View-Volume Network for Semantic Scene Completion from a Single Depth Image

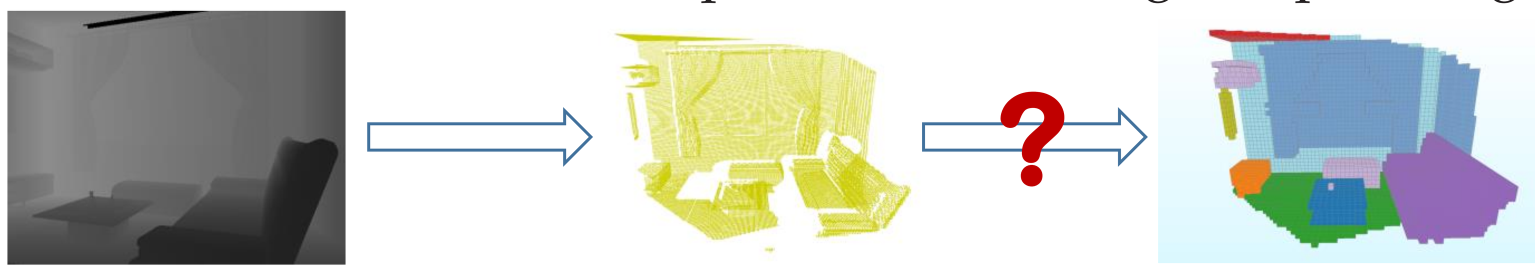
Yuxiao Guo[1], Xin Tong[2]

[1] University of Electronic Science and Technology of China

[2] Microsoft Research Asia

Introduction

Problem: Semantic scene completion from a single depth image

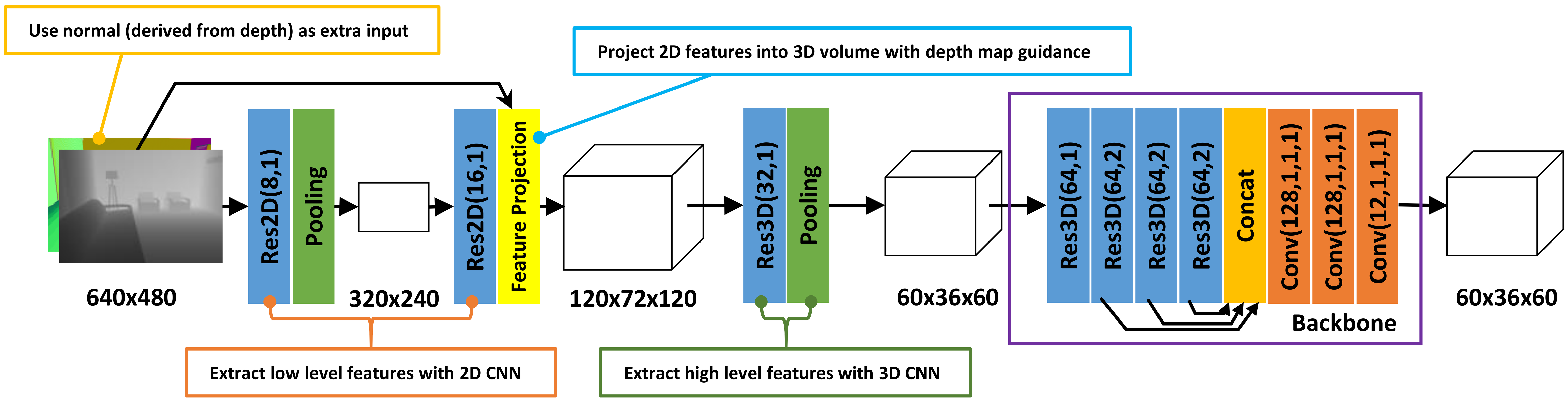


Challenge: 2D CNN is hard to perform geometry completion task while 3D CNN is limited by computation resource and memory to deal with high-resolution scenes.

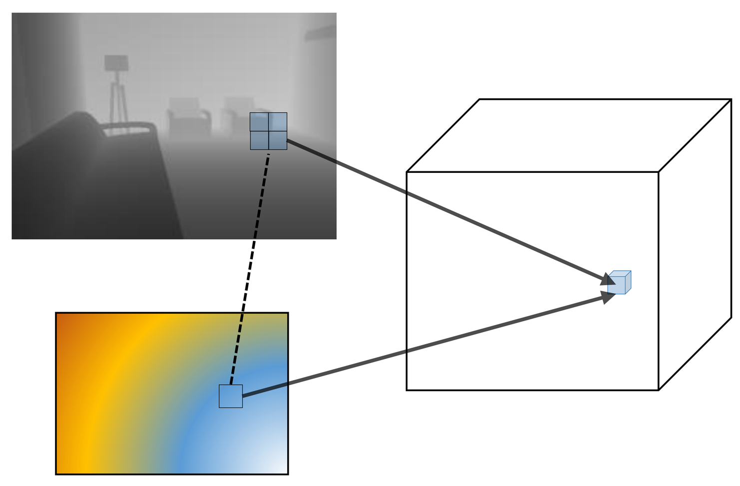
Contribution

- Propose a differential feature projection operation to fuse 2D features into 3D volume space with depth map guidance.
- Propose the view-volume hybrid architecture to efficiently organize 2D and 3D CNNs in semantic scene completion task, with flexible network choices.
- Outperform the state-of-art methods in both synthetic and real datasets, with much better accuracy and 3-10 times speedup.

Framework

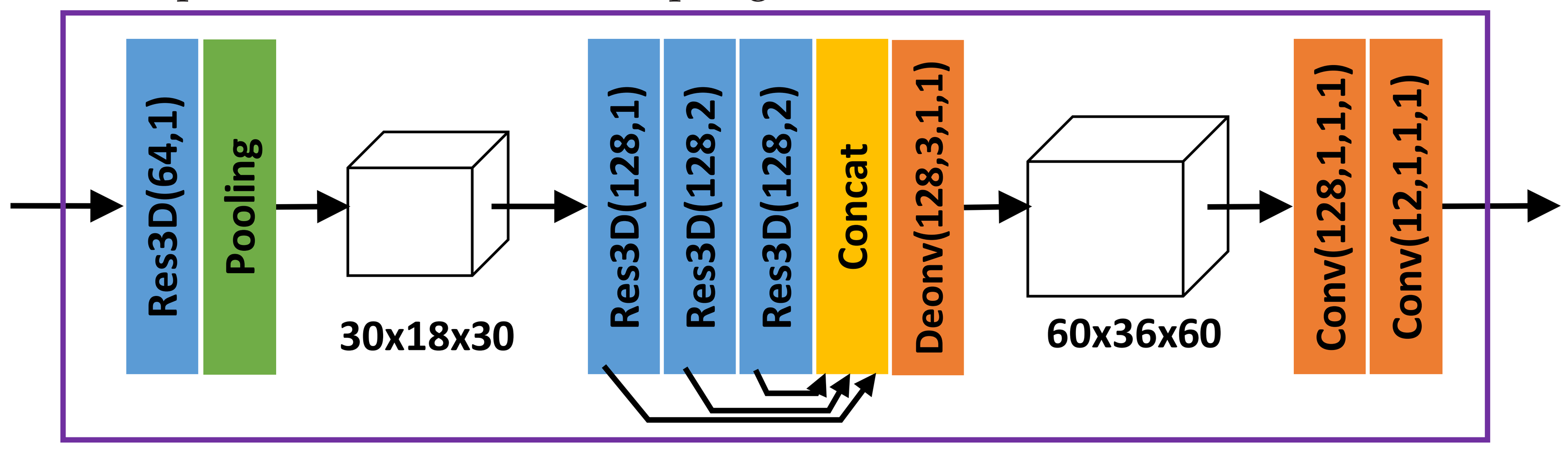


Feature Projection



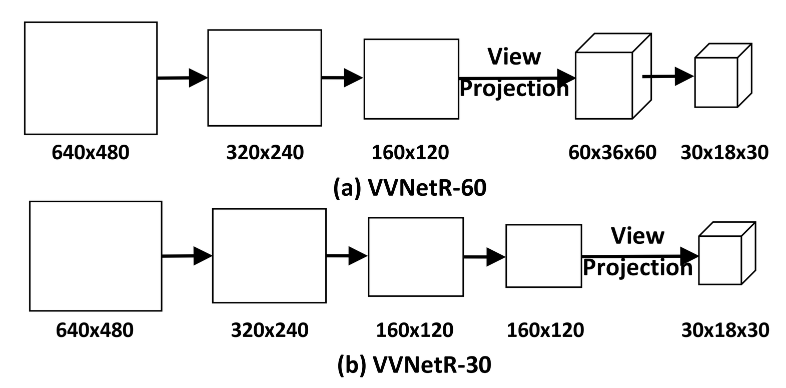
Reception Field Enhancement Backbone

Double reception field via down-sampling volume to a lower resolution



Tradeoffs

View-volume architecture provides flexible 2D-3D fusion choices



Results

Table 1: Performances of different variant VVNet design on the SUNCG dataset. **half** refers to the network that takes half-resolution image as input. **depth** refers to the network that use depth only as input.

	scene completion			semantic scene completion											
Network	prec.	recall	IoU	ceil.	floor	wall	win.	chair	bed	sofa	table	TVs	furn.	objs.	avg.
SSCNet	76.3	**95.2**	73.5	96.3	84.9	56.8	28.2	21.3	56.0	52.7	33.7	10.9	44.3	25.4	46.4
SSCNet*	90.4	89.7	82.0	97.8	**88.2**	59.4	37.3	39.2	77.9	68.9	48.3	31.5	56.8	44.9	59.1
SSCNet*-half	90.5	89.5	81.9	97.8	88.0	60.8	34.8	39.8	77.5	69.5	47.8	29.8	56.0	44.8	58.8
VVNet-120-half	90.7	89.6	82.1	97.9	85.2	59.4	47.5	44.2	77.4	71.1	49.3	34.2	58.2	49.0	61.3
VVNet-120-depth	90.6	89.6	82.0	97.6	84.8	58.6	44.5	44.8	77.6	70.7	48.8	33.2	57.8	46.2	60.4
VVNet-120	**90.8**	90.0	82.5	97.9	85.4	58.6	49.2	45.3	79.2	71.8	50.3	37.3	62.0	50.9	62.5
VVNetR-120	**90.8**	91.7	84.0	**98.4**	87.0	**61.0**	**54.8**	**49.3**	**83.0**	**75.5**	**55.1**	**43.5**	**68.8**	**57.7**	**66.7**
VVNetR-60	90.6	92.5	**83.7**	97.6	86.7	60.2	54.4	47.2	80.7	75.0	53.8	39.4	66.9	56.1	65.3
VVNetR-30	88.8	90.2	81.0	98.0	86.4	55.6	**54.8**	41.8	78.0	72.1	48.7	31.6	63.2	51.8	62.0

Table 2: The performances of different scene completion methods on the NYU dataset.

	scene completion			semantic scene completion											
Method	prec.	recall	IoU	ceil.	floor	wall	win.	chair	bed	sofa	table	TVs	furn.	objs.	avg.
Lin et al., 2013	58.5	49.9	36.4	0.0	11.7	13.3	14.1	9.4	29.0	24.0	6.0	7.0	16.2	1.1	12.0
Geiger et al., 2015	65.7	58.0	44.4	10.2	62.5	19.1	5.8	8.5	40.6	27.7	7.0	6.0	22.6	5.9	19.6
SSCNet	59.3	**92.9**	56.6	15.1	94.6	24.7	10.8	17.3	53.2	45.9	15.9	13.9	31.1	12.6	30.5
SSCNet*	69.7	81.3	59.8	16.1	**94.8**	27.0	10.1	**20.6**	53.2	50.1	16.7	**14.3**	35.5	13.0	31.9
VVNet-120	68.4	83.2	60.0	19.2	94.4	27.2	**13.8**	19.1	54.0	49.3	17.1	11.2	35.3	12.4	32.1
VVNetR-120	**69.8**	83.1	**61.1**	19.3	**94.8**	28.0	12.2	19.6	**57.0**	50.5	**17.6**	11.9	**35.6**	15.3	32.9
VVNetR-60	68.3	85.1	60.9	**21.6**	94.5	**28.6**	12.9	19.7	56.3	**51.0**	17.2	10.4	35.2	**15.6**	**33.0**

Table 3: Memory footprints and computational time of different networks for model training and inference.

	training		inference
Network	memory	speed	speed
SSCNet*	852M	912ms	578ms
VVNet-120	846M	386ms	75ms
VVNetR-120,	712M	375ms	74ms
VVNetR-60,	336M	194ms	51ms
VVNetR-30,	246M	156ms	45ms

Table 4: Performance of different methods on NYUCAD dataset.

Method	prec.	recall	IoU
Zheng et al., 2013	60.1	46.7	34.6
Firman et al., 2016	66.5	69.7	50.8
SSCNet	75.0	96.0	73.0
SSCNet*	83.2	92.7	78.0
VVNet-120	83.3	93.1	78.5
VVNetR-120	86.4	92.0	80.3
VVNetR-60	85.6	91.5	79.2

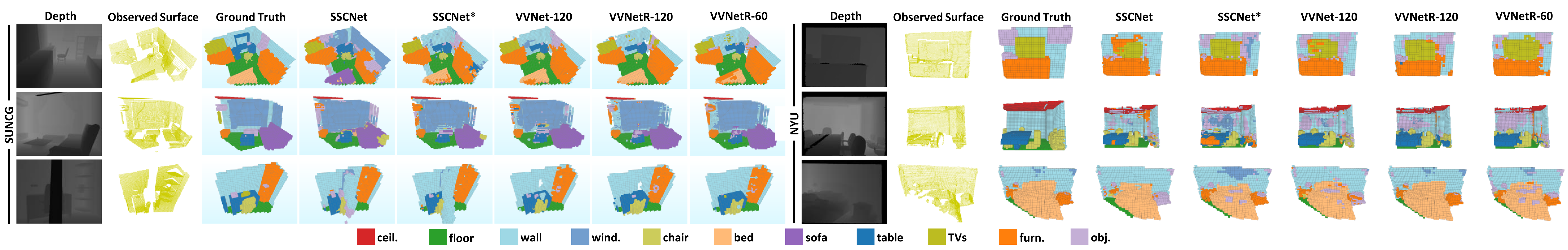


Figure 1: Semantic scene completion results generated by different methods for SUNCG and NYU datasets.