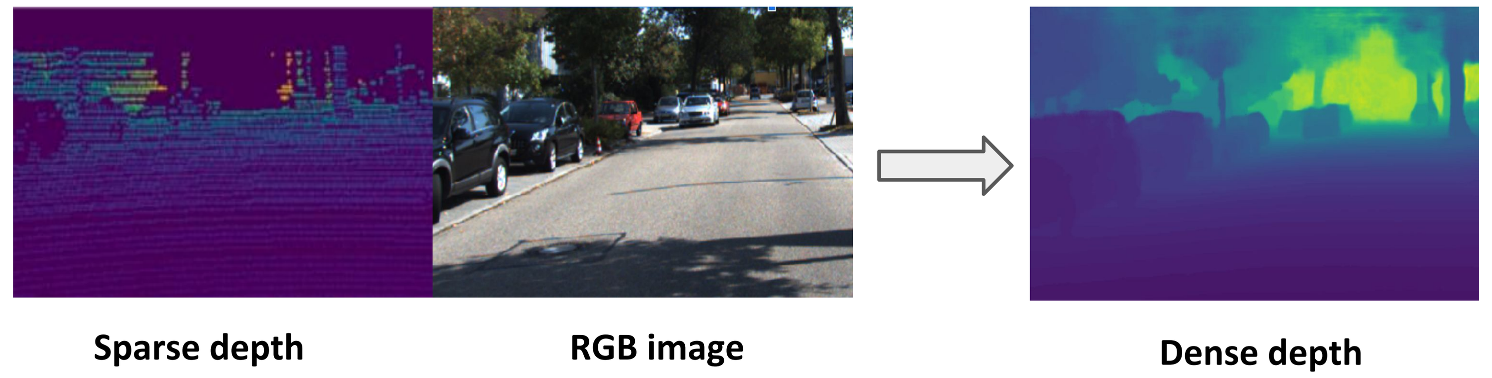
Learning Joint 2D-3D Representations for Depth Completion

Yun Chen[1], Bin Yang[1,2], Ming Liang[1], Raquel Urtasun[1,2]
[1]Uber ATG, [2]University of Toronto

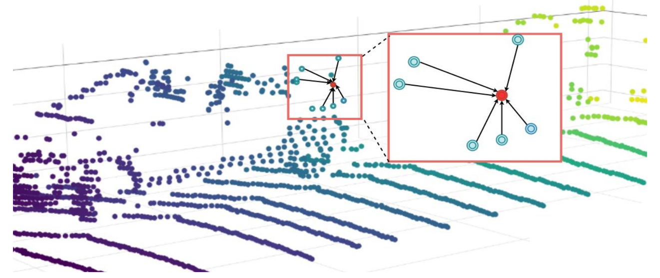
Introduction

- **Depth completion task** is to generate dense depth predictions from RGB image and sparse depth.



- Depth completion is important in self-driving for downstream perception tasks (e.g., 3D detection) because it can remedy the sparsity of lidar observation.
- Related Work:
  - Learns representations solely in 2D image space by applying 2D convolution on RGBD features.
  - Has difficulty capturing fine-grained 3D geometry as 3D structure is ignored in feature learning
  - Resorts to multi-task learning for better performance (e.g., exploiting supervision from semantic segmentation, surface normal estimation)
- By performing convolution on both **2D and 3D** spaces, the model is able to learn better feature representation.
  - **2D representation:** It provides dense appearance cue, but the grid neighbors may cover both **foreground** and **background** objects.

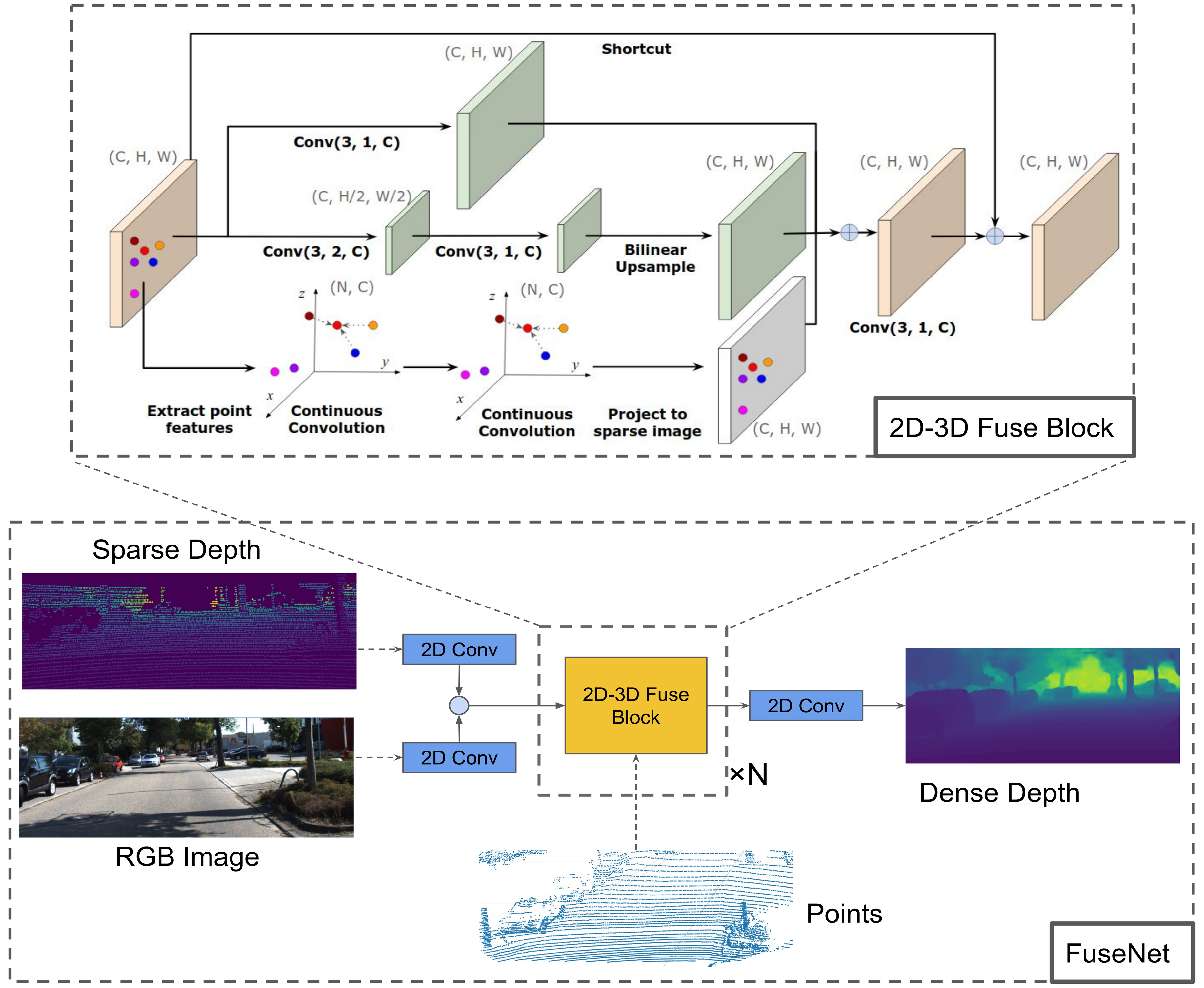
  - **3D representation:** Object has clear boundary. The K nearest neighbors of the center point are based on the exact 3D geometric correlation.



- We build our depth completion model, **FuseNet**, by stacking a set of proposed **2D-3D Fuse Blocks**.
- Our method achieves new **state-of-the-art** on KITTI depth completion benchmark.

Model

- The proposed block can fully exploit both 2D appearance and 3D geometric features. It contains 3 components:
  - **2D Branch:** exctract multi-scale 2D convolution on RGBD feature map.
  - **3D Branch:** index point features from the image with projection matrix and then adopt *continuous convoluton* to extract 3D geometric features.
  - **Fusion:** fuse the point features back to image plane using unprojection and sum all features.



- FuseNet is built with proposed blocks plus a few 2D convolution layers at the input and output stages.
- We can modify the number of channels and blocks to get a better **trade-off between accuracy and model size**.
- FuseNet is **trained from scratch** without using any additional data or pretrain weight.

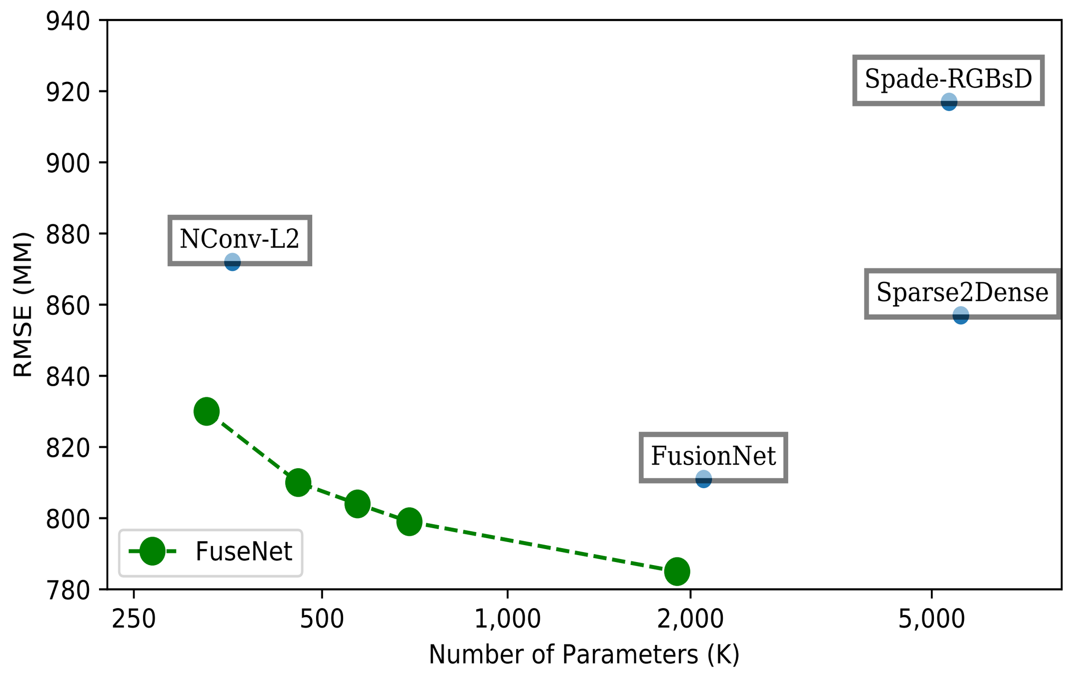
Experiment Results

Our method achieves state-of-the-art in the KITTI benchmark with much better accuray-size trade-off.

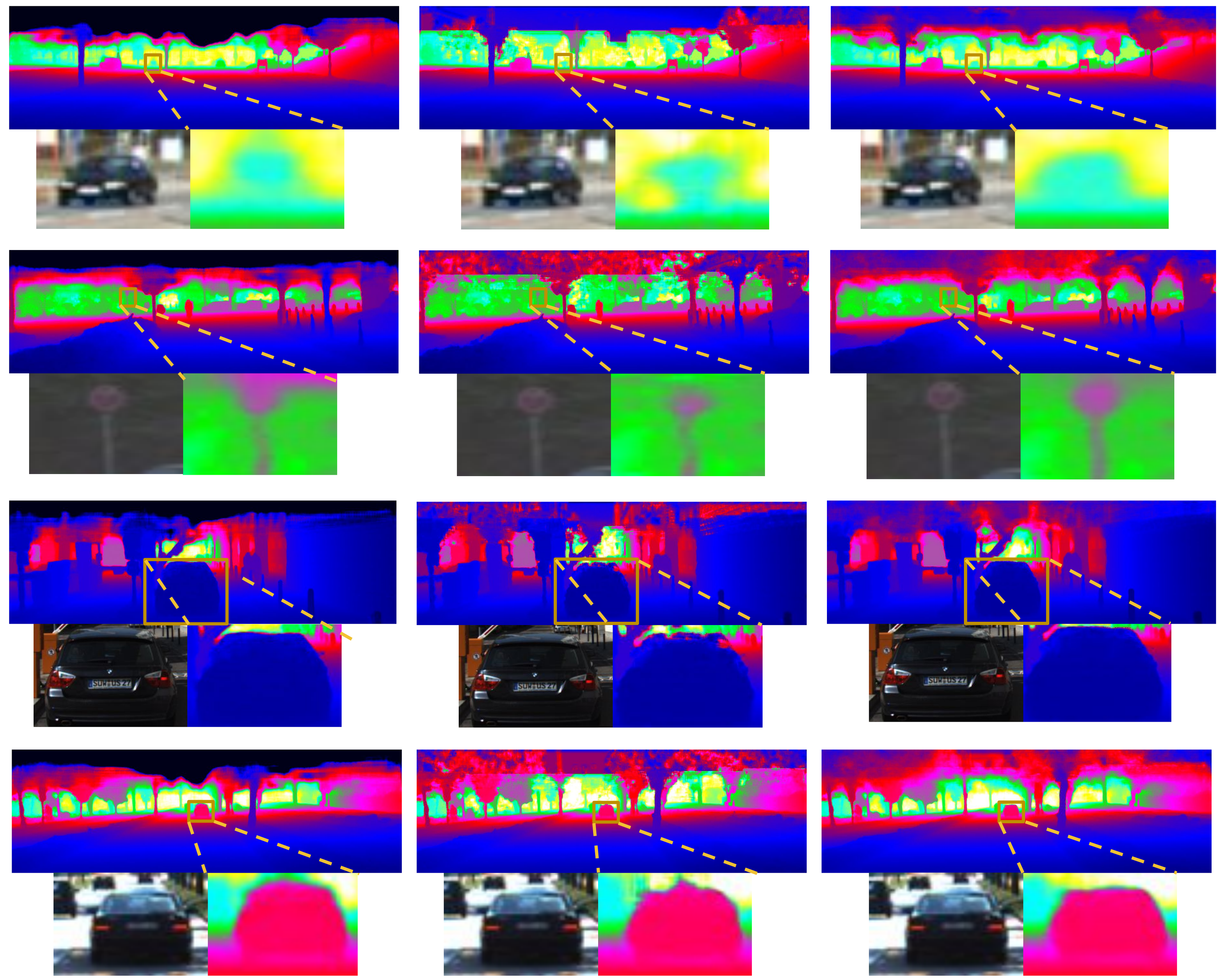
Results on KITTI test set

Method	RMSE (mm)	MAE (mm)
SparseConvs	1601.33	481.27
CSPN	1019.64	279.46
DDP	832.94	**203.96**
NConv-CNN	829.98	233.26
Sparse2Dense	814.73	249.95
PwP	777.05	235.17
DeepLiDAR	775.52	245.28
FusionNet	772.87	215.02
Our FuseNet	**752.88**	221.19

Trade-off between accuracy and model size



Qualitative Results



Sparse2Dense NConv Ours (FuseNet)