Alternatives in robotic perception for self-driving cars:

what can academics add to a robust industrial research area?

Matthew Johnson-Roberson
University of Michigan

The world around us is changing - industry has started pumping billions of dollars into ML, CV and more recently Robotics

What is the purpose of a research university when

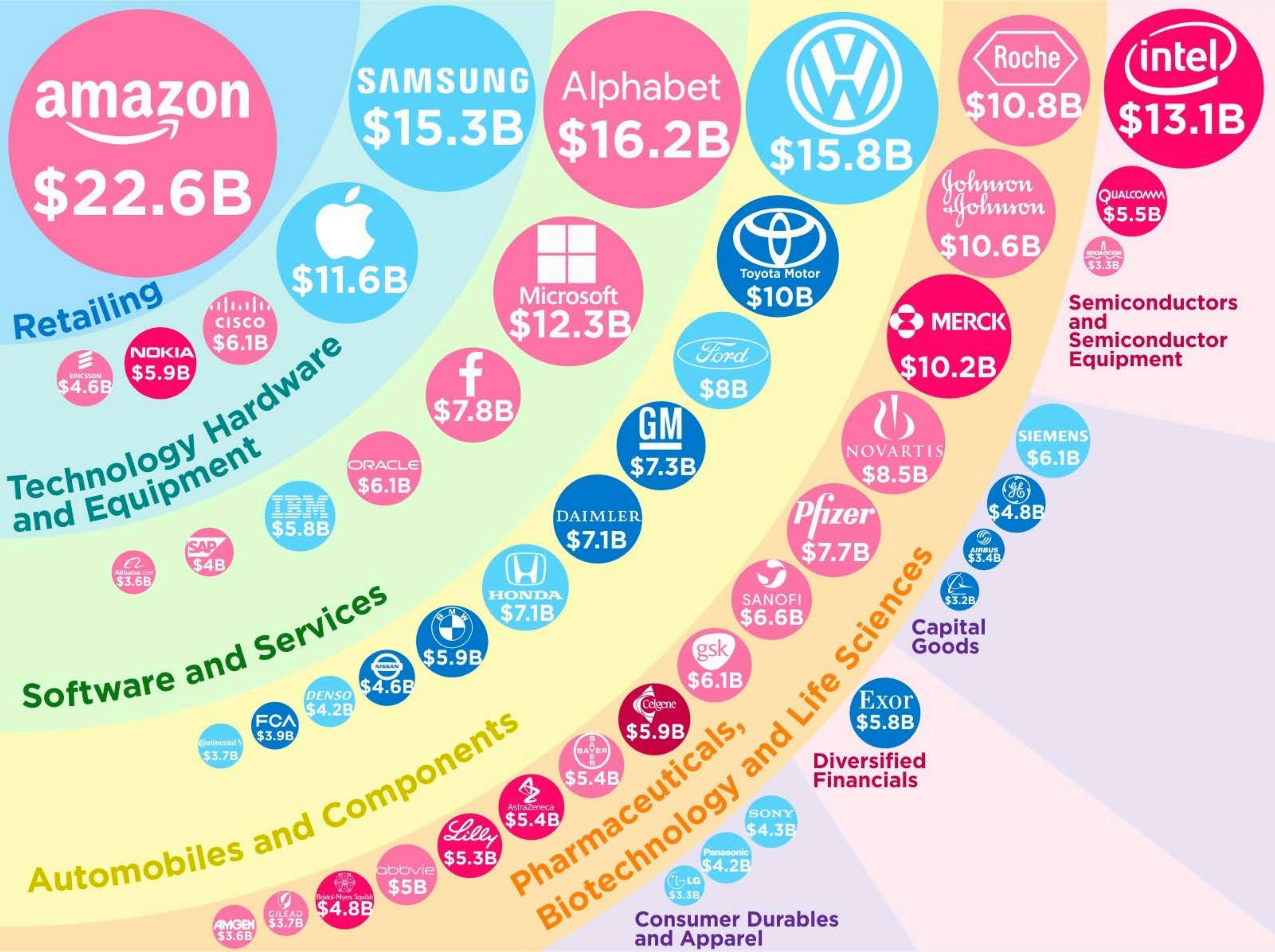
"industry research labs"

look increasingly like universities?

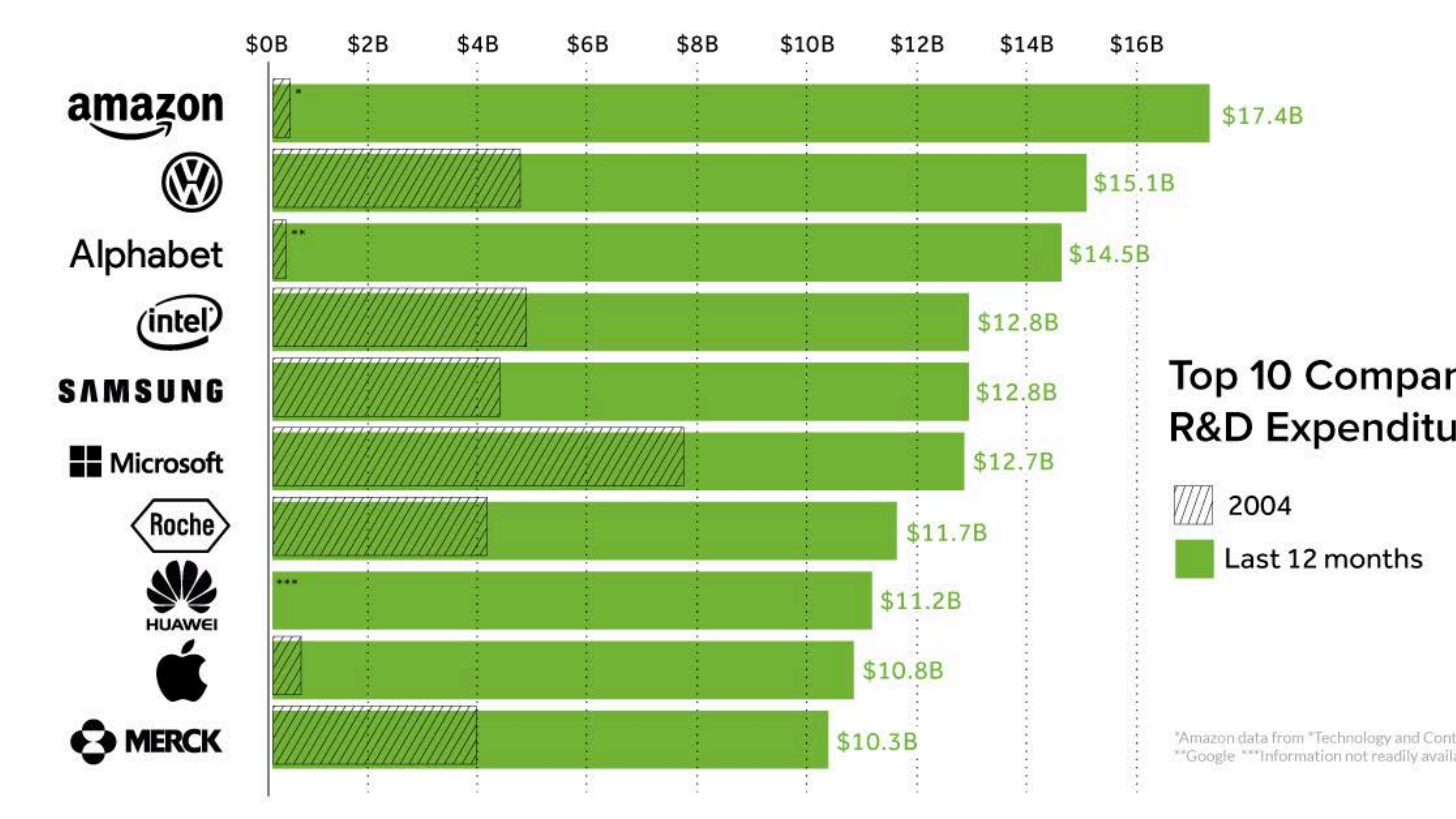
A company's ultimate allegiance is to its shareholders

a research university's ultimate allegiance is to knowledge

Economist.com

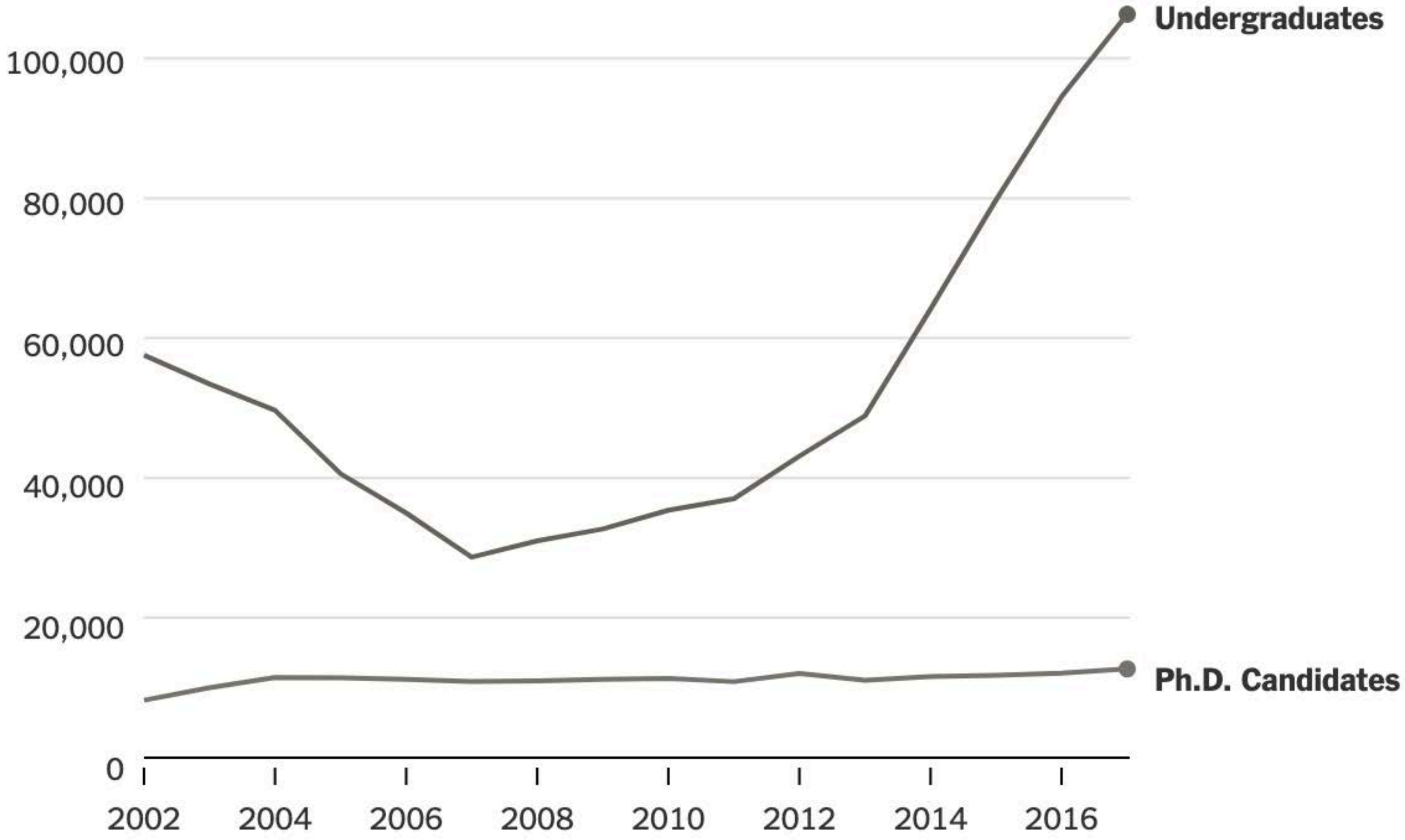


amazon
$22.6B
SAMSUNG
$15.3B
Alphabet
$16.2B
$15.8B
Roche
$10.8B
intel
$13.1B
$11.6B
Microsoft
$12.3B
Toyota Motor
$10B
Johnson & Johnson
$10.6B
QUALCOMM
$5.5B
BROADCOM
$3.3B
Retailing
CISCO
$6.1B
NOKIA
$5.9B
ERICSSON
$4.6B
Ford
$8B
MERCK
$10.2B
Semiconductors and Semiconductor Equipment
$7.8B
Technology Hardware and Equipment
GM
$7.3B
NOVARTIS
$8.5B
SIEMENS
$6.1B
ORACLE
$6.1B
IBM
$5.8B
DAIMLER
$7.1B
Pfizer
$7.7B
$4.8B
SAP
$4B
AIRBUS
$3.4B
Alibaba.com
$3.6B
HONDA
$7.1B
SANOFI
$6.6B
$3.2B
Software and Services
$5.9B
Capital Goods
gsk
$6.1B
NISSAN
$4.6B
DENSO
$4.2B
Exor
$5.8B
FCA
$3.9B
Celgene
$5.9B
Continental
$3.7B
Automobiles and Components
BAYER
$5.4B
Diversified Financials
Pharmaceuticals, Biotechnology and Life Sciences
AstraZeneca
$5.4B
SONY
$4.3B
Lilly
$5.3B
Panasonic
$4.2B
abbvie
$5B
LG
$3.3B
Bristol-Myers Squibb
$4.8B
GILEAD
$3.7B
AMGEN
$3.6B
Consumer Durables and Apparel


Top 10 Compar
R&D Expenditu
2004
Last 12 months
$0B
$2B
$4B
$6B
$8B
$10B
$12B
$14B
$16B
amazon
$17.4B
$15.1B
Alphabet
$14.5B
intel
$12.8B
SAMSUNG
$12.8B
Microsoft
$12.7B
Roche
$11.7B
HUAWEI
$11.2B
$10.8B
MERCK
$10.3B
*Amazon data from "Technology and Cont
**Google ***Information not readily availa

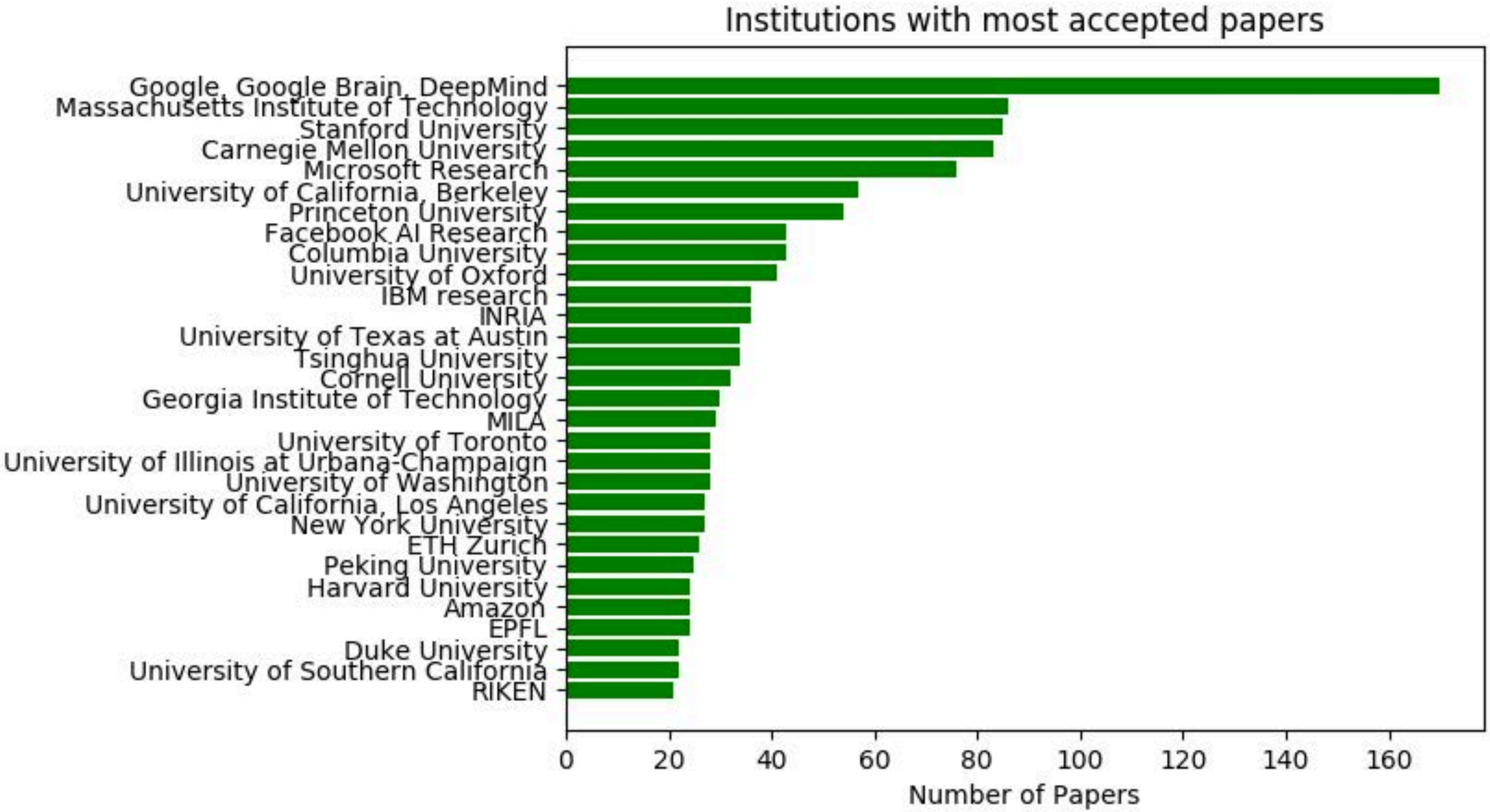
“I had a faculty member who came in with an offer from a bank, and they were told that, with their expertise, the starting salary would be $1 million to $4 million,”

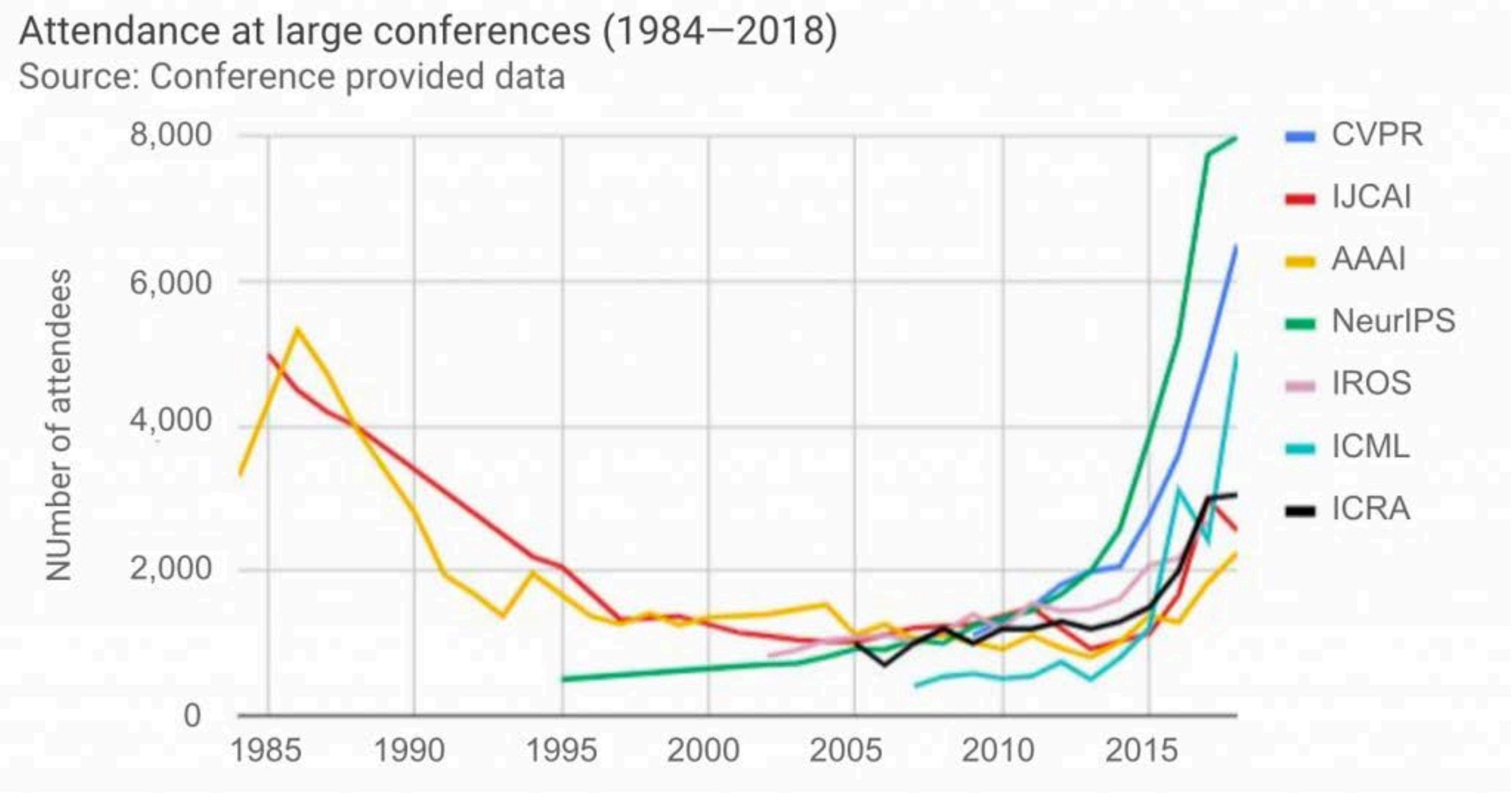
Greg Morrisett, dean of computing and information science at Cornell University - Nytimes Jan. 24, 2019



Greg Morrisett, dean of computing and information science at Cornell University - Nytimes Jan. 24, 2019

NeurIPS 2019




Attendance at large conferences (1984—2018)
Source: Conference provided data
NUmber of attendees
8,000
6,000
4,000
2,000
0
1985
1990
1995
2000
2005
2010
2015
CVPR
IJCAI
AAAI
NeurIPS
IROS
ICML
ICRA

CVPR 2020

Number of submissions and admission rate

Year	Number of submitted papers	Number of accepted papers	Acceptance rate
2015	2123	602	28.35%
2016	2145	643	29.97%
2017	2620	783	29.88%
2018	3303	979	29.63%
2019	5160	1294	25.07%
2020	6656	1476	22.17%

Institutions ranked by number of contributions (top 20)

Name of institution	Number of accepted papers
Google	88
Chinese Academy of Sciences	84
Tsinghua University	60
Microsoft	59
SenseTime	56
Peking University	54
Chinese University of Hong Kong	45
Peng Cheng Laboratory	45
Huawei	44
Facebook	42
Carnegie Mellon University	41
ETH Zürich	39
University of Science and Technology of China	38
Adobe	37
Nanyang Technological University	35
Massachusetts Institute of Technology	32
Nanjing University	32

What can we do that a company cannot?

Interdisciplinary Work

Work for Social Good

Ethics and Accountability

Fundamental Research

Education

What from a technical perspective?

- Weird sensors
- Less brute-force approaches
- Simulation
- Orthogonal

Pixel-Wise Motion Deblurring of Thermal Videos

Manikandasriram S.R

Pixel-Wise Motion Deblurring of Thermal Videos (S. Manikandasriram, R. Vasudevan, M. Johnson-Roberson), *In Robotics: Science and Systems*, 2020

Small exposure time eliminates motion blur

Visible Image captured at **30fps** while panning

Thermal Image captured at **200fps** while panning

Microbolometers work differently

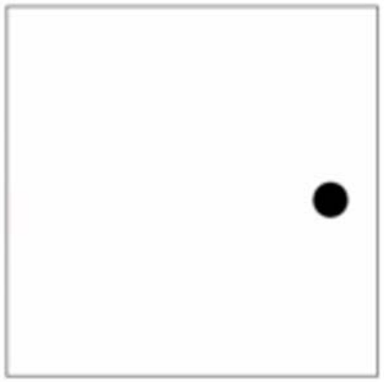
Visible Cameras

- Controllable exposure time
- Frame is a snapshot

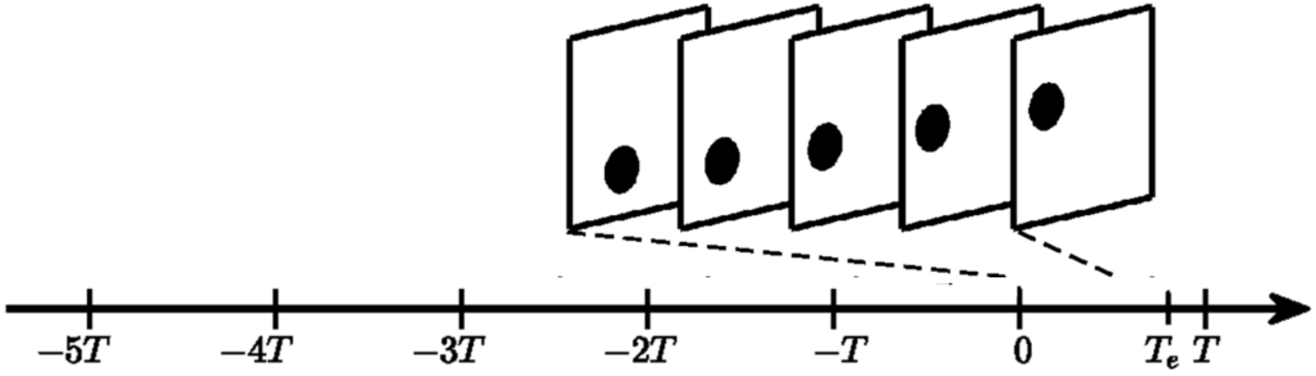
Microbolometers

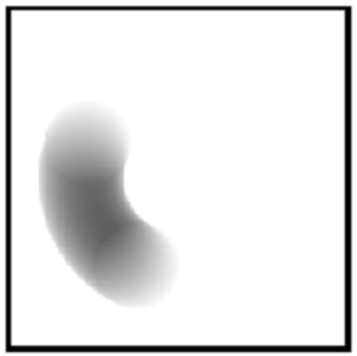
- Always exposed
- Does not reset to zero

Physics behind Motion Blur

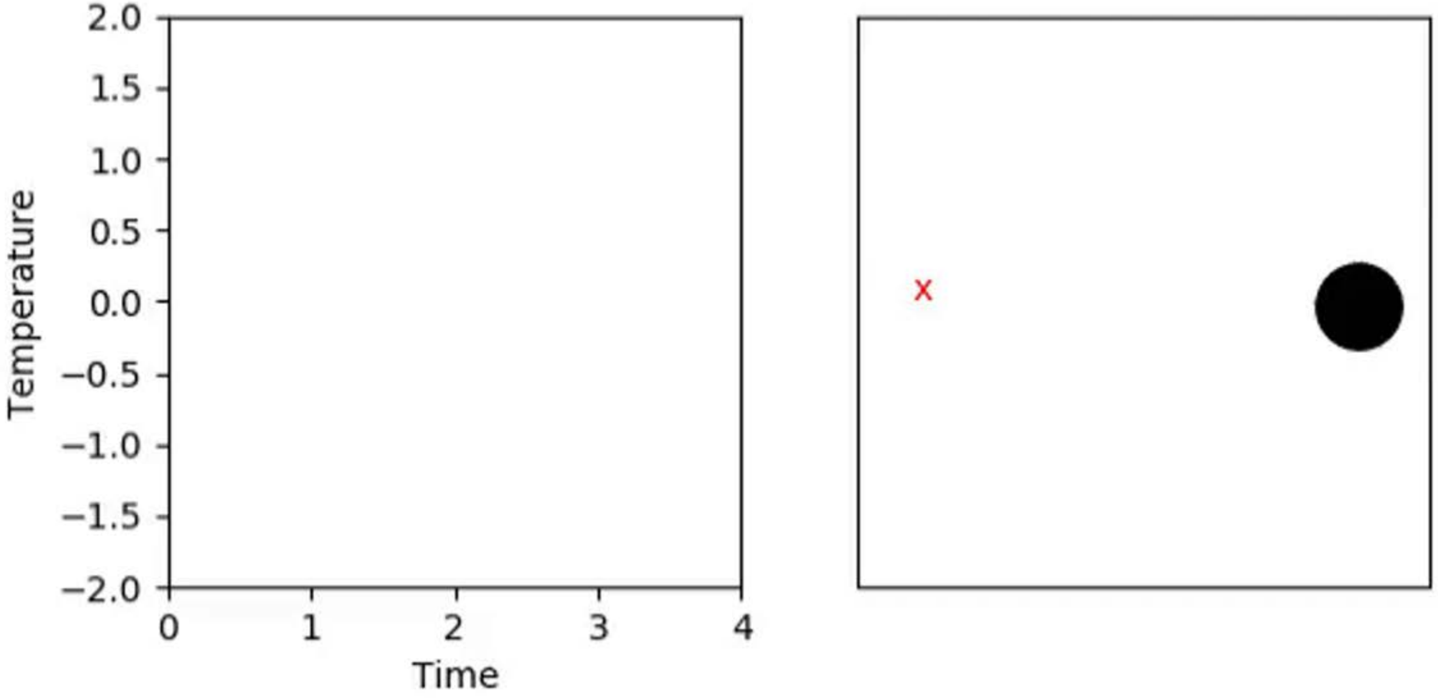


Motion blur in Visible camera

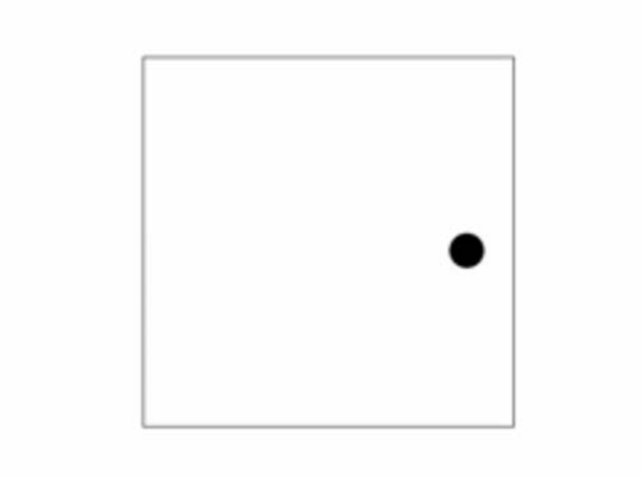




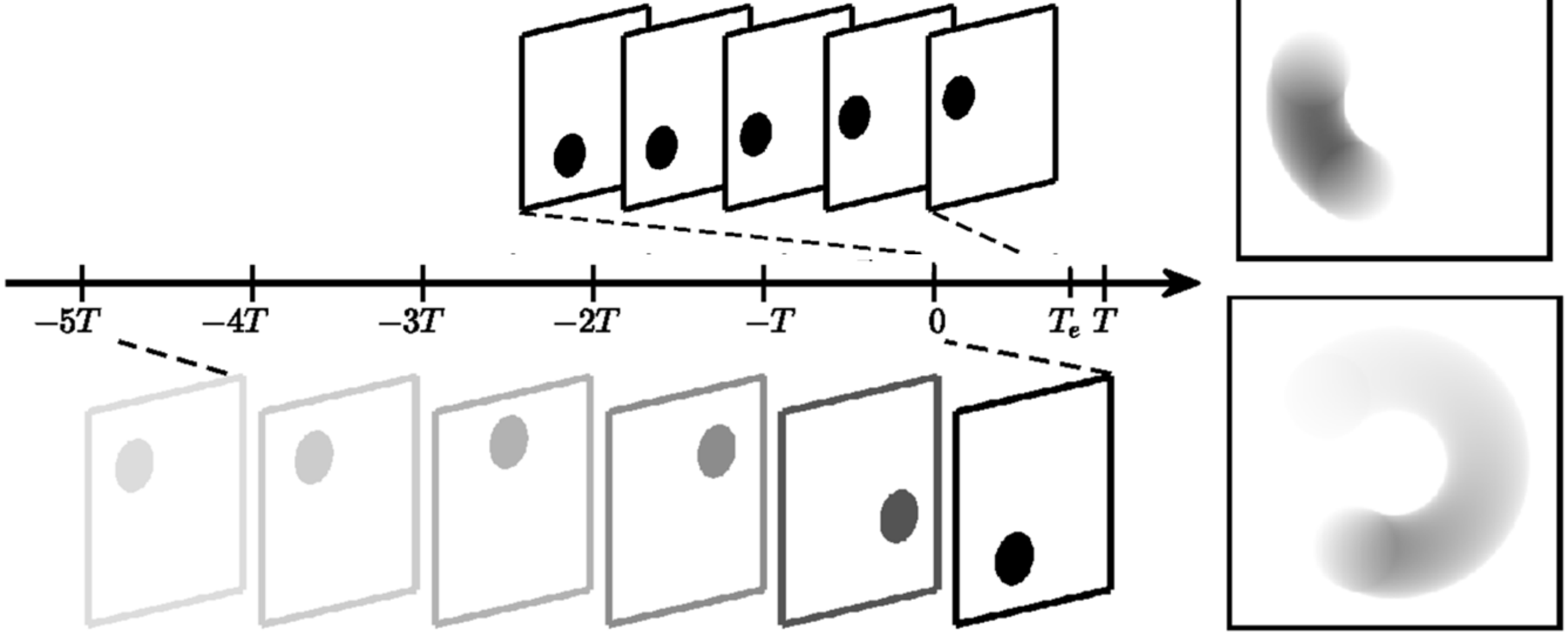
Microbolometer pixel is like a resistor-capacitor circuit



Physics behind Motion Blur



Motion blur in Visible camera



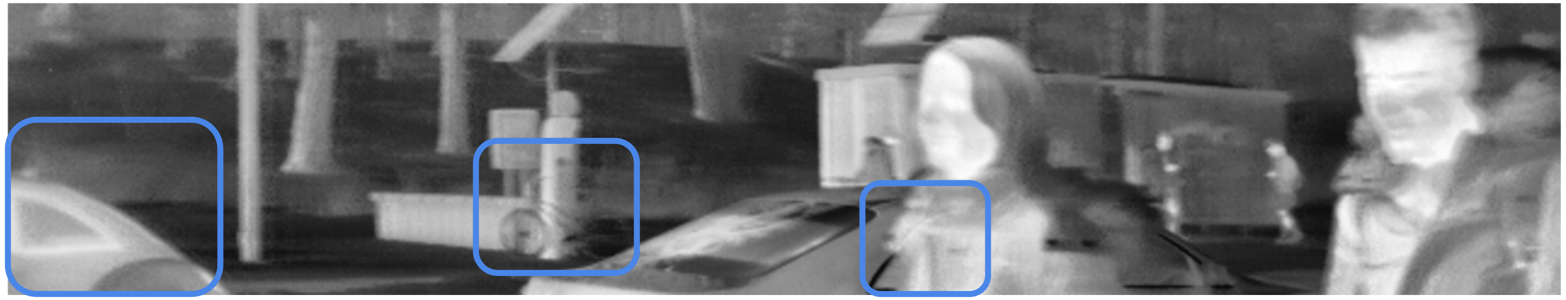
Motion blur in Microbolometer

Qualitative results

Blurred input

DeblurGANv2

Ours



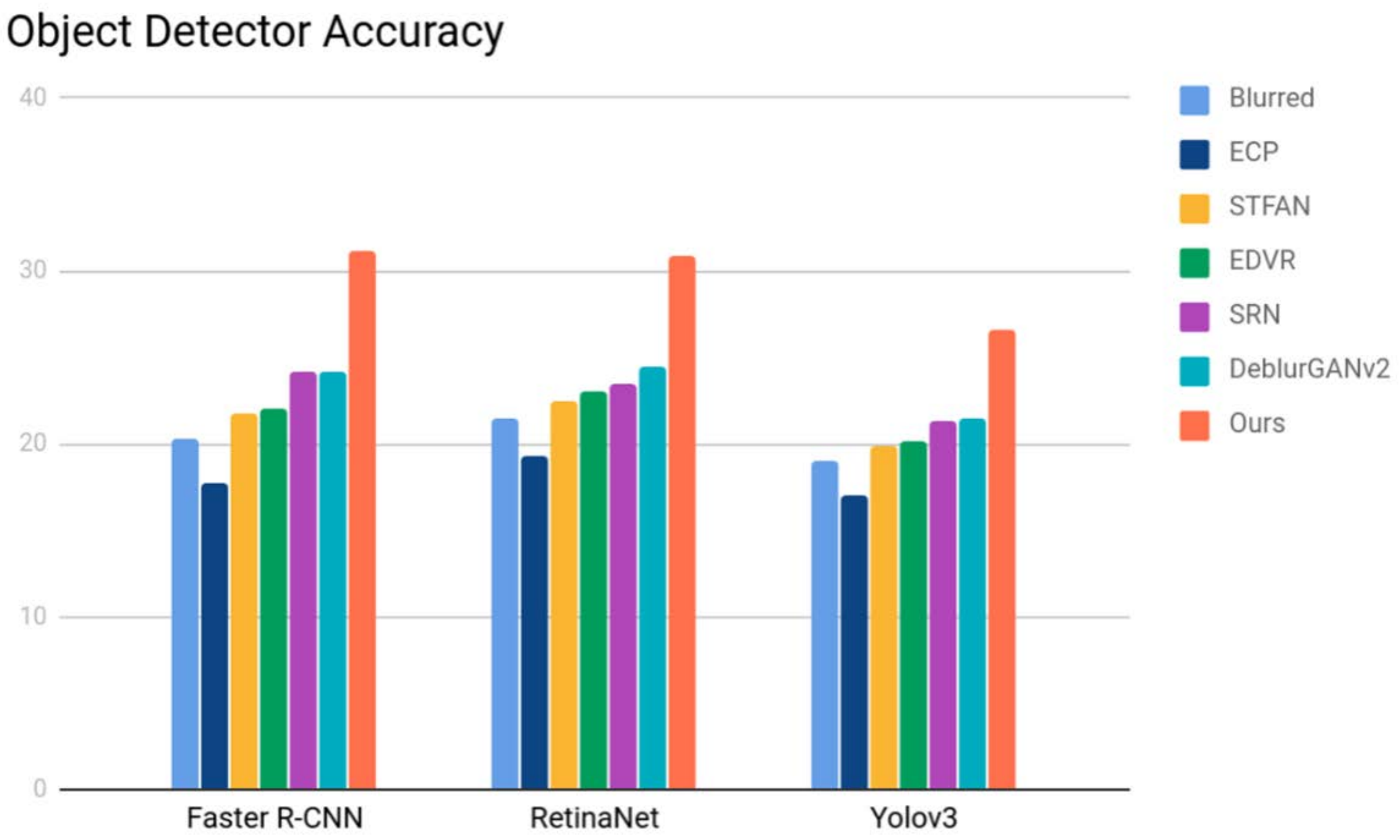
Qualitative results

Blurred input

DeblurGANv2

Ours

Quantitative evaluation



Key Contributions

- Our model-based algorithm
  - Respects microbolometer physics
  - Handles arbitrary camera motions
  - Handles arbitrary scene dynamics
  - Achieves state-of-the-art performance

Motion Deblurring

Literature (Image-Wise)

$$I(i,j) = \iint H(i-x, j-y) L(x,y)\, dx\, dy$$

x, y	Pixel coordinates
H	Point Spread Function
L	Latent image
I	Observed image

Motion Deblurring

Literature (Image-Wise)

$$I(i,j) = \iint H(i-x, j-y) L(x,y)\, dx\, dy$$

- H models relative motion
- Both H and L are unknown

Ours (Pixel-Wise)

$$I(i,j) = \frac{1}{\tau} \int_{-\infty}^{t} e^{\frac{s-t}{\tau}} L(s) ds$$

t	Time
τ	Thermal time constant
L	Latent image
I	Observed image

Motion Deblurring

Literature (Image-Wise)

$$I(i,j) = \iint H(i-x, j-y) L(x,y)\, dx\, dy$$

- H models relative motion
- Both H and L are unknown

Ours (Pixel-Wise)

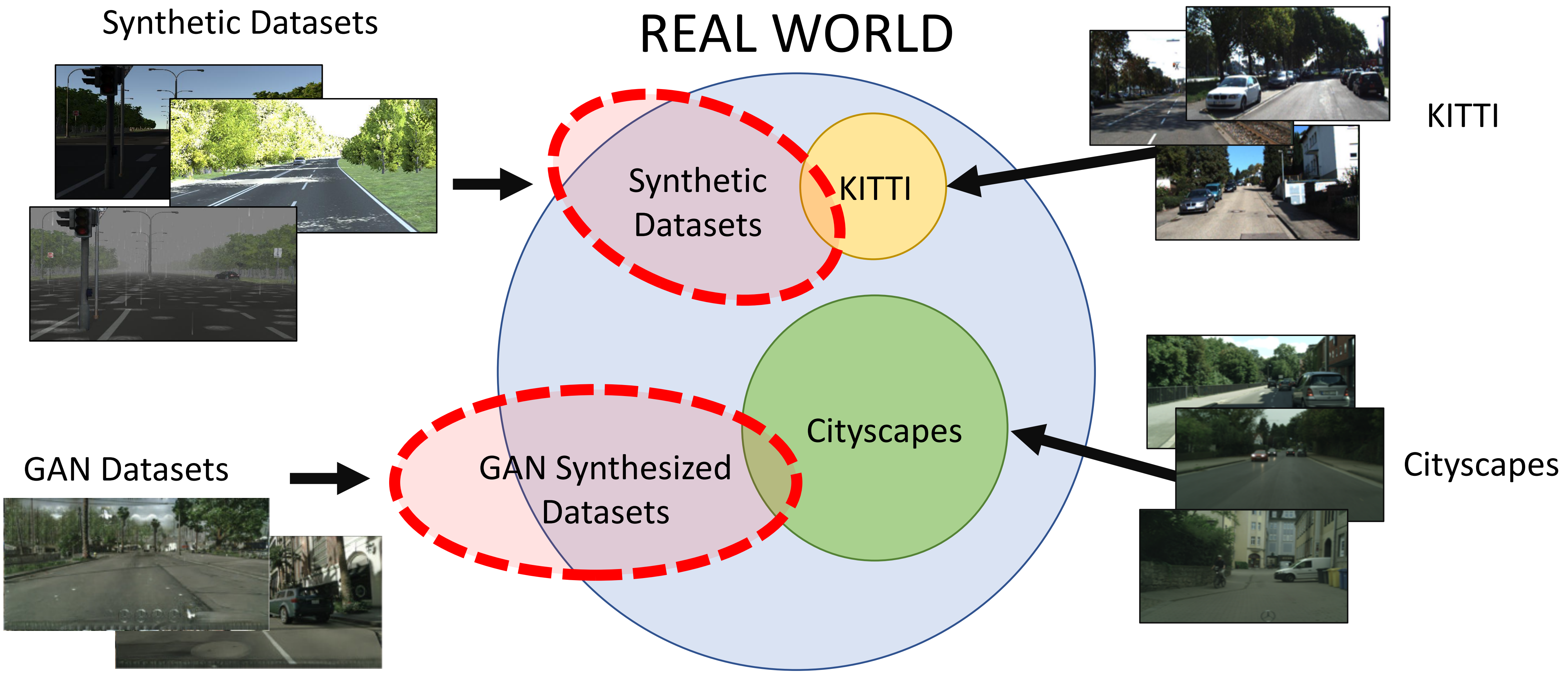
$$I(i,j) = \frac{1}{\tau} \int_{-\infty}^{t} e^{\frac{s-t}{\tau}} L(s) ds$$

- τ is fixed and can be calibrated
- Only L is unknown

Physically-based Augmentation Techniques to overcome Domain Adaptation

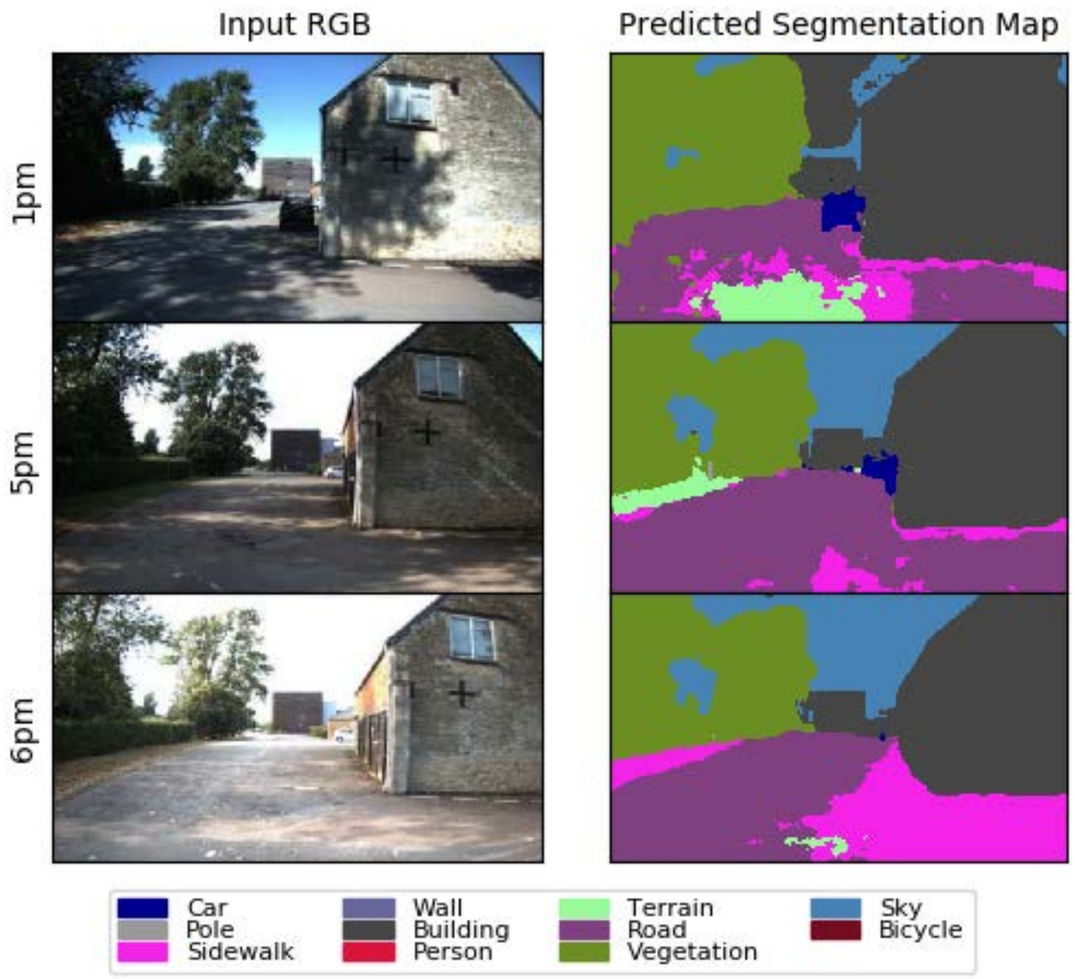
Alexa Carlson

Introduction: The domain shift in Rendered and Real Datasets



Prior Work: Illumination effects degrade performance (and contribute to Domain Shift!)

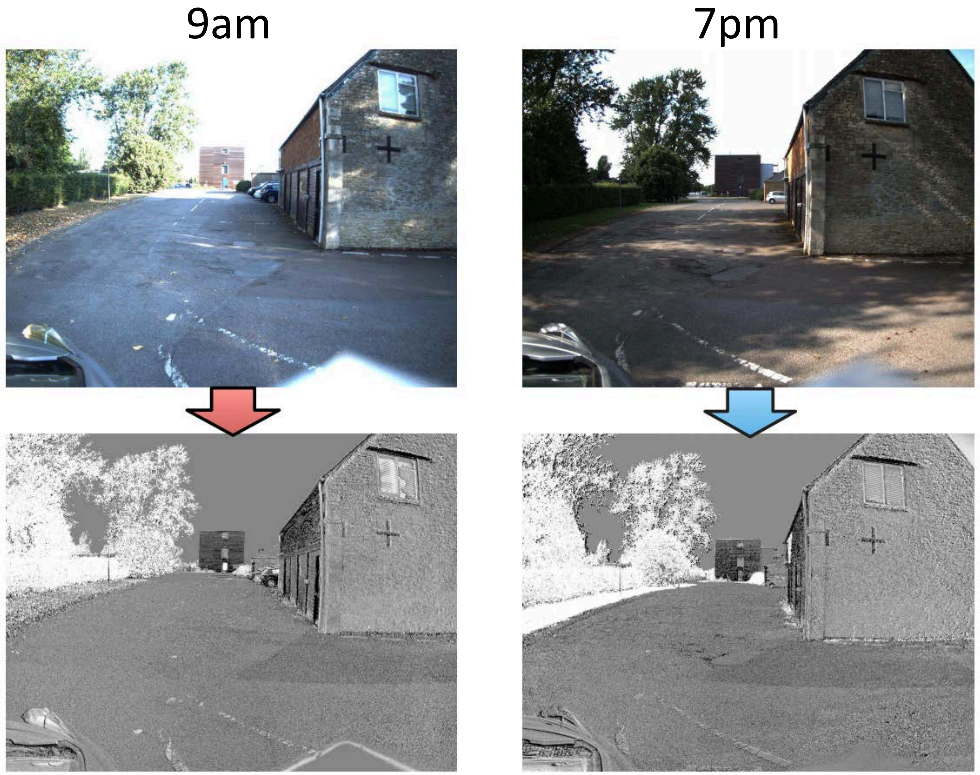
- By considering changes in illumination, we consider a huge variety of visual effects:
  - Specular highlights, reflections
  - Overexposure/saturation, underexposure
  - Soft and hard shadows, shading
  - Color changes
- Environmental lighting cause severe prediction errors for deep learning algorithms trained for object tracking, detection and segmentation tasks



Maddern et al, *Illumination Invariant Imaging: Applications in Robust Vision-based Localisation, Mapping and Classification for Autonomous Vehicles*, ICRA 2014

Prior Work: **Physically-based data augmentation**

- Illumination Invariant Color spaces[1,2,3]



1 Alshammari et al, *On the Impact of Illumination-Invariant Image Pre-transformation for Contemporary Automotive Semantic Scene Understanding*, IV 2018
2 Alshammari et al, *Multi-Task Learning for Automotive Foggy Scene Understanding via Domain Adaptation to an Illumination-Invariant Representation,* arxiv 2019
3 Maddern et al, *Illumination Invariant Imaging: Applications in Robust Vision-based Localisation, Mapping and Classification for Autonomous Vehicles*, ICRA 2014

Brief overview of past approach

Proposed Approach: *Shadow Transfer Network*

- *We cast as a multi-domain transfer problem, where the goal is to transfer illumination effects between times of day*

- Learns an illumination model via a deep neural encoder-decoder framework that operates upon input that is easily obtained from a car-mounted RGB camera

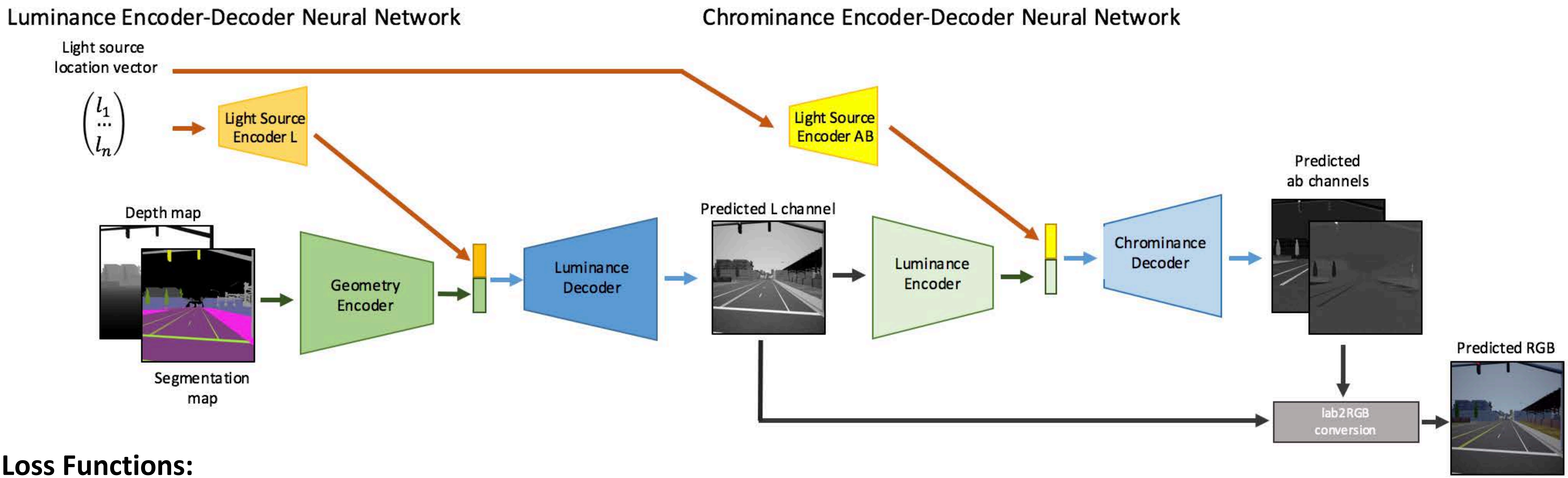
- Designed to be self-supervised, removing the need for labeling illumination features in images, like shadows, brightness or global color temperature

Contributions

- To learn a deep illumination model that can relight a given image, and use this model to better understand the failure modes of detection and segmentation DNNs

Brief overview of past approach

Shadow Transfer Network Architecture



Loss Functions:

- L1 loss on predicted L and ab channels
- Standard Perceptual and Style loss on predicted RGB
- Sun Estimation Perceptual loss on Predicted RGB

Brief overview of past approach

Sun Estimation Perceptual loss on Predicted RGB

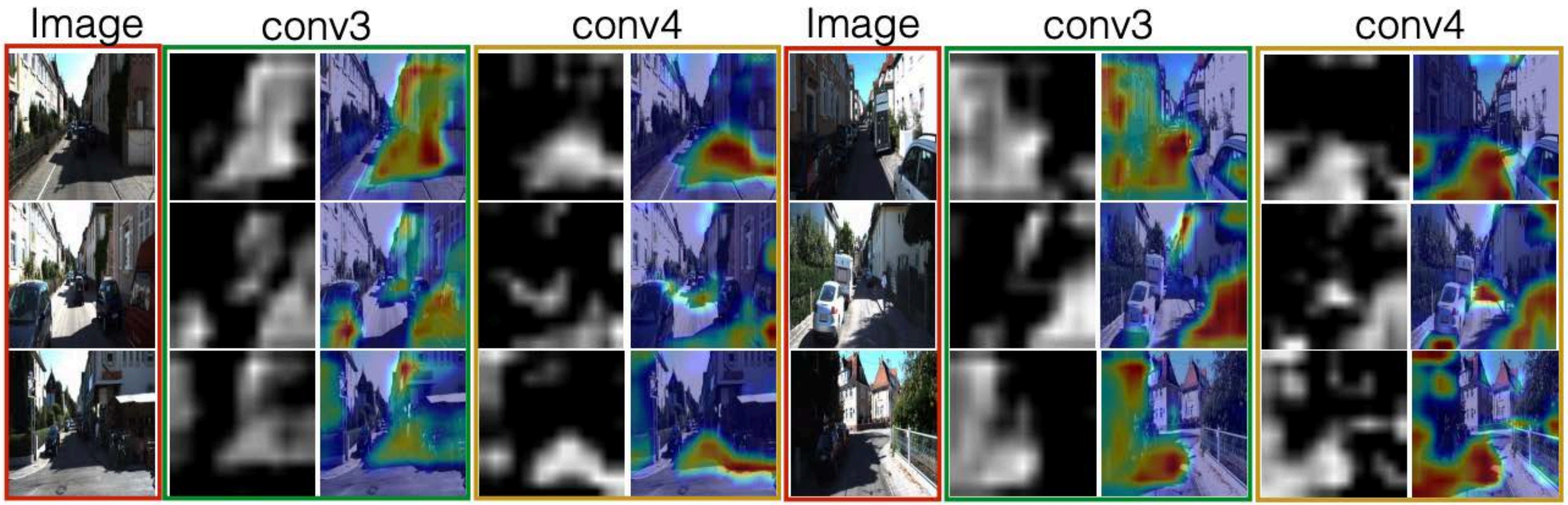
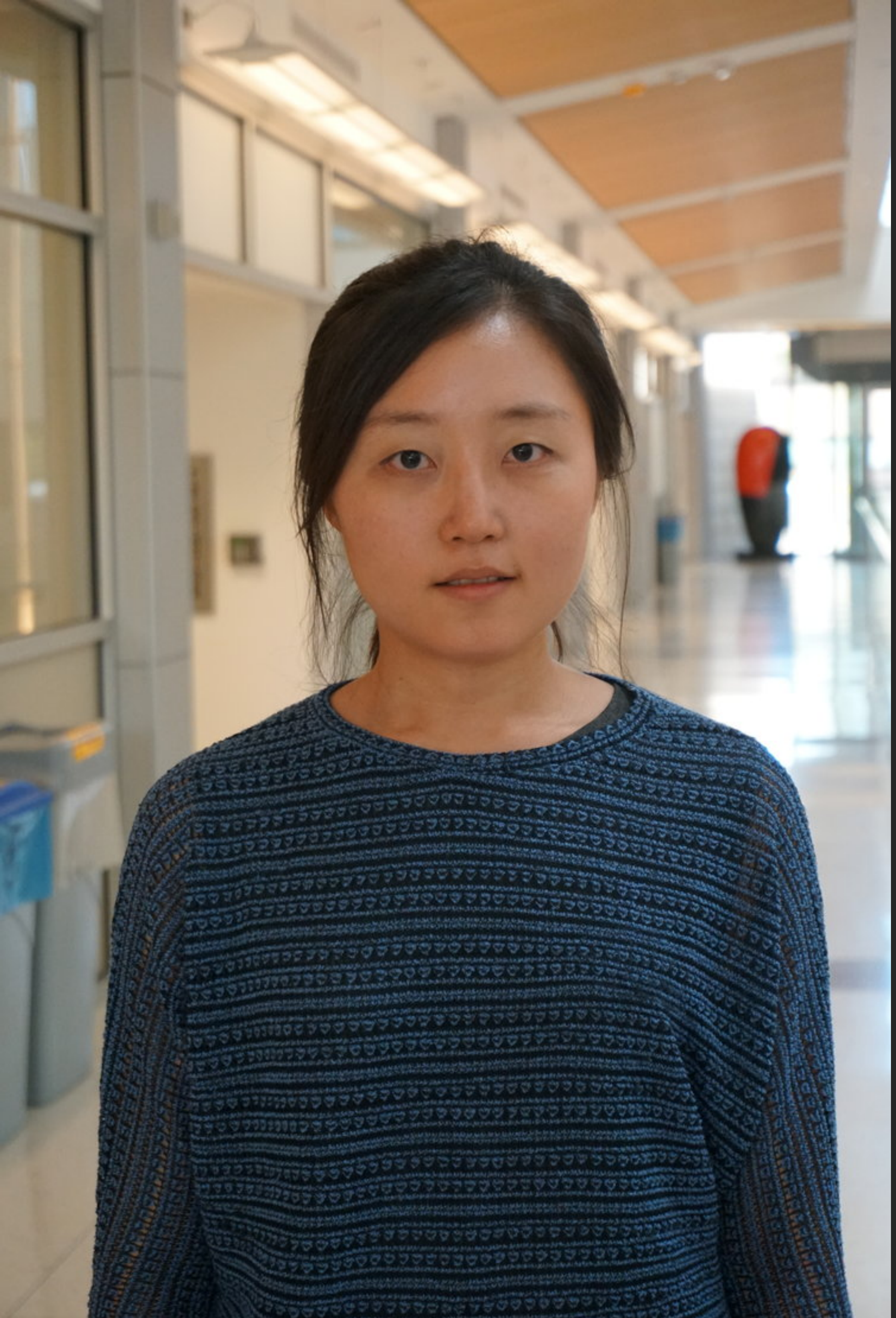


Figure 5: **Shading/shadow detectors emerge in Sun-CNN:** Test images and the corresponding activation maps of certain units in conv3 and conv4 layers of Sun-CNN. Despite being trained on *image-level* label (the relative sun position), our Sun-CNN automatically learns to fire on shadings (conv3) and shadows (conv4).

Ma et al, *Find your Way by Observing the Sun and Other Semantic Cues,* arxiv 2016

Brief overview of past approach

<u>Results: Real Dataset *KITTI-sun*</u>



ParametricX: 3D Reconstruction of Urban Intersections to Bridge the Gap Between Real and Synthetic Data

Wonhui Kim

Capturing Data at Urban Intersections

How do we capture full dynamics of the entire urban intersections?

Previously in **PedX**, we parked our capture vehicle at the curb
⇒ Limited perspective RGB images and LiDAR point clouds with occlusions

A moving vehicle passes through the intersection, and
after STOP sign it needs to choose a single route (Left turn/ straight/ right turn)
⇒ Not enough time to fully observe the surroundings,
⇒ Limited perspective data

Bird's-eye view data of the intersection is good to obtain trajectories,
⇒ Limited view data
⇒ Lack of data other than trajectories

Dense 3D Reconstruction of Intersections

Bridging gaps between real and synthetic data:

Real *trajectories of dynamic agents* + ***Synthetic*** *reconstruction of static/dynamic components* + ***Real*** *scene geometry*

Urban intersection consists of many scene components.

Background	Static objects	Dynamic objects
Ground Lanes Sidewalks Crosswalks Parking lots	Buildings Trees Lampposts Road poles Traffic signs Trash bins Bike racks	Pedestrians Vehicles

3D Model Fitting

Scene backgrounds are modeled based on plane fitting and manual labeling using *Blender*.

Static scene objects are reconstructed by fitting 3D CAD models from *ObjectNet3D dataset.*

Pedestrians from *PedX dataset* were adjusted to be consistent with other scene models.

Vehicles are reconstructed by fitting 3D models with the following steps:
- LiDAR point cloud segmentation
- Global trajectory fitting
- Optimization to determine vehicle pose (translation, heading orientation)

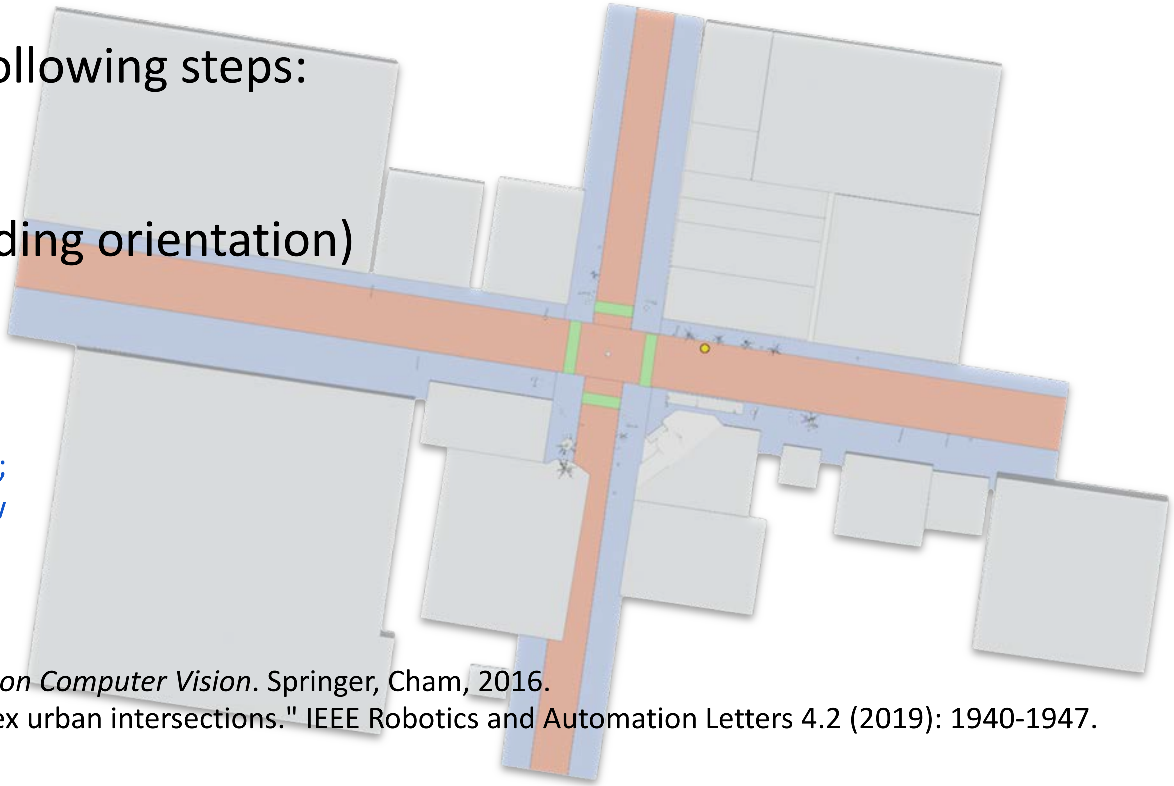
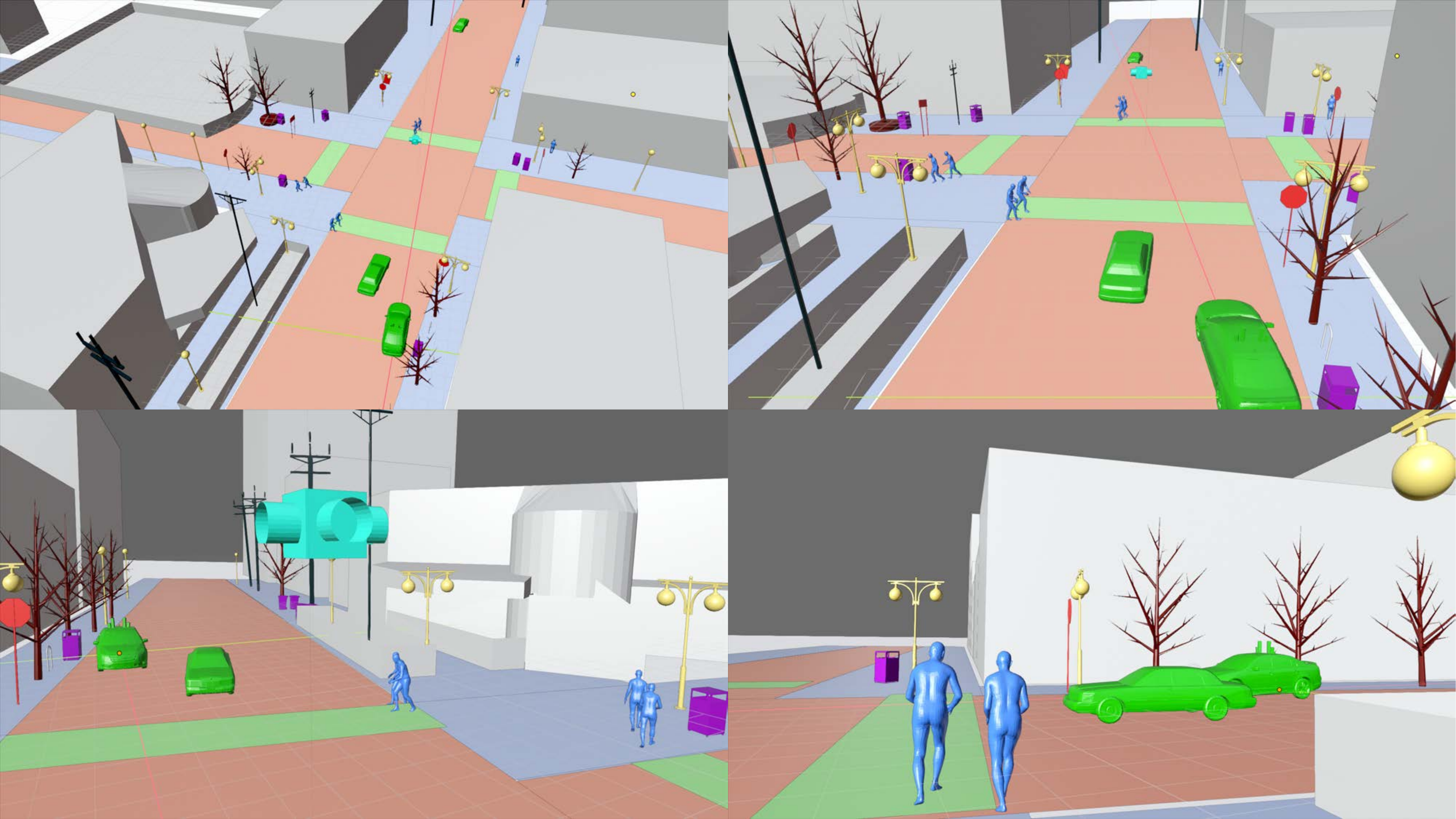


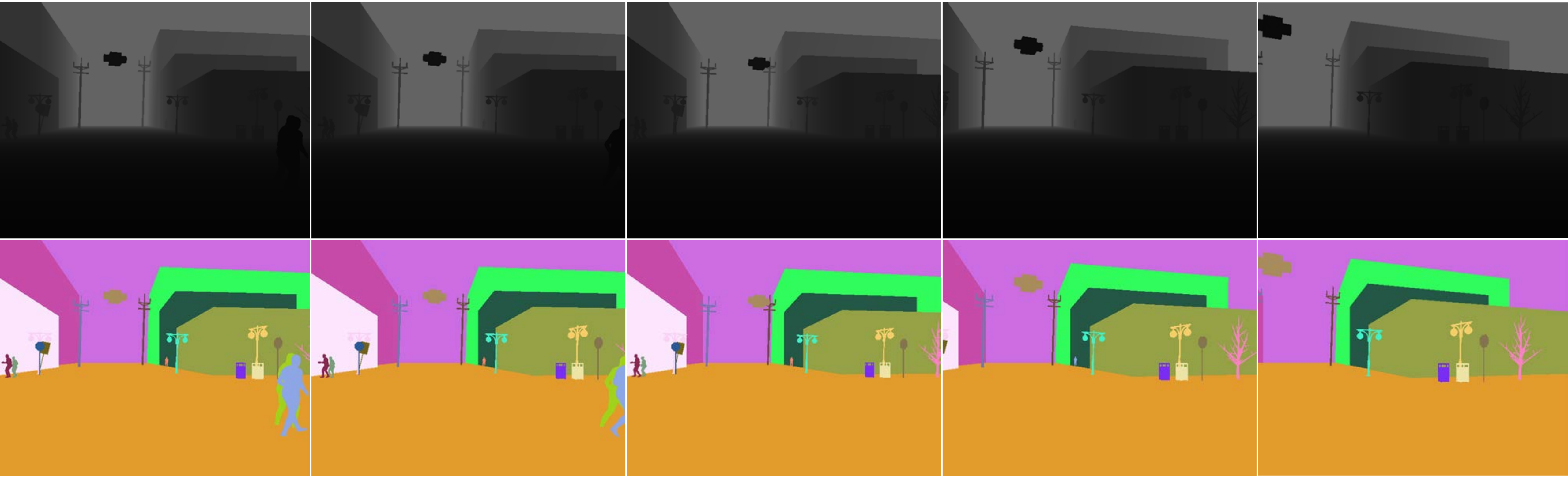
Figure: Lanes, sidewalks, crosswalks, buildings are shown; Rendered from a bird's eye view

"**Blender** - a 3D modelling and rendering package", http://www.blender.org, 2018.
Xiang, Yu, et al. "**Objectnet3d**: A large scale database for 3d object recognition." *European Conference on Computer Vision*. Springer, Cham, 2016.
Kim, Wonhui, et al. "**PedX**: Benchmark dataset for metric 3-D pose estimation of pedestrians in complex urban intersections." IEEE Robotics and Automation Letters 4.2 (2019): 1940-1947.



Generating Depth and Label Images from Simulation

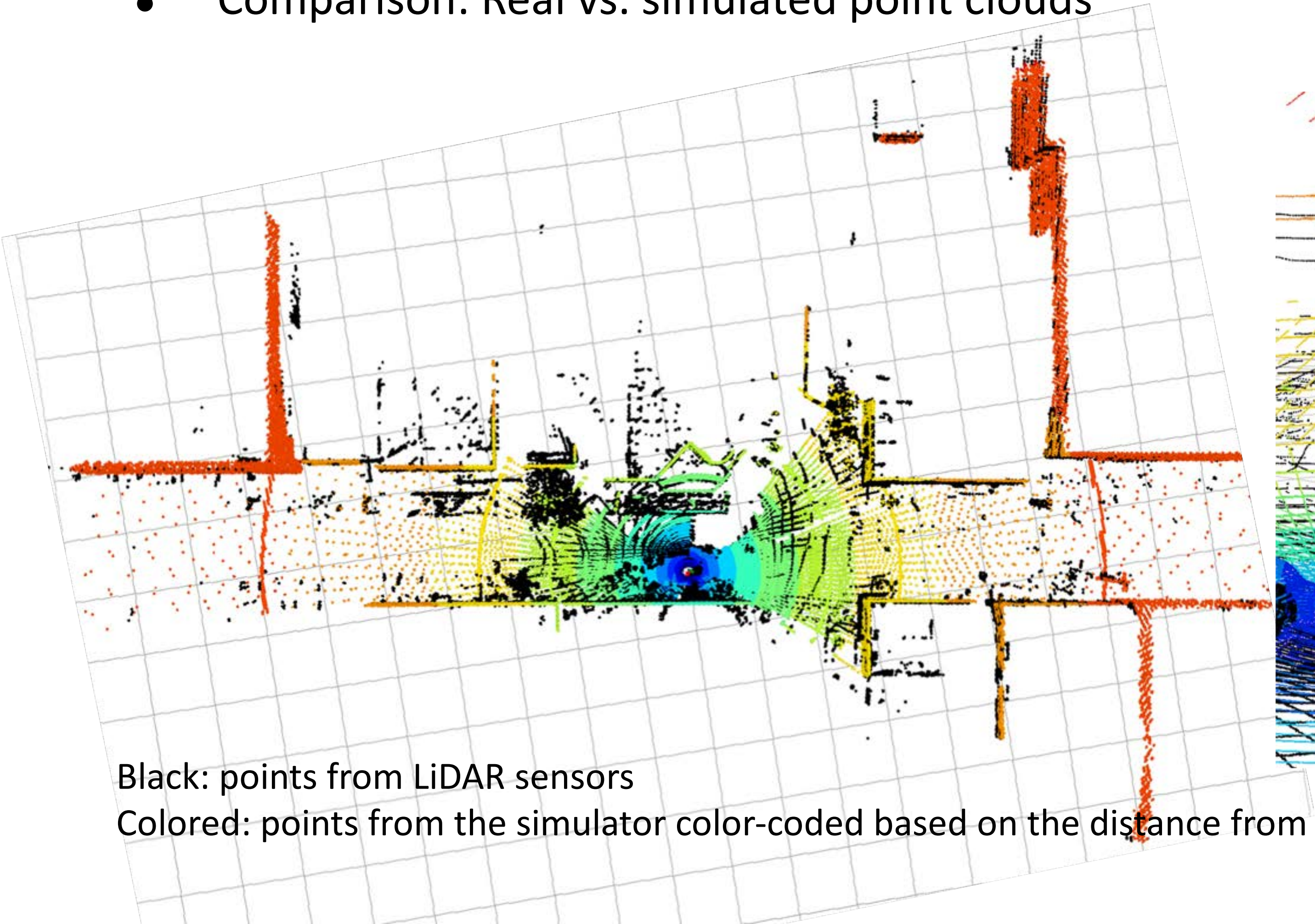
- A virtual camera is placed at a vehicle turning right after the STOP line.
- The trajectory is from the real data capture.

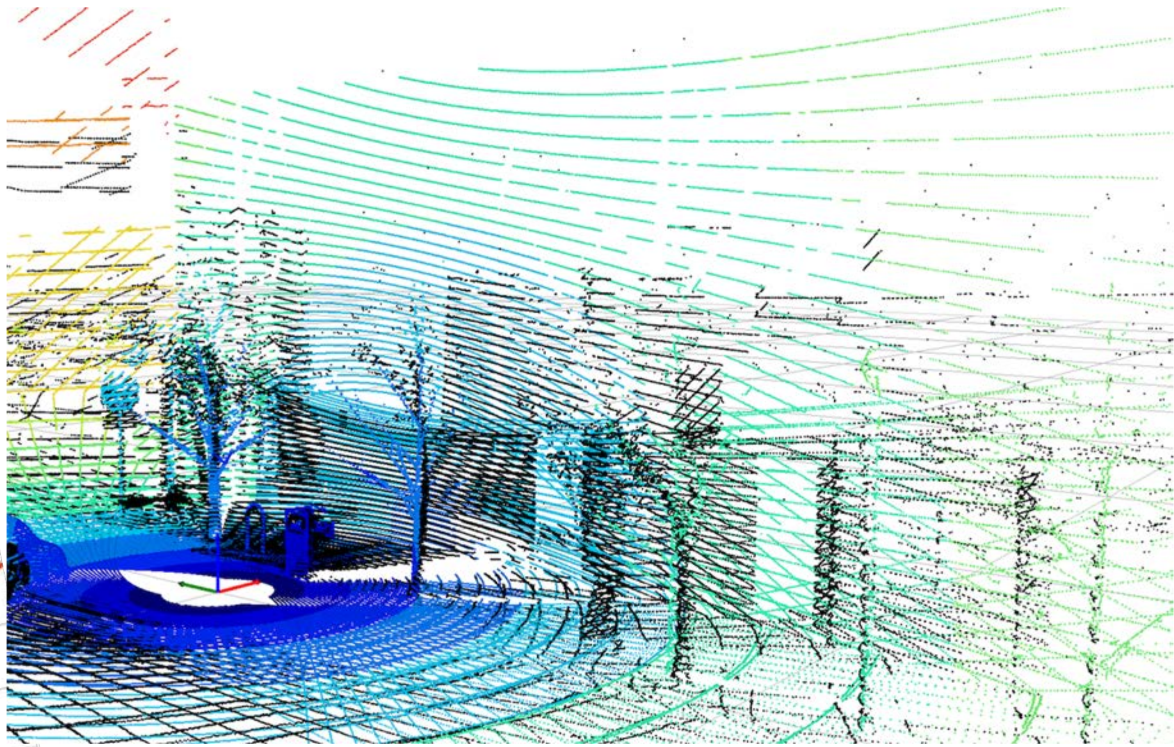


Depth maps (top) / Instance-level label images (bottom)

Generating LiDAR Point Clouds from Simulation

- Virtual LiDARs are placed on the roof of the capture vehicle as in the real configuration.
- Comparison: Real vs. simulated point clouds





Black: points from LiDAR sensors
Colored: points from the simulator color-coded based on the distance from the LiDAR origin.

Generating Trajectories from Prediction

Cyrus Anderson

Off The Beaten Sidewalk: Pedestrian Prediction In Shared Spaces For Autonomous Vehicles (Cyrus Anderson, Ram Vasudevan, M. Johnson-Roberson), *In IEEE Robotics and Automation Letters (RA-L) Special Issue on Long-Term Human Motion Prediction*, 2020

Generating Trajectories from Prediction

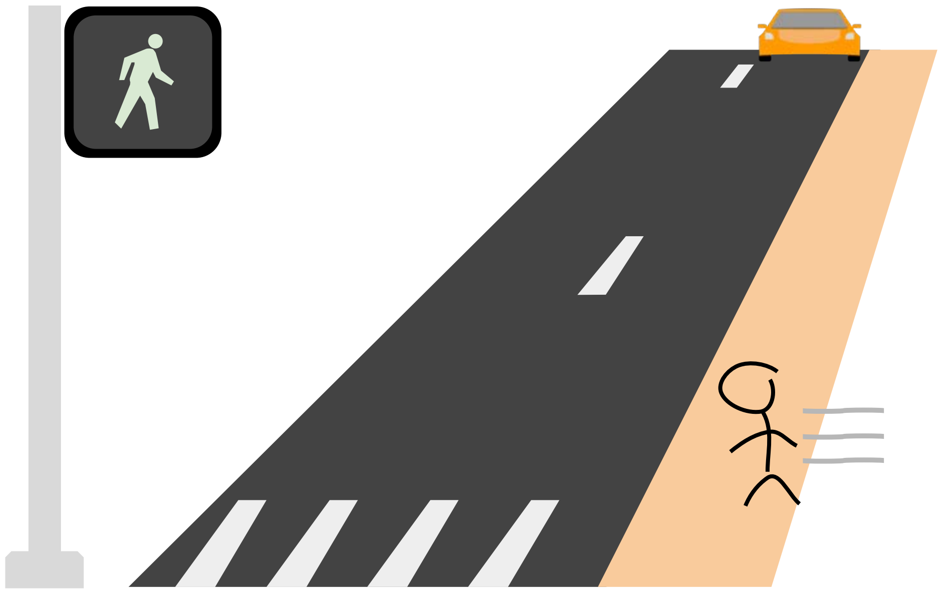
- How to get data for training pedestrian prediction algos

- Anderson, Cyrus, et al. "Stochastic Sampling Simulation for Pedestrian Trajectory Prediction." arXiv preprint arXiv:1903.01860 (2019).
- Du, Xiaoxiao, Ram Vasudevan, and Matthew Johnson-Roberson. "Bio-lstm: A biomechanically inspired recurrent neural network for 3-d pedestrian pose and gait prediction." IEEE Robotics and Automation Letters 4.2 (2019): 1501-1508.
- Yao, Yu, et al. "BiTraP: Bi-directional Pedestrian Trajectory Prediction with Multi-modal Goal Estimation." arXiv preprint arXiv:2007.14558 (2020).
- Zhao, Tianyang, et al. "Multi-agent tensor fusion for contextual trajectory prediction." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.
- Ma, Yuexin, et al. "Trafficpredict: Trajectory prediction for heterogeneous traffic-agents." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33. 2019.
- Xue, Hao, Du Q. Huynh, and Mark Reynolds. "SS-LSTM: A hierarchical LSTM model for pedestrian trajectory prediction." 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2018.

Standard Pedestrian Prediction

Key ingredients:

- Pedestrian
- Vehicles



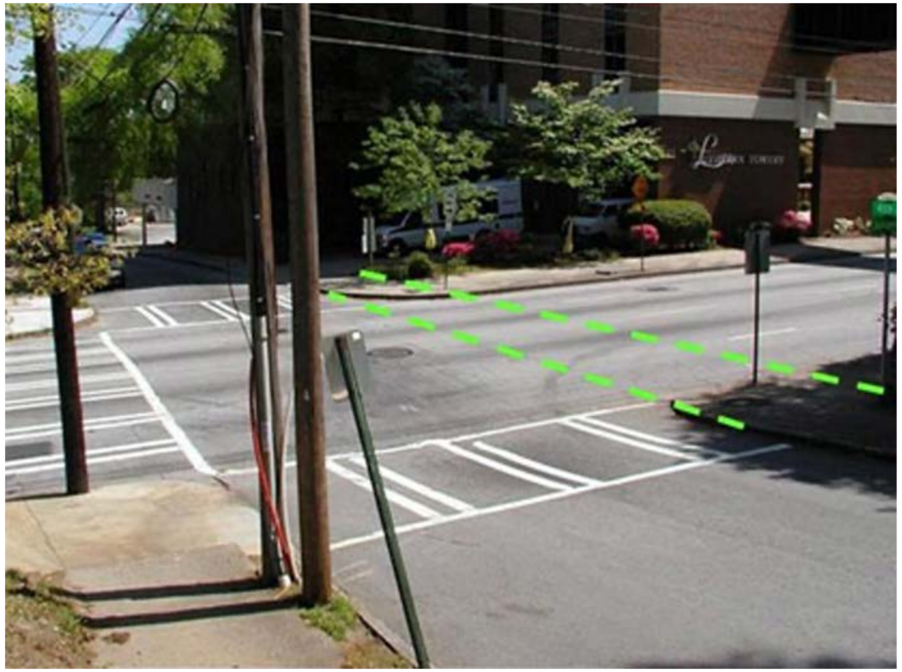
Infrastructure:

- Curbs
    - [Kooiji et al., IJCV ‘19]
- Marked crosswalk
    - [Blaiotta, RA-L ‘19]
    - [Jayaraman et al., ICRA ‘20]
- Signalized intersection
    - [Hashimoto et al., ITS ‘15]

Shared Space

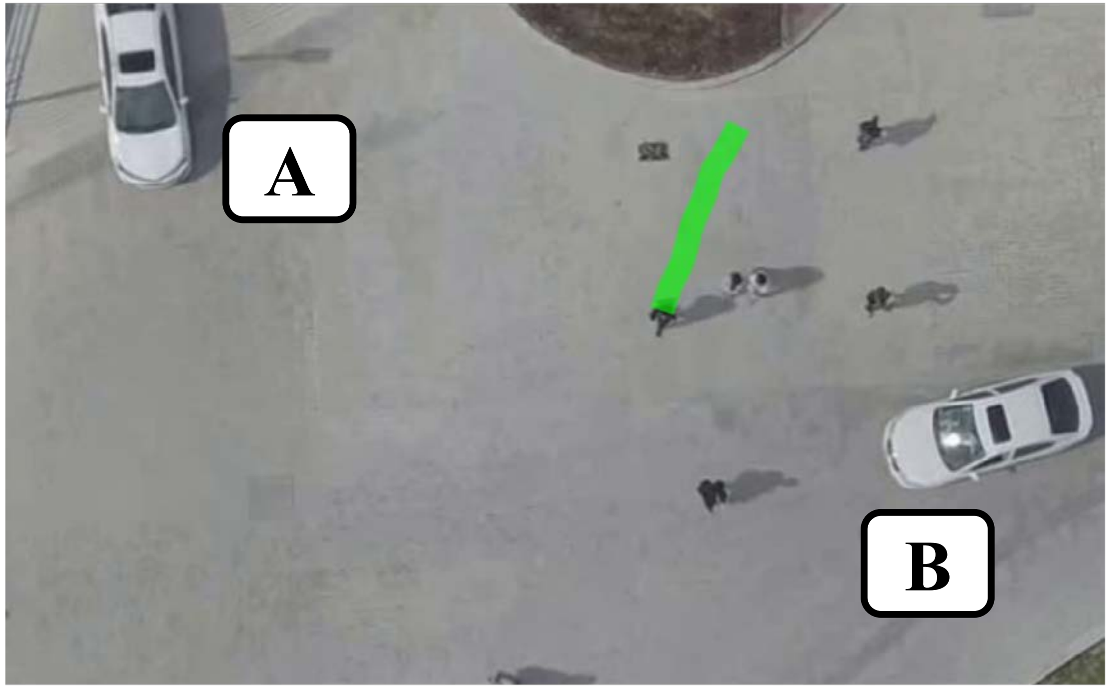
Predictions in more general scenes

- Infrastructure
  - Unmarked crosswalks
- Pedestrian behavior
  - May change across scenes
  - Less than 100% adherence to traffic rules

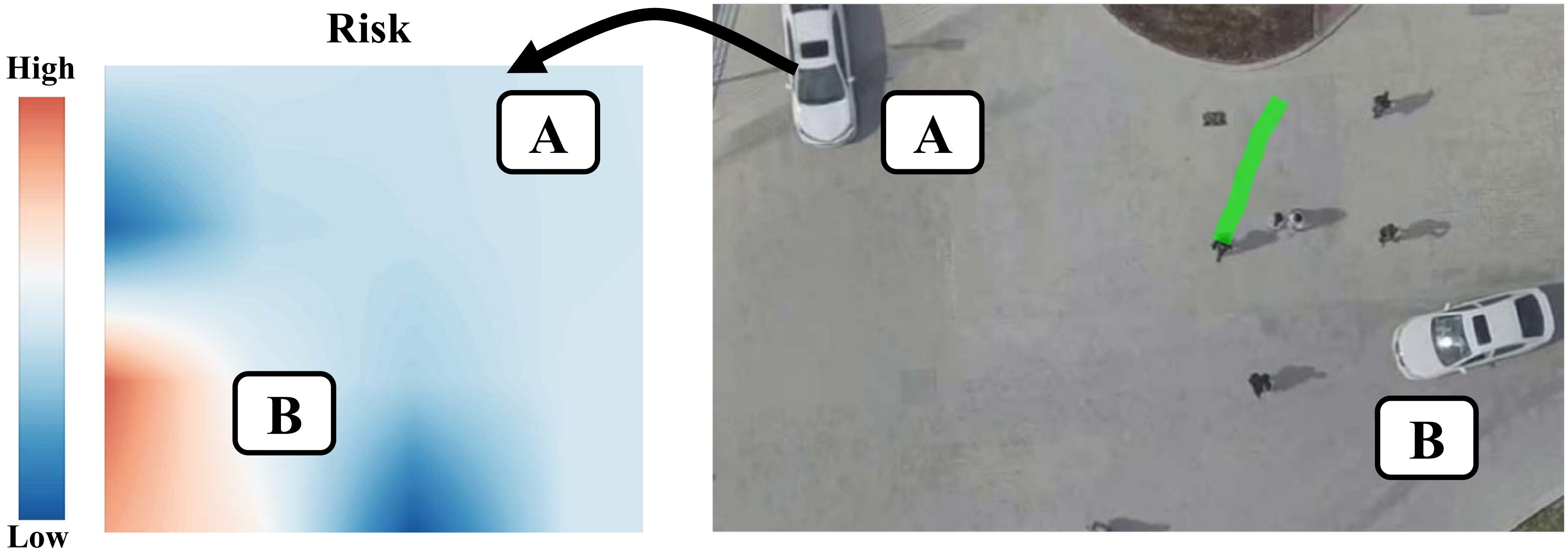


Predictions off the sidewalk

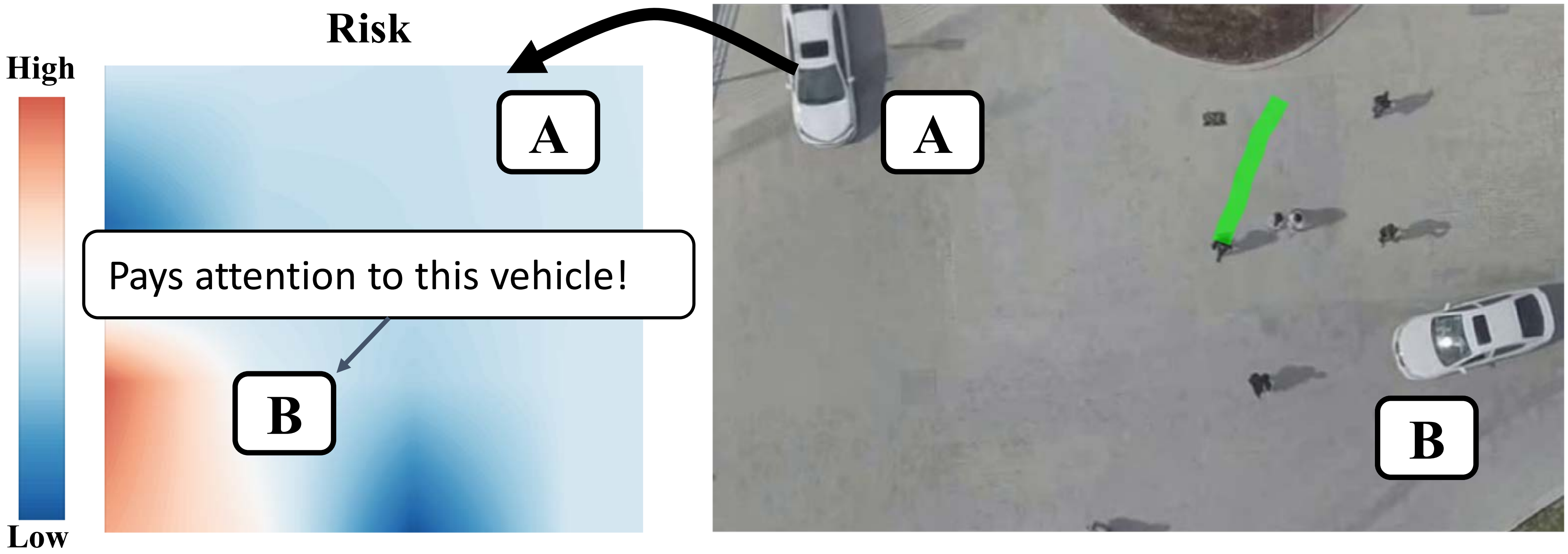
Predictions off the sidewalk



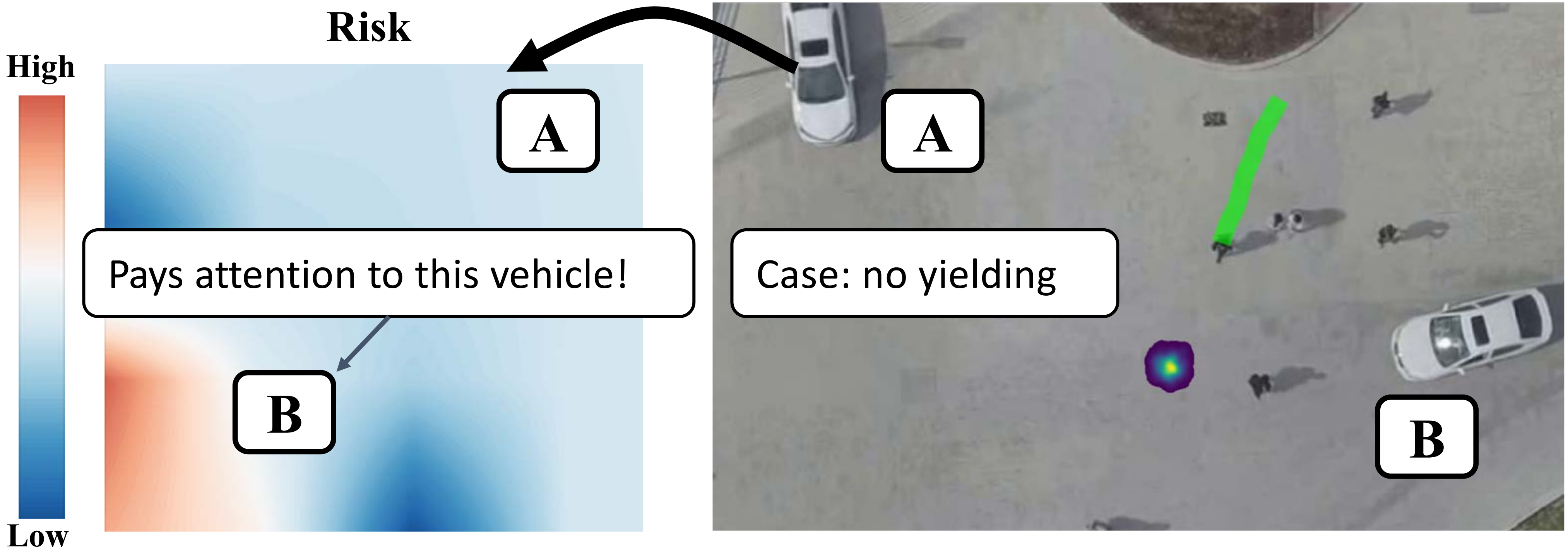
Predictions off the sidewalk



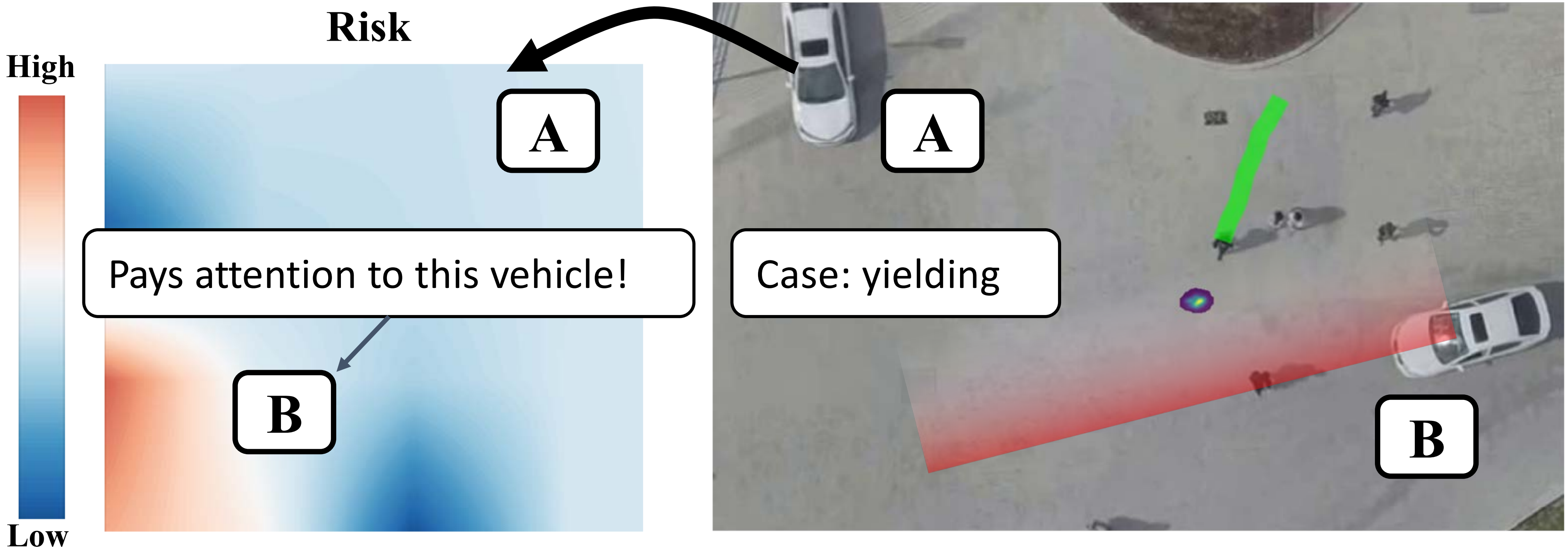
Predictions off the sidewalk



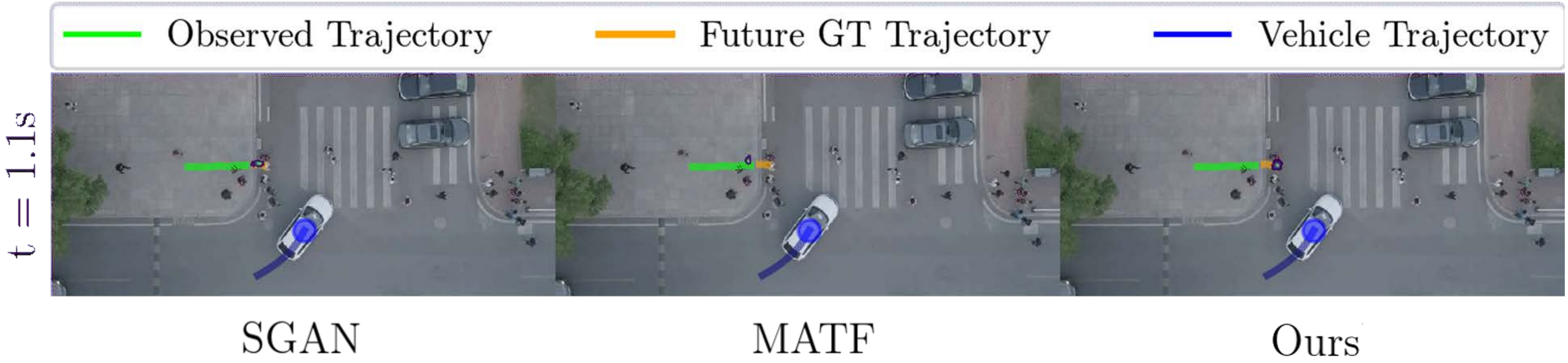
Predictions off the sidewalk

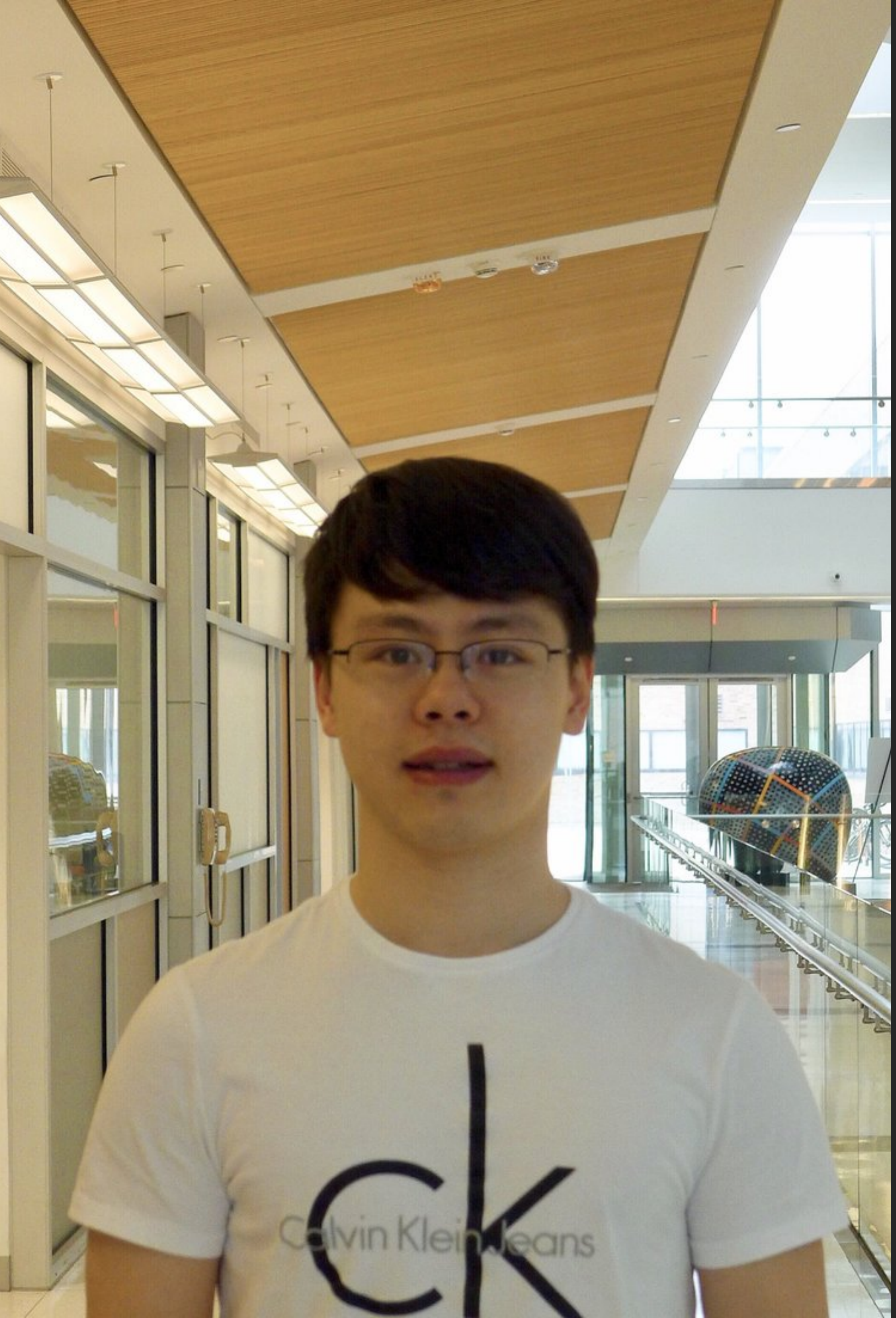


Predictions off the sidewalk



Predicted distributions - DUT



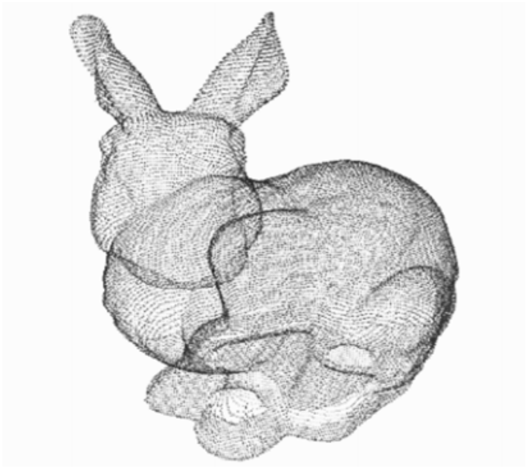


Point Set Voting for Partial Point Cloud Analysis

Junming Zhang

Motivation

Depth Sensors



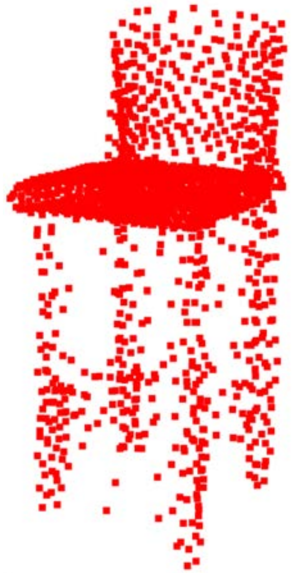
Point clouds

Point clouds are easily generated by depth sensors

Motivation

CAD models from ShapeNet

Sample

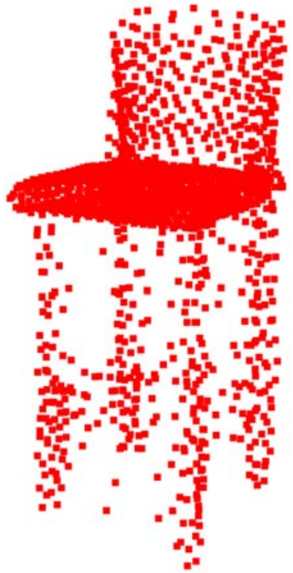


Synthetic Point clouds

Synthetic point clouds are generated by sampling from CAD models

Motivation

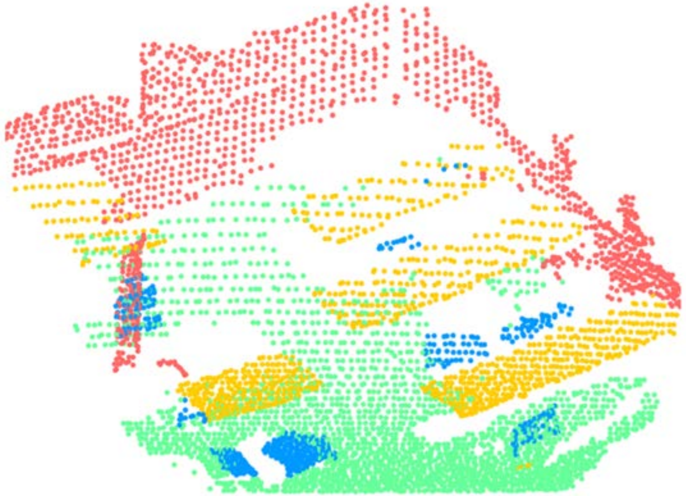
- RS-CNN [Liu, et al.]
- DG-CNN [Wang, et al.]
- SF-CNN [Rao, et al.]
- Pointnet [Qi, et al.]
- Pointnet++ [Qi, et al.]



Synthetic Point clouds

Many methods developed for analyzing point clouds are based on synthetic dataset

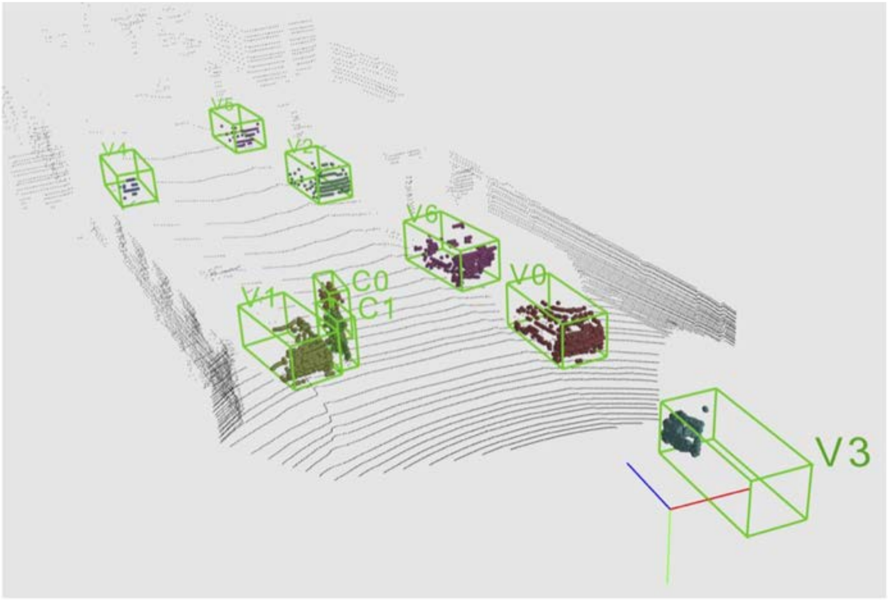
Motivation



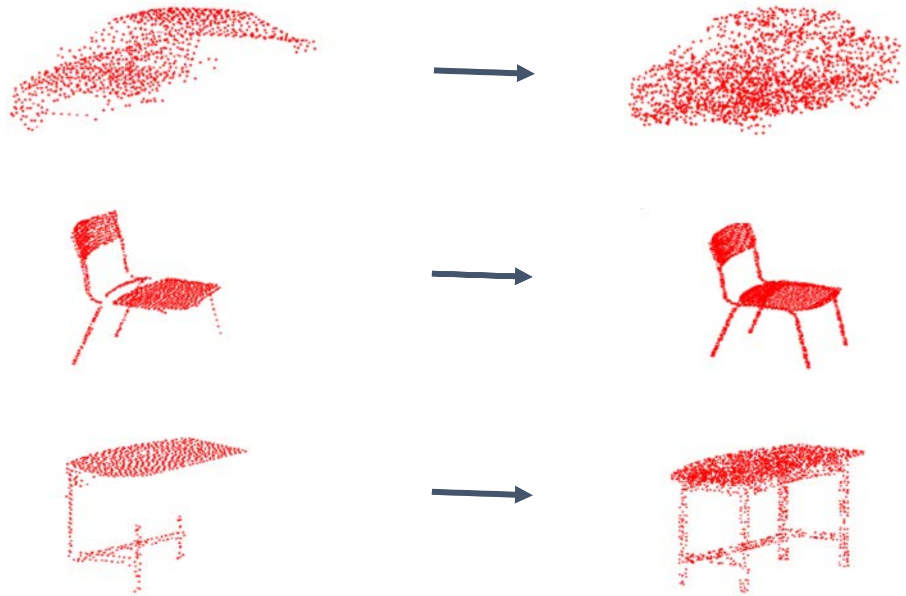
Real-world point clouds are usually incomplete

Motivation

Training on incomplete point clouds

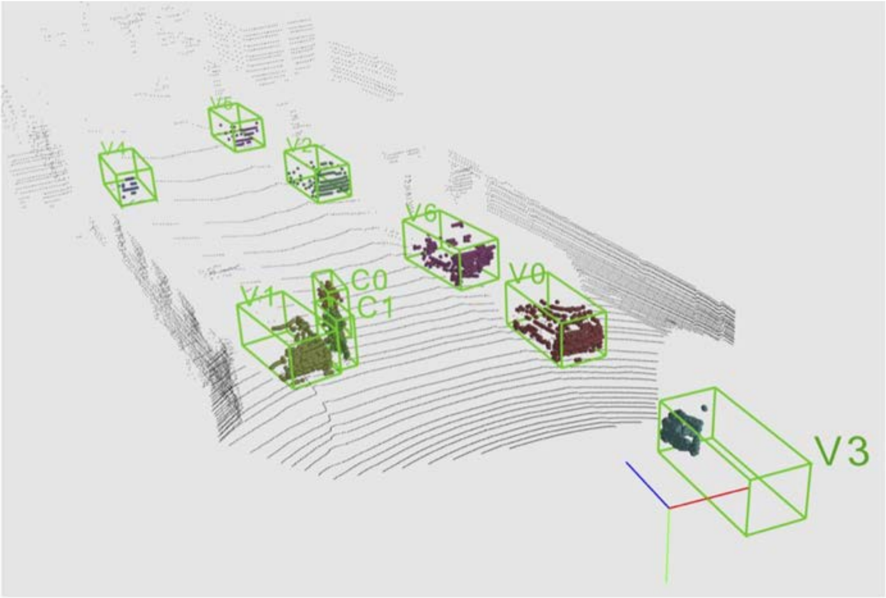


Complete partial point clouds



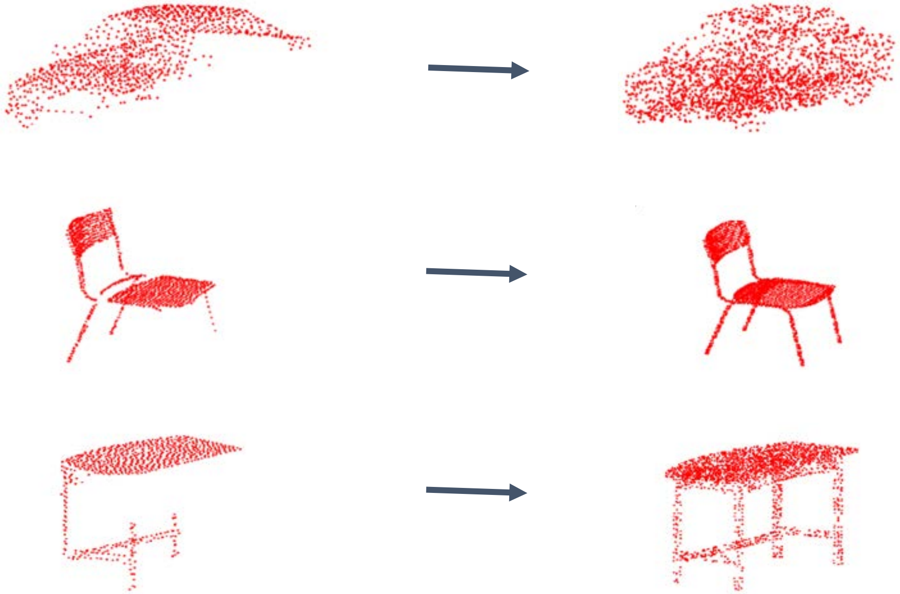
Motivation

Training on incomplete point clouds



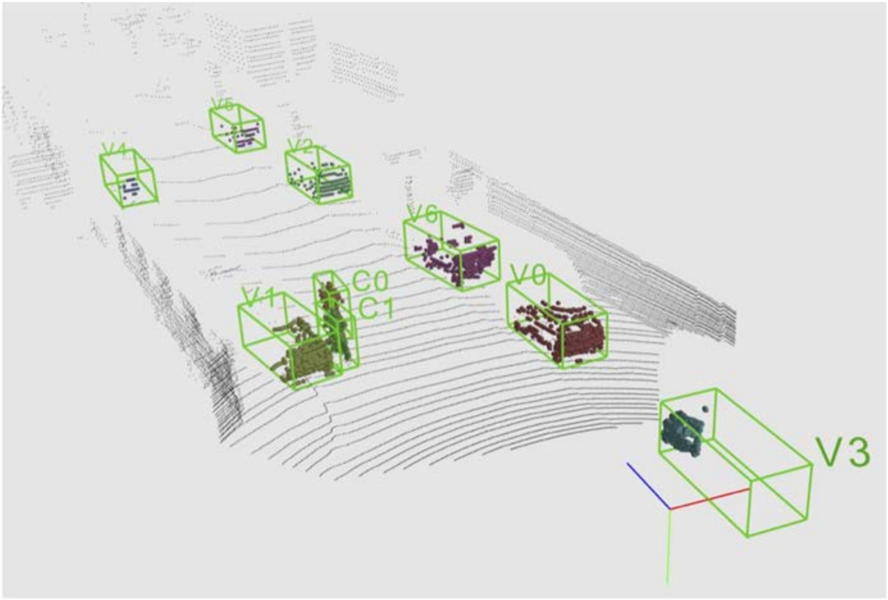
Annotation is expensive

Complete partial point clouds



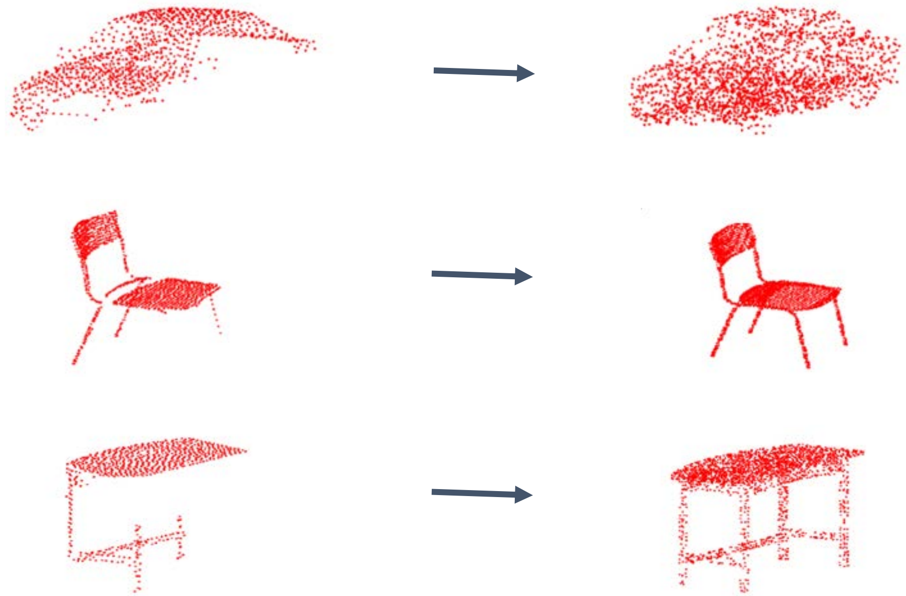
Motivation

Training on incomplete point clouds



Annotation is expensive

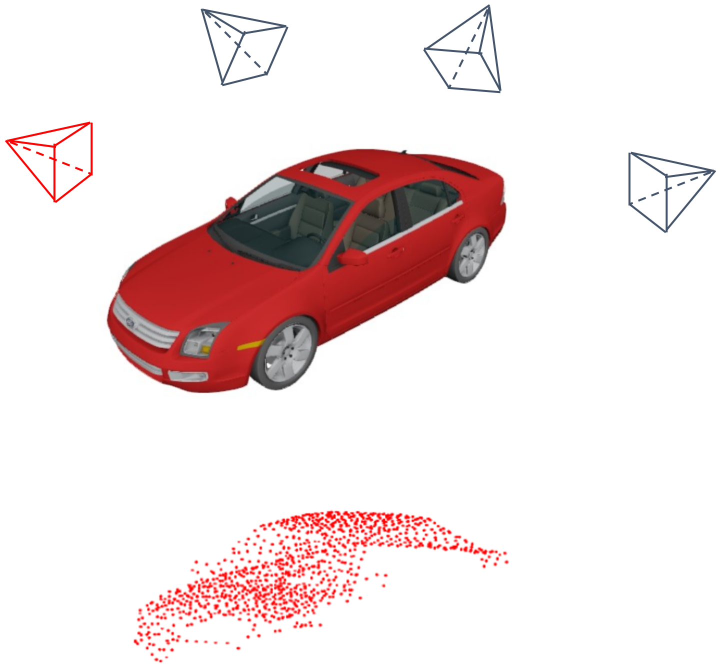
Complete partial point clouds

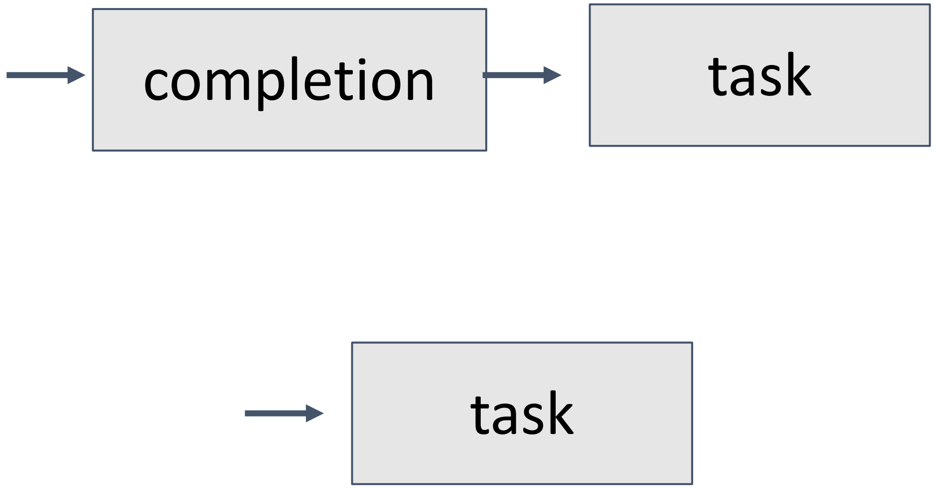


Limitations

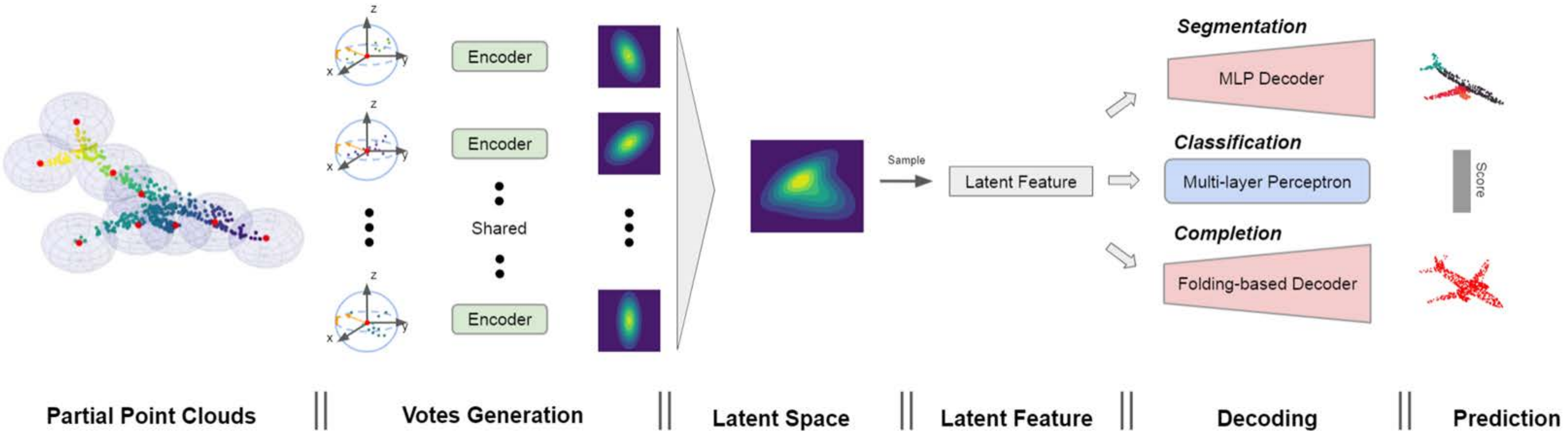
Motivation

Point clouds completion



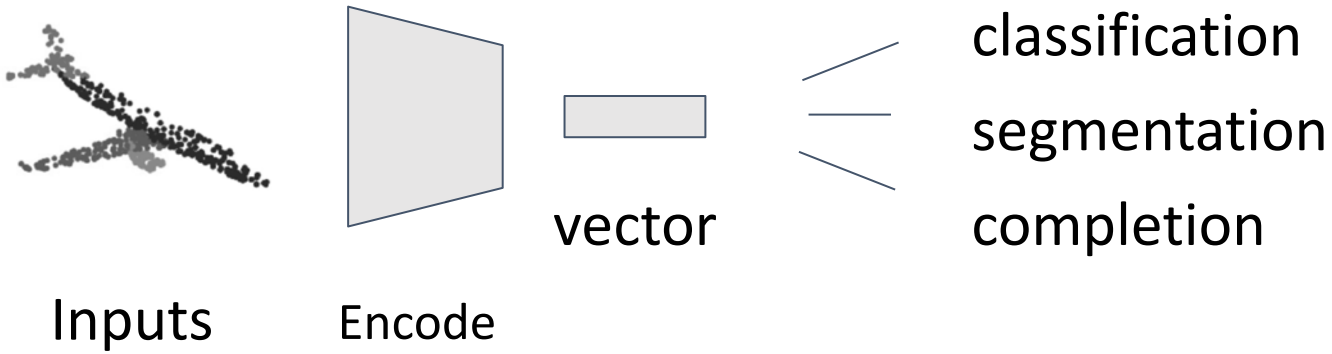


Method



One-stage model for any partial point clouds analysis

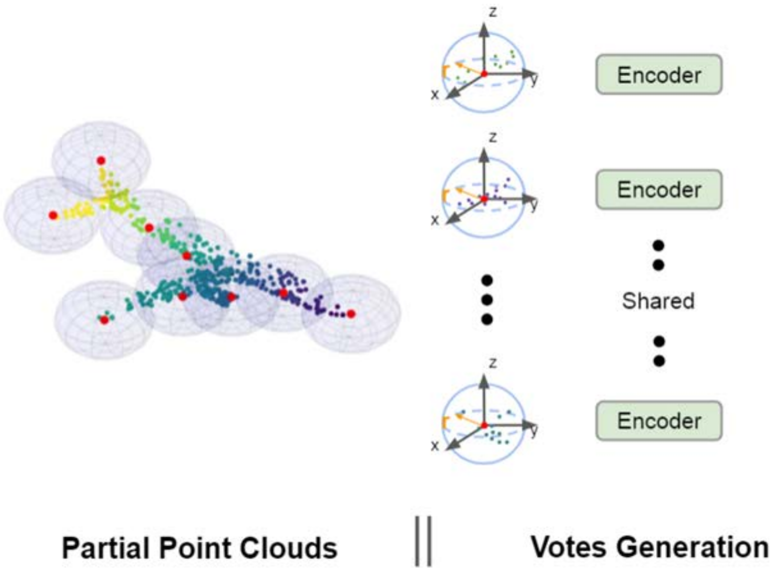
Method

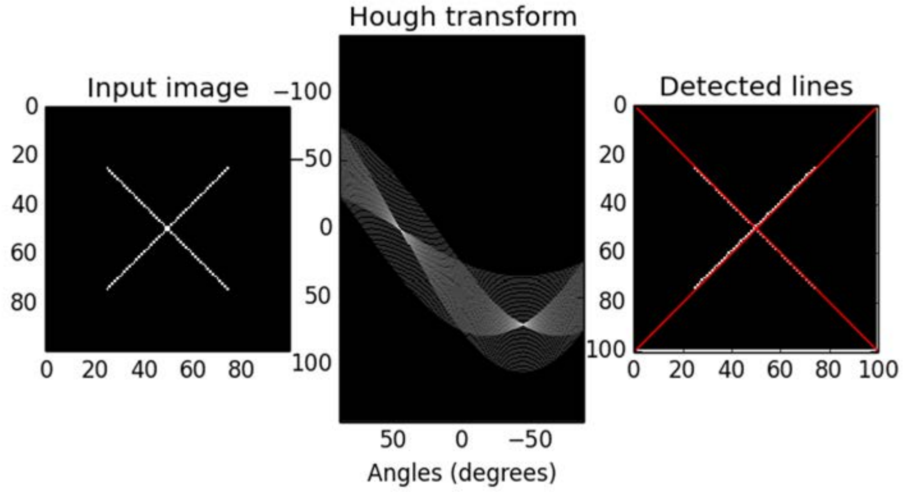


1. Mapping different inputs into the same feature vector
2. Not able to transfer to other incomplete point clouds

Method

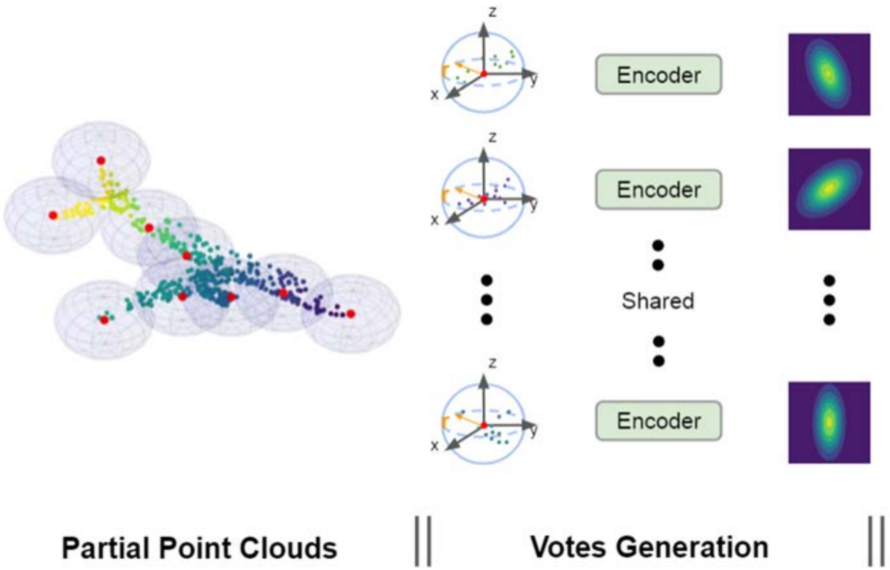
Hough transform





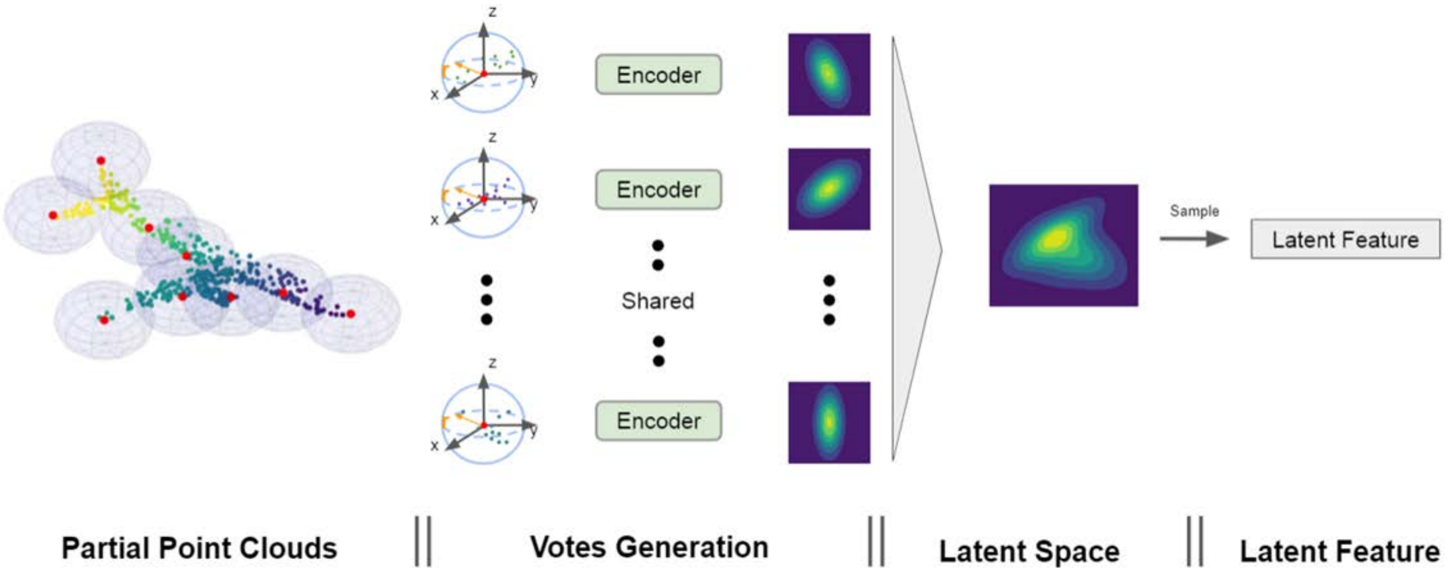
Propose voting strategy to infer the feature for encoding complete PC

Method



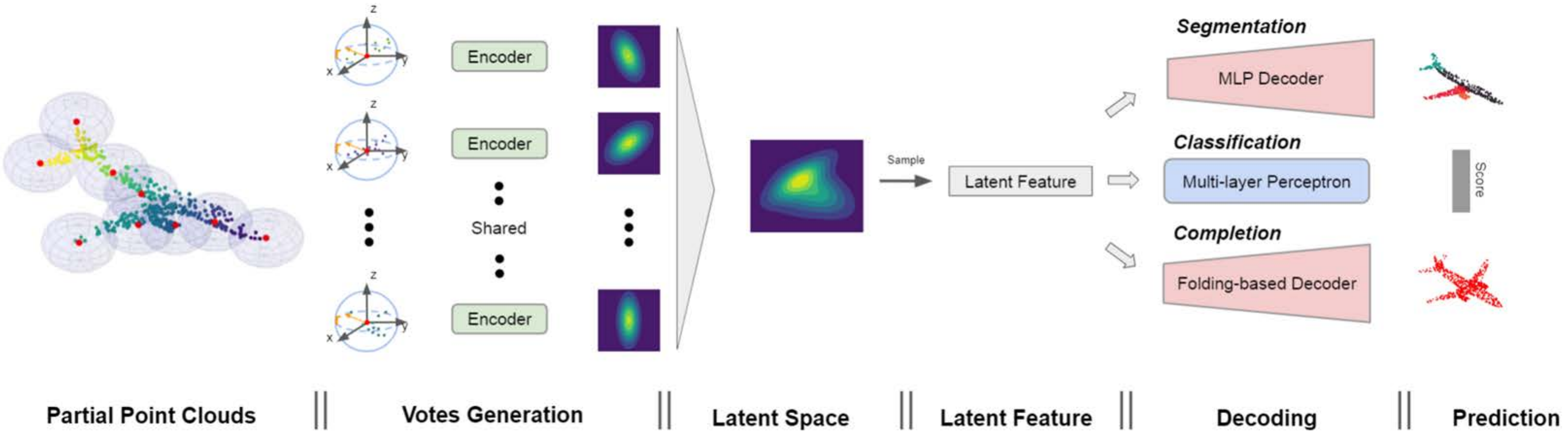
Each vote is a distribution in the latent space

Motivation



Latent feature is sampled from constructed latent space

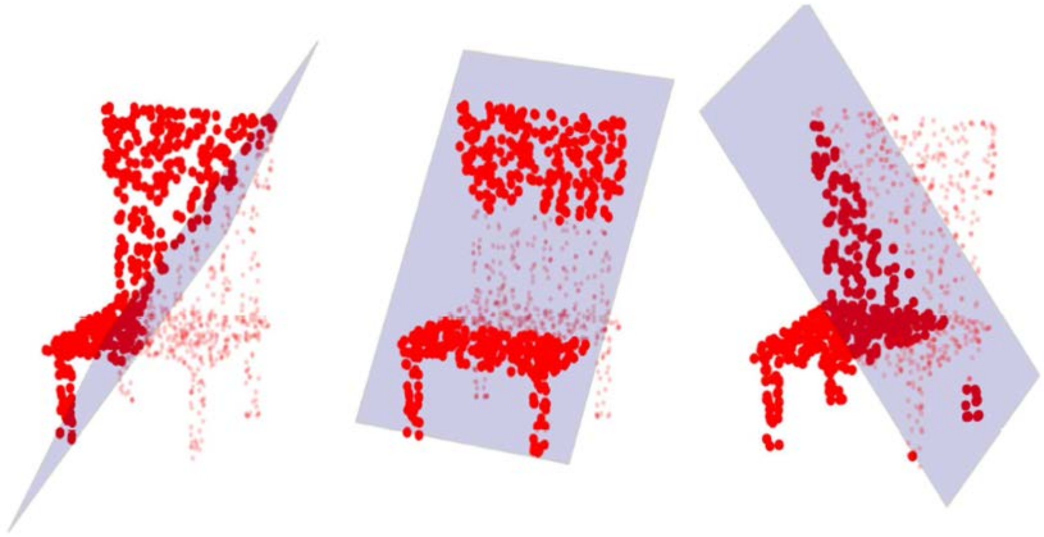
Motivation



Latent feature is passed to decoding modules

Results

Method	Input	Complete	Partial
PointNet [33]	xyz	88.8	20.9
PointNet++ [35]	xyz	91.0	61.5
RS-CNN [27]	xyz	92.3	43.3
DG-CNN [47]	xyz	**92.9**	51.5
Ours	xyz	91.4	**86.4**



Shape classification on ModelNet40

Results

Results on complete point clouds in ShapeNet
from models trained on ShapeNet

Part Segmentation trained on ShapeNet

Results

Results on simulated partial point clouds from models trained on ShapeNet

Part Segmentation trained on ShapeNet

Results

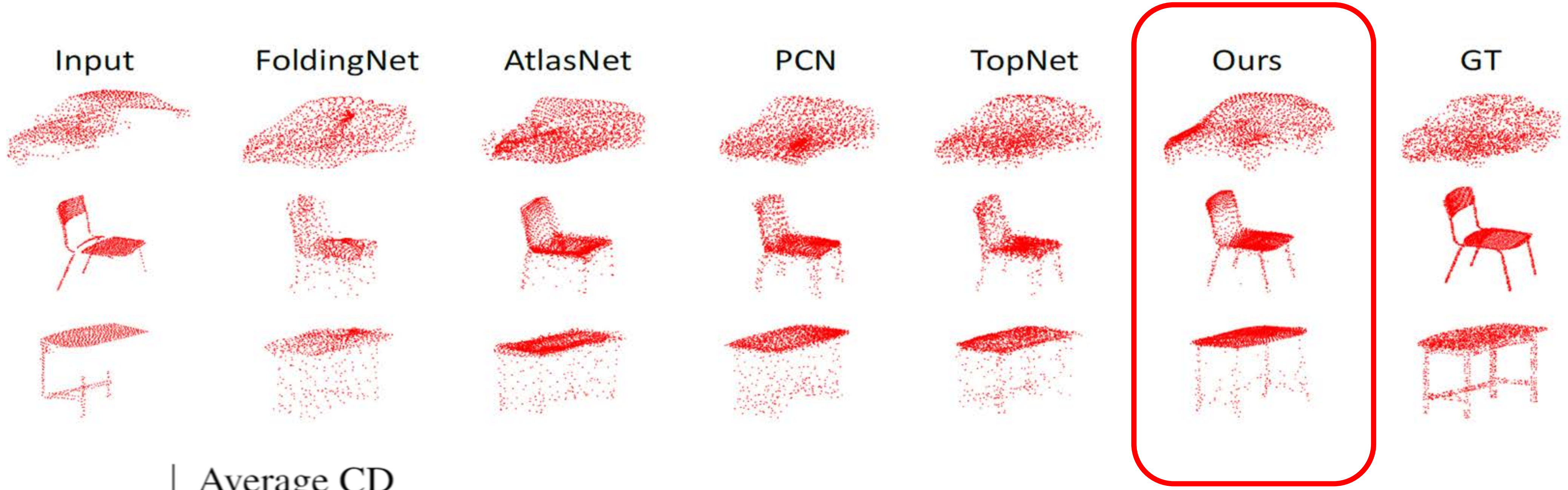
Results on point clouds in Completion3D from models trained on ShapeNet

Part Segmentation trained on ShapeNet

Results

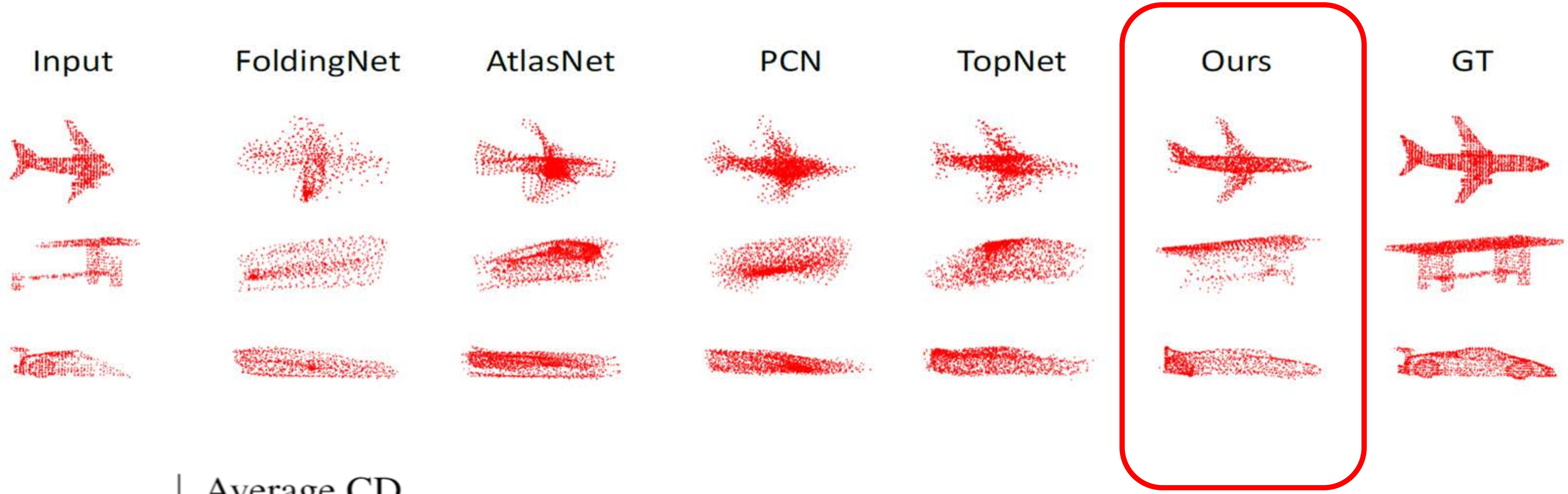
Method	Input	Complete	Partial
PointNet [33]	xyz	80.5	29.9
PointNet++ [35]	xyz	82.0	30.9
DG-CNN [47]	xyz	82.3	29.8
RS-CNN [27]	xyz	**82.4**	30.6
Ours	xyz	79.0	**78.1**

Part Segmentation trained on ShapeNet



Model	Average CD
FoldingNet [51]	19.07
PCN [52]	18.22
AtlasNet [13]	17.77
TopNet [45]	**14.25**
Ours	18.18

Point clouds completion



Model	Average CD
FoldingNet [51]	34.56
PCN [52]	34.93
AtlasNet [13]	39.73
TopNet [45]	31.87
Ours	**17.22**

Point clouds completion

To current students

Questions?