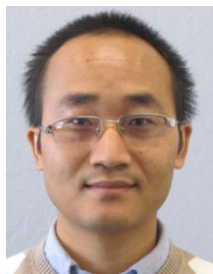


Building Blocks for Visual 3D Scene Understanding towards Autonomous Driving

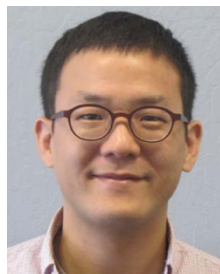
Media Analytics,
NEC Labs America



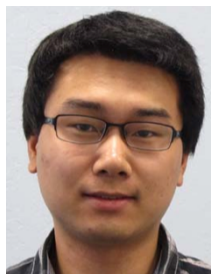
Manmohan
Chandraker



Xiaoyu
Wang



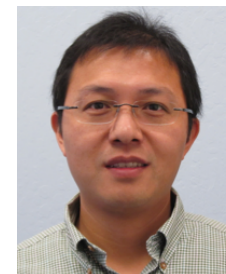
Wongun
Choi



Shiyu
Song



Shiliang
Zhang

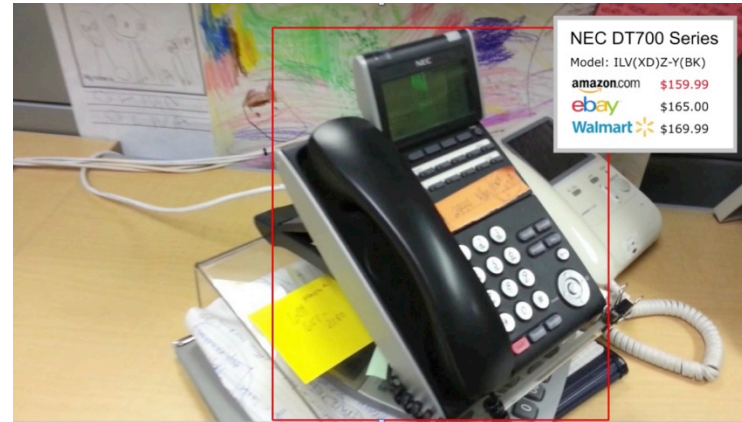
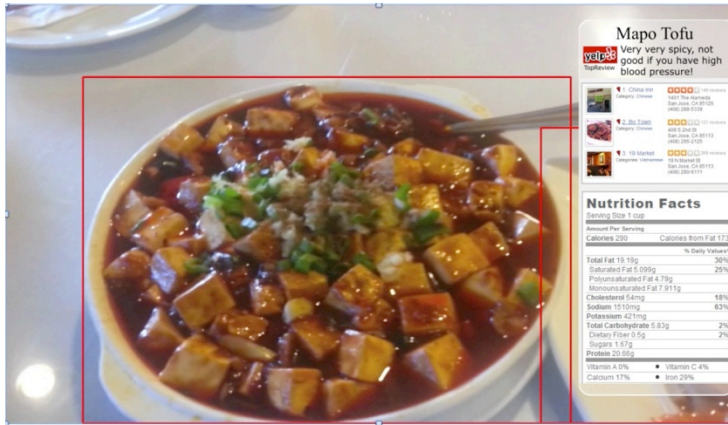


Yuanqing
Lin

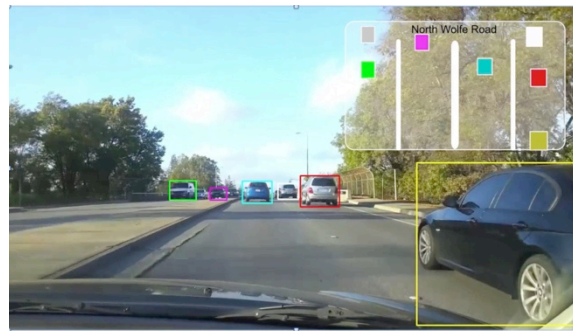


An overview of research directions in our group

- **Image recognition:** recognize things of interest on a mobile-cloud platform -- up to fine-grained identity information



- **Visual 3D scene understanding** – for example, for autonomous driving



- 3D dense reconstruction

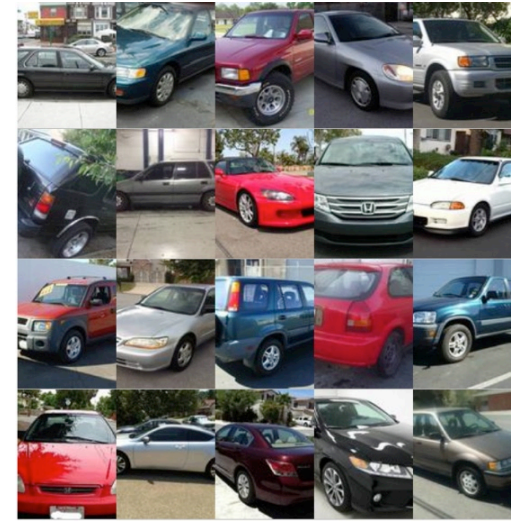
A couple of more words -- our research on image recognition



Recognizing >1000 types of flowers on a company's catalog. An iPhone app on this is coming to App store in one week.



Recognizing as "which restaurant which dish". As the first batch, covering 10 restaurants around Cupertino.



Is this a "Honda Accord Sedan 2010"? Covering all models/years from Nissan, Honda, Toyota, Ford and Chevrolet since 1990

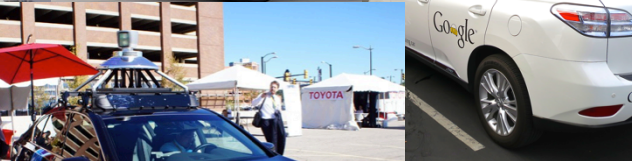
- Amazon's Firefly recognizes book covers, CD covers, bar codes. **We target for more generic objects.**
- "Very deep" into each vertical domain, but with research focus on **generic recognition algorithms.**
- More: all Toy"r"us toys, faces, scene texts, shoes, ...

Image recognition -- research portfolio

- **Metric learning**
 - Very fast algorithm for high-dimension large-scale data
- **Deep learning**
 - State-of-the-art systems, research to tailor it for fine-grained image recognition
- **Boosting**
 - Another way for supervised feature learning
- **Object detection (object centric pooling)**
 - To overcome clutter background
- We are building very rich research portfolio – aiming for **the best way** to solve the fine-grained image recognition problem.
- It is a very fun direction to work on – **things are moving so fast!**

Building Blocks for Visual 3D Scene Understanding towards Autonomous Driving

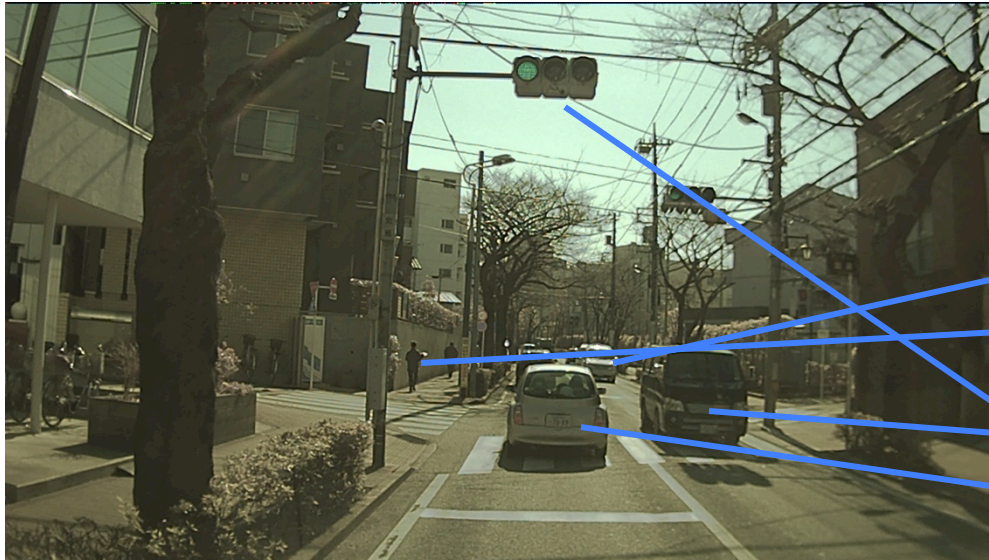
Autonomous driving – a big new trend for the automobile industry



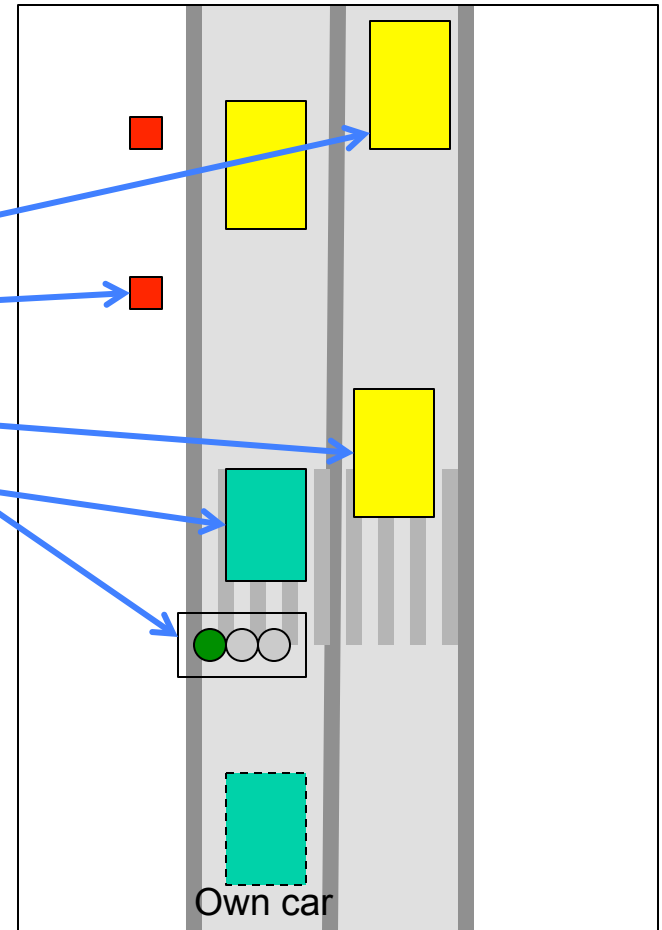
- Autonomous driving: we only focus on sensing → **visual sensing**, or we call it **visual 3D scene understanding**

Visual 3D scene understanding

From: video frames

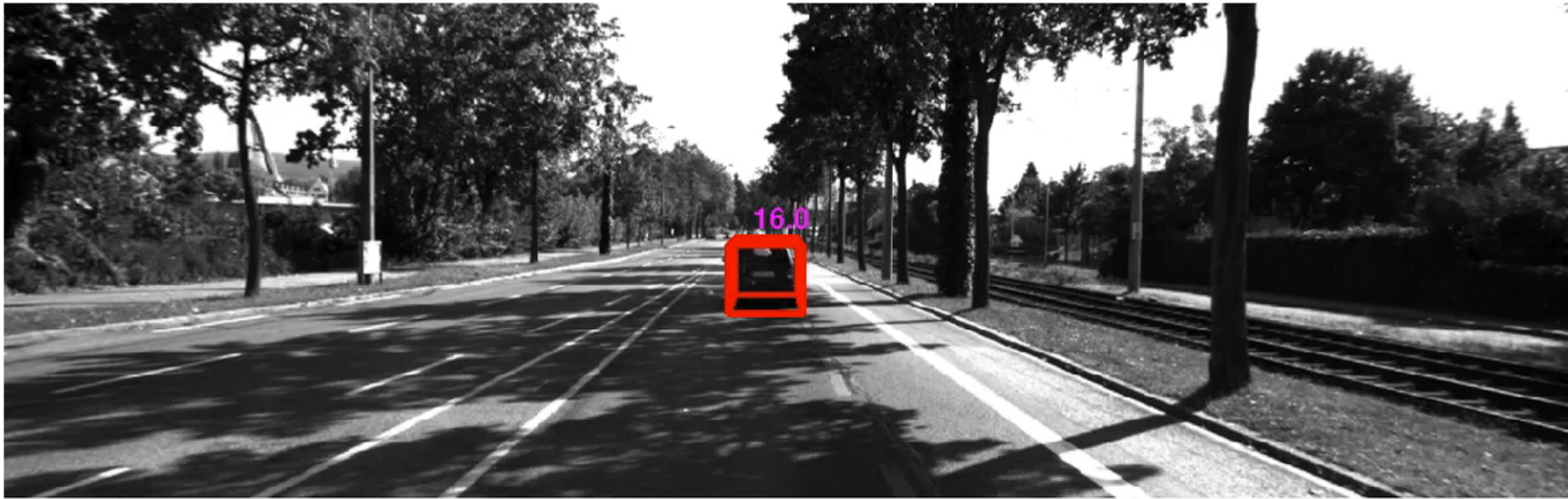


Output: 3D localization of objects with scene consistency



Visual 3D driving scene understanding: for sensing the driving environments.

Visual 3D scene understanding (3D object localization for this demo)



Our group is focused on a monocular system



LIDAR



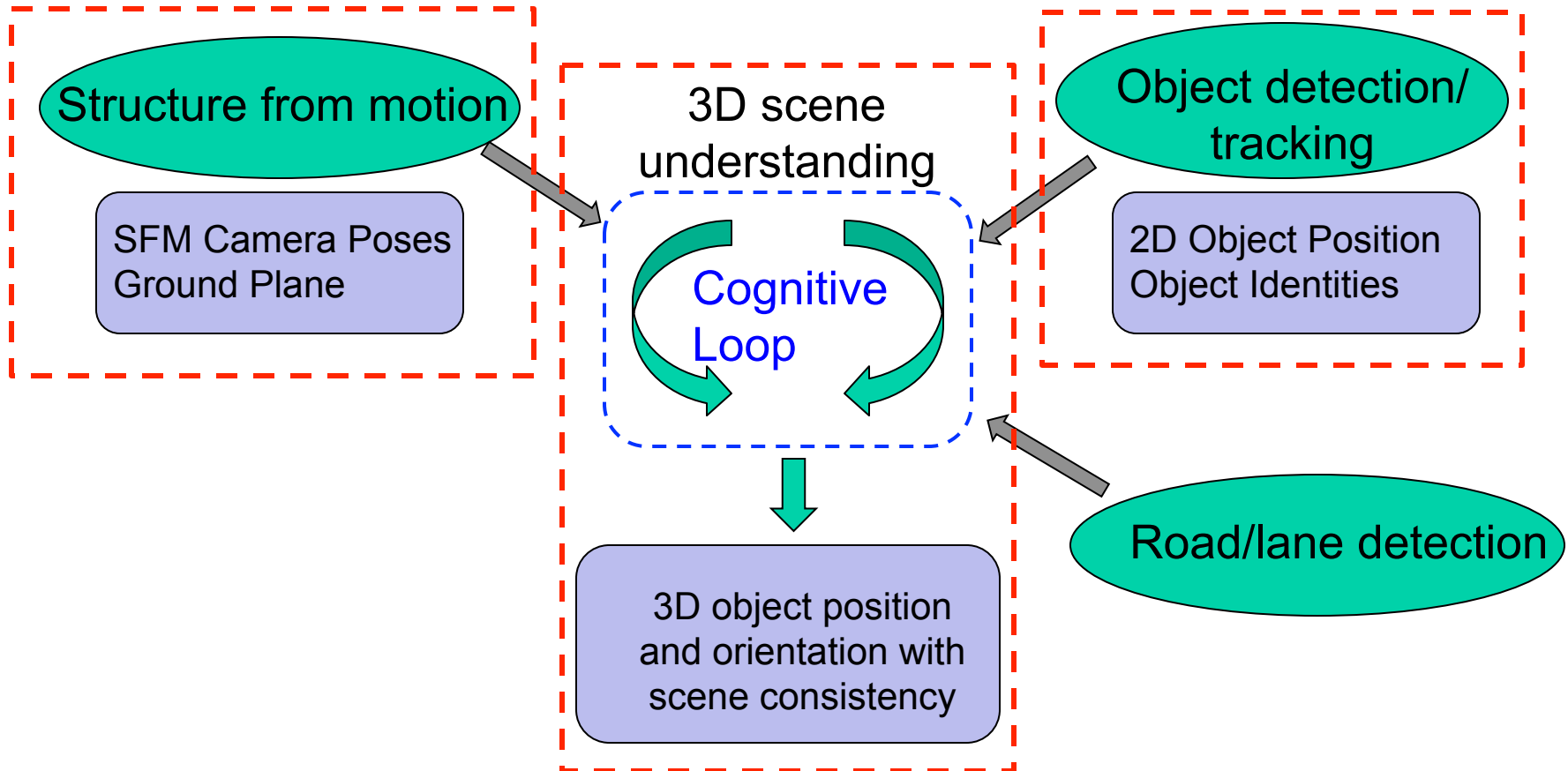
Stereo cameras



Monocular camera

- (Almost) All existing systems: stereo camera or LIDAR is **a must**.
- Our monocular system: **radically simpler** hardware.
- **Our goal**: develop a stand-alone monocular camera based sensing system.
- Working closely with Japan car makers.

Building Blocks for Visual 3D Scene Understanding



- 3D scene understanding: **4 major functional blocks**

KITTI Evaluation Benchmark

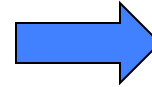
KITTI dataset: Geiger *et al.*, CVPR 2012, <http://www.cvlibs.net/datasets/kitti/>

- Real-world driving sequences
- City, countryside, highway, crowds,
- Speeds 0 to 90 kmph
- SFM Benchmark: 22 sequences, 50 km of driving
- Benchmark for object detection, tracking, road/lane detection

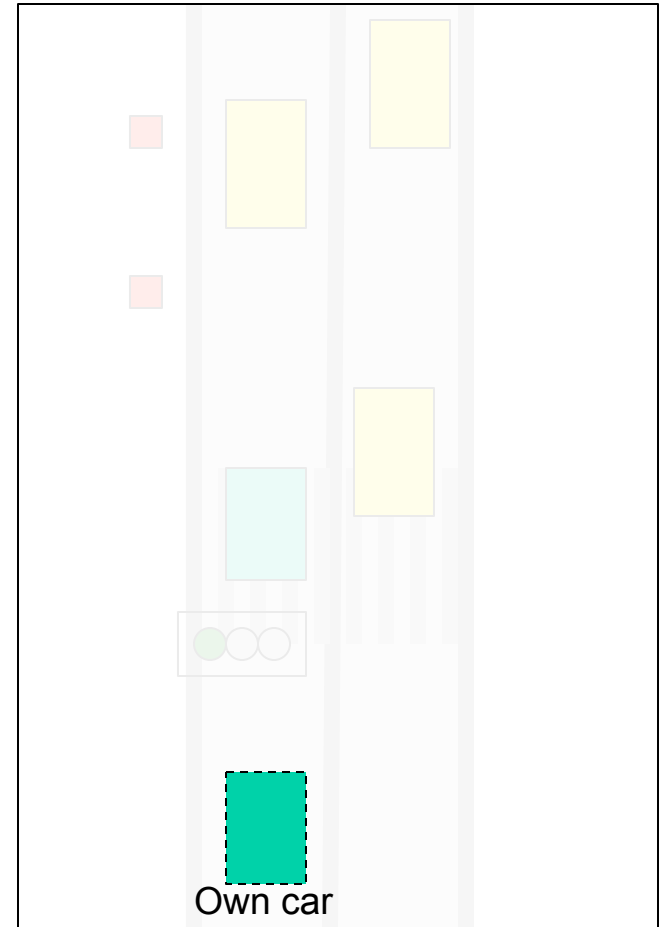


Structure from motion (SFM)

From: video frames
(from a monocular camera)

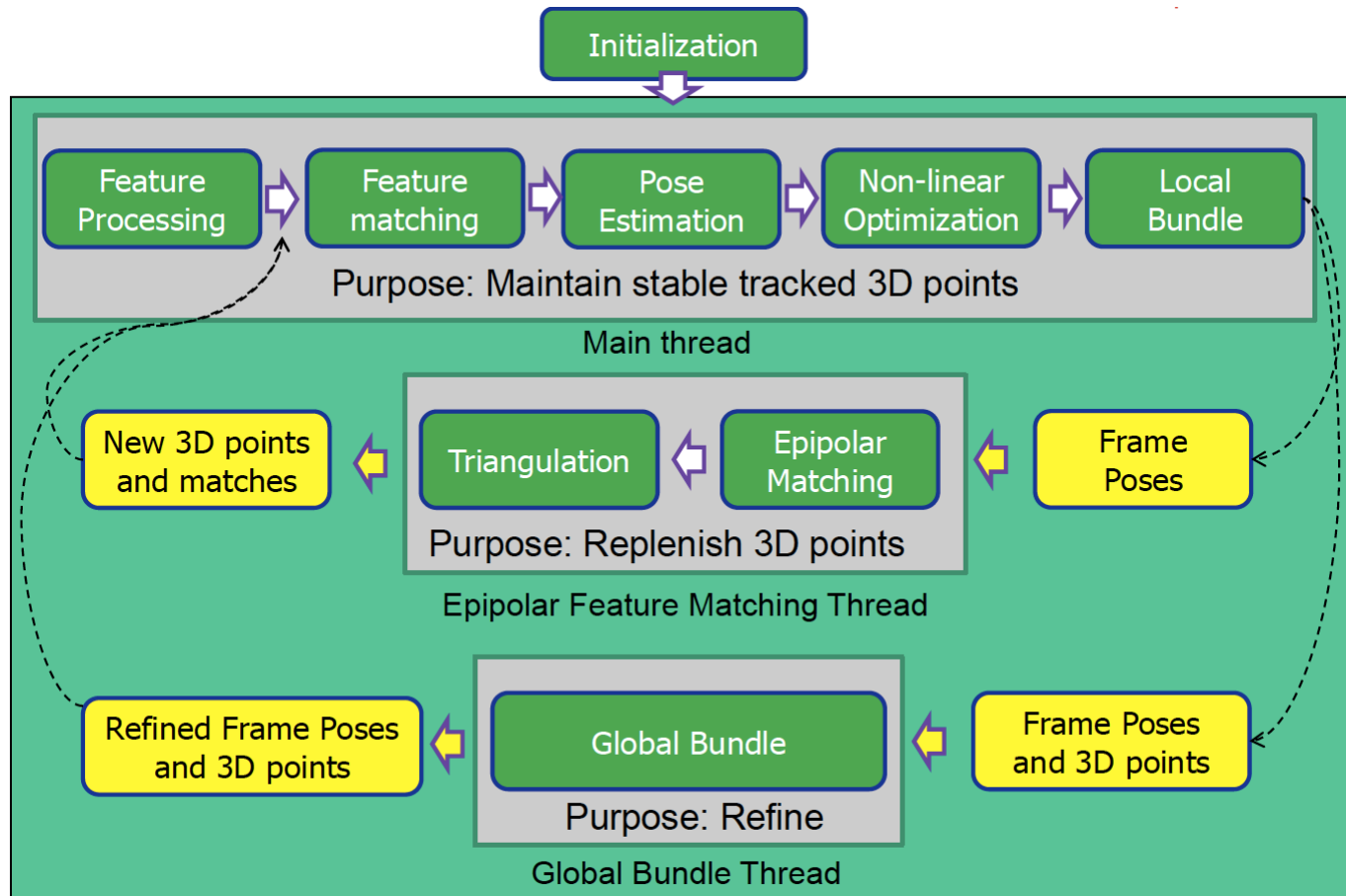


Output: the pose of own car
in 3D world-coordinate



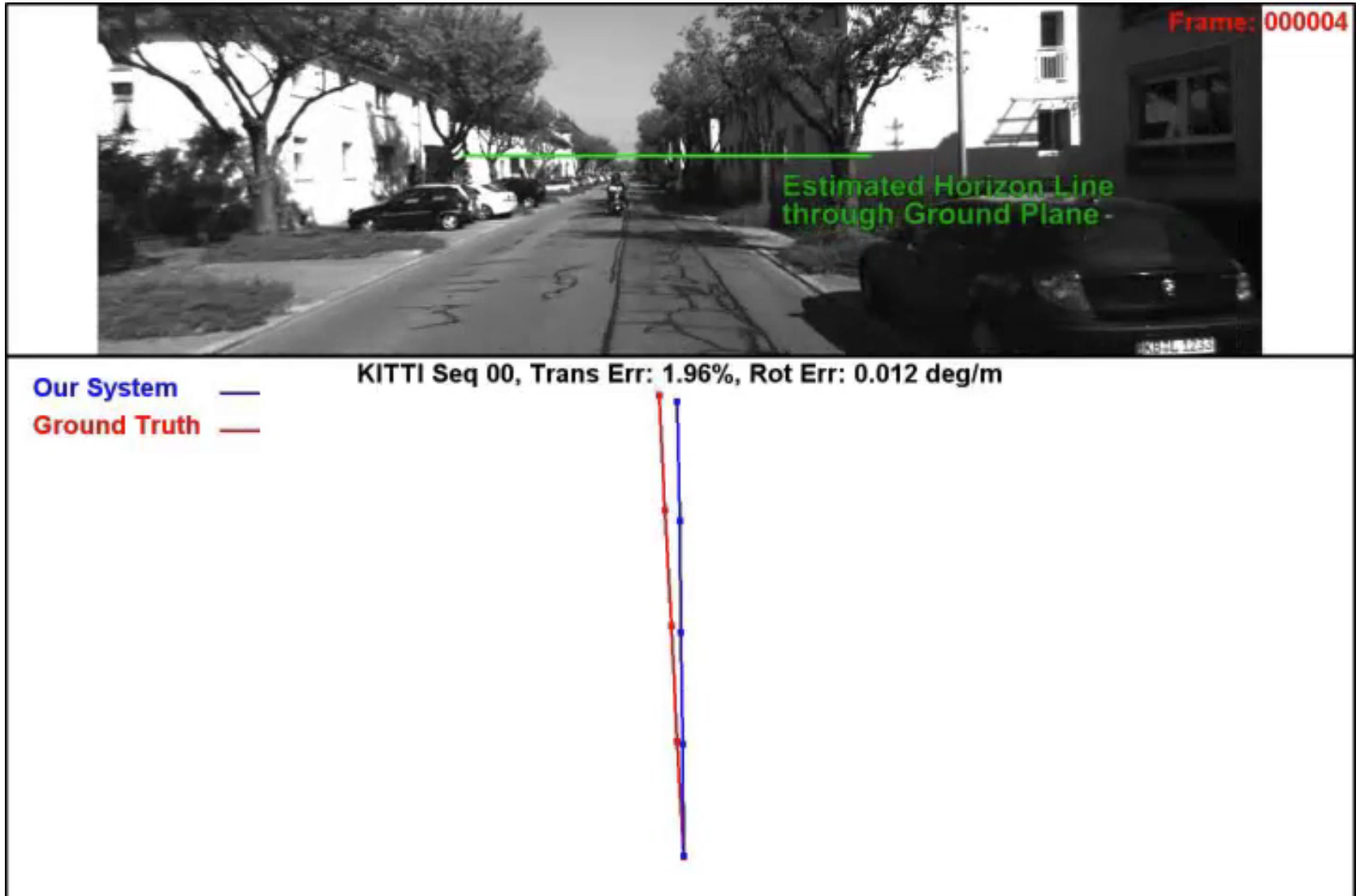
- SFM: compute the **3D pose of the own car** (or the camera).
- Why need camera self-pose: **need to refer to the camera** to get the 3D positions of objects in the world coordinate.

Our monocular SFM system



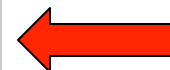
- **Multi-thread system:** ensures robust feature matching
- **SFM + road plane estimation:** yield absolute distance

SFM demo



SFM results

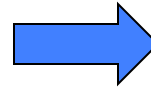
Methods	Rot (deg/m)	Trans (%)	Running time (second)
VISOs-M (Geiger, 2012)	0.0234	11.94	0.1
Ours (Oct 2012)	0.0119	6.42	0.03
Ours (Jan 2013)	0.0104	4.07	0.03
Ours (Jan 2014)	0.0054	3.21	0.03
Ours (now)	0.0057	2.54	0.03
D6DVO (stereo)	0.0051	2.04	0.03
MFI (stereo)	0.003	1.30	0.1



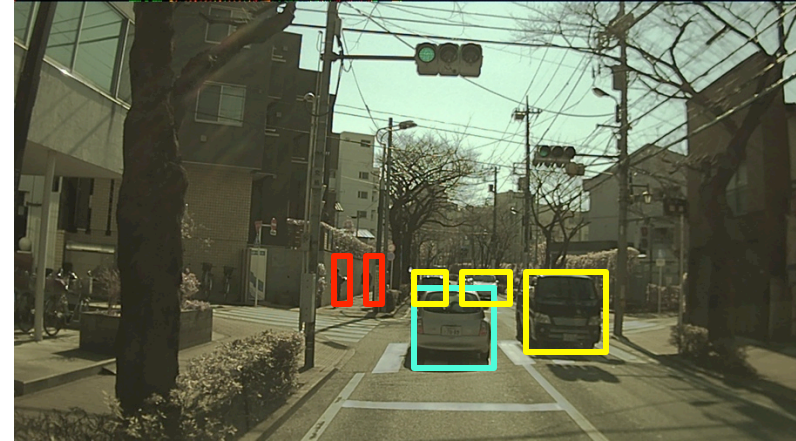
- **Accuracy: dramatically better** than previous state-of-the-art monocular system, **similar performance** as state-of-the-art stereo systems

Object detection + tracking (2D)

From: video frames
(from a monocular camera)



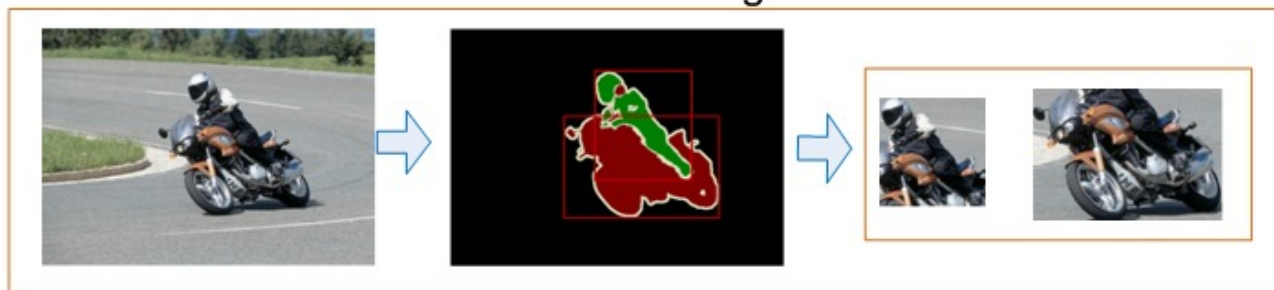
Output: 2D bounding boxes +
object ID



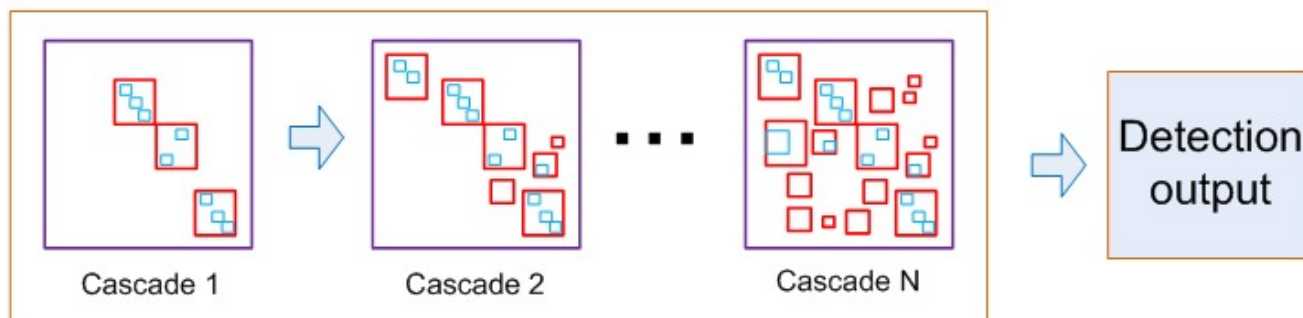
Object detection and tracking: figure out the **position of TPs** (like pedestrians, cars, vans, bikes, etc.) **in each video frame (2D)**

Regionlet for object detection

Generate candidate detection bounding boxes



Boosting classifier cascades



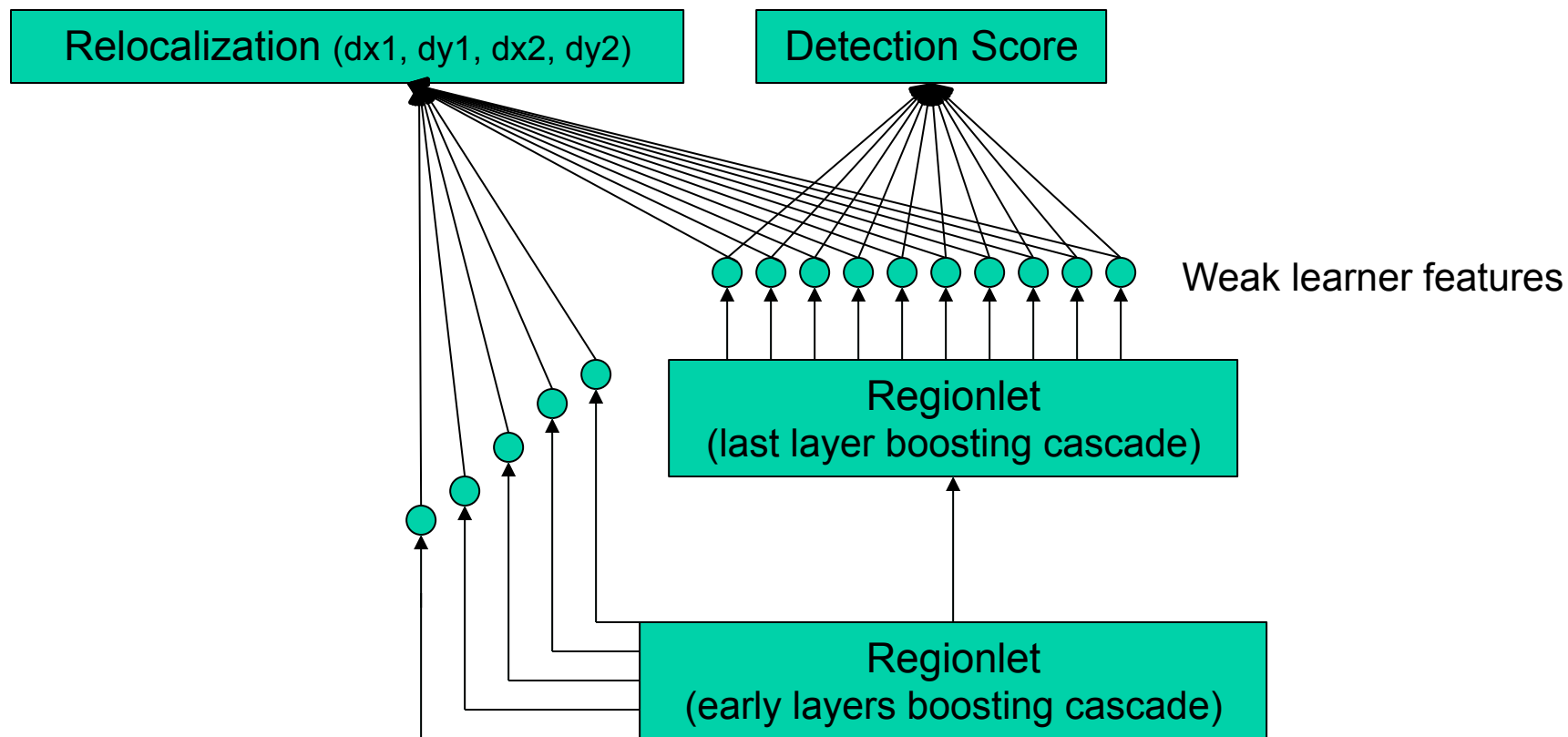
Regionlet

Region

Weak classifier




- **Regionlet approach:** radically different from deformable part model (DPM) system
- **The key:** feature learning through boosting

Regionlet with relocalization



- **Relocalization: very cheap to compute** but with significant performance boost.

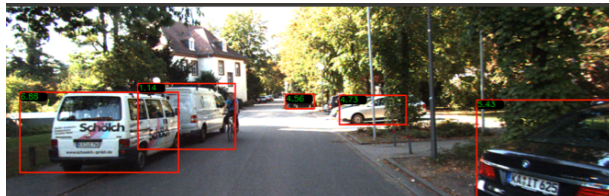
Detection Results on PASCAL07

Methods	Accuracy (mAP)	
DPM (Felzenszwalb, 2010)	26.7%	
DPM (Felzenszwalb, 2013)	33.7%	
DPM + context (Felzenszwalb, 2013)	35.4%	
DPM + context (Song, 2011)	37.7%	
Selective search (Van de Sande, 2011)	33.8%	
Regionlet (Ours, May 2013)	41.6%	
Regionlet (Ours, now)	44.1%	
R-CNN (Girshick, 2014, using outside data)	58.5%	

- **Regionlet: dramatically outperforms DPM**

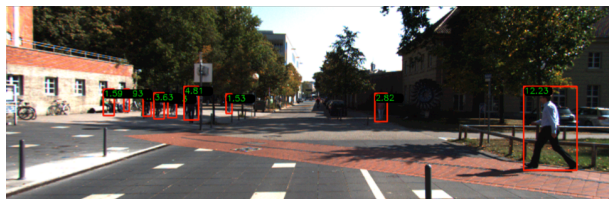
Detection results (AP) on KITTI

Car



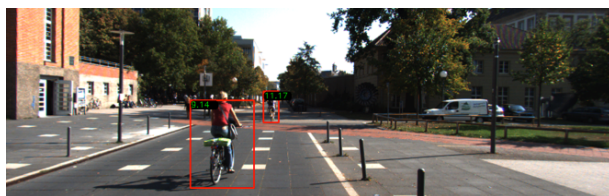
Methods	Easy	Moderate	Hard
DPM (Felzenszwalb, 2010)	66.53%	55.42%	41.04%
The best of all others	81.94%	67.49%	55.60%
Regionlet (Ours)	84.27%	75.58%	59.20%

Pedestrian



Methods	Easy	Moderate	Hard
DPM (Felzenszwalb, 2010)	45.50%	38.35%	34.78%
The best of all others	65.26%	54.49%	48.60%
Regionlet (Ours)	68.79%	55.01%	49.75%

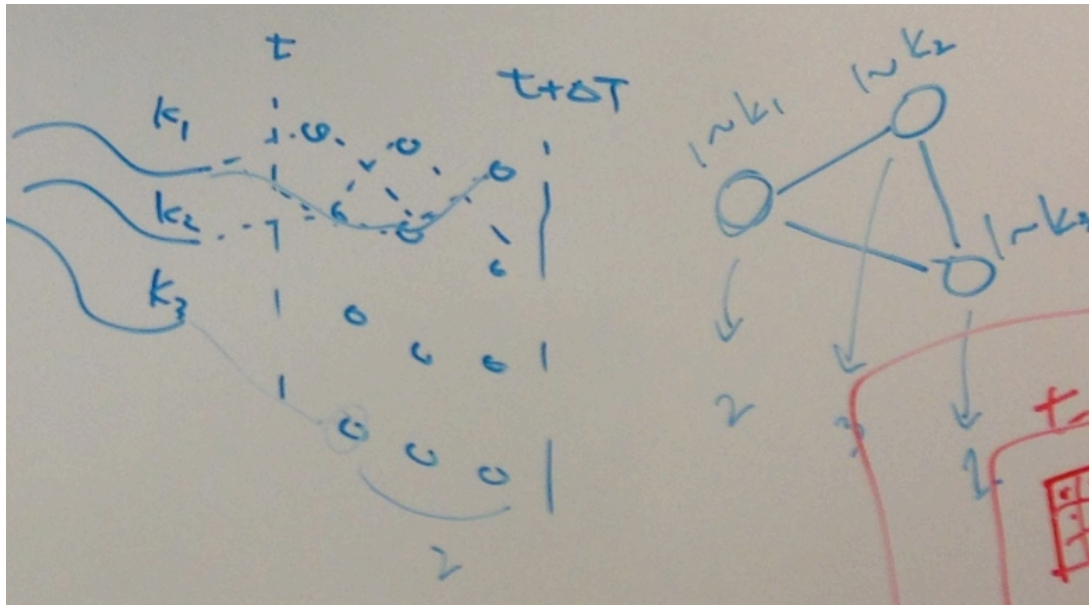
Cyclist



Methods	Easy	Moderate	Hard
DPM (Felzenszwalb, 2010)	38.84%	29.88%	27.31%
The best of all others	51.62%	38.03%	33.38%
Regionlet (Ours)	56.96%	44.65%	39.05%

- **Regionlet: outperforms all competing methods on every case, mostly 15-20% better than DPM**

Object tracking (work in progress)



- Generate track hypothesis using some features
- Decision may be delayed until more cues coming in or when you have to make many decisions
- Work in progress – already achieve very good performance

Preliminary tracking results on KITTI

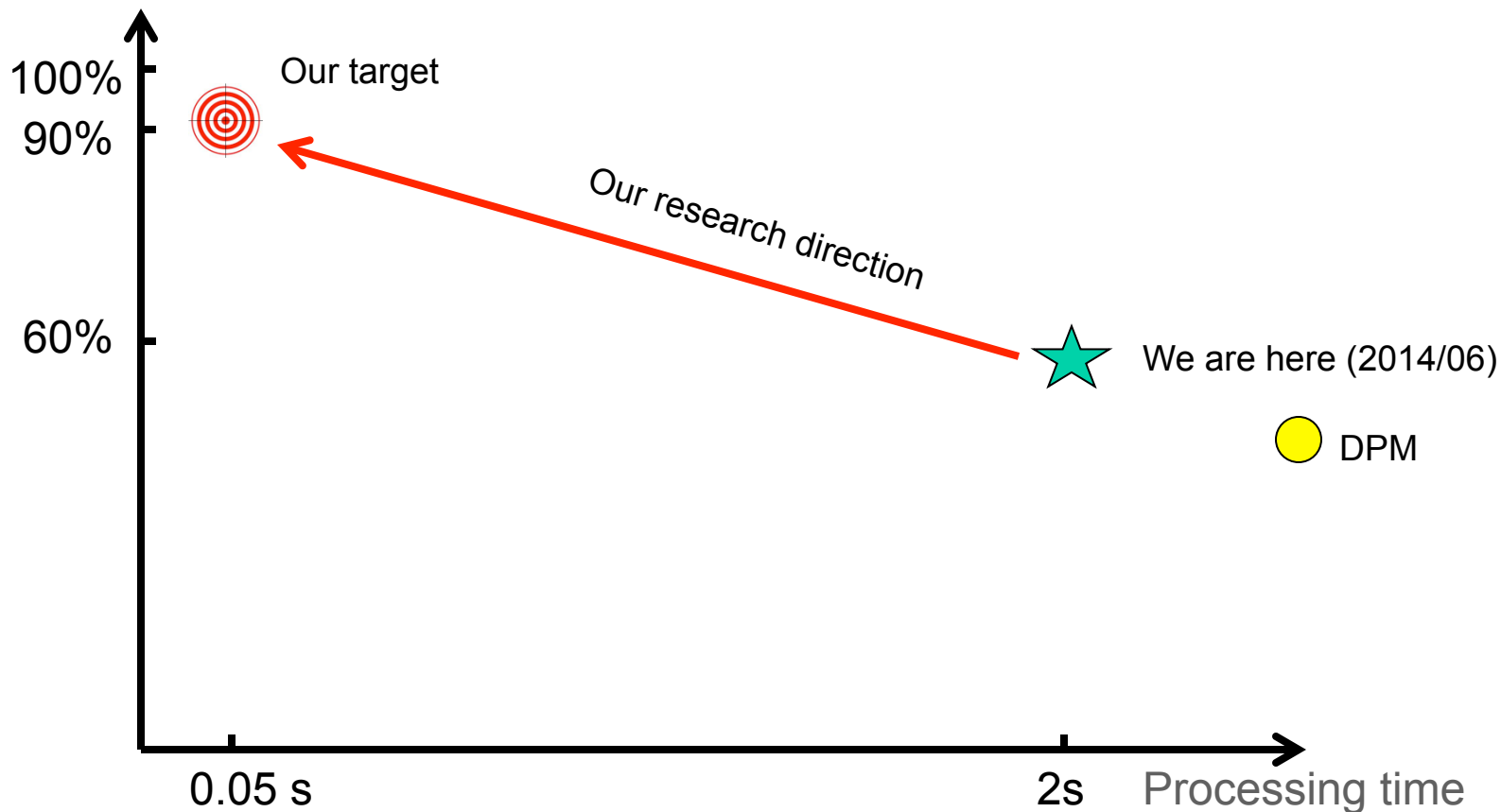
Car

Methods	MOTA	MOTP	MT	ML	IDS	FRAG
The best of the rest	54.17%	78.49%	20.33%	30.35%	12	401
NONT (Anonymous)	58.82%	79.01%	29.44%	26.10%	81	290
Ours	60.88%	78.92%	30.05%	27.62%	33	227

- We achieve similar best performance on car tracking, with much less identity switch.
- For fair comparison, we used the detection results provided by the KITTI

Our goal in detection/tracking – solve the problem

Accuracy (mAP)

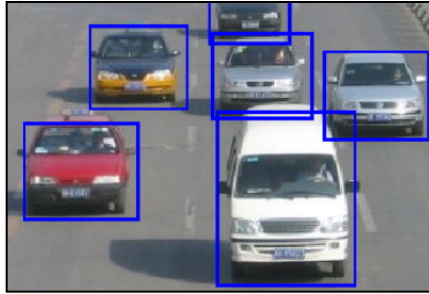


■ Closing the gap (very challenging):

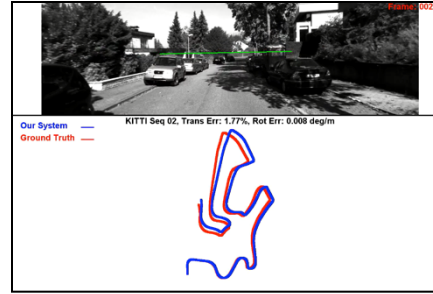
- large-scale training data (collecting > 1 million of labels each class);
- radically more light-weight algorithms but rich enough model (learning with large-scale data);
- exploit the properties of videos (like 3D cues from SFM, dense tracking, etc.).

Putting them together: 3D localization

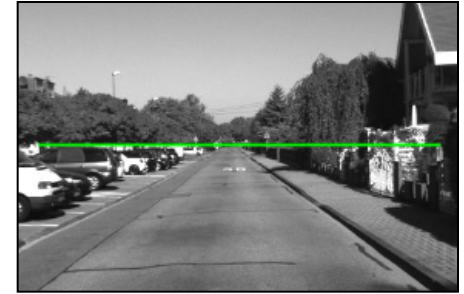
Input



Detection

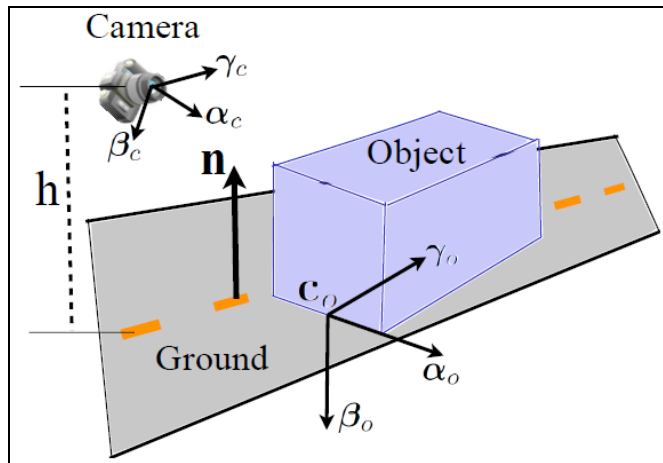


SFM: Camera Motion
 + 3D Tracks on Object



SFM: Ground Plane

Output



Putting things together

Monocular SFM + Detection:
 gives ground plane

SFM + Detection + Ground plane:
 gives object position

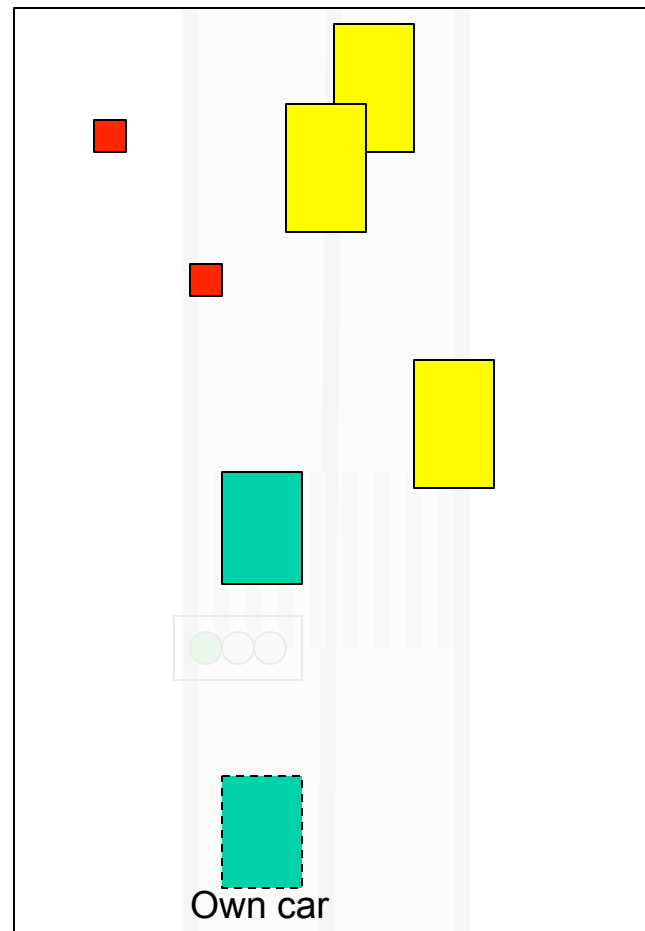
Object SFM + Ground plane:
 gives 3D object bounding box

3D object localization

From: video frames
(from a monocular camera)



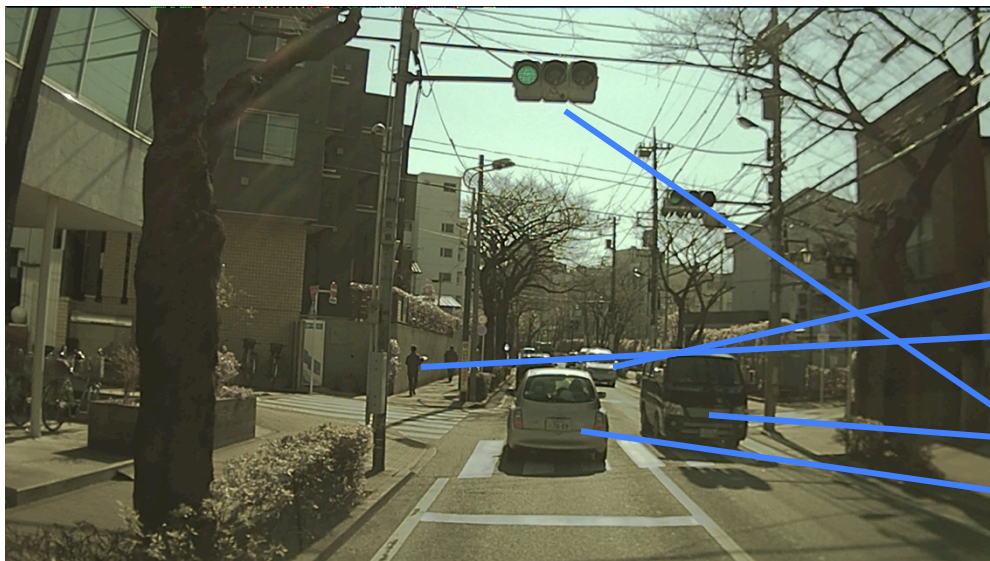
Output: the 3D pose of TPs



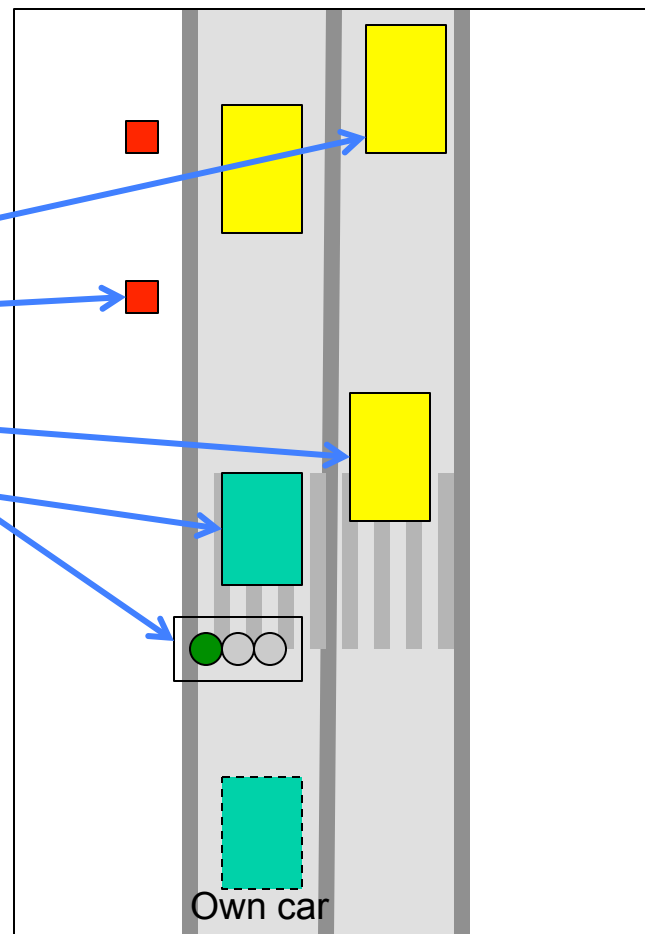
- 3D localization: provide the **3D coordinate of each object** (or in 2D bird-eye view)
- **No constraints from TP-TP relation or TP-scene relations**: due to localization errors, different objects may overlap in 3D (not possible in reality), car may be slightly on sidewalk...

Visual 3D scene understanding

From: video frames



Output: 3D localization of objects with scene consistency



3D driving scene understanding: need scene components like lane/road, traffic sign/signals; provide 3D pose estimation consistent with scene components and among TPs. For example, a driving car is likely to be in the middle of a lane; two objects should not occupy a same 3D space, etc.

Lane detection (preliminary results)



Methods	PRE	F1	HR	PRE	F1	HR	PRE	F1	HR
	-20	-20	-20	-30	-30	-30	-40	-40	-40
The best of others	98.1	97.3	96.6	96.9	96.0	94.3	91.2	88.4	76.0
Ours	98.4	97.2	94.7	97.8	94.7	90.0	91.4	79.3	68.4

Summary

- Autonomous driving is an exciting new opportunity for computer vision. It requires “**research**” to solve some of the fundamental problems in computer vision.
- Our group has achieved state-of-the-art results on 3 KITTI benchmark tasks: monocular SFM, object detection and tracking. We are catching up on road/lane detection.
- These strong building blocks will enable us to build a powerful visual 3D scene understanding system -- **based on monocular camera**.
- Please don't get me wrong: the research is not about the numbers -- rather, it is **the excitement of solving fundamental computer vision problems** that get us very passionate! We go beyond KITTI dataset.
- We are hiring 😊