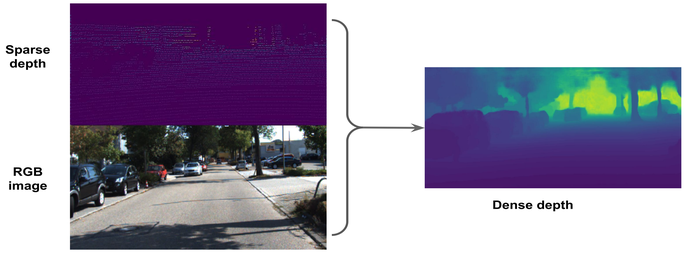
Learning Joint 2D-3D Representations for Depth Completion

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** given sparse depth (from LiDAR) + dense RGB image, estimate dense depth.Important for 3D reconstruction and downstream multi-sensor perception and prediction tasks.
- **Related work** applies 2D CNN on RGBD images or resorts to multi-task learning for better performance (semantic/surface normal).
- **Ours** take advantage of both 2D and 3D representations and design a simple yet effective architecture that fuses the information between these representations at multiple levels. We achieves new SOTA on KITTI with much fewer parameters.



Motivation

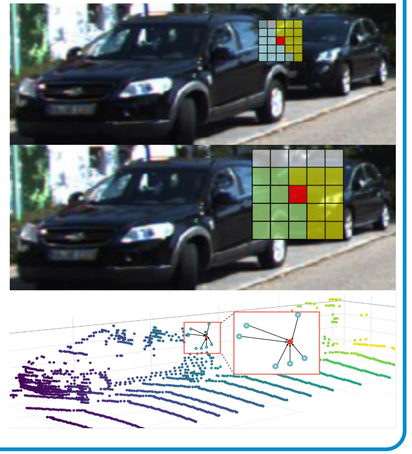
Comparison of *image* and *points*

- **Image** contains dense colorful appearance feature, but lack geometric relation.
- **Points** provide 3D info but is sparse.

Comparison of *grid convolution* and *points convolution*

- **Grid convolution:** the grid neighbors are not necessarily close to each other in 3D space, may cover both **foreground** and **background** objects.
- **Continuous convolution:** the K neighbors (x_k) of x_i are based on the exact 3D geometric correlation.

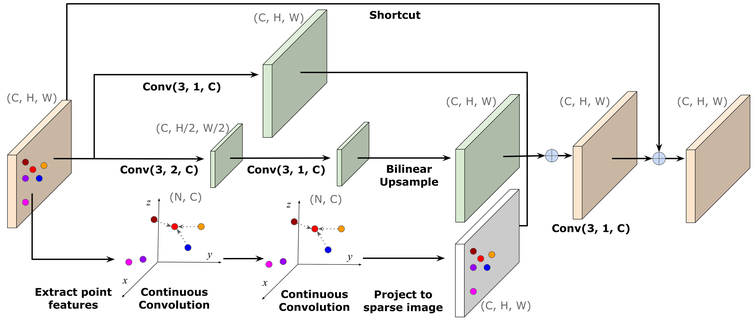
$$\mathbf{h}_i = W(\sum_k \mathrm{MLP}(\mathbf{x}_i - \mathbf{x}_k) \odot \mathbf{f}_k)$$



2D-3D Fuse Block

We propose a building block to fully exploit both 2D appearance and 3D geometry features.

- **2D Image Branch:** Multi-scale 2D convolution on RGBD feature map.
- **3D point Branch:** First index point feature from the image with project matrix and then adopt *continuous convoluton* to extract 3D geometry feature.
- **Fusion:** Fuse the points feature back to image plane using unprojection and sum all features.

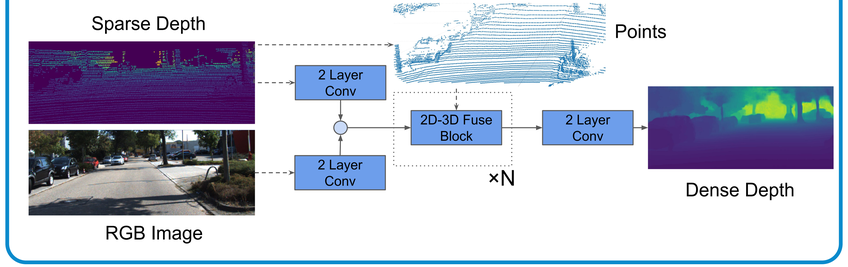


Stacking such blocks together leads to hierarchical joint representation learning that fully exploits the complementary information between the two sensor modalities.

FuseNet

We build our depth completion model **FuseNet** by simply stacking a set of proposed blocks plus a few 2D convolution layers at the input and output stages.

- The 2D-3D fused network takes image and sparse depth as input and predicts dense depth output. The main part of the network is the stacking of N 2D-3D fuse blocks.
- We can tune the number of channels (C) and blocks to get better trade-off of speed-param.
- The model is training from **scratch** without using any additional data or pretrain weight.



Experiment Results

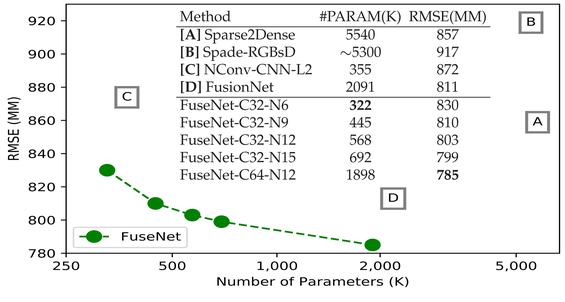
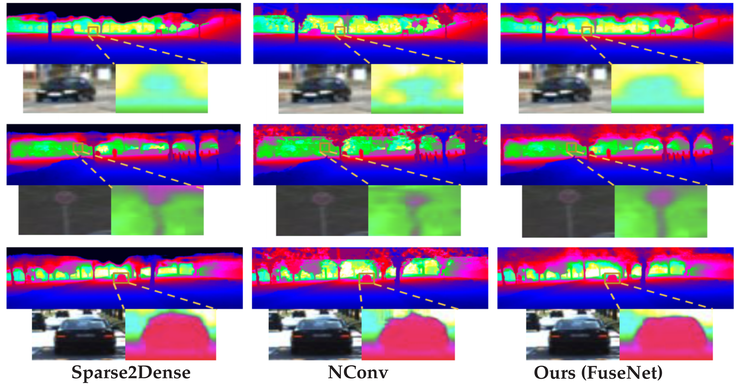
Our method rank **1*st*** in the KITTI benchmark at submission. Previous methods usually combine results of two models or use additional data/pretrain weight.

Results on KITTI test server

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
DeepLiDAR	775.52	245.28
FusionNet	772.87	215.02
Our FuseNet	**752.88**	221.19

Ablation Studies on block and loss

stride_1 conv	stride_2 conv	cont. conv	RMSE (mm)
	✓	✓	840
✓		✓	826
✓	✓		817
✓	✓	✓	**803**

Loss	RMSE	MAE	iRMSE	iMAE
ℓ_2	790	232	2.51	1.16
smooth ℓ_1	839	**197**	**2.23**	**0.91**
ℓ_2+smooth ℓ_1	**785**	217	2.36	1.08

Trade-off between accuracy and model size