Deep Learning for Robot Perception and Navigation

Wolfram Burgard

UNI FREIBURG

AiS Autonomous Intelligent Systems

Deep Learning for Robot Perception and Navigation

Lifeng Bo, Claas Bollen, Thomas Brox, Andreas Eitel, Dieter Fox, Gabriel L. Oliveira, Oier Mees, Luciano Spinello, Jost Tobias Springenberg, Martin Riedmiller, Michael Ruhnke, Abhinav Valada, Jingwei Zhang, …

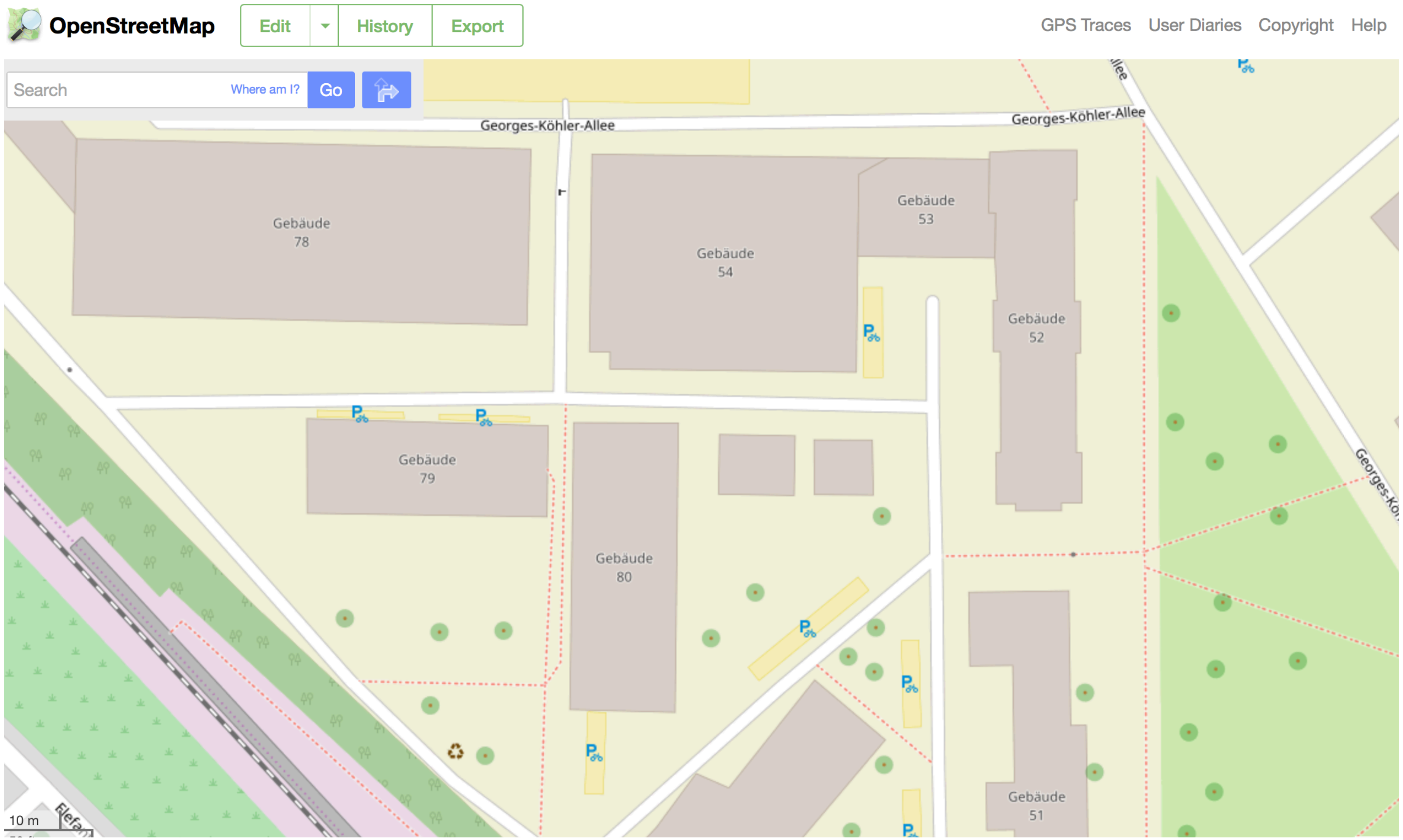
UNI FREIBURG

Autonomous Intelligent Systems

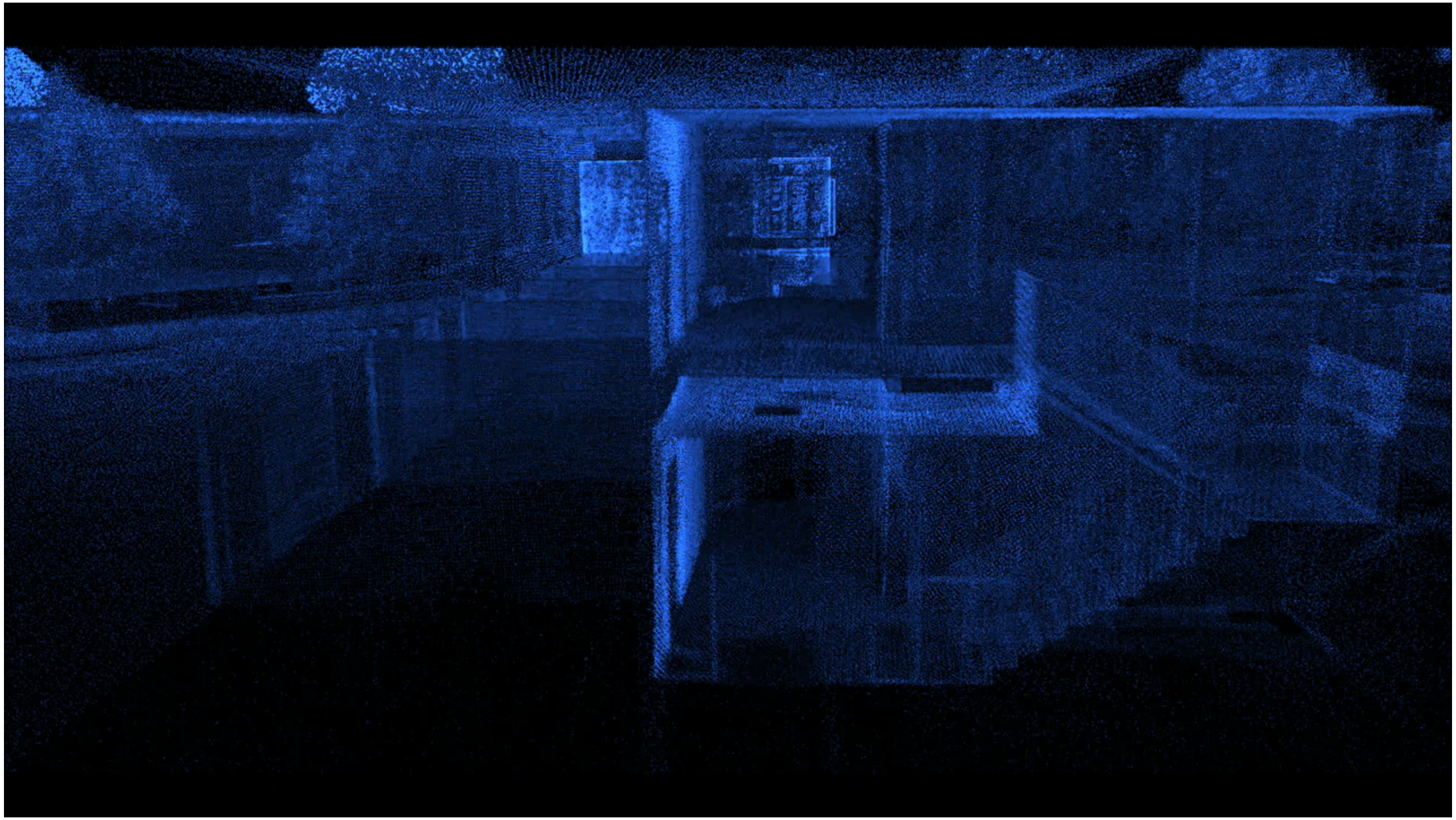
Some facts about the AIS Lab

- 3 senior researchers
- 32 Ph.D. students
- >350 publications
- 1 technician
- 1 secretary
- 1 project manager
- Head of the DFG Cluster of Excellence BrainLinks-BrainTools
- Hosted in the Integrated Robotics Center

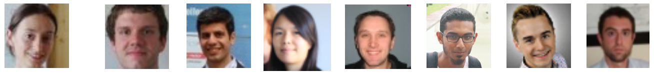
First Building Added by Robots to OpenStreetMap

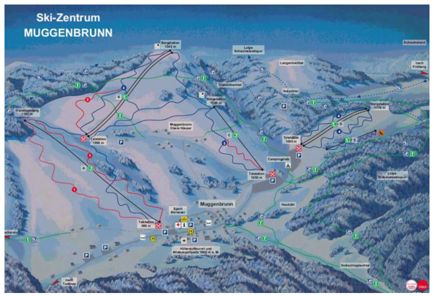


The IRC Mapped by Robots



Freiburg is a Great Place

- Great team  ...
- Great advisor 
- Great city
- Great environment
- Great food
- Great wine
- Great 29 days of vacations
- Great 10 national holidays
- Great massage places



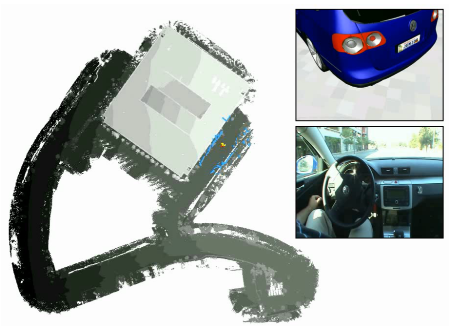
In case of any doubt: apply!

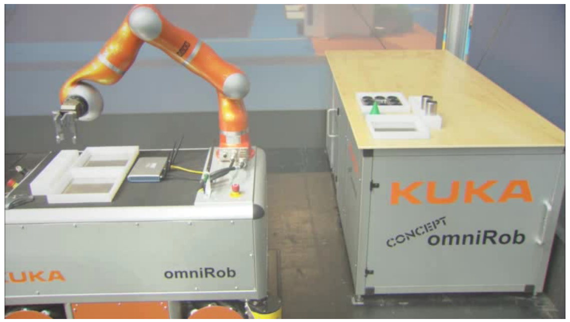
Fields of Research

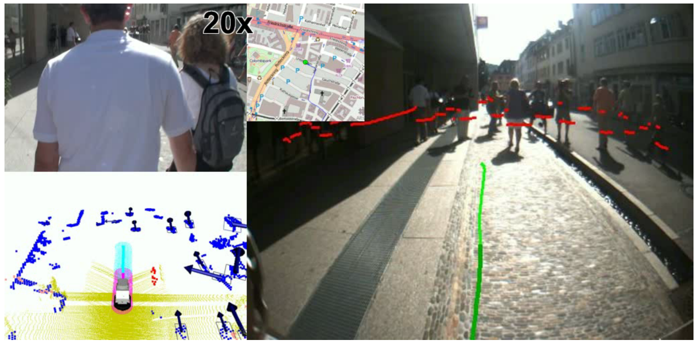
- Mobile robotics
- State estimation and modeling
- Mapping
- Decision-theoretic approaches
- Adaptive techniques and learning
- Scene understanding
- Mobile manipulation
- Multi-robot coordination
- Robots and embedded systems
- Autonomous cars
- Flying vehicles
- …
- **Probabilistic robotics**

Autonomous Robots

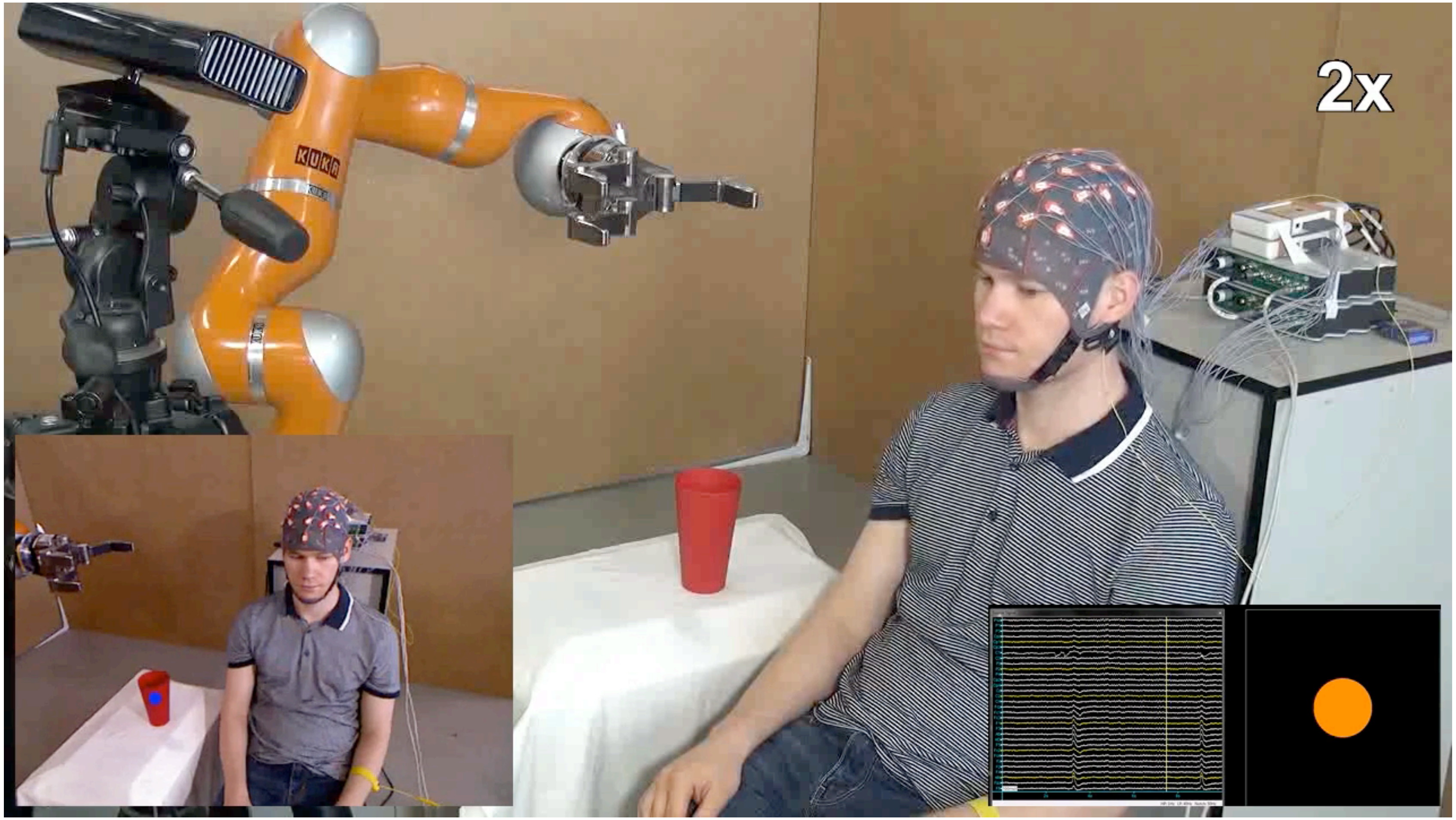
Robots that reliably fulfill their tasks in real-world environments







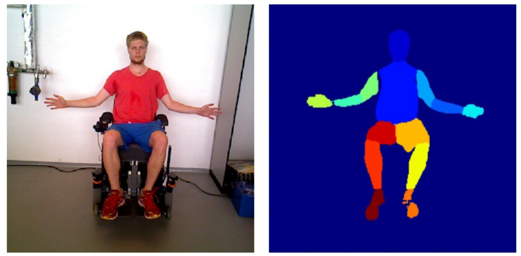
Neurobots

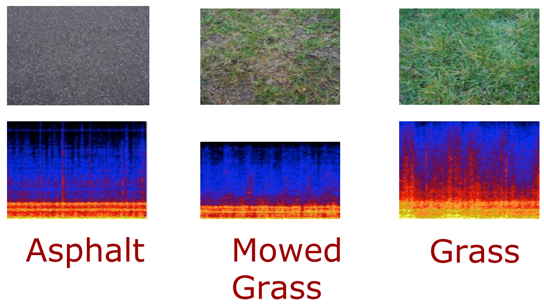


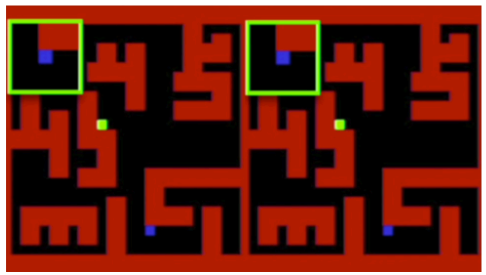
The Tagesthemen-Report

What our Robots Should do ...

- Perception
  - object recognition
  - human detection
  - sensor interpretation
- Navigation
- Manipulation







Why Deep Learning?

- **Multiple layers of abstraction** provide an advantage for **solving complex** (pattern recognition) **problems**
- Highly **successful** in **computer** vision and **pattern recognition**
- Can serve **wide range of fields and applications**
- **End-to-end systems**

Deep Learning in Robotics

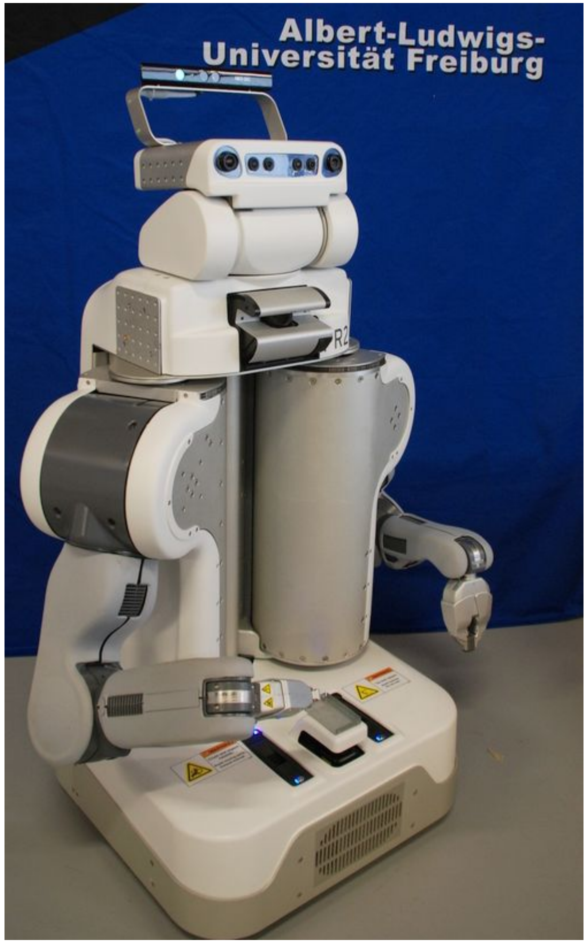
- Robot perception is a challenging problem and involves many different aspects such as
  - Scene understanding
  - Object detection
  - Detection of humans
- Opportunities
  - improving perception,
  - manipulation
  - navigation

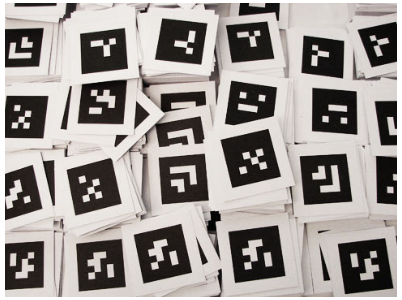
Multimodal Deep Learning for Robust RGB-D Object Recognition

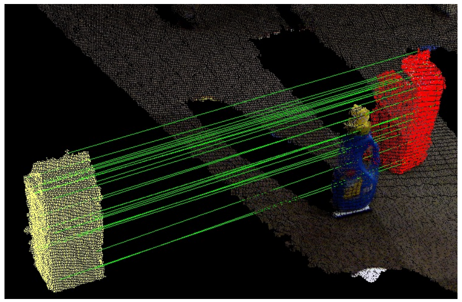
Andreas Eitel, Jost Tobias Springenberg, Martin Riedmiller, Wolfram Burgard

[IROS 2015]

RGB-D Object Recognition







Often too little Data for Deep Learning Solutions

Deep networks are hard to train and require large amounts of data

- Lack of sufficient amount of labeled training data for RGB-D domain
- How to deal with limited sizes of available datasets?

Data often too Clean for Deep Learning Solutions

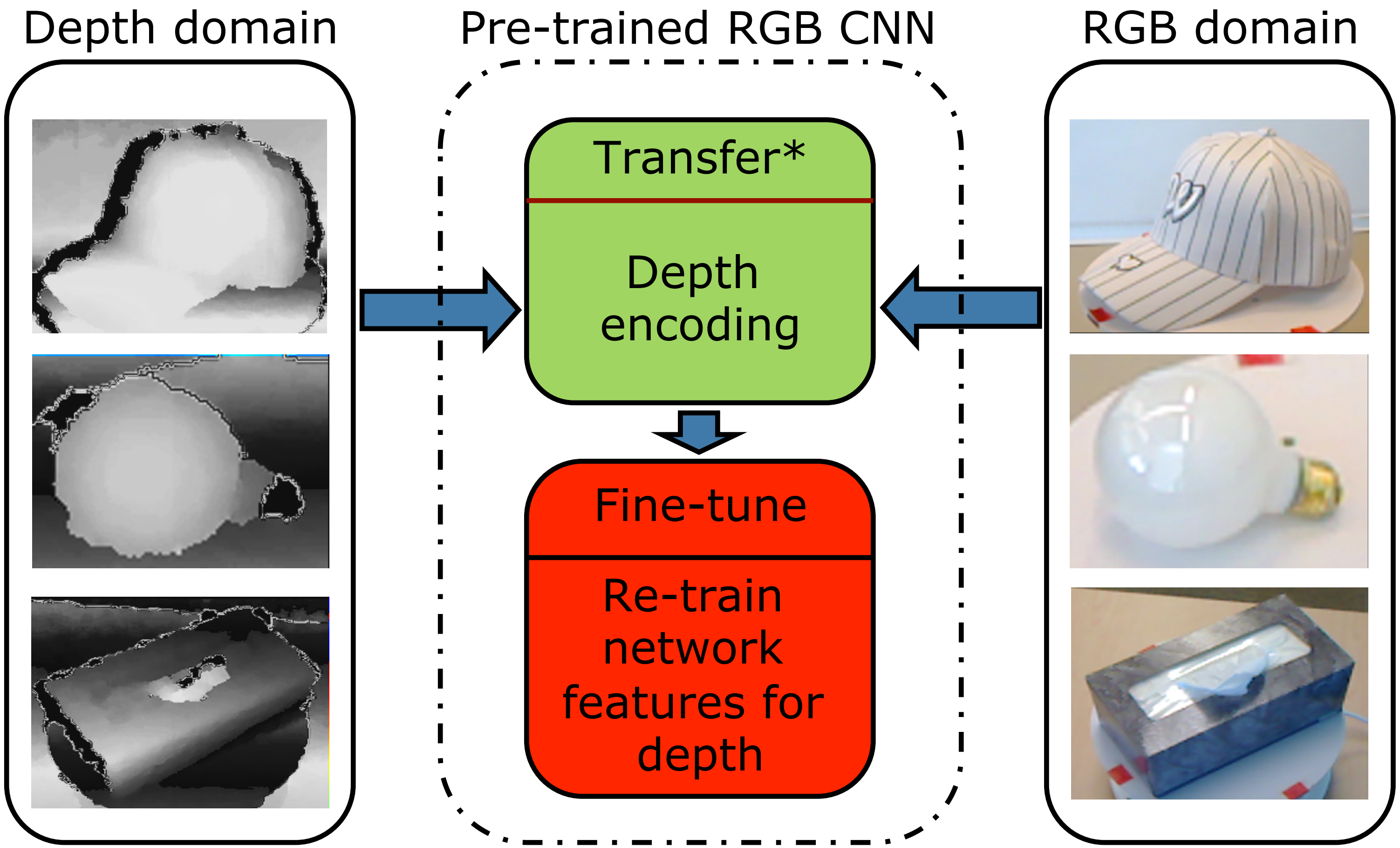
Large portion of RGB-D data is recorded under controlled settings

- How to improve recognition in real-world scenes when the training data is "clean"?
- How to deal with sensor noise from RGB-D sensors?

Solution: Transfer Deep RGB Features to Depth Domain

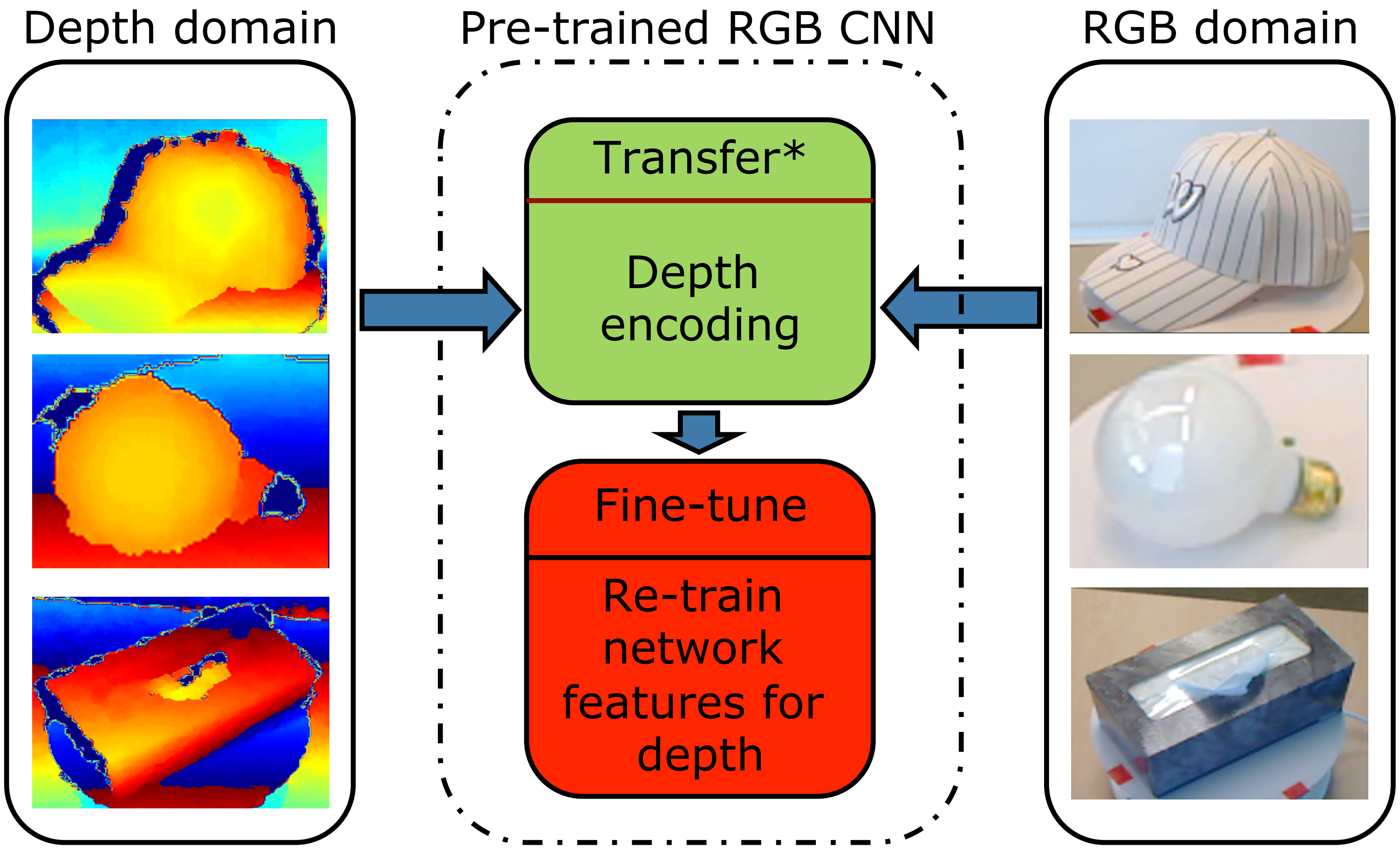
Both domains share similar features such as edges, corners, curves, …

Solution: Transfer Deep RGB Features to Depth Domain



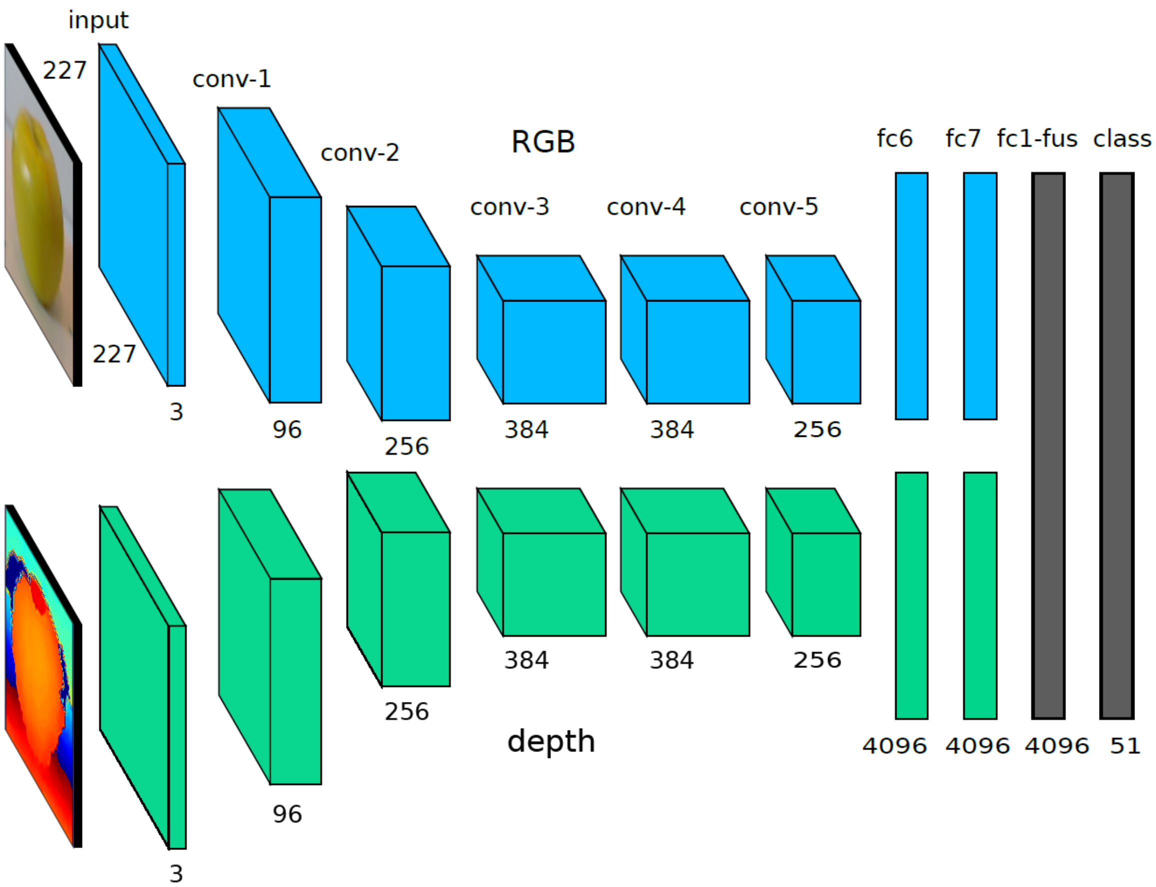
* Similar to [Schwarz et. al 2015, Gupta et. al 2014]

Solution: Transfer Deep RGB Features to Depth Domain



* Similar to [Schwarz et. al 2015, Gupta et. al 2014]

Multimodal Deep Convolutional Neural Network



- Two input modalities
- Late fusion network
- 10 convolutional layers
- Max pooling layers
- 4 fully connected layers
- Softmax classifier

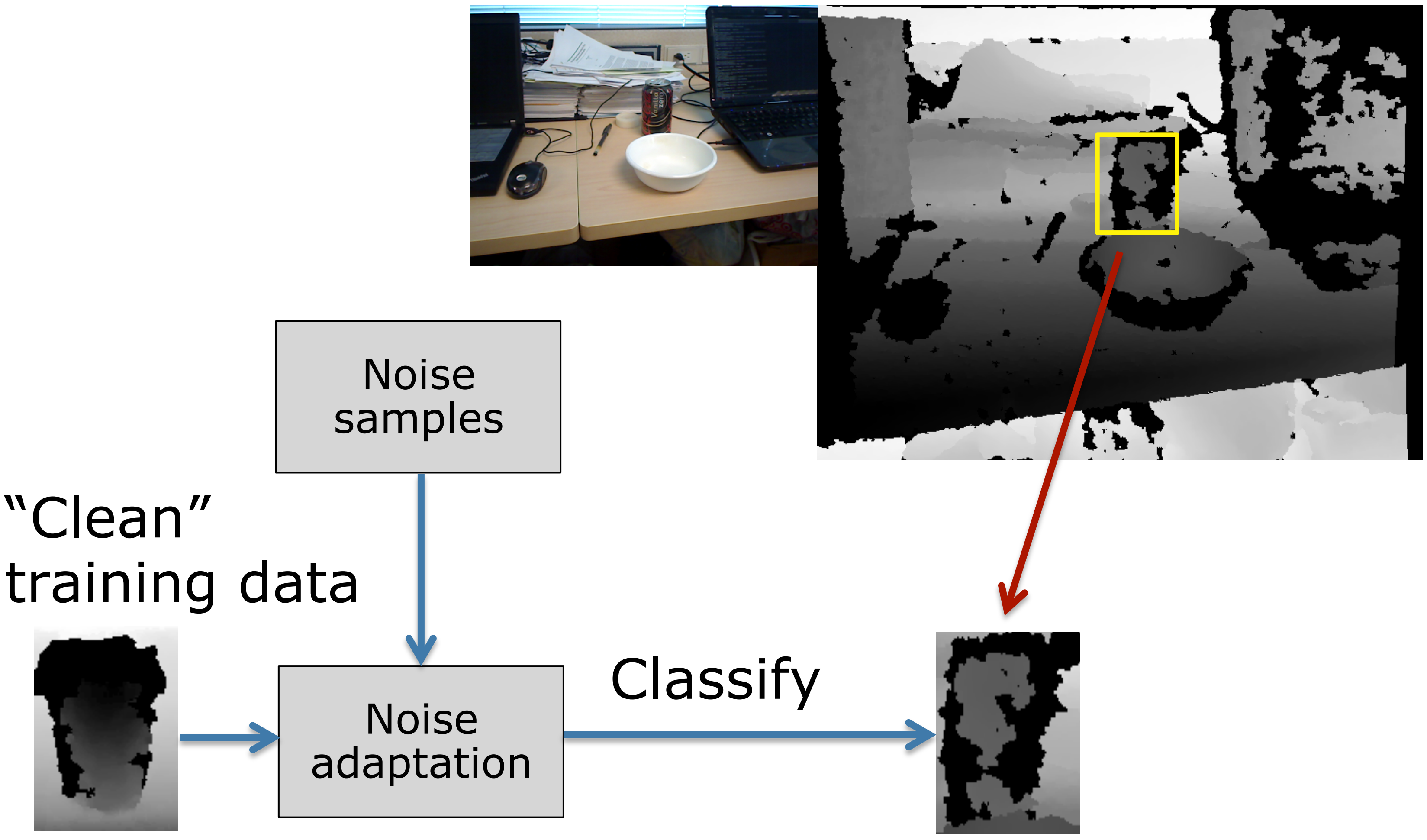
2xAlexNet + fusion net

How to Encode Depth Images?

- Distribute depth over color channels
  - Compute min and max value of depth map
  - Shift depth map to min/max range
  - Normalize depth values to lie between 0 and 255
  - Colorize image using jet colormap (red = near , blue = far)
- Depth encoding improves recognition accuracy by 1.8 percentage points

RGB | Raw depth | Colorized depth

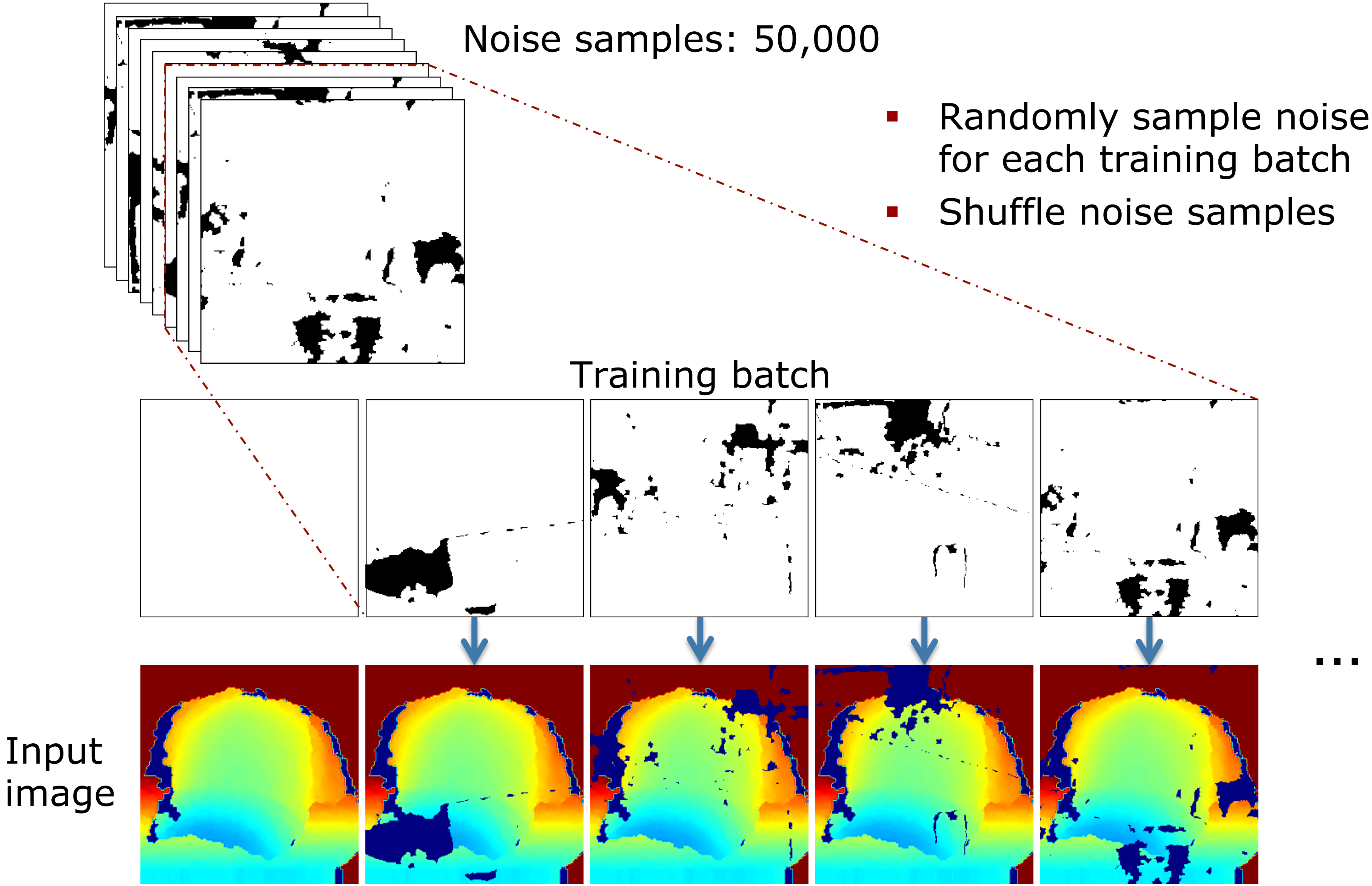
Solution: Noise-aware Depth Feature Learning



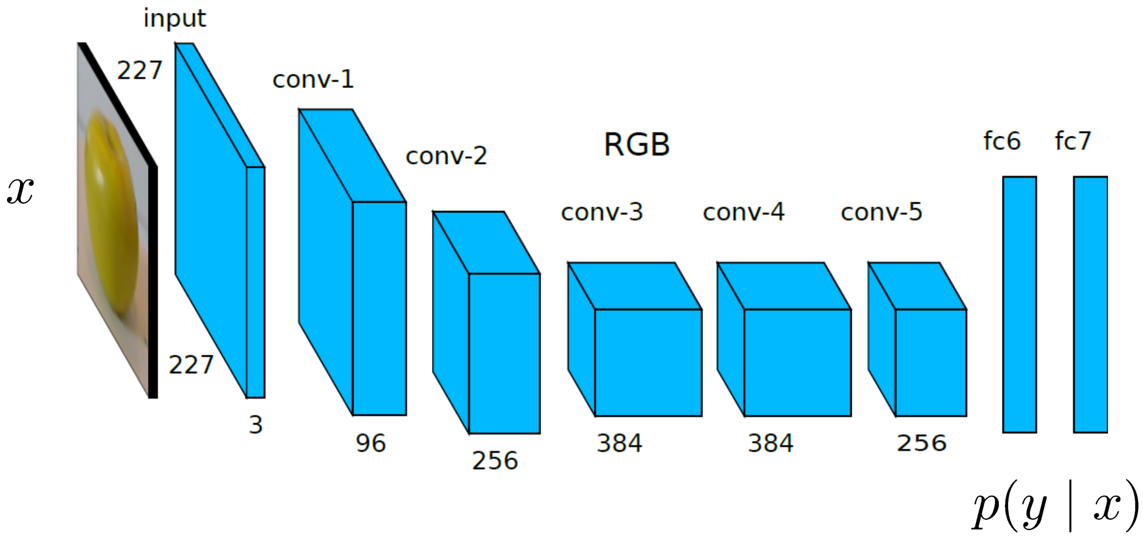
Training with Noise Samples

Noise samples: 50,000

- Randomly sample noise for each training batch
- Shuffle noise samples

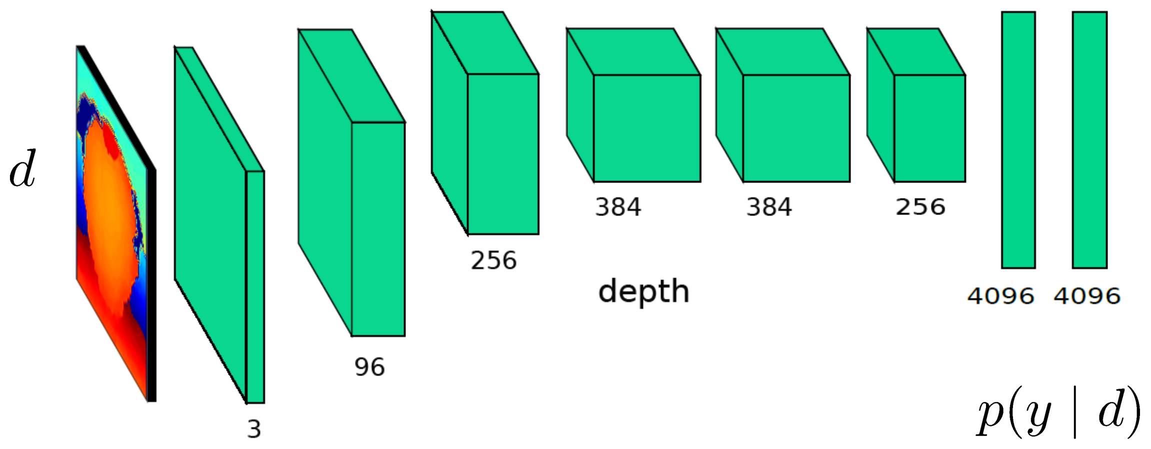


RGB Network Training



- Maximum likelihood learning
- Fine-tune from pre-trained AlexNet weights

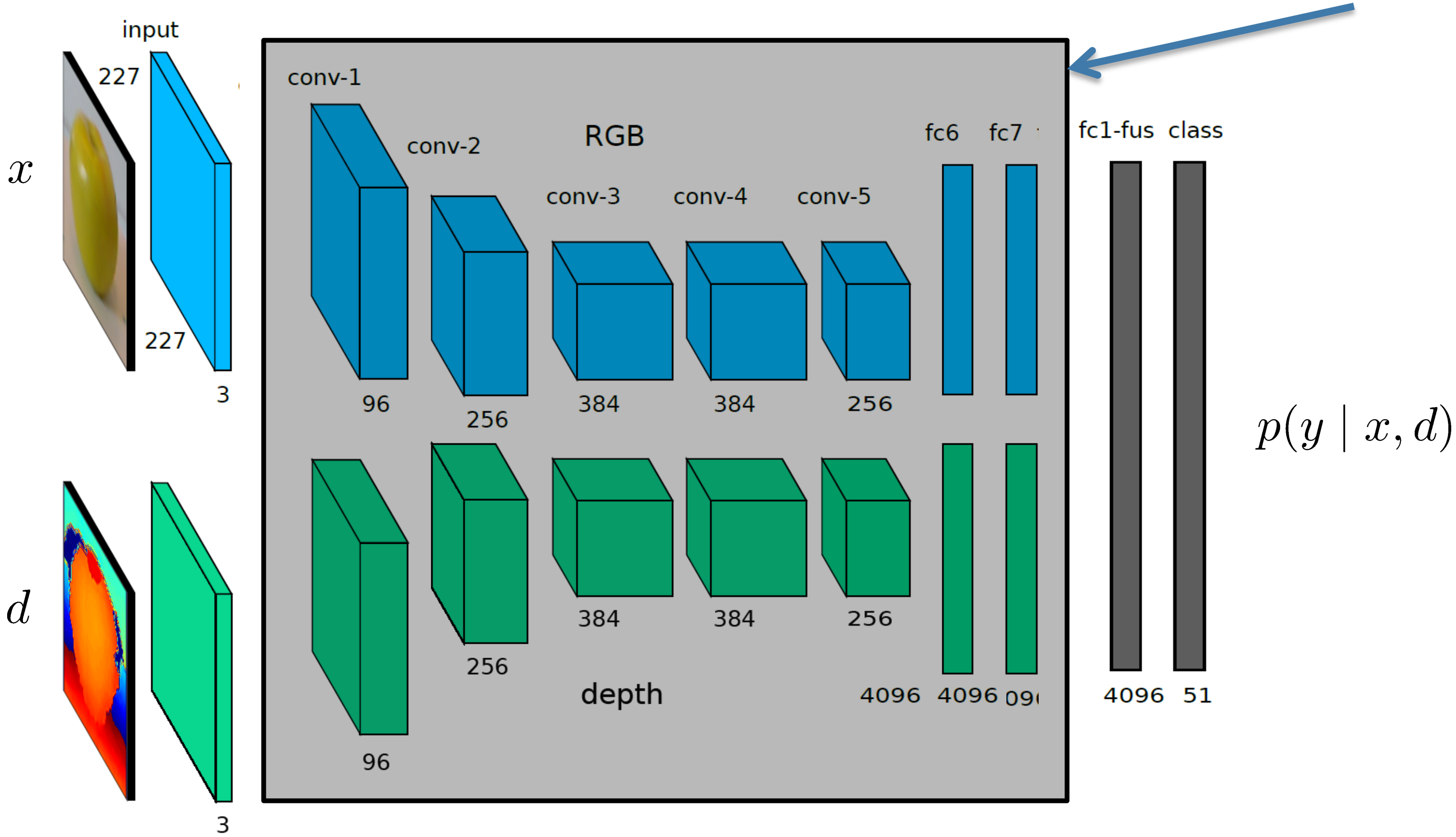
Depth Network Training



- Maximum likelihood learning
- Fine-tune from pre-trained AlexNet weights

Fusion Network Training

- Fusion layers automatically learn to combine feature responses of the two network streams
- During training, weights in first layers stay fixed



UW RGB-D Object Dataset

[Lai et. al, 2011]

Category-Level Recognition [%] (51 categories)

Method	RGB	Depth	RGB-D
CNN-RNN	80.8	78.9	86.8
HMP	82.4	81.2	87.5
CaRFs	N/A	N/A	88.1
CNN Features	83.1	N/A	89.4

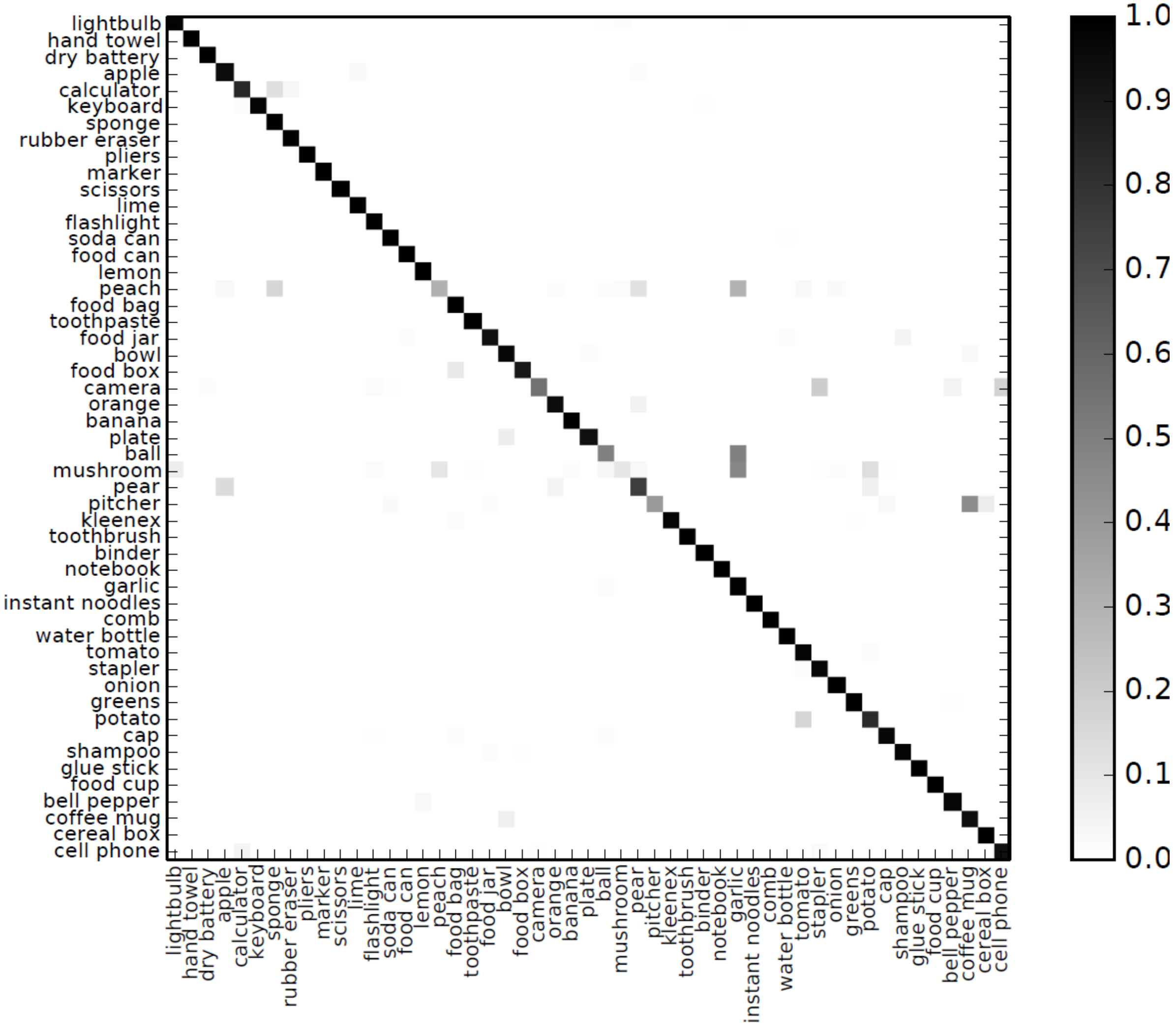
UW RGB-D Object Dataset

[Lai et. al, 2011]

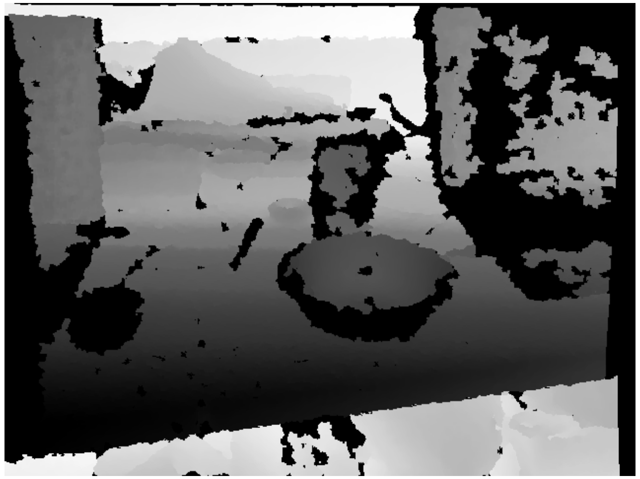
Category-Level Recognition [%] (51 categories)

Method	RGB	Depth	RGB-D
CNN-RNN	80.8	78.9	86.8
HMP	82.4	81.2	87.5
CaRFs	N/A	N/A	88.1
CNN Features	83.1	N/A	89.4
This work, Fus-CNN	**84.1**	**83.8**	**91.3**

Confusion Matrix

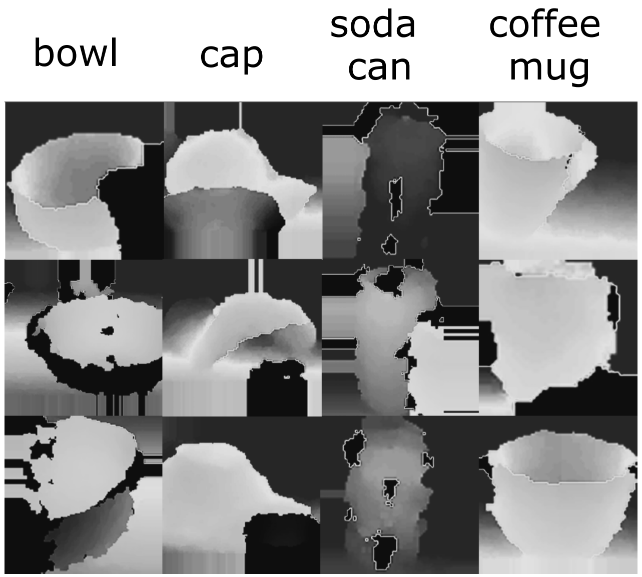


Recognition in Noisy RGB-D Scenes



Recognition using

annotated bounding boxes



Noise adapt. = correct prediction

No adapt. = false prediction

Category-Level Recognition [%] depth modality (6 categories)

Noise adapt.	flash-light	cap	bowl	soda can	cereal box	coffee mug	class avg.
-	**97.5**	68.5	66.5	66.6	96.2	79.1	79.1
√	96.4	**77.5**	**69.8**	**71.8**	**97.6**	**79.8**	**82.1**

Deep Learning for RGB-D Object Recognition

- Novel RGB-D object recognition for robotics
- Two-stream CNN with late fusion architecture
- Depth image transfer and noise augmentation training strategy
- State of the art on UW RGB-D Object dataset for category recognition: 91.3%
- Recognition accuracy of 82.1% on the RGB-D Scenes dataset

Choosing Smartly: Adaptive Multimodal Fusion for Object Detection in Changing Environments

Oier Mees,

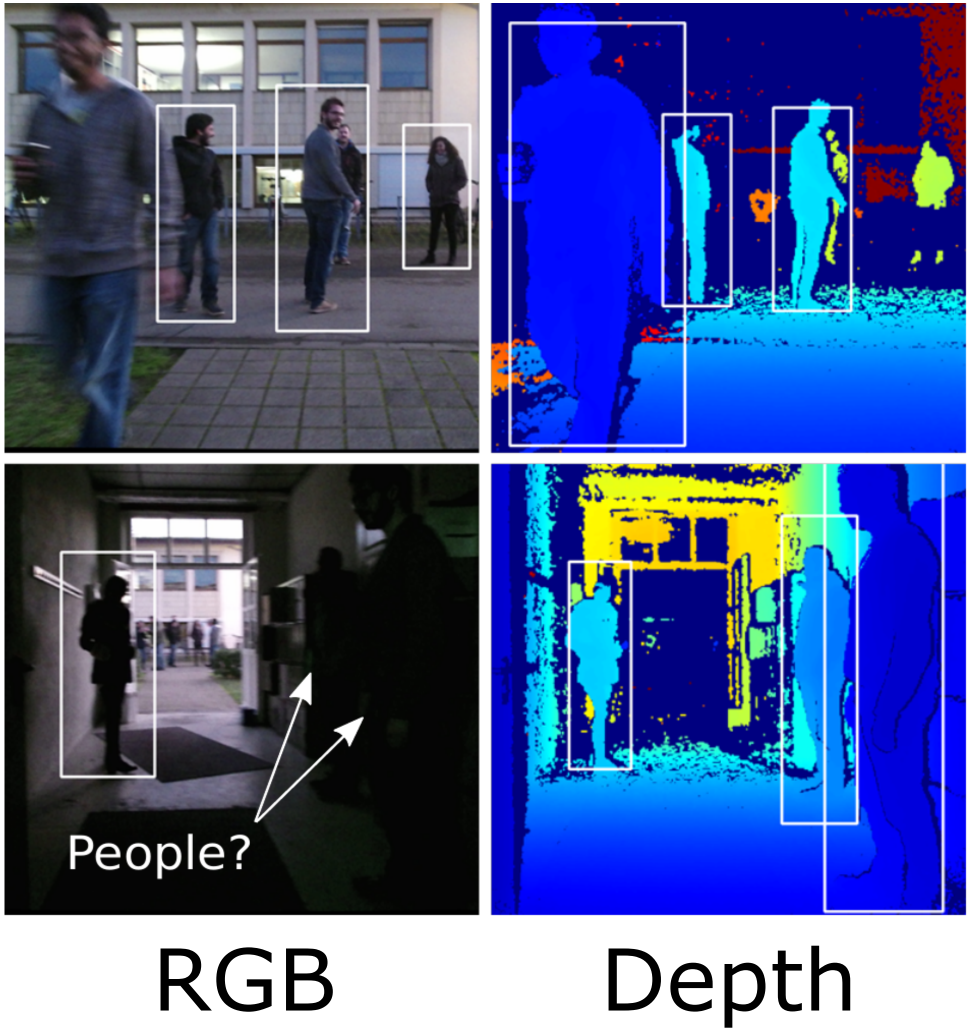
Andreas Eitel, Wolfram Burgard

UNI FREIBURG

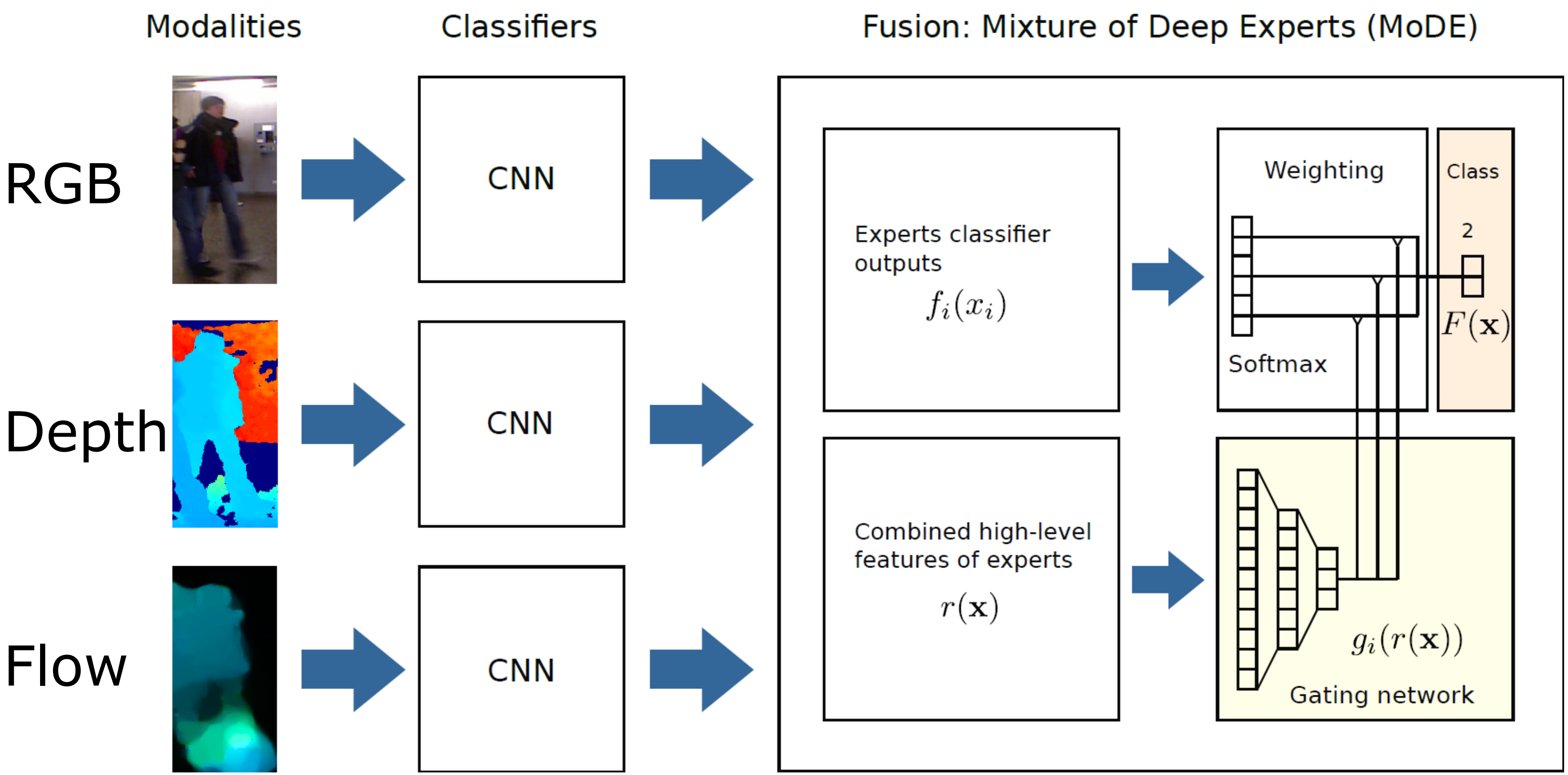
AIS Autonomous Intelligent Systems

Object Detection in Changing Environments

- How to com-bine different modalities for detection
- Sensor noise changes with the environ-ment



Mixture of Deep Neural Networks for People Detection



General learning approach to fuse different modalities such as RGB + depth + optical flow

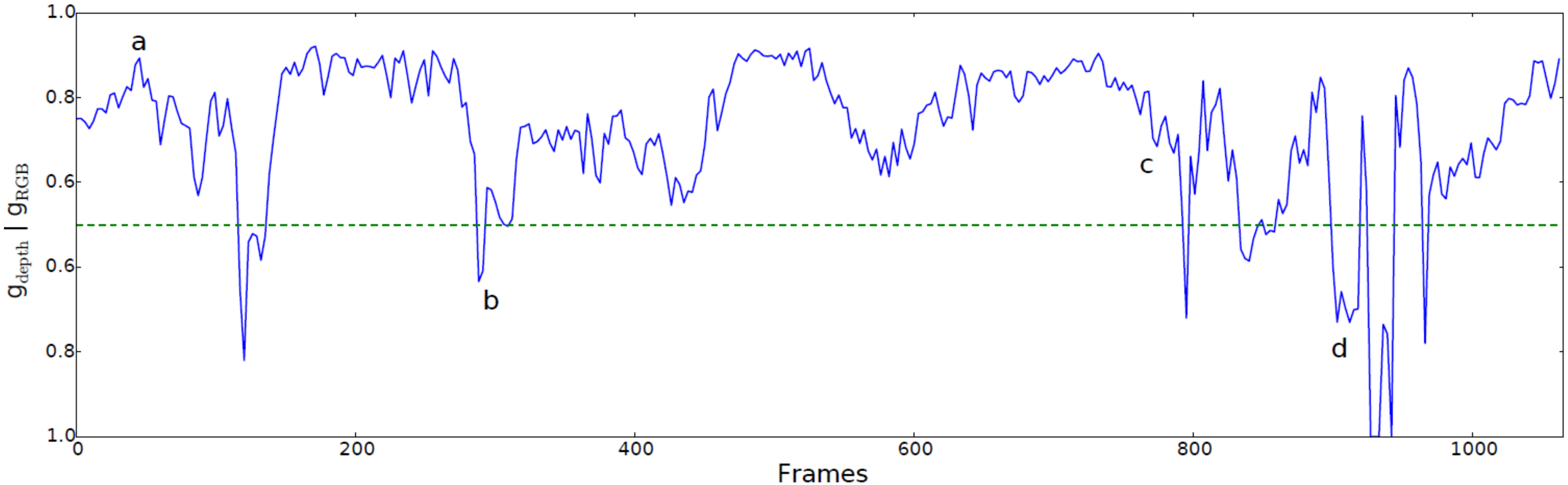
Quantitative Results

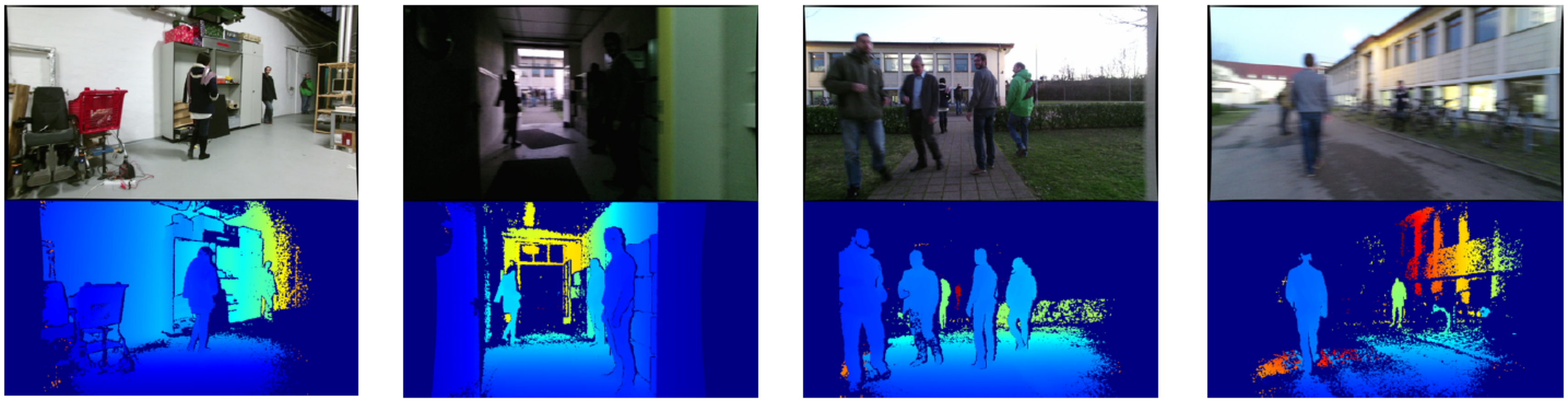
- Comparison of fusion approaches
  - Late fusion approach with additional two-layer fusion network on top of expert networks
  - Hierarchical mixture of experts
  - Our Mixture of Experts approach (MoDE)
- Adaptive fusion improves performance

Input	Method	AP/Recall	EER
Depth	HOD [Spinello et al. 2012]	-	56.3
RGB-D	HGE [Spinello et al. 2012]	-	87.4
RGB-D-Flow	Ours, CifarNet late fusion	88.0/88.4	88.2
RGB-D-Flow	Ours, MoDE	**88.6/90.0**	**89.3**

Adaptive Weighting in Test Set

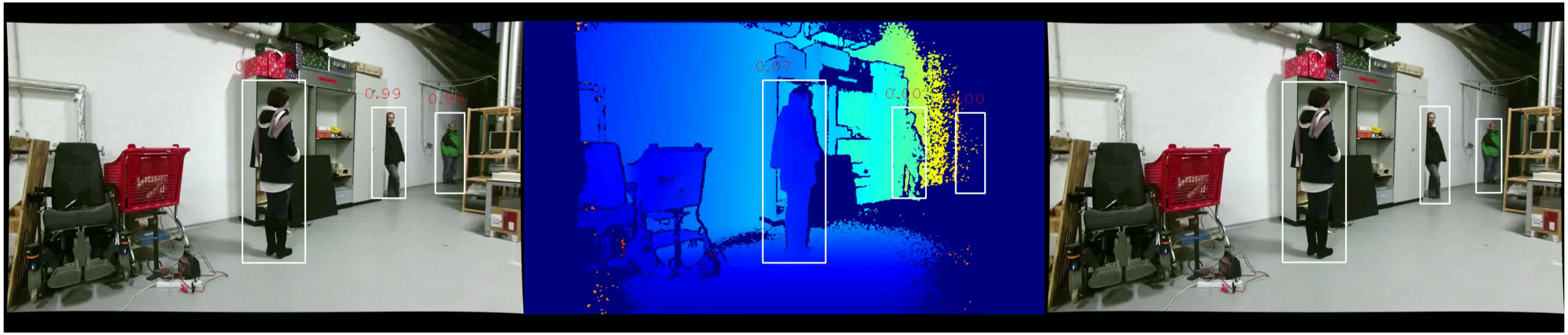
Mean gating assignments per frame





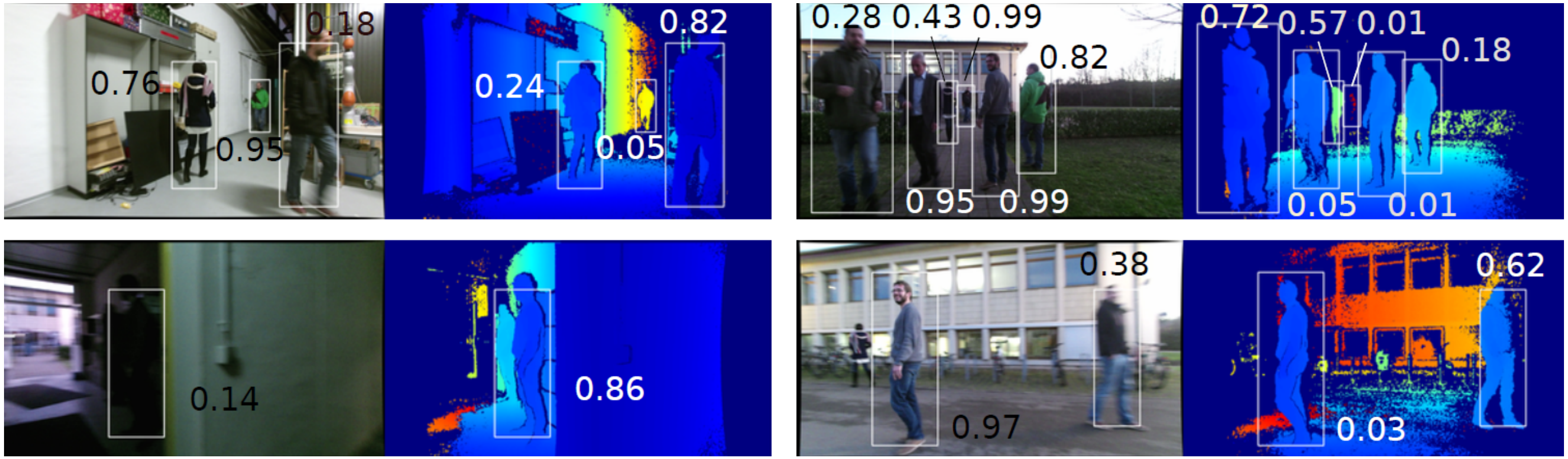
a b c d

Example Application



Qualitative Results

Gating assignments per bounding box



Quantitative Results

Performance of single and multimodal networks

Input	Method	AP/Recall
RGB	GoogLeNet-xxs	70.0/79.6
RGB	CifarNet	55.3/62.9
Depth	GoogLeNet-xxs	71.6/78.9
RGB-D	GoogLeNet-xxs average	71.1/73.9
RGB-D	GoogLeNet-xxs late fusion	72.0/76.3
RGB-D	GoogLeNet-xxs MoDE	**80.4/81.1**

Deep Learning for Human Part Discovery in Images

Gabriel L. Oliveira, Abhinav Valada, Claas Bollen, Wolfram Burgard, Thomas Brox

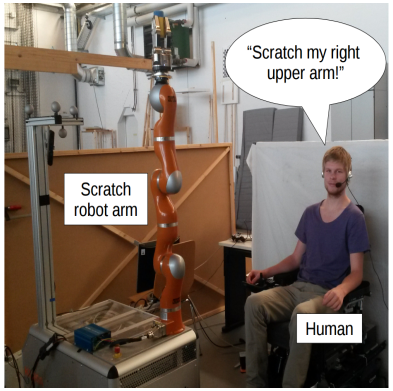
UNI FREIBURG

AIS Autonomous Intelligent Systems

[to be presented at ICRA 2016]

Deep Learning for Human Part Discovery in Images

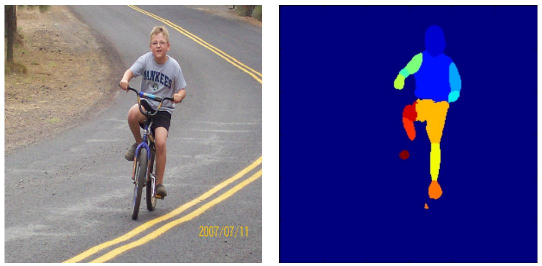
- Human-robot interaction



- Robot rescue

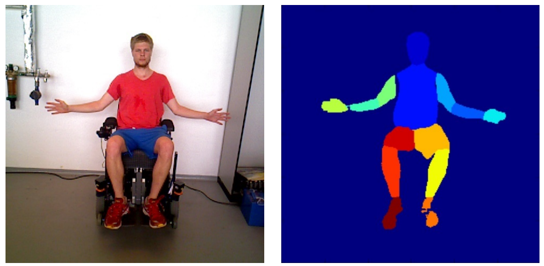
Deep Learning for Human Part Discovery in Images

- Dense prediction can provide pixel classification of the image
- Human part segmentation is naturally challenging due to
  - Non-rigid aspects of the body
  - Occlusions



PASCAL Parts

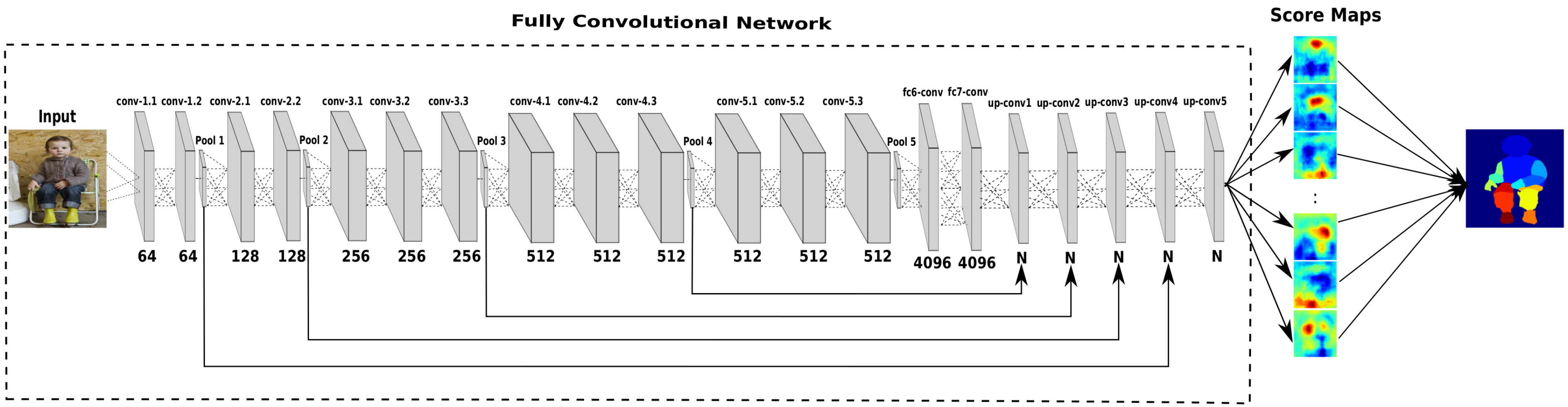
MS COCO



Freiburg Sitting

Network Architecture

- Fully convolutional network
  - Contraction and expansion of network input
  - Up-convolution operation for expansion
- Pixel input, pixel output



Experiments

- Evaluation of approach on
  - Publicly available computer vision datasets
  - Real-world datasets with ground and aerial robots
- Comparison against state-of-the-art semantic segmentation approach: FCN proposed by Long et al. [1]

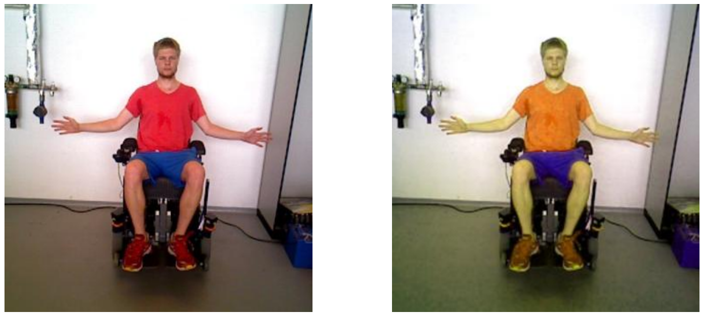
[1] John Long, Evan Shelhamer, Trevor Darrell, CVPR 2015

Data Augmentation

Due to the low number of images in the available datasets, augmentation is crucial

- Spatial augmentation (rotation + scaling)

- Color augmentation



PASCAL Parts Dataset

- PASCAL Parts, 4 classes, IOU

Method	Head	Torso	Arms	Legs	All
FCN	70.74	60.62	48.44	50.38	57.35
Ours	75.08	64.81	55.61	56.72	63.03
Ours (Spatial)	80.49	74.39	67.17	70.39	73.00
Ours (Spatial + Color)	**83.24**	**79.41**	**73.73**	**76.52**	**78.23**

- PASCAL Parts, 14 classes, IOU

Method	Head	Torso	L U arm	L LW arm	L hand	R U hand	R LW arm	R hand	R U leg	R LW leg	R foot	L U leg	L LW leg	L foot	Mean
FCN	74.0	66.2	56.6	46.0	34.1	58.9	44.1	31.0	49.3	44.5	40.8	48.5	47.6	41.2	53.1
Ours (Spatial)	81.8	78.0	69.5	63.1	59.0	71.2	63.0	58.7	65.4	60.6	52.0	67.9	60.3	50.0	66.9
Ours (Spatial+Color)	**84.0**	**81.5**	**74.1**	**68.0**	**64.0**	**75.4**	**67.4**	**61.9**	**72.4**	**67.1**	**56.9**	**73.0**	**66.1**	**57.7**	**71.7**

R = right, L = left, U = upper, LW = lower.

Freiburg Sitting People Part Segmentation Dataset

- We present a novel dataset for human part segmentation in wheelchairs

Input Image

Ground Truth

Segmentation mask

Method	Accuracy	IOU
FCN	59.69	43.17
Ours (Trained on PASCAL)	78.04	59.84
Ours (2 people train - 4 people test)	**81.78**	**64.10**

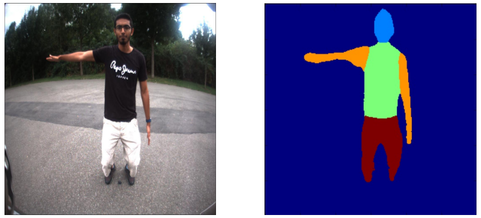
Robot Experiments

- Range experiments with ground robot
- Aerial platform for disaster scenario (Segmentation under severe body occlusions)

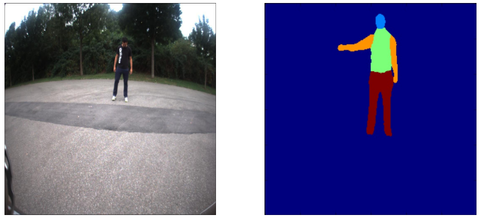
Range Experiments

Recorded using Bumblebee camera

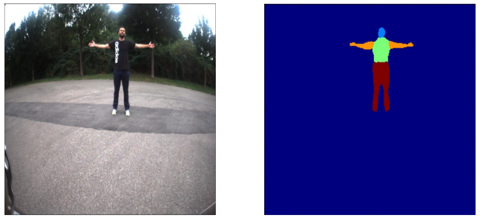
- Robust to radial distortion
- Robust to scale



(a) 1.0 meter



(b) 2.0 meters

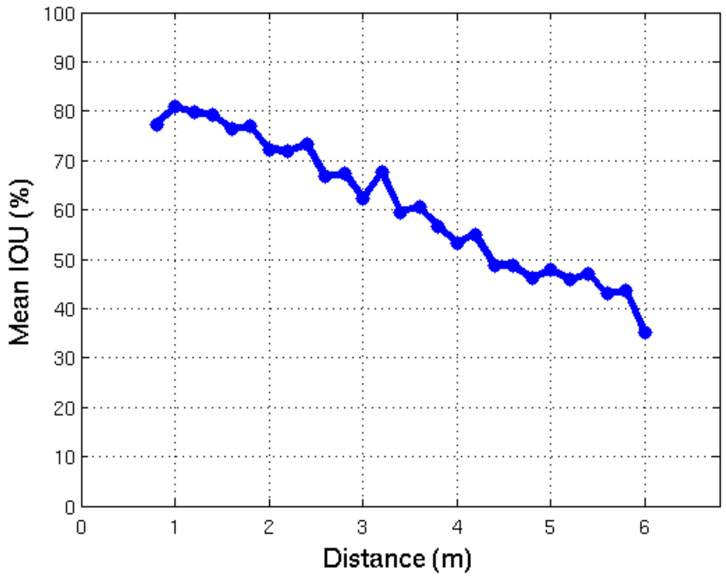


(c) 3.0 meters

(d) 4.0 meters

(e) 5.0 meters

(f) 6.0 meters



Freiburg People in Disaster

Dataset designed to test severe occlusions

Input Image

Ground Truth

Segmentation mask

Method	Head	Torso	Arms	Legs	IOU
FCN	52.71	62.49	35.04	43.25	43.20
Ours	**80.56**	**79.45**	**63.93**	**64.91**	**71.99**

Application to Obelix Data

Efficient Deep Models for Monocular Road Segmentation

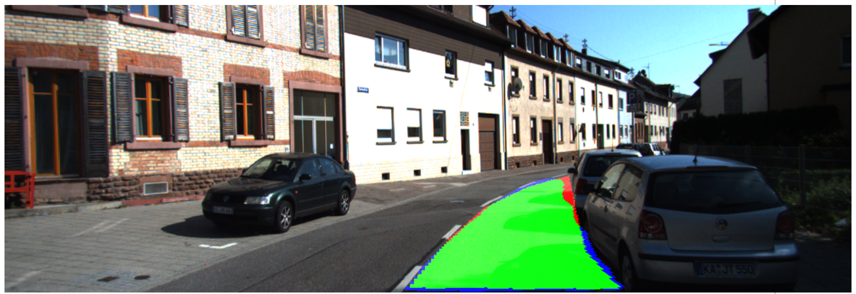
Gabriel Leivas Oliveira

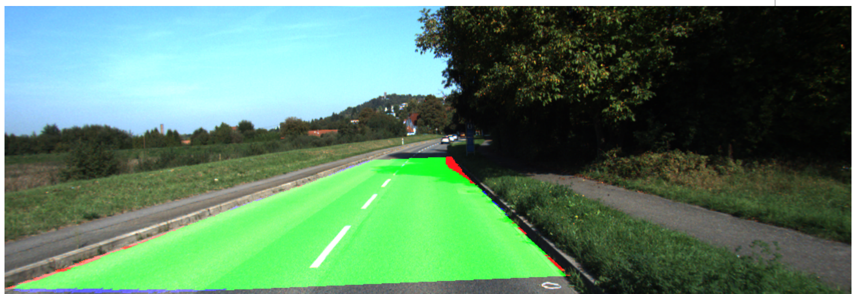
Wolfram Burgard

Thomas Brox

University of Freiburg, Germany

Motivation

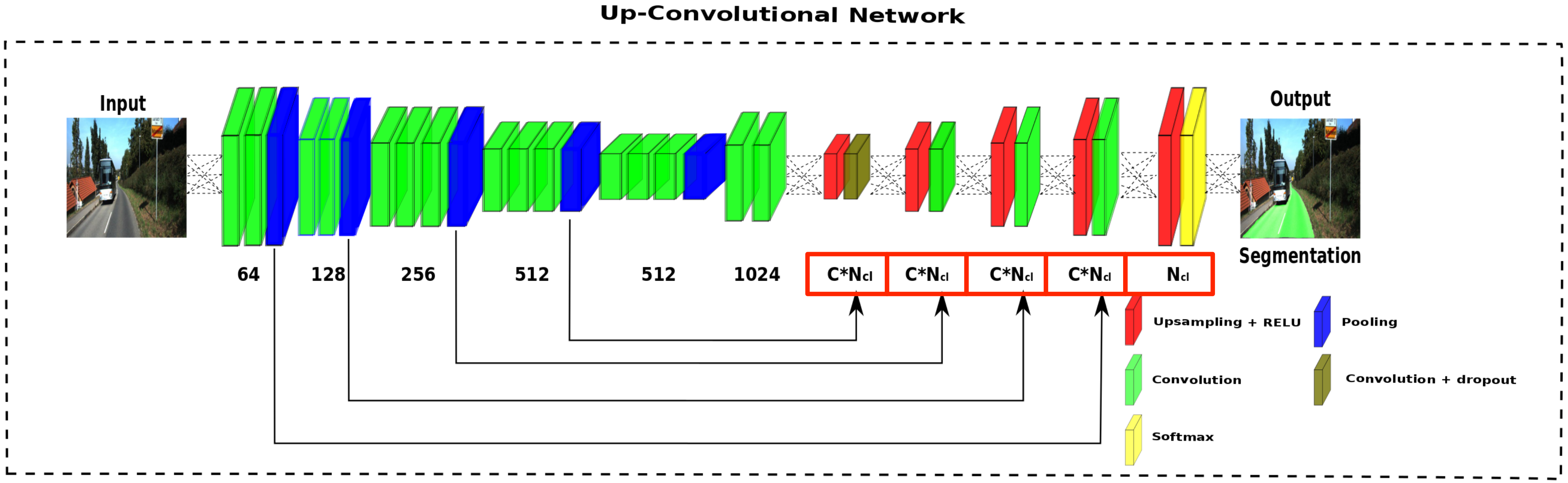




Architecture

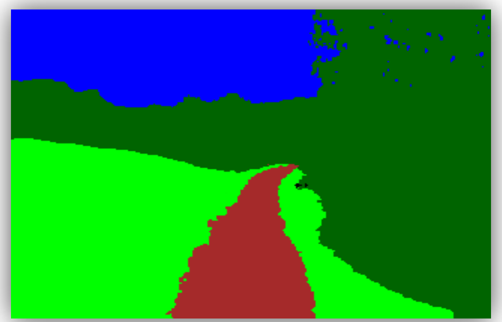
More parameters expansion

1-to-C*Ncl filters per refinement

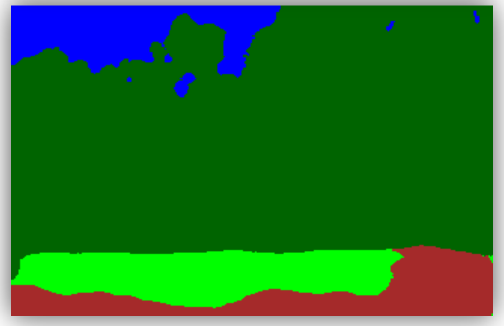


Terrain Classification using a Late Fusion DCNN Architecture

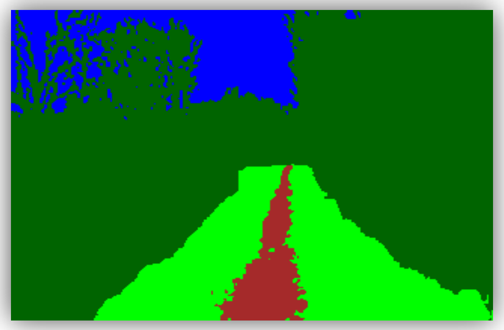
Snow



Glare

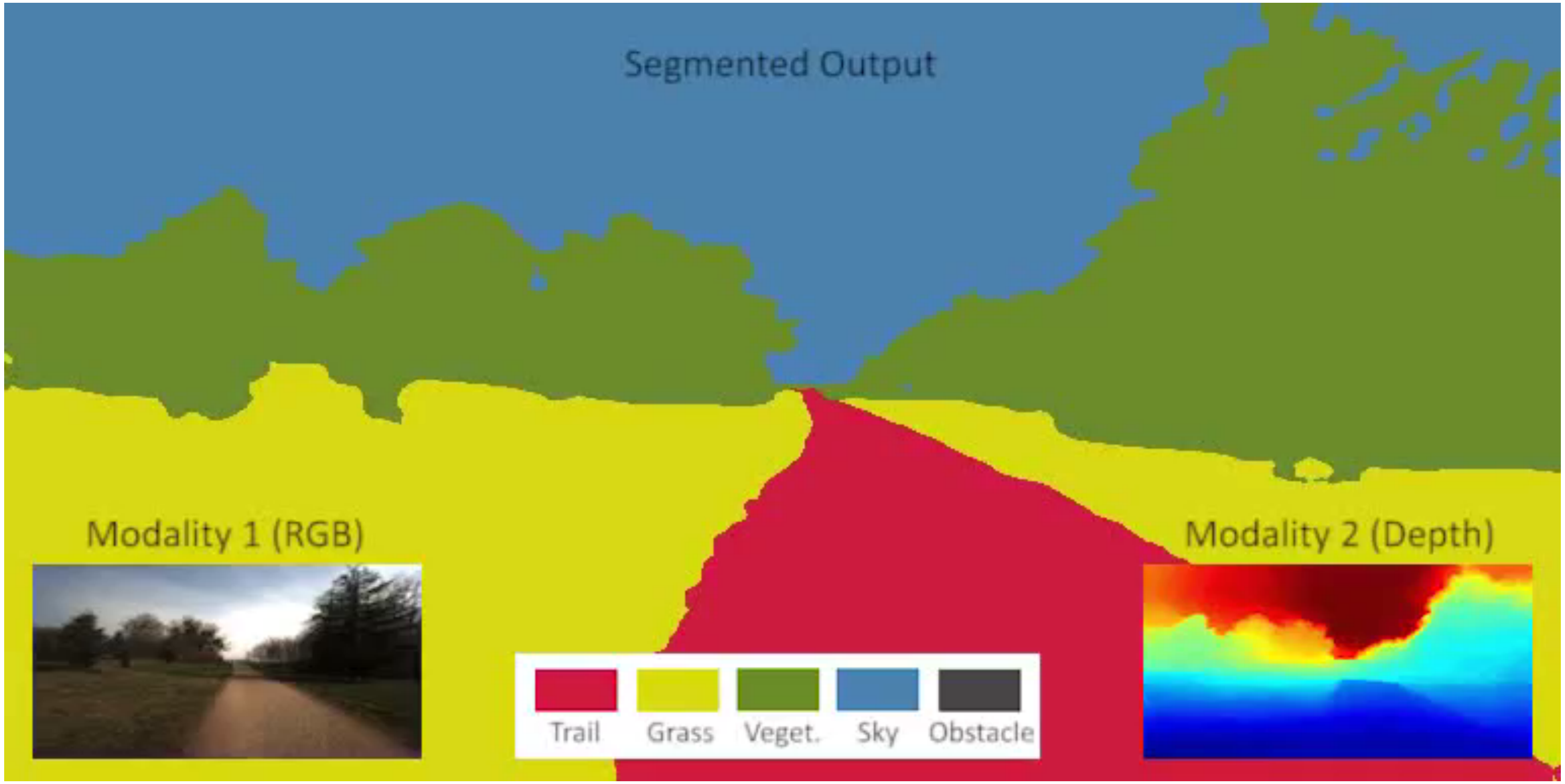


Low Lighting



Autonomous Navigation in Outdoor Areas

Terrain Classification using a Late Fusion DCNN Architecture



Semantic Segmentation of Moving Objects using Convolutional Neural Networks

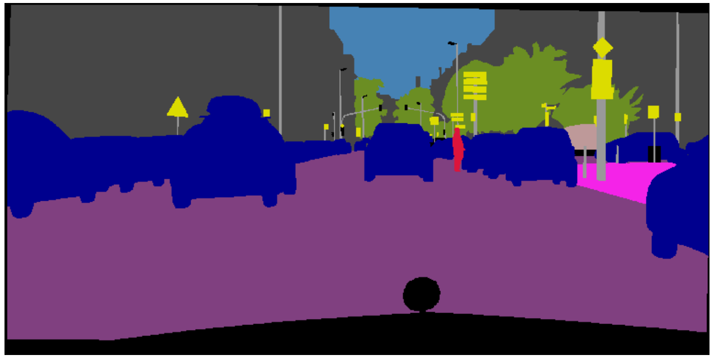
Johan Vertens, Abhinav Valada, Wolfram Burgard

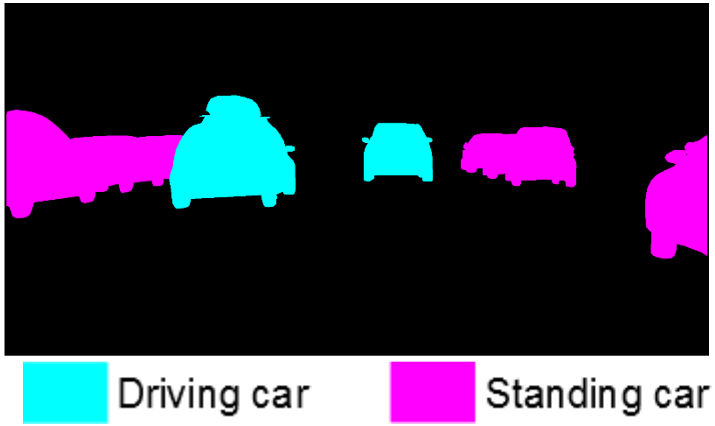
UNI FREIBURG

AIS Autonomous Intelligent Systems

Goal

1. Robust and fast semantic segmentation of driving scenarios
2. Semantic motion segmentation



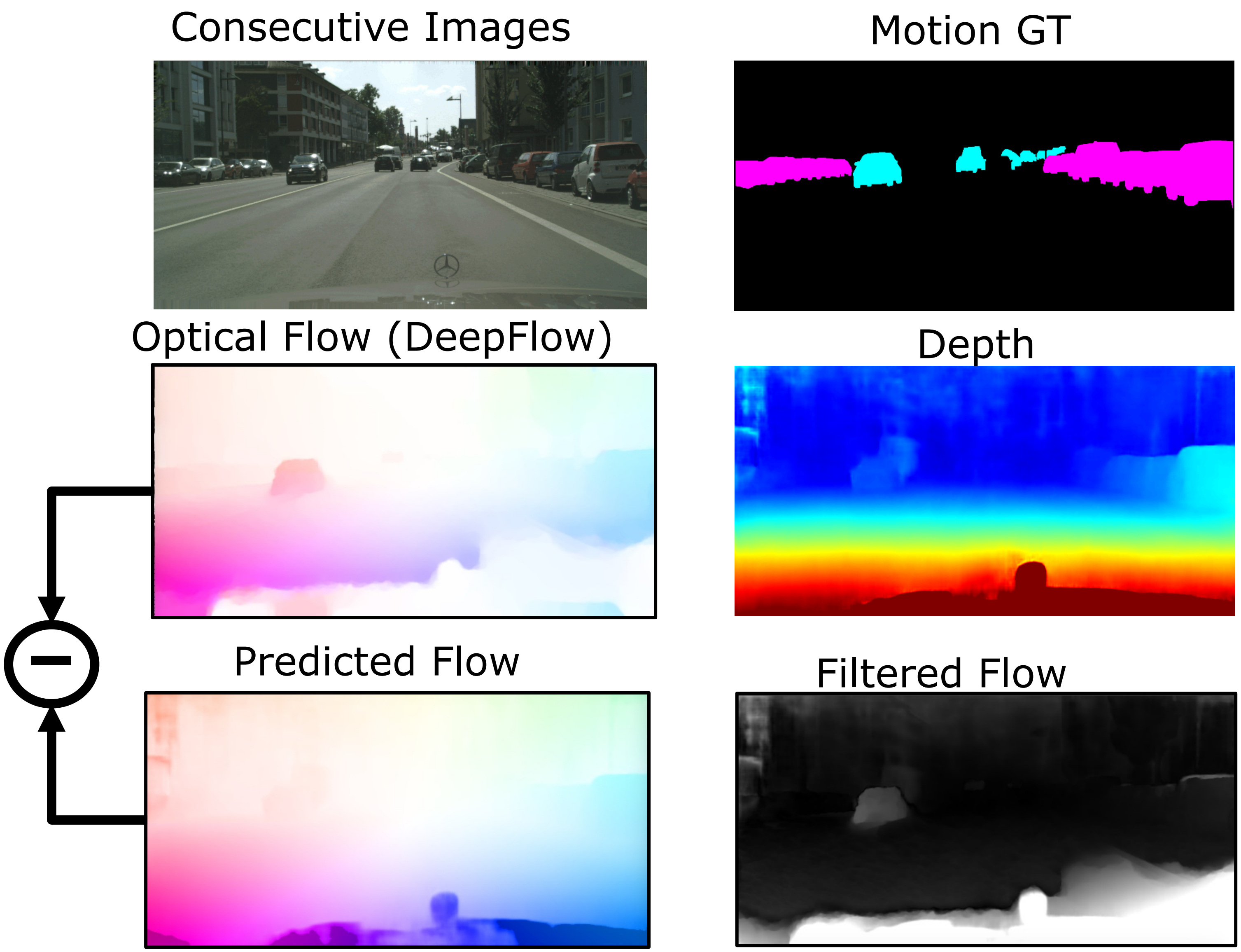


For semantic motion segmentation we consider the “car”-class.

Semantic Motion Segmentation

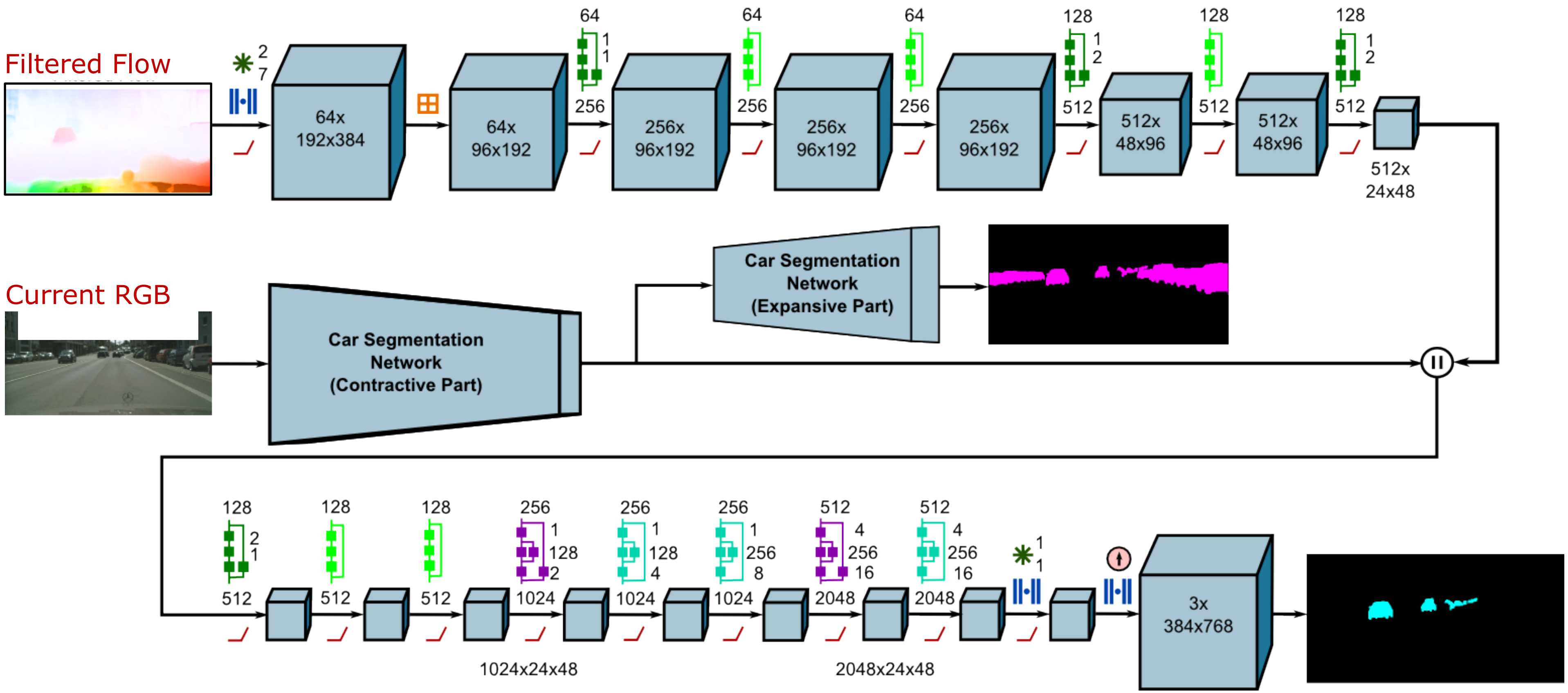
- Fuse semantic features and generate motion features within a CNN
- Two architectures:
  1. FiltFlow-Net: Takes precomputed motion features
  2. Siamese-Net: Motion features are learned entirely

FiltFlow-Net

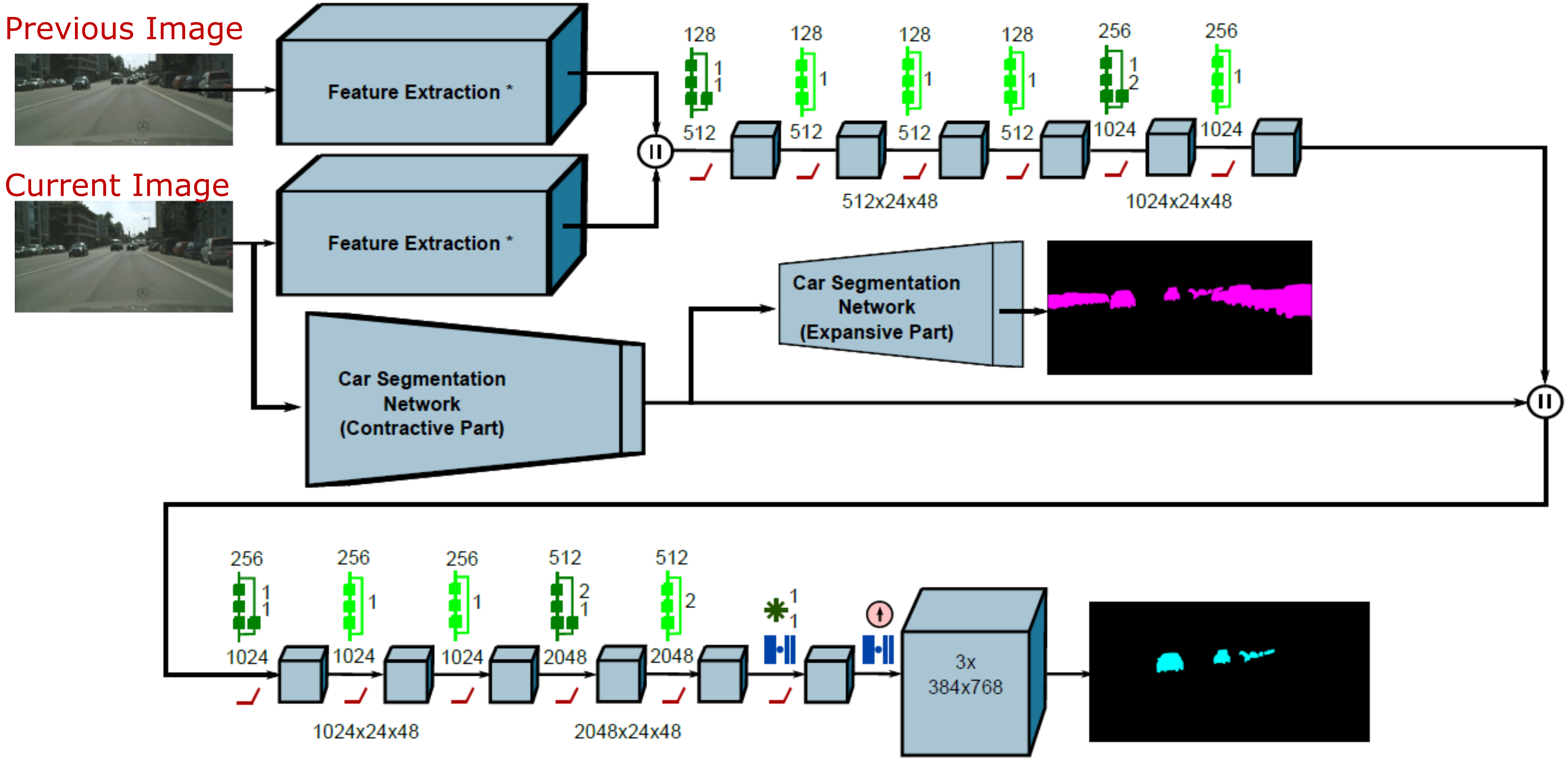


FiltFlow-Net: Architecture

- Embedded MultiNet
- Predicts moving cars



Siamese-Net: Architecture



Comparison of Architectures

- FiltFlow-Net achieves an improvement of 6.26 IoU over Siamese-Net

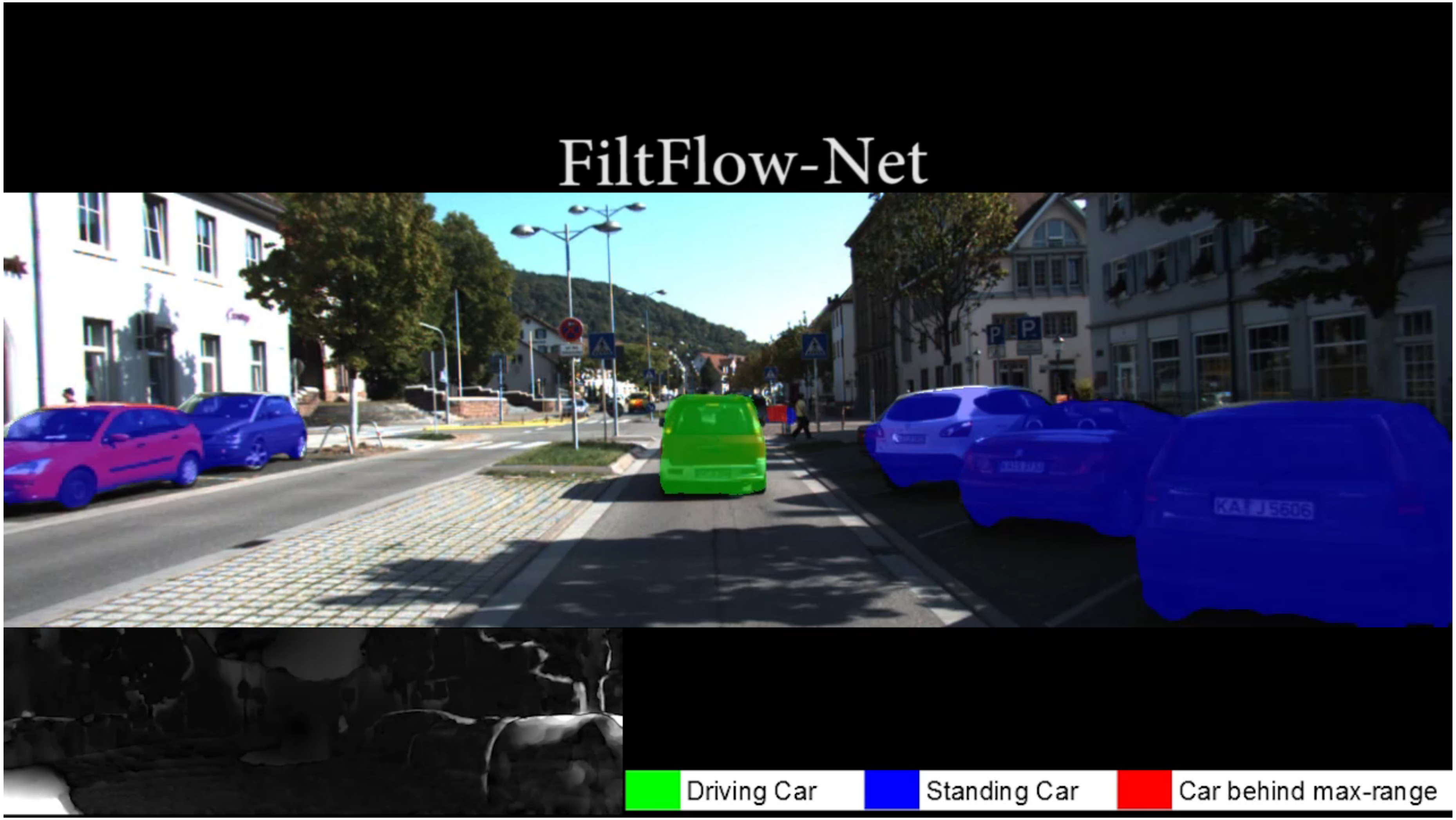
Approach	IoU	AP	FPR	FNR
FiltFlow-Net	**83.44**	**94.67**	**04.39**	**11.41**
Siamese-Net	77.18	89.64	09.10	13.68

Comparison of Inference Time

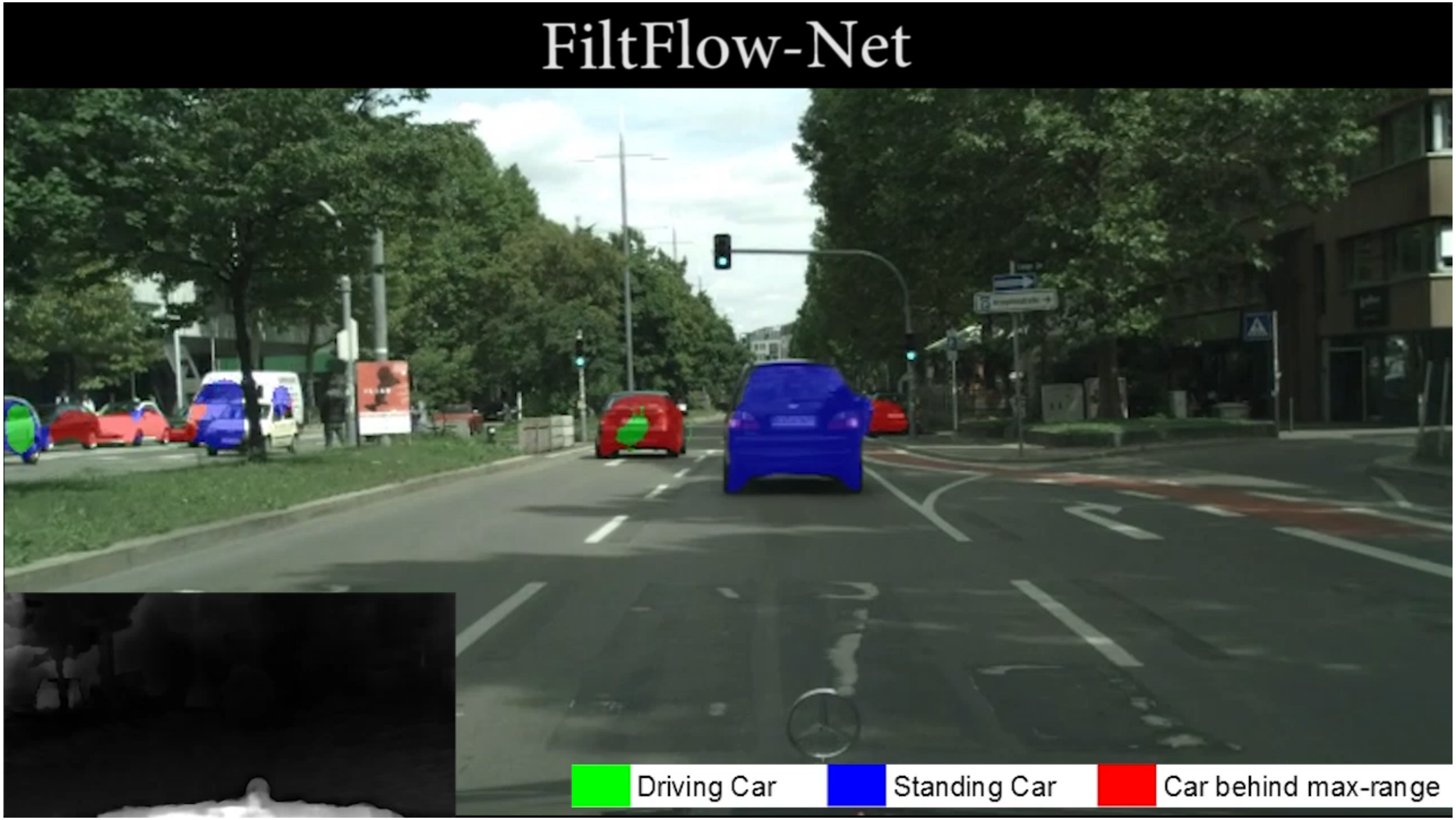
- Siamese-Net has much lower inference time

	FiltFlow-Net	Siamese-Net
Optical Flow (DeepFlow)	11.2s	-
Predicted Optical Flow	112ms	-
Neural Network	87.4ms	83.3ms
Total	11.4s	**83.3**ms

KITTI Motion Segmentation



Cityscapes Motion Segmentation



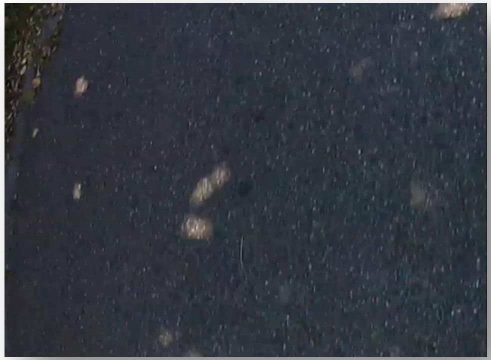
Deep Feature Learning for Acoustics-based Terrain Classification

Abhinav Valada, Luciano Spinello, Wolfram Bugard

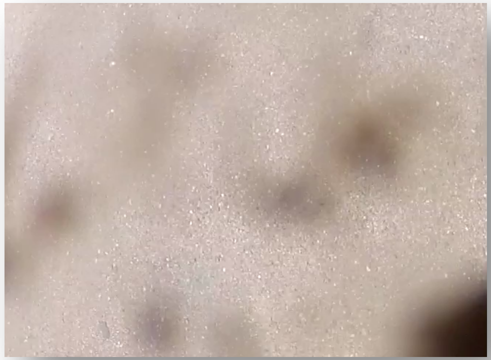
[ISRR 2015]

Motivation

Lighting Variations



Shadows



Dirt on Lens

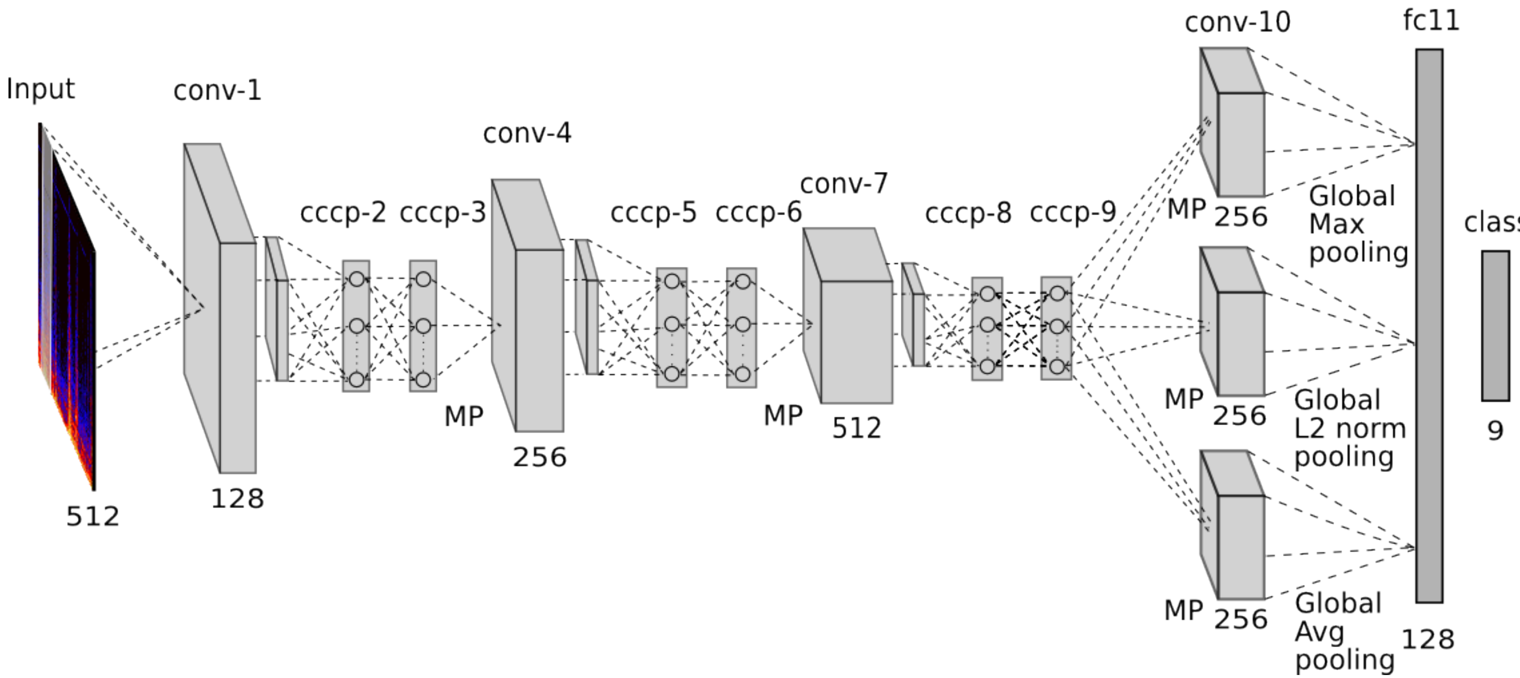
Optical sensors are highly sensitive to visual changes

Motivation

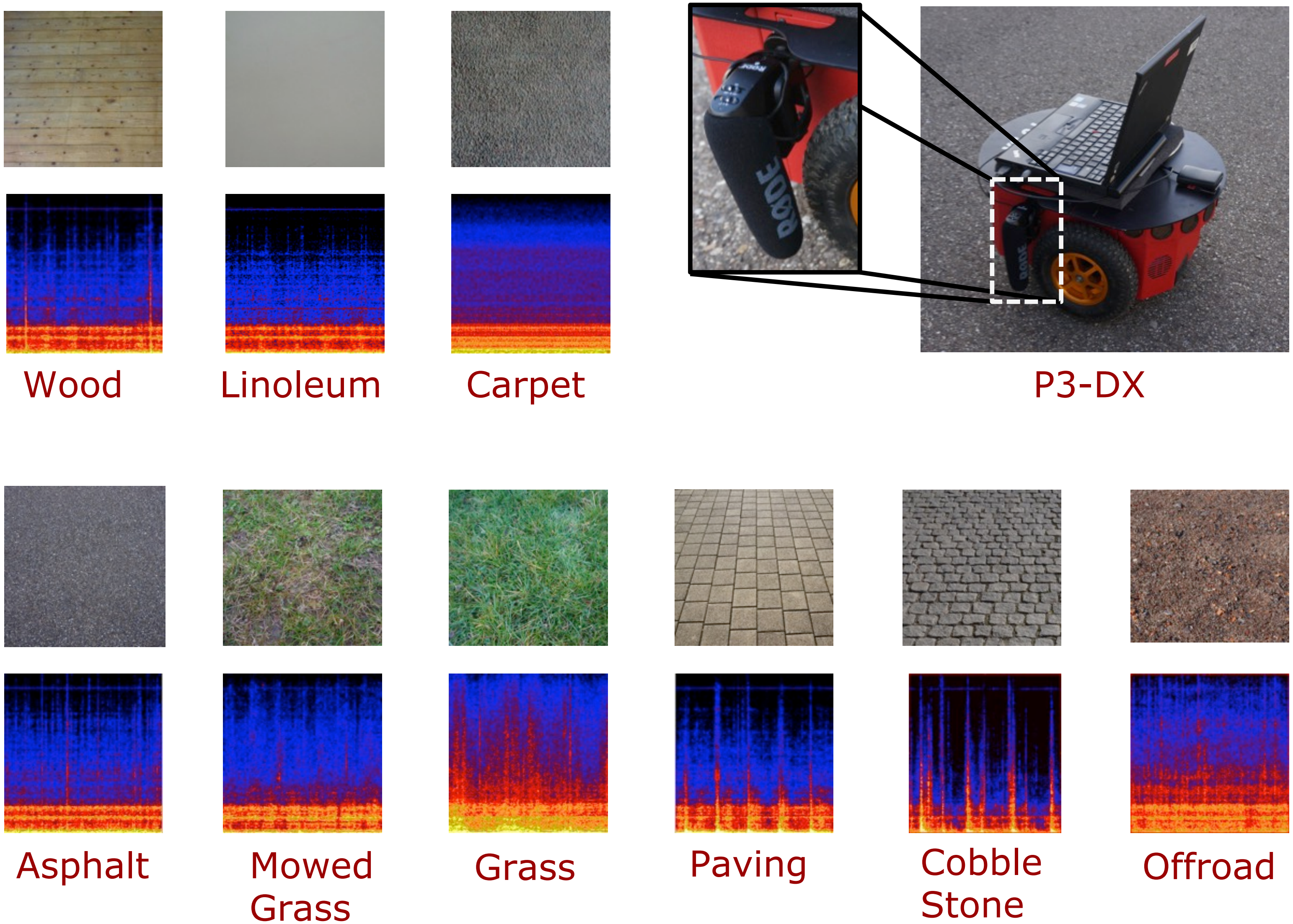
Use sound from vehicle-terrain interactions to classify terrain

Network Architecture

- Novel architecture designed for unstructured sound data
- Global pooling gathers statistics of learned features across time



Data Collection



Results - Baseline Comparison

(300ms window)

Features	SVM Linear	SVM RBF	k-NN
Ginna [1]	44.87 ± 0.70	37.51 ± 0.74	57.26 ± 0.60
Spectral [2]	84.48 ± 0.36	78.65 ± 0.45	76.02 ± 0.43
Ginna & Shape [3]	85.50 ± 0.34	80.37 ± 0.55	78.17 ± 0.37
MFCC & Chroma [4]	88.95 ± 0.21	$\mathbf{88.55 \pm 0.20}$	88.43 ± 0.15
Trimbral [5]	89.07 ± 0.12	86.74 ± 0.25	84.82 ± 0.54
Cepstral [6]	$\boxed{\mathbf{89.93 \pm 0.21}}$	78.93 ± 0.62	$\mathbf{88.63 \pm 0.06}$

[illegible]% improvement over the previous state of the art

[illegible]% using a 500ms window

[1] T. Giannakopoulos, K. Dimitrios, A. Andreas, and T. Sergios, SETN 2006

[2] M. C. Wellman, N. Srour, and D. B. Hillis, SPIE 1997.

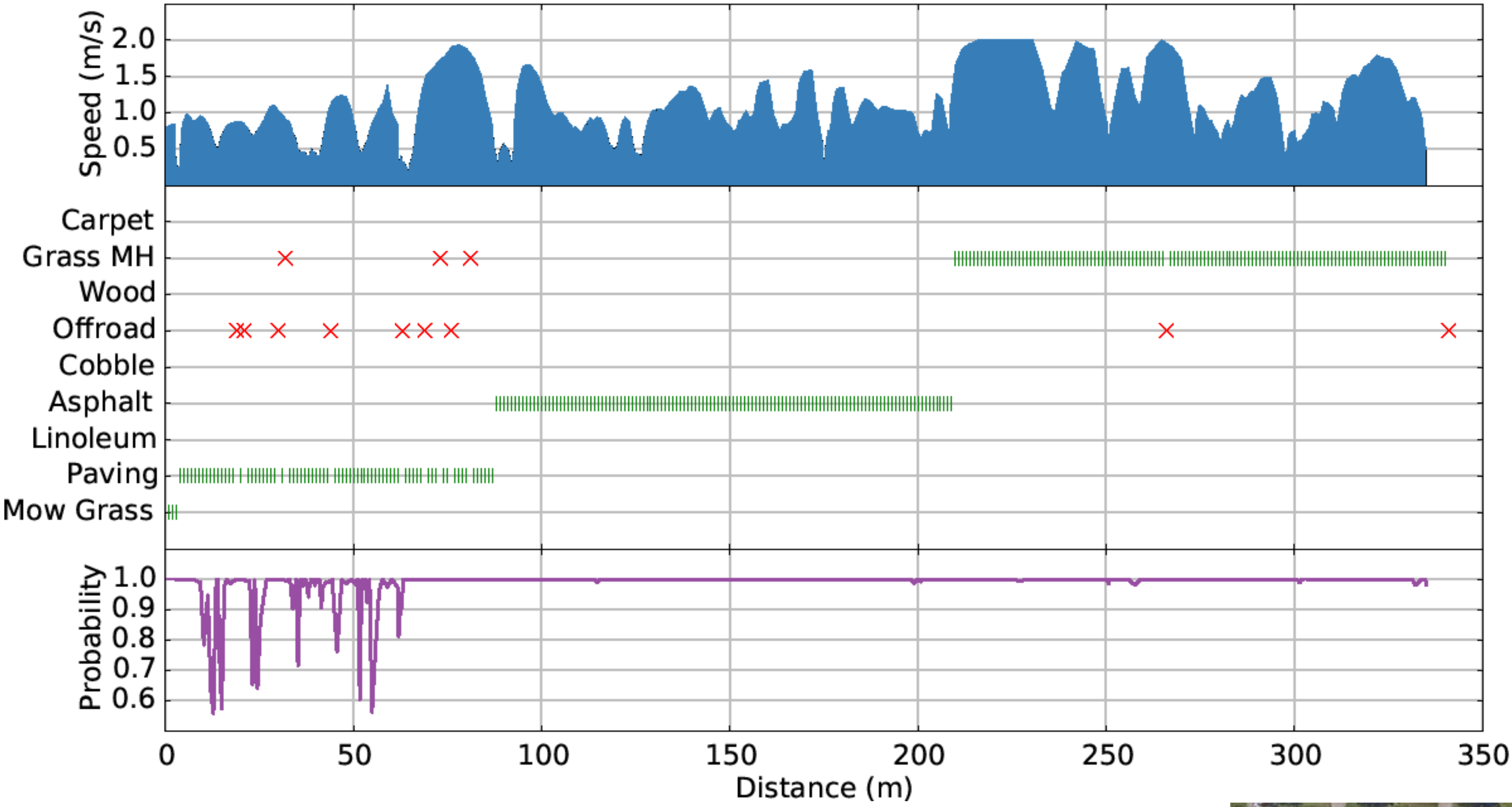
[3] J. Libby and A. Stentz, ICRA 2012

[4] D. Ellis, ISMIR 2007

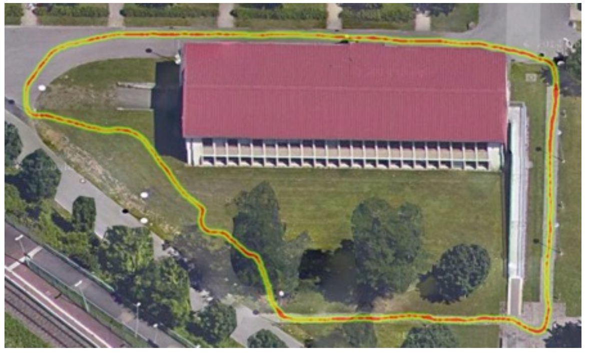
[5] G. Tzanetakis and P. Cook, IEEE TASLP 2002

[6] V. Brijesh , and M. Blumenstein, Pattern Recognition Technologies and Applications 2008

Real-World Stress Testing



| - True Positives
× - False Positives



Avg. accuracy of **98.54%**

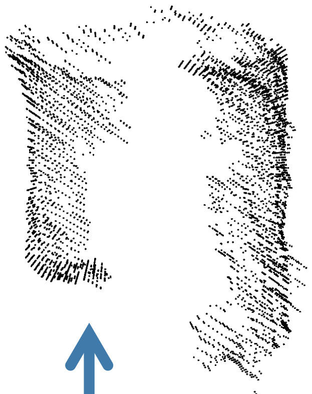
Can You Guess the Gerrain?

Social Experiment

- Avg. human performance = 24.66%
- Avg. network performance = 99.5%
- Go to **deepterrain.cs.uni-freiburg.de**
- Listen to five sound clips of a robot traversing on different terrains
- Guess what terrain they are

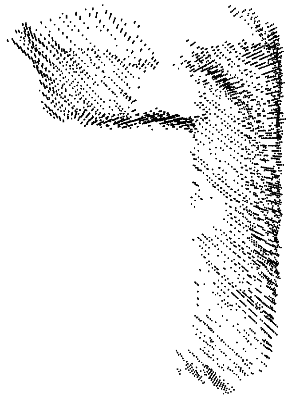
Liquid Height Detection in Cups using RGB-D Data is Hard

Transparent Liquid: Water



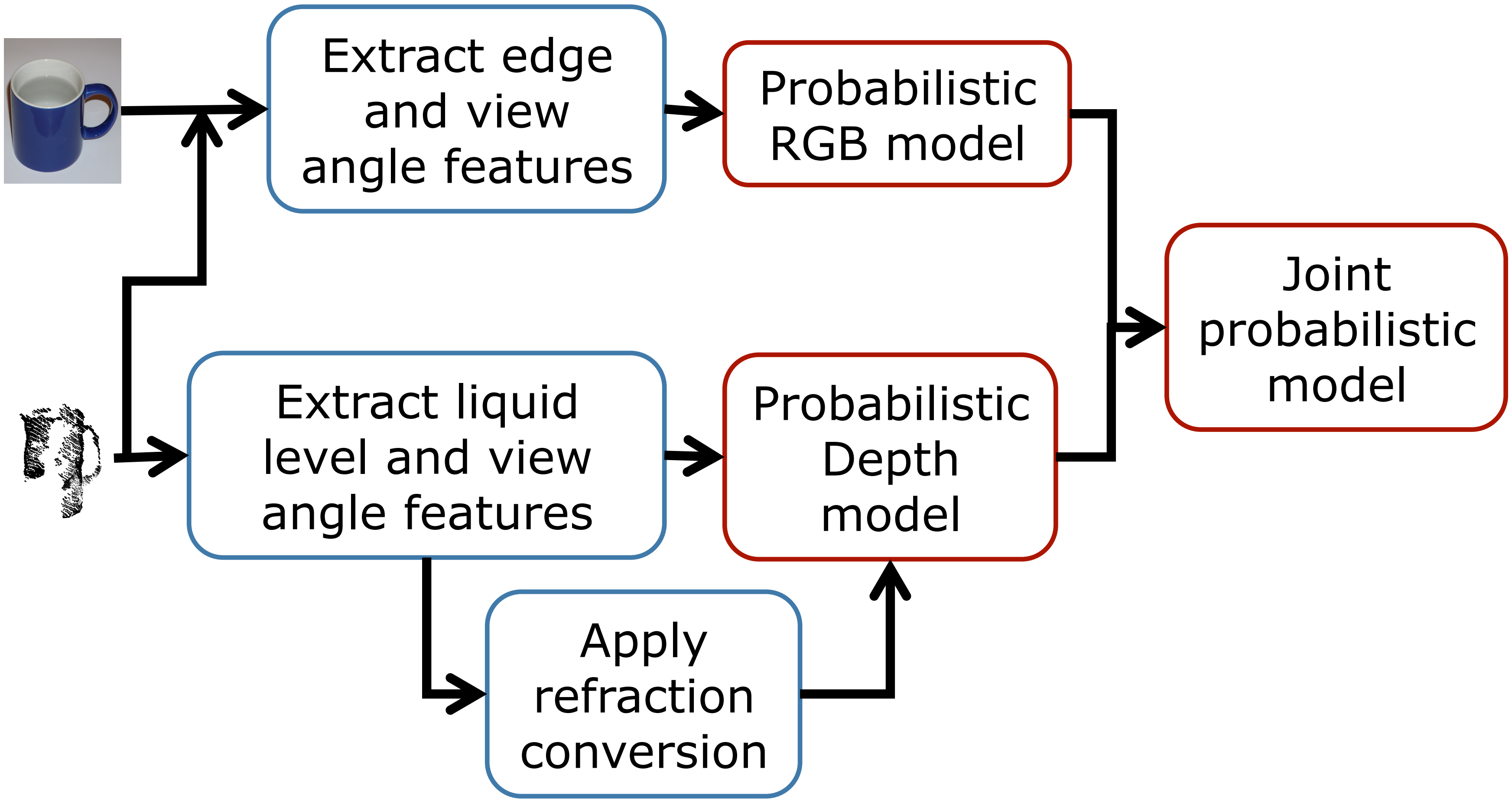
Refracted bottom

Opaque Liquid: Orange juice

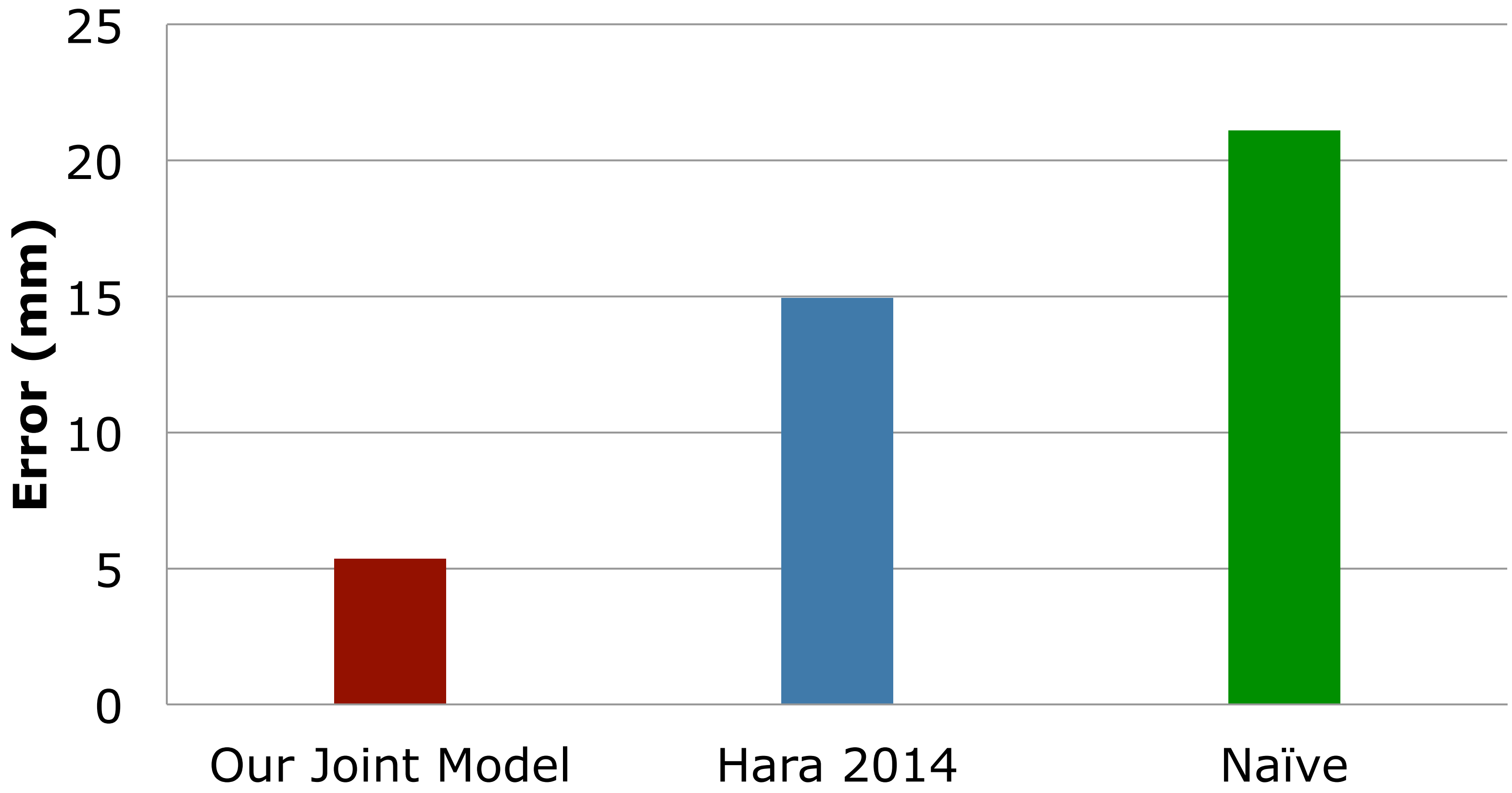


Reflected liquid height

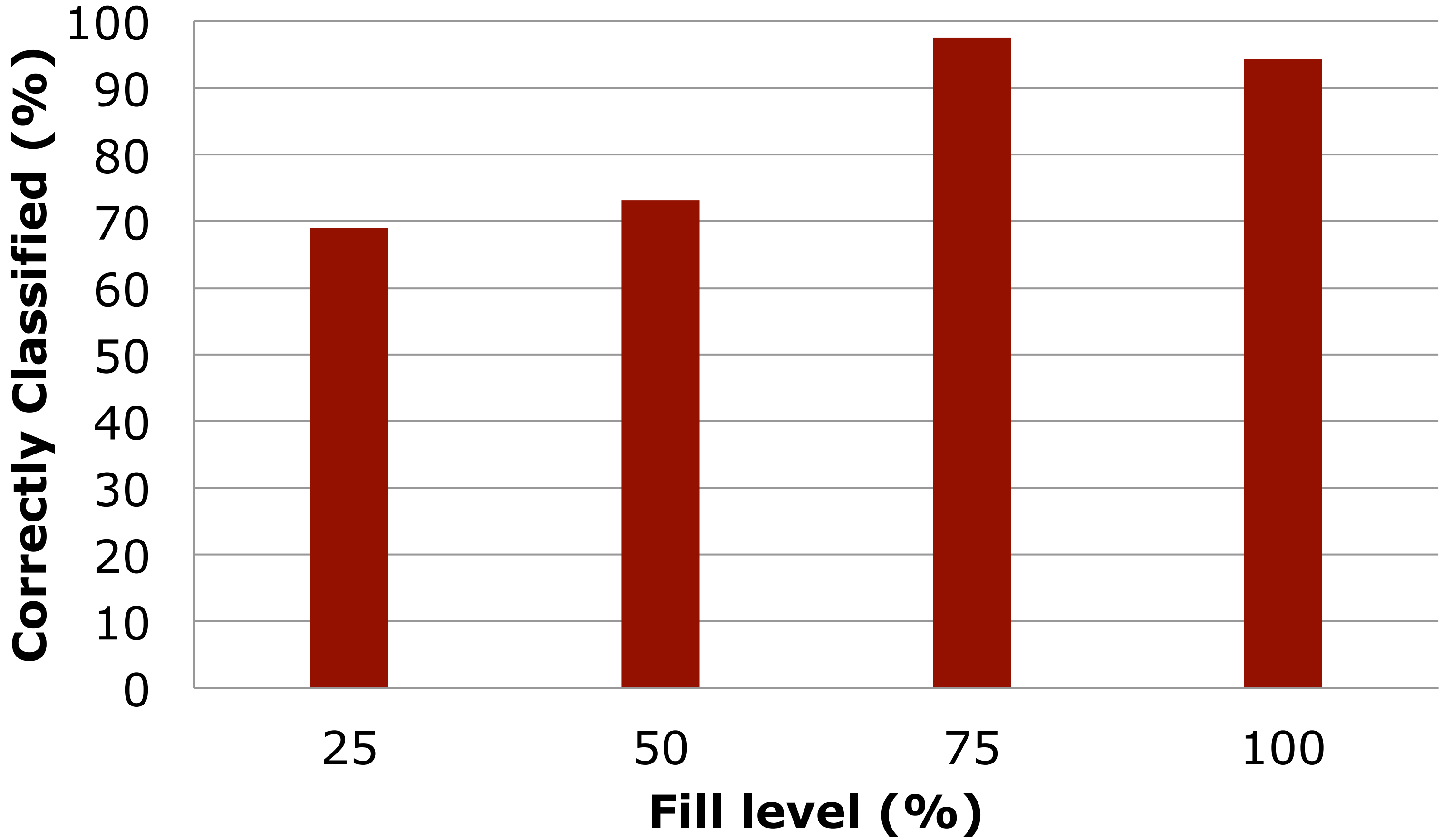
Approach Overview



Average Fluid Level Error



Opaque vs. Transparent Classification



Pouring

Liquid level detection assumes:

- Liquid is present in the cup
- Can take multiple views of the liquid at the same height

Pouring challenges:

- Starting with empty cup so that no initial liquid to obtain information from
- Only single view of cup so that it is awkward to take multiple views while pouring

Assumption:

- Liquid type is known

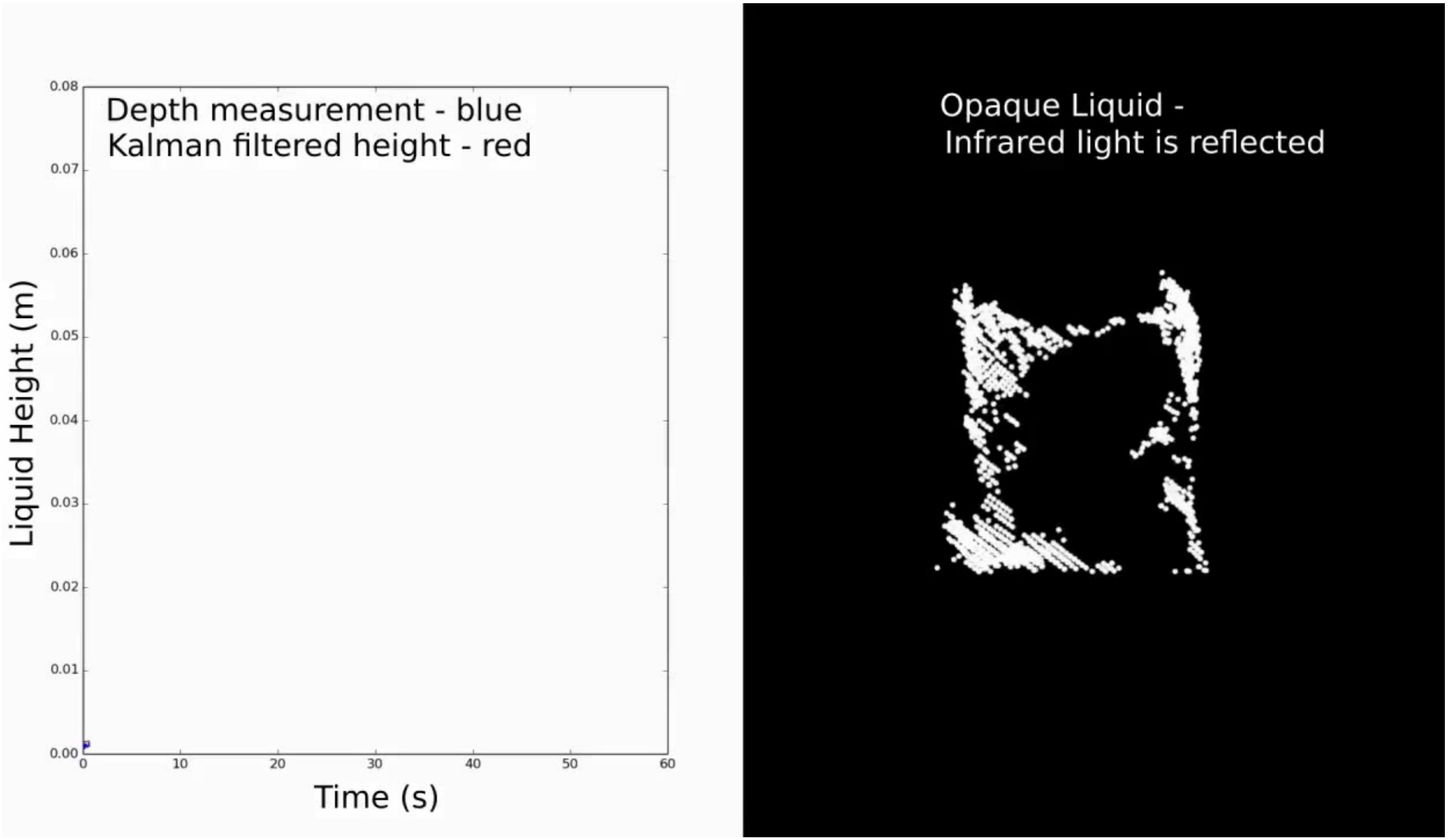
Liquid level Estimation during Pouring

- **Opaque Liquid**: Use raw measured depth value

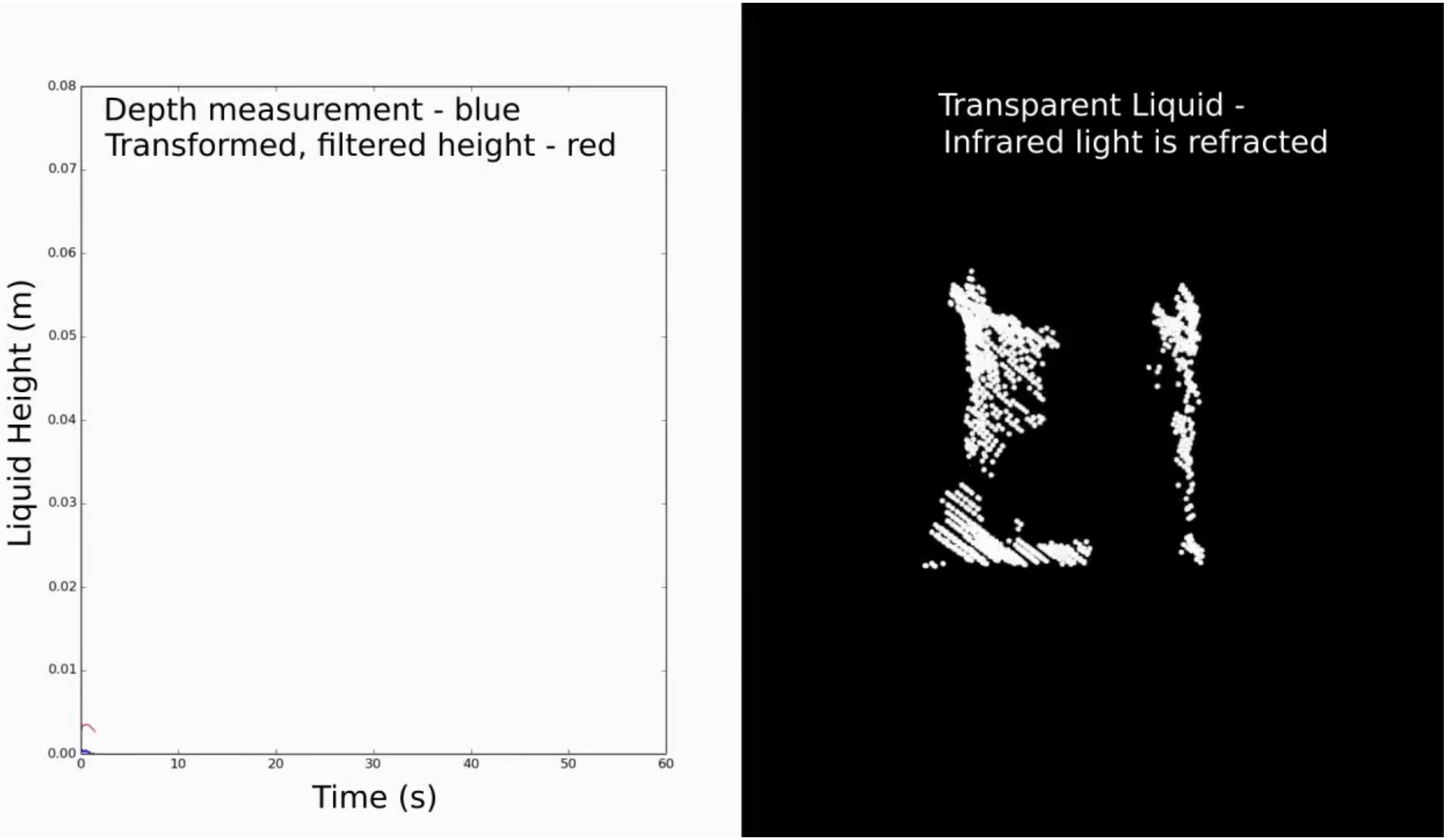
- **Transparent Liquid**: Liquid height can be determined from

$$h_r = \left(1 - \frac{\cos(\alpha)}{\sqrt{n_l^2 - 1 + \cos^2(\alpha)}}\right) h$$

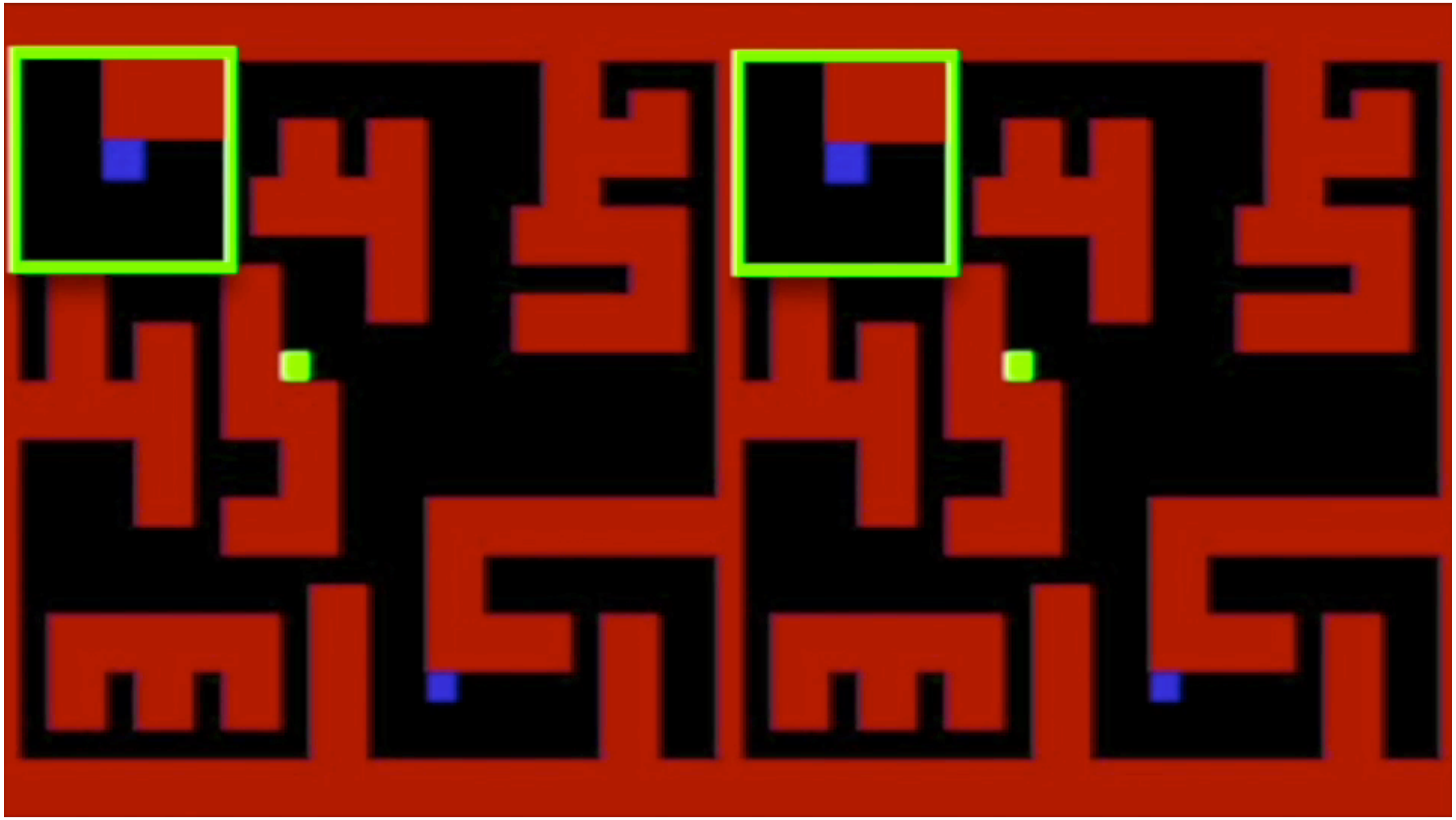
Tracking the Liquid Level using a Kalman Filter



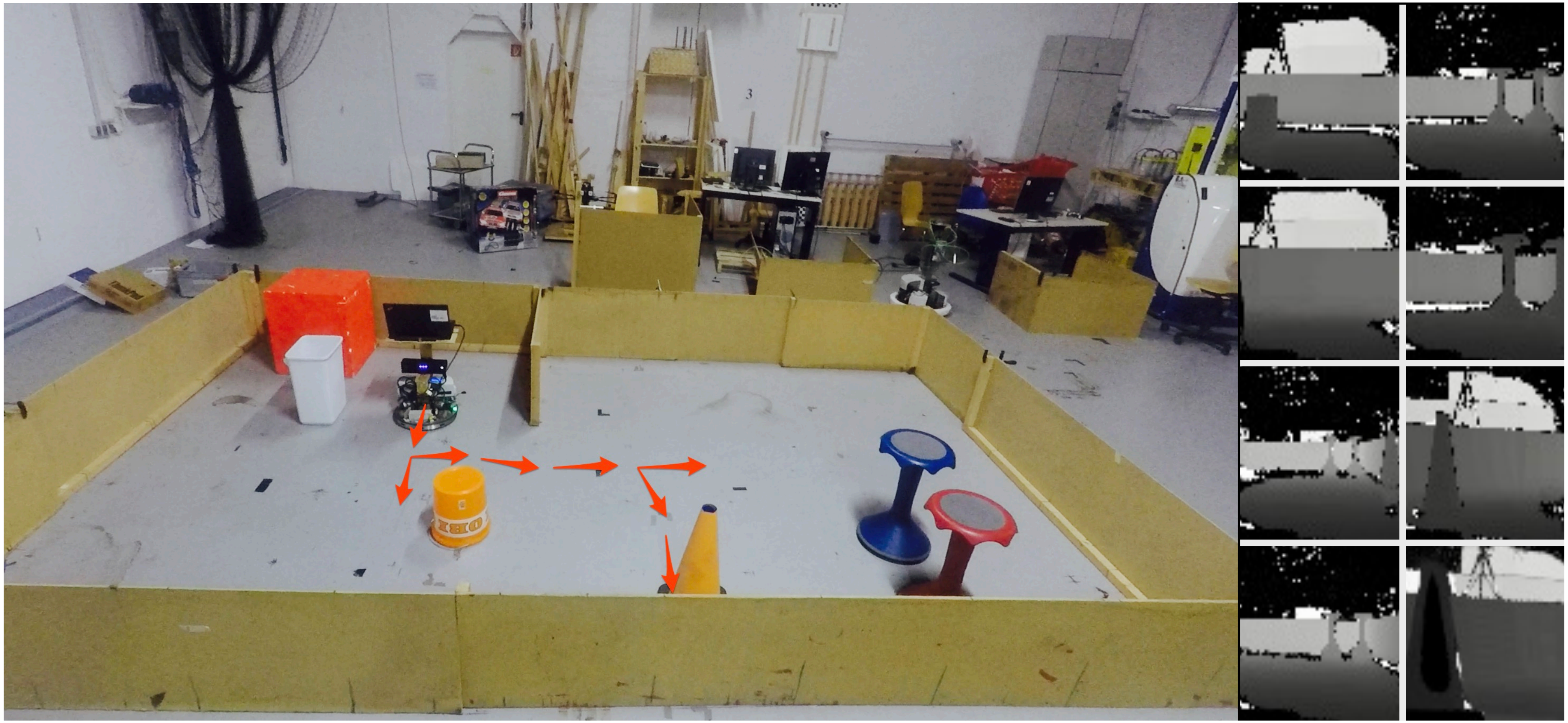
Tracking the Liquid Level using a Kalman Filter



… and End to End Navigation



Deep Reinforcement Learning with Successor Features for Navigation across Similar Environments



Motivation

- Finding a solution for navigation that:
  - Does not require explicit SLAM, localization and path planning procedures
  - Can adapt to new situations (new navigation goals and environments)
- Aim for the agent:
  - Capable of solving all tasks by the end of training
  - Using minimal interaction time for each task

Transfer learning between navigation tasks

Transfer learning scenarios:

1. Multiple goal positions: same environment and transition dynamics but different reward function.
2. Multiple environments: changes in the maze structure or robot dynamics

Successor Feature RL with Task Transfer (SF-RL-Transfer)

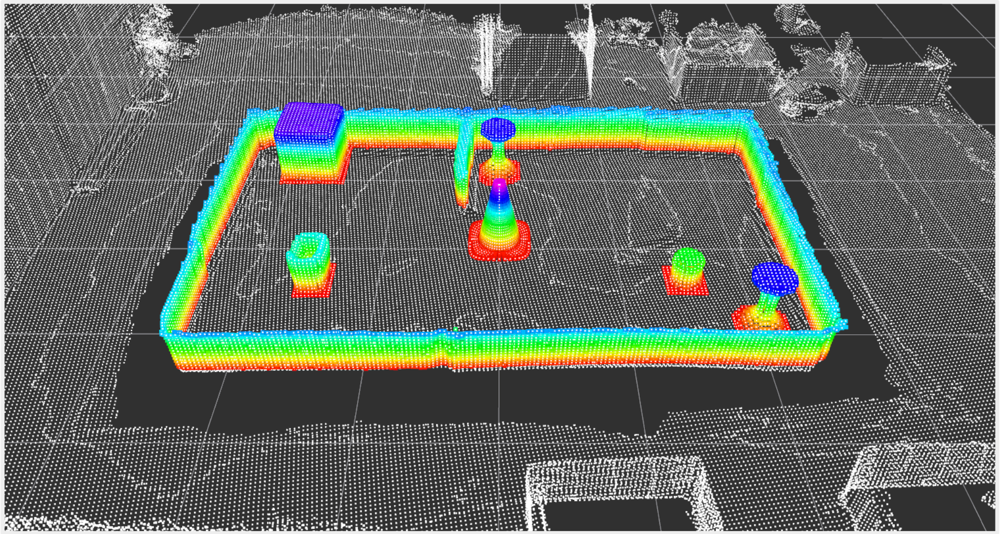
$$\phi^k_{\mathbf{s}_t} = \phi^k(\mathbf{s}_t, \theta_{\phi^k}) \tag{1}$$

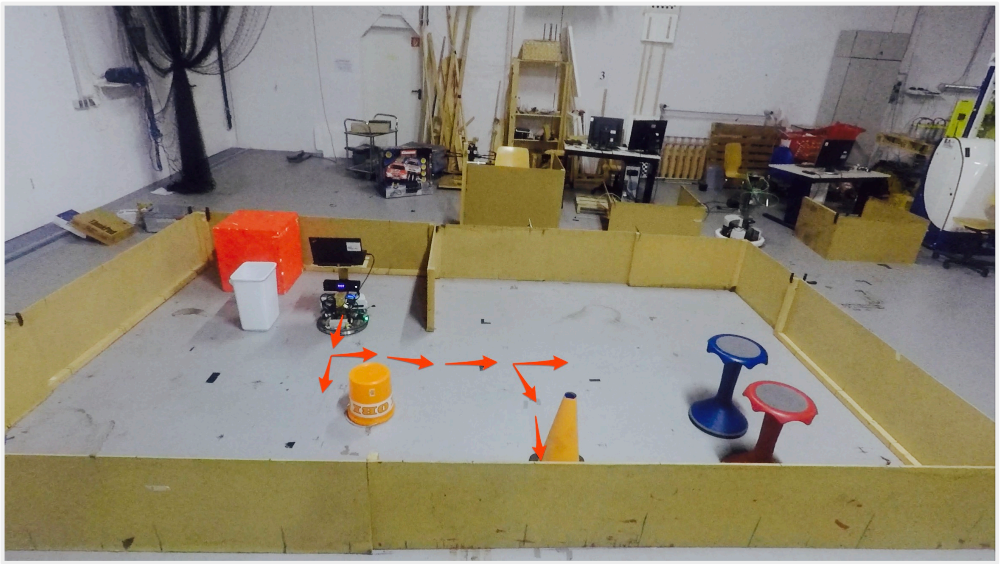
$$Q^\pi_k(\mathbf{s}, \mathbf{a}) \approx \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \phi^k_{\mathbf{s}_t} \middle| \mathbf{s}_0 = \mathbf{s}, \mathbf{a}_0 = \mathbf{a}, \pi^k\right] \cdot \omega^k \tag{2}$$

$$\phi^i_{\mathbf{s}} = \mathcal{B}^i \phi^k_{\mathbf{s}}, \mathcal{B}^k = I, \forall i \leq k \tag{3}$$

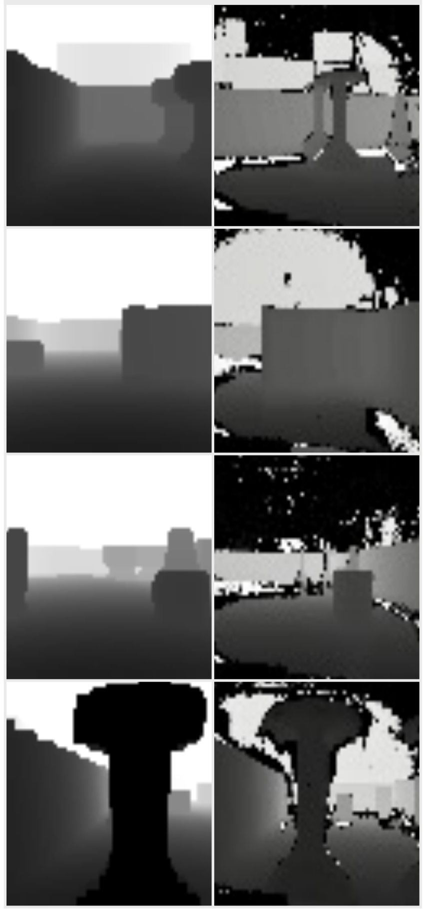
$$\begin{aligned}
Q^\pi_i(\mathbf{s}, \mathbf{a}) &\approx \mathbb{E}\left[\sum_{t=0}^{\infty} \mathcal{B}^i \gamma^t \phi^k_{\mathbf{s}_t} \middle| \mathbf{s}_0 = \mathbf{s}, \mathbf{a}_0 = \mathbf{a}, \pi^i\right] \cdot \omega^i \\
&= \mathcal{B}^i \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \phi^k_{\mathbf{s}_t} \middle| \mathbf{s}_0 = \mathbf{s}, \mathbf{a}_0 = \mathbf{a}, \pi^i\right] \omega^i \\
&= \mathcal{B}^i \psi^{\pi^i}\left(\phi^k_{\mathbf{s}_t}, \mathbf{a}\right)^T \omega^i && (4) \\
&= \psi^{\pi^i}\left(\mathcal{B}^i \phi^k_{\mathbf{s}_t}, \mathbf{a}\right)^T \omega^i. && (5)
\end{aligned}$$

Training Setup

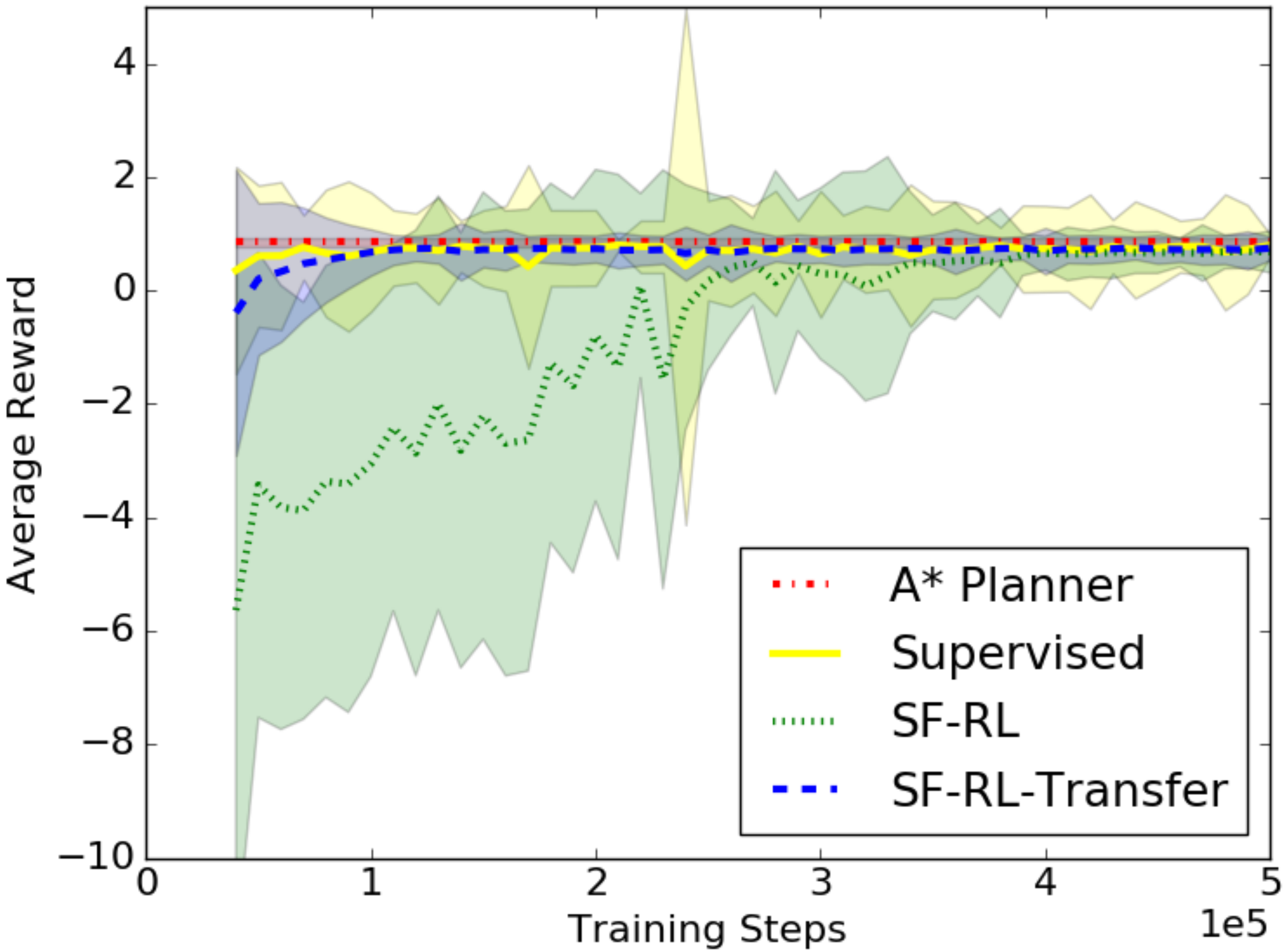




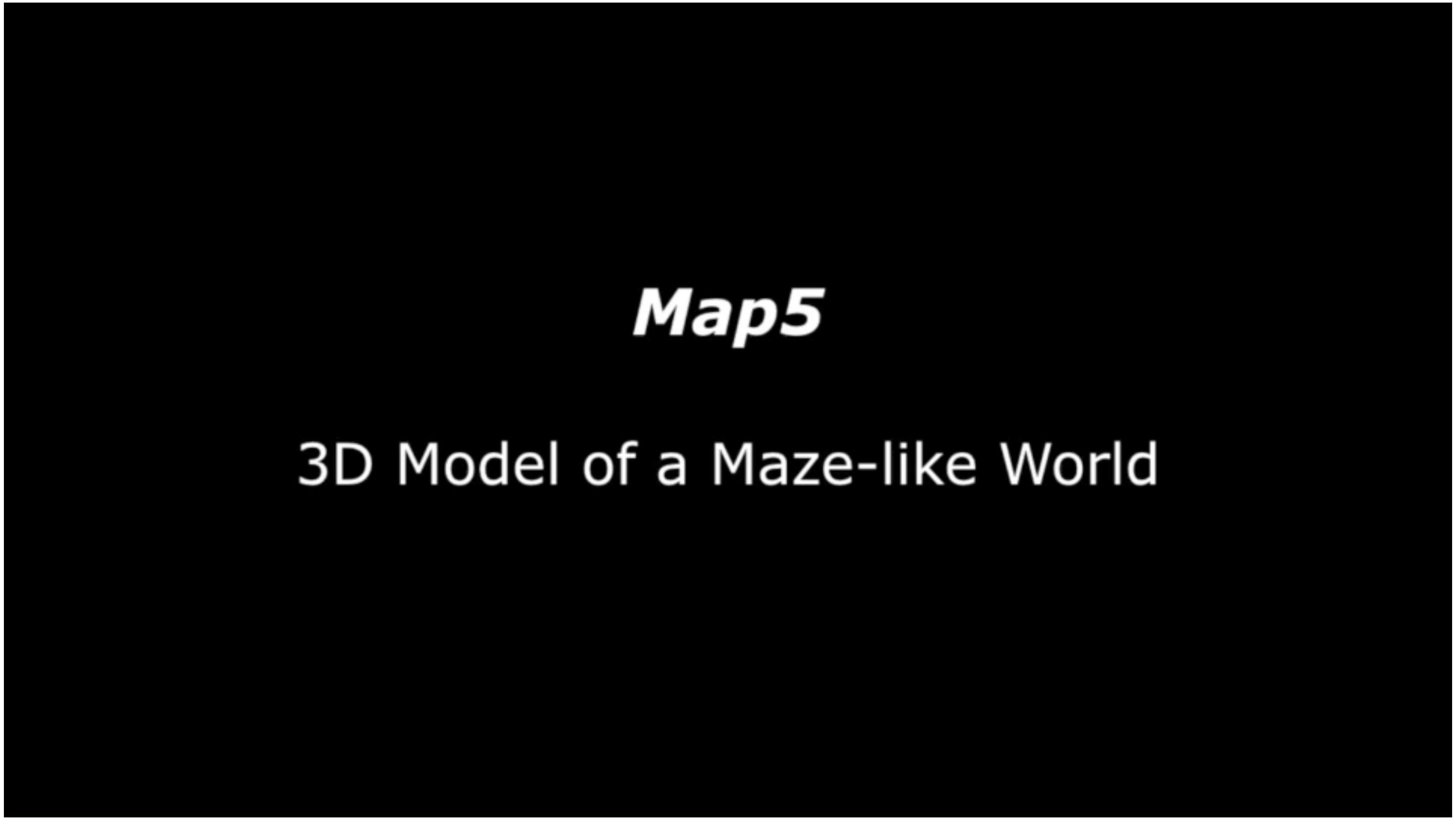
Synthetic Depth Images	Real Depth Images



SF-RL-Transfer



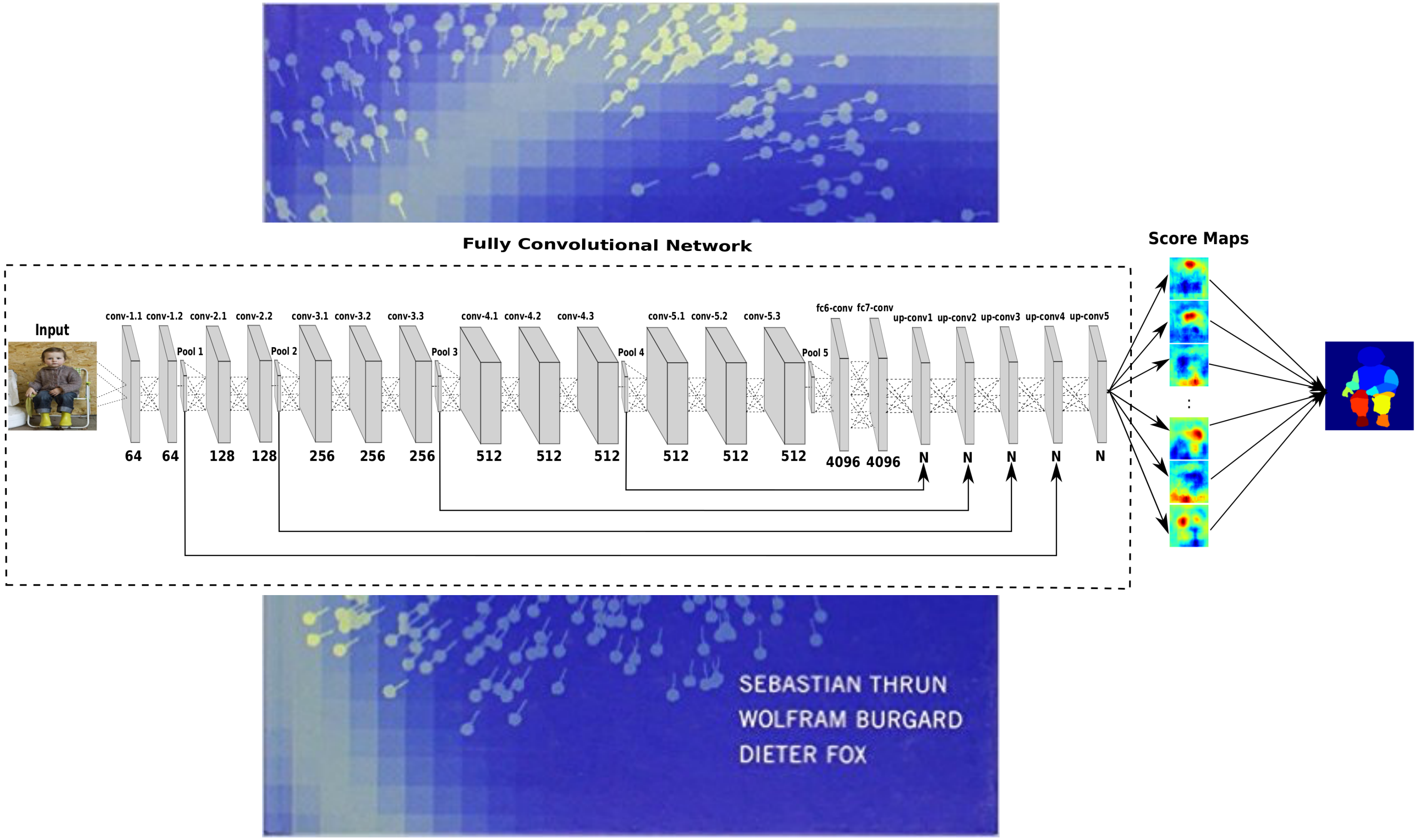
Real-world Experiments



Overall Conclusions

- Deep networks are a promising approach to solve complex perception problems in robotics
- The key challenges are
  - finding the proper architecture
  - using proper data augmentation strategies
- Goal: Achieving end-to-end learning of complex (navigation) tasks.

What is the Future of Probabilistic Robotics?



Thank you for your attention!