Towards Affordable Self-Driving Cars

Raquel Urtasun

Computer Science
UNIVERSITY OF TORONTO

Some "Scary" Statistics: Traffic Fatalities

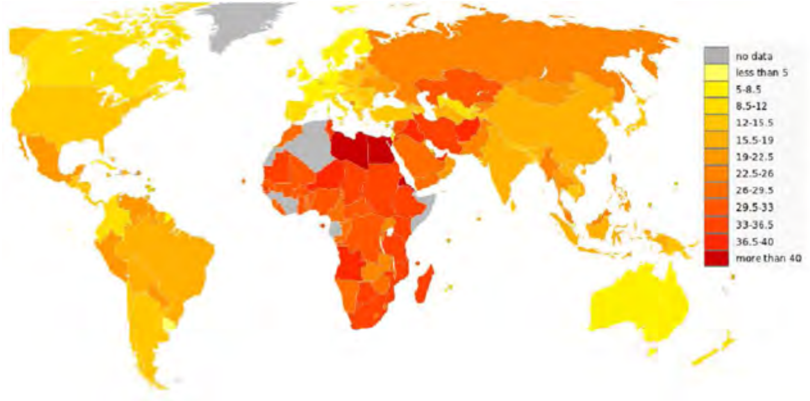


Figure : Road Fatalities per 100,000 inhabitants and year

In total (2010): USA (36,166), Canada (2,075), World (1.24 million!)

Benefits of Autonomous Driving

1. Lower the risk of accidents

Benefits of Autonomous Driving

1. Lower the risk of accidents
2. Provide mobility for the elderly and people with disabilities
    - In the US 45% of people with disabilities still work

Benefits of Autonomous Driving

1. Lower the risk of accidents
2. Provide mobility for the elderly and people with disabilities
   - In the US 45% of people with disabilities still work
3. Decrease pollution for a more healthy environment

Benefits of Autonomous Driving

1. Lower the risk of accidents
2. Provide mobility for the elderly and people with disabilities
   - In the US 45% of people with disabilities still work
3. Decrease pollution for a more healthy environment
4. New ways of Public Transportation

Boring life of a car

- 95% of the time a car is parked

Figure from http://theoatmeal.com/blog/google_self_driving_car

Autonomous Driving

DARPA GRAND CHALLENGE

DARPA URBAN CHALLENGE

Autonomous Driving

State of the art

- Localization, path planning, obstacle avoidance

Autonomous Driving

State of the art

- Localization, path planning, obstacle avoidance
- Heavy usage of Velodyne and detailed (recorded) maps

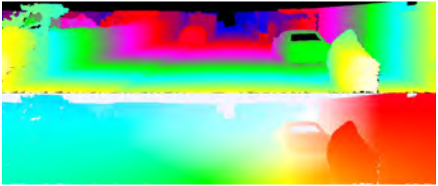
Autonomous Driving

State of the art

- Localization, path planning, obstacle avoidance
- Heavy usage of Velodyne and detailed (recorded) maps

Goal: autonomous driving cheap sensors and little prior knowledge

Autonomous Driving



State of the art

- Localization, path planning, obstacle avoidance
- Heavy usage of Velodyne and detailed (recorded) maps

Goal: autonomous driving cheap sensors and little prior knowledge

Problems for computer vision

- Stereo, optical flow, visual odometry, structure-from-motion

Autonomous Driving

State of the art

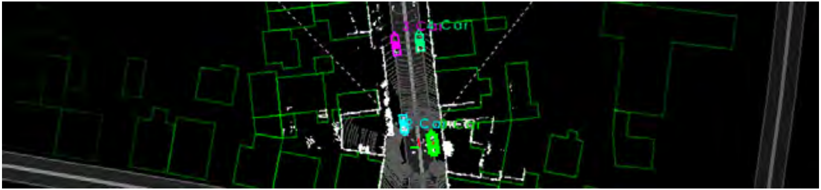
- Localization, path planning, obstacle avoidance
- Heavy usage of Velodyne and detailed (recorded) maps

Goal: autonomous driving cheap sensors and little prior knowledge

Problems for computer vision

- Stereo, optical flow, visual odometry, structure-from-motion
- Object detection, recognition and tracking

Autonomous Driving



State of the art

- Localization, path planning, obstacle avoidance
- Heavy usage of Velodyne and detailed (recorded) maps

Goal: autonomous driving cheap sensors and little prior knowledge

Problems for computer vision

- Stereo, optical flow, visual odometry, structure-from-motion
- Object detection, recognition and tracking
- 3D scene understanding

What do we need?

- Data: not anyone has an autonomous driving platform!

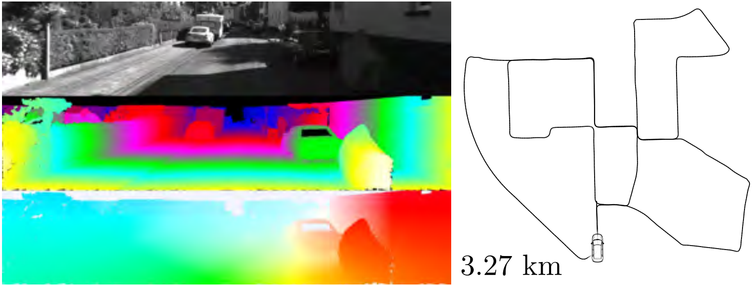
- Holistic Models that can capture the complex dependencies between the different tasks
- Learning algorithms that are efficient and can learn good representations that are useful for many tasks.
- Efficient inference algorithms (realtime on CPU, GPU or other HW accelerators)

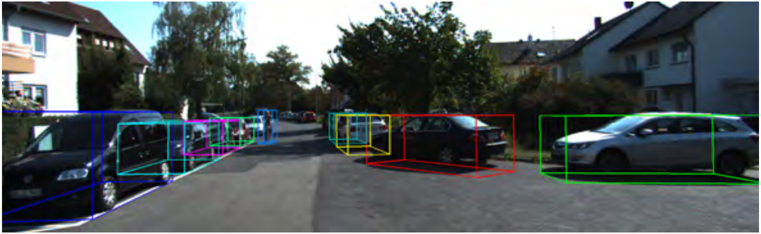
Collecting Big Data

Benchmarks: KITTI Big Data Collection

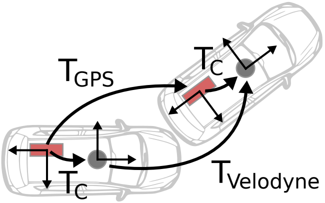
- Two stereo rigs (1392 × 512 px, 54 cm base, 90° opening)
- Velodyne laser scanner, **GPS+IMU** localization
- 6 hours at 10 frames per second → 3Tb

[A. Geiger, P. Lenz, R. Urtasun, In CVPR 2012]



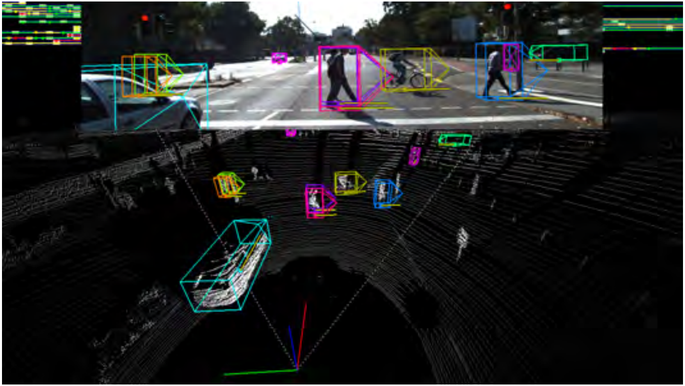


First Difficulty: Sensor Calibration



- Camera calibration [Geiger et al., ICRA 2012]
- Velodyne $\leftrightarrow$ Camera registration
- GPS+IMU $\leftrightarrow$ Velodyne registration

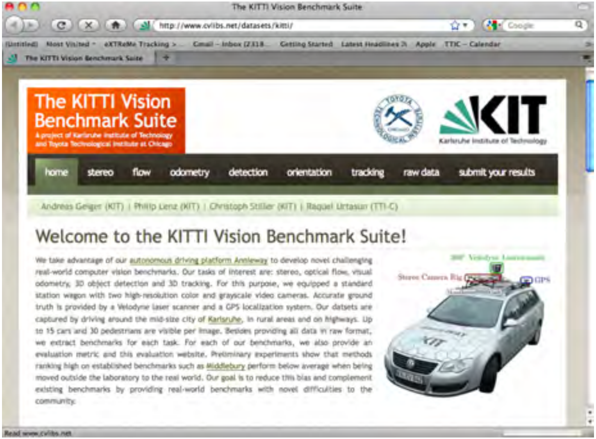
Second Difficulty: Object Annotation



- 3D object labels: Annotators (undergrad students from KIT working for months)
- Occlusion labels: Mechanical Turk

Big Success!

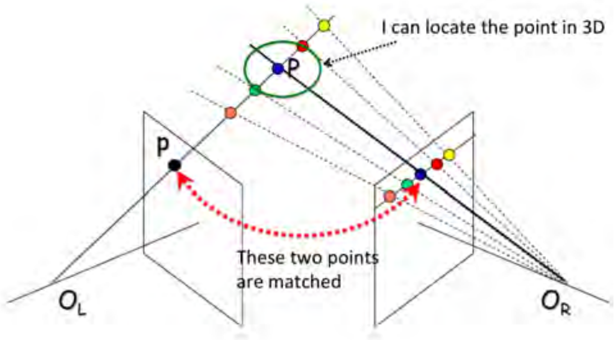
[A. Geiger, P. Lenz, R. Urtasun, In CVPR 2012]



- More than 500 submissions, 20,000 downloads since June 2012!

Reconstructing the 3D World

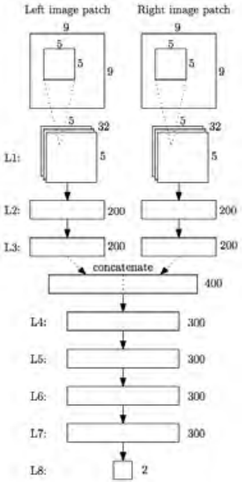
Stereo Estimation



Desired Properties:

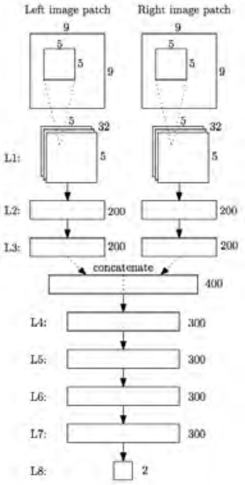
- Robust to saturation, shadows, repetitive patterns, specularities, etc
- Good enough to detect obstacles precisely
- Fast: current accurate techniques are too slow
- Trainable with only a few images, i.e., 100

Matching Networks: geometry-aware CNN



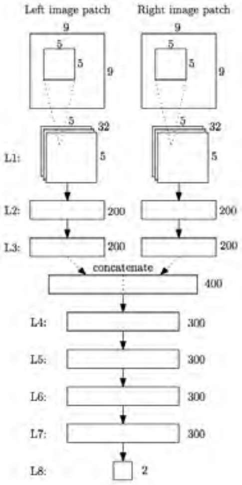
- Current approaches use a siamese network
- Combine the two branches via concatenation follow by further processing
- Too slow: 1 minute of computation on the GPU for KITTI!

Matching Networks: geometry-aware CNN



- Current approaches use a siamese network
- Combine the two branches via concatenation follow by further processing
- Too slow: 1 minute of computation on the GPU for KITTI!
- We solve this problem by learning features that already capture the similarity

Matching Networks: geometry-aware CNN



- Current approaches use a siamese network
- Combine the two branches via concatenation follow by further processing
- Too slow: 1 minute of computation on the GPU for KITTI!
- We solve this problem by learning features that already capture the similarity
- Uncertainty estimates by building probability distributions over all possible solutions

Quantitative Matching Results

[W. Luo, A. Schwing and R. Urtasun, In CVPR 2016]

	> 2 pixel		> 3 pixel		> 4 pixel		> 5 pixel		End-Point		Runtime(s)
	Non-Occ	All	Non-Occ	All	Non-Occ	All	Non-Occ	All	Non-Occ	All	
MC-CNN-acrt	15.02	16.92	12.99	14.93	12.04	13.98	11.38	13.32	4.39 px	5.21 px	20.13
Ours(19)	**10.87**	**12.86**	**8.61**	**10.64**	**7.62**	**9.65**	**7.00**	**9.03**	**3.31** px	**4.2** px	0.14

Table : KITTI 2012 validation set.

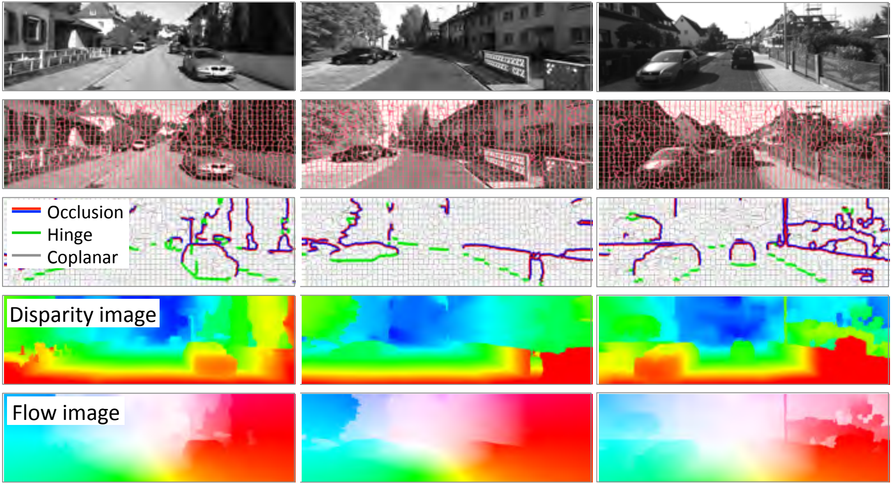
	> 2 pixel		> 3 pixel		> 4 pixel		> 5 pixel		End-Point		Runtime(s)
	Non-Occ	All	Non-Occ	All	Non-Occ	All	Non-Occ	All	Non-Occ	All	
MC-CNN-acrt	15.20	16.83	12.45	14.12	11.04	12.72	10.13	11.80	4.01 px	4.66 px	22.76
Ours(37)	**9.96**	**11.67**	**7.23**	**8.97**	**5.89**	**7.62**	**5.04**	**6.78**	**1.84** px	**2.56** px	0.34

Table : KITTI 2015 validation set.

- Our approach produces much more accurate matches, 2-orders of magnitude faster than competing approaches [Zbontar & LeCunn, CVPR 2015]

Results on KITTI

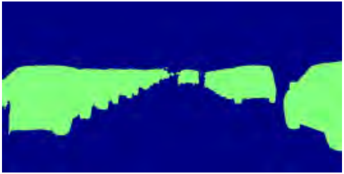
[W. Luo, A. Schwing and R. Urtasun, In CVPR 2016]



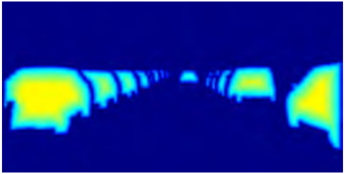
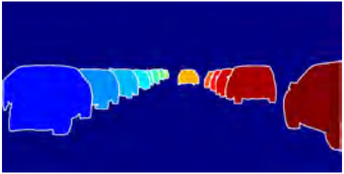
Instance Segmentation

Deep Watershed Transform For Instance Segmentation

- Combine deep learning with classical grouping methods
- Exploit Semantic Segmentation to focus only on important regions

- End-to-End trainable to predict the energy of the system
- Inference via a forward pass follow by Watershed transform

- Extremely good performance

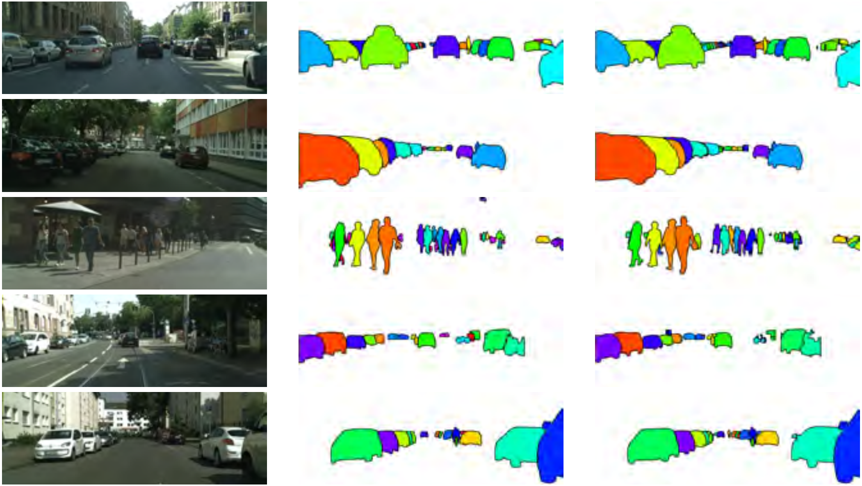
Quantitative Results

[M. Bai and R. Urtasun, In ArXiv'16]

	mAP	mAP(50%)	mAP(100m)	mAP(50m)
van den Brand et al. 16	2.3	3.7	3.9	4.9
R-CNN + MCG	4.6	12.9	7.7	10.3
Uhrig et al. 16	8.9	21.1	15.3	16.7
Ours	**15.6**	**30.0**	**26.2**	**31.8**

Table : **Cityscapes Test Set**: Our approach outperforms the state-of-the-art by a large margin. Results are averaged over classes (person, rider, car, truck, bus, train, motorcycle, bicycle)

Sample Predictions



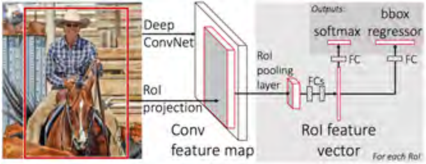
Input Image | Our Prediction | Ground Truth

3D Object Detection and Tracking

Object Detection

- Current approaches to object detection typically work in two steps:
  1. Generate object proposals, e.g, bottom-up grouping
  2. Score the most promising ones with sophisticated CNNs

Fast RCNN

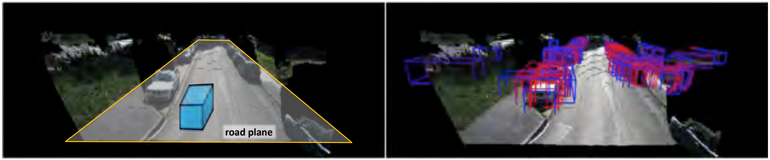


[Girshick et al. 14]

- Unfortunately this works poorly in autonomous driving scenarios
- Furthermore, for autonomous driving we need to know distance to obstacle
- 3D allow us to have better priors, and directly get distances

Our 3D Object Detection

- Use structure prediction to learn to propose object candidates in 3D



- Use deep learning to do final detection

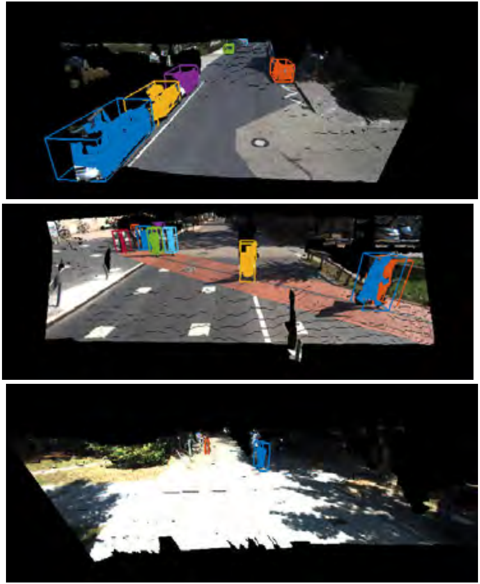
KITTI Detection Results

[X. Chen, K. Kundu, Y. Zhu, H. Ma, S. Fidler and R. Urtasun, NIPS'15]

	Cars			Pedestrians			Cyclists		
	Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
LSVM-MDPM-sv	68.02	56.48	44.18	47.74	39.36	35.95	35.04	27.50	26.21
SquaresICF	-	-	-	57.33	44.42	40.08	-	-	-
DPM-C8B1	74.33	60.99	47.16	38.96	29.03	25.61	43.49	29.04	26.20
MDPM-un-BB	71.19	62.16	48.43	-	-	-	-	-	-
DPM-VOC+VP	74.95	64.71	48.76	59.48	44.86	40.37	42.43	31.08	28.23
OC-DPM	74.94	65.95	53.86	-	-	-	-	-	-
AOG	84.36	71.88	59.27	-	-	-	-	-	-
SubCat	84.14	75.46	59.71	54.67	42.34	37.95	-	-	-
DA-DPM	-	-	-	56.36	45.51	41.08	-	-	-
Fusion-DPM	-	-	-	59.51	46.67	42.05	-	-	-
R-CNN	-	-	-	61.61	50.13	44.79	-	-	-
FilteredICF	-	-	-	61.14	53.98	49.29	-	-	-
pAUCEnsT	-	-	-	65.26	54.49	48.60	51.62	38.03	33.38
MV-RGBD-RF	-	-	-	70.21	54.56	51.25	54.02	39.72	34.82
3DVP	87.46	75.77	65.38	-	-	-	-	-	-
Regionlets	84.75	76.45	59.70	73.14	61.15	55.21	70.41	58.72	51.83
Faster R-CNN	86.71	81.84	71.12	78.86	65.90	61.18	72.26	63.35	55.90
Ours	**93.04**	**88.64**	**79.10**	**81.78**	**67.47**	**64.70**	**78.39**	**68.94**	**61.37**

Table : Average Precision (AP) (in %) on the test set of the KITTI Object Detection Benchmark (at the time of paper published)

Video



Tracking

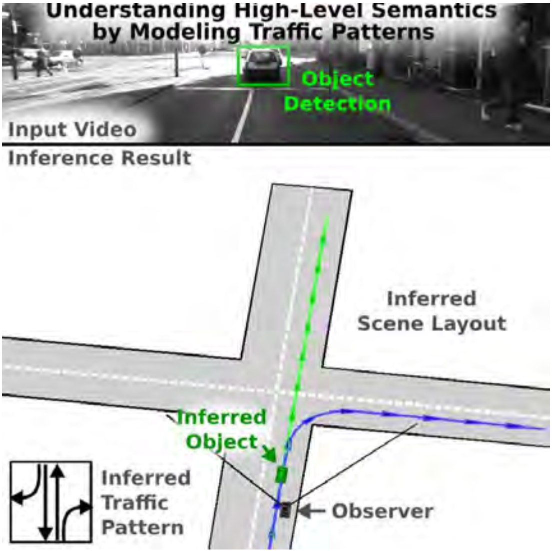
[D. Frossard and R. Urtasun, In ArXiv soon]

- End-to-end detection and tracking with a deep structured model

Holistic Models

Semantic Scene Understanding

[H. Zhang, A. Geiger and R. Urtasun, ICCV 2013]



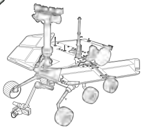
The vehicle has to self-localize

Motivation

- Localization is crucial for autonomous systems







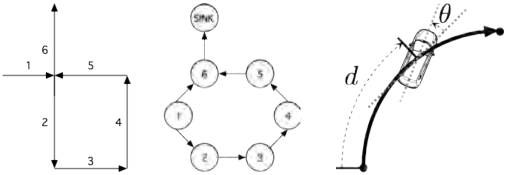
- GPS has limitations in terms of reliability and availability
- Place recognition techniques use image features or depth maps and a database of previously collected images (e.g., Google car)
- We develop an inexpensive technique for localizing to 3m in unseen regions

Humans as an inspiration

- Humans are able to use a map, combined with visual input and exploration, to localize effectively
- Detailed, community developed maps are freely available (OpenStreetMap)
- How can we exploit maps, combined with visual cues, to localize a vehicle?

Probabilistic Localization using Visual Odometry

- Maps can be considered as a graph
  - Nodes of the graph represent street segments
  - Edges represent intersections and transitions between these segments
- Position is defined by the current street and the distance travelled **d**, and orientation θ on that street



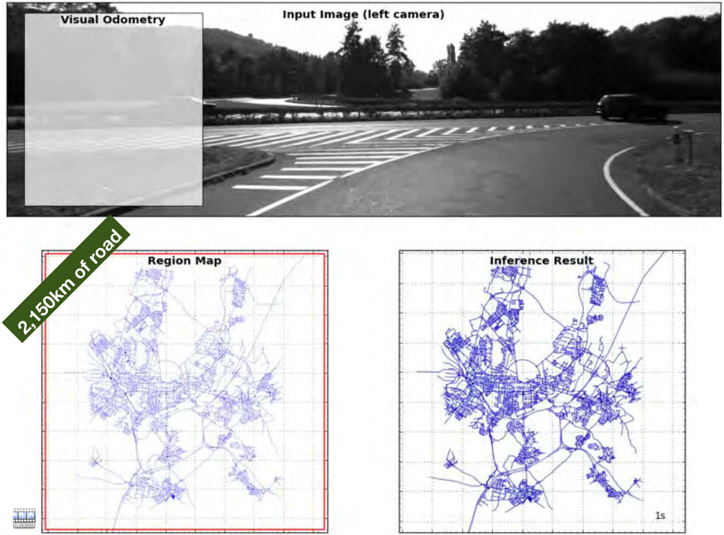
- Localization is formulated as posterior inference $p(u_t, \mathbf{s}_t | \mathbf{y}_{1:t})$

$$\propto \underbrace{p(\mathbf{y}_t | u_t, \mathbf{s}_t)}_{\text{likelihood}} \sum_{u_{t-1}} \int \underbrace{p(u_t | u_{t-1}, \mathbf{s}_{t-1})}_{\text{street transition}} \underbrace{p(\mathbf{s}_t | u_t, u_{t-1}, \mathbf{s}_{t-1})}_{\text{pose transition}} \underbrace{p(u_{t-1}, \mathbf{s}_{t-1} | \mathbf{y}_{1:t-1})}_{\text{previous posterior}} d\mathbf{s}_{t-1}$$

with u_t street segment and $\mathbf{s}_t$ the location and orientation in the segment

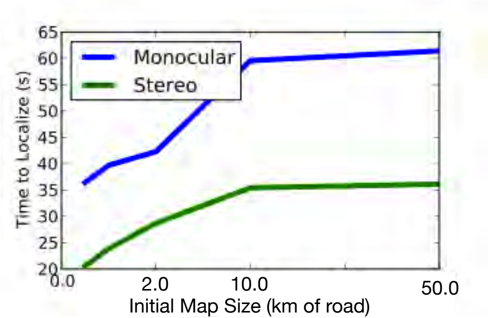
Self-localization

[M. Brubaker, A. Geiger and R. Urtasun, CVPR'13 best paper runner up award]



Quantitative Experiments

Average	Stereo Odometry	Monocular Odometry	Map Projection
Position Error	3.1m	18.4m	1.4m
Heading Error	1.3°	3.6°	-
Localization Time	36s	62s	-



Better maps will make autonomous driving easier

Building Road Maps of the World

- Companies like HERE maps use dedicated vehicles with many sensors to do mapping
- This has small coverage, and its expensive!
- How can we have large coverage and cost 0$?

View of an Intelligent Vehicle

- A single car has a narrow view of the world

What can we do?

- "Big brother" knows everything about what we are doing

See through the clouds

drones

See through the clouds

UAVs

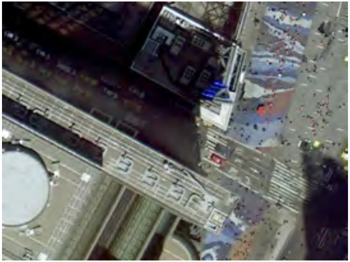
See through the clouds

planes

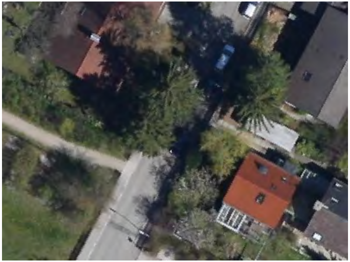
See through the clouds

satellites

Challenges of Aerial/Satellite Imagery

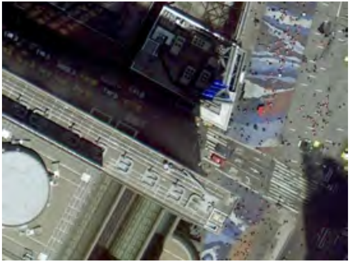


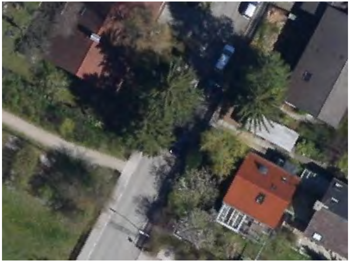
shadows



occlusion

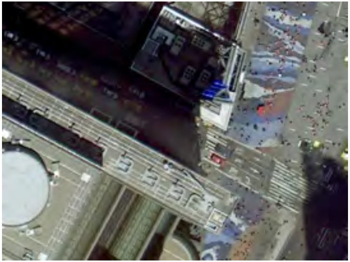
Challenges of Aerial/Satellite Imagery


shadows

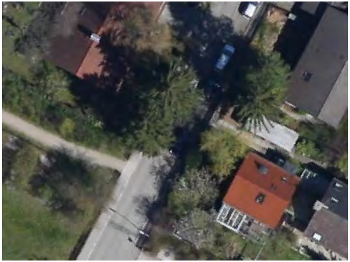

occlusion

- Typically framed as semantic segmentation
  - We can use all the tricks we learned from standard images
  - How can we obtain topology?

Challenges of Aerial/Satellite Imagery



shadows



occlusion

- Typically framed as semantic segmentation
  - We can use all the tricks we learned from standard images
  - How can we obtain topology?
- We don't need to start from scratch

Using OpenStreetMaps

- More than half the world is already mapped

- Typically only contain the road centerline
- Trick: Use OSM topology to define the model

Ground and Aerial Views

- Ground and aerial views are very complementary, so we should use both
- They do not need to overlap everywhere

- We need to estimate the alignment between aerial and ground imagery
  - GPS is not good enough

Large Coverage HD Maps

[G. Mattyus, S. Wang, S. Fidler and R. Urtasun, In CVPR 2016]

- Fine-grained categorization

(a) Intersection with tram line

(b) Small town

(c) A road with three lanes

(d) Two roads with tram stop in between

Next Big Challenge: Large Scale Semantic 3D

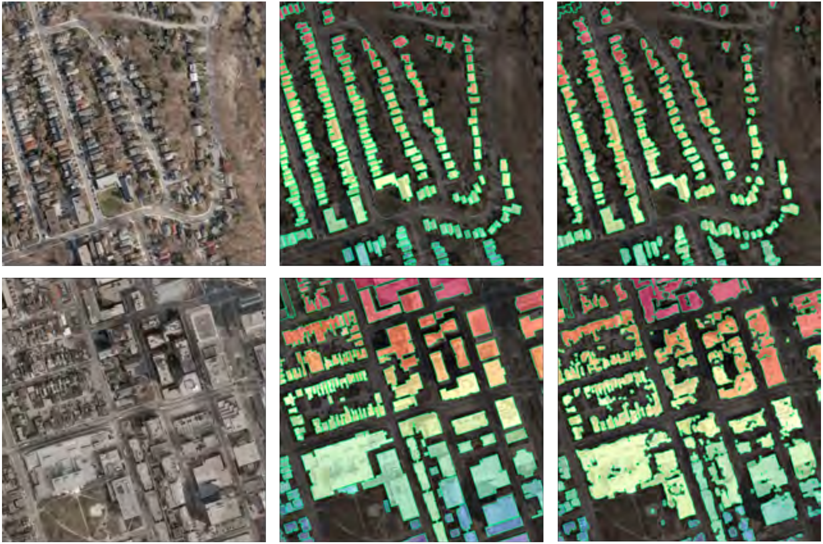
TorontoCity Benchmark

[S. Wang, M. Bai, G. Mattyus, H. Chu, W. Luo, B. Yang, J. Liang, J. Cheverie, S. Fidler and R. Urtasun, In Arxiv'16]

- Full coverage of $712.5km^2$ with $397,846$ buildings and $8439km$ of road

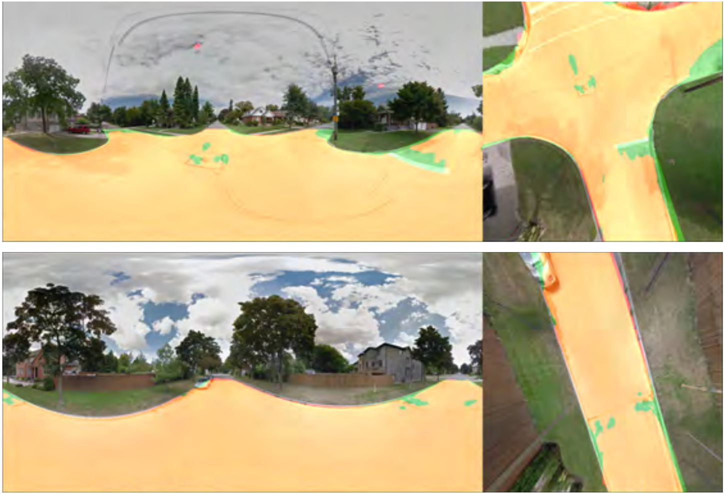
Input Ground-truth Ours

Instance Segmentation



Input Ground-truth Ours

Groundview road segmentation



Yellow: Ground-truth and prediction agree.

Green: Ground-truth is road and prediction is non-road

Red: Prediction is road and ground-truth is non-road

Estimating Road Topology from Aerial Images

[G. Mattyus, W. Luo and R. Urtasun, Soon in Arxiv]

Ours

Ground Truth

Conclusions

- Affordable self-driving cars
  - Sensing: stereo, flow
  - Perception: detection, holistic models
  - Localization
  - Mapping
- Next big benchmark
- Still lots of research to be done!

Acknowledgment

Faculty

Sanja Fidler

Postdocs

Gellert Mattyus

Graduate Students

Shenlong Wang

Kaustav Kundu

Wenjie Luo

Min Bai

Hang Chu

Justin Liang

Bin Yang

Joel Cheverie

Davi Frossard