From HD Maps to No Maps

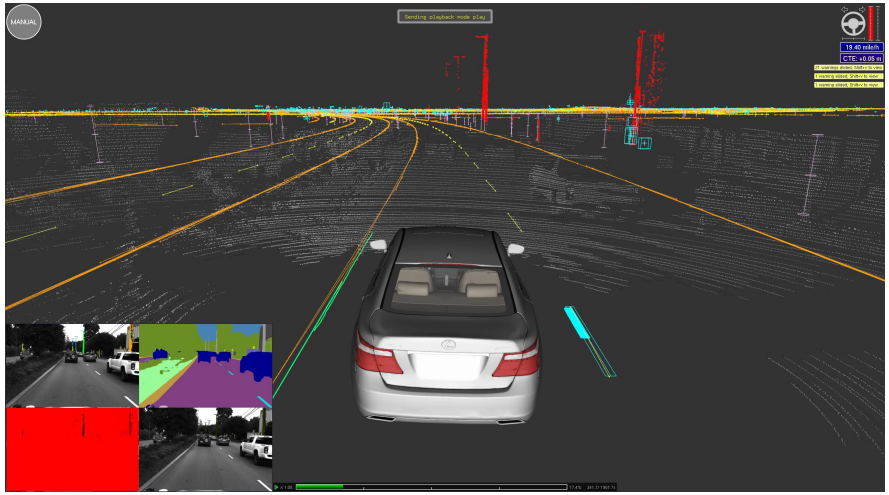
Wolfram Burgard

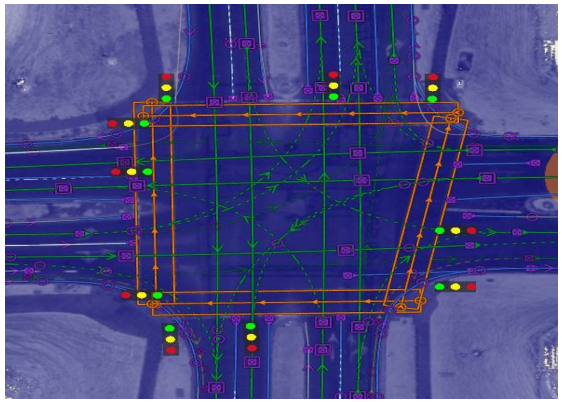
Joint work with: David Pannen, Martin Liebner, Jannik Zürn, Johan Vertens, Kshitij Sirohi, Juana Valeria Hurtado, Rohit Mohan, Daniel Büscher, Abhinav Valada …

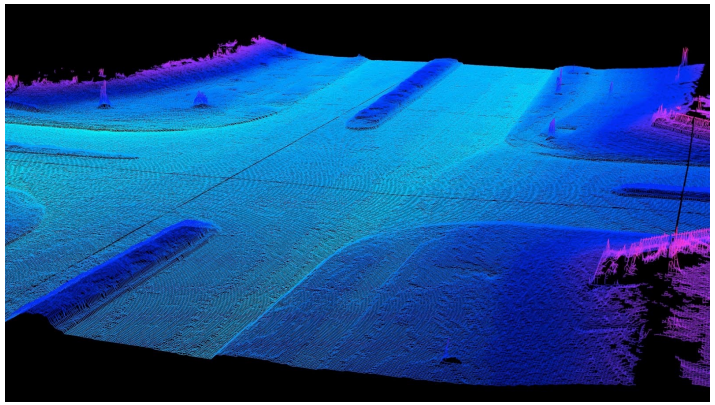
Maps in Automated Driving and Robotics

Useful for

- Perception
- Tracking
- Localization
- Prediction
- Planning
- Control
- …

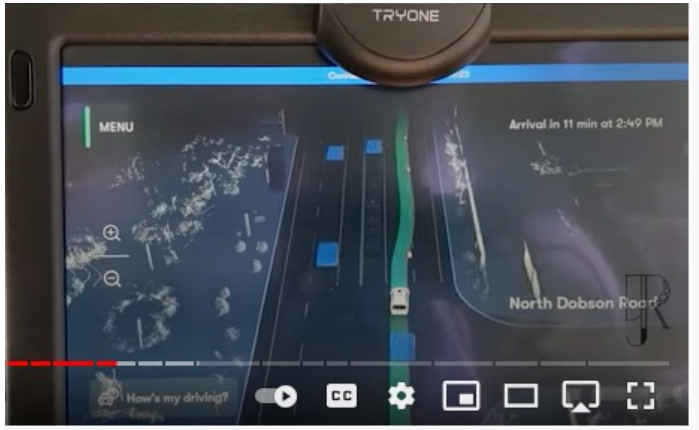






Challenges for HD Maps

- Expensive to acquire
- Assumptions about availability of features
- Change detection
- Expensive to update
- L5 barrier
- …



In this Talk

- Change detection through **crowdsorcing**
- **Learning** for on-the-fly estimation of HD map information
  - Semantics
  - Tracking
  - Topology

Change Detection for HD Maps

2019 International Conference on Robotics and Automation (ICRA)
Palais des congres de Montreal, Montreal, Canada, May 20-24, 2019

HD Map Change Detection with a Boosted Particle Filter

David Pannen[1], Martin Liebner[1] and Wolfram Burgard[2]

***Abstract*— In this paper, we present a change detection algorithm that can run in real time as part of a backend-based stream processing pipeline. It can process the floating car data collected by series-production vehicles to detect changes in an automotive high definition digital (HD) map used for automated driving. The algorithm uses a particle filter approach with odometry, GNSS and landmark readings to localize the vehicle within the digital map. While all particles together represent the probability distribution for the vehicle's position at a given time, each individual particle also serves as a hypothesis about the vehicle's position. This is used to compute various metrics for how well the current sensor readings match the world model encoded in the HD map. The different metrics are evaluated by a number of weak classifiers that are used as input for a trained Adaboost classifier. The achievable detection rate of a single vehicle is then compared to that of a simple crowd-based approach, where each vehicle votes on whether or not the current section of the road has changed.**

I. Introduction

To achieve the goal of safe, comfortable and efficient automated driving, high definition digital maps (HD maps) may be used to augment the vehicle's on-board sensors with prior beliefs about the environment and thus help to provide localization and sensor outlier detection. HD maps can be part of a fall-back strategy in case of sensor failure, but they also contribute to an increased automated driving comfort by enhancing the vehicle's foresight. Finally, they contain the baseline knowledge about drivable lanes, lane connectivity and applicable traffic regulations that is necessary for automated driving.

The current state of the art for creating HD maps uses mapping vehicles equipped with a suite of high accuracy sensors such as light detection and ranging (LIDAR) sensors or high accuracy global navigation satellite system (GNSS) sensors [1]. The problem with this approach is the high cost and resulting limited number of mapping vehicles. The mapping company HERE, for example, at the end of 2015, used a fleet of just over 200 cars worldwide [1]. As those vehicles can cover only a small fraction of the road network per day, the frequency of visits for each road segment is low and changes cannot be incorporated into the map in time.

This poses a high risk for automated driving, as a faulty map could lead to dangerous situations or even accidents. Inconsistencies between the map and reality must therefore be detected as fast as possible. Due to the reasons stated above, this is not achievable with conventional mapping vehicles. Car manufacturers like BMW have therefore started to collect anonymized floating car data (FCD) from their vehicles [2–5] which allows them to continuously monitor the quality of the HD map. Due to bandwidth limitations, only higher-level features such as lane markings and traffic sign positions are transmitted. Based on those features, the quality and up-to-dateness of a given map can be monitored. If a mismatch between the collected data and the HD map is detected, the affected area can be closed for automated driving functions and a mapping vehicle can be alerted to update the map.

In this paper, we present a change detection algorithm that can run in real time as part of a backend-based stream-processing pipeline that processes the FCD. It uses a particle filter based on odometry, GNSS and landmark detections to localize the vehicle within the digital map. While all particles together represent the probability distribution for the vehicle's position at a given time, each individual particle also serves as a hypothesis about the vehicle's position that is used to compute various metrics for how well the current sensor readings match the world model that is encoded in the HD map. The different metrics are evaluated by a number of weak classifiers that are used as input for a trained Adaboost classifier [6]. The achievable detection rate of a single vehicle is then compared to that of a simple crowd-based approach, where each vehicle has a vote of whether or not the current section of the road has changed.

For evaluation purposes, we employ the developed algorithm to detect changes caused by motorway construction sites. This is primarily due to the fact that information about motorway construction sites is publicly available in Germany [7, 8] which facilitates obtaining ground-truth information. Also, motorway construction sites are of particular interest for the deployment of automated driving, as motorways are the primary target of most early automated driving systems. Since most of those systems cannot handle construction sites yet, especially without a valid map, changes in their location must be detected quickly in order to close affected sections for automated driving functions.

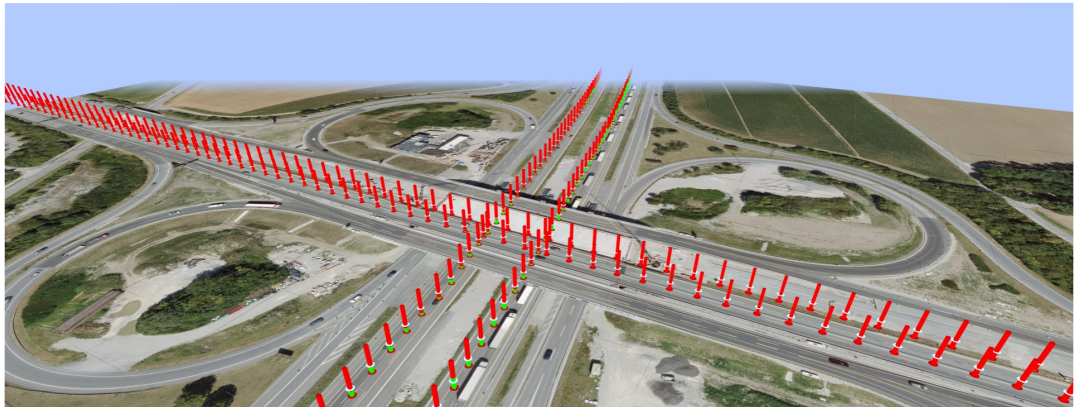
A. Related Work

Detecting changes in a given map is an important problem with regards to life-long mapping and thus has been subject to extensive research in the robotics community. The approaches to solve this problem can be divided into two major categories: handling dynamic changes and handling semi-static changes. Dynamic changes refer to occlusions by fast moving objects such as pedestrians that can be detected directly from measurements as the time scale of their move-

[1]D. Pannen and M. Liebner are with BMW Group, 80788 Munich, Germany. firstname.lastname@bmw.de
[2]W. Burgard is with the Department of Computer Science, University of Freiburg, Georges-Köhler-Allee 080, 79110 Freiburg i. Br., Germany. burgard@informatik.uni-freiburg.de

978-1-5386-6027-0/19/$31.00 ©2019 IEEE

HD Map Change Detection with a Boosted Particle Filter. David Pannen, Martin Liebner and Wolfram Burgard, ICRA 2019



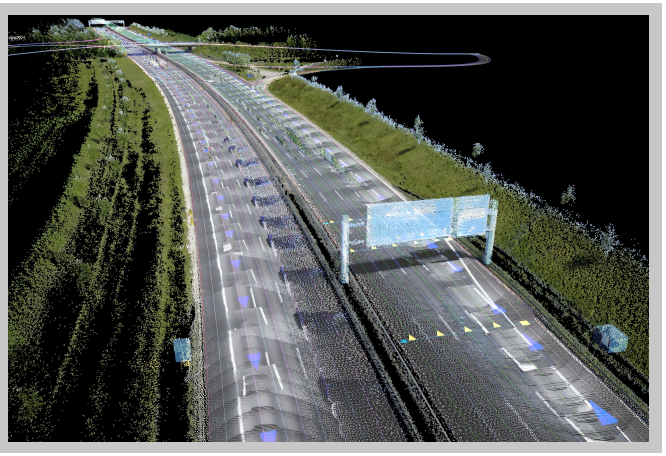
Creating HD Maps

Automated driving promises increased

- Safety
- Comfort
- Efficiency

HD maps enable

- Sensor augmentation
- Extended foresight
- Fallback for sensor failures

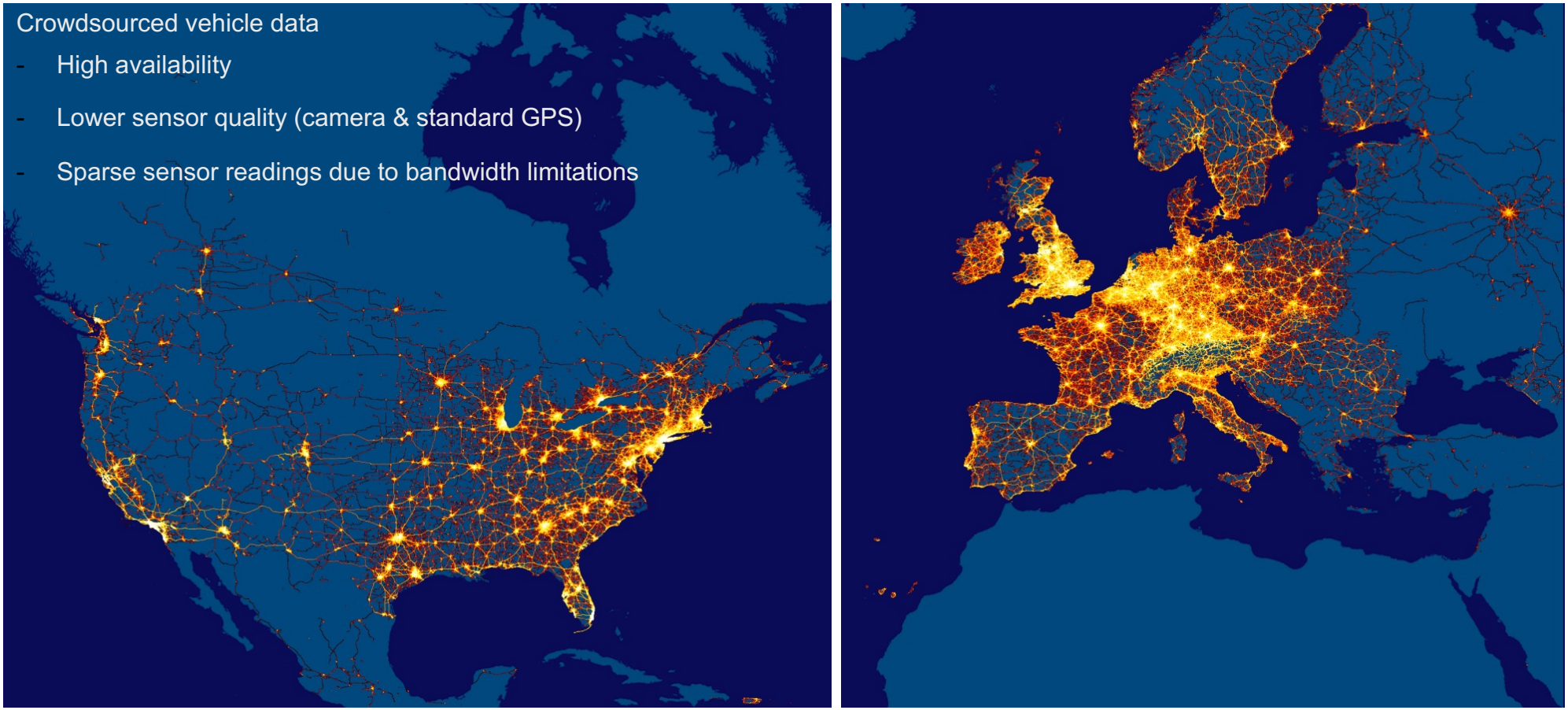


Environmental changes frequently occur and make change detection necessary

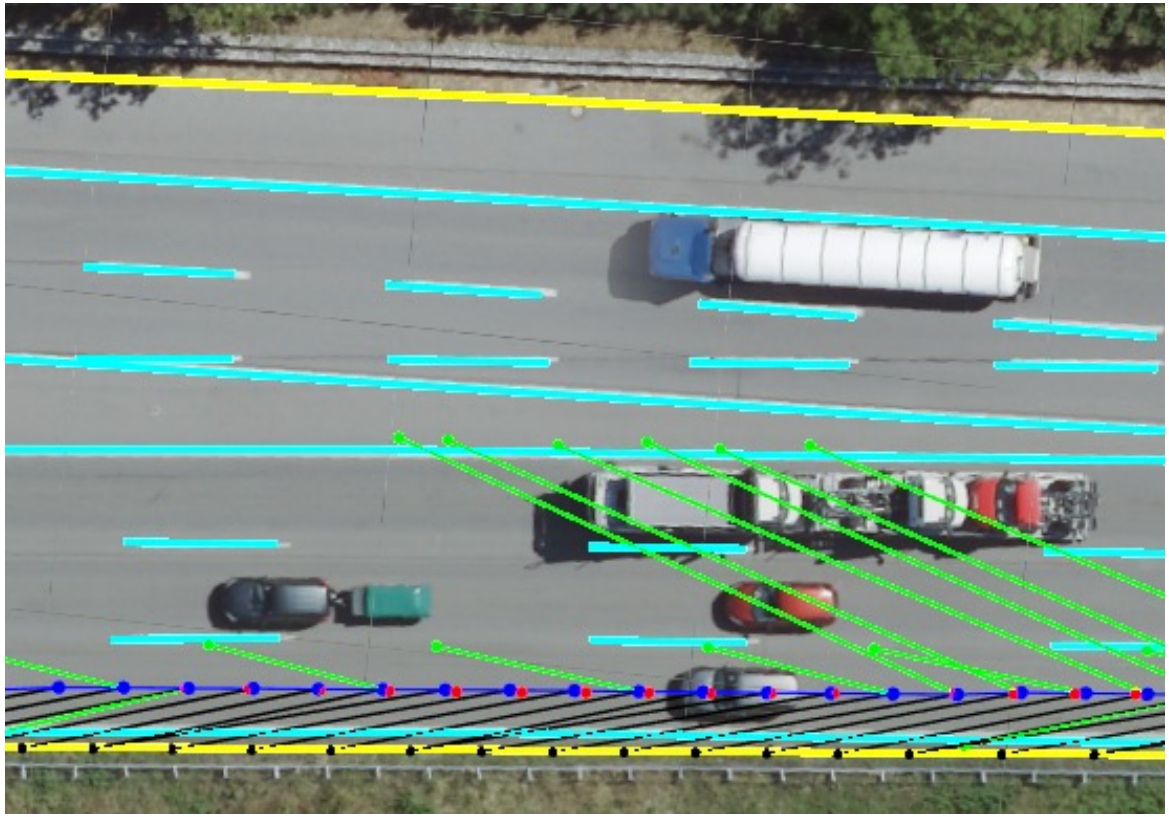
State of the art mapping vehicles

- Expensive sensors
- Small #vehicles
- Small frequency of visits

With Crowdsourced Vehicle Data



HD Map Representation & Measurements



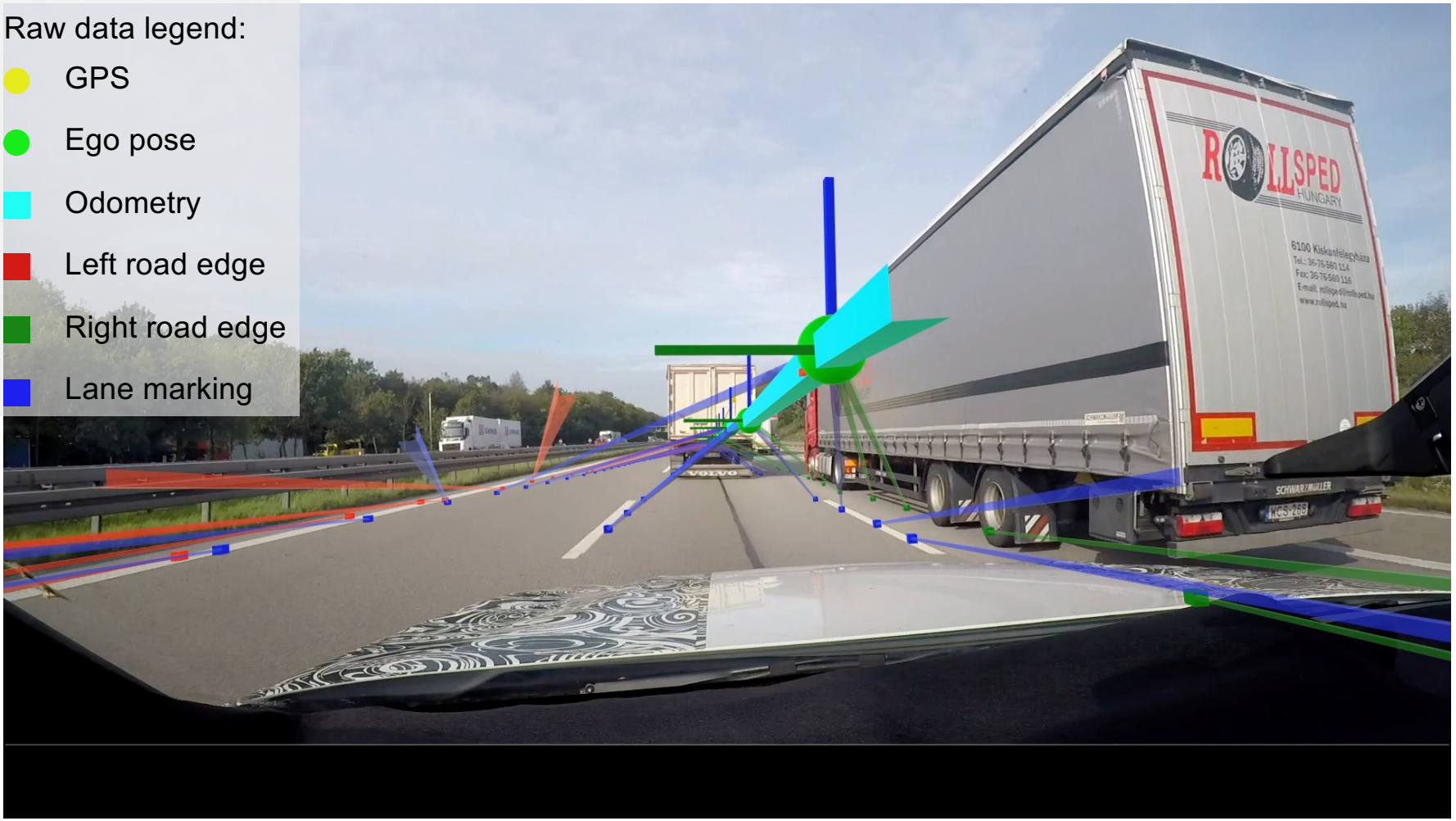
Crowdsourced data

- Vehicle position
- Odometry measurement
- GNSS measurement
- Lane marking observation
- Road edge observation

HD map

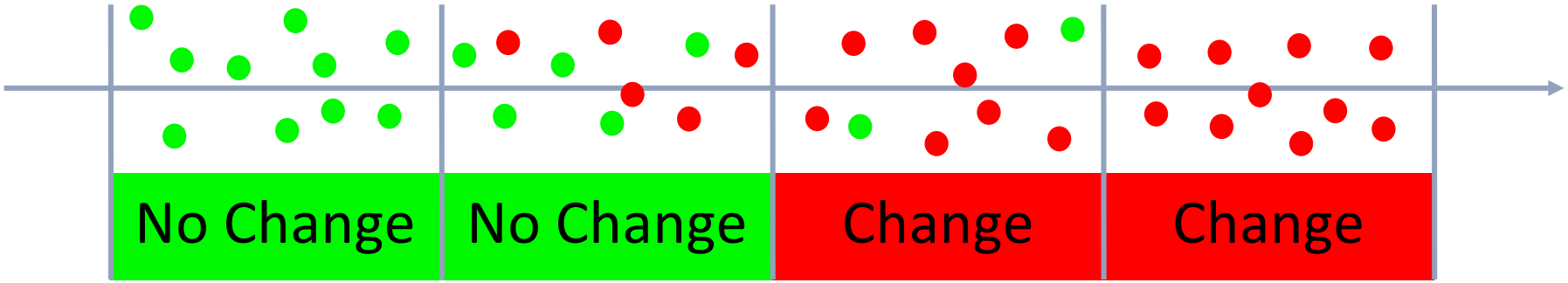
- Lane marking
- Road Edge

Features

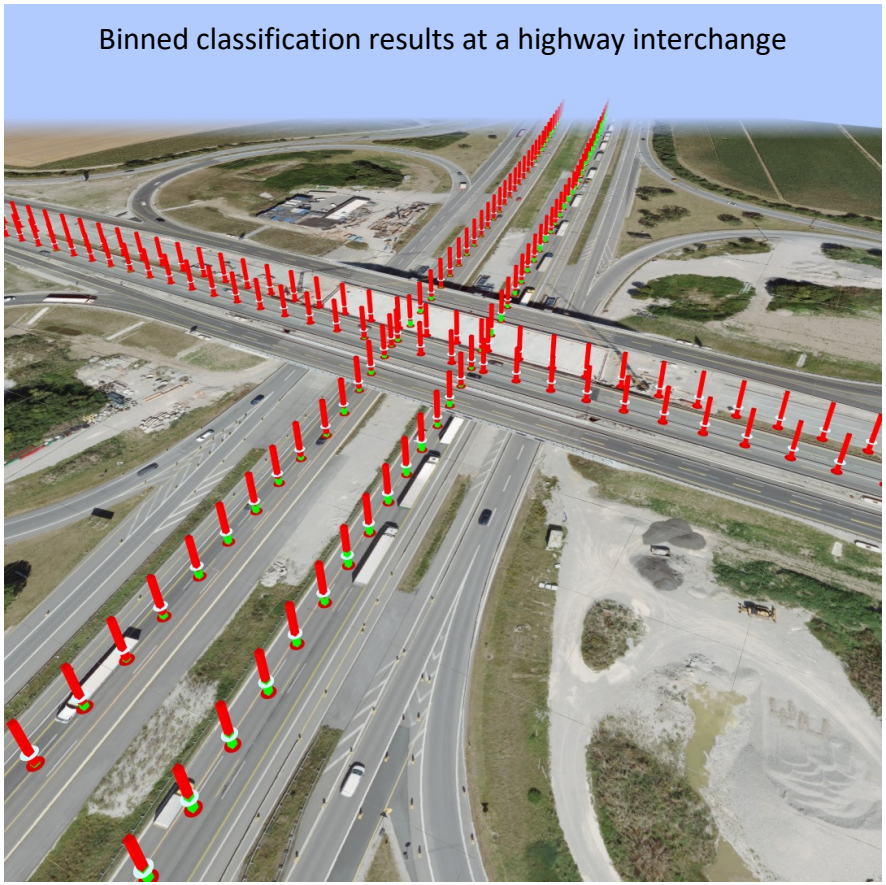

Raw data legend:
GPS
Ego pose
Odometry
Left road edge
Right road edge
Lane marking
ROLLSPED
HUNGARY

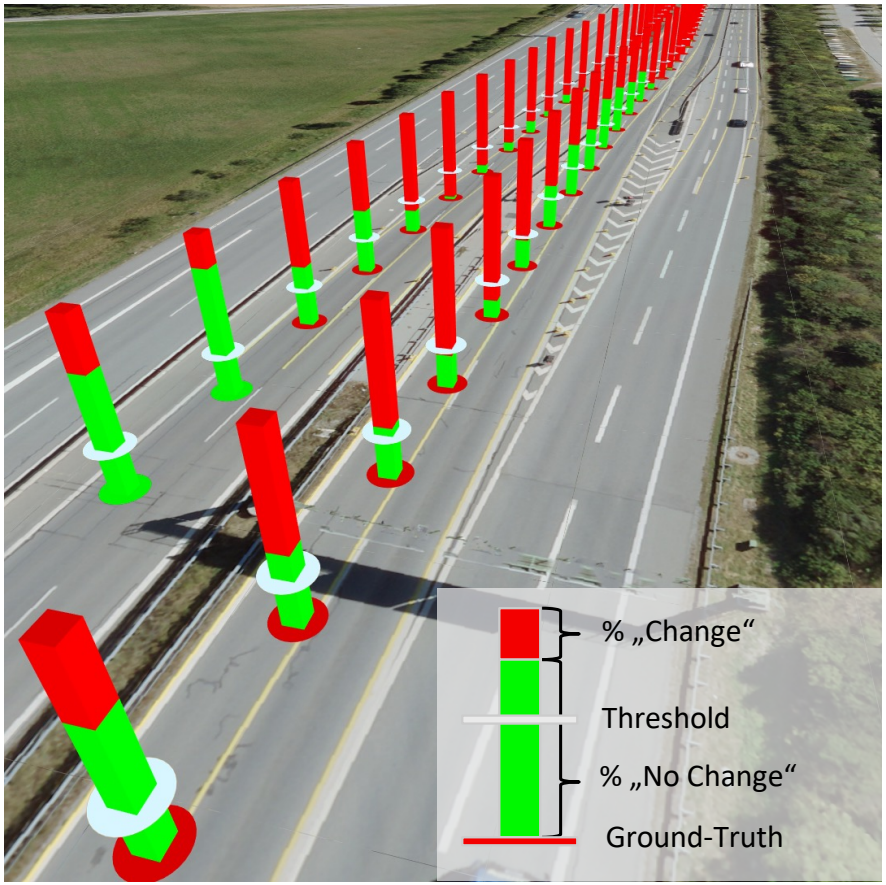
Particle Filters to Detect Innovation

- Odometry
- GNSS
- Lane markings
- Road edges
- Adaboost for classification
- Accumulation of classification results in bins along the road



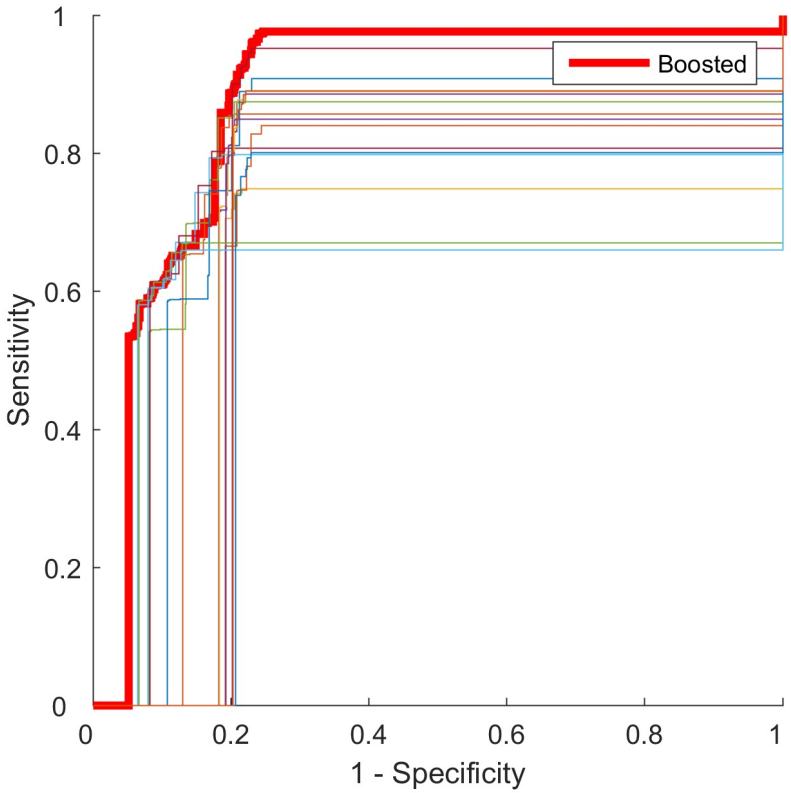
Qualitative Results



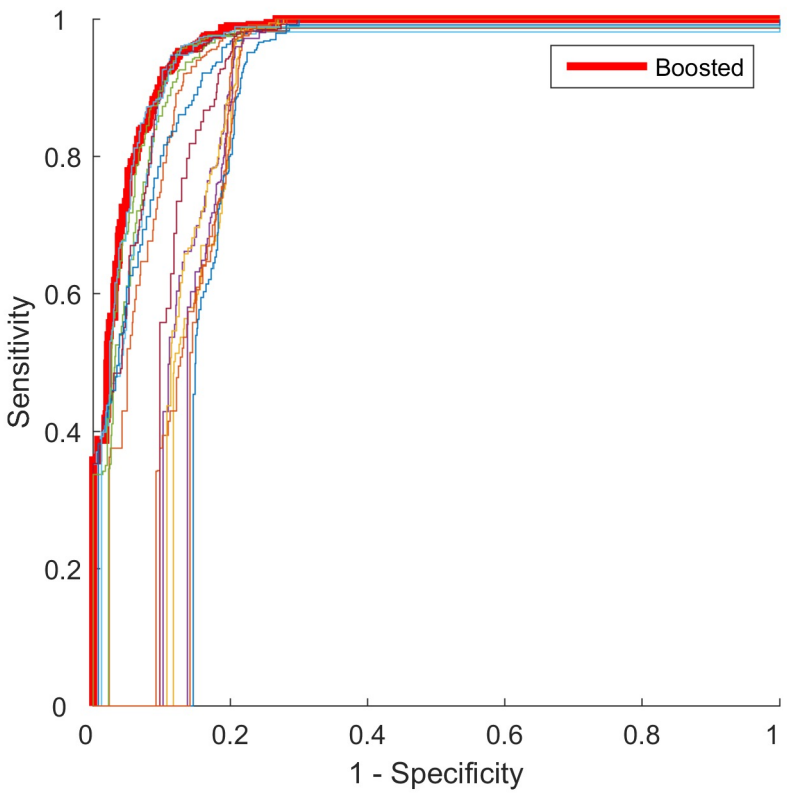


Results

ROC for single trace classification

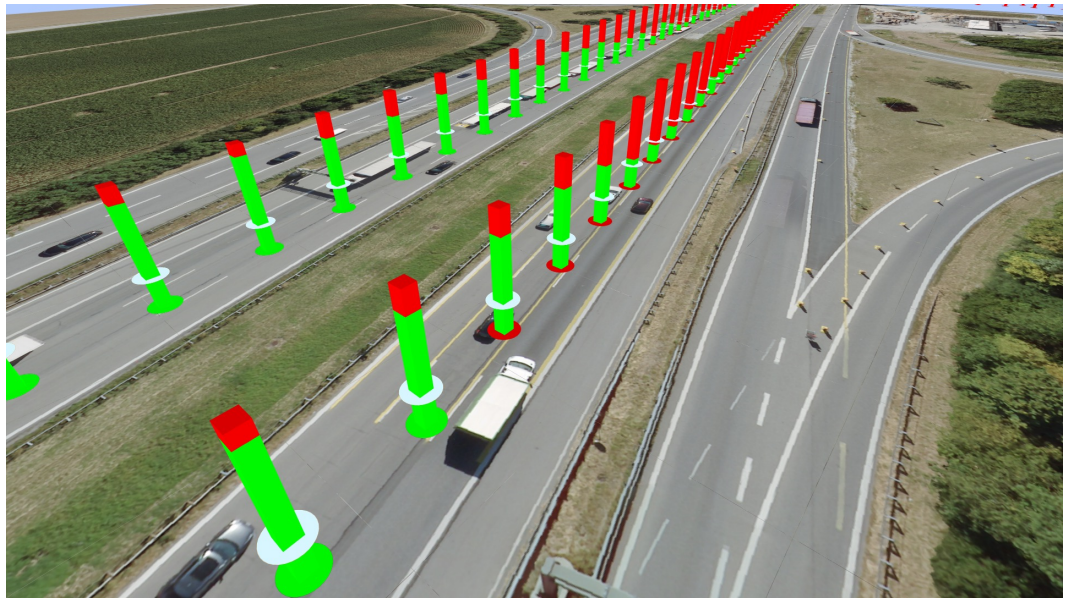


ROC for multi trace classification



Contribution

- New HD map change detection approach based on particle filter localization and boosting
- Using comparably sparse and noisy sensor data
- Crowdsourcing significantly improves single traversal classification with a sensitivity of 98.7% and specificity of 81.2%
- Evaluation showed successful detection of real-world changes in construction sites



Challenges for Learning Maps on the Fly

- Semantics
- Data association
  - Tracking
  - Traffic light associations
  - …
- Topology
  - Road structure
  - Intersection topology
  - …

Learning Semantics & Dynamics

EfficientLPS: Efficient LiDAR Panoptic Segmentation

Kshitij Sirohi*,1, Rohit Mohan*,1, Daniel Büscher[1], Wolfram Burgard[1,2], and Abhinav Valada[1]

***Abstract*—Panoptic segmentation of point clouds is a crucial task that enables autonomous vehicles to comprehend their vicinity using their highly accurate and reliable LiDAR sensors. Existing top-down approaches tackle this problem by either combining independent task-specific networks or translating methods from the image domain ignoring the intricacies of LiDAR data and thus often resulting in sub-optimal performance. In this paper, we present the novel top-down Efficient LiDAR Panoptic Segmentation (EfficientLPS) architecture that addresses multiple challenges in segmenting LiDAR point clouds including distance-dependent sparsity, severe occlusions, large scale-variations, and re-projection errors. EfficientLPS comprises of a novel shared backbone that encodes with strengthened geometric transformation modeling capacity and aggregates semantically rich range-aware multi-scale features. It incorporates new scale-invariant semantic and instance segmentation heads along with the panoptic fusion module which is supervised by our proposed panoptic periphery loss function. Additionally, we formulate a regularized pseudo labeling framework to further improve the performance of EfficientLPS by training on unlabelled data. We benchmark our proposed model on two large-scale LiDAR datasets: nuScenes, for which we also provide ground truth annotations, and SemanticKITTI. Notably, EfficientLPS sets the new state-of-the-art on both these datasets.**

***Index Terms*—Scene Understanding, Semantic Segmentation, Instance Segmentation, Panoptic Segmentation.**

I. Introduction

AUTONOMOUS vehicles are required to operate in challenging urban environments that consist of a wide variety of agents and objects, making comprehensive perception a critical task for robust and safe navigation. Typically, perception tasks are focused on independently reasoning about the semantics of the environment and recognition of object instances. Recently, panoptic segmentation [1] which unifies semantic and instance segmentation has emerged as a popular scene understanding problem that aims to provide a holistic solution. Panoptic segmentation simultaneously segments the scene into 'stuff' classes that comprise of background objects or amorphous regions such as road, vegetation, and buildings, as well as 'thing' classes that represent distinct foreground objects such as cars, cyclists, and pedestrians. Panoptic segmentation has been extensively studied in the image domain [1]–[4], facilitated by the ordered structure of images being supported by well-researched convolutional networks. However, only a handful of methods have been proposed for panoptic segmentation of LiDAR point clouds [5], [6]. LiDARs have become an indispensable sensor for autonomous vehicles due to their illumination independence and geometric description of the scene, making scene understanding using LiDAR point clouds an essential capability. However, the typical unordered, sparse, and irregular structure of point clouds pose several unique challenges.

To this end, deep learning methods that rely on grid based convolutions to address these challenges typically follow two different directions. They either project the point cloud into the 3D voxel space and employ 3D convolutions on them [7], [8], or they project the point cloud into the 2D space [6], [9], [10] and employ the well-researched 2D Convolutional Neural Networks (CNNs). While voxel-based method achieve high accuracy, they are computationally more expensive and require substantial memory to store the voxelized point clouds. The 2D projection based methods on the other hand, yield a more denser representation and require comparatively lesser computational resources, but they suffer from information loss during projection, blurry CNN outputs, and incorrect label assignment to the occluded points during re-projection. Therefore, there is a need to bridge this gap with a method that has the advantages of fast and memory-efficient 2D convolutions while mitigating the problems due to the projection.

In this work, we present the novel Efficient LiDAR Panoptic Segmentation (EfficientLPS) architecture that effectively addresses the aforementioned challenges by employing a 2D CNN for the task while explicitly utilizing the unique 3D information provided by point clouds. EfficientLPS consists of a shared backbone comprising our novel Proximity Convolution

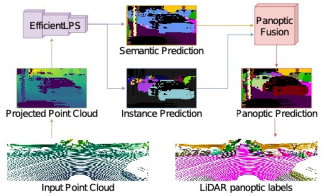
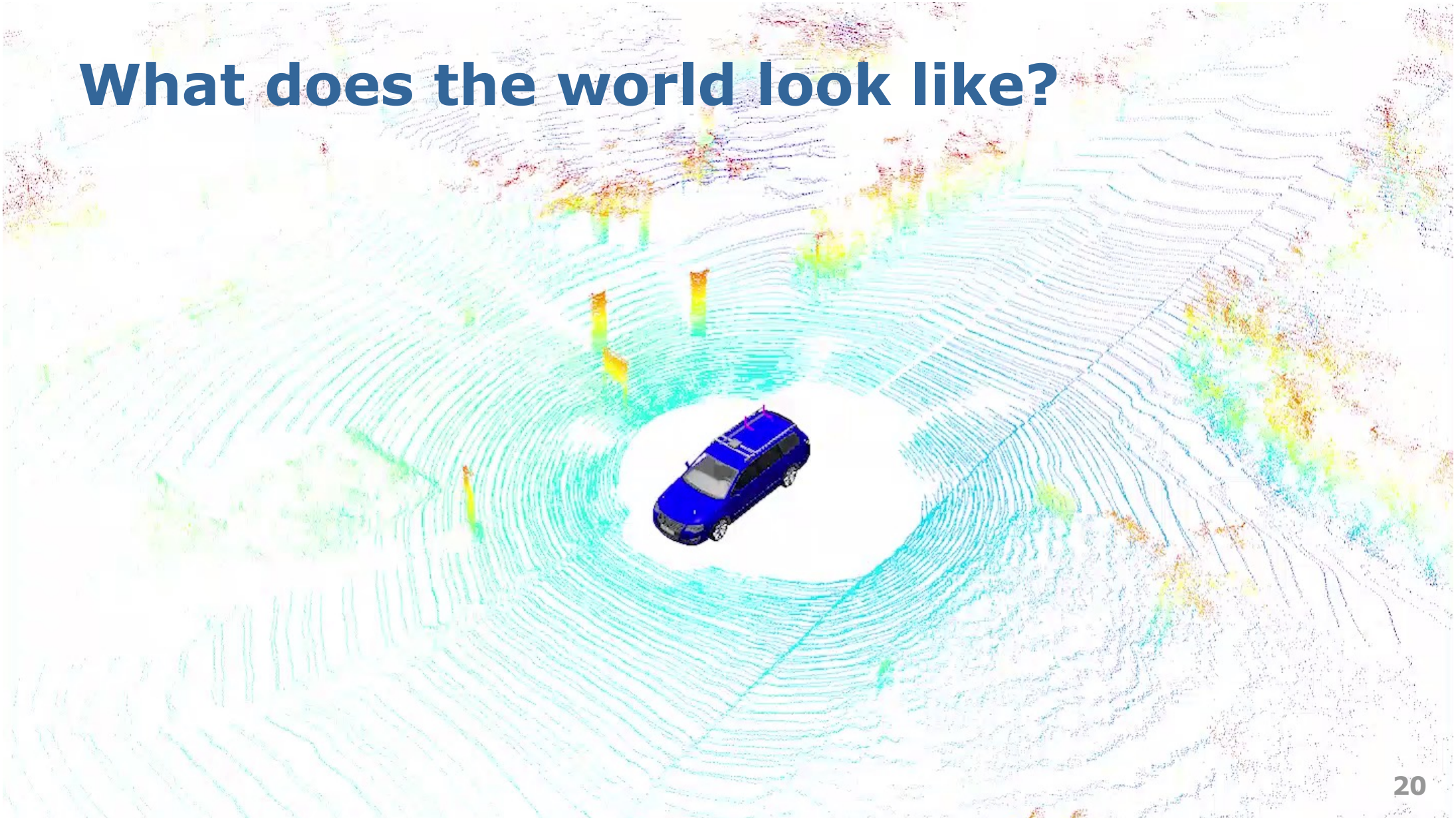


Fig. 1: Overview of the top-down EfficientLPS architecture that consists of a shared backbone to learn spatially-aware features from the projected point cloud and individual heads to learn semantic and instance specific features which are fused in the panoptic fusion module. The network explicitly utilizes the range information in the backbone, semantic head and fusion module to mitigate the problems due to the projection and distance-dependent sparsity of LiDAR point clouds.

*These authors contributed equally.
[1]Department of Computer Science, University of Freiburg, Germany
[2]Toyota Research Institute, Los Altos, USA.

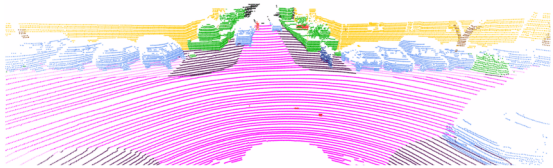
EfficientLPS: Efficient LiDAR Panoptic Segmentation. Kshitij Sirohi, Rohit Mohan, Daniel Büscher, Wolfram Burgard, Abhinav Valada

What does the world look like?

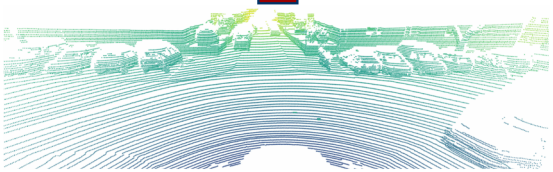


Semantic Tasks

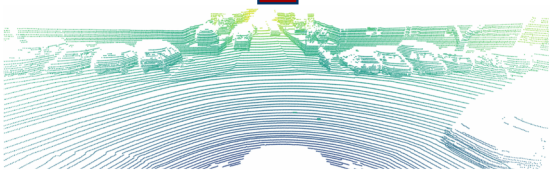
Semantic Segmentation



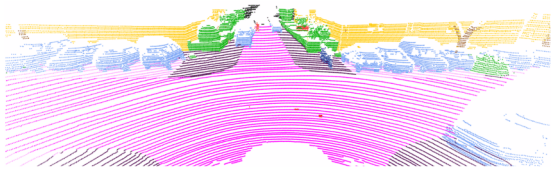
No reasoning about instances



Semantic Tasks

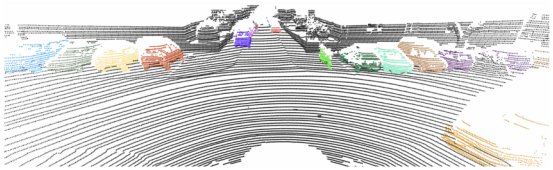


Semantic Segmentation



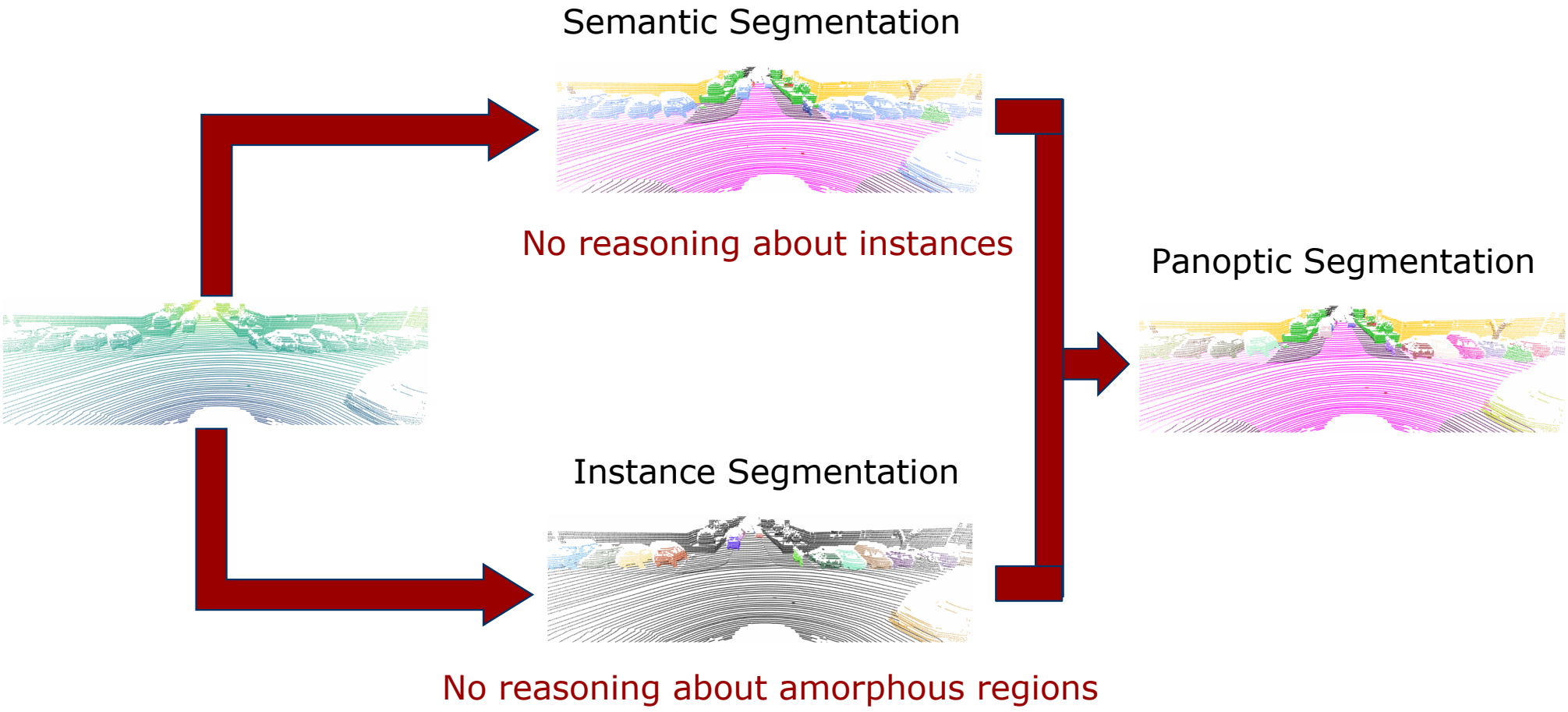
No reasoning about instances

Instance Segmentation



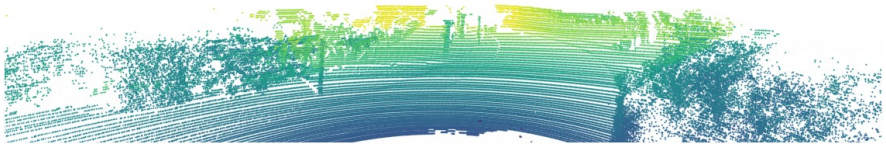
No reasoning about amorphous regions

Semantic Tasks



Challenges in using LiDAR data

- LiDARs provide an accurate illumination independent geometric description
- Unordered and irregular structure of point clouds
- Distance dependent sparsity
- Memory and time intensive 3D convolutions
- Information loss during point cloud projection into a 2D space



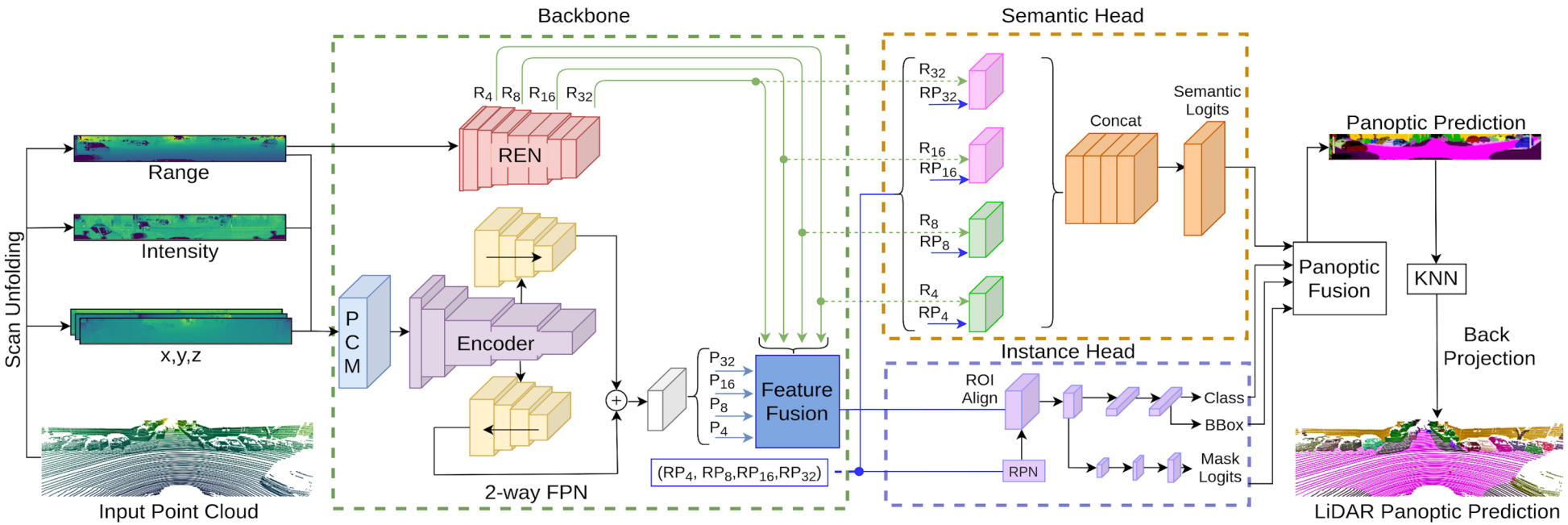
LiDAR Scan Projection

Spherical Projection

Scan Unfolding Projection

- Spherical projection (Milioto et al., 2019) suffers from systematic point occlusions
- Scan unfolding projection (Triess et al., 2020) reduces systematic occlusion artifacts
- Improves performance

EfficientLPS Architecture



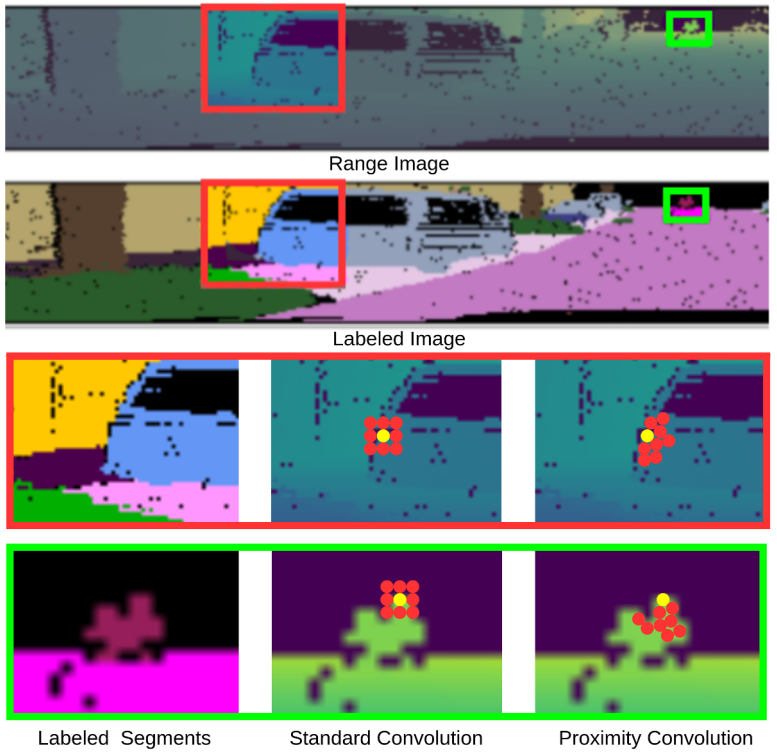
- Scan unfolding projection
- Backbone: PCM + Encoder + REN + 2-way FPN
- Semantic Head, Instance Head, Panoptic Fusion Module
- Reprojection into 3D using kNNs

Proximity Convolution Module

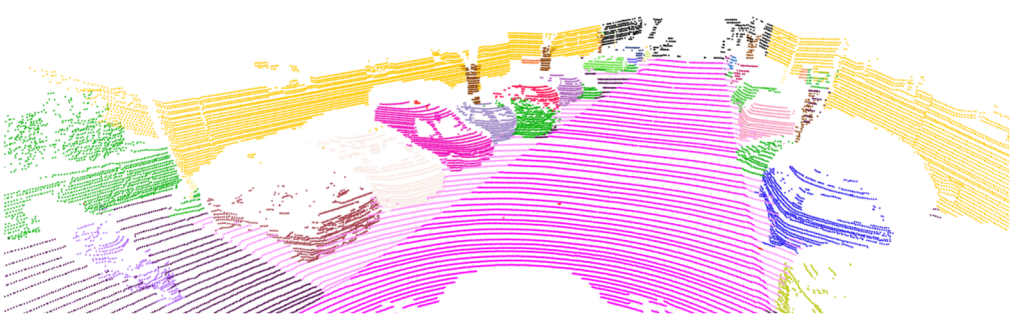
- Exploits range information to augment spatial sampling locations
- Enhances the geometric modeling capabilities of convolutions
- Effectively captures contextual information from sparse points

Standard Convolution 3x3

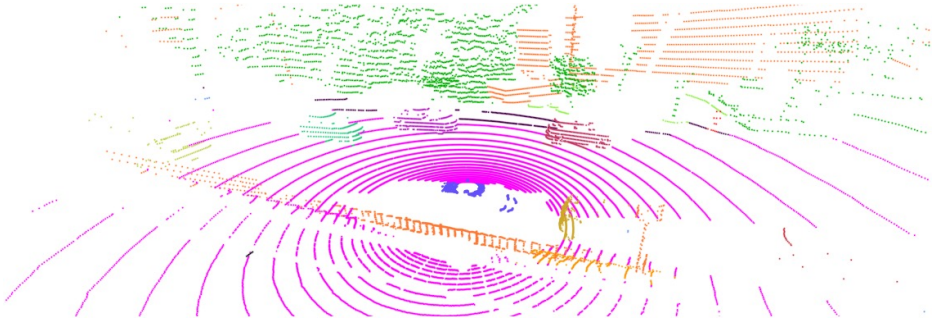
3x3 Convolution on range filtered neighbors



Evaluation Datasets



SemanticKITTI



nuScenes

Benchmarking Results

- Ranked #1 on SemanticKITTI and nuScenes datasets for panoptic segmentation
- Novel regularized pseudo labeling framework further improves the performance

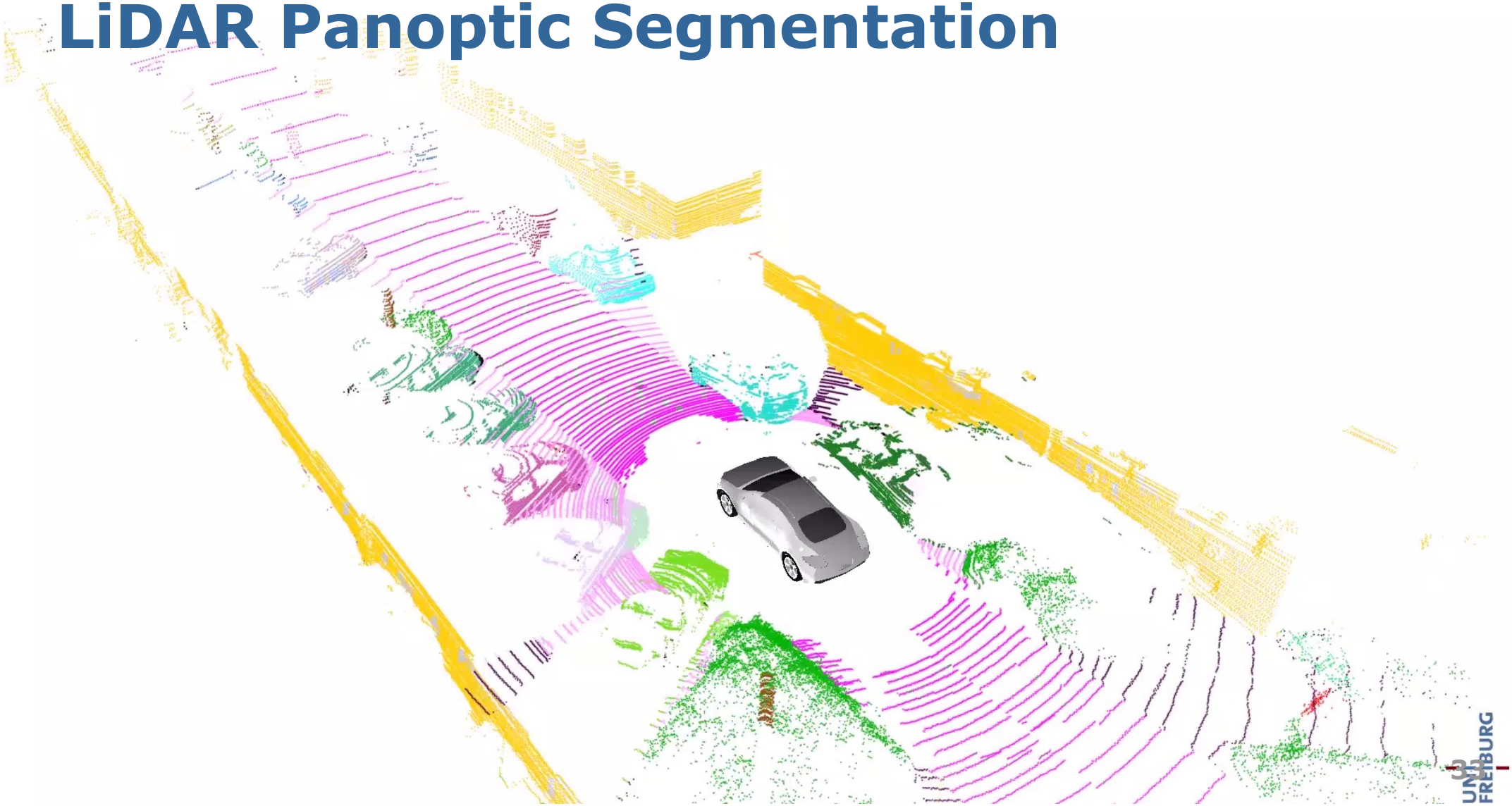
SemanticKITTI Leaderboard

Method	PQ
RangeNet++ [23] + PointPillars [43]	37.1
KPConv [21] + PointPillars [43]	44.5
LPSAD [6]	38.0
PanopticTrackNet [10]	43.1
Panoster [5]	52.7
EfficientLPS (ours)	**57.4**

nuScenes Validation Set

Method	PQ
RangeNet++ [23] + Mask R-CNN [38]	45.2
PanopticTrackNet [10]	50.0
KPConv [21] + Mask R-CNN [38]	50.1
EfficientLPS (ours)	**59.2**

LiDAR Panoptic Segmentation



How Do We Go Beyond?

Panoptic Segmentation

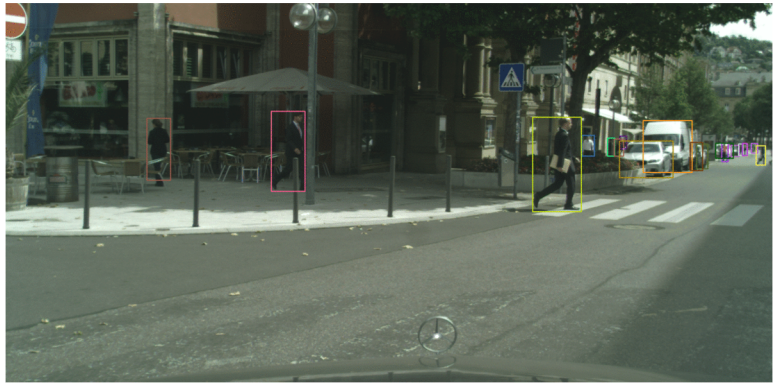
Frame-by frame, inconsistent over time

How Do We Go Beyond?

Panoptic Segmentation

Frame-by frame, inconsistent over time

Multi-Object Tracking

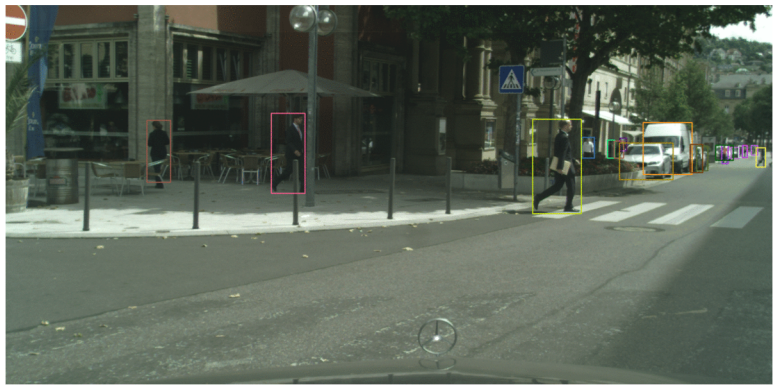


Bbox tracking performance is saturating

How Do We Go Beyond?

Panoptic Segmentation

Multi-Object Tracking



Multi-Object Panoptic Tracking

Instance IDs are temporally associated

Multi-Object Panoptic Tracking (MOPT)

MOPT: Multi-Object Panoptic Tracking

Juana Valeria Hurtado Rohit Mohan Wolfram Burgard Abhinav Valada
University of Freiburg

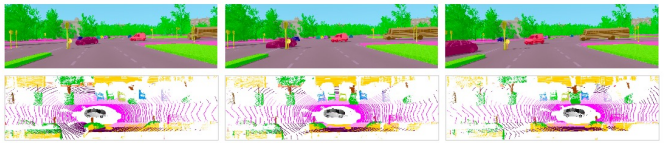


Figure 1: MOPT output overlaid on the image\LiDAR input from the Virtual KITTI 2 (top row) and SemanticKITTI (bottom row) datasets. MOPTS unifies semantic segmentation, instance segmentation and multi-object tracking to yield segmentation of stuff classes and segmentation of thing classes with temporally consistent instance IDs. Observe that the tracked thing instances retain their color-code temporally in subsequent timesteps.

Abstract

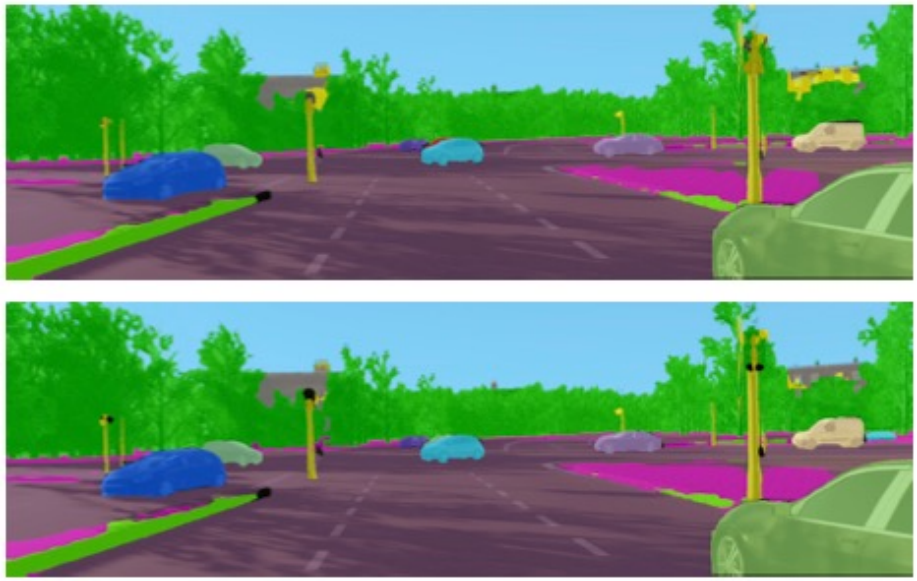
Comprehensive understanding of dynamic scenes is a critical prerequisite for intelligent robots to autonomously operate in their environment. Research in this domain, which encompasses diverse perception problems, has primarily been focused on addressing specific tasks individually rather than modeling the ability to understand dynamic scenes holistically. In this paper, we introduce a novel perception task denoted as multi-object panoptic tracking (MOPT), which unifies the conventionally disjoint tasks of semantic segmentation, instance segmentation, and multi-object tracking. MOPT allows for exploiting pixel-level semantic information of thing and stuff classes, temporal coherence, and pixel-level associations over time, for the mutual benefit of each of the individual sub-problems. To facilitate quantitative evaluations of MOPT in a unified manner, we propose the soft panoptic tracking quality (sPTQ) metric. As a first step towards addressing this task, we propose the novel PanopticTrackNet architecture that builds upon the state-of-the-art top-down panoptic segmentation network EfficientPS by adding a new tracking head to simultaneously learn all sub-tasks in an end-to-end manner. Additionally, we present several strong baselines that combine predictions from state-of-the-art panoptic segmentation and multi-object tracking models for comparison. We present extensive quantitative and qualitative evaluations of both vision-based and LiDAR-based MOPT that demonstrate encouraging results.

1. Introduction

Comprehensive scene understanding is a critical challenge that requires tackling multiple tasks simultaneously to detect, localize, and identify the scene elements as well as to understand the occurring context, dynamics, and relationships. These fundamental scene comprehension tasks are crucial enablers of several diverse applications [21, 26, 22, 25, 30] including autonomous driving, robot navigation, augmented reality and remote sensing. Typically, these problems have been addressed by solving distinct perception tasks individually, i.e., image\pointcloud recognition, object detection and classification, semantic segmentation, instance segmentation, and tracking. The state of the art in these tasks have been significantly advanced since the advent of deep learning approaches, however their performance is no longer increasing at the same groundbreaking pace [18]. Moreover, as most of these tasks are required to be performed simultaneously in real-world applications, the scalability of employing several individual models is becoming a limiting factor. In order to mitigate this emerging problem, recent works [8, 27, 28, 17] have made efforts to exploit common characteristics of some of these tasks by jointly modeling them in a coherent manner.

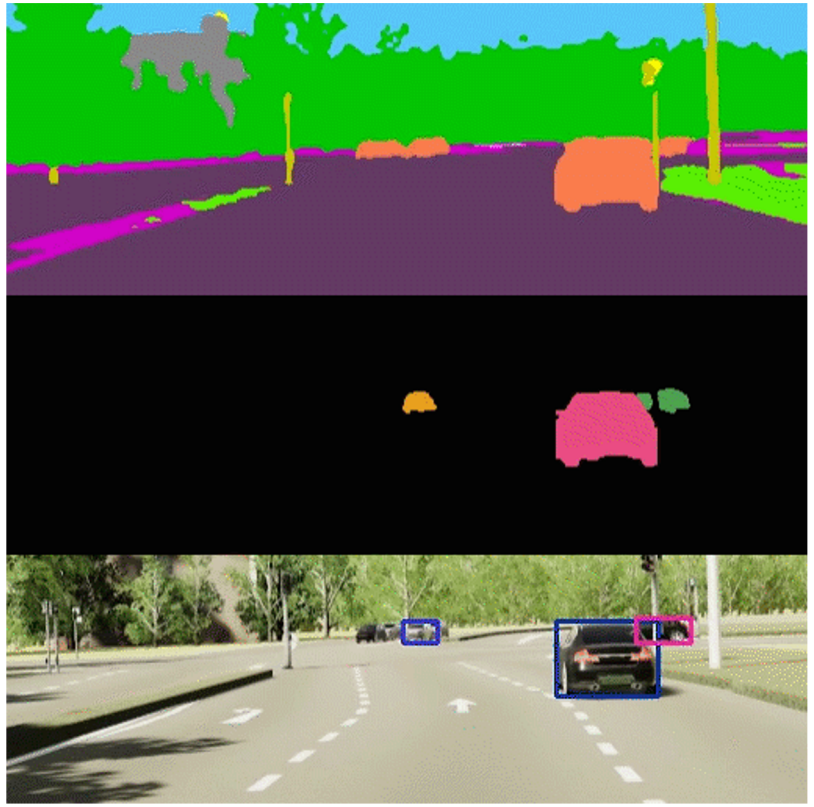
Two such complementary tasks have gained a substantial amount of interest in the last few years due to the availability of public datasets [1, 3] and widely adopted benchmarks [28, 1]. The first of which is panoptic segmentation [8] that unifies semantic segmentation of stuff classes which consist of amorphous regions and instance segmentation of thing classes which consist of countable objects. While the second task is Multi-Object Tracking and Segmentation (MOTS) [28] which extends multi-object tracking to the pixel level by unifying with instance segmentation of thing classes. Since the introduction of these tasks, considerable advances have been made in both panoptic segmentation [7, 31, 15, 14] and MOTS [16, 12, 11, 29] which

MOPT: Multi-Object Panoptic Tracking.
Juana Valeria Hurtado, Rohit Mohan,
Wolfram Burgard, Abhinav Valada

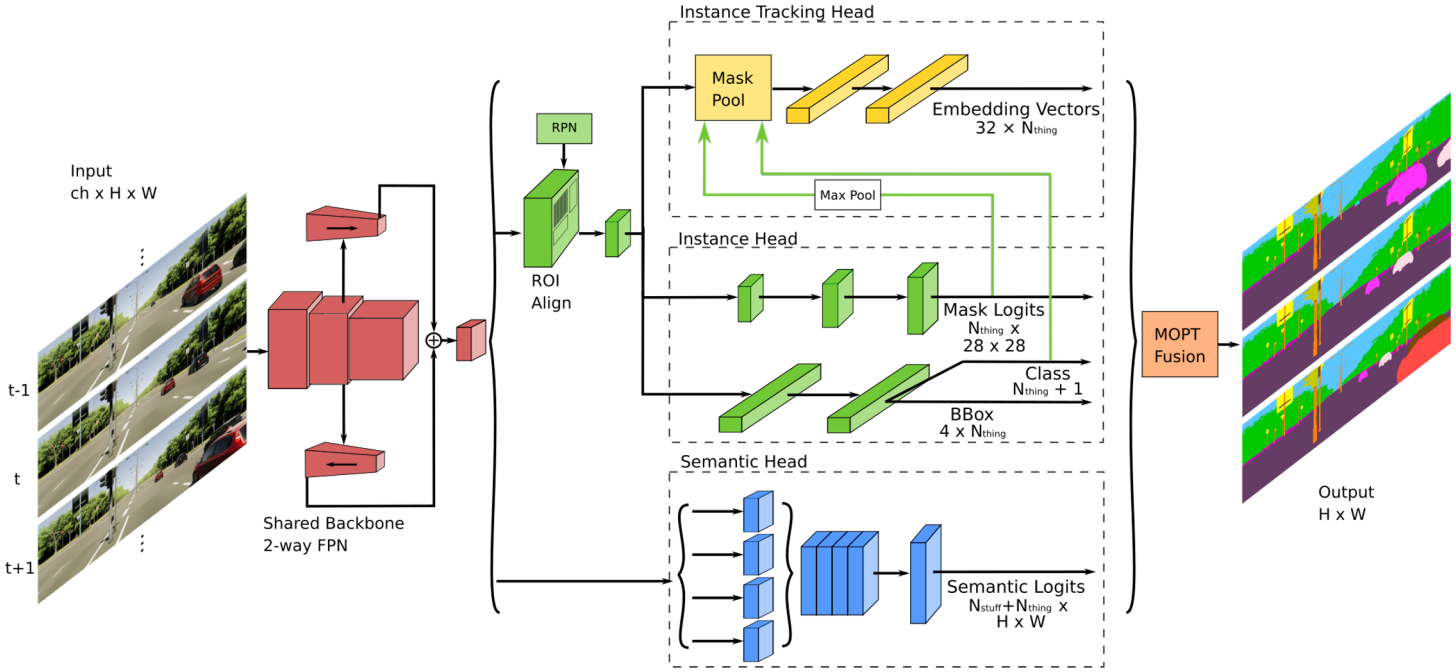


MOPT: Multi-Object Panoptic Tracking

- Encourage holistic modeling of dynamic scenes
- Unifies semantic segmentation, instance segmentation, multi-object tracking
- Exploit temporal consistency for learning better semantics
- Efficient and scalable model for dynamic scene understanding

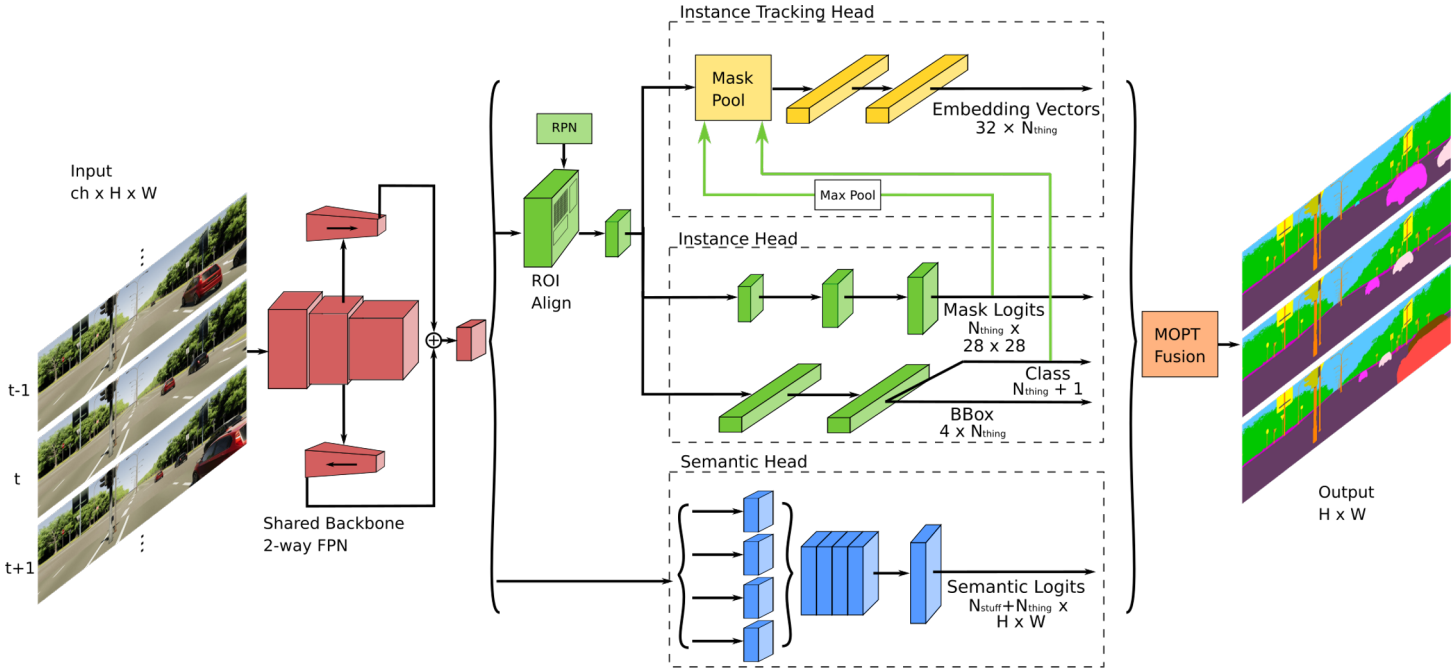


PanopticTrackNet Architecture



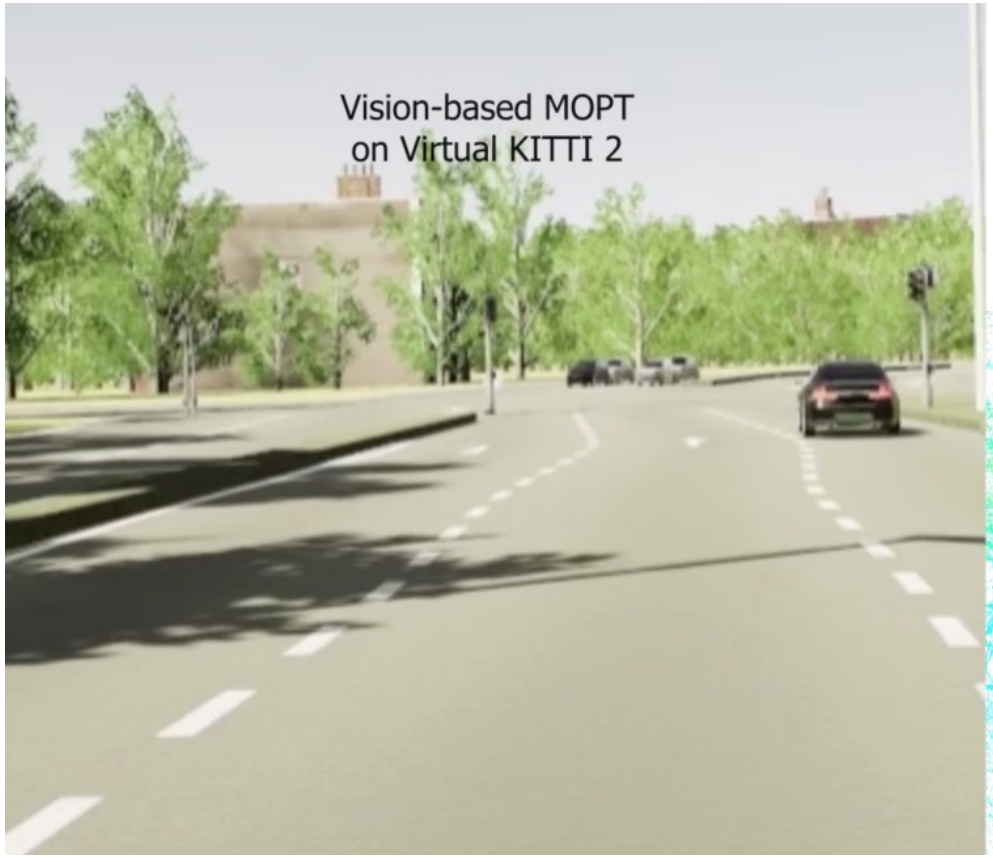
- Built upon our EfficientPS architecture
- Three task specific heads for semantic segmentation, instance segmentation, and multi-object tracking

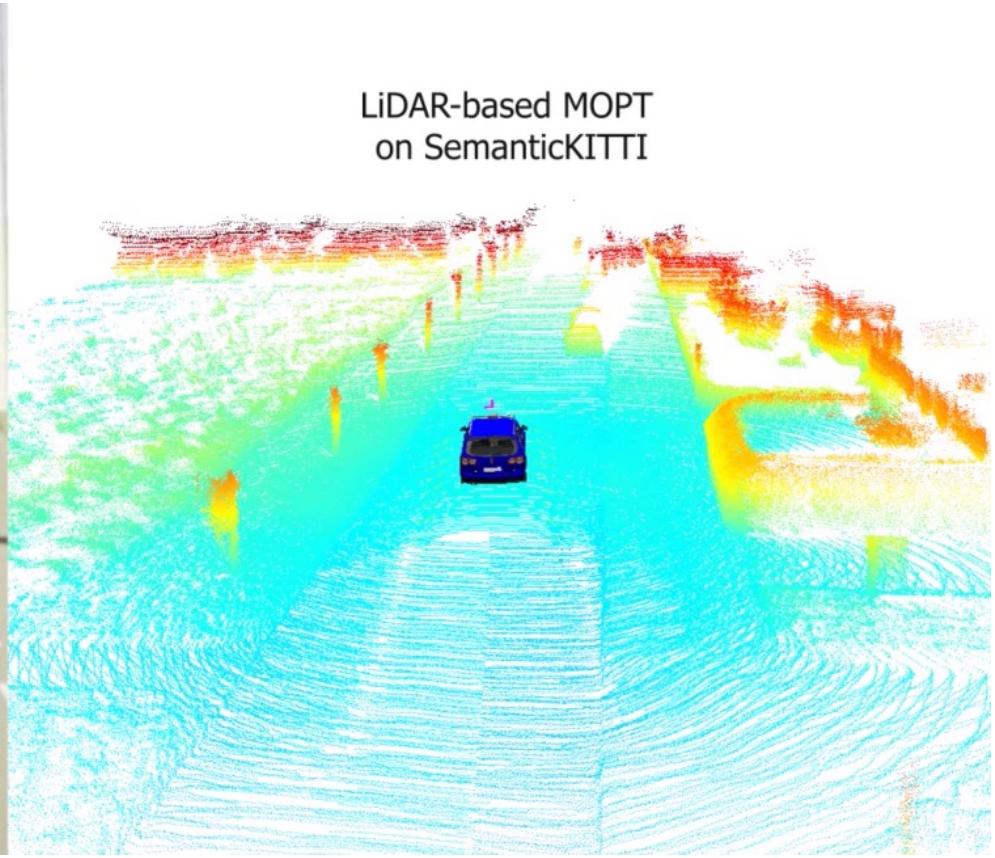
PanopticTrackNet Architecture



We adaptively fuse task-specific outputs in our fusion module to yield the panoptic output with temporally tracked instances

Evaluation Datasets





Results: Vision-Based MOPT

Network	sPTQ (%)	PTQ (%)	Params. (M)	FLOPs (B)	Time (ms)
Seamless [1] + Track R-CNN [2]	45.66	45.28	79.91	273.96	115
EfficientPS [3] + Track R-CNN [2]	46.68	46.3	69.91	232.80	115
EfficientPS [3] + MaskTrack R-CNN [4]	46.17	45.99	120.57	224.43	117
PanopticTrackNet (ours)	**47.27**	**46.67**	**45.08**	**167.40**	**114**

Results: LiDAR-Based MOPT

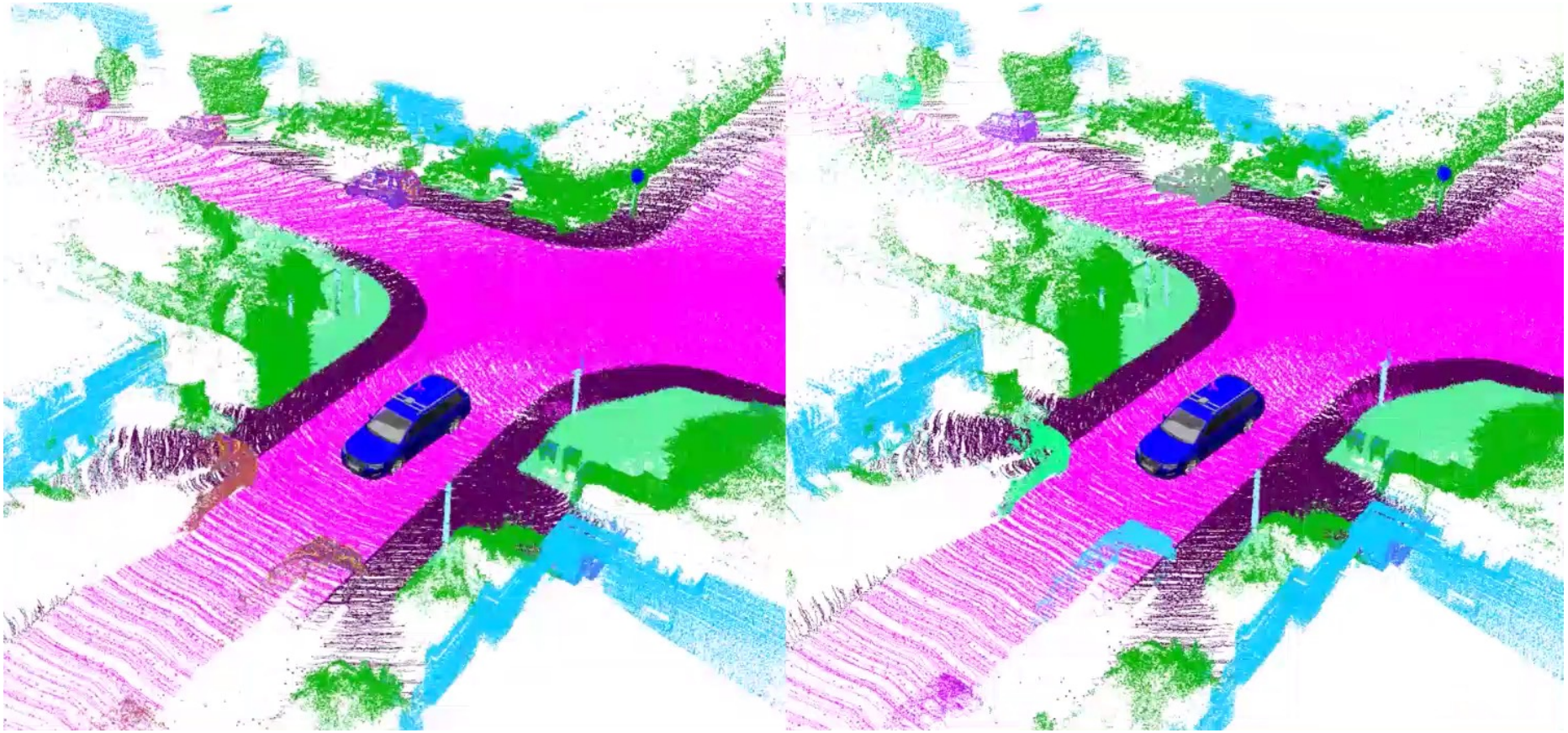
Network	sPTQ (%)	PTQ (%)	Params. (M)	FLOPs (B)	Time (ms)
RangeNet++ [1] + PointPillars [2] + Track R-CNN [3]	42.22	41.94	110.74	695.51	409
KPConv [4] + PointPillars [2] + Track R-CNN [3]	46.04	45.50	89.96	438.34	514
EfficientPS [5] + Track R-CNN [3]	44.50	43.96	79.02	379.73	148
EfficientPS [5] + MaskTrack R-CNN [6]	44.03	43.72	120.64	445.90	151
PanopticTrackNet (ours)	**48.23**	**47.89**	**45.13**	**300.81**	**146**

Qualitative Results: Vision-Based MOPT

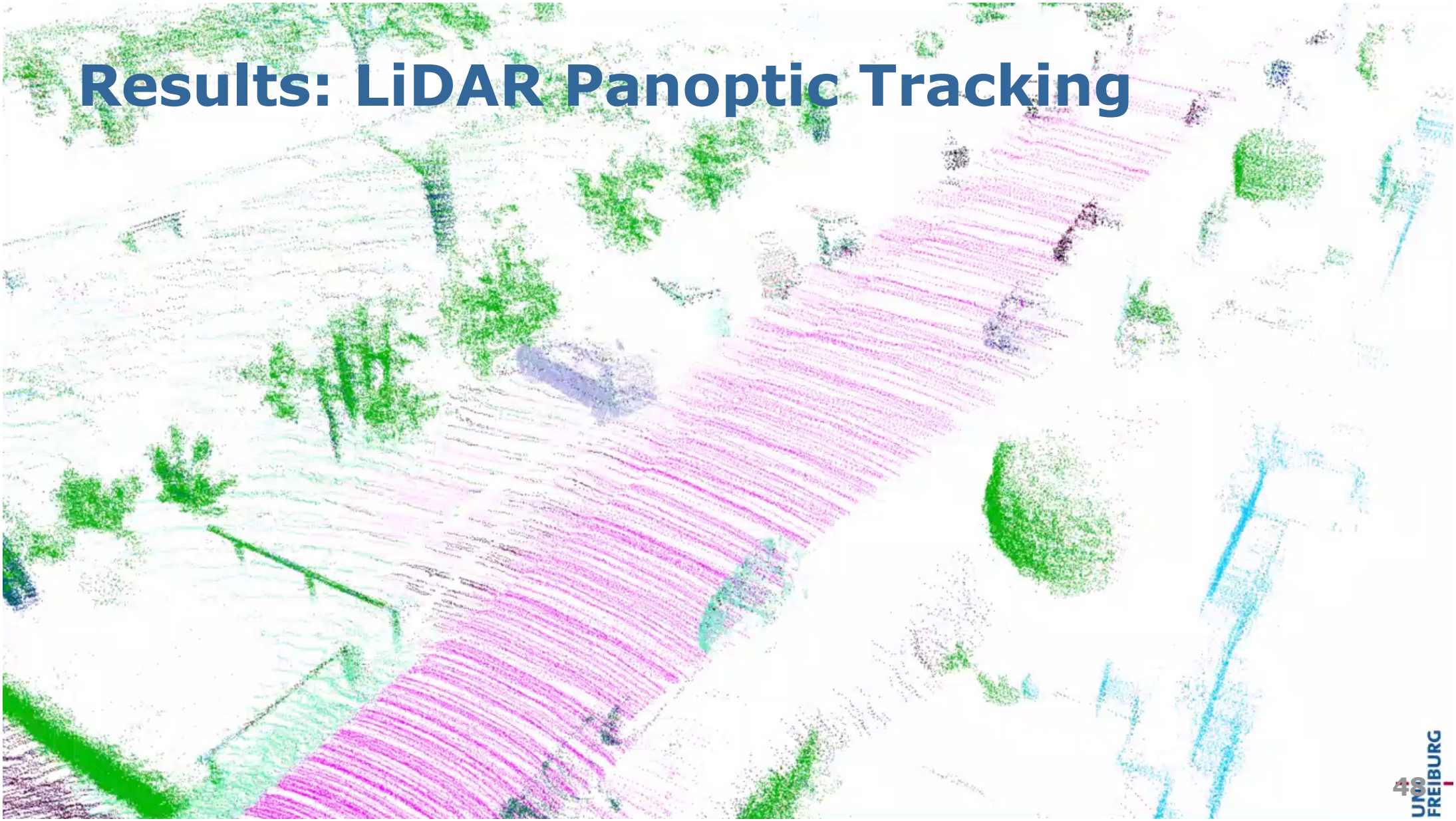
Results: LiDAR Panoptic Tracking

Panoptic Segmentation

MOPT



Results: LiDAR Panoptic Tracking



Robust Semantic Segmentation by Domain Adaption

HeatNet: Bridging the Day-Night Domain Gap in Semantic Segmentation with Thermal Images

Johan Vertens*, Jannik Zürn*, and Wolfram Burgard

***Abstract*— The majority of learning-based semantic segmentation methods are optimized for daytime scenarios and favorable lighting conditions. Real-world driving scenarios, however, entail adverse environmental conditions such as nighttime illumination or glare which remain a challenge for existing approaches. In this work, we propose a multimodal semantic segmentation model that can be applied during daytime and nighttime. To this end, besides RGB images, we leverage thermal images, making our network significantly more robust. We avoid the expensive annotation of nighttime images by leveraging an existing daytime RGB-dataset and propose a teacher-student training approach that transfers the dataset's knowledge to the nighttime domain. We further employ a domain adaptation method to align the learned feature spaces across the domains and propose a novel two-stage training scheme. Furthermore, due to a lack of thermal data for autonomous driving, we present a new dataset comprising over 20,000 time-synchronized and aligned RGB-thermal image pairs. In this context, we also present a novel target-less calibration method that allows for automatic robust extrinsic and intrinsic thermal camera calibration. Among others, we employ our new dataset to show state-of-the-art results for nighttime semantic segmentation.**

Fig. 1. Our multimodal segmentation network leverages both nighttime and daytime images. We transfer relevant knowledge from a large-scale unimodal daytime dataset for semantic segmentation with a teacher model to our multimodal HeatNet and simultaneously adapt our model to the nighttime domain by unsupervised domain adaptation.

I. INTRODUCTION

Robust and accurate semantic segmentation of urban scenes is one of the enabling technologies for autonomous driving in complex and cluttered driving scenarios. Recent years have shown great progress in RGB image segmentation for autonomous driving [36], [5], which were predominantly demonstrated in favorable daytime illumination conditions. While the reported results demonstrate high accuracies on benchmark datasets [5], [18], these models tend to generalize poorly to adverse weather conditions and low illumination levels present at nighttime. This constraint becomes especially apparent in rural areas where artificial lighting is weak or scarce. In autonomous driving, to ensure safety and situation awareness, robust perception in these conditions is a vital prerequisite.

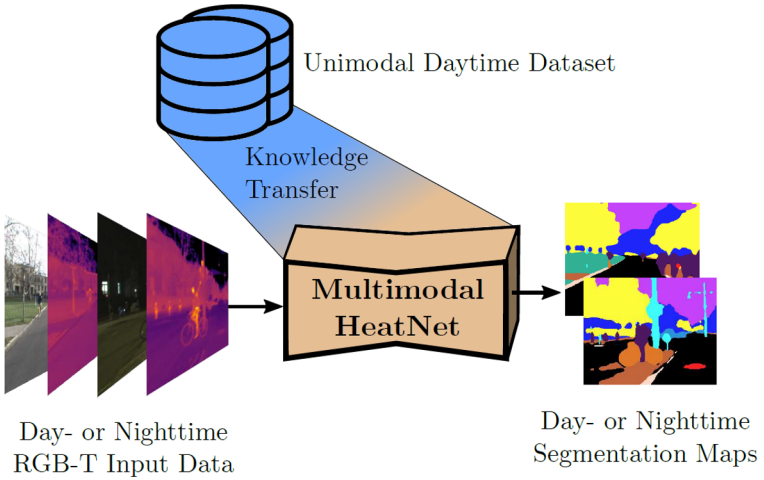
Transfer learning and domain adaptation approaches aim at narrowing the domain gap between a source domain, where supervised learning from labelled data is possible, to a target domain, where labelled data is either sparse or not available. Such approaches, as demonstrated in [28] or [35], allow to adapt a given segmentation model to a different domain. These approaches, however, do not leverage a complementary modality such as thermal infrared images that can contain more relevant information to solve a given task in certain environmental conditions than a single modality would provide.

In order to perform similarly well in challenging illumination conditions, it is beneficial for autonomous vehicles to leverage modalities complementary to RGB images [29], [30]. Encouraged by prior work in thermal image processing for object detection [31], object tracking [14], and semantic segmentation [9], [25], we investigate leveraging thermal images for nighttime semantic segmentation of urban scenes. Thermal images contain accurate thermal radiation measurements with a high spatial density. Furthermore, thermal radiation is much less influenced by sunlight illumination changes and is less sensitive to adversary conditions. Existing RGB-thermal datasets for semantic image segmentation such as [9] are not as large-scale as their RGB-only counterparts. Thus, models trained on such datasets generalize poorly to challenging real-world scenarios.

In contrast to previous works, we utilize a semantic segmentation network for RGB daytime images as a teacher model to provide labels for the RGB daytime images in our dataset. We project the thermal images into the viewpoint of the RGB camera images using extrinsic and intrinsic camera parameters that we determine using our novel target-less camera calibration approach. Afterwards, we can reuse labels from this teacher model to train a multimodal semantic segmentation network on our daytime RGB-thermal image pairs. In order to encourage day-night invariant segmentation of scenes, we simultaneously train a feature discriminator

*These authors contributed equally. All authors are with the University of Freiburg, Germany. Wolfram Burgard is also with the Toyota Research Institute, Los Altos, USA. Corresponding author: vertensj@informatik.uni-freiburg.de

HeatNet: Bridging the Day-Night Domain Gap in Semantic Segmentation with Thermal Images. Johan Vertens, Jannik Zürn and Wolfram Burgard, IROS 2020



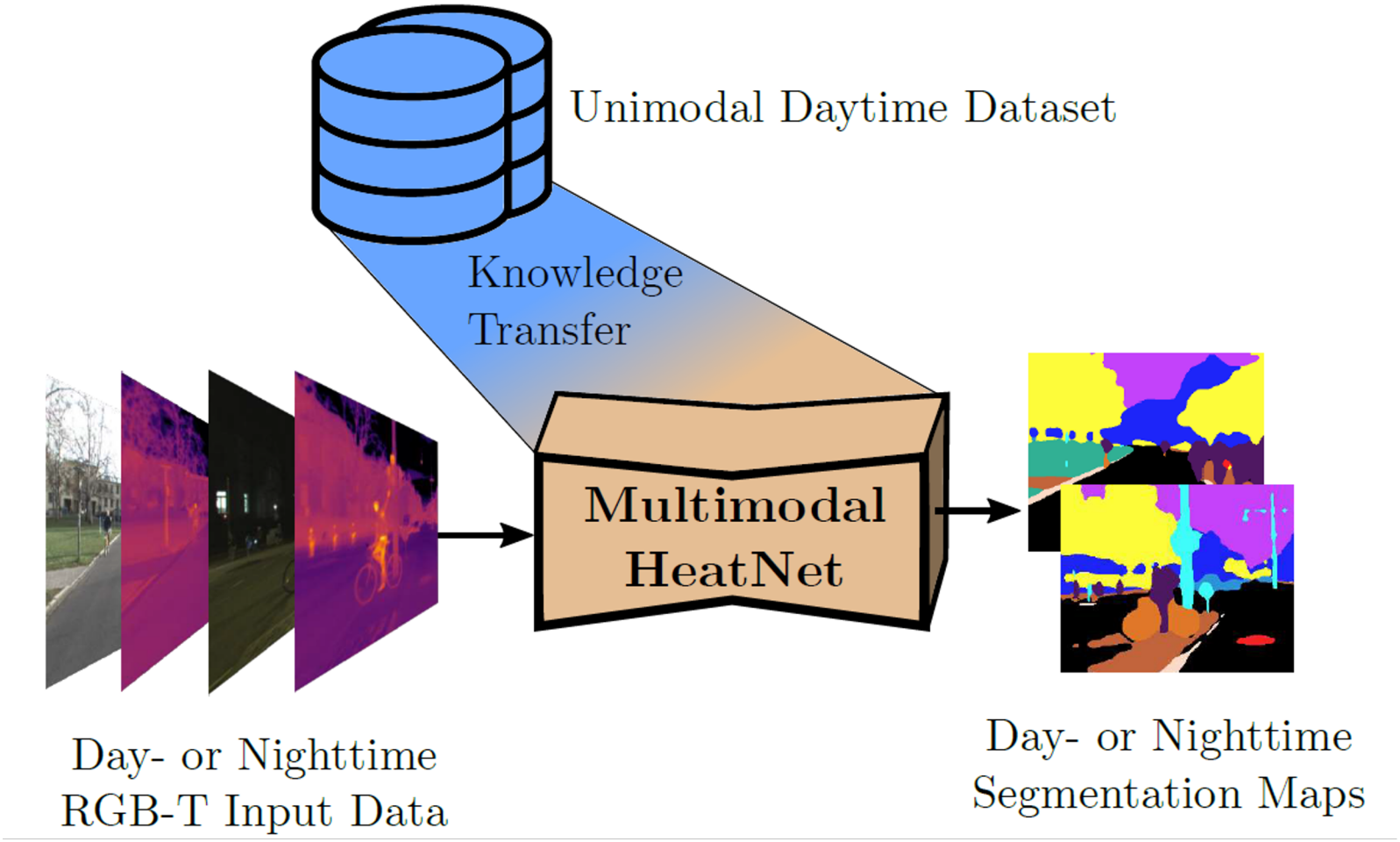
Motivation

Semantic segmentation prediction during nighttime using conventional CNN trained on publicly available datasets

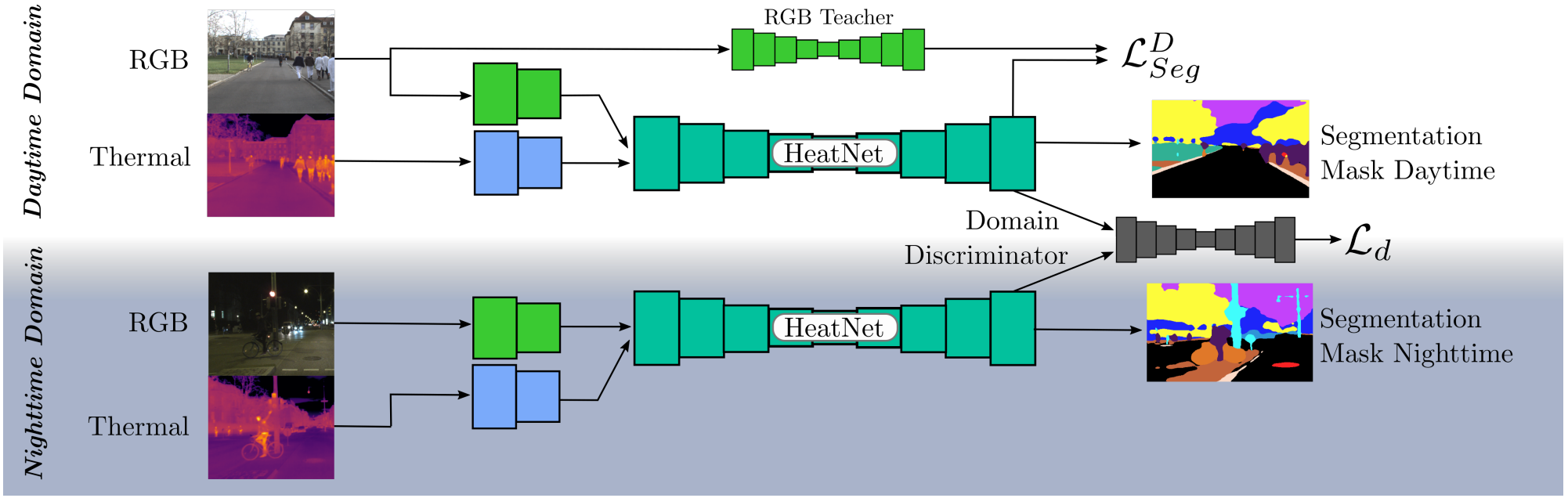
Motivation

Thermal infrared images exhibit small domain gap between daytime and nighttime

Approach


Unimodal Daytime Dataset
Knowledge
Transfer
Multimodal
HeatNet
Day- or Nighttime
RGB-T Input Data
Day- or Nighttime
Segmentation Maps

Approach

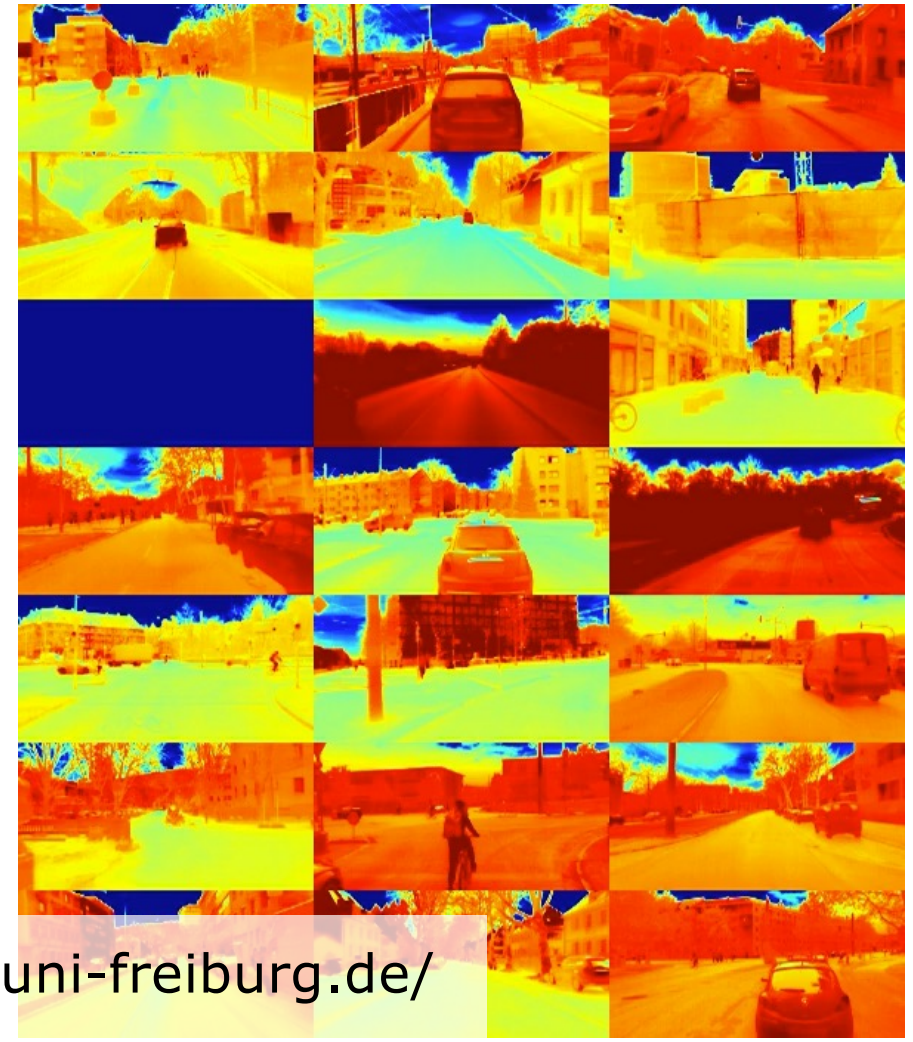


$$\mathcal{L}_{p_1} = \mathcal{L}_s^D + \lambda[0 - C(S_N)]^2,$$

$$\mathcal{L}_{p_2} = \frac{1}{HW} \sum_{h,w} \begin{cases} [0 - C(S_X)]^2, & \text{if } X = D \\ [1 - C(S_X)]^2, & \text{if } X = N \end{cases}$$

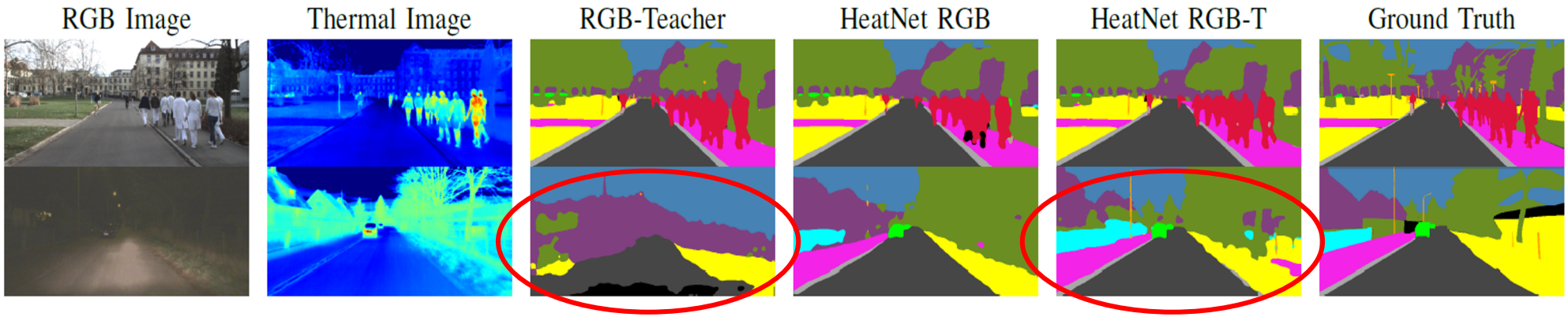
Dataset

Dataset



http://thermal.cs.uni-freiburg.de/

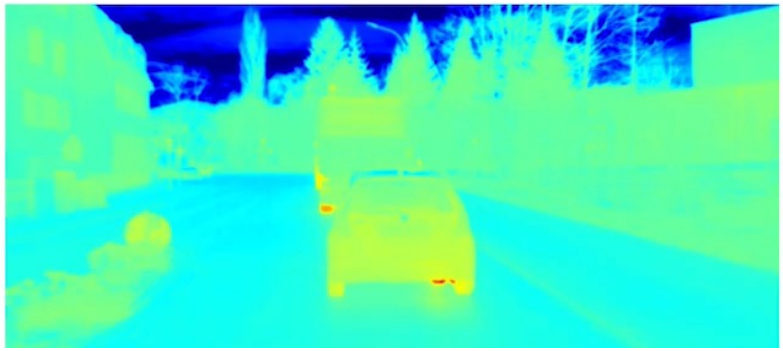
Experimental Results

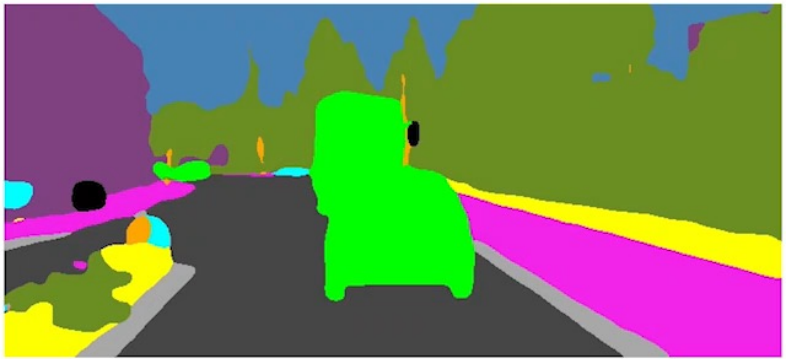


Experimental Results

RGB

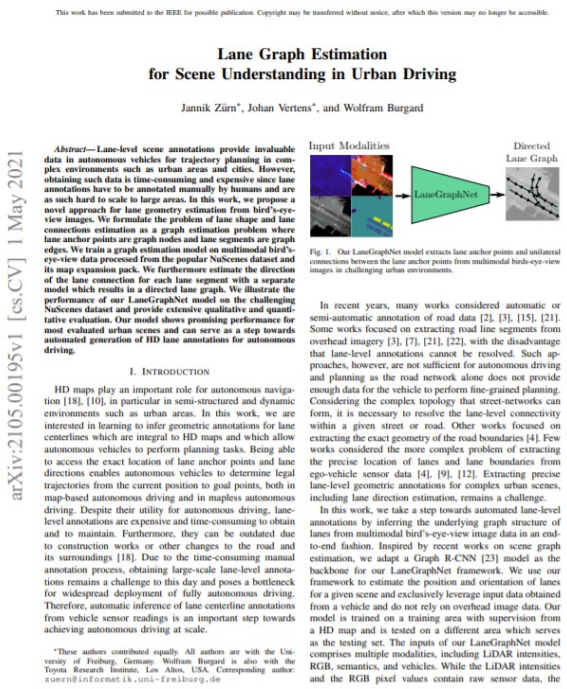
Thermal





HeatNet

Topology (Lane Graph) Estimation



Lane Graph Estimation for Scene Understanding in Urban Driving

Jannik Zürn*, Johan Vertens*, and Wolfram Burgard

***Abstract*— Lane-level scene annotations provide invaluable data in autonomous vehicles for trajectory planning in complex environments such as urban areas and cities. However, obtaining such data is time-consuming and expensive since lane annotations have to be annotated manually by humans and are as such hard to scale to large areas. In this work, we propose a novel approach for lane geometry estimation from bird's-eye-view images. We formulate the problem of lane shape and lane connections estimation as a graph estimation problem where lane anchor points are graph nodes and lane segments are graph edges. We train a graph estimation model on multimodal bird's-eye-view data processed from the popular NuScenes dataset and its map expansion pack. We furthermore estimate the direction of the lane connection for each lane segment with a separate model which results in a directed lane graph. We illustrate the performance of our LaneGraphNet model on the challenging NuScenes dataset and provide extensive qualitative and quantitative evaluation. Our model shows promising performance for most evaluated urban scenes and can serve as a step towards automated generation of HD lane annotations for autonomous driving.**

I. Introduction

HD maps play an important role for autonomous navigation [18], [10], in particular in semi-structured and dynamic environments such as urban areas. In this work, we are interested in learning to infer geometric annotations for lane centerlines which are integral to HD maps and which allow autonomous vehicles to perform planning tasks. Being able to access the exact location of lane anchor points and lane directions enables autonomous vehicles to determine legal trajectories from the current position to goal points, both in map-based autonomous driving and in mapless autonomous driving. Despite their utility for autonomous driving, lane-level annotations are expensive and time-consuming to obtain and to maintain. Furthermore, they can be outdated due to construction works or other changes to the road and its surroundings [18]. Due to the time-consuming manual annotation process, obtaining large-scale lane-level annotations remains a challenge to this day and poses a bottleneck for widespread deployment of fully autonomous driving. Therefore, automatic inference of lane centerline annotations from vehicle sensor readings is an important step towards achieving autonomous driving at scale.

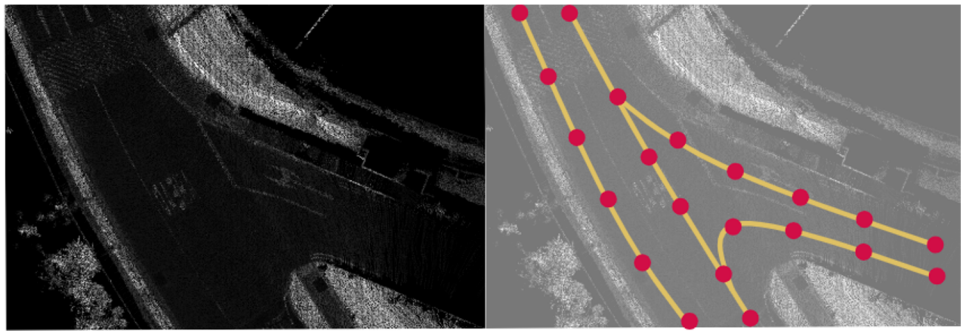
*These authors contributed equally. All authors are with the University of Freiburg, Germany. Wolfram Burgard is also with the Toyota Research Institute, Los Altos, USA. Corresponding author: zuern@informatik.uni-freiburg.de

Fig. 1. Our LaneGraphNet model extracts lane anchor points and unilateral connections between the lane anchor points from multimodal birds-eye-view images in challenging urban environments.

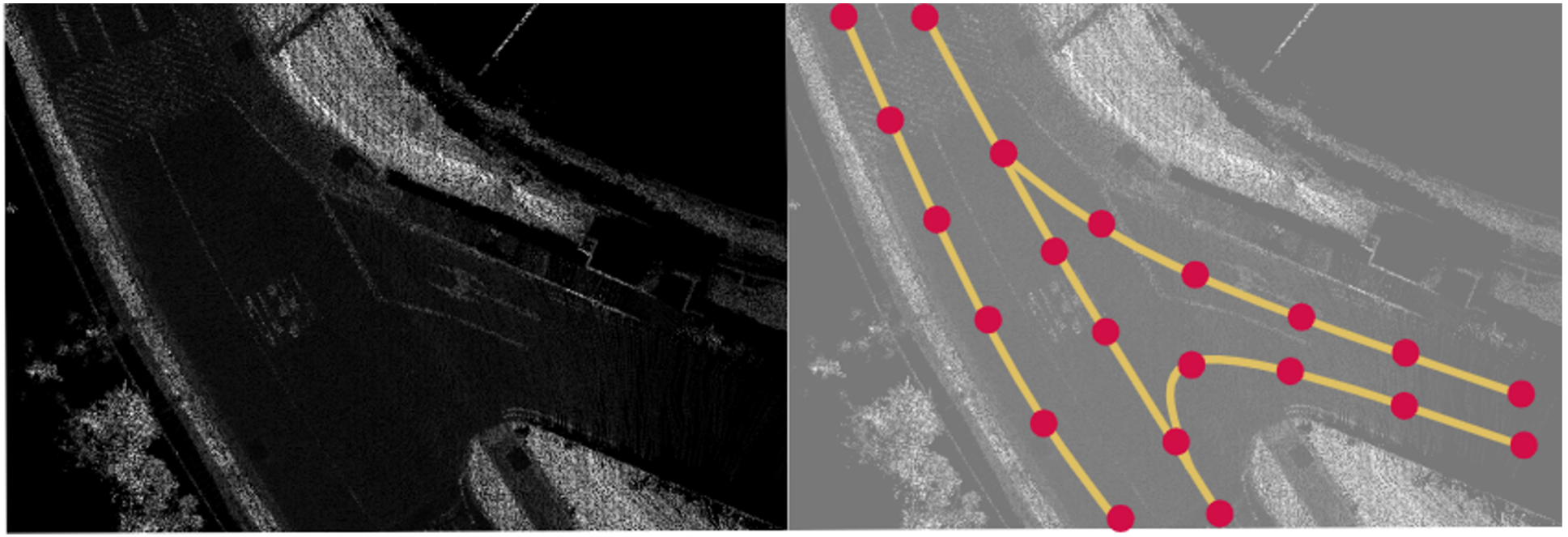
In recent years, many works considered automatic or semi-automatic annotation of road data [2], [3], [15], [21]. Some works focused on extracting road line segments from overhead imagery [3], [7], [21], [22], with the disadvantage that lane-level annotations cannot be resolved. Such approaches, however, are not sufficient for autonomous driving and planning as the road network alone does not provide enough data for the vehicle to perform fine-grained planning. Considering the complex topology that street-networks can form, it is necessary to resolve the lane-level connectivity within a given street or road. Other works focused on extracting the exact geometry of the road boundaries [4]. Few works considered the more complex problem of extracting the precise location of lanes and lane boundaries from ego-vehicle sensor data [4], [9], [12]. Extracting precise lane-level geometric annotations for complex urban scenes, including lane direction estimation, remains a challenge.

In this work, we take a step towards automated lane-level annotations by inferring the underlying graph structure of lanes from multimodal bird's-eye-view image data in an end-to-end fashion. Inspired by recent works on scene graph estimation, we adapt a Graph R-CNN [23] model as the backbone for our LaneGraphNet framework. We use our framework to estimate the position and orientation of lanes for a given scene and exclusively leverage input data obtained from a vehicle and do not rely on overhead image data. Our model is trained on a training area with supervision from a HD map and is tested on a different area which serves as the testing set. The inputs of our LaneGraphNet model comprises multiple modalities, including LiDAR intensities, RGB, semantics, and vehicles. While the LiDAR intensities and the RGB pixel values contain raw sensor data, the

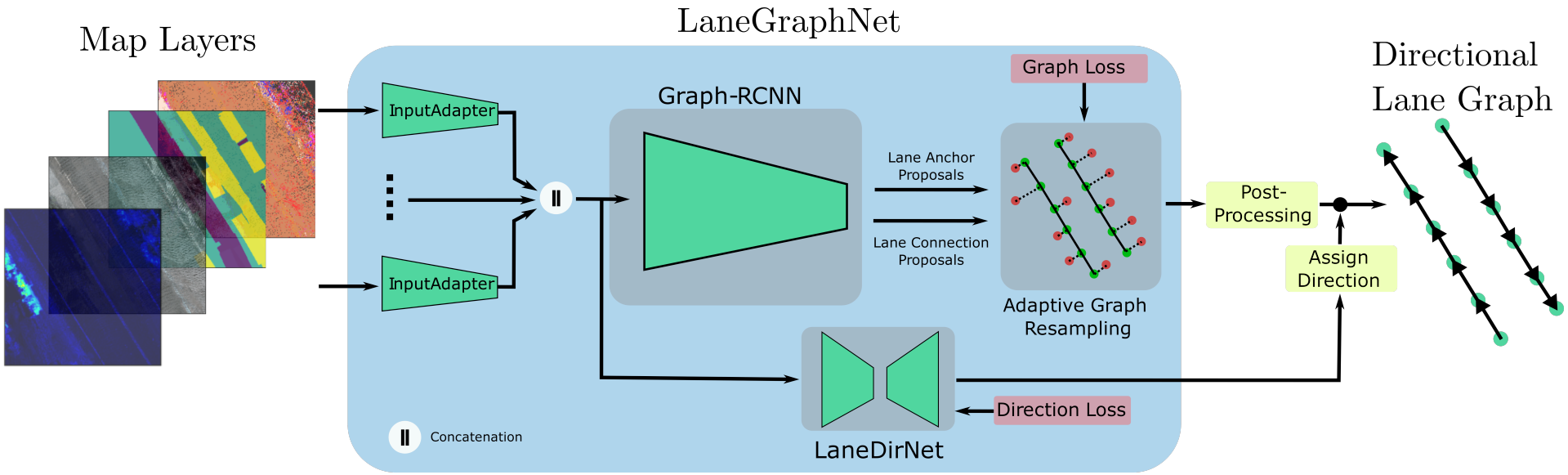
Lane Graph Estimation for Scene Understanding in Urban Driving. Jannik Zürn, Johan Vertens and Wolfram Burgard, Submitted to RA-L



Can we Learn HD-maps from BEV Sensor Data?

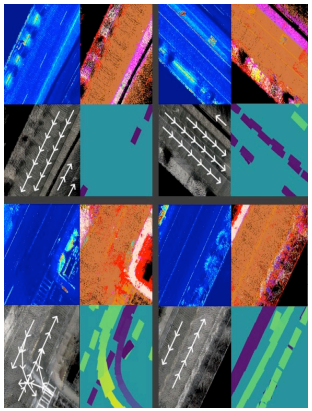


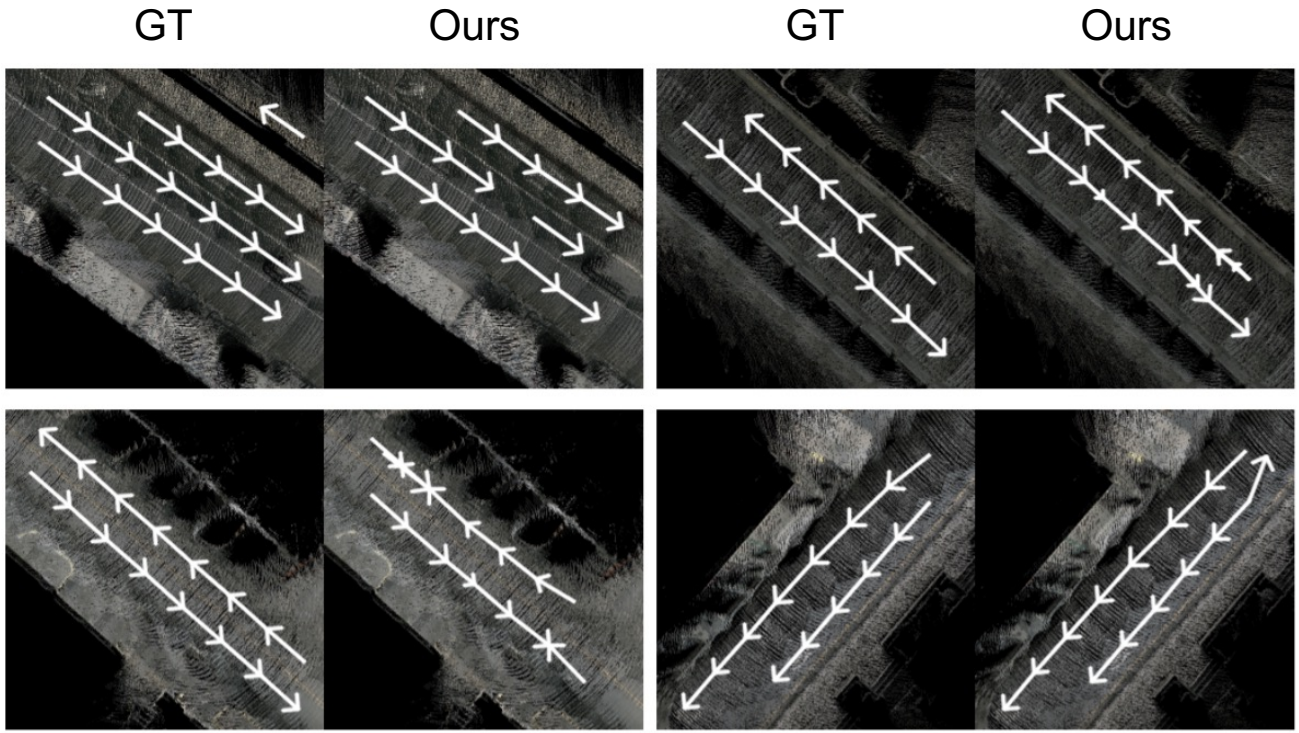
Approach


Map Layers
LaneGraphNet
InputAdapter
InputAdapter
Graph-RCNN
Lane Anchor Proposals
Lane Connection Proposals
Graph Loss
Adaptive Graph Resampling
Post-Processing
Assign Direction
Directional Lane Graph
Direction Loss
LaneDirNet
Concatenation

Approach, Qualitative Results

BEV-Modalities:
Lidar, RGB, Semantics,
Observed Car Poses





Summary

- HD maps are useful but limiting
- If we commit to HD maps we need to have proper approachs for change detection
- If we want to get rid of HD maps we need substantially extend the perception capabilities

- The key technology for this will be **Machine Learning**