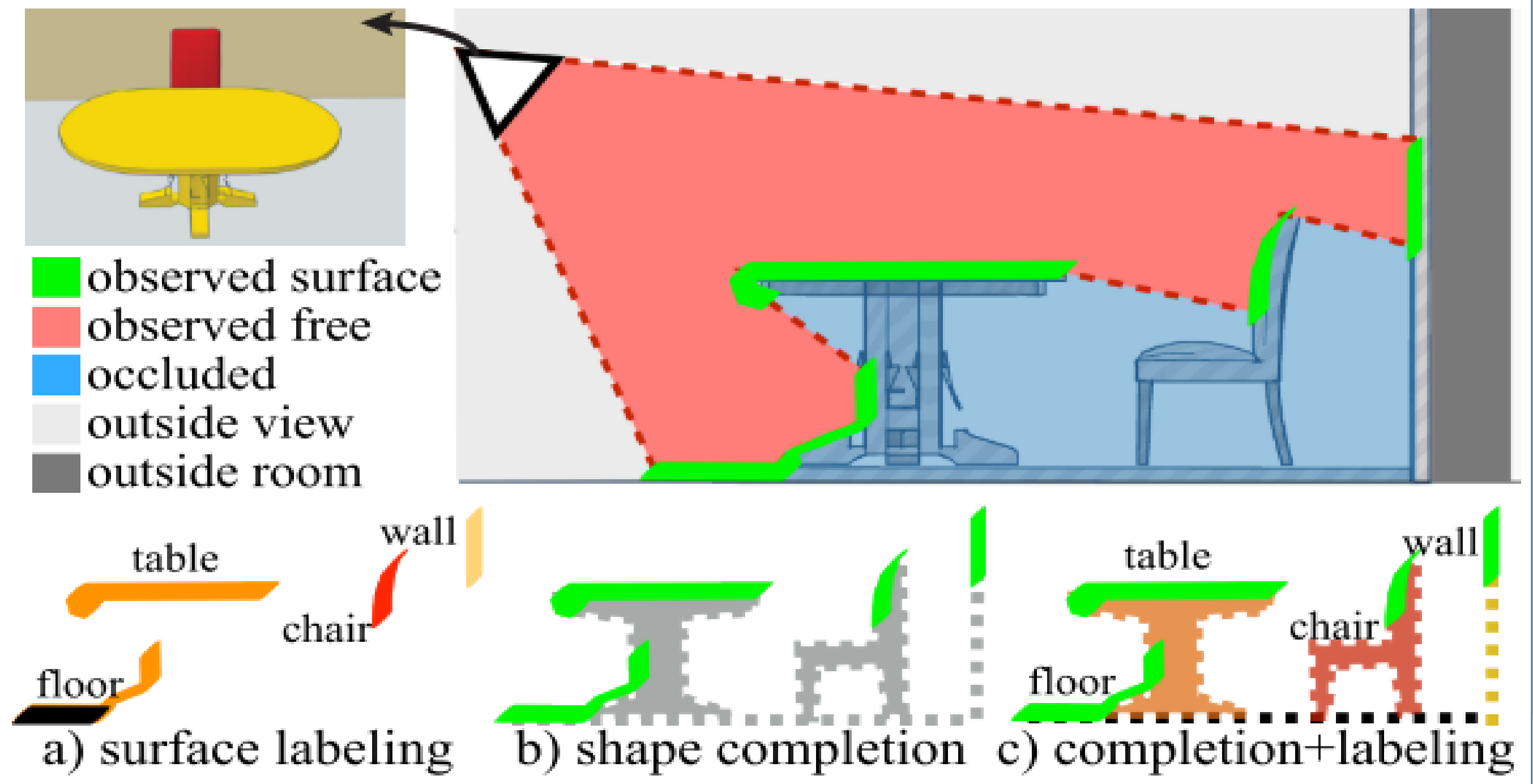
NeurIPS | 2018

See and Think: Disentangling Semantic Scene Completion

Shice Liu, Yu Hu, Yiming Zeng, Qiankun Tang, Beibei Jin, Yinhe Han, Xiaowei Li
{liushice, huyu, zengyiming, tangqiankun, jinbeibei, yinhes, lxw}@ict.ac.cn
State Key Laboratory of Computer Architecture, Institute of Computing Technology, Chinese Academy of Sciences
University of Chinese Academy of Sciences

1. Problem Statement



a) surface labeling b) shape completion c) completion+labeling

(The figure is extracted from [1].)

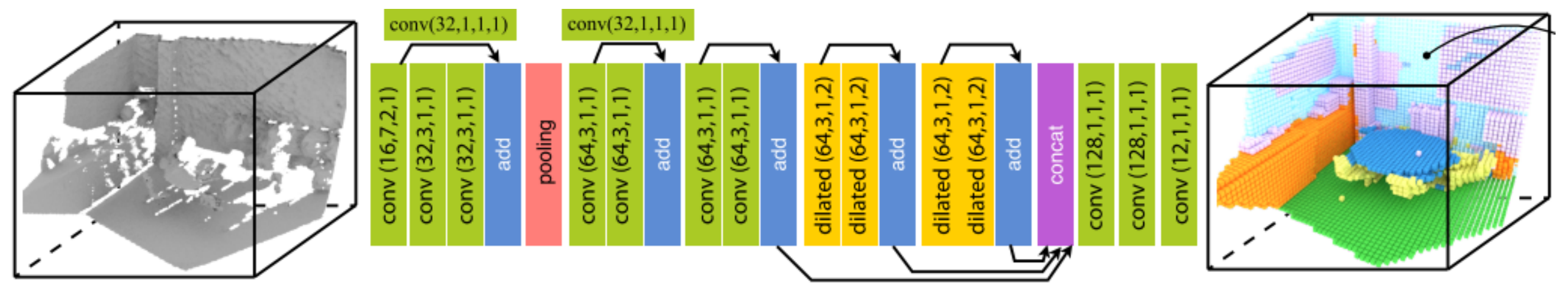
Semantic Scene Completion:

- Introduced by Song et al. in CVPR 2017.
- To predict volumetric occupancy and object category of a 3D scene, simultaneously.

2. Related Works

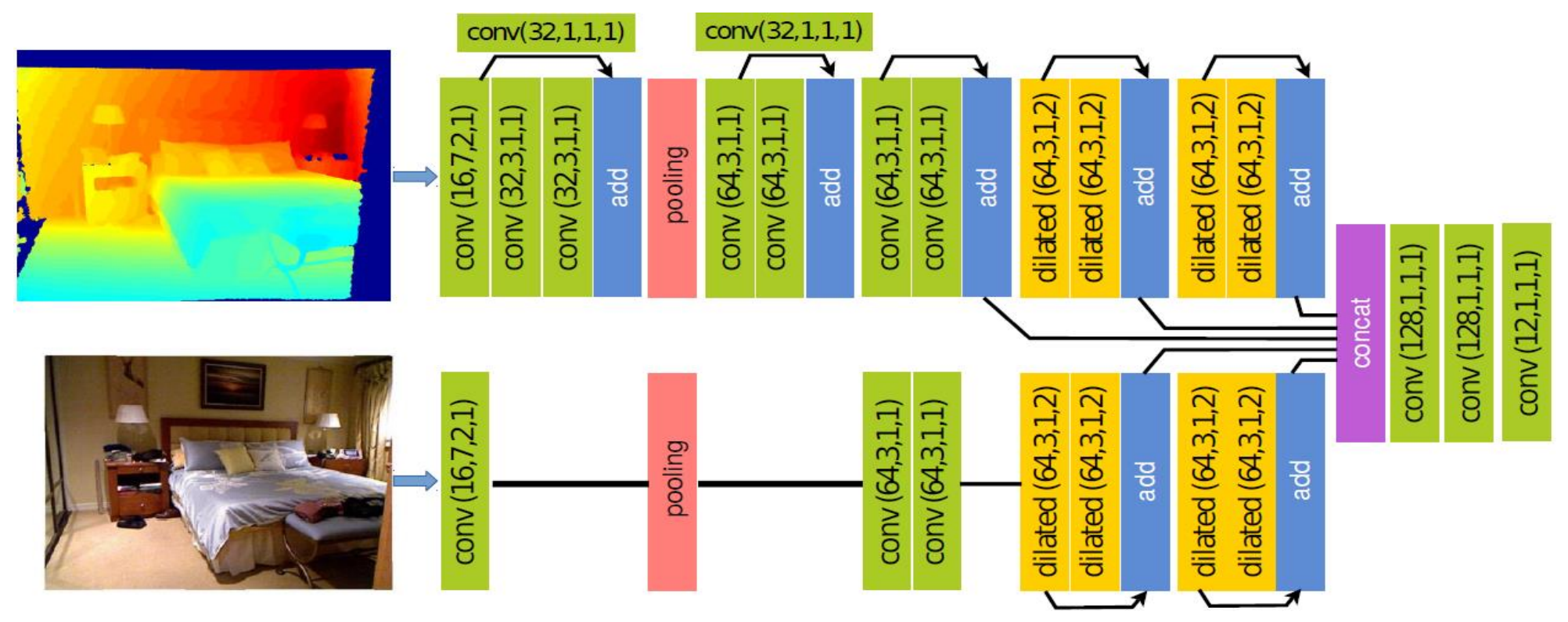
SSCNet [1]:

- The depth is encoded to a 3D tensor by flipped-TSDF.
- 3D convolutions are used for semantic completion.



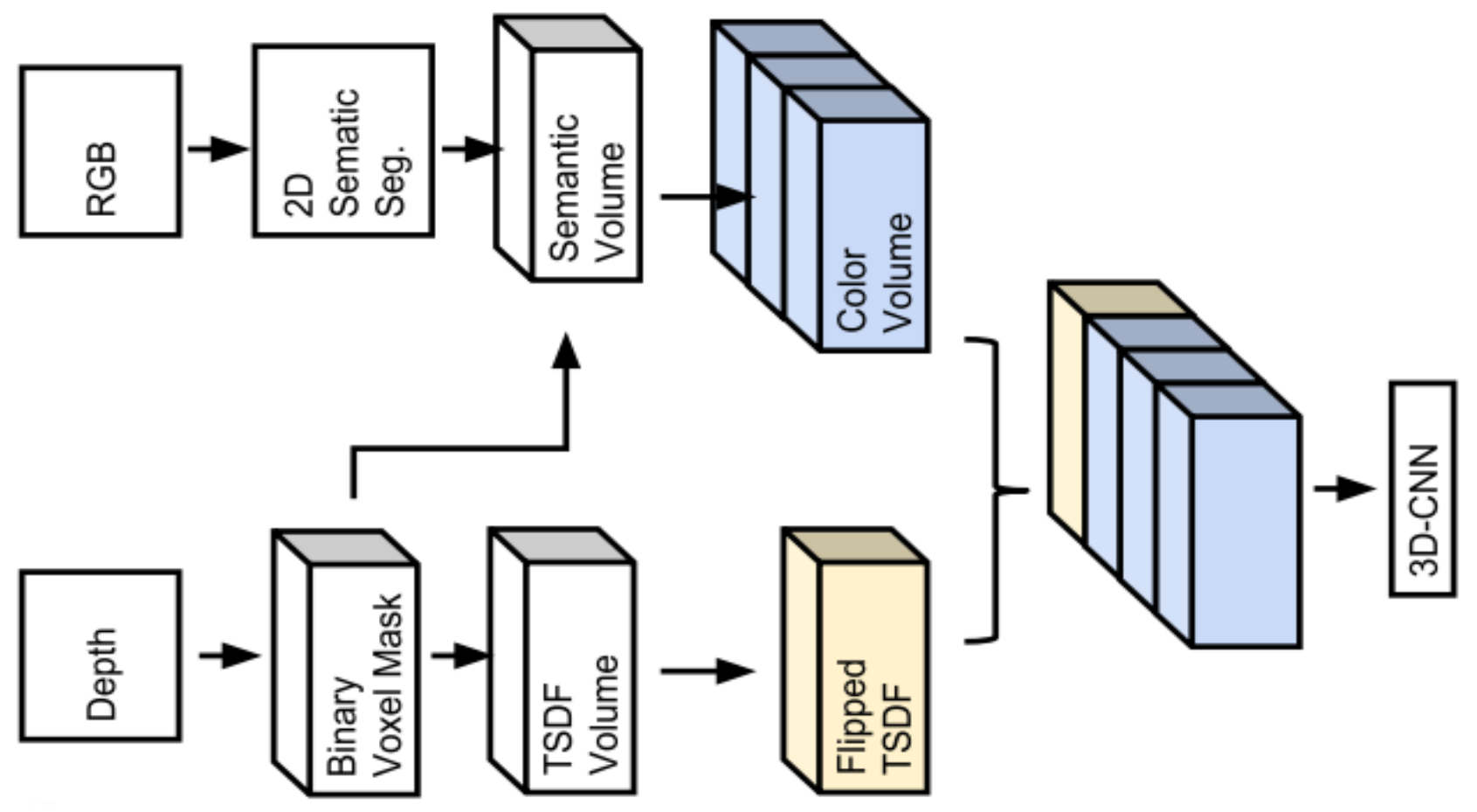
Colour-SSCNet [2]:

- The RGB image is projected to a 3D tensor.
- 3D convolutions for RGB are parallel to those for depth.



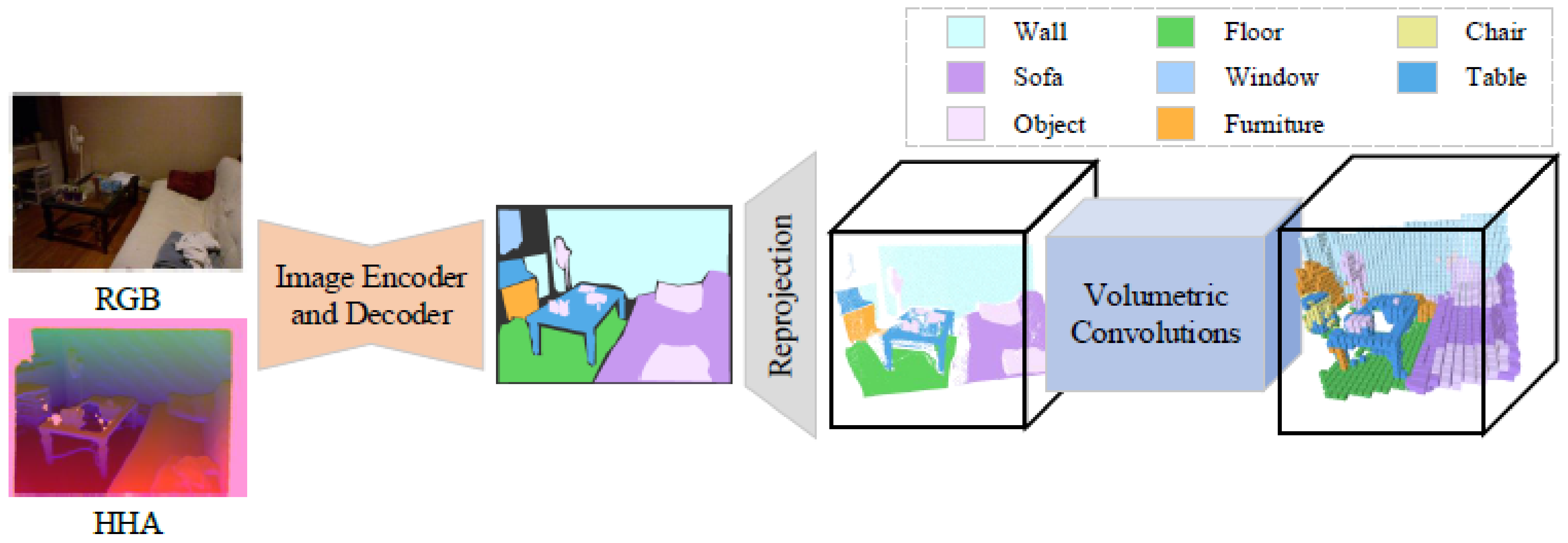
Two-stream SSCNet [3]:

- The semantic segmentation is introduced to RGB images.
- The segmentation result is projected to a 3D tensor.



3. Method Details

See And Think Network (SATNet) consists of three modules: (a) SNet, (b) a reprojection layer, and (c) TNet, sequentially carrying out 2D semantic segmentation, 2D-3D reprojection and 3D semantic scene completion.



(a) SNet: 2D Semantic Segmentation (b) Reprojection Layer (c) TNet: 3D Semantic Scene Completion

(a) SNet: 2D Semantic Segmentation

- 2D convolutions are organized as an encoder-decoder architecture with skip connections.
- Whichever input is given, SNet maintains the same architecture.

(b) 2D-3D Reprojection Layer

- The 2D semantic segmentation results are reprojected into 3D space.

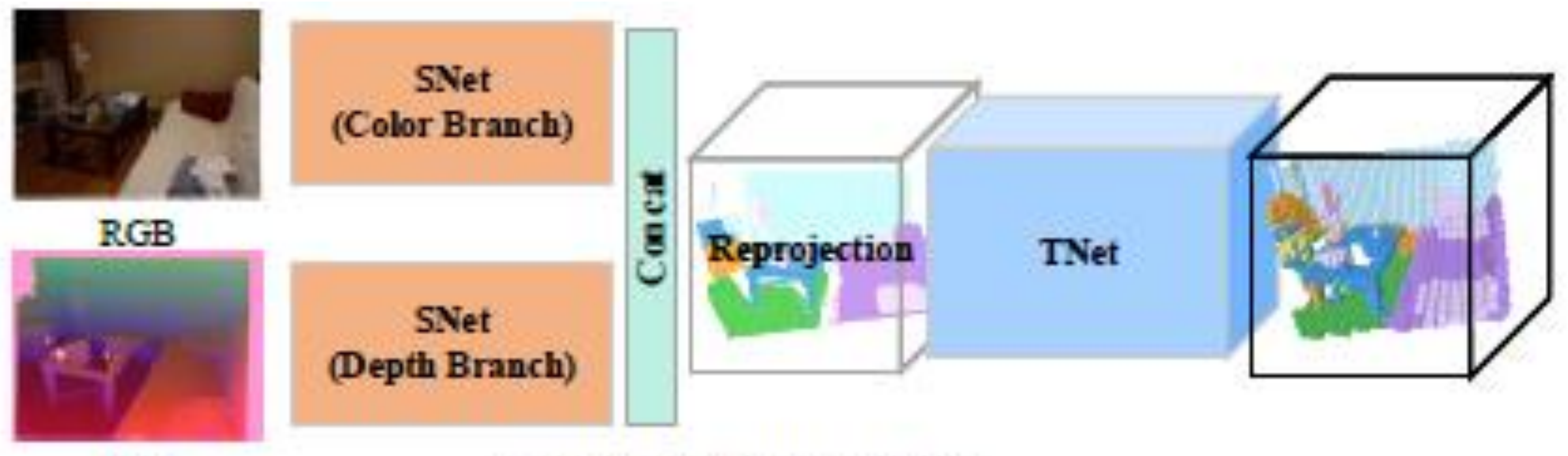
(c) TNet: 3D Semantic Scene Completion

- 3D convolutions are utilized to produce semantic scene completions by the 3D semantic surfaces of the scene.
- TNet cares only about the semantic segmentation results, instead of the types of inputs.

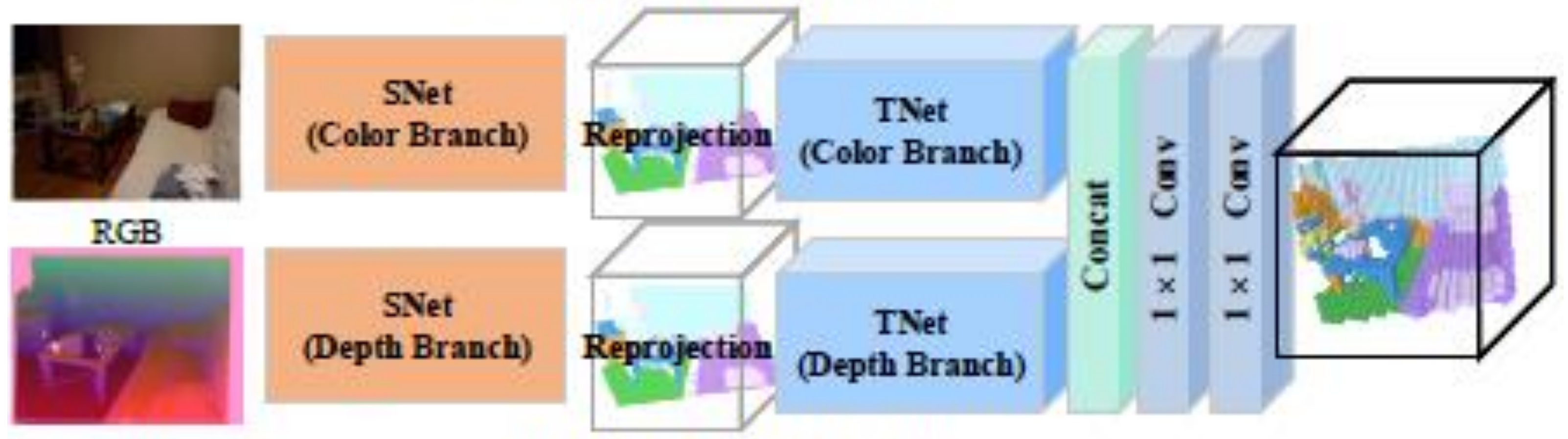
Double-branch RGB-D Fusion

(a) RGB-D Fusion in SNet: The semantic segmentation of RGB-D images is introduced for semantic scene completion.

(b) RGB-D Fusion in TNet: The two semantic scene completions generated by RGB and depth are integrated at the end of the TNet.



(a) RGB-D Fusion in SNet



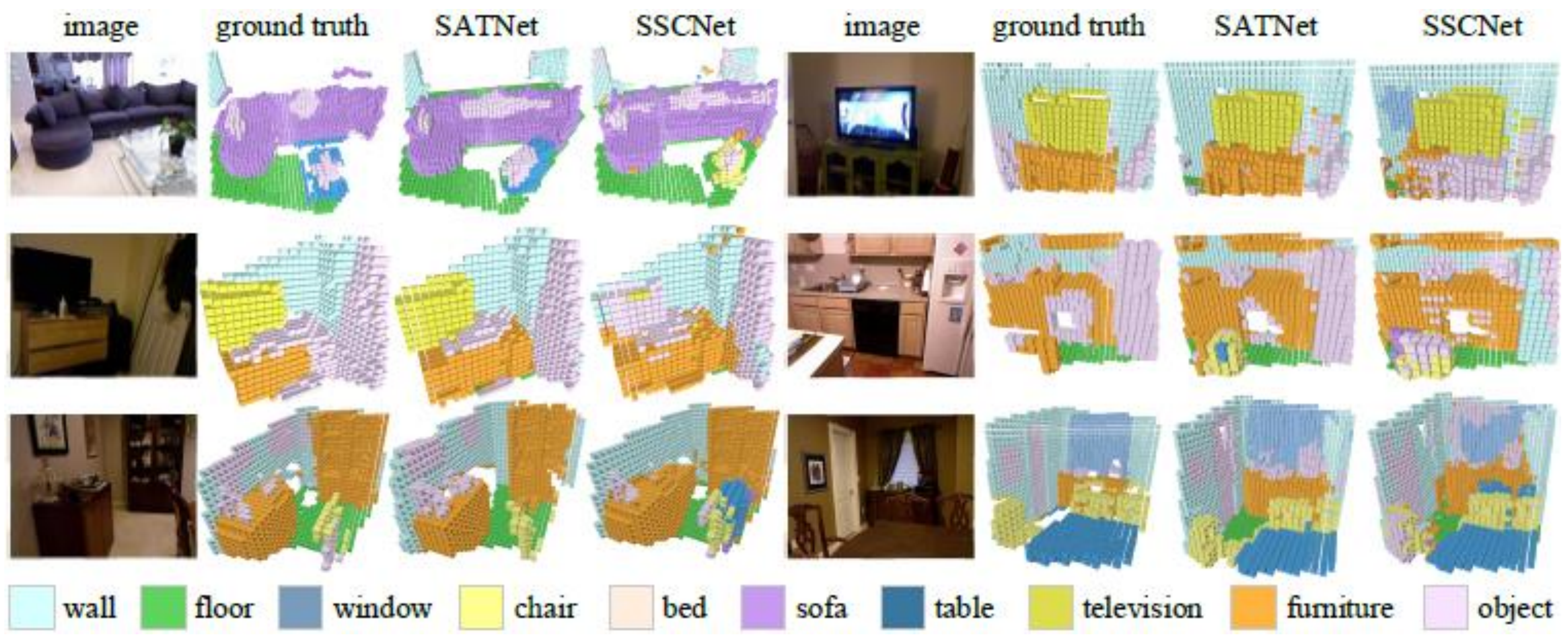
(b) RGB-D Fusion in TNet

4. Experimental Results

Datasets:

- **SUNCG:** A synthetic dataset with 40k~100k training samples.
  - SUNCG-D: Composed of depth images and annotations.
  - SUNCG-RGBD: Composed of RGB images, depth images and annotations.
- **NYUv2:** A real dataset whose ground truths are annotated with CAD model library.

Qualitative Results:



Quantitative Results on NYUv2:

	Scene completion			Semantic scene completion											
method	prec.	recall	IoU	ceil.	floor	wall	win.	chair	bed	sofa	table	tvs.	furn.	objs.	avg.
Lin [29]	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 [26]	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
Song [1]	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
Guedes [6][1]	62.5	82.3	54.3	-	-	-	-	-	-	-	-	-	-	-	27.5
Garbade [24]	**69.5**	82.7	**60.7**	12.9	92.5	25.3	20.1	16.1	56.3	43.4	17.2	10.4	33.0	14.3	31.0
Depth	66.8	86.6	60.6	20.6	91.3	27.0	9.2	**19.5**	56.9	**54.7**	16.9	15.2	37.1	15.7	33.1
RGBD	69.2	81.2	59.5	**22.5**	87.0	**30.0**	**21.1**	17.9	52.4	44.5	15.1	**19.5**	36.0	17.3	33.0
SNetFuse	67.6	85.9	**60.7**	22.2	91.0	28.6	18.2	19.2	56.2	51.2	16.2	12.2	37.0	17.4	33.6
TNetFuse	67.3	85.8	60.6	17.3	92.1	28.0	16.6	19.3	**57.5**	53.8	**17.7**	18.5	**38.4**	**18.9**	**34.4**

Quantitative Results on SUNCG-D:

	Scene completion			Semantic scene completion											
method	prec.	recall	IoU	ceil.	floor	wall	win.	chair	bed	sofa	table	tvs.	furn.	objs.	avg.
Song [1]	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
Depth	**80.7**	**96.5**	**78.5**	**97.9**	82.5	**57.7**	**58.5**	**45.1**	**78.4**	**72.3**	**47.3**	**45.7**	**67.1**	**55.2**	**64.3**

Quantitative Results on SUNCG-RGBD:

	Scene completion			Semantic scene completion											
method	prec.	recall	IoU	ceil.	floor	wall	win.	chair	bed	sofa	table	tvs.	furn.	objs.	avg.
Song [1]	43.5	90.7	41.5	64.9	60.1	**57.6**	25.2	25.5	40.4	37.9	23.1	29.8	45.7	4.7	37.7
Depth	52.3	92.7	50.2	62.5	57.8	48.6	**58.5**	24.4	46.5	50.4	26.9	**41.1**	40.7	20.2	43.4
RGBD	49.8	94.3	48.3	59.0	45.0	46.0	50.6	24.9	42.0	49.0	26.8	40.8	**46.6**	22.4	41.2
SNetFuse	**56.7**	91.7	**53.9**	**65.5**	**60.7**	50.3	56.4	26.1	**47.3**	43.7	**30.6**	37.2	44.9	**30.0**	44.8
TNetFuse	53.9	**95.2**	52.6	60.6	57.3	53.2	52.7	**27.4**	46.8	**53.3**	28.6	**41.1**	44.1	29.0	**44.9**

5. Main References

[1] Shuran Song et al. Semantic Scene Completion from a Single Depth Image. In CVPR, 2017.

[2] Andre B. S. Guedes et al. Semantic Scene Completion Combining Colour and Depth: Preliminary Experiments. arXiv:1802.04735, 2018.

[3] Martin Garbade et al. Two Stream 3D Semantic Scene Completion. arXiv:1804.03550, 2018.