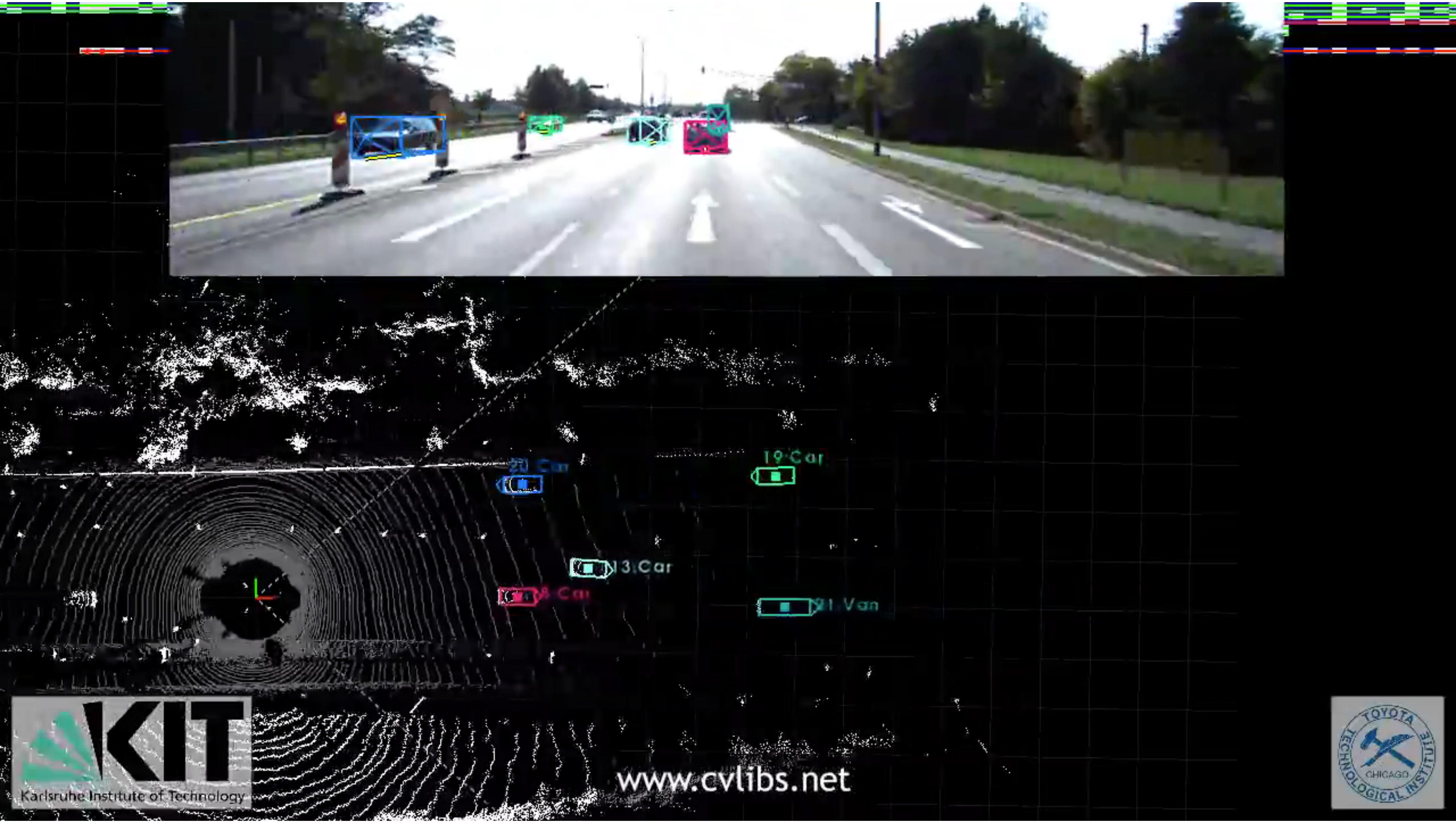


Karlsruhe Institute of Technology
www.cvlibs.net
TOYOTA TECHNOLOGICAL INSTITUTE
CHICAGO

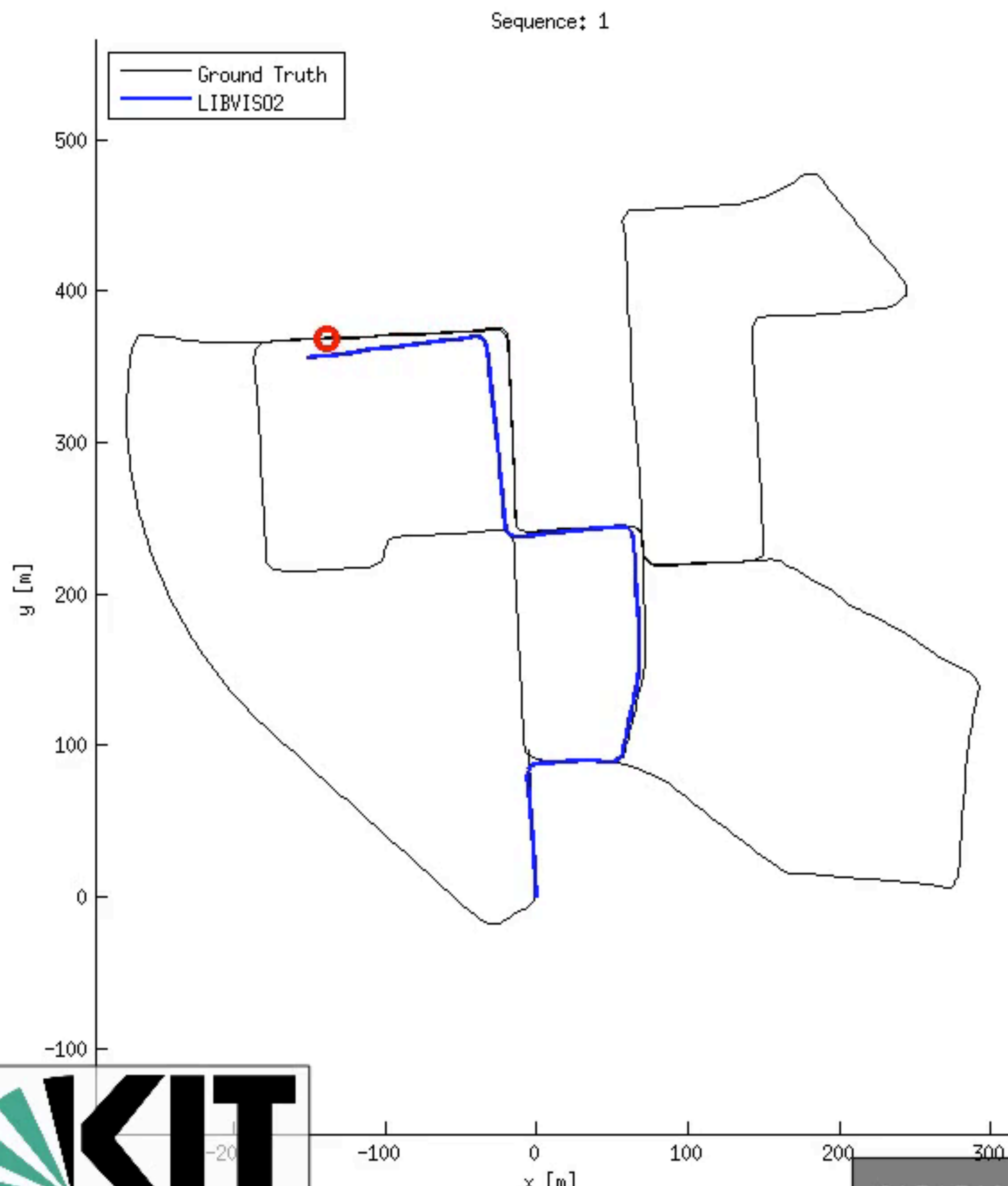
Dataset

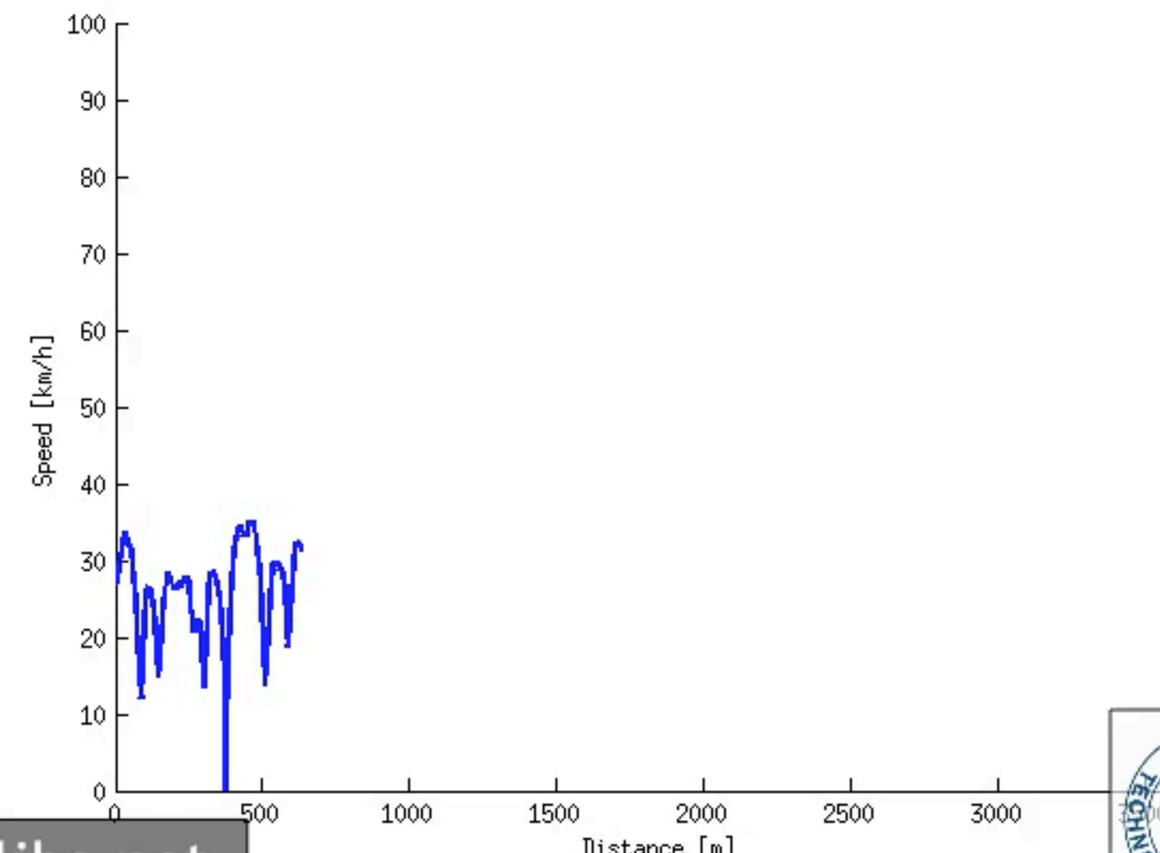

Sequence: 1
Ground Truth
LIBVISO2
y [m]
x [m]
Speed [km/h]
Distance [m]
www.cvlibs.net
KIT
Karlsruhe Institute of Technology
TOYOTA TECHNOLOGICAL INSTITUTE
CHICAGO

Dataset


19 Car
13 Car
Van
Karlsruhe Institute of Technology
www.cvlibs.net
TOYOTA TECHNOLOGICAL INSTITUTE
CHICAGO

Dataset





Karlsruhe Institute of Technology

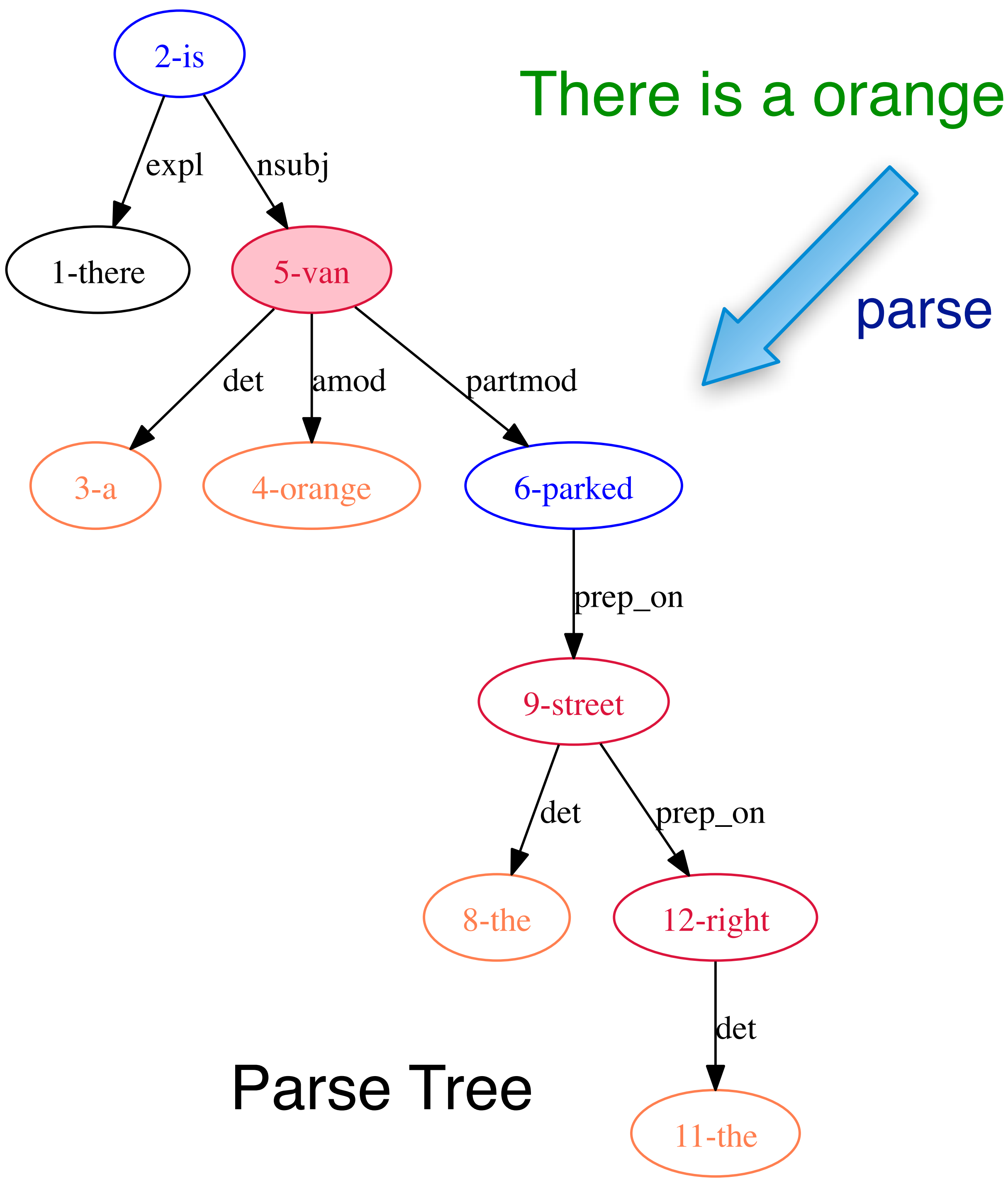
www.cvlibs.net

Dataset

Sequence: 1
Ground Truth
LIBVISO2
y [m]
x [m]
Speed [km/h]
Distance [m]
www.cvlibs.net
KIT
Karlsruhe Institute of Technology
TOYOTA TECHNOLOGICAL INSTITUTE
CHICAGO

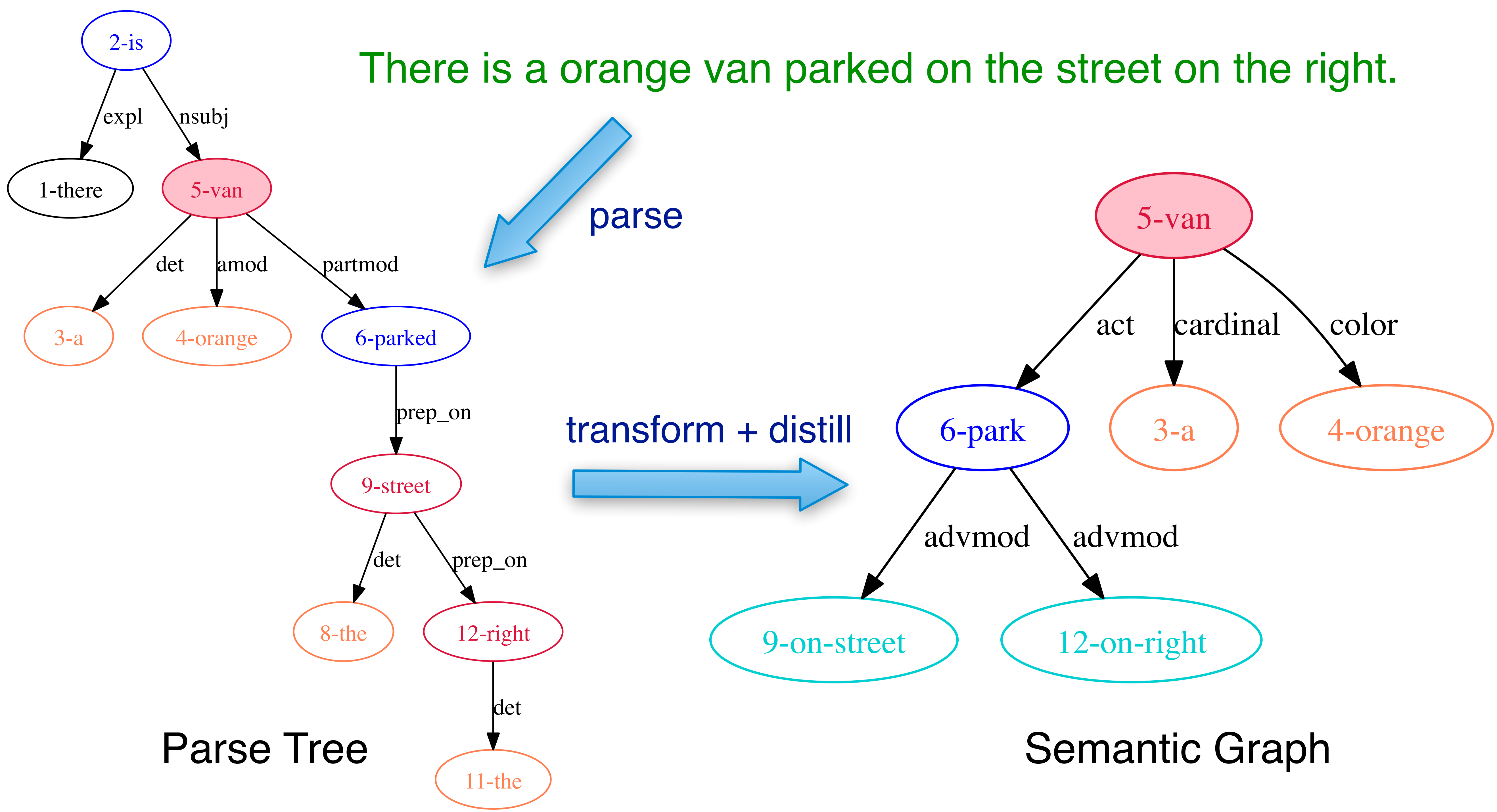
Proposed Solution

There is a orange van parked on the street on the right.



Parse Tree

Proposed Solution



Proposed Solution

Matching Text and Video Segments

$$\max_{\mathbf{y}} \sum_{uv} h_{uv} y_{uv} \tag{1}$$

Proposed Solution

Matching Text and Video Segments

$$\max_{\mathbf{y}} \sum_{uv} h_{uv} y_{uv} \tag{1}$$

$$\text{s.t.} \quad \sum_{v} y_{uv} = s_u, \quad \forall u = 1, \ldots, m$$

Proposed Solution

Matching Text and Video Segments

$$
\begin{aligned}
\max_{\mathbf{y}} \quad & \sum_{uv} h_{uv} y_{uv} \\
\text{s.t.} \quad & \sum_{v} y_{uv} = s_u, \quad \forall u = 1, \ldots, m \\
& \sum_{u} y_{uv} \leq t_v, \quad \forall v = 1, \ldots, n
\end{aligned} \tag{1}
$$

Proposed Solution

Matching Text and Video Segments

$$
\begin{aligned}
\max_{\mathbf{y}} \quad & \sum_{uv} h_{uv} y_{uv} \\
\text{s.t.} \quad & \sum_{v} y_{uv} = s_u, \quad \forall u = 1, \ldots, m \\
& \sum_{u} y_{uv} \leq t_v, \quad \forall v = 1, \ldots, n \\
& 0 \leq y_{uv} \leq 1, \quad \forall u = 1, \ldots, m, \;\; v = 1, \ldots, n-1
\end{aligned} \tag{1}
$$

Proposed Solution

Matching Text and Video Segments

$$\begin{aligned}\max_{\mathbf{y}} \quad & \sum_{uv} h_{uv} y_{uv} \\ \text{s.t.} \quad & \sum_{v} y_{uv} = s_u, \quad \forall u = 1, \ldots, m \\ & \sum_{u} y_{uv} \leq t_v, \quad \forall v = 1, \ldots, n \\ & 0 \leq y_{uv} \leq 1, \quad \forall u = 1, \ldots, m, \;\; v = 1, \ldots, n-1 \end{aligned} \tag{1}$$

$$h_{uv} = \sum_{k=1}^{K} w_k f_{uv}^{(k)} = \mathbf{w}^T \mathbf{f}_{uv}. \tag{2}$$

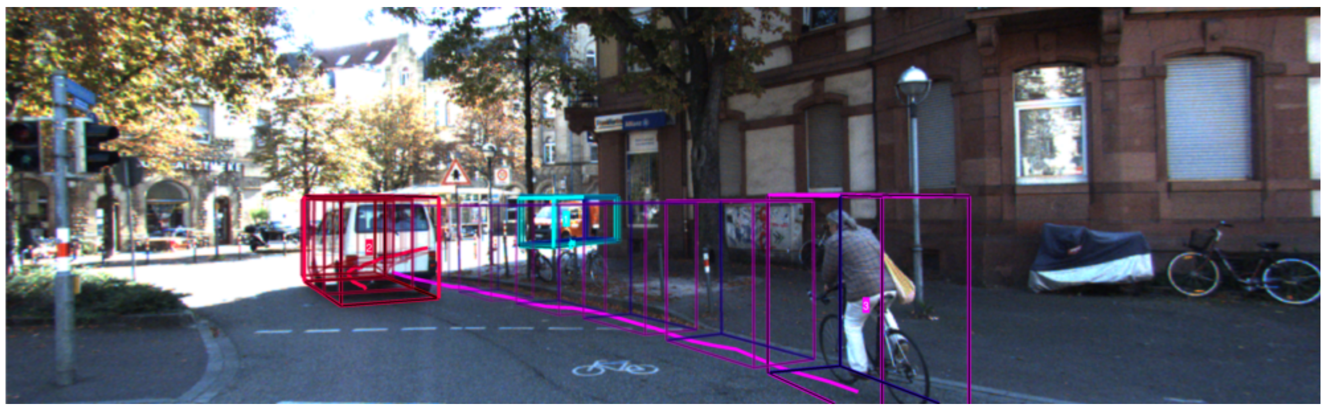
Proposed Solution

Learning

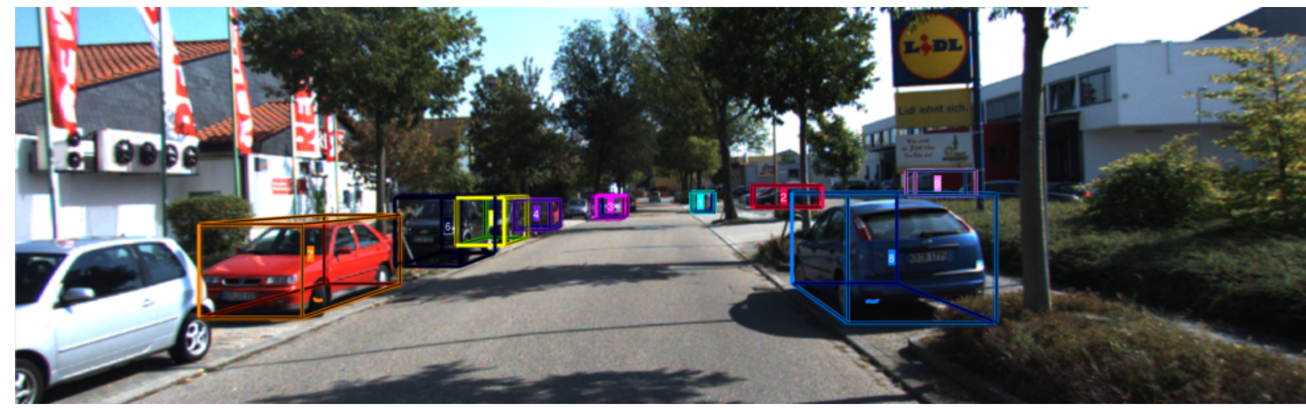
$$\min_{\xi, \mathbf{w}} \ \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i \tag{3}$$
$$\text{s.t.} \quad \xi_i \geq \mathbf{w}^T(\boldsymbol{\phi}_i(\mathbf{y}) - \boldsymbol{\phi}_i(\mathbf{y}^{(i)})) + \Delta(\mathbf{y}, \mathbf{y}^{(i)}), \ \ \forall \mathbf{y} \in \mathcal{Y}^{(i)}$$
$$\xi_i \geq 0, \ \ \forall i = 1, \dots, N.$$

$$\boldsymbol{\phi}_i(\mathbf{y}) = [\phi_i^{(1)}(\mathbf{y}), \dots, \phi_i^{(K)}(\mathbf{y})], \text{ with } \phi_i^{(k)} = \sum_{uv} f_{uv}^{(ik)} y_{uv}$$

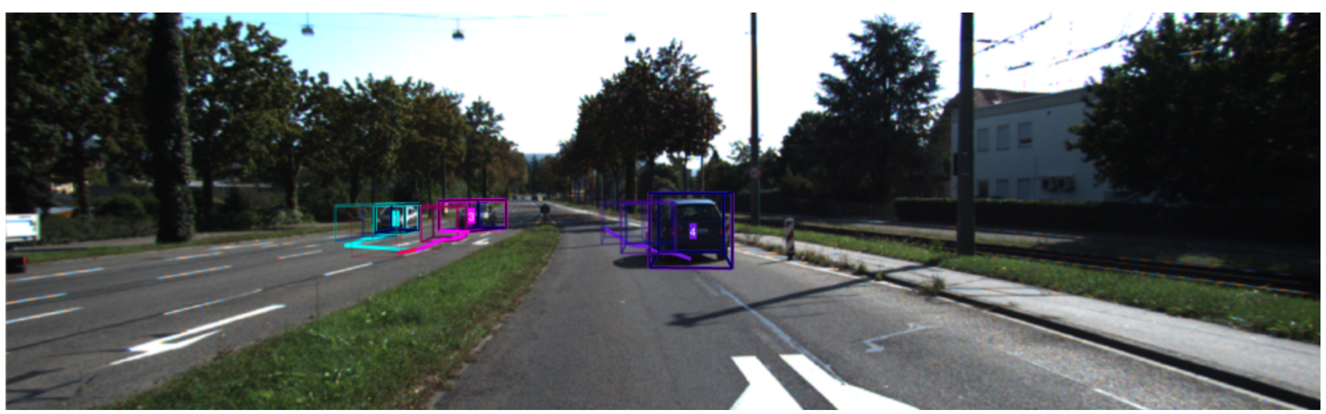
Results



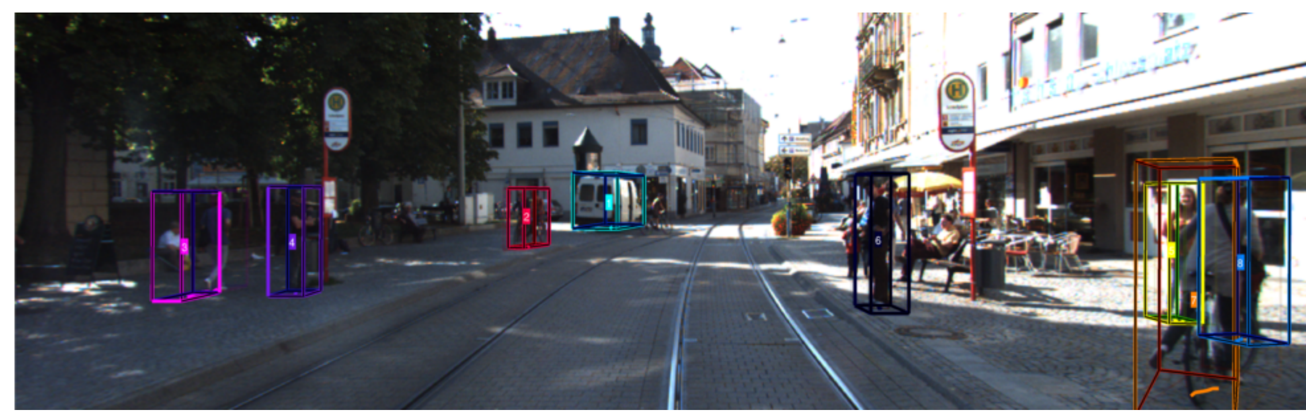
A **bicyclist** is biking on the road, to the right of my car.
A **white van** is driving at safe distance in front of me.



There are multiple **cars** parked on the left side of the street and one blue **car** parked on the right side of the street.

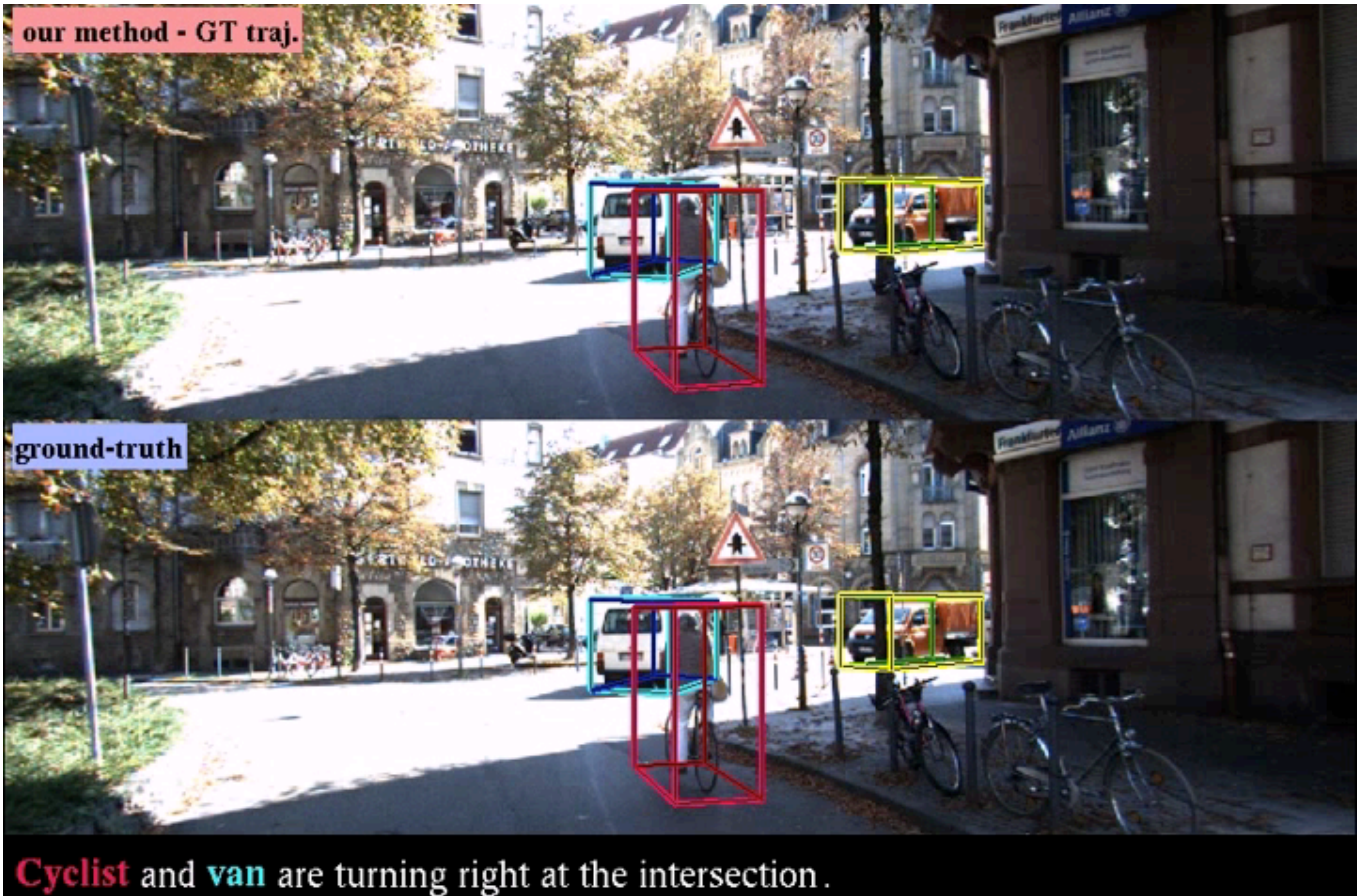


There is **a car** in front of us.
A couple of cars are in the opposite street.

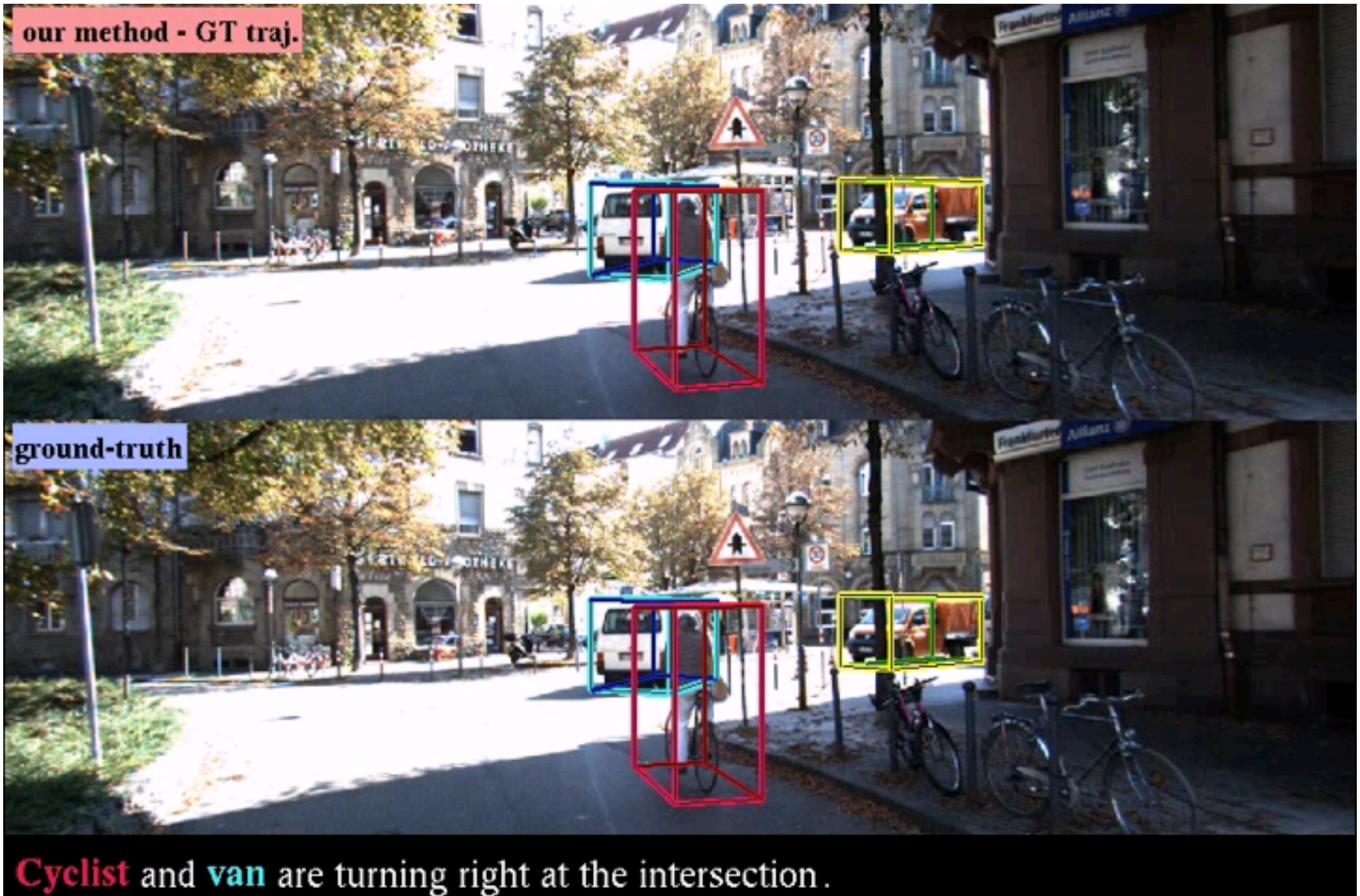


Some people are sitting and **some pedestrians** are on right sidewalk.
Some pedestrians on left sidewalk, and **a van** is parked.
And I see **a cyclist**.

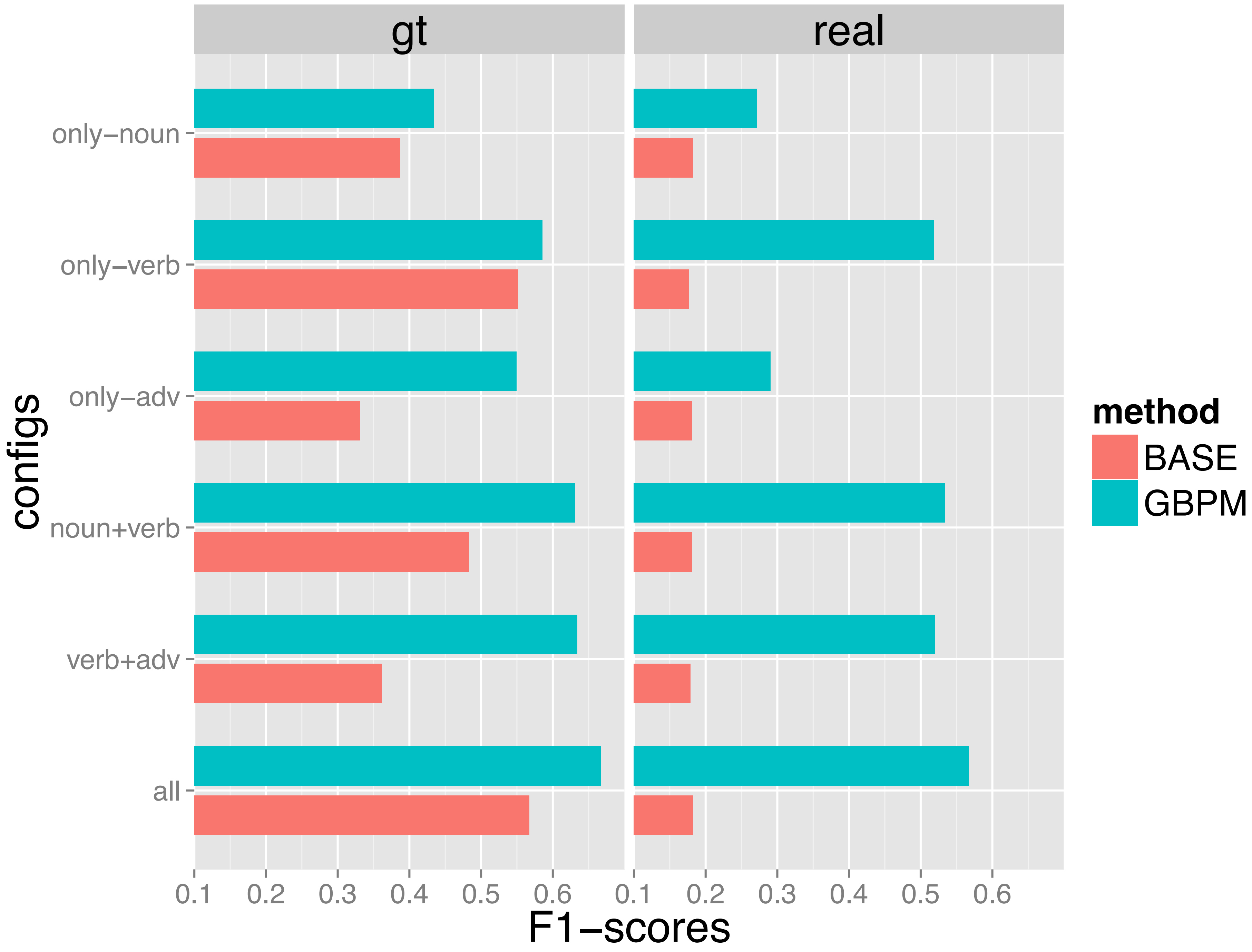
Results


our method - GT traj.
ground-truth
Cyclist and van are turning right at the intersection.

Results


our method - GT traj.
ground-truth
Cyclist and van are turning right at the intersection.

Results



gt
real
only−noun
only−verb
only−adv
noun+verb
verb+adv
all
configs
0.1
0.2
0.3
0.4
0.5
0.6
F1−scores
method
BASE
GBPM

Results

		BASE						REAL					
		noun	verb	adv	n.+v.	v.+a.	all	noun	verb	adv	n.+v.	v.+a.	all
GT	recall	.8777	.5897	.2170	.6884	.2485	.6726	.4379	.5700	.5562	.6391	.6430	.6765
	prec.	.2483	.5182	.7006	.3721	.6632	.4906	.4302	.6021	.5434	.6243	.6257	.6583
	F1	.3871	.5517	.3313	.4830	.3615	.5674	.4340	.5856	.5497	.6316	.6342	.6673
real	recall	.5301	.5137	.5246	.5246	.5191	.5301	.3251	.4563	.3497	.5328	.4754	.5710
	prec.	.1102	.1068	.1091	.1091	.1080	.1102	.2333	.6007	.2485	.5357	.5743	.5633
	F1	.1825	.1769	.1806	.1806	.1787	.1825	.2717	.5186	.2906	.5342	.5202	.5672

Results

	K	rand	noun	verb	adv	n.+v.	v.+a.	all
GT	1	.0397	.0613	.0873	.0967	.1061	.1274	.1486
	2	.0794	.1250	.1533	.1651	.1910	.2288	.2335
	3	.1191	.1840	.2052	.2217	.2712	.3160	.3467
	5	.1985	.3042	.3443	.3514	.4057	.4481	.4693
real	1	.0425	.0755	.0566	.0889	.0836	.1078	.0943
	2	.0849	.1375	.1132	.1321	.1429	.1698	.1779
	3	.1274	.1914	.1752	.1698	.2022	.2264	.2399
	5	.2123	.2722	.2857	.2722	.3181	.3342	.3208

Table 3. Average hit rates of video segment retrieval.

Results

	K	rand	noun	verb	adv	n.+v.	v.+a.	all
GT	1	.0397	.0613	.0873	.0967	.1061	.1274	.1486
	2	.0794	.1250	.1533	.1651	.1910	.2288	.2335
	3	.1191	.1840	.2052	.2217	.2712	.3160	.3467
	5	.1985	.3042	.3443	.3514	.4057	.4481	.4693
real	1	.0425	.0755	.0566	.0889	.0836	.1078	.0943
	2	.0849	.1375	.1132	.1321	.1429	.1698	.1779
	3	.1274	.1914	.1752	.1698	.2022	.2264	.2399
	5	.2123	.2722	.2857	.2722	.3181	.3342	.3208

Table 3. Average hit rates of video segment retrieval.

	K	rand	noun	verb	adv	n.+v.	v.+a.	all
GT	1	.1673	.2571	.3029	.2800	.3286	.3429	.3629
	2	.1673	.2686	.2771	.2600	.3400	.3386	.3557
	3	.1673	.2790	.2714	.2610	.3410	.3267	.3533
	5	.1673	.2749	.2640	.2589	.3280	.3109	.3383
real	1	.1673	.2680	.2484	.2876	.2810	.2941	.2941
	2	.1673	.2647	.2304	.2484	.2843	.2680	.2908
	3	.1673	.2702	.2462	.2495	.2898	.2800	.3017
	5	.1673	.2686	.2444	.2477	.2784	.2758	.2869

Table 4. Average relevance of video segment retrieval.

Point of Strength

Point of Strength

- Efficient learning procedure (simplified learning).
- Robustness to tracking errors.
- Free-form complex language queries.

Point of Weakness

Point of Weakness

- Features extraction (preprocessing) might be slow to compute (e.g., visual scores).
- Features are engineered - learned features could improve results.

Contributions

Contributions

- Matching individual words in the query to specific objects, as opposed to find a video given a query.
- Collected a new dataset for semantic retrieval.
- Developed a new framework for semantic video search.

Conclusion

Conclusion

- We are getting closer to “real” AI, as perceived by most people.
- The proposed method is heading exactly that way.
- Interesting and a hard problem, with proposed method demonstrating effectiveness.

Thanks!

Thanks!

Questions?