Learning-based Multi-Modal Perception

Timo Class

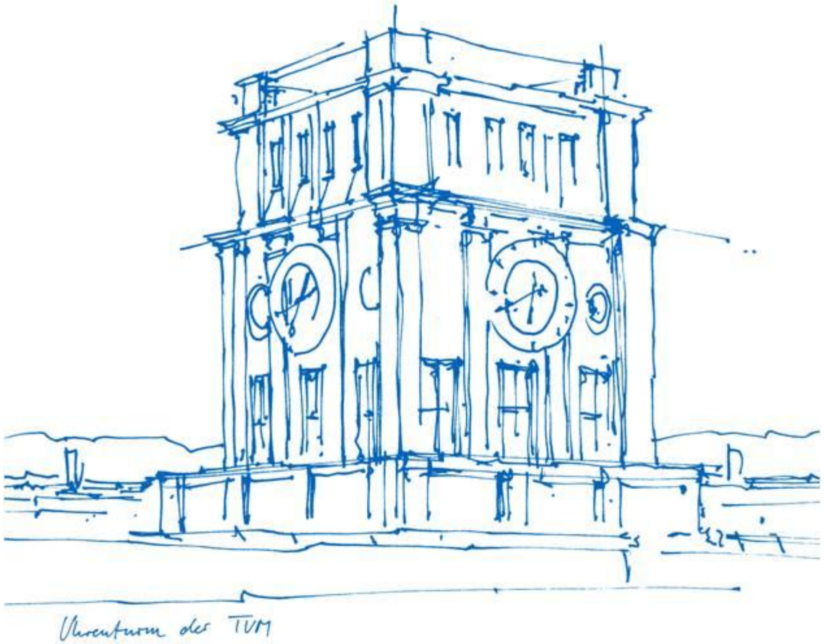
Technical University Munich

TUM School of Computation, Information and Technology

Smart Robotics Lab

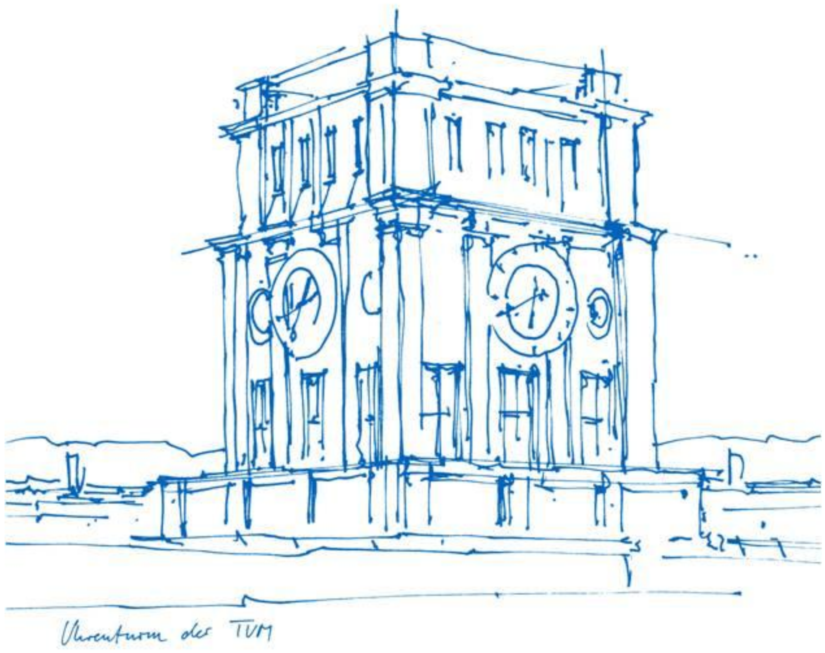
Seminar: Robot Perception & Intelligence

03 December 2024

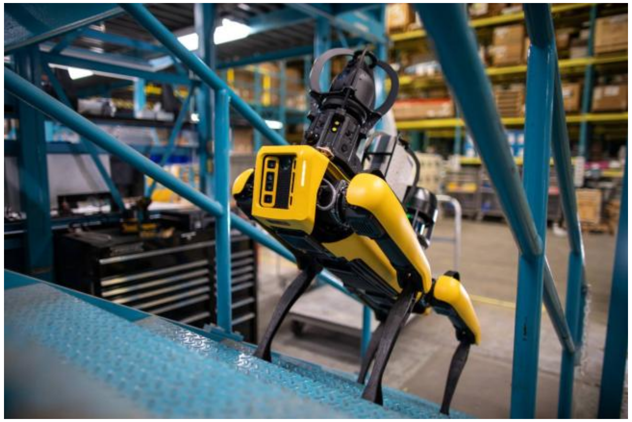


Structure

- Motivation
- Overview
- Focused research
    - Related work
    - Method descriptions
    - Experiments and results
    - Shortcomings and future work
- Conclusion



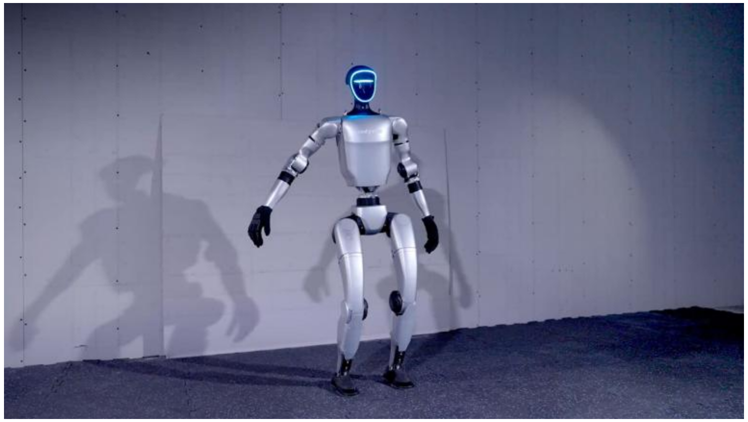
Motivation



Boston Dynamics, Spot

Waymo, Autonomous vehicle

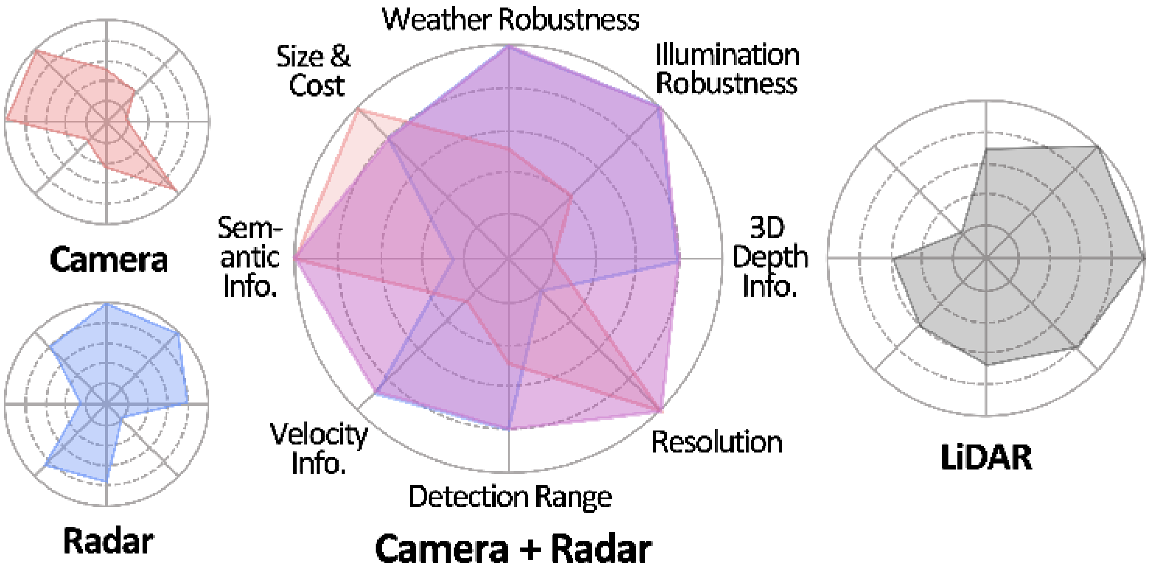
Starship Technologies, Delivery robot



Unitree Robotics, Humanoid Robot G1

Motivation

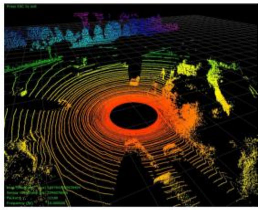
Challenges:



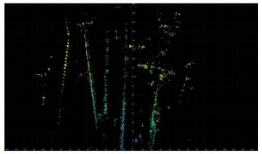
Apoorv Singh, Vision-RADAR fusion for Robotics BEV Detections: A Survey

Motivation

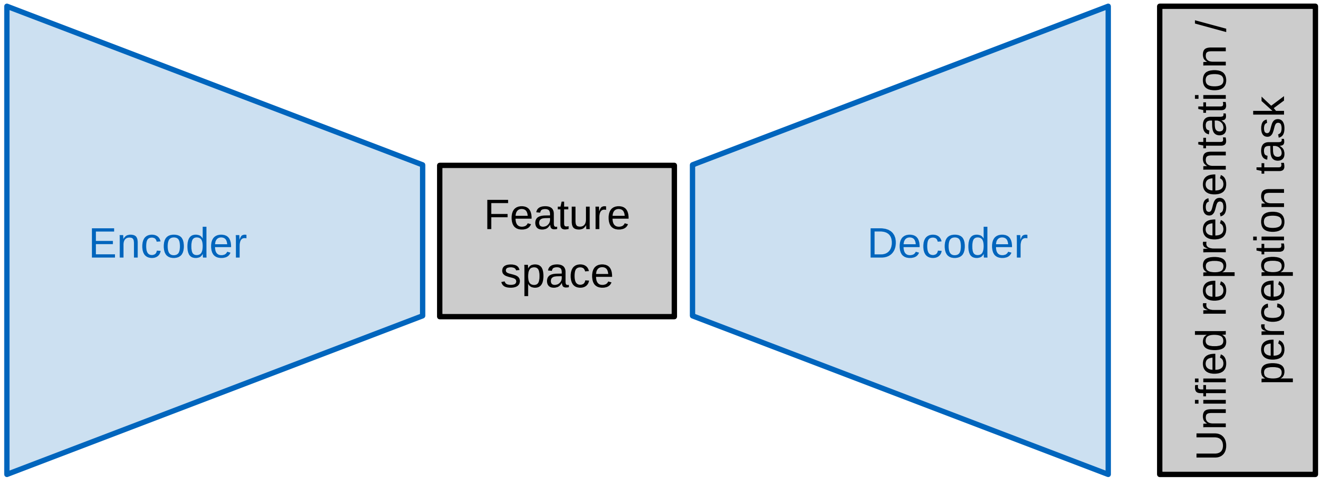
Camera images



LiDAR point cloud



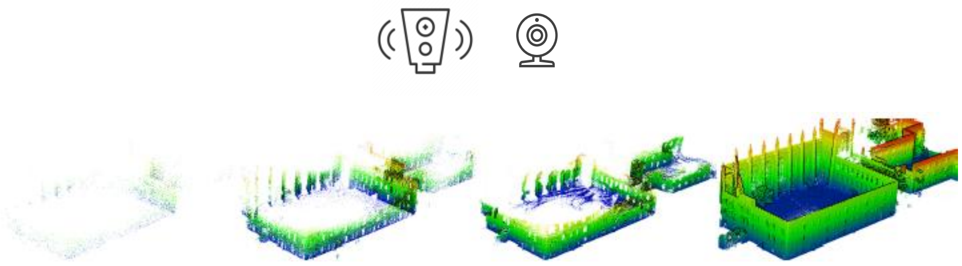
Radar point cloud



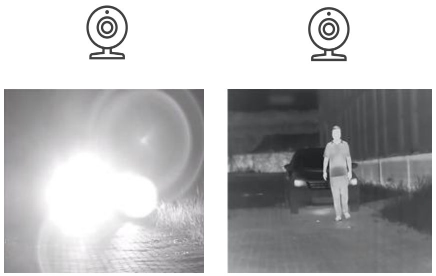
Encoder-decoder network structure

Overview

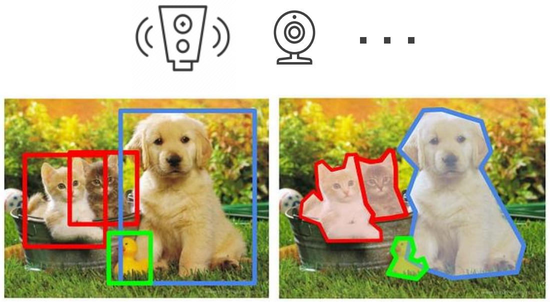
3D LiDAR Reconstruction



RTFNet

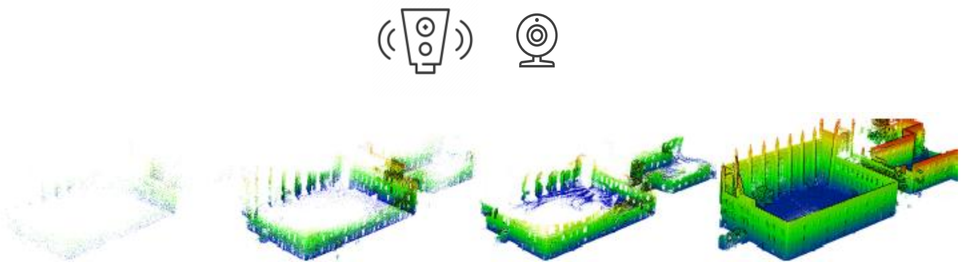


BEV Fusion

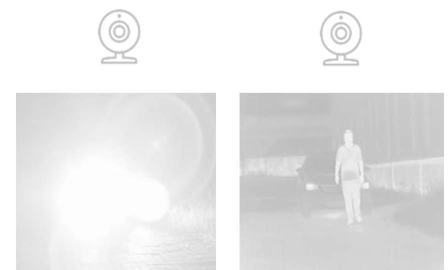


Overview

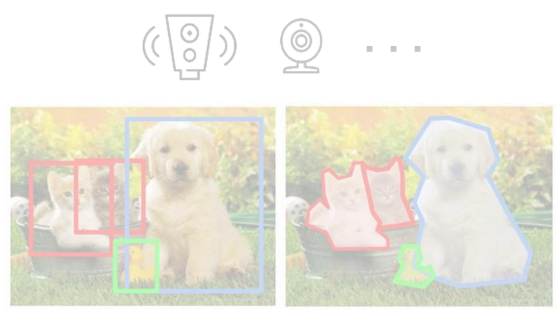
3D LiDAR Reconstruction



RTFNet



BEV Fusion

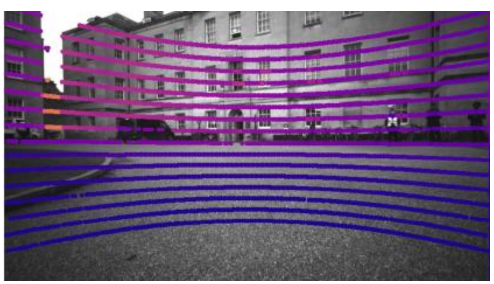


3D Lidar Reconstruction

Method description

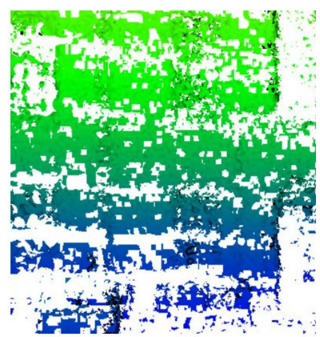
- **Problem:**
  - 3D reconstruction with sparse measurements results in incomplete reconstruction
    → path planning and free-space estimation for autonomous navigation may fail

Legged robot scanning a building



Camera image with 16-channel LiDAR overlay

Sparse reconstruction →

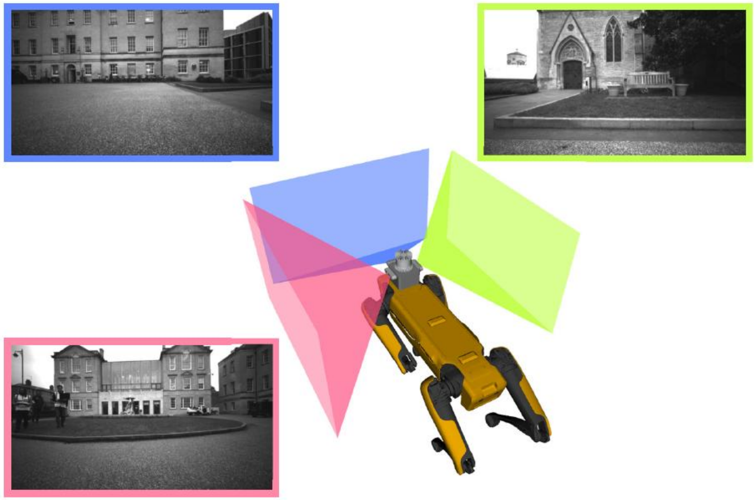


16-channel LiDAR 3D reconstruction

3D Lidar Reconstruction

Method description

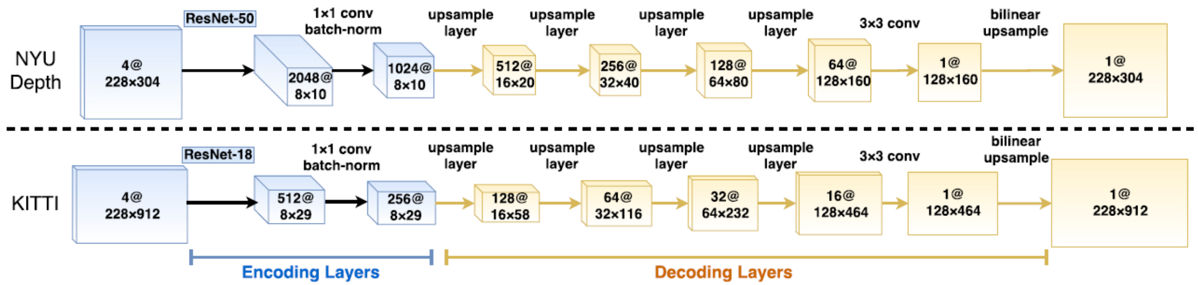
- **Solution:**
  - Learning-based dense depth completion
  - Incorporate RGB images from three-camera setup (270°)
  - Incorporate learning-based depth uncertainty predictions



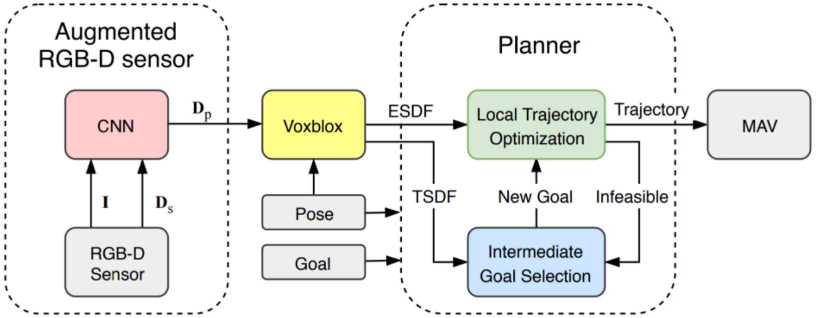
Camera setup on walking robot

3D Lidar Reconstruction

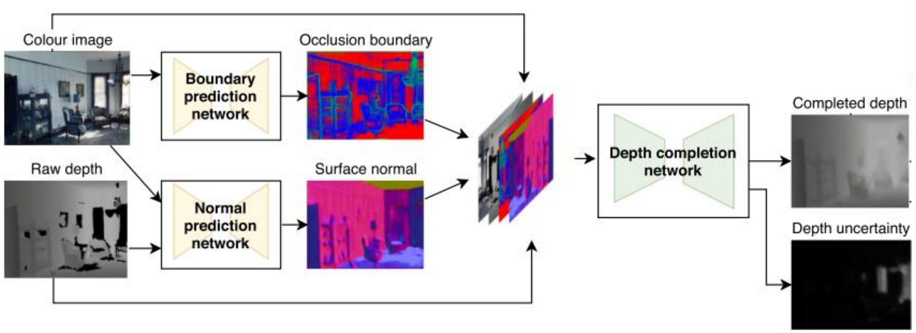
Related work



Fangchang Ma, Sparse-to-Dense



Fehr Marius, Predicting Unobserved Space For Planning via Depth Map Augmentation



Marija Popović, Volumetric Occupancy Mapping

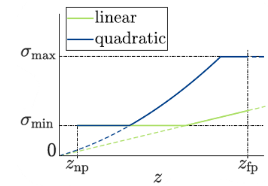
3D Lidar Reconstruction

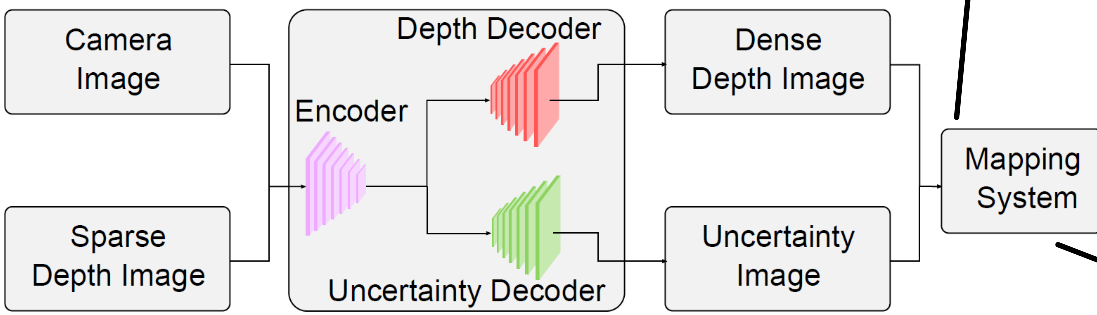
Method description

Linear sensor model: $\sigma(d)$

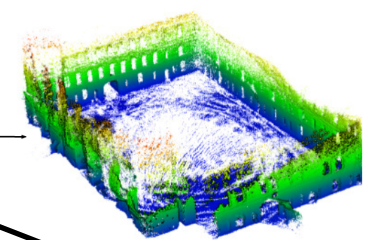
Depth filter:

$\sigma_{pred} < 2\sigma(d)$

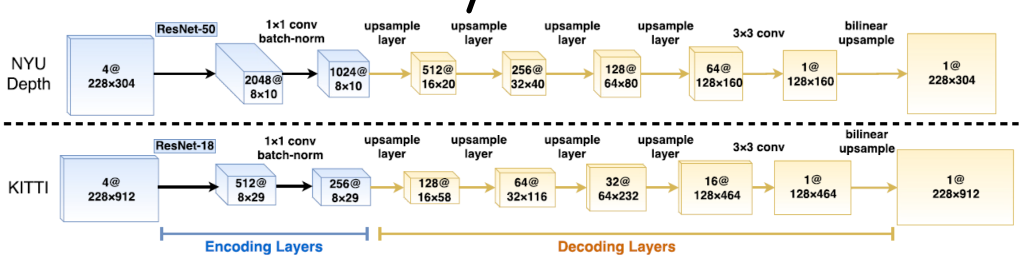




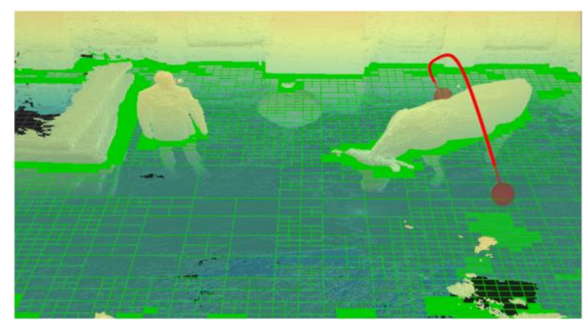
Dense Reconstruction



System pipeline for 3D reconstruction



Fangchang Ma, Sparse-to-Dense (Depth prediction CNN network S2D)



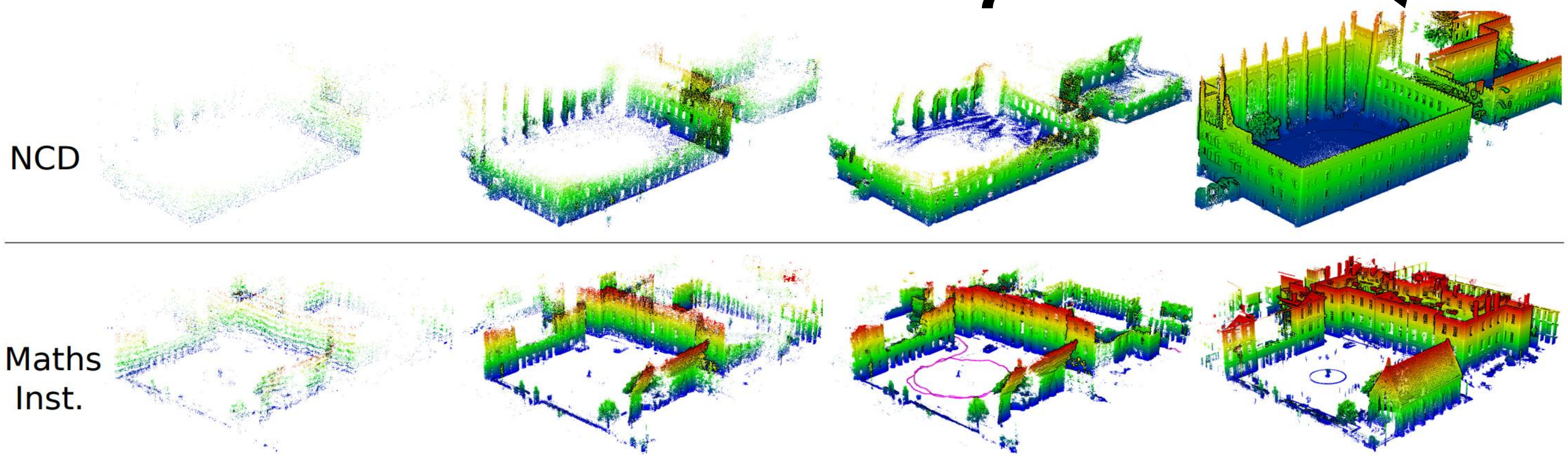
Nils Funk, Multi-Resolution 3D Mapping

3D Lidar Reconstruction

Experiments and results

Reconstruction & free-space estimation

Average point-to-point error: 0.2m

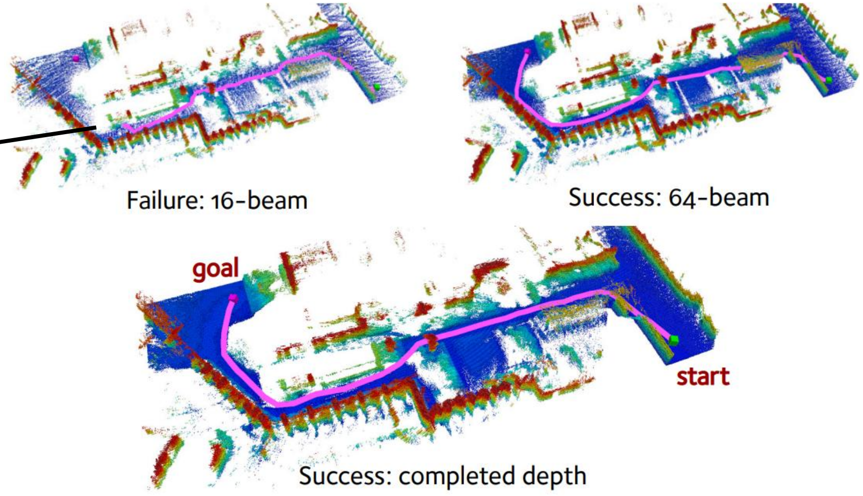


3D Reconstruction of NCD & Maths Inst. dataset for (left) 16-channel LiDAR, 64-channel LiDAR, Depth Completion, Ground Truth

3D Lidar Reconstruction

Experiments and results

Narrow passage

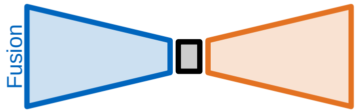


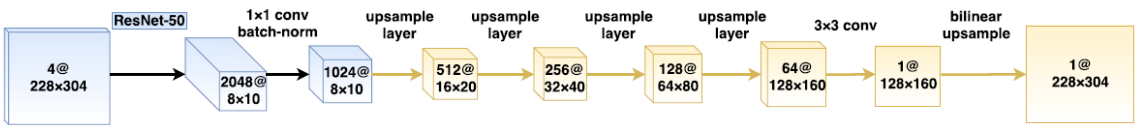
RRT* path planning test

3D Lidar Reconstruction

Shortcomings and future work

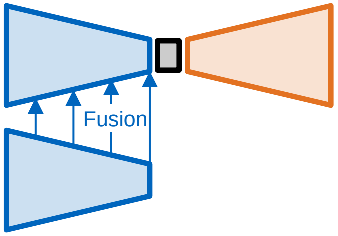
- **Shortcoming:**
  - Imbalanced modality representation in Encoder input
  - RGB image potentially dominates input due to higher channel size





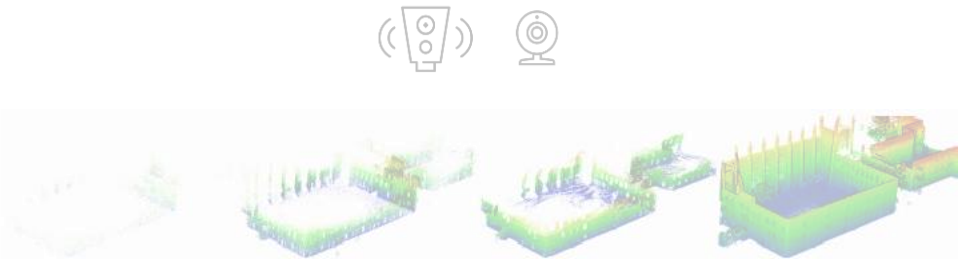
Fangchang Ma, Sparse-to-Dense (Depth prediction CNN network S2D)

- **Solution:**
  - Add additional encoder
  - Separate feature extraction of RGB and depth image
  - Fuse features in final layer or simultaneously at each layer

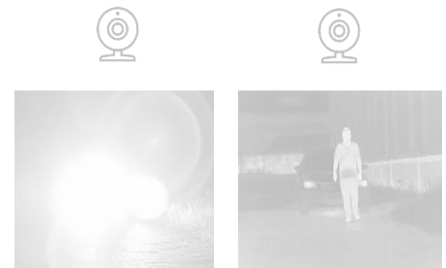


Overview

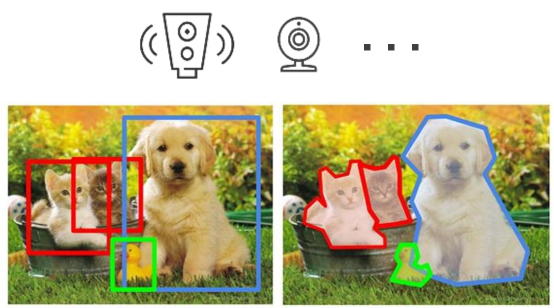
3D LiDAR Reconstruction



RTFNet



BEV Fusion



BEVFusion: Multi-Task Multi-Sensor Fusion

Method description

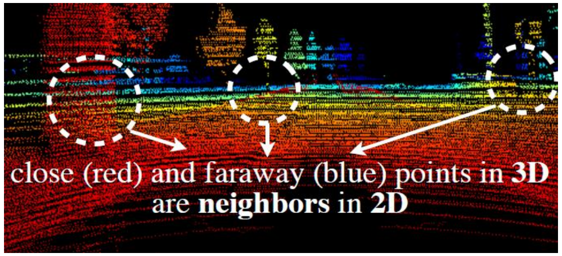
- **Problem:**
  - Various sensors entail different data modalities
  - Various tasks entail different requirements

- **Solution:**
  - Transformation into unified representation

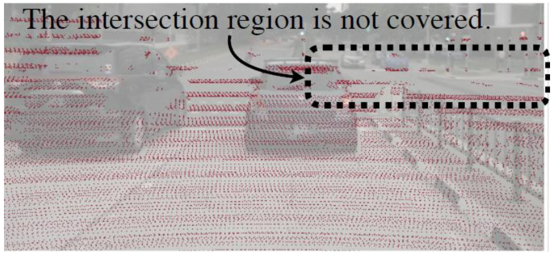
BEVFusion: Multi-Task Multi-Sensor Fusion

Method description

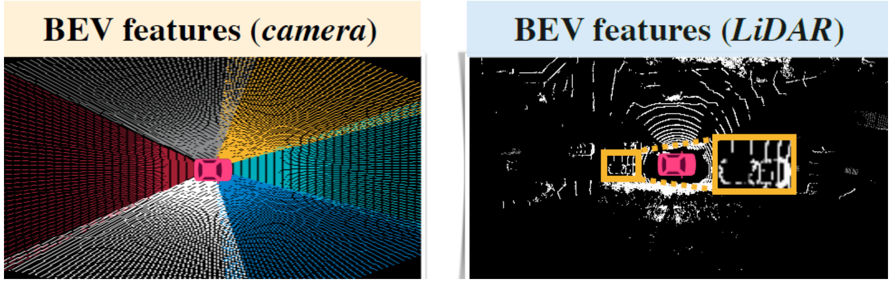
- **Solution:**
  - To camera → geometric-lossy



  - To LiDAR → semantic-lossy

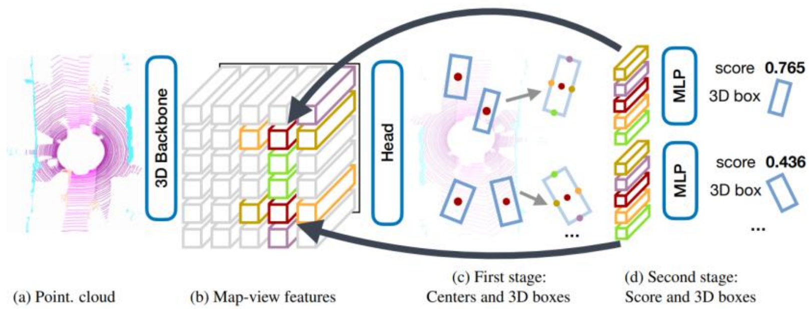


  - To Birds-Eye-View
    → preserves geometric information
    → preserves semantic information

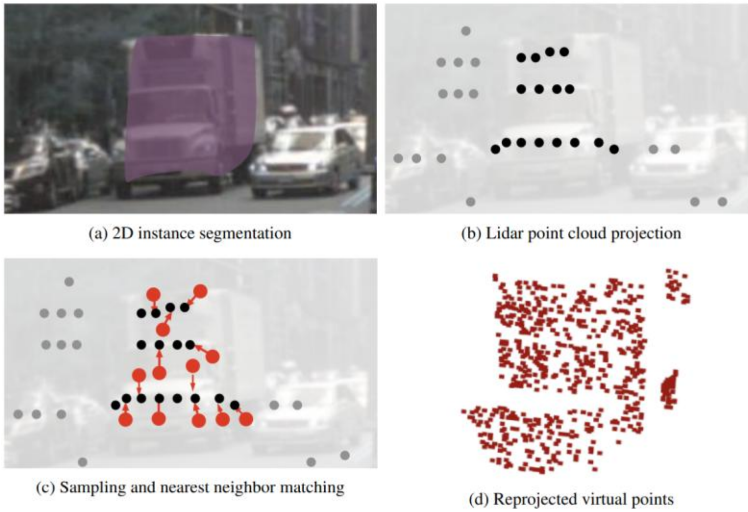


BEVFusion: Multi-Task Multi-Sensor Fusion

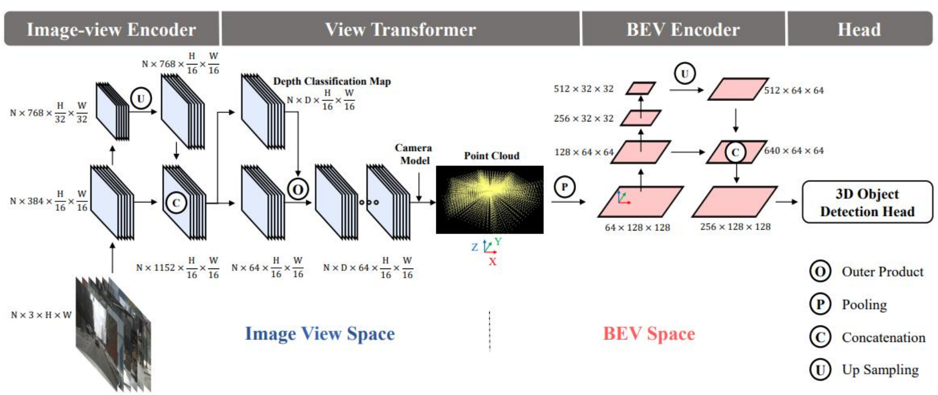
Related work



Tianwei Yin, Center-based 3D Object Detection and Tracking



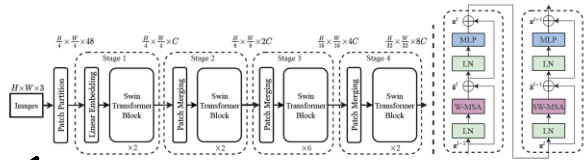
Tianwei Yin, Multimodal Virtual Point 3D Detection



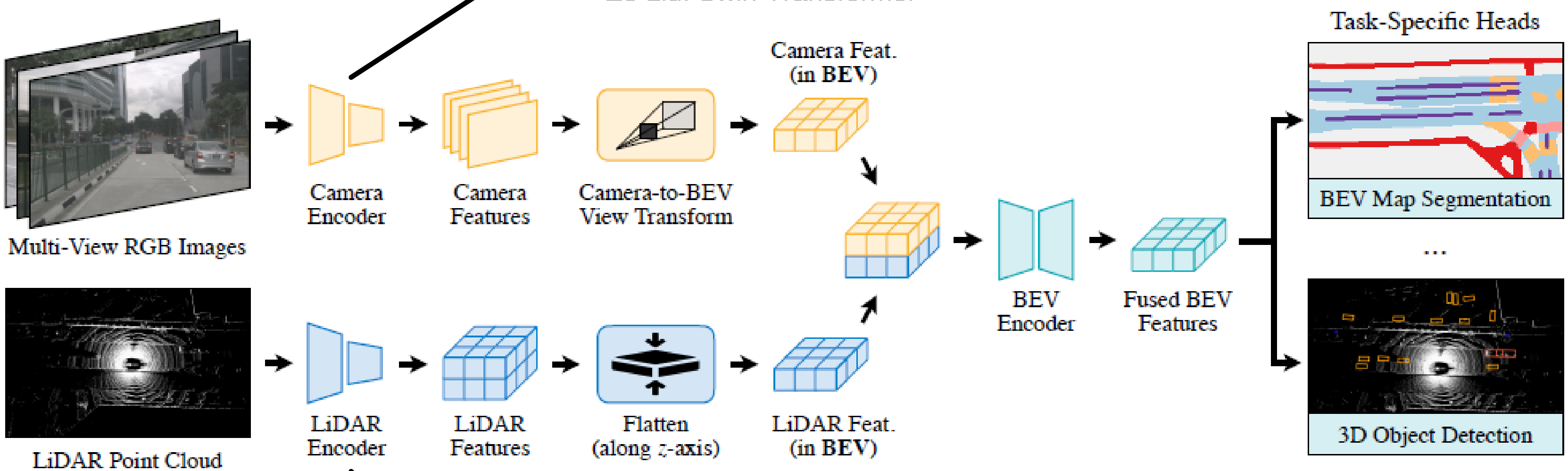
Junjie Huang, BEVDet

BEVFusion: Multi-Task Multi-Sensor Fusion

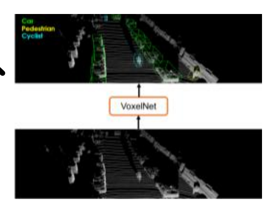
Method description



Ze Liu. Swin Transformer



BEVFusion system pipeline



Yin Zhou, VoxelNet

BEVFusion: Multi-Task Multi-Sensor Fusion

Experiments and results

	Modality	Sunny mAP	Sunny mIoU	Rainy mAP	Rainy mIoU	Day mAP	Day mIoU	Night mAP	Night mIoU
CenterPoint	L	62.9	50.7	59.2	42.3	62.8	48.9	35.4	37.0
BEVDet/LSS*	C	32.9	59.0	33.7	50.5	33.7	57.4	13.5	30.8
MVP	C+L	65.9 (+3.0)	51.0 (-8.0)	66.3 (+7.1)	42.9 (-7.6)	66.3 (+3.5)	49.2 (-8.2)	38.4 (+3.0)	37.5 (+6.7)
BEVFusion	C+L	68.2 (+5.3)	65.6 (+6.6)	69.9 (+10.7)	55.9 (+5.4)	68.5 (+5.7)	63.1 (+5.7)	42.8 (+7.4)	43.6 (+12.8)

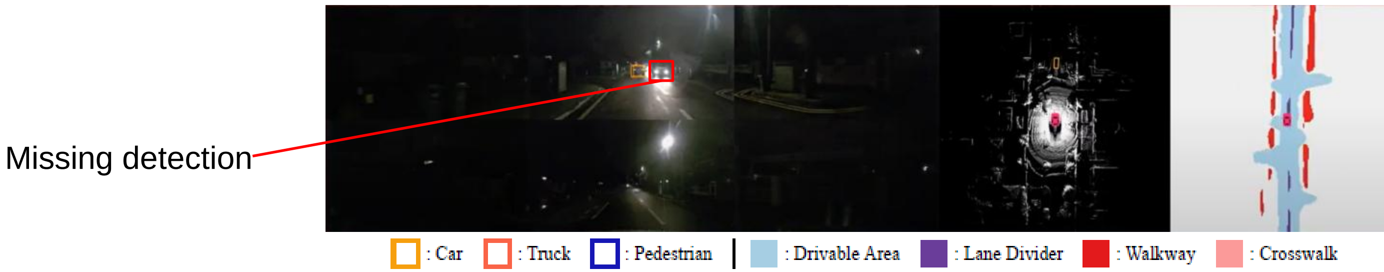
=> Improved performance at nighttime

=> Improved performance in rainy weather

Performance analysis of BEVFusion under different weather and lightning conditions

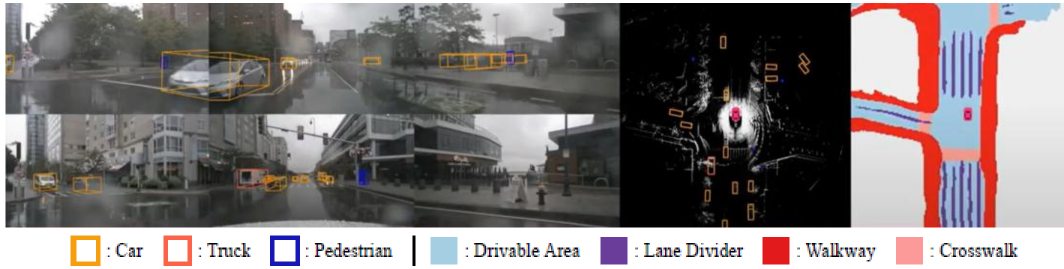
BEVFusion: Multi-Task Multi-Sensor Fusion

Experiments and results



Camera-only baseline

LiDAR-only baseline

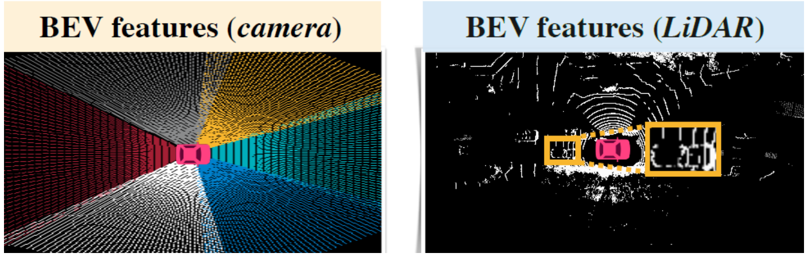


BEVFusion

BEVFusion: Multi-Task Multi-Sensor Fusion

Shortcomings and future work

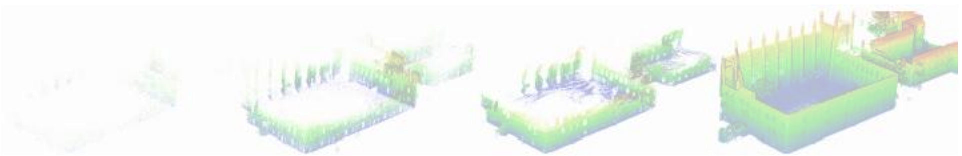
- **Shortcoming:**
  - Loss of z-dimension due to transformation to Bird‘s-Eye View space



  - **Solution:**
    - Multi-Scale BEV representation
    - Hybrid representation (BEV & 3D front-view features)
    - Volumetric representation (Voxelization)

Overview

3D LiDAR Reconstruction



RTFNet

BEV Fusion

RTFNet: RGB-Thermal Fusion Network

Method description

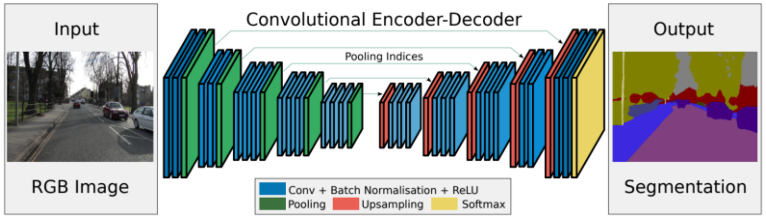
- **Problem:**
  - RGB camera performance is prone to lightning condition
  - Worse performance in total darkness or glaring situations

- **Solution:**
  - Incorporate thermal camera image

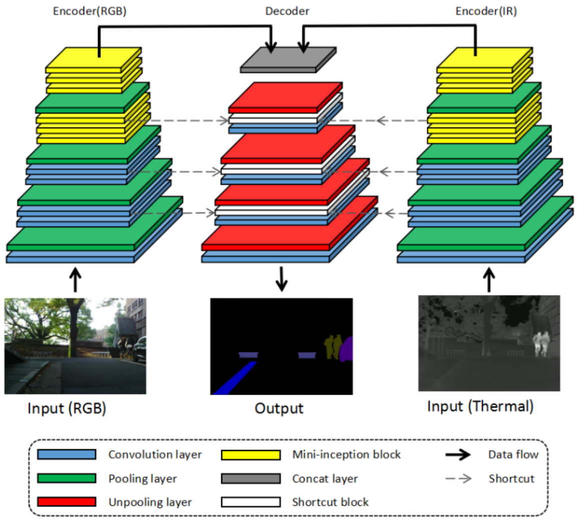
Comparison of RGB- and thermal image in a bright scene

RTFNet: RGB-Thermal Fusion Network

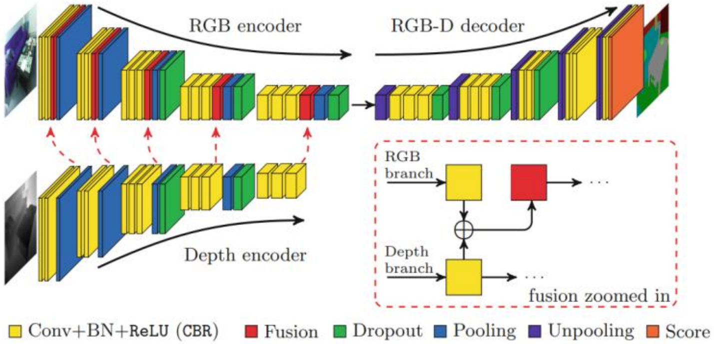
Related work



Vijay Badrinarayanan, SegNet



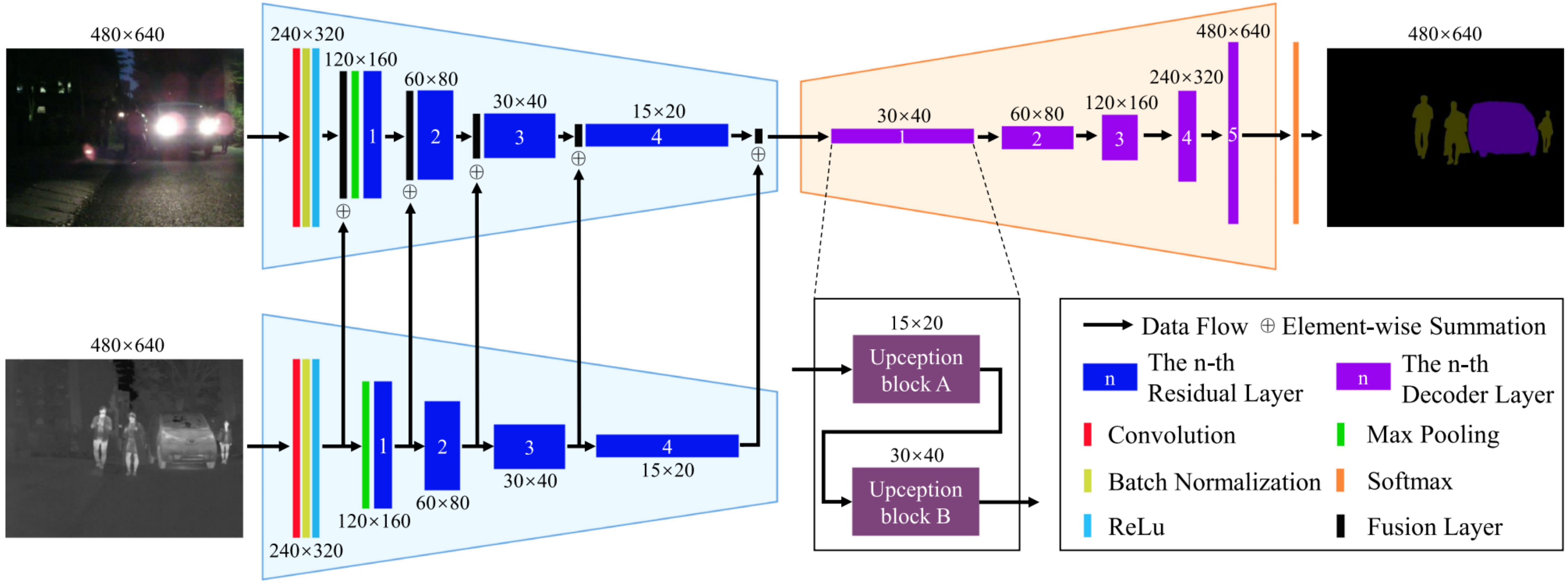
Qishen Ha, MFNet



Caner Hazirbas, FuseNet

RTFNet: RGB-Thermal Fusion Network

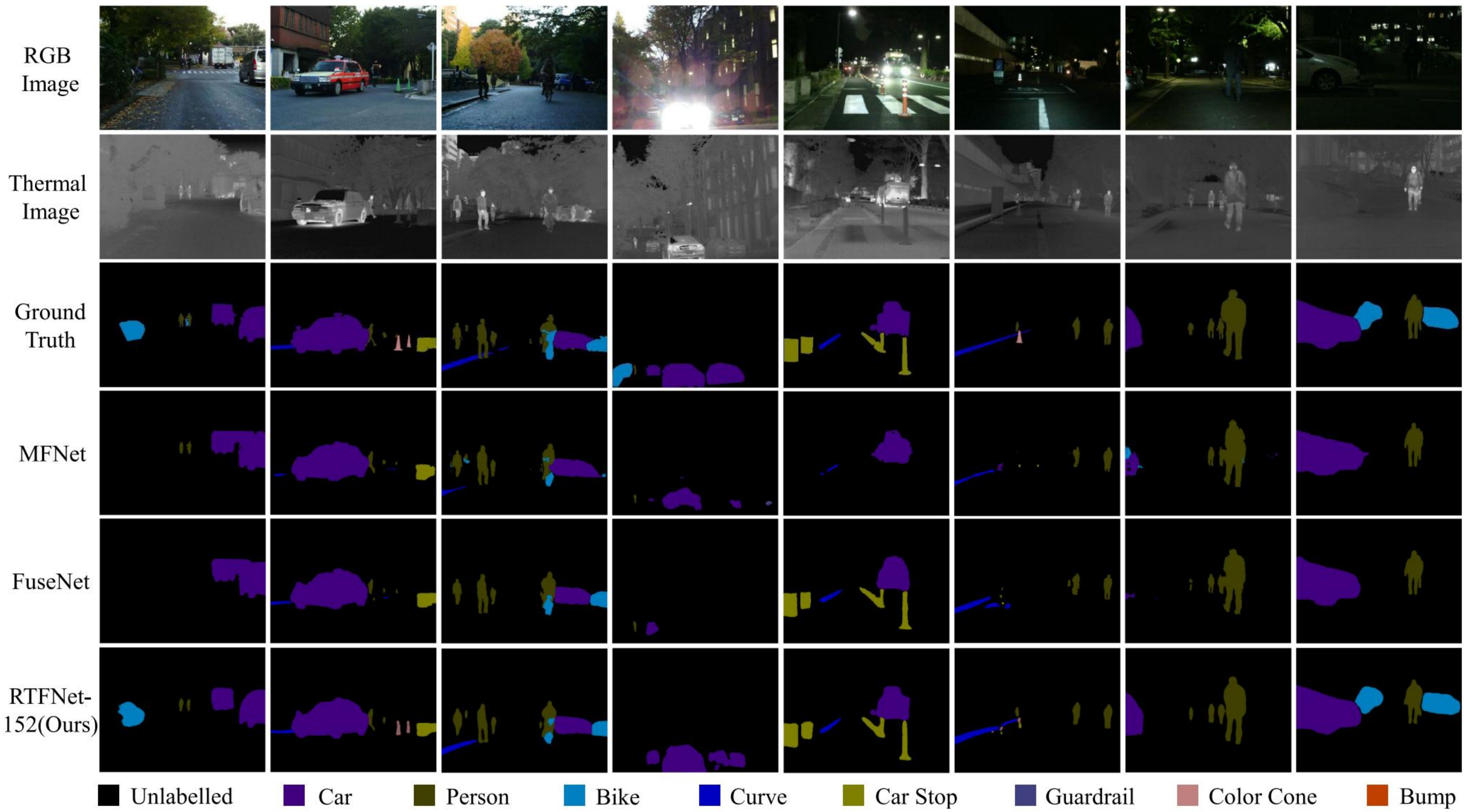
Method description



RTFNet network architecture

RTFNet: RGB-Thermal Fusion Network

Experiments and results



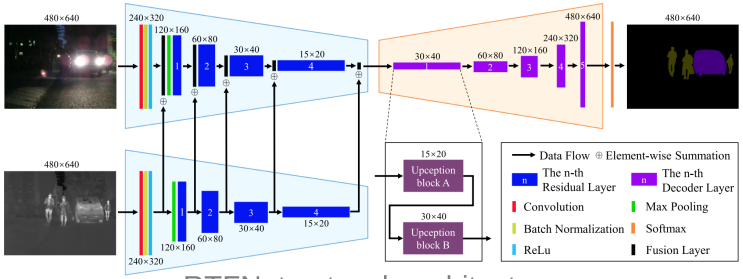
Qualitative comparison of data-fusion networks and RTFNet in different lightning conditions

RTFNet: RGB-Thermal Fusion Network

Shortcomings and future work

- **1st Shortcoming:**
  - Slow inference speed on embedded platforms
  - Large Encoder network

- **1st Solution:**
  - Reduce network size (especially Encoder)

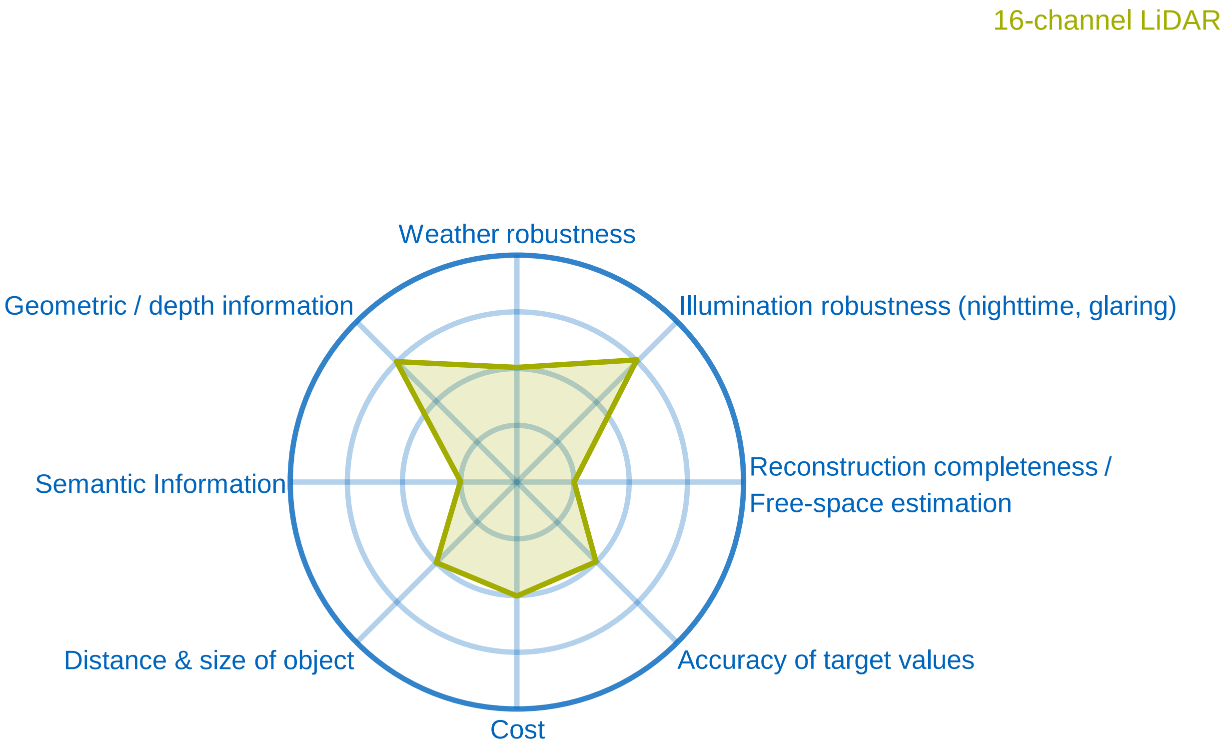


RTFNet network architecture

- **2nd Shortcoming:**
  - Thermal images less informative when near objects share similar temperature

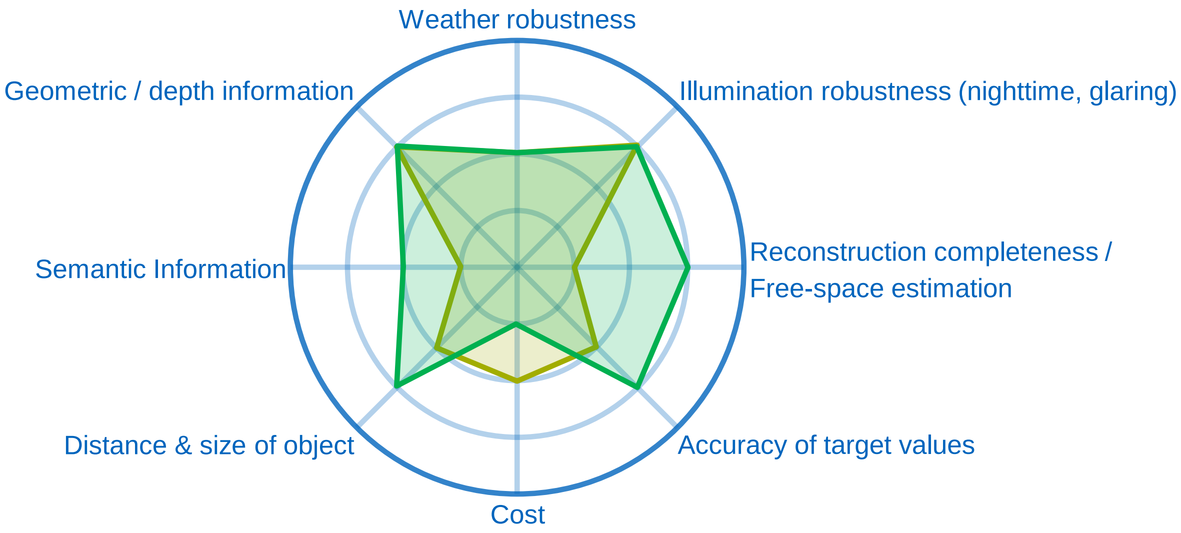
- **2nd Solution:**
  - Develop mechanisms to identify data that is more informative

Conclusion


16-channel LiDAR
Weather robustness
Illumination robustness (nighttime, glaring)
Reconstruction completeness /
Free-space estimation
Accuracy of target values
Cost
Distance & size of object
Semantic Information
Geometric / depth information

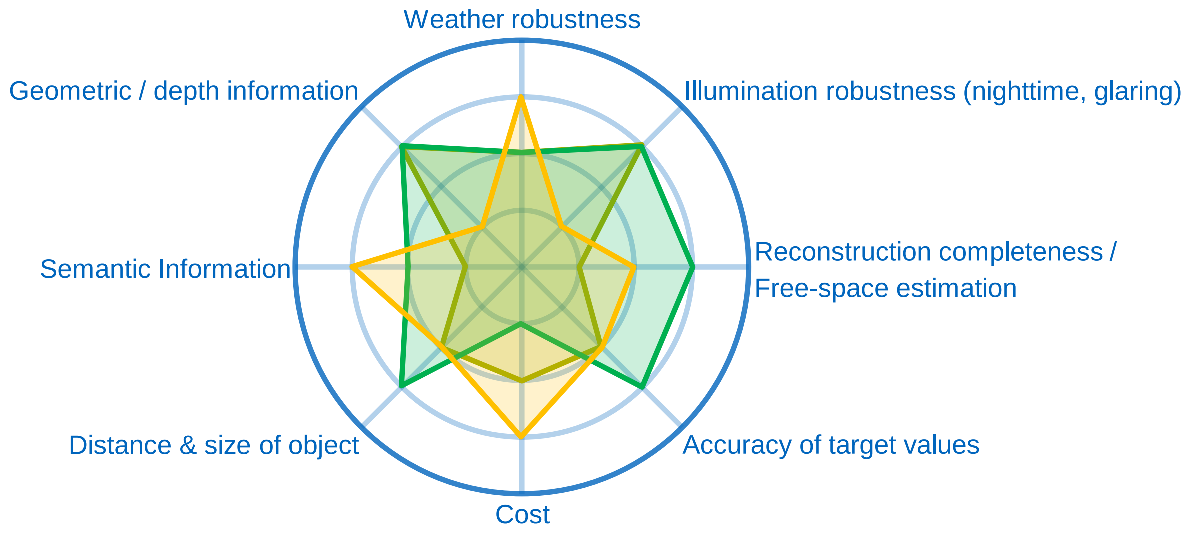
Conclusion

16-channel LiDAR
64-channel LiDAR

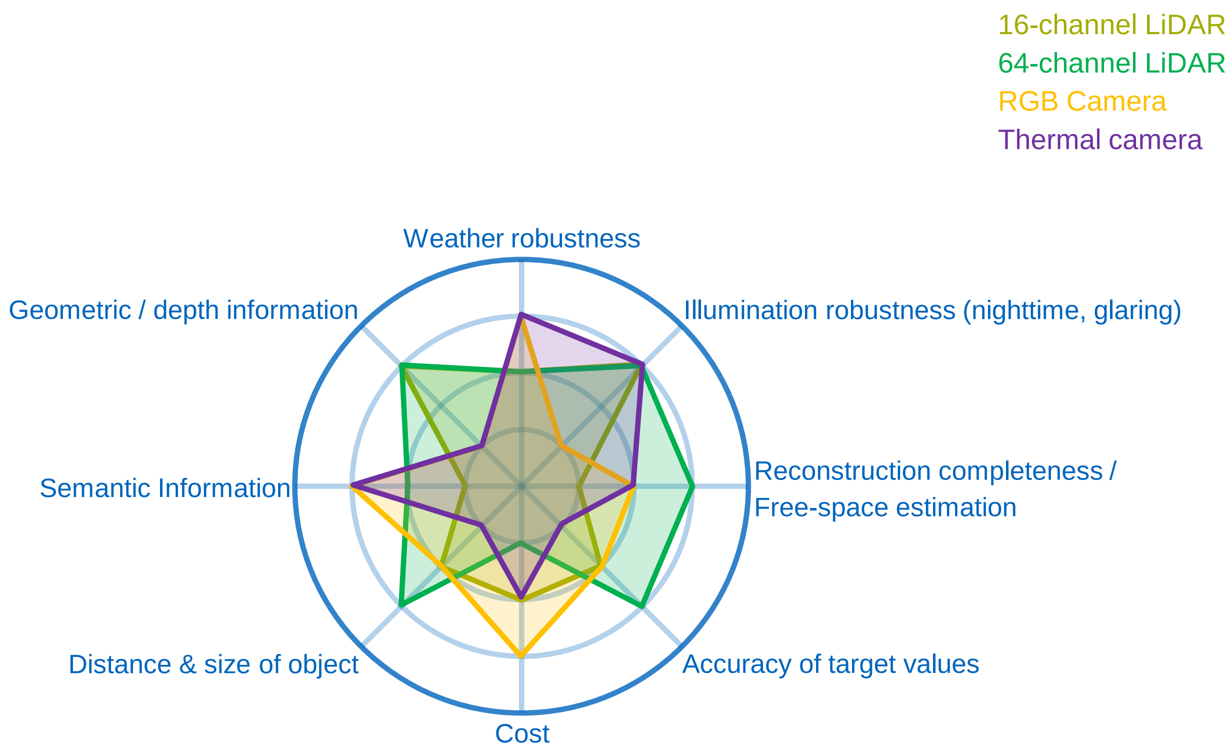


Conclusion

16-channel LiDAR
64-channel LiDAR
RGB Camera

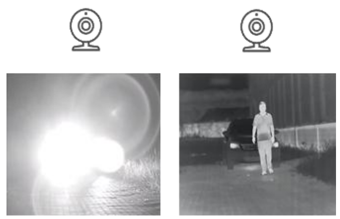


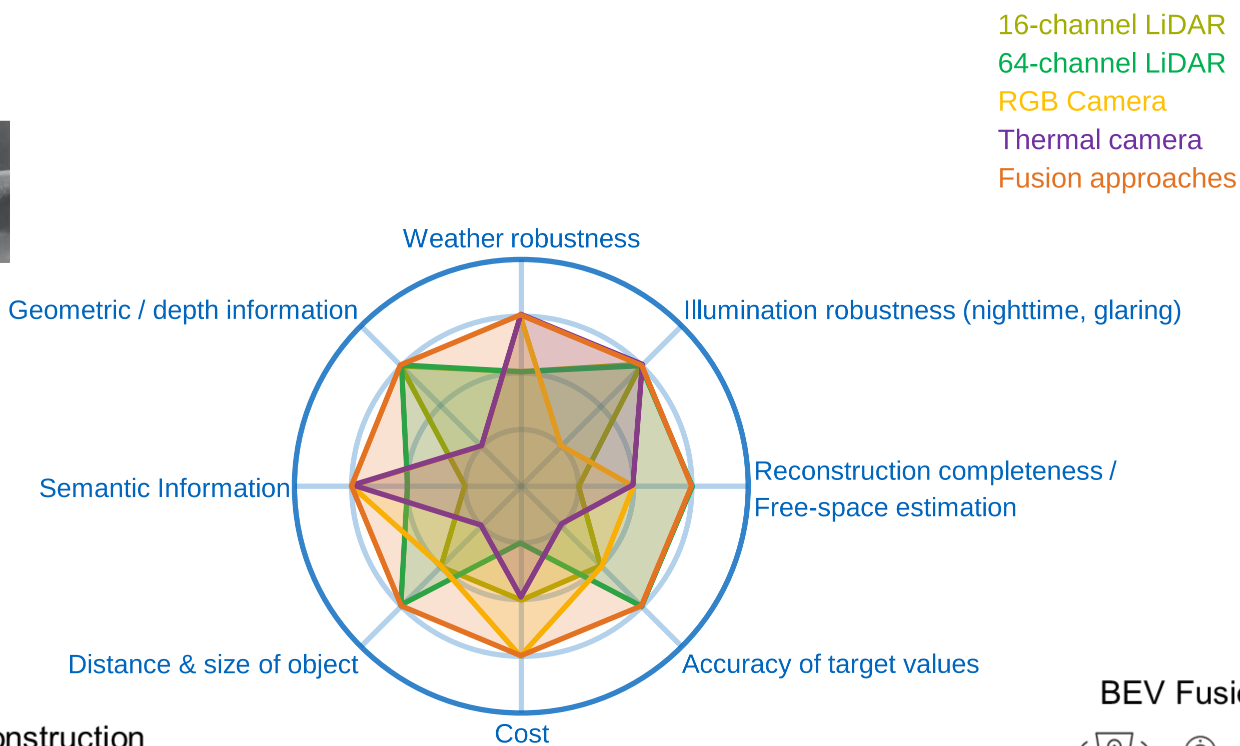
Conclusion


16-channel LiDAR
64-channel LiDAR
RGB Camera
Thermal camera
Weather robustness
Illumination robustness (nighttime, glaring)
Reconstruction completeness /
Free-space estimation
Accuracy of target values
Cost
Distance & size of object
Semantic Information
Geometric / depth information

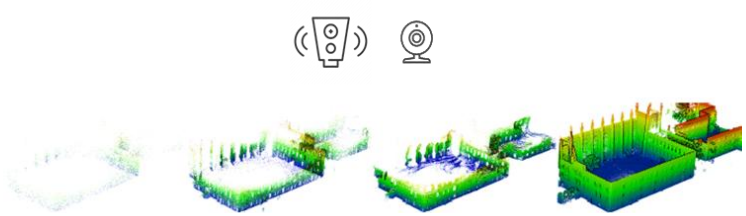
Conclusion

RTFNet





3D LiDAR Reconstruction



BEV Fusion