Hands-on Intro To Developing Vision and LiDAR Classifier

Formula Student Driverless Workshop

powered by

Introduction

Sibo Zhu

- Perception Lead at MIT Driverless
- Research Assistant at MIT HAN Lab

Zhijian Liu

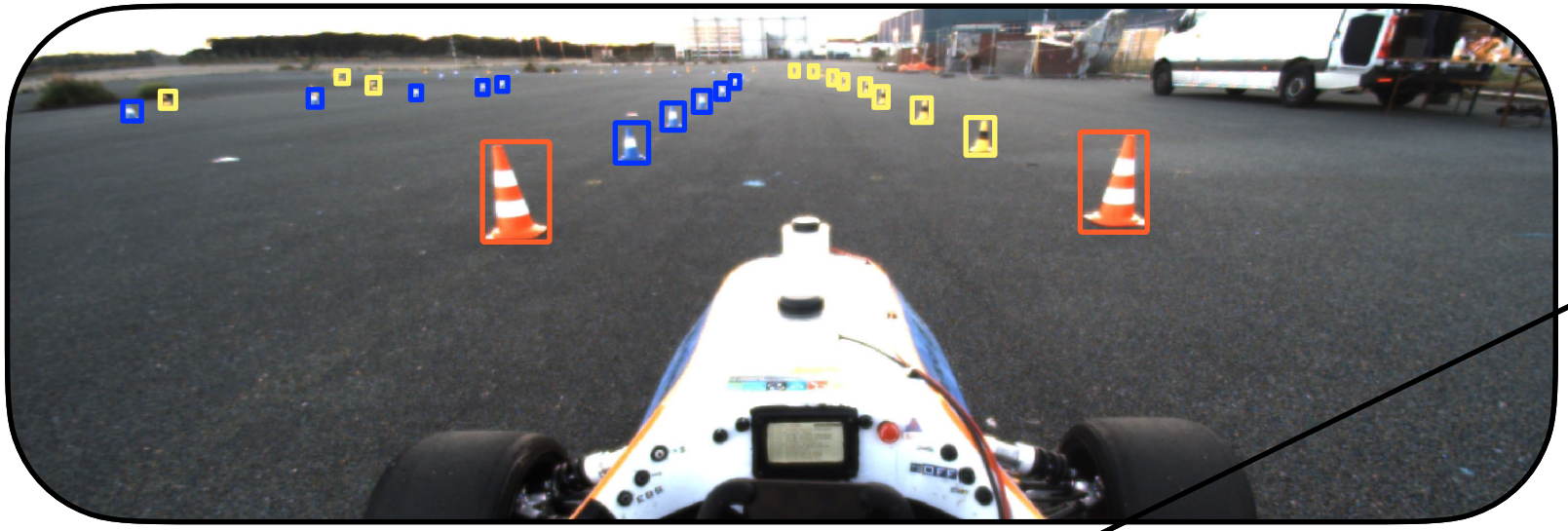
- Perception Lead at MIT Driverless
- PhD student at MIT HAN Lab

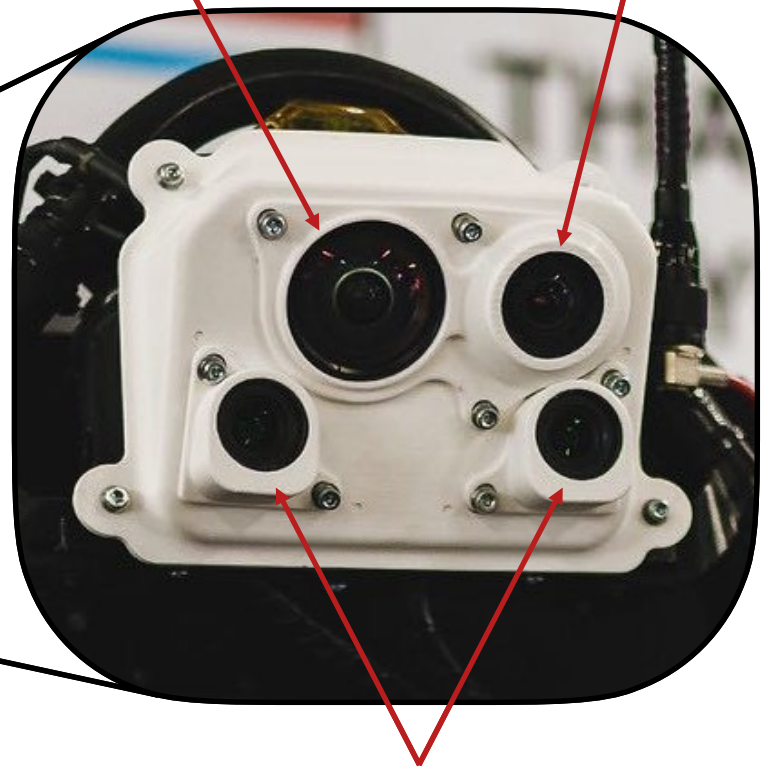
Haotian Tang

- Perception Lead at MIT Driverless
- PhD student at MIT HAN Lab

Introduction

Vision Perception Task



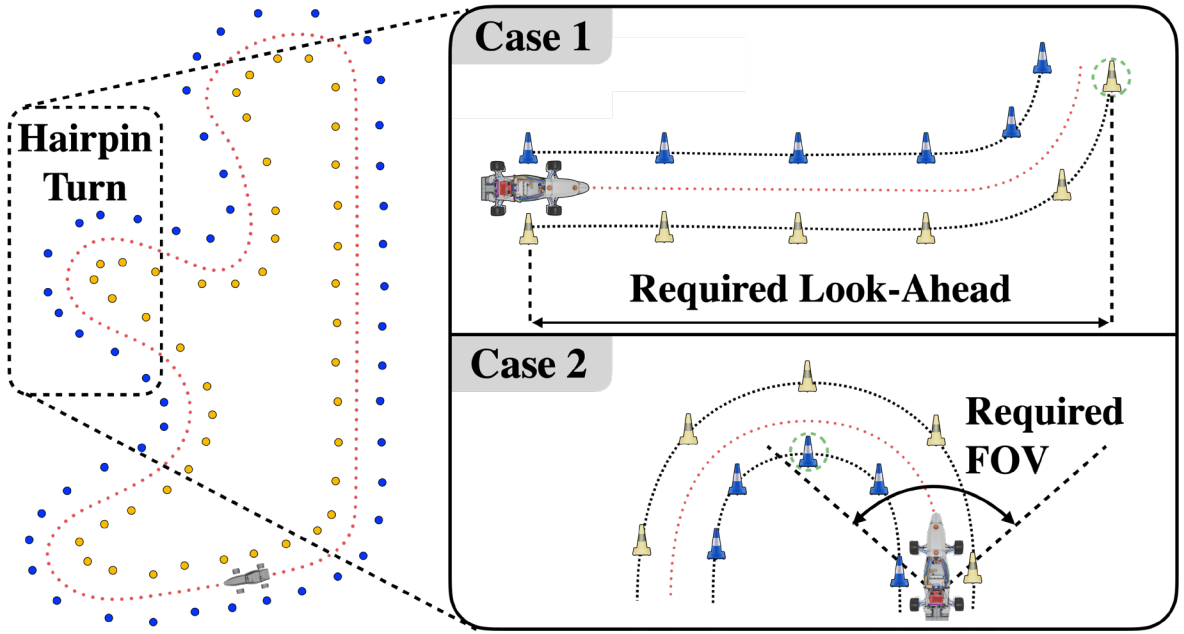


Wide Angle Camera

VIO Camera

Stereovision Pair

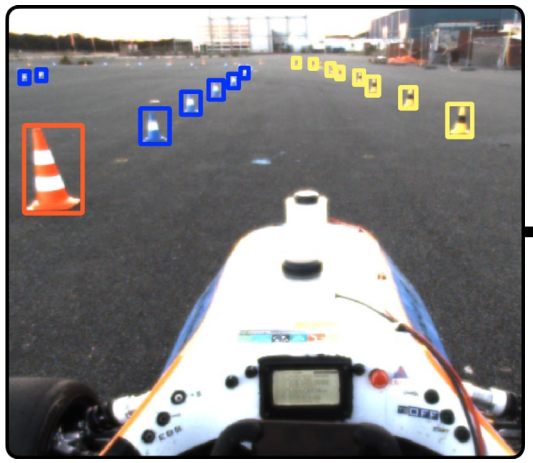
System Requirement



- *Latency*
  - Maximum view-to-actuation time for emergency stop from top speed during an acceleration run
- *Mapping Accuracy*
  - Driven by downstream mapper
- *Horizontal Field-of-View (FOV)*
  - perceive landmarks on the inside apex of a hairpin turn
- *Look-ahead Distance*
  - depends on the full-stack-latency and vehicle deceleration rate

Depth Estimation

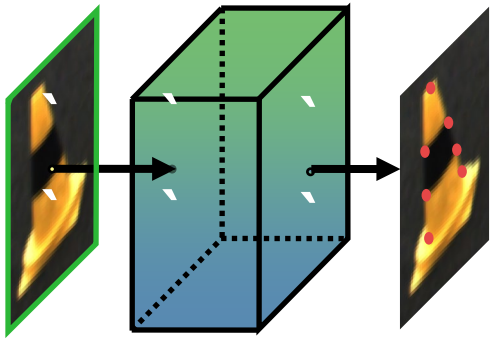
Step 1



Object Detection

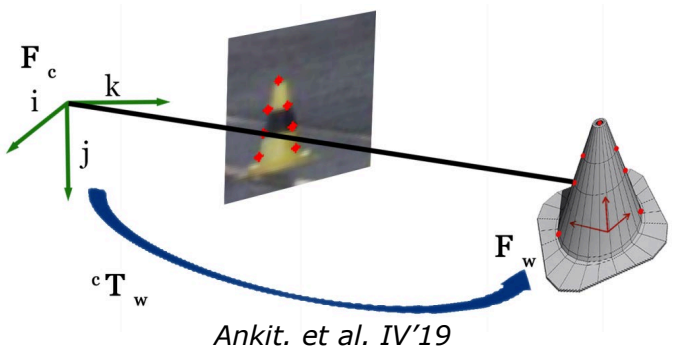
Mono

Step2



Keypoints Detection

Step3

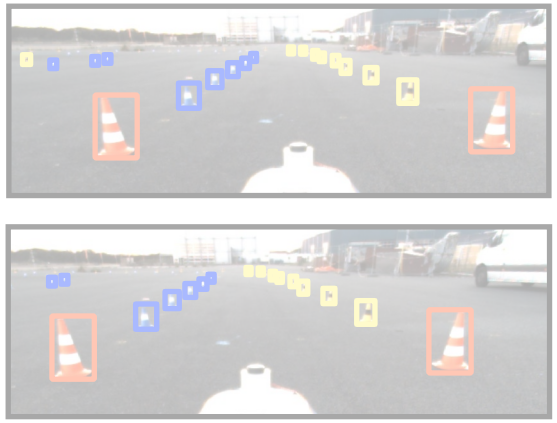


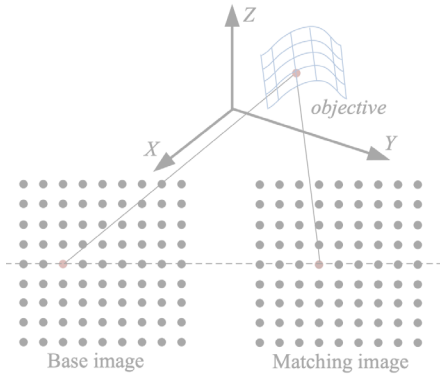
Ankit. et al. IV'19

Perspective-n-Point (PnP)

Stereo

Step2





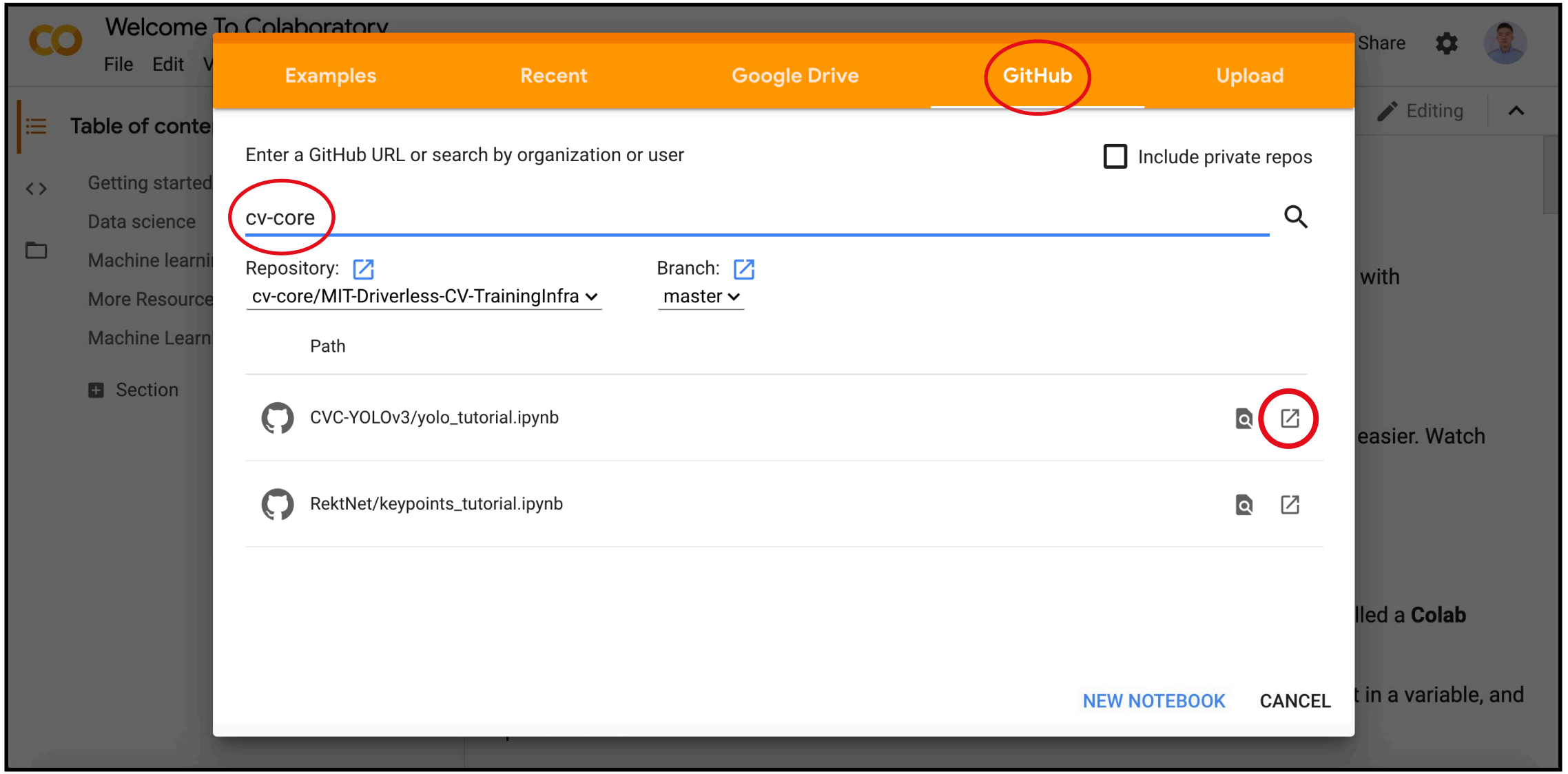
Stereo Matching Algorithm

AI

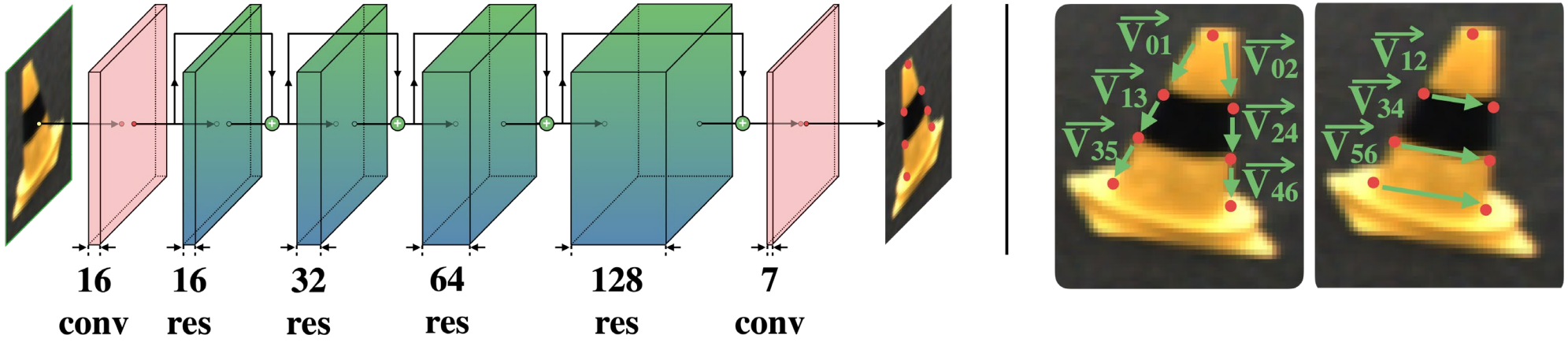
Hands-on Tutorial To Train Your Own Cone Detection Network

colab.research.google.com

Colab Tutorial : colab.research.google.com



Software Design - Keypoints Detection

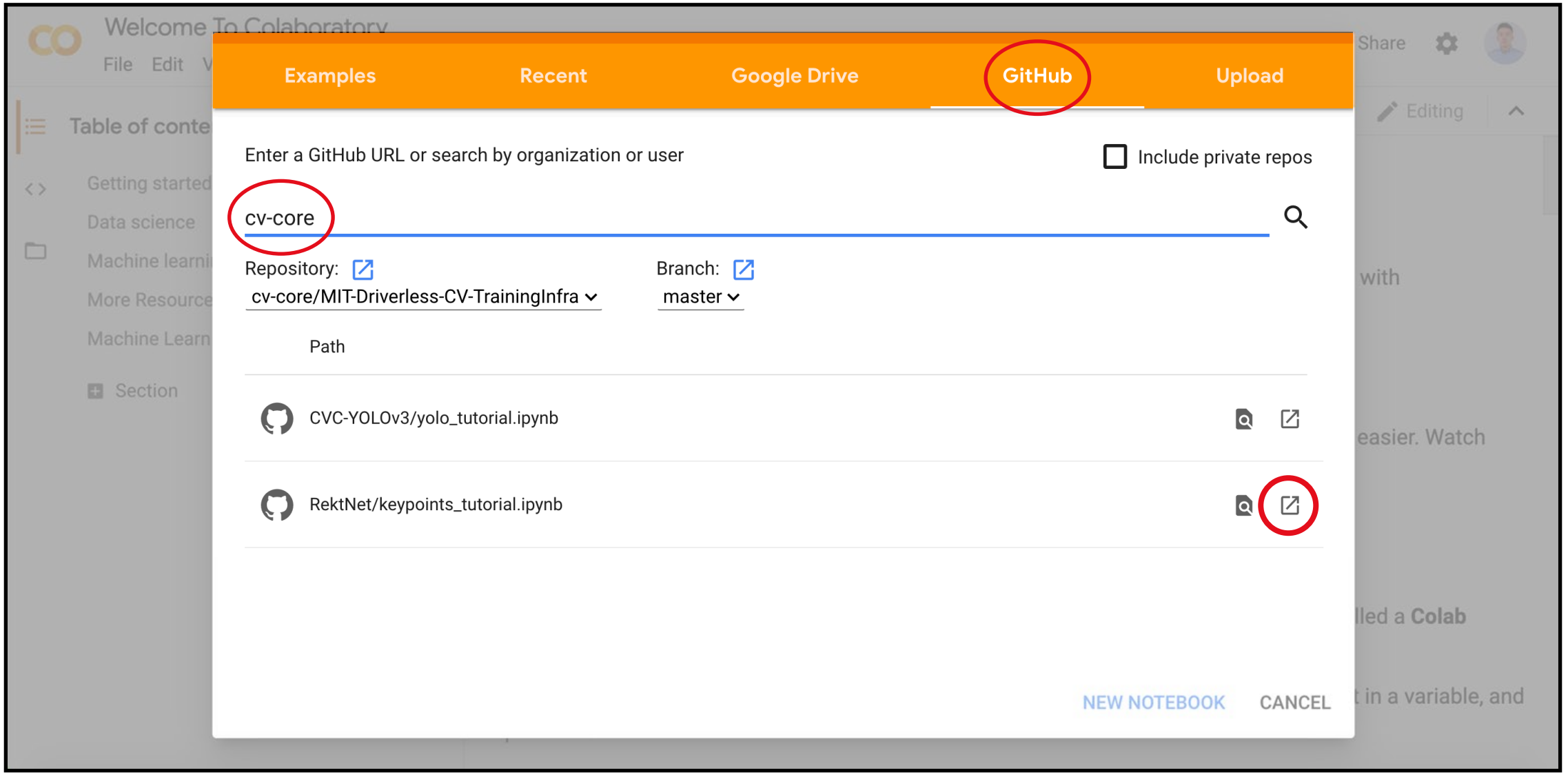


- Detects seven keypoints on each YOLOv3 detection
- A residual NN that leverages the geometric relationship between keypoints
- Seven detected key points will be then used in a Perspective-n-Point (PnP) to get depth

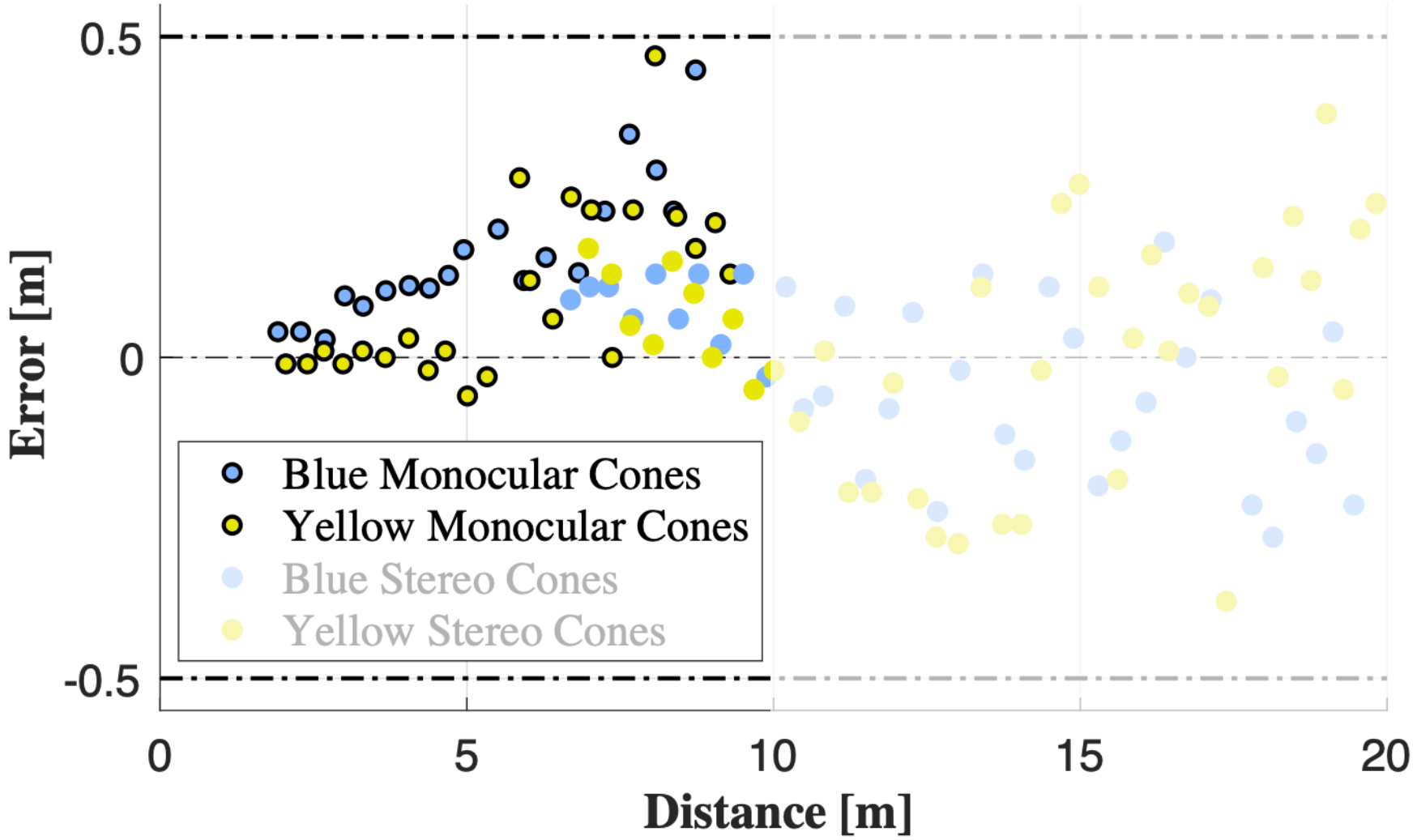
Hands-on Tutorial To Train Your Own Keypoints Detection Network

colab.research.google.com

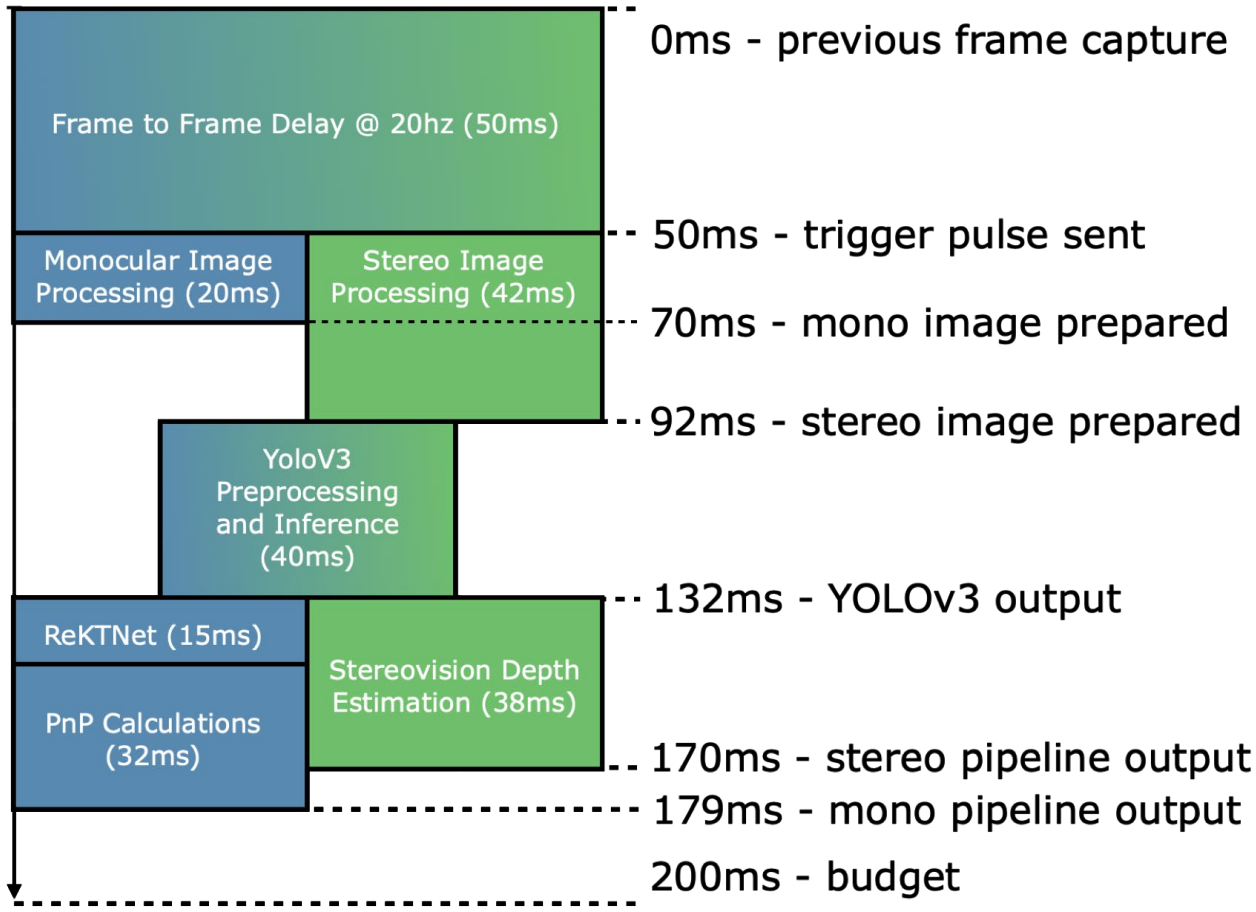
Colab Tutorial : colab.research.google.com



Validation - Accuracy



Validation - Latency



Open Sourced here: github.com/cv-core

Beyond This Computer Vision Tutorial...

- Change the object detection backbone from YOLOv3 to YOLOv4/ EfficientNet/etc
- Adding temporal information for more stable and accurate detection
  - Temporal Shift Module: hanlab.mit.edu/projects/tsm/
- Inference with TensorRT in C++
  - Open sourced here: github.com/cv-core
- Prune the full YOLO architecture for cone detection task
- Quantization (int8) for even faster inference

LiDAR Perception

Thank you for your attention

Hands-on Intro to Developing Vision and LiDAR Classifiers

Formula Student Driverless Workshop

powered by

Speaker Introduction

Sibo Zhu

- RA at MIT HAN Lab
- Perception Lead at MIT Driverless

Zhijian Liu

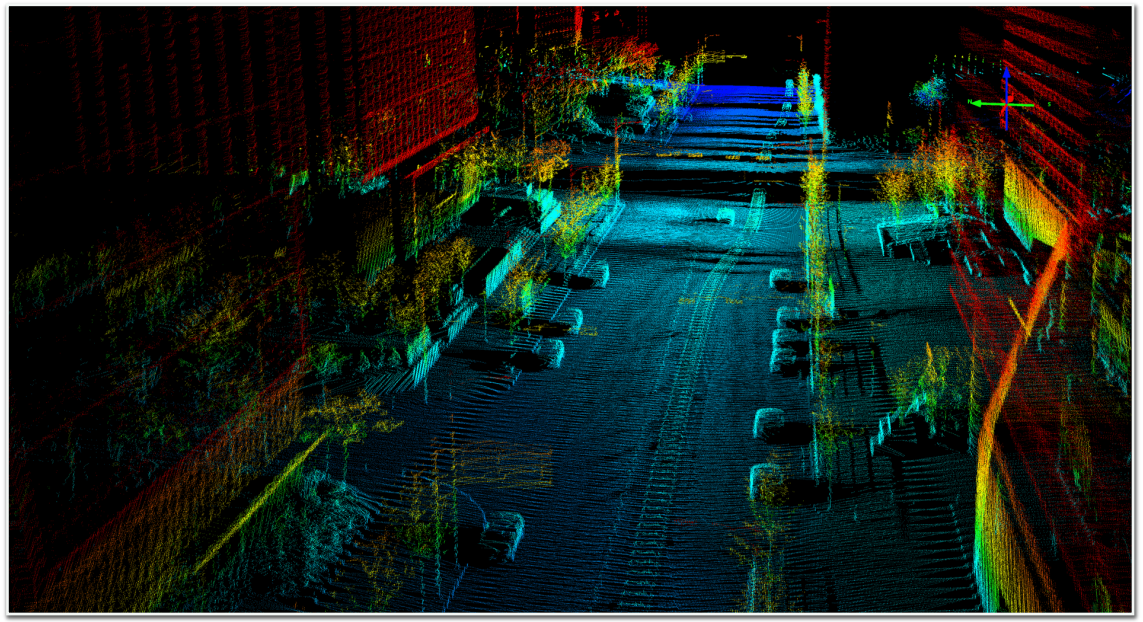
- PhD student at MIT
- Perception Lead at MIT Driverless

Haotian Tang

- PhD student at MIT
- Perception Lead at MIT Driverless

LiDAR Perception

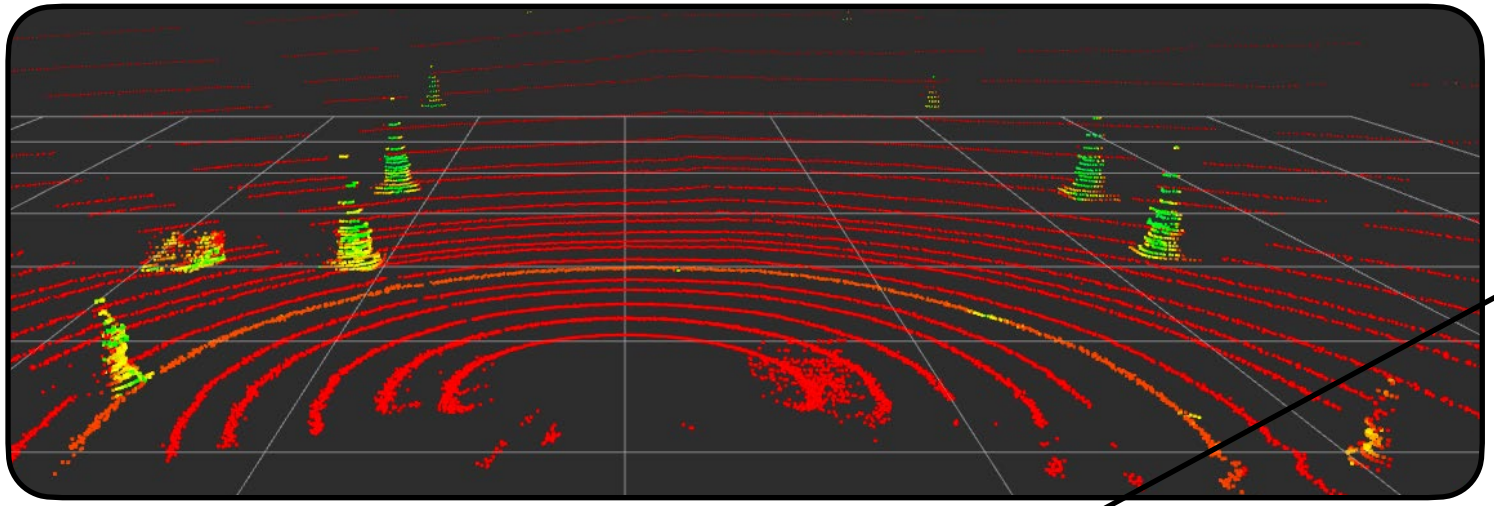
3D LiDAR Sensor

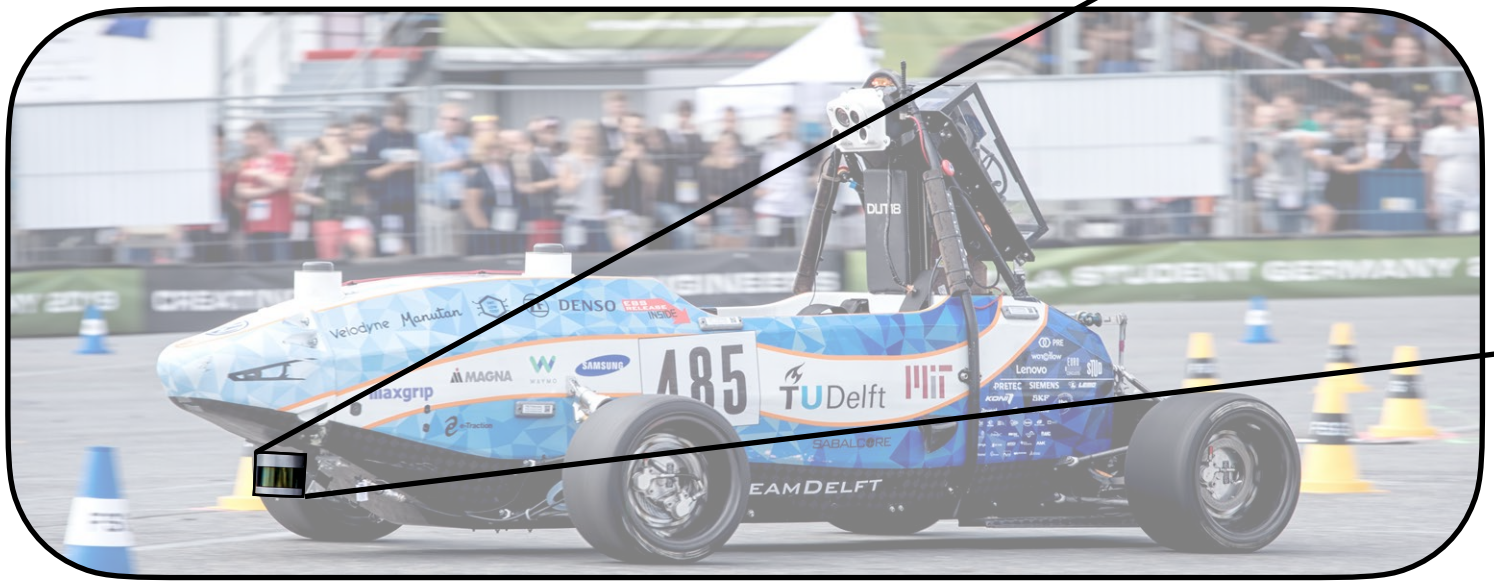


3D Point Cloud

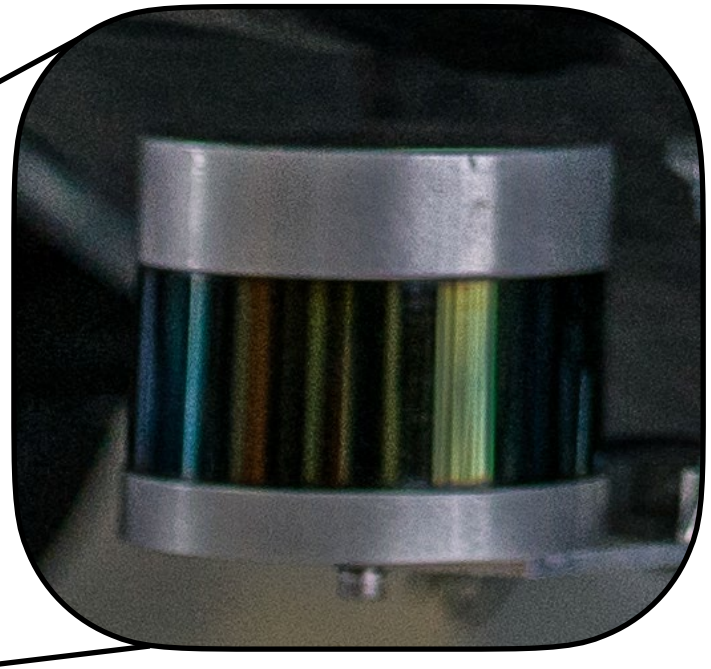
500k+ points: (x, y, z, intensity)

LiDAR Perception Task





Velodyne 32C LiDAR



Autonomous Racing Vehicle: Objectives

Low Latency
(Drive Faster)

High Accuracy
(Prevent Collisions)

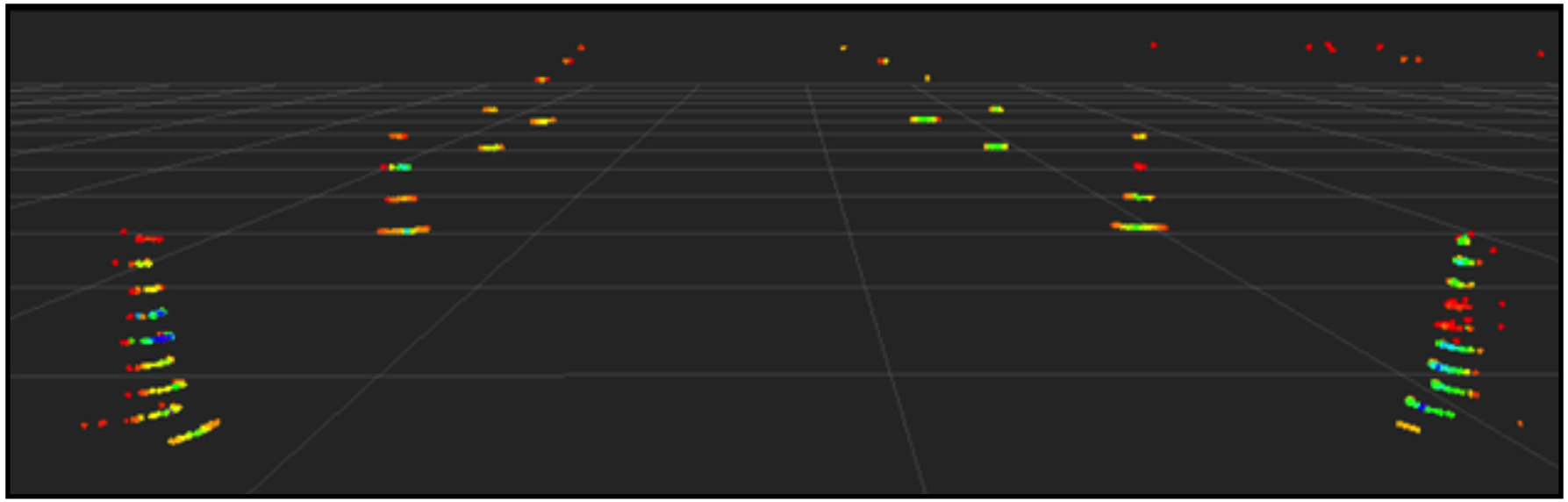
Autonomous Racing Vehicle: Challenges

Self-Driving Cars

A whole trunk of computers!

We need more efficient algorithms that do not consume intensive computations.

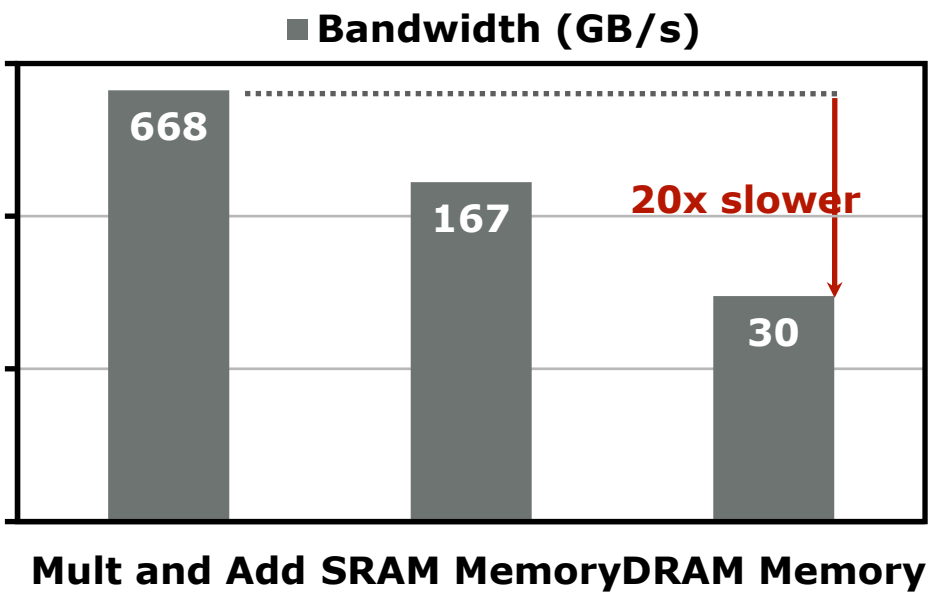
Autonomous Racing Vehicle: Challenges



Longer Distance	→ Fewer Laser Rings on Objects	→ Fewer Laser Points
Shorter Distance	→ Too many Laser Rings on Objects	→ Too many Laser Points

Author: Sibo Zhu, Zhijian Liu, Haotian Tang

Efficient LiDAR Perception: Bottleneck



Sequential Memory Access

1	2	3	4	5	6	7	8

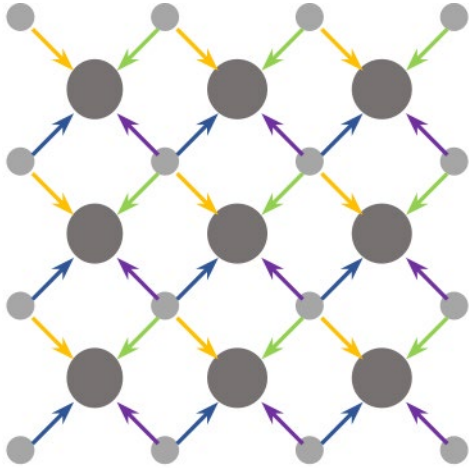
Random Memory Access

7	5	2	4	6	1	8	3

Off-chip DRAM access is much more expensive than arithmetic operation!

Random memory access is inefficient due to the potential bank conflicts!

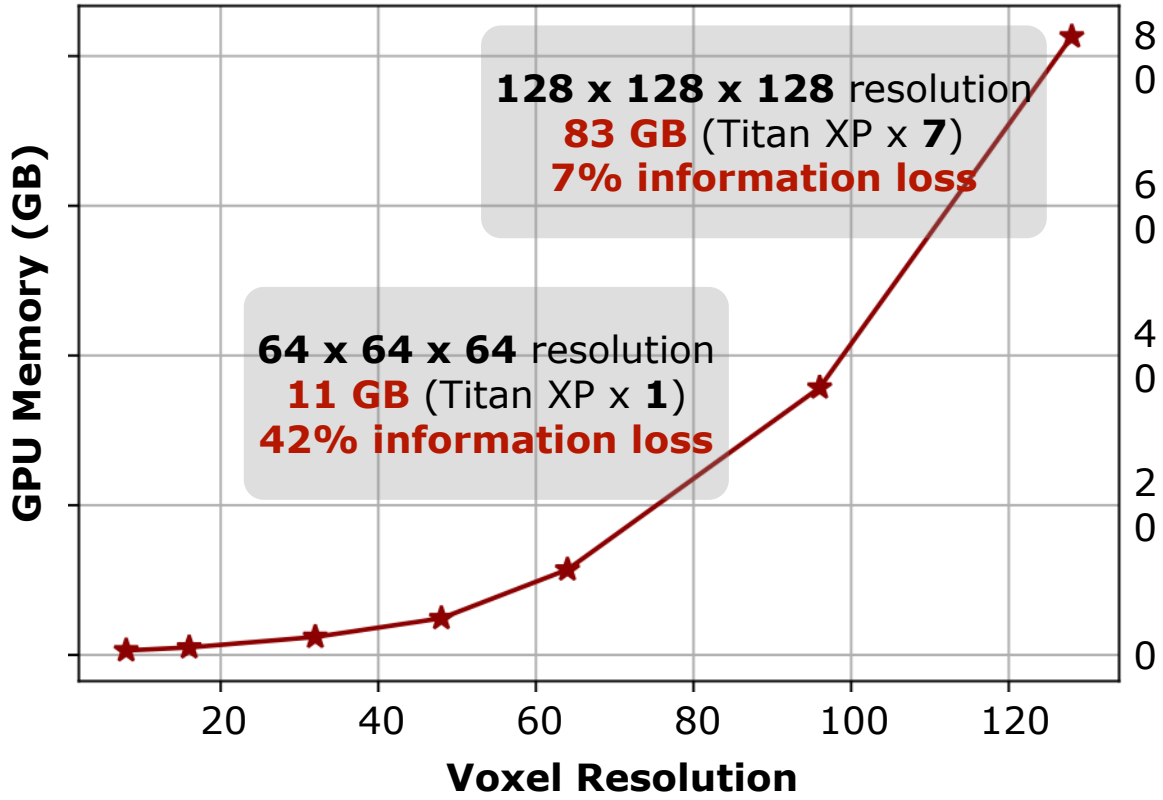
Voxel-Based Models: Cubically-Growing Memory



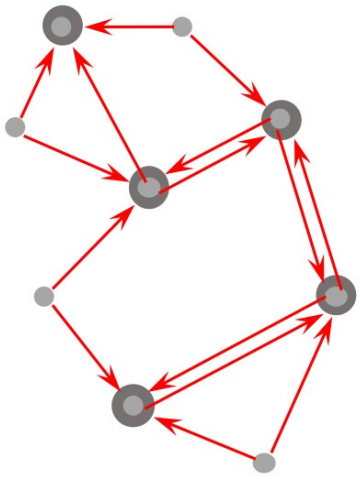
3D ShapeNets [CVPR'15]

VoxNet [IROS'15]

3D U-Net [MICCAI'16]



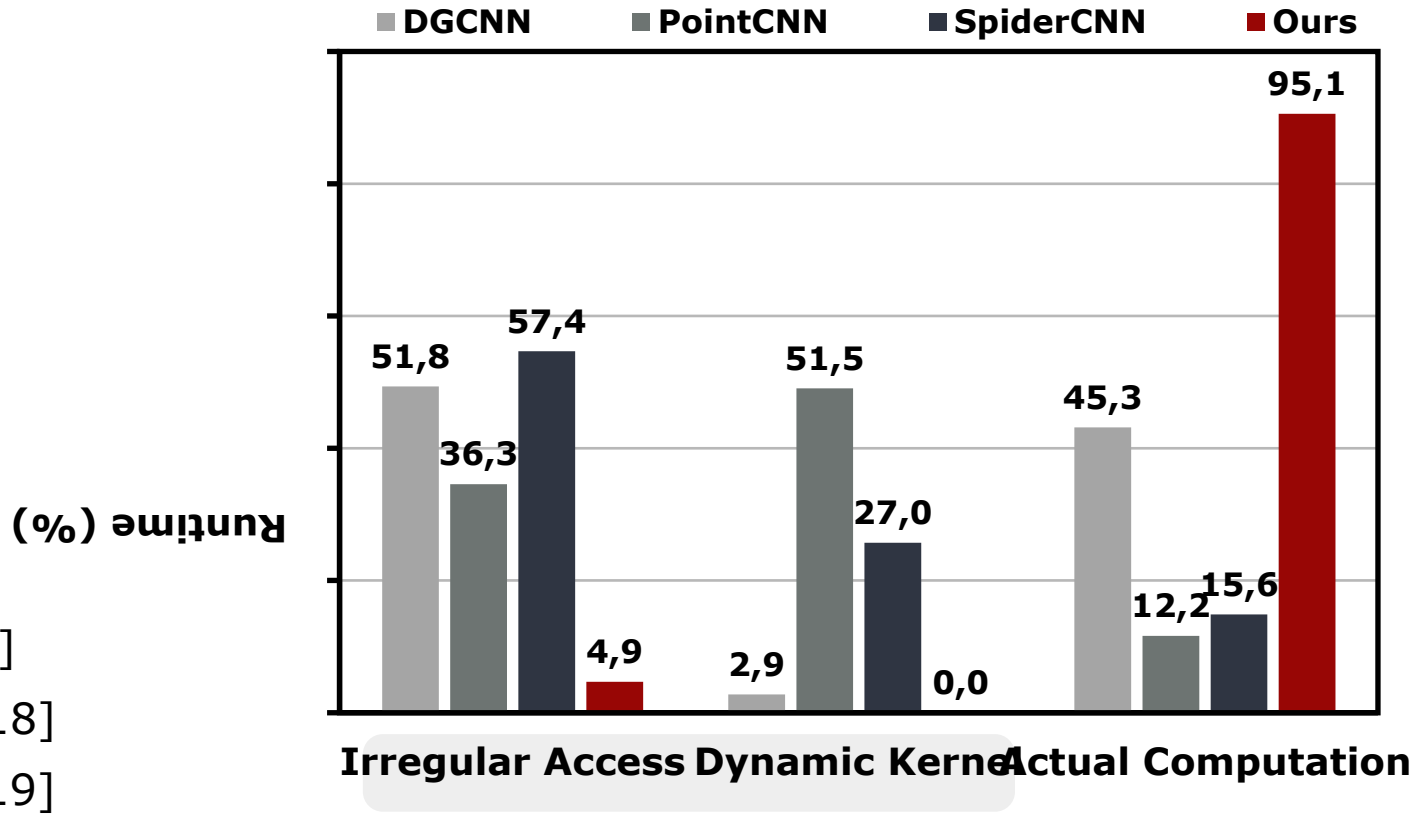
Point-Based Models: Sparsity Overheads



PointNet [CVPR'17]

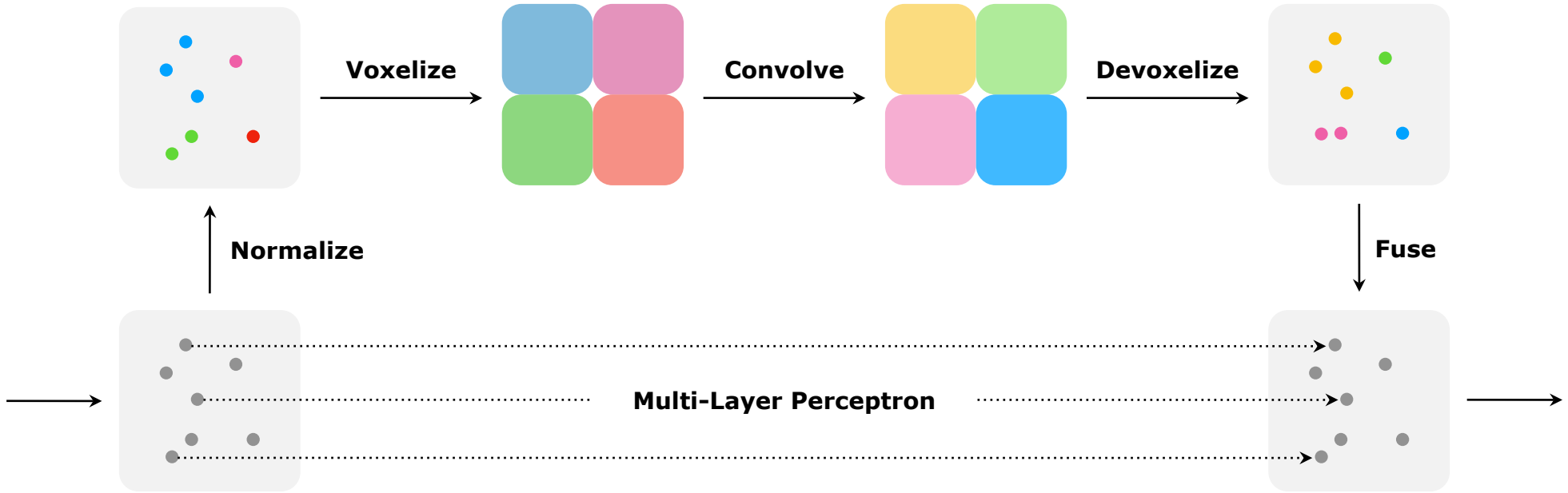
PointCNN [NeurIPS'18]

DGCNN [SIGGRAPH'19]



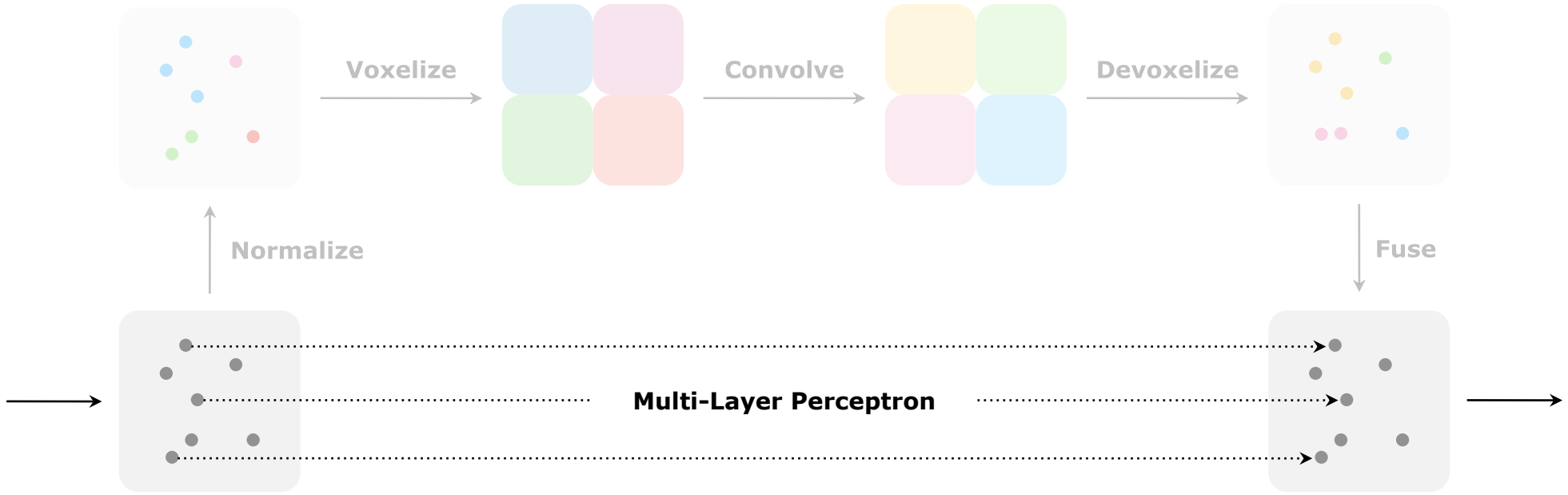
Author: Sibo Zhu, Zhijian Liu, Haotian Tang

Point-Voxel Convolution (PVConv)



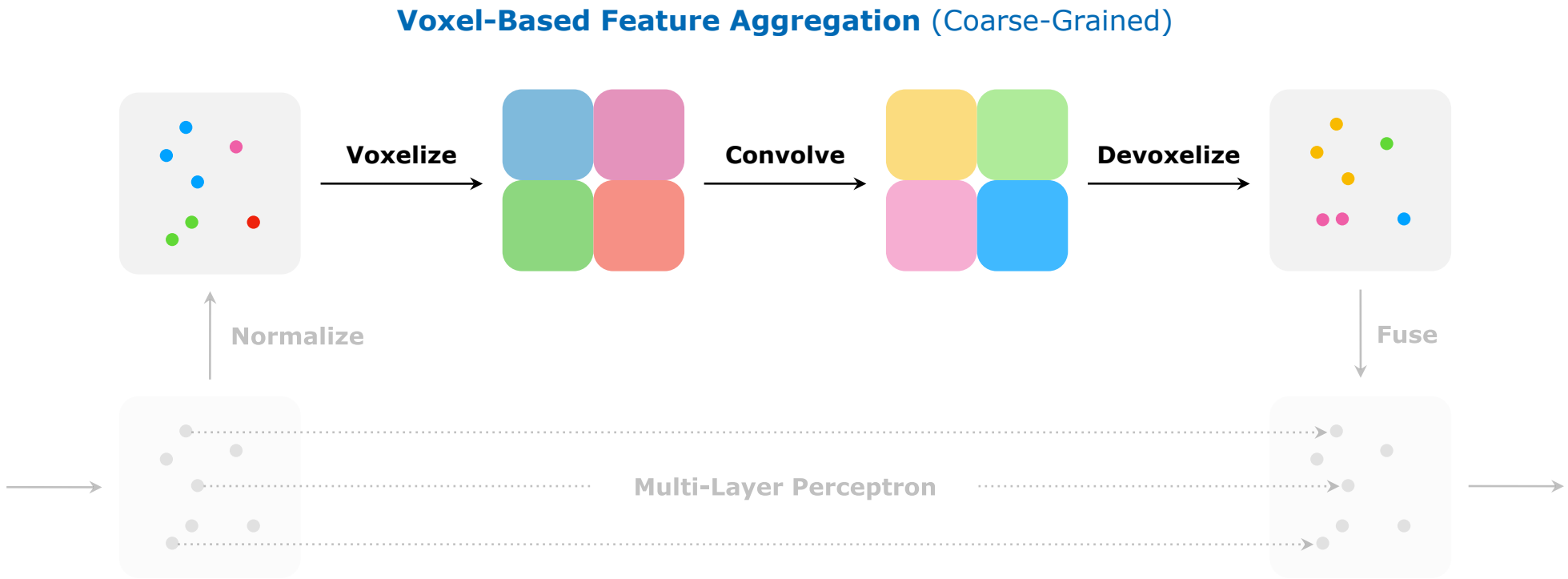
Author: Sibo Zhu, Zhijian Liu, Haotian Tang

Point-Voxel Convolution (PVConv)



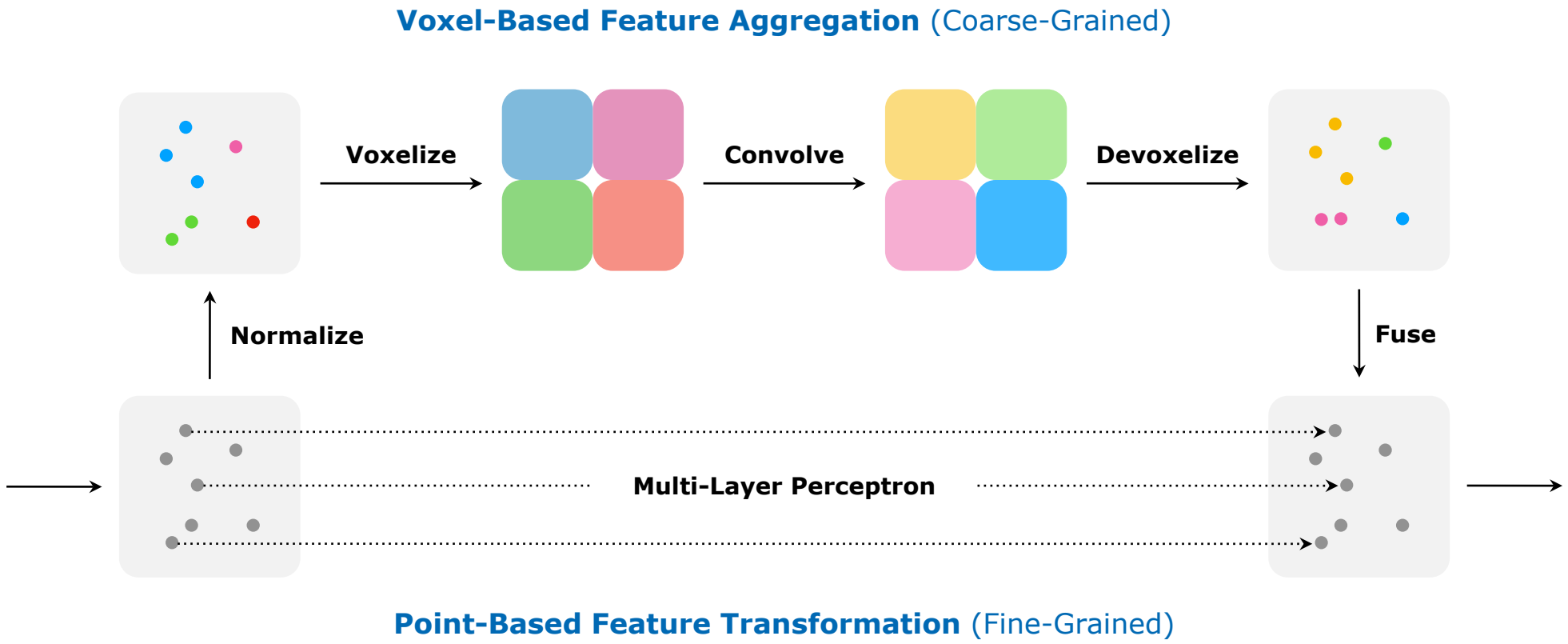
Point-Based Feature Transformation (Fine-Grained)

Point-Voxel Convolution (PVConv)



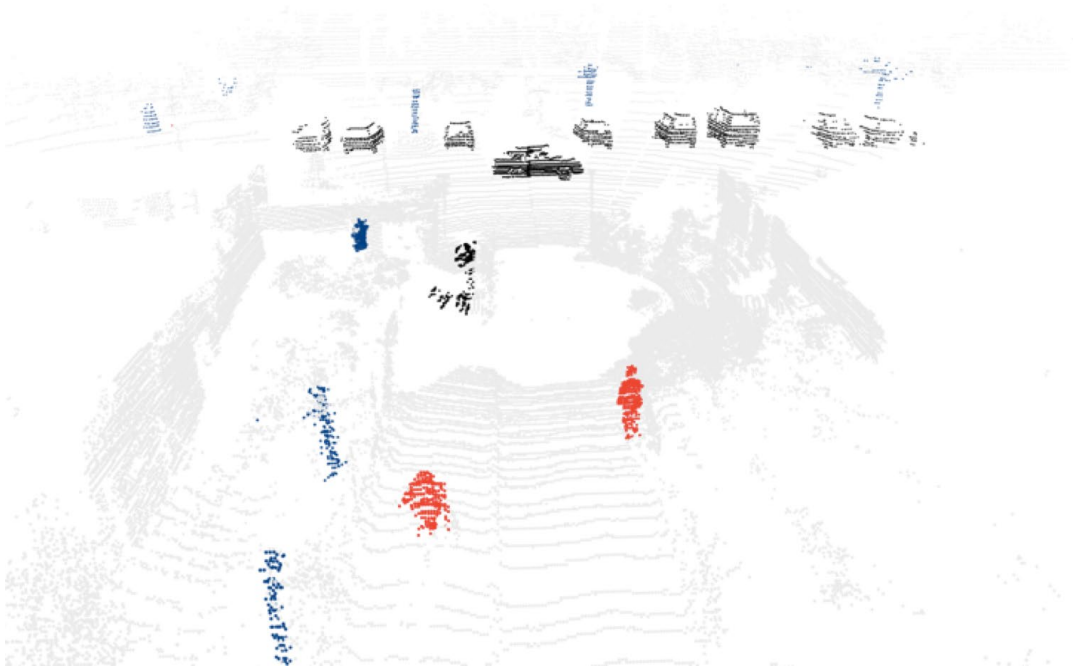
Author: Sibo Zhu, Zhijian Liu, Haotian Tang

Point-Voxel Convolution (PVConv)

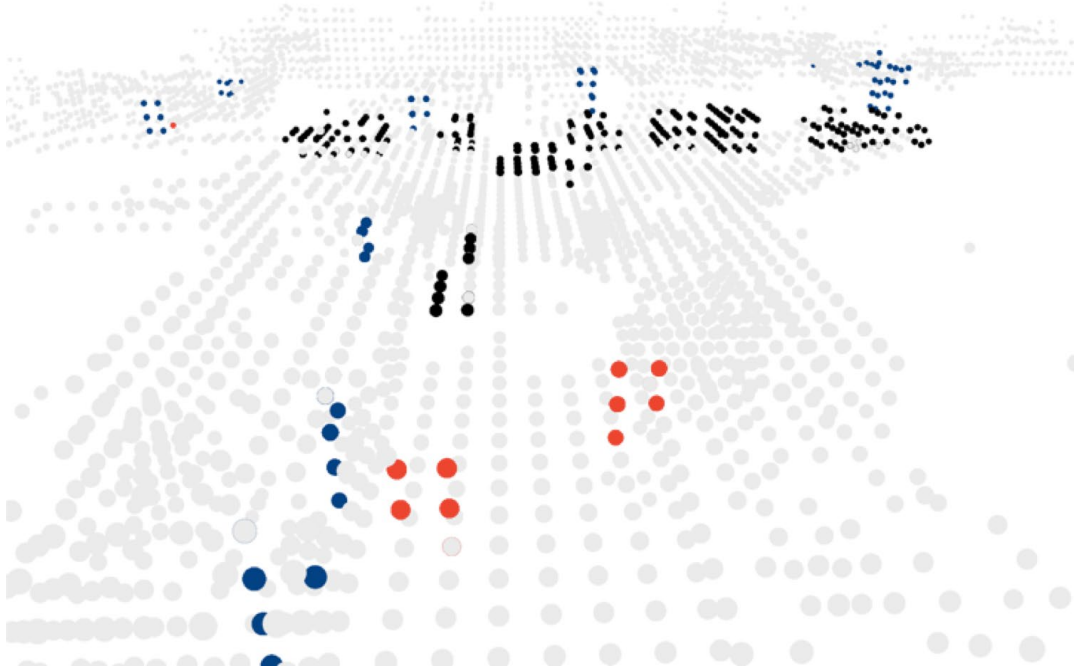


Author: Sibo Zhu, Zhijian Liu, Haotian Tang

Low Resolution with Constrained Memory

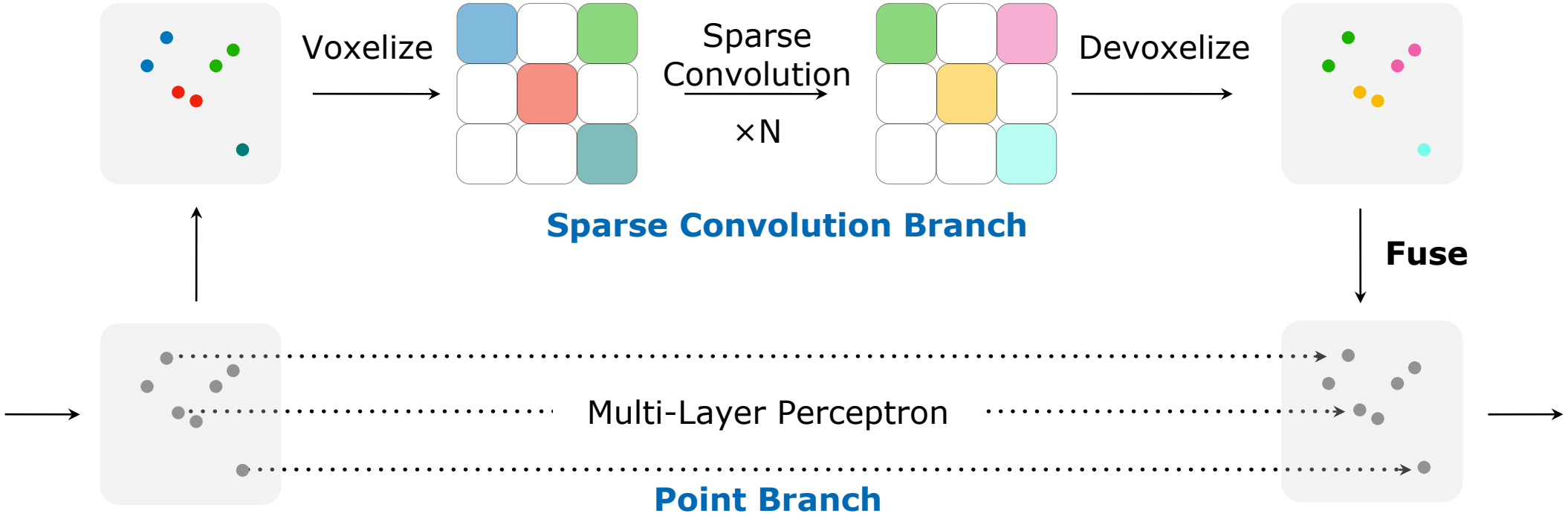


Original Scene



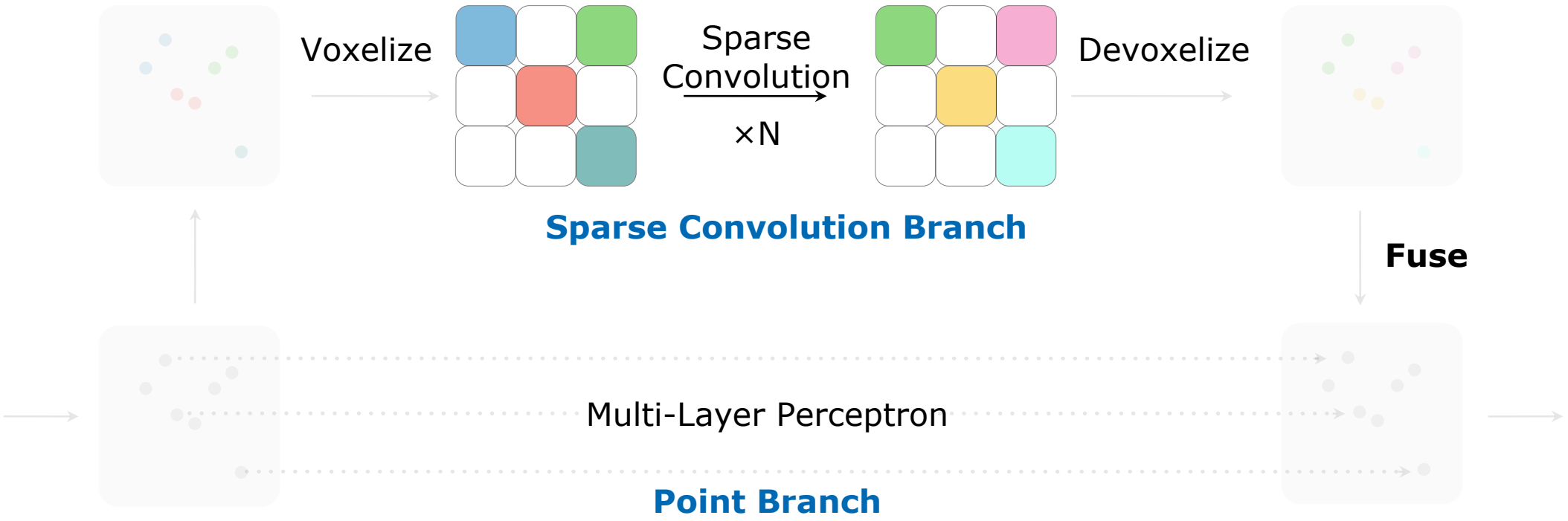
Downsampled Scene

Sparse Point-Voxel Convolution (SPVConv)



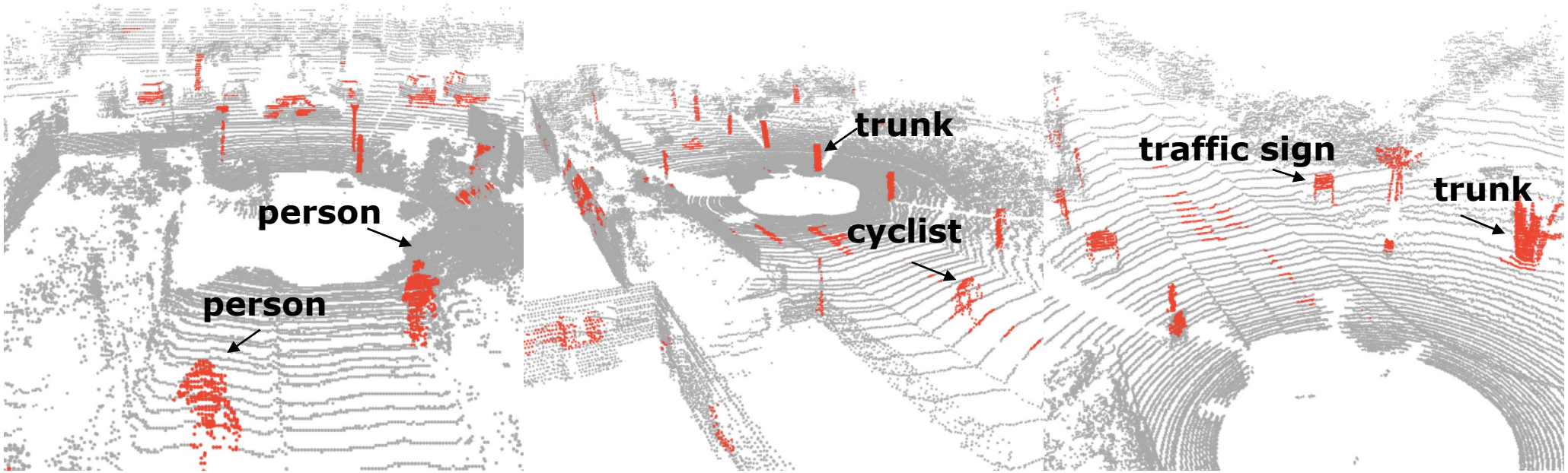
Author: Sibo Zhu, Zhijian Liu, Haotian Tang

Sparse Point-Voxel Convolution (SPVConv)

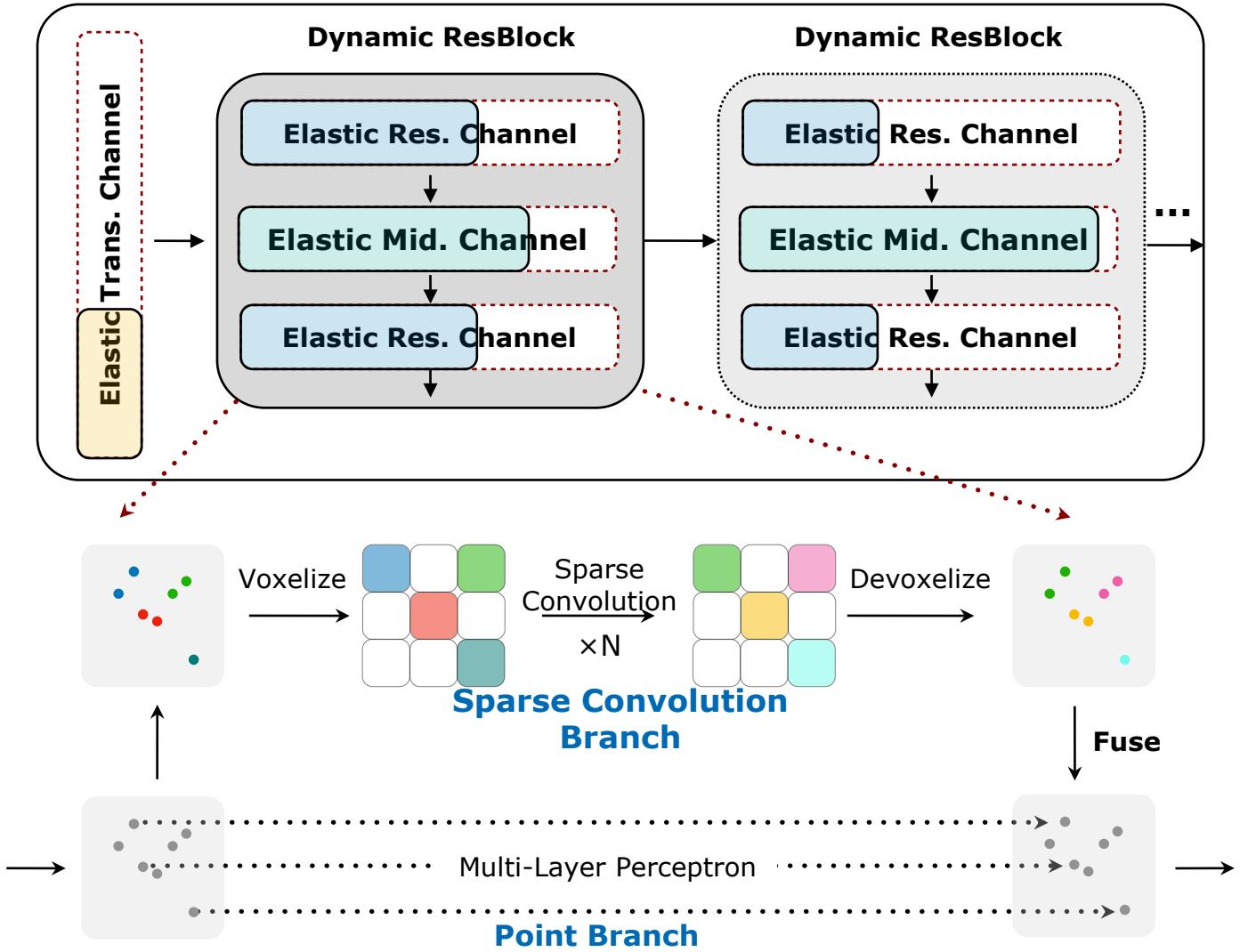


Author: Sibo Zhu, Zhijian Liu, Haotian Tang

Designing Efficient 3D Modules (SPVConv)

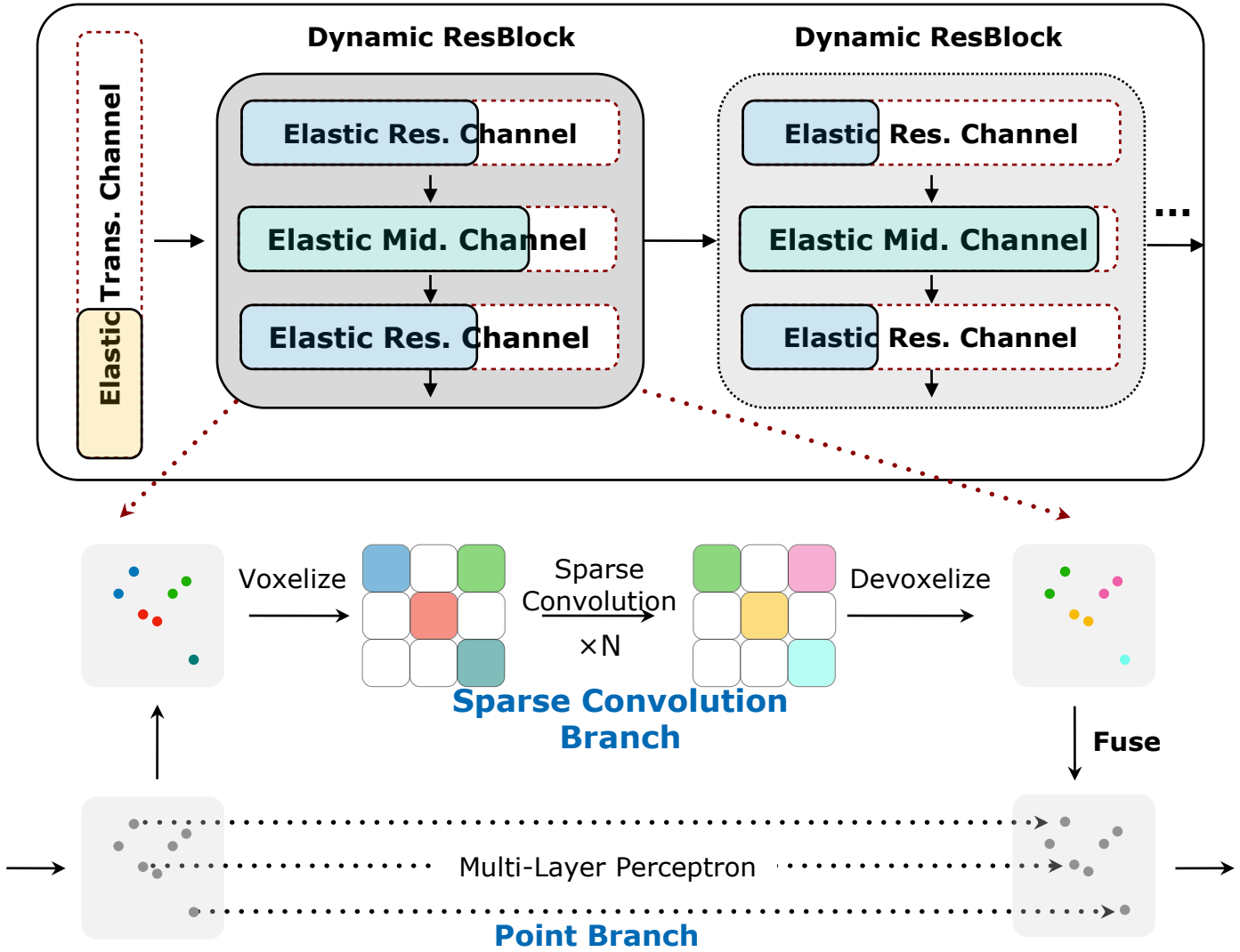


Searching Efficient 3D Architectures (3D-NAS)



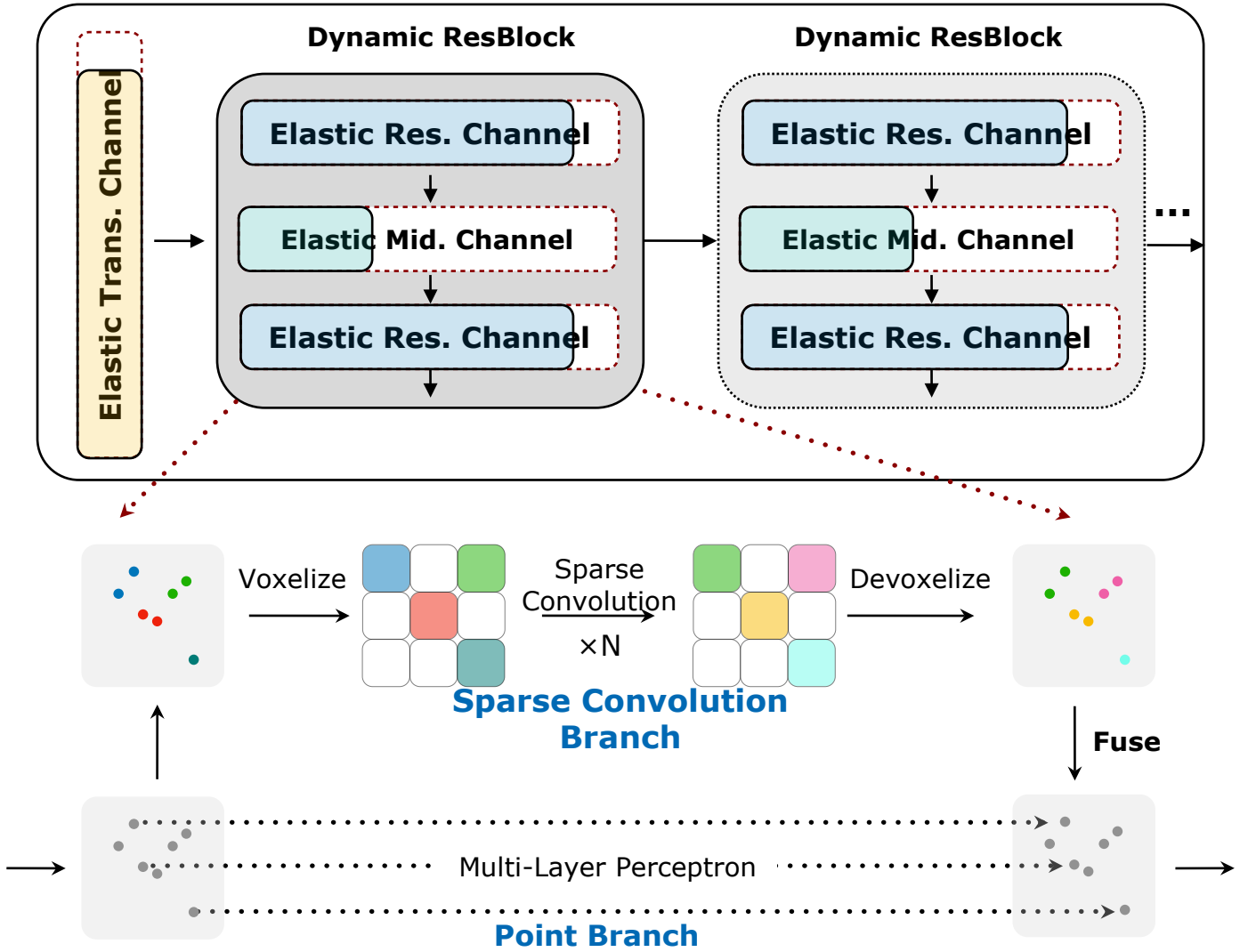
Author: Sibo Zhu, Zhijian Liu, Haotian Tang

Searching Efficient 3D Architectures (3D-NAS)



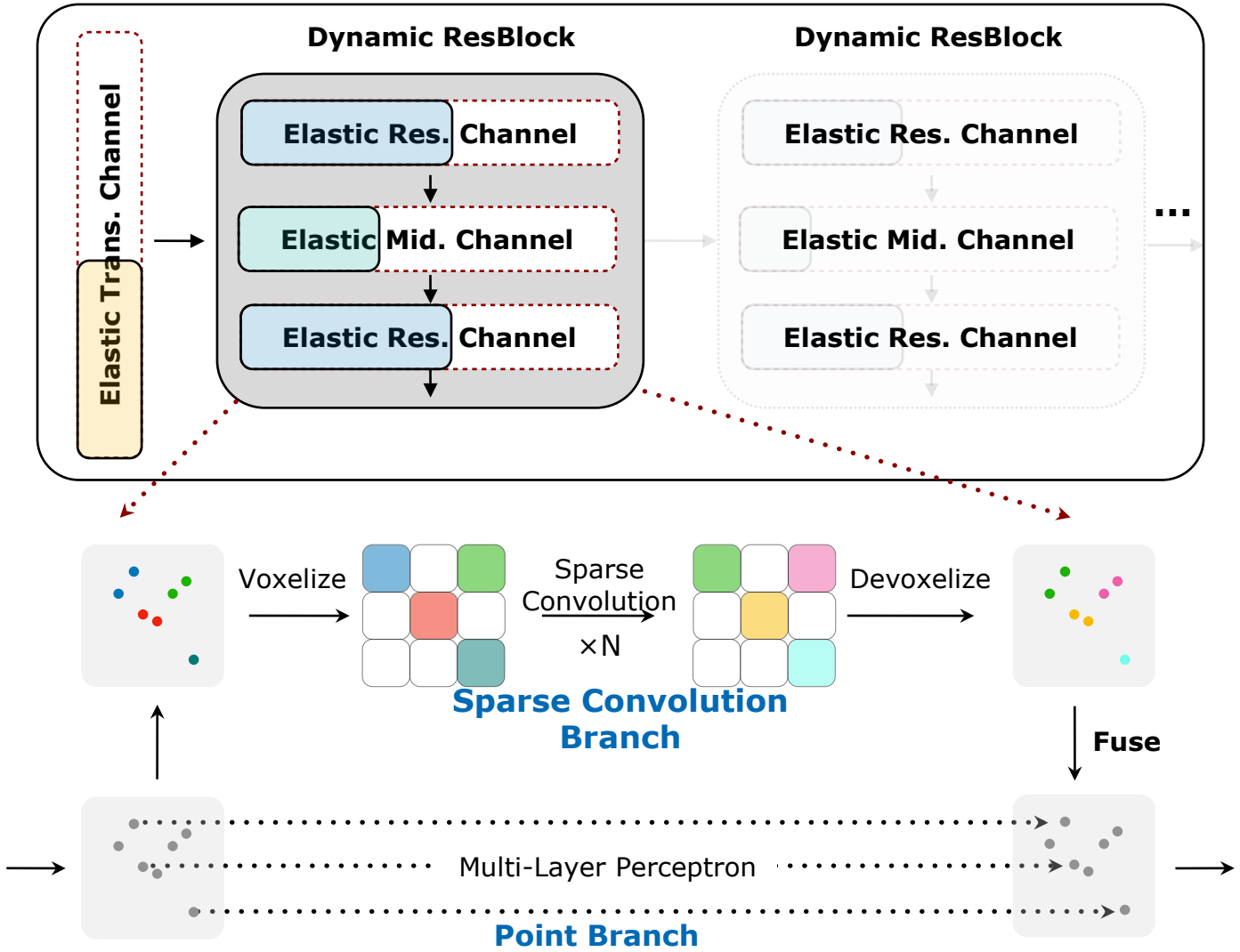
Author: Sibo Zhu, Zhijian Liu, Haotian Tang

Searching Efficient 3D Architectures (3D-NAS)



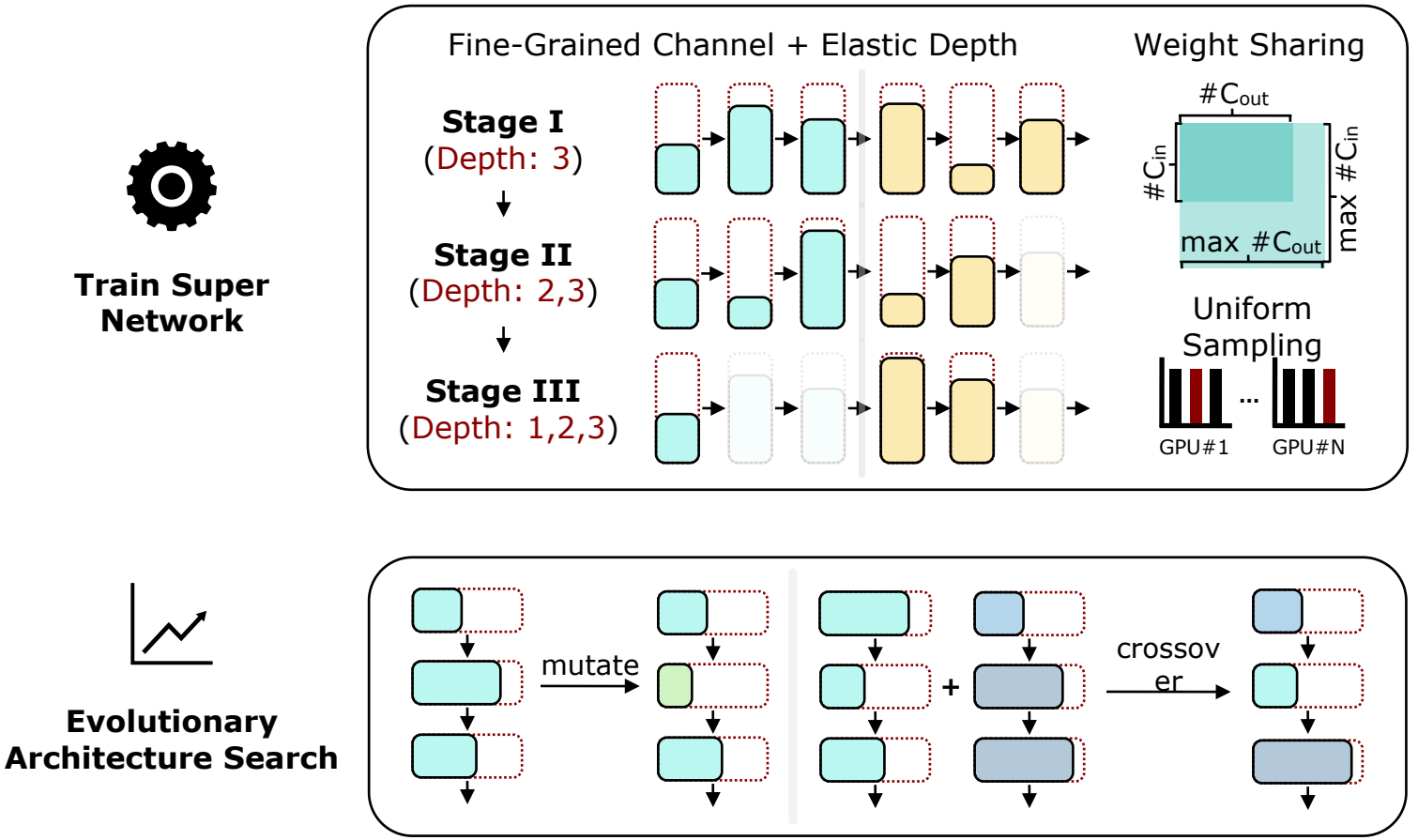
Author: Sibo Zhu, Zhijian Liu, Haotian Tang

Searching Efficient 3D Architectures (3D-NAS)



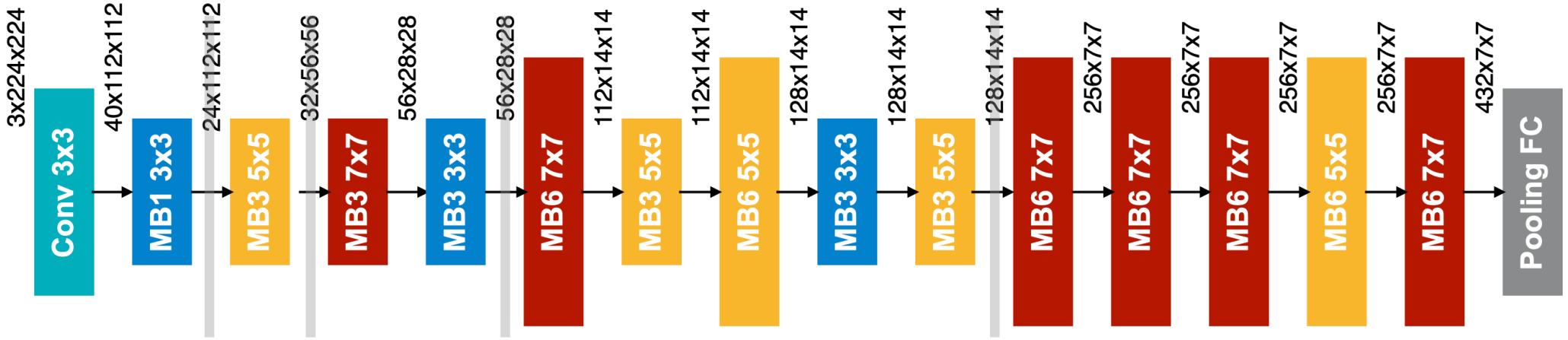
Author: Sibo Zhu, Zhijian Liu, Haotian Tang

Searching Efficient 3D Architectures (3D-NAS)



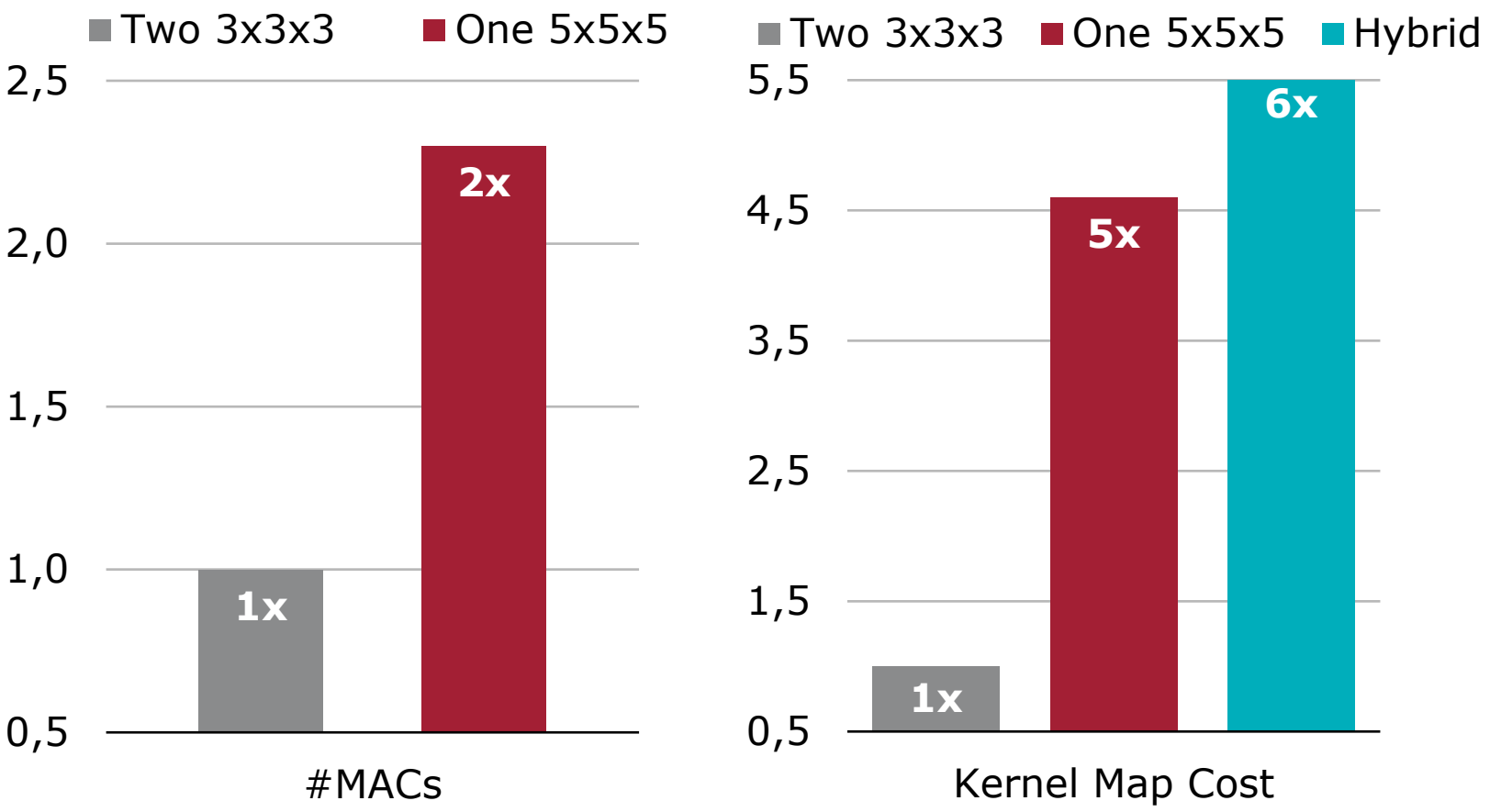
Author: Sibo Zhu, Zhijian Liu, Haotian Tang

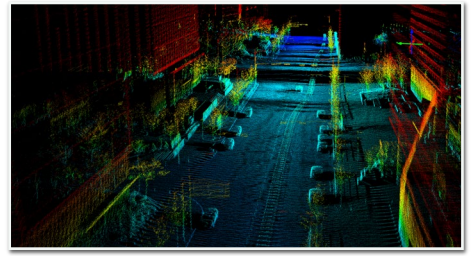
Details: Small Kernel Matters



Large kernels are efficient in 2D-NAS

Details: Small Kernel Matters





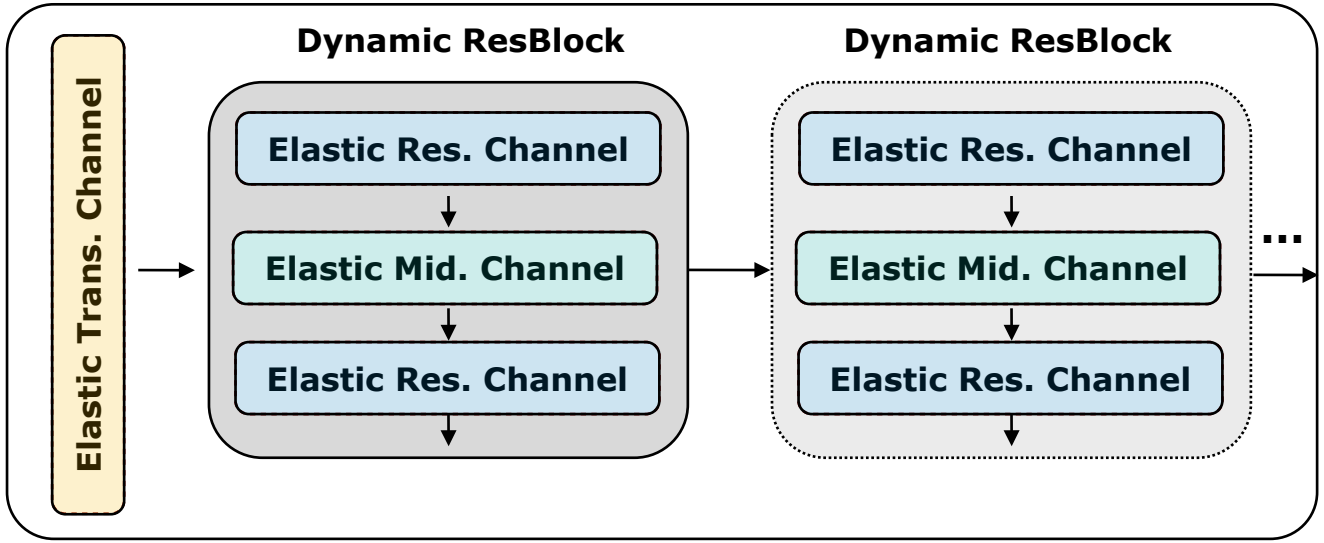
3D Deep Learning

Small Kernels

Large Kernels

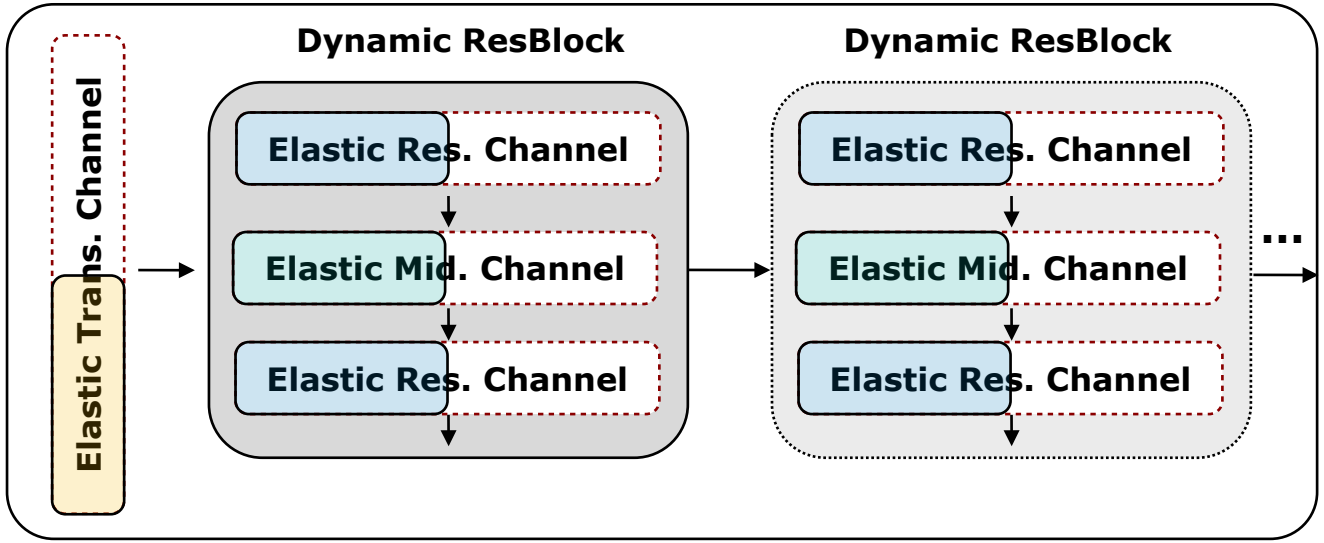
Cost of large kernels in 3D deep learning is more prohibitive than 2D.

Details: Elastic Network Depths



Author: Sibo Zhu, Zhijian Liu, Haotian Tang

Details: Elastic Network Depths

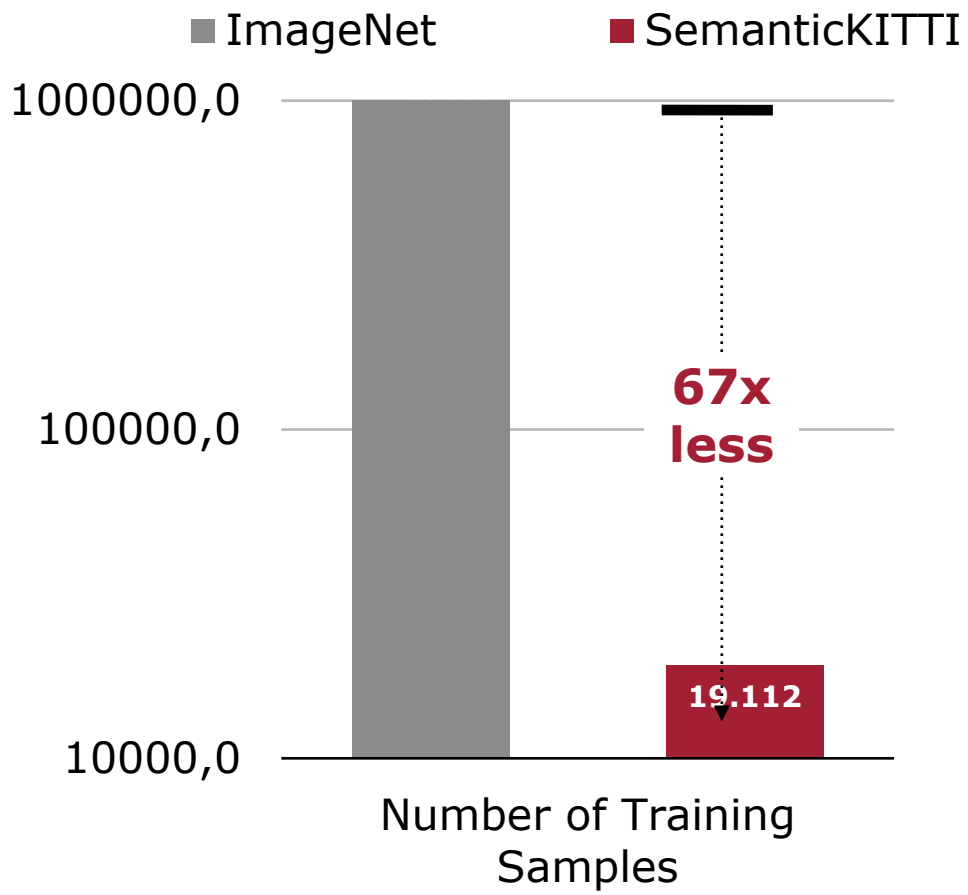


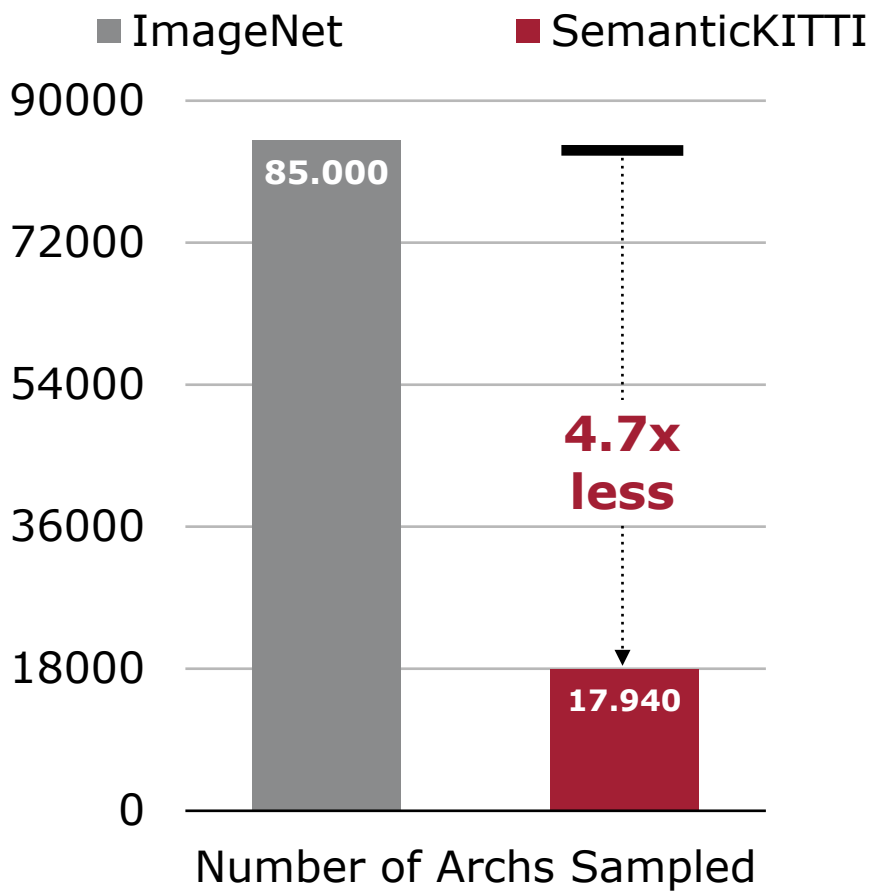
FLOPs: 7.5G → 1.9G(4x)

Latency: 105 ms → 96 ms(1.1x)

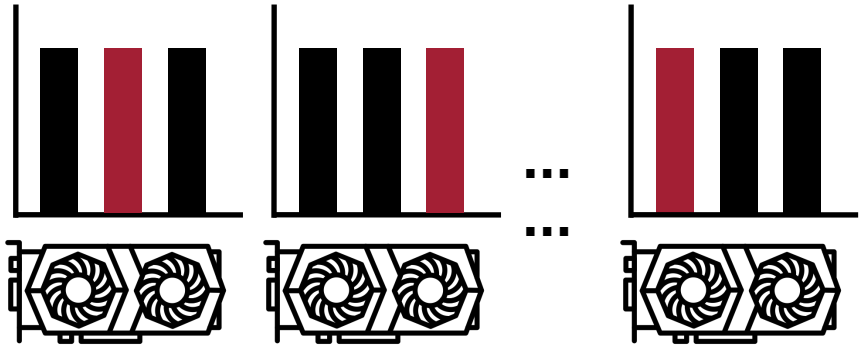
Scaling network channels only cannot result in efficient models.

Challenge: Sample Efficiency



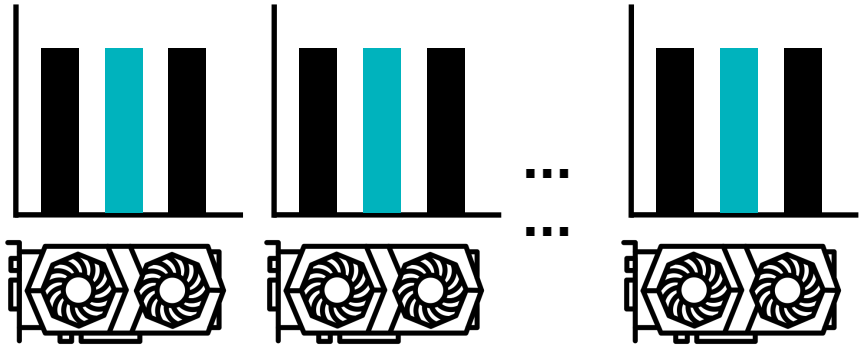


Solution: Distributed Sampling Across GPUs



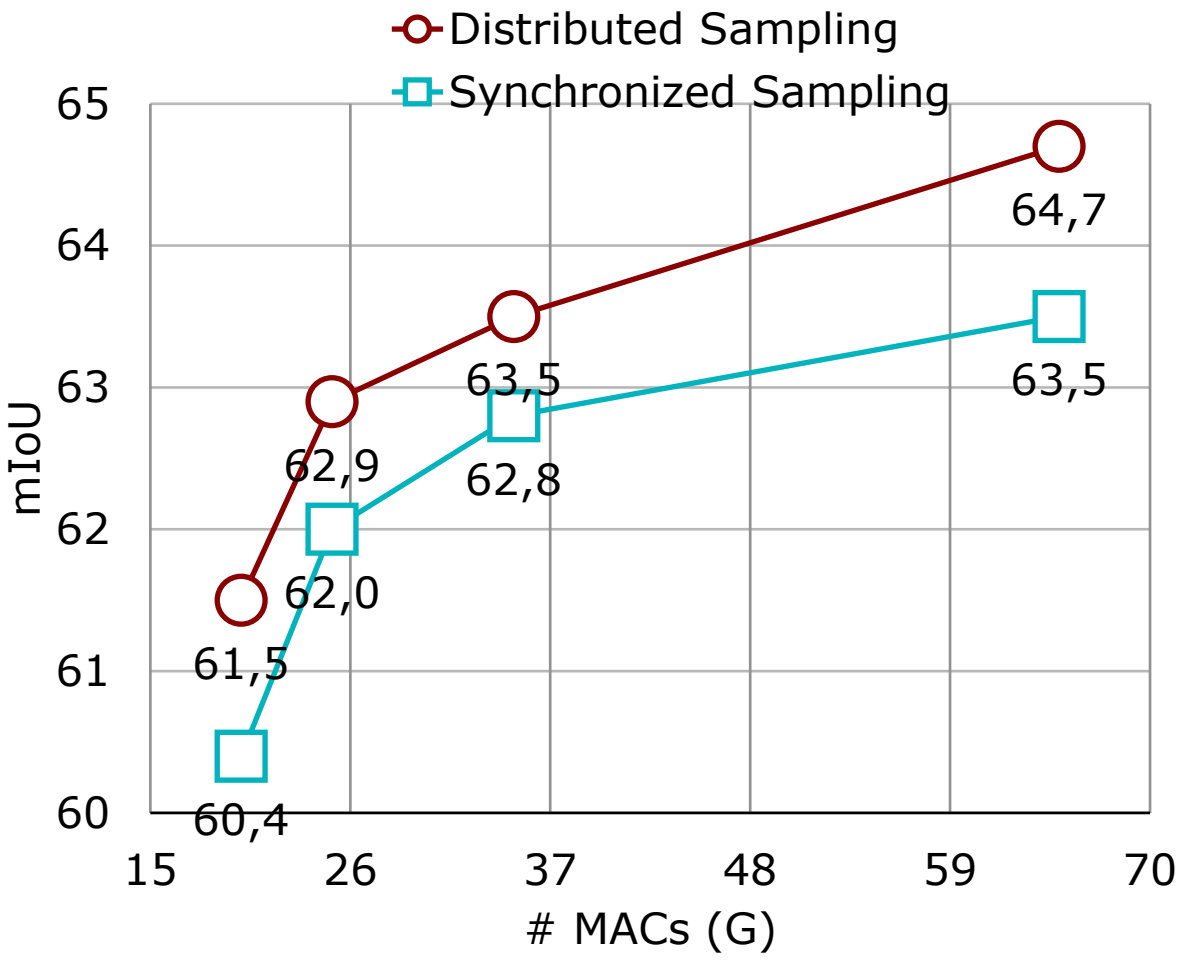
Distributed Sampling

Different sub-networks on different GPUs



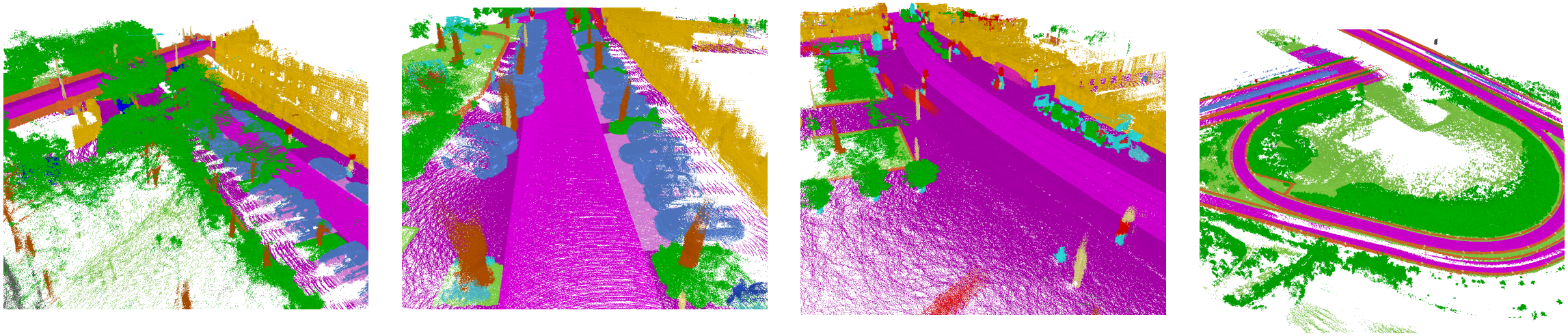
Synchronized Sampling

The same sub-networks on different GPUs



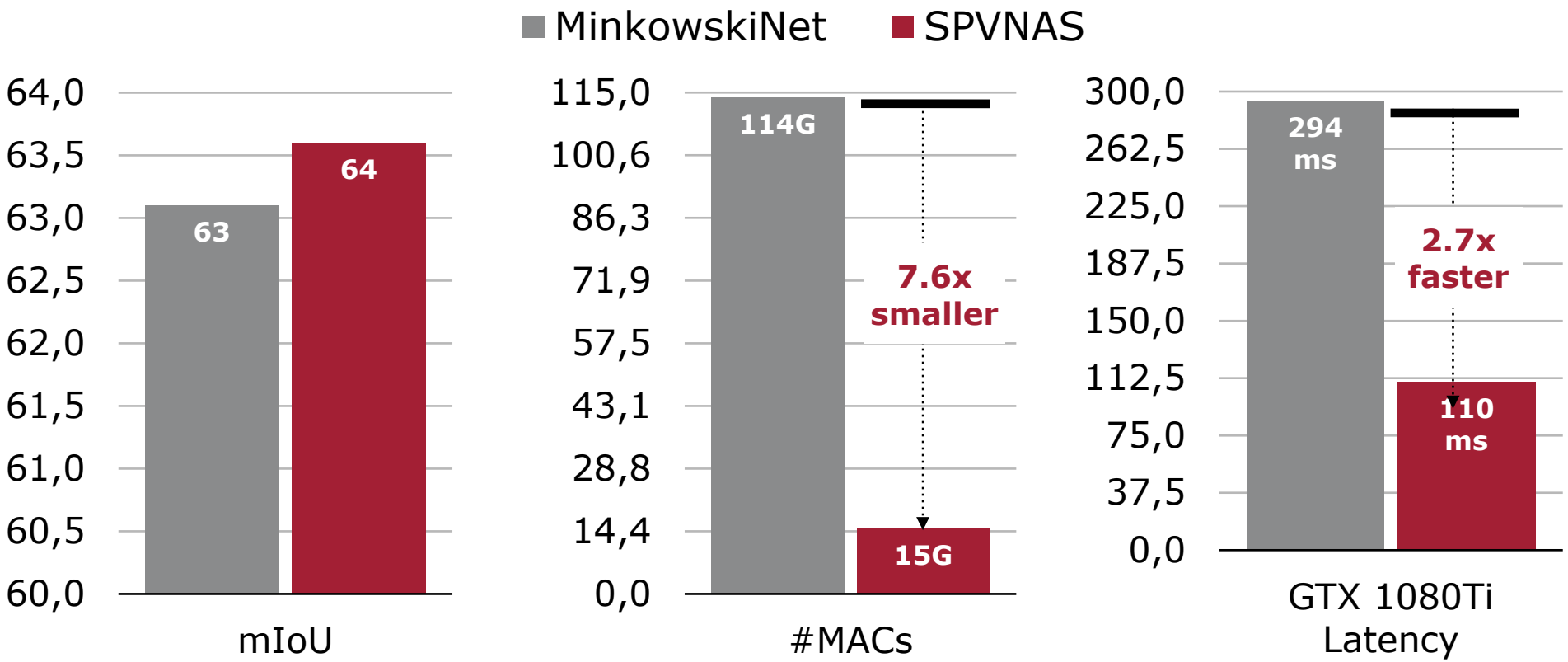
Author: Sibo Zhu, Zhijian Liu, Haotian Tang

Results: 3D Semantic Scene Segmentation



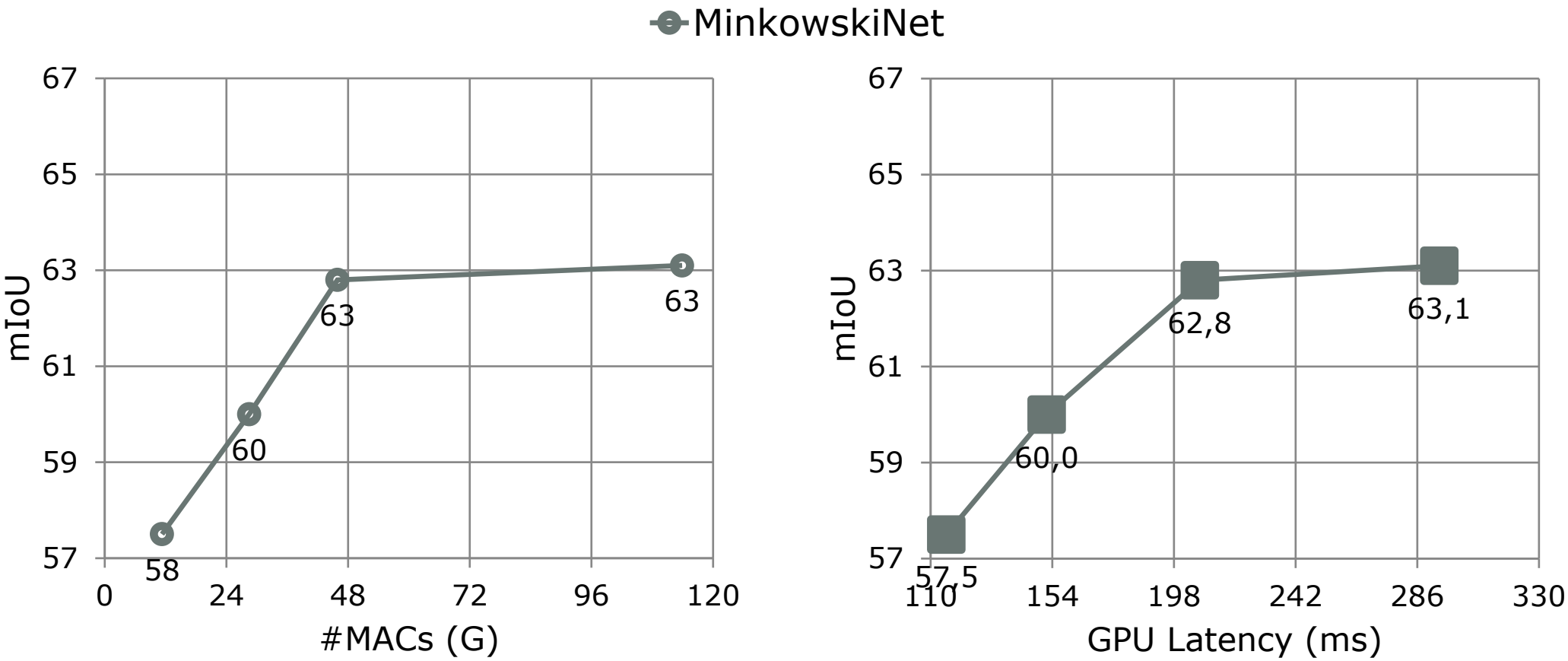
- SemanticKITTI is the **largest** 3D point cloud semantic segmentation dataset. It is **29x** larger than ScanNet, **160x** larger than S3DIS.
- SemanticKITTI is collected from **real driving scenarios**, and provides **point-level** annotation for **video sequences**.

Results: 3D Semantic Scene Segmentation



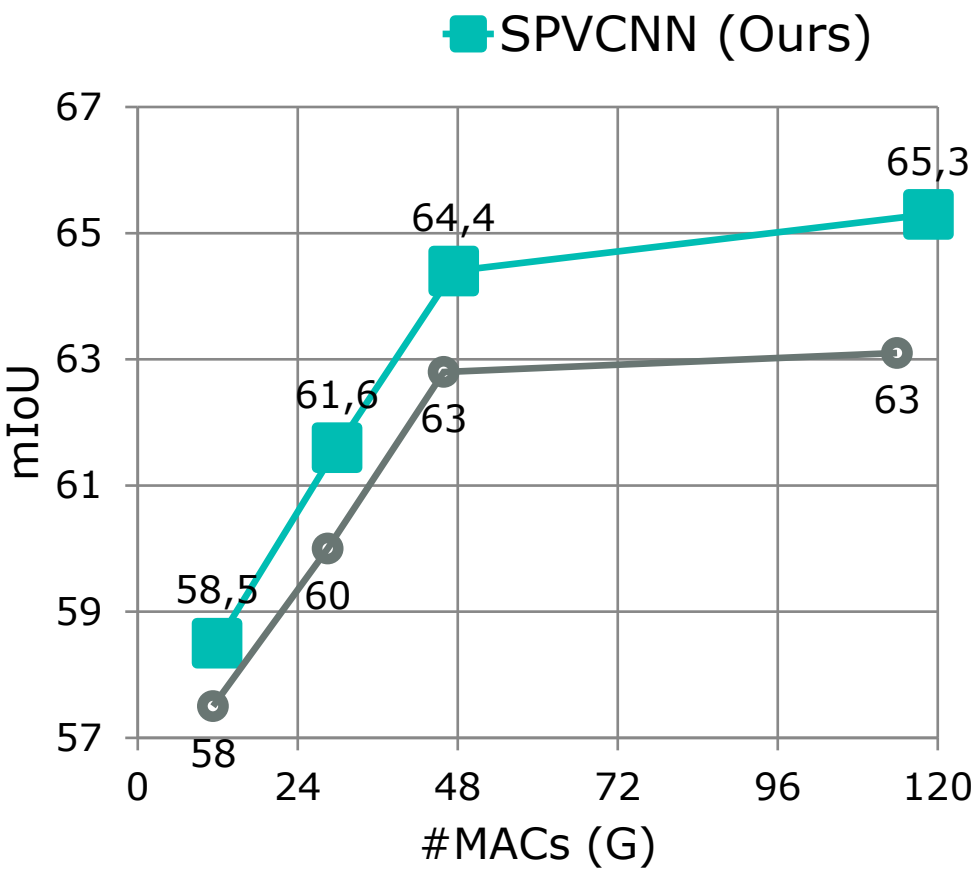
We achieve **8x** MACs reduction and **3x** speedup over MinkowskiNet with **SPVNAS**

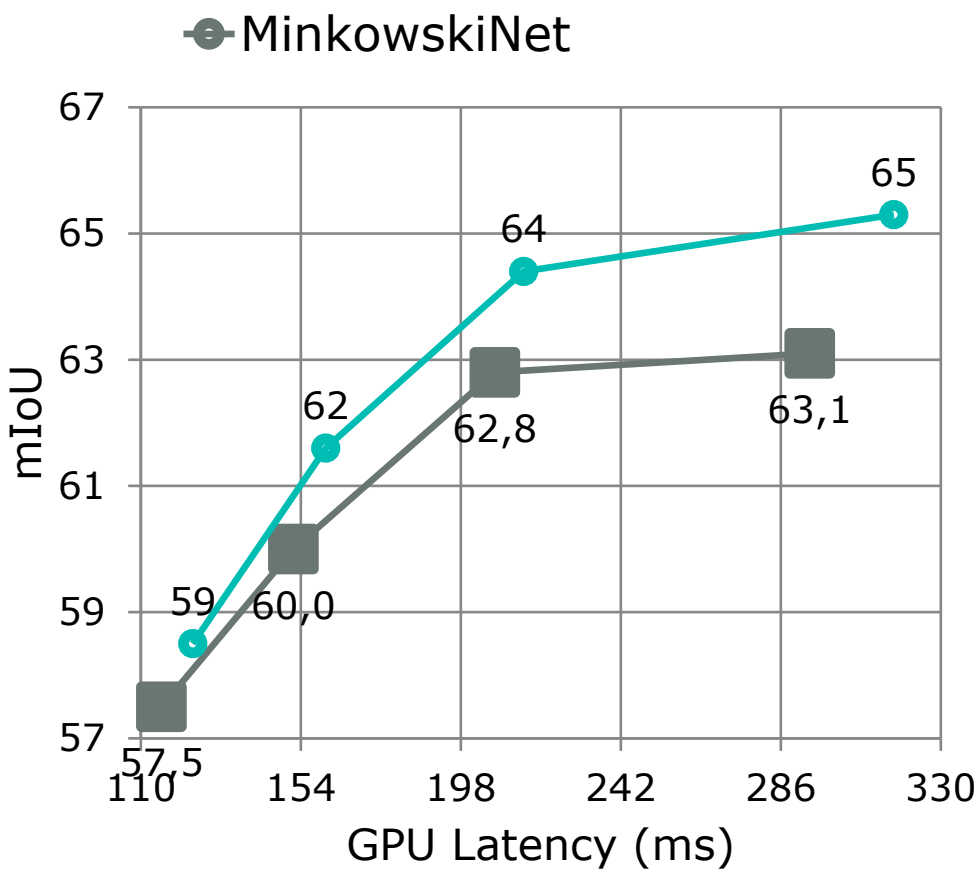
Results: 3D Semantic Scene Segmentation



Both a better module (**SPVConv**) and **3D-NAS** improve the performance of MinkowskiNet.

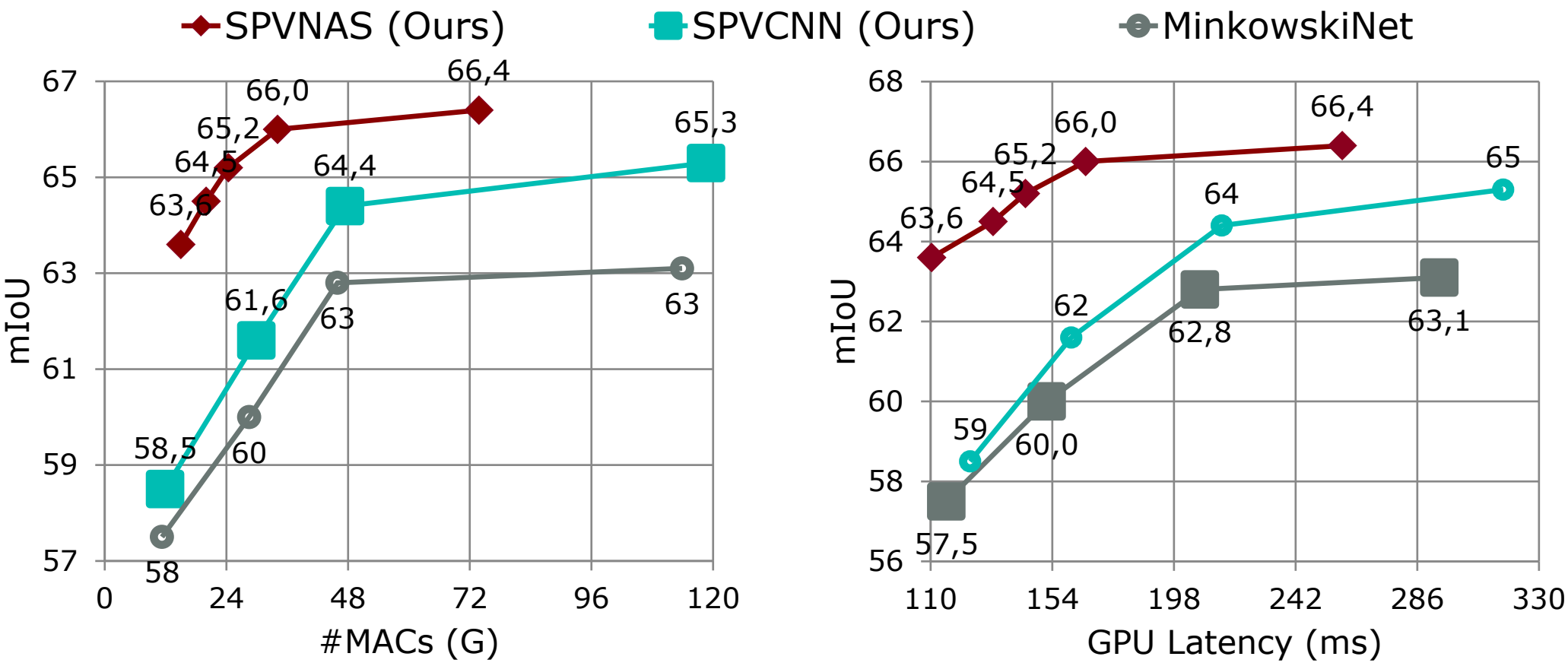
Results: 3D Semantic Scene Segmentation





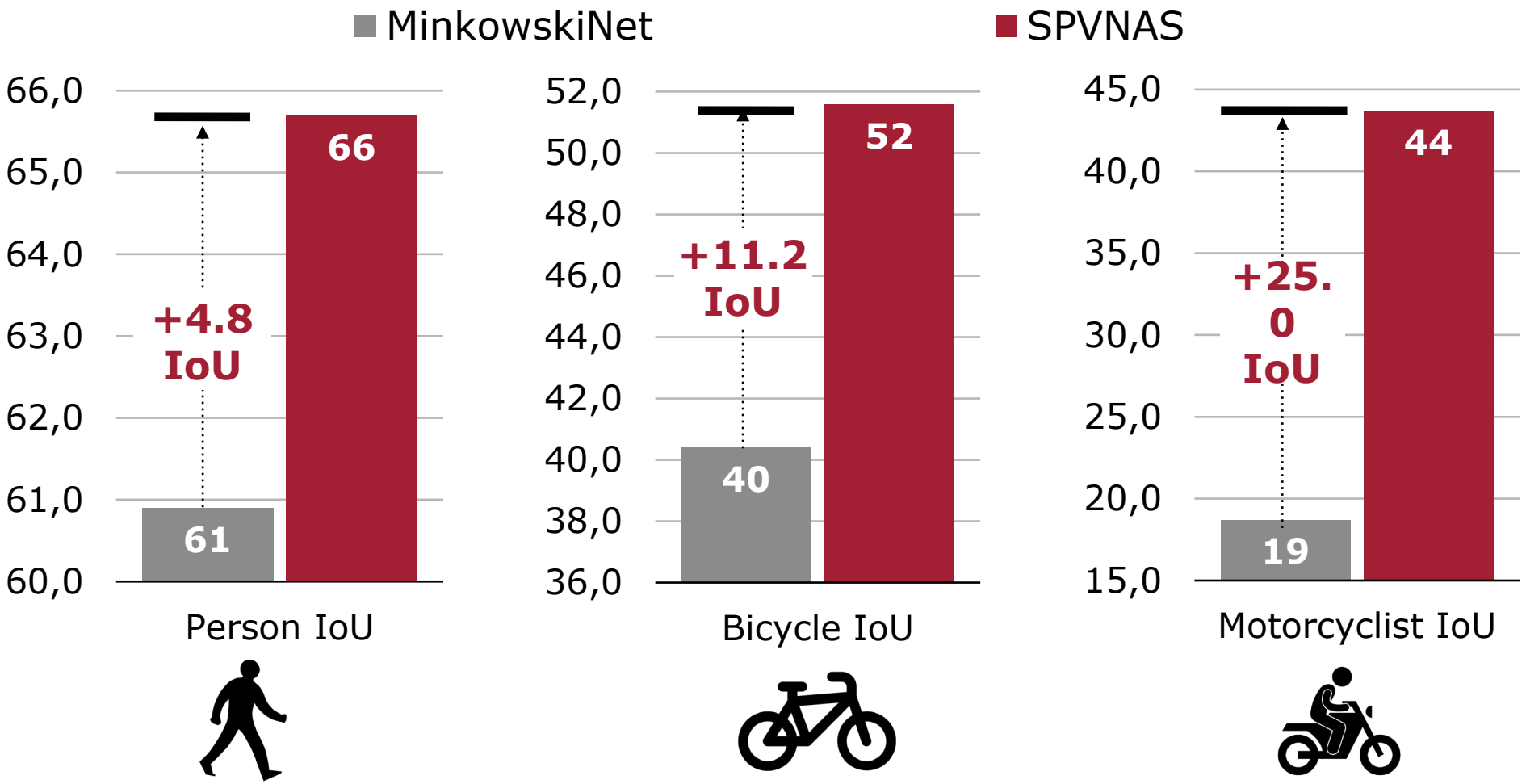
Both a better module (**SPVConv**) and **3D-NAS** improve the performance of MinkowskiNet.

Results: 3D Semantic Scene Segmentation



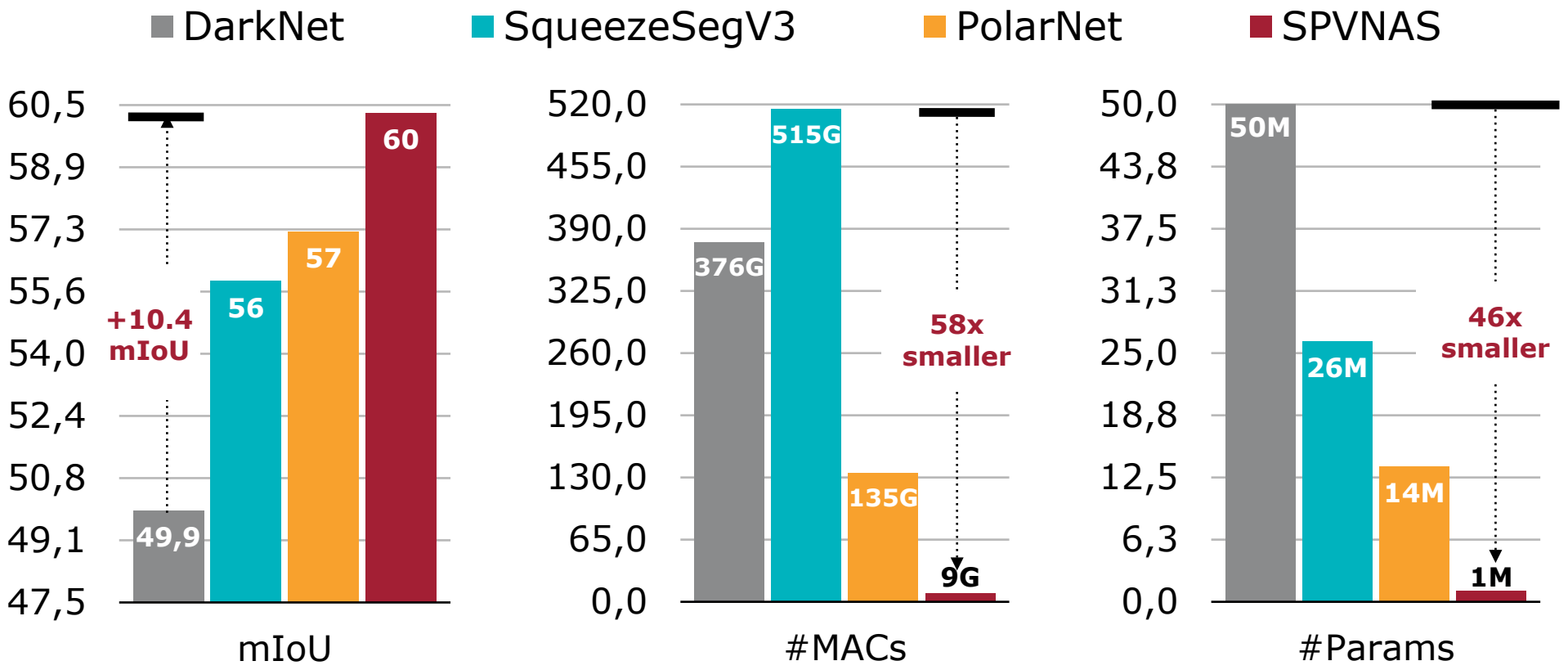
Both a better module (**SPVConv**) and **3D-NAS** improve the performance of MinkowskiNet.

Results: 3D Semantic Scene Segmentation



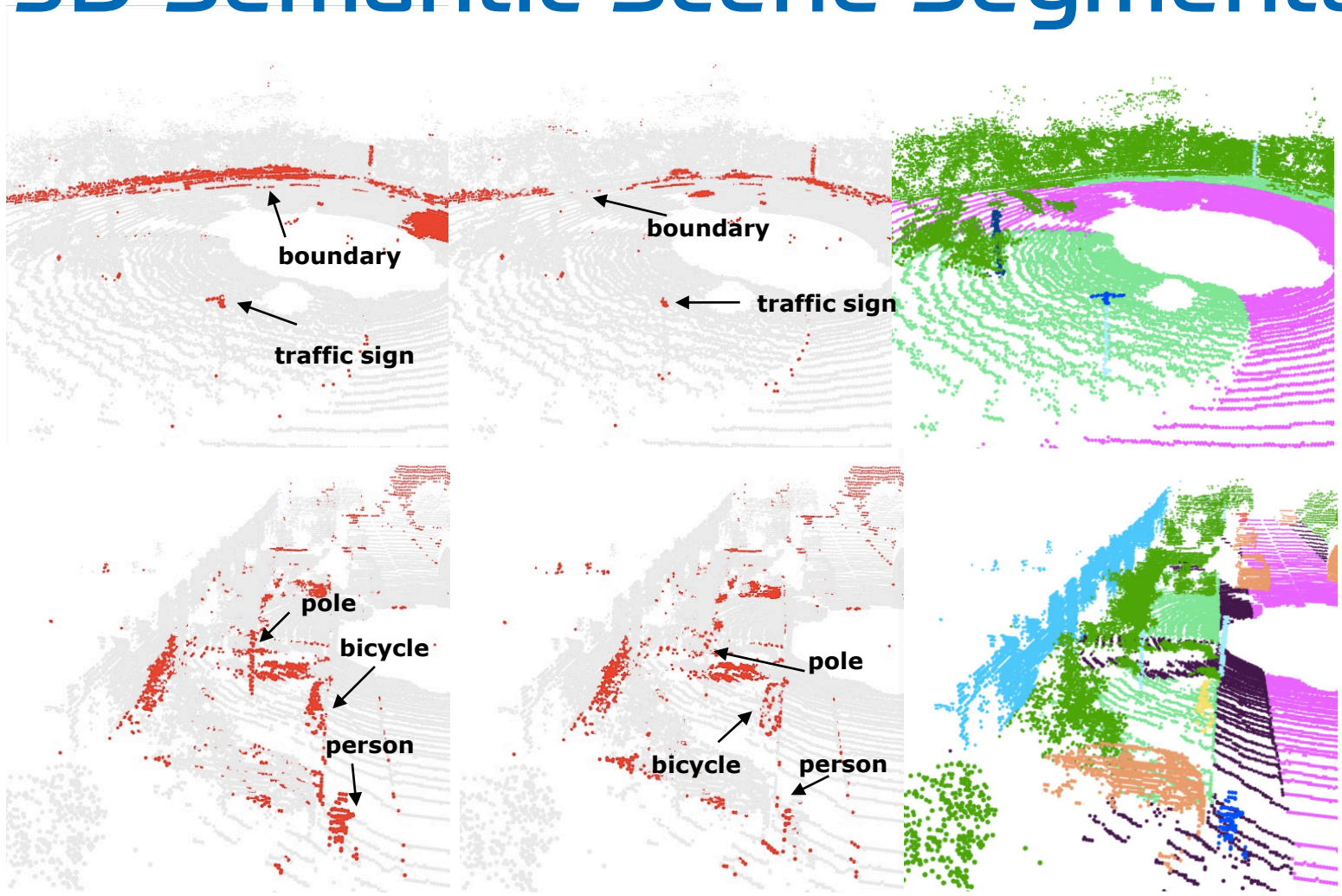
We achieve up to **25 mIoU** improvement on **safety-critical** small objects.

Results: 3D Semantic Scene Segmentation



We achieve up to **58x** MACs reduction and **46x** params reduction over projection-based methods.

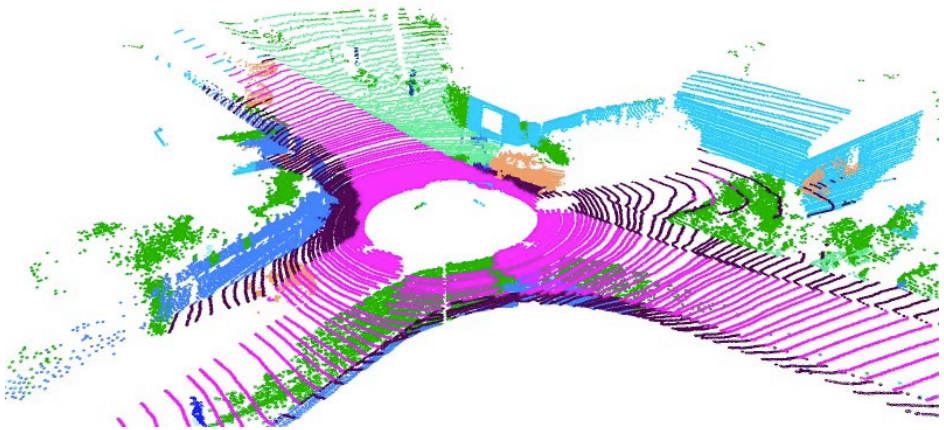
Results: 3D Semantic Scene Segmentation



MinkowskiNet | SPVNAS (**Ours**) | Ground Truth

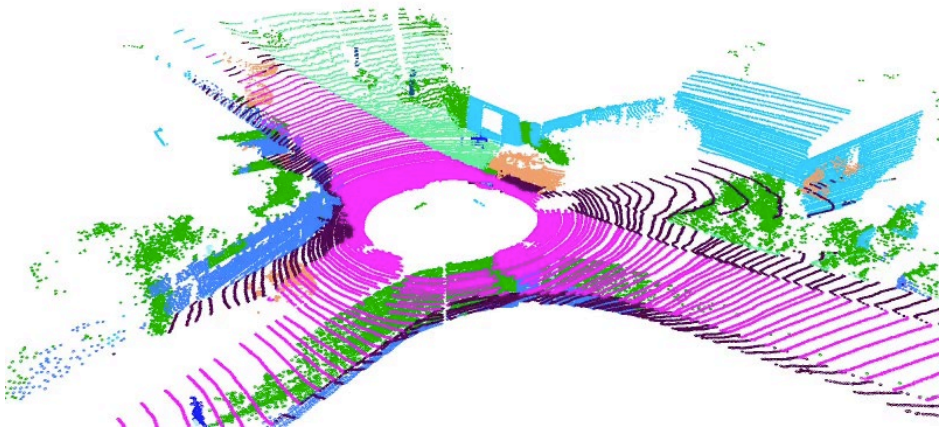
Demo: Significantly Faster than MinkowskiNets

MinkowskiNet



Mean IoU: **63.1** Throughput: **3.4** FPS
(**21.7M** Params **114.0G** FLOPs)

SPVNAS (Ours)



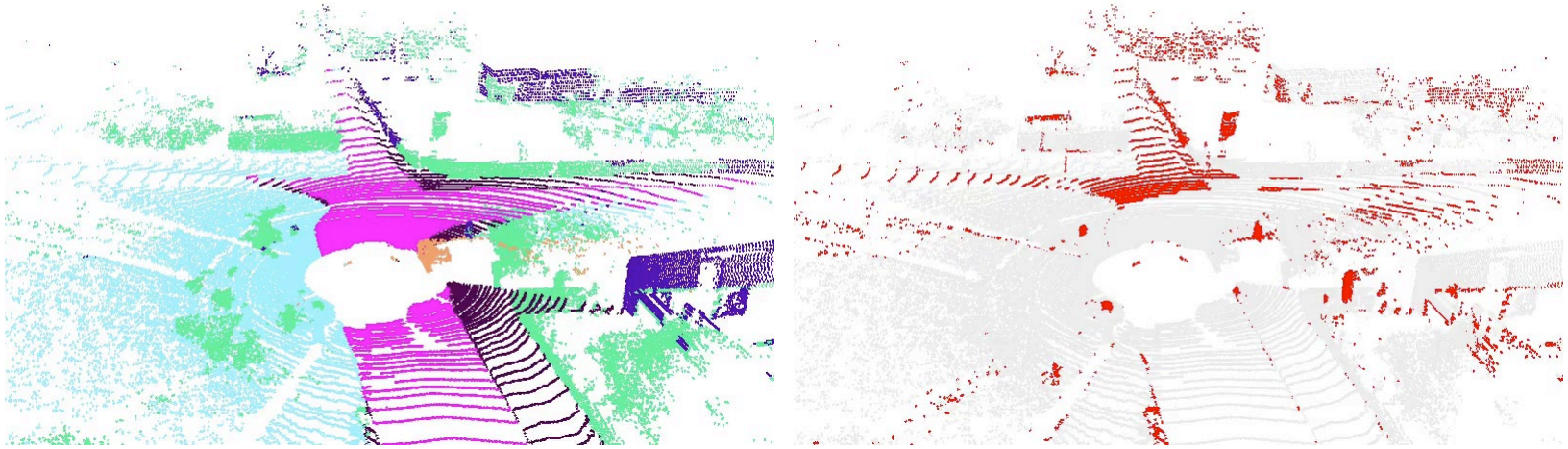
Mean IoU: **63.6** Throughput: **9.1** FPS
(**2.6M** Params **15.0G** FLOPs)

SPVNAS outperforms the state-of-the-art MinkowskiNet
(with **3x** measured speedup and **8x** model size reduction).

Demo: Faster than 2D, Accuracy of 3D

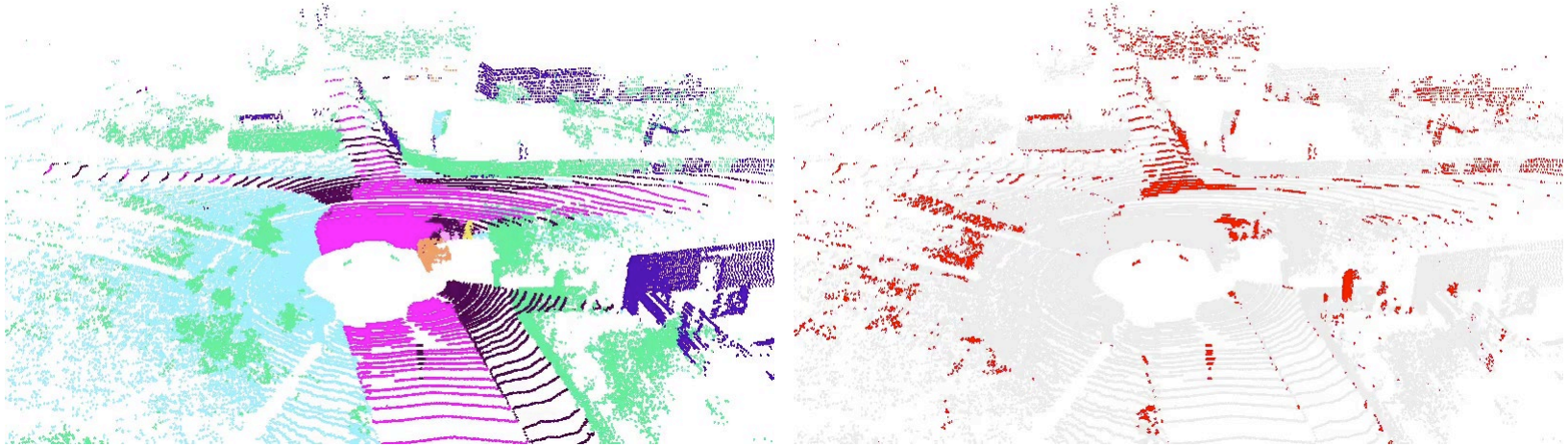
DarkNet53Seg

Mean IoU: **49.9**
Throughput: **9.7** FPS
50.4M Params
376.3G FLOPs

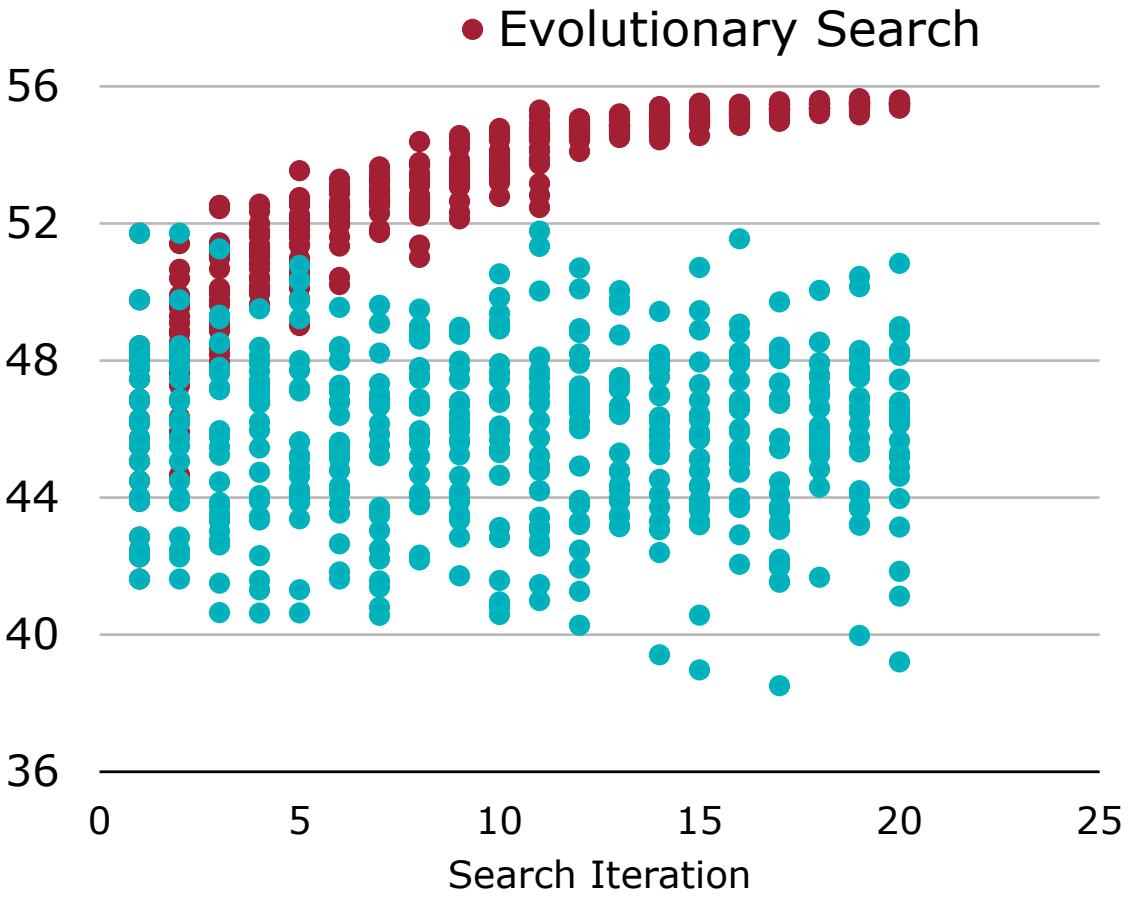


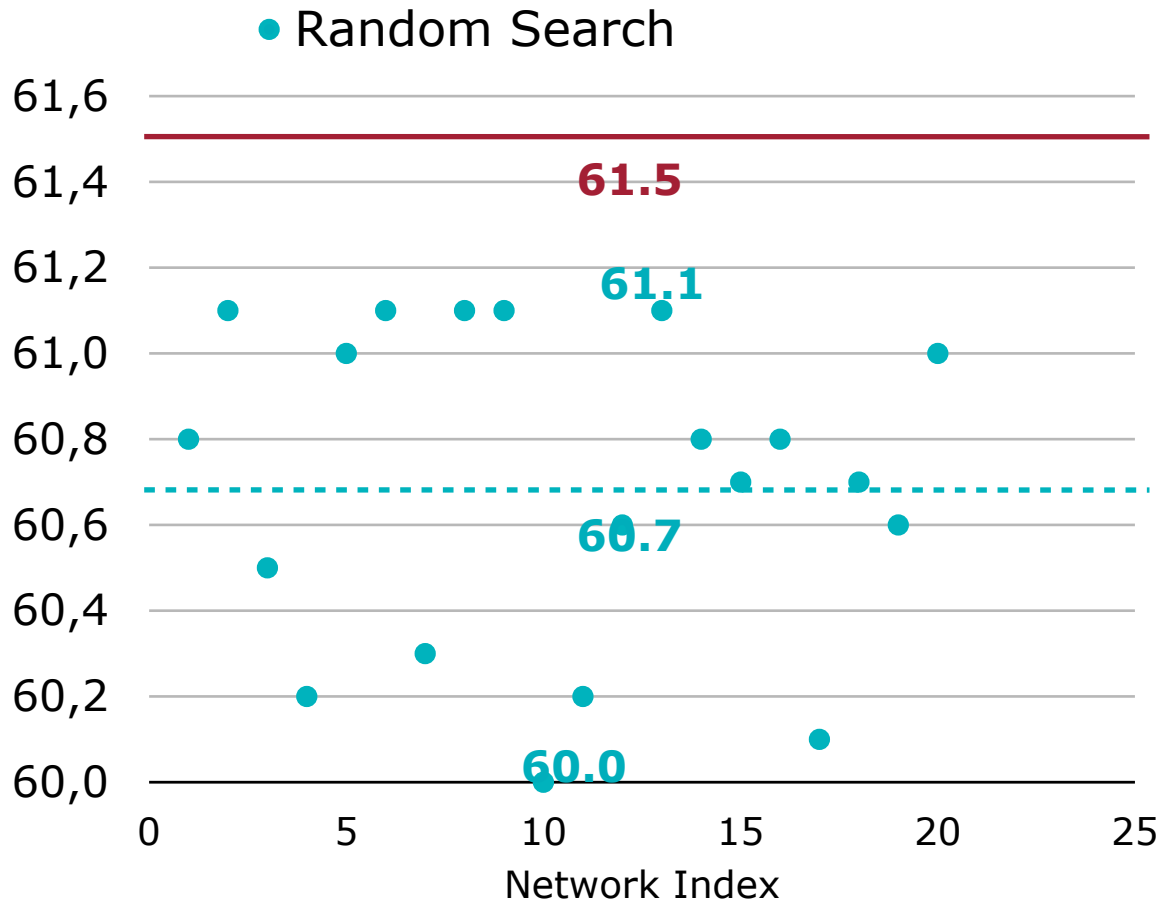
SPVNAS (Ours)

Mean IoU: **60.3** (> **KPConv**) Throughput: **11.2** FPS
1.1M Params
8.9G FLOPs



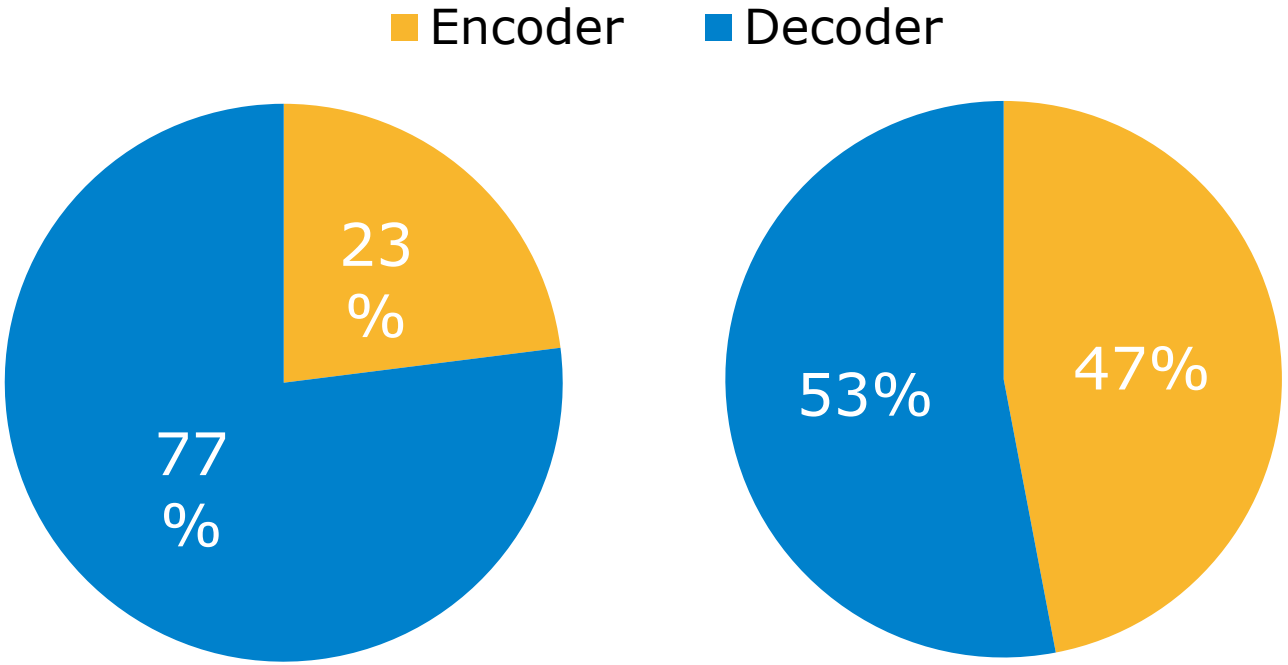
Results: Evolutionary Architecture Search





Author: Sibo Zhu, Zhijian Liu, Haotian Tang

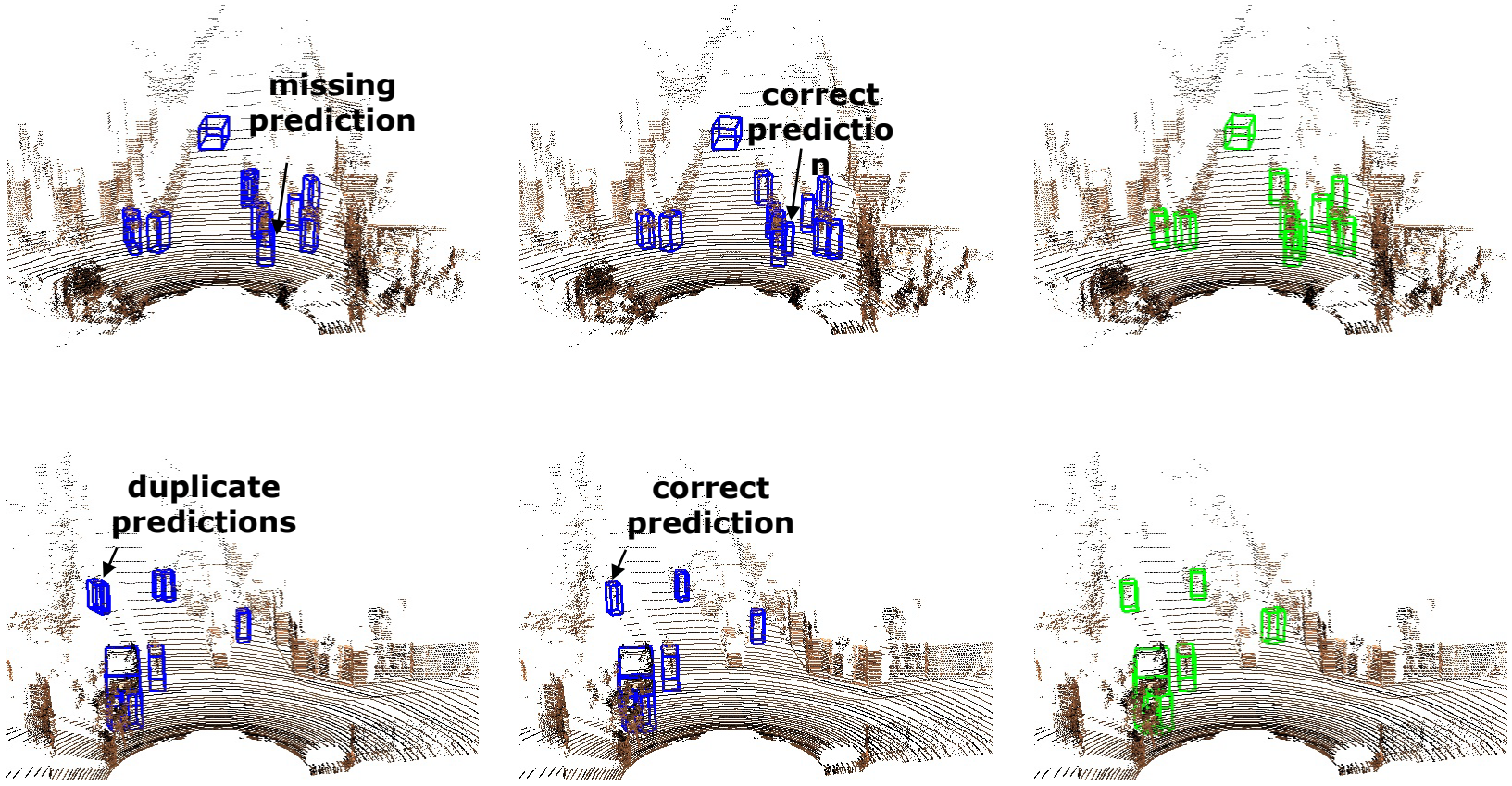
Results: Evolutionary Architecture Search



MinkowskiNet **SPVNAS - 20G**

SPVNAS balances the encoder / decoder computation ratio.

Results: 3D Object Detection



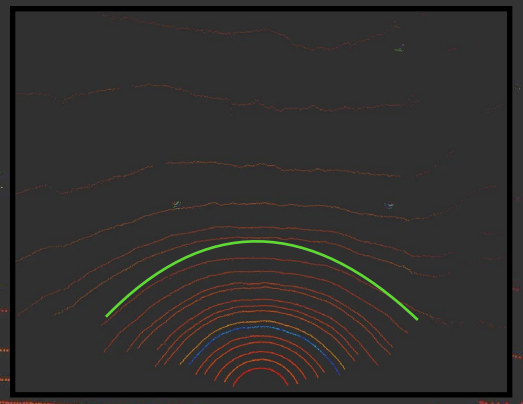
SECOND SPVCNN Ground Truth

Original Solution

Accuracy: **95%**

Range: **8 meters**

Latency: **2 ms/object**

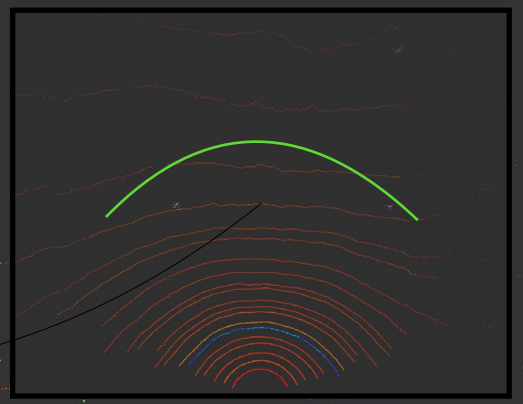


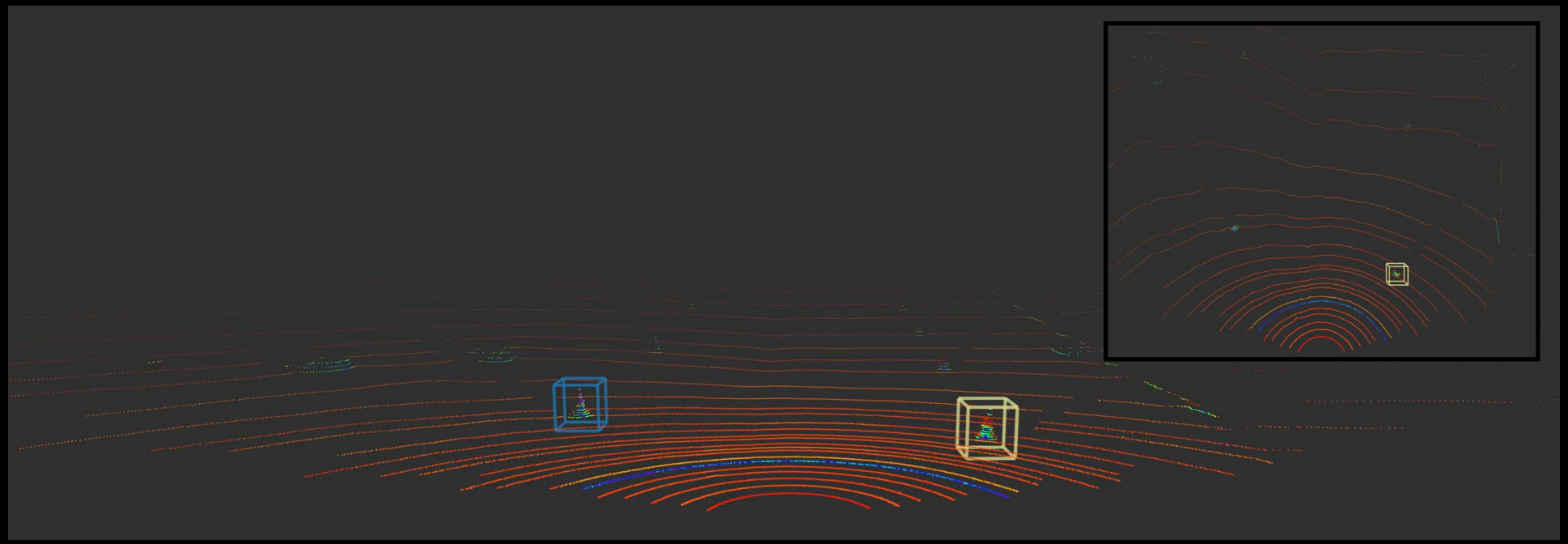
PVCNN

Accuracy: **99.93%**

Range: **12 meters**

Latency: **1.25 ms/object**





Author: Zhijian Liu, Haotian Tang

Thank you!