EdgeNet: Semantic Scene Completion from RGB-D images

Aloisio Dourado, Teofilo Emidio de Campos
University of Brasilia
Brasilia, Brazil
t.decampos@st-annes.oxon.org

Hansung Kim, Adrian Hilton
University of Surrey
Surrey, UK
(h.kim, a.hilton)@surrey.ac.uk

Abstract

Semantic scene completion is the task of predicting a complete 3D representation of volumetric occupancy with corresponding semantic labels for a scene from a single point of view. Previous works on Semantic Scene Completion from RGB-D data used either only depth or depth with colour by projecting the 2D image into the 3D volume resulting in a sparse data representation. In this work, we present a new strategy to encode colour information in 3D space using edge detection and flipped truncated signed distance. We also present EdgeNet, a new end-to-end neural network architecture capable of handling features generated from the fusion of depth and edge information. Experimental results show improvement of 6.9% over the state-of-the-art result on real data, for end-to-end approaches.

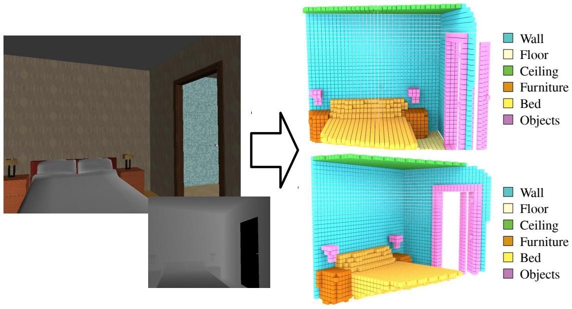


Figure 1: Semantic scene completion. Given an RGB-D image, the goal is to infer a complete 3D occupancy grid with associated semantic labels. For instance, part of the furniture is occluded by the bed, but its 3D reconstruction should handle occlusions.

1. Introduction

The ability of reasoning about scenes in 3D is a natural task for humans, but remains a challenging problem in Computer Vision [7]. Knowing the complete 3D geometry of a scene and the semantic labels of each 3D voxel has several applications, like robotics, surveillance, assistive computing, augmented reality and many others.

Given a partial 3D scene model generated from a single RGB-D image, the goal of scene completion is to generate a complete 3D voxelized volumetric representation where each voxel is labelled as occupied by some object or free space. In addition, for occupied voxels, the goal of *semantic* scene completion is to assign a label that indicates to which class of object it belongs, as shown on the right part of Figure 1.

Our work focuses on semantic scene segmentation using depth and colour. In order to address the RGB data sparsity issue, we introduce a new strategy for encoding information extracted from RGB image after projection from 2D to 3D. We also present and evaluate a new end-to-end 3D CNN architecture to deal with all the features gathered after fusion of colour and depth. We propose a lightweight framework and pipeline to train deep 3D semantic scene completion CNNs with lower memory and time requirements than previous implementations. Comprehensive experiments are conducted to evaluate the proposed solution. Results show that our solution is superior to previous works.

To summarise, our main contributions are:

- a new strategy for encoding colour into a 3D volume to address the sparsity problem of RGB data;
- a new end-to-end convolutional network architecture that benefits from the fusion of depth and colour for semantic scene segmentation;
- a lightweight framework to train deep 3D CNNs.

2. Related Work

Scene Semantic Completion (SSC) in 3D is a problem that was established quite recently and has a high computational cost due to the volume of 3D data.

Song *et al*. [10] used a large synthetic dataset (SUNCG) to generate approximately 140 thousand depth maps that were used to train a typical contracting fully convolutional CNN with 3D dilated convolutions, called SSCNet. They showed that jointly training for segmentation and completion leads to better results, as both tasks are inherently intertwined. Zhang *et al.* [11] used Spatial Group Convolution (SGC) and a U-Net shaped network [8] for accelerating the computation of 3D dense prediction. SGC was used in the encoding branch of the U-Net, while regular 3D convolutions and 3D transposed convolutions were used in the decoding branch. Guedes *et al.* [2] reported preliminary results obtained by adding colour to an SSCNet-like architecture [10]. They used three extra projected volumes, corresponding to the channels of the RGB image, with no encoding, resulting in 3 sparse cubes. The authors reported no significant improvement using the colour information in this sparse manner. Liu *et al.* [6] used depth maps and RGB information as input of an encoder-decoder 2D segmentation CNN. The encoder branch of the 2D CNN is a ResNet-101 [5] and the decoder branch contains a series of dense upsampling convolutions. The generated features from the 2D CNN are then reprojected to 3D using camera parameters, before being fed into a 3D CNN.

Using 2D segmentation maps on 3D SCC brings an additional complexity to the training phase which is training and evaluating the 2D segmentation network prior to the 3D CNN training. In this work, we focus on end-to-end approaches, where the whole network can be trained and evaluated as a whole.

3. Our Approach

We introduce a new strategy to fuse colour appearance and depth information for 3D SSC. Our approach consists on detecting edges in the image, which gives a 2D binary representation of the scene that can highlight flat objects on flat surfaces. For instance, a poster on a wall is expected to be invisible in a depth map, especially after down-sampling. On the other hand, RGB edges highlight the presence of that object. We use the standard Canny edge detector [1] to perform edge detection. Each edge location is projected to a point in the 3D space using its depth information and the camera calibration matrix. The resulting point cloud is voxelised in the same way as the depth point clout, resulting in a sparse volume of 240 x 144 x 240 voxels.

The main advantage of extracting edges and projecting them to 3D is the possibility to apply F-TSDF on both edges and surface volumes, as they are both binary, thus providing two meaningful input signals to the 3D CNNs.

In order to better capture and aggregate information from both depth and edges, we present a new 3D Semantic Segmentation CNN that we call **EdgeNet**. Our proposed network architecture is a deeper 3D CNN inspired by the U-Net design [8] which has been successfully used in many 2D semantic segmentation problems, and is presented in Figure 2. We address the degradation problem of deeper networks, by replacing simple convolutional blocks of U-Net by ResNet modules [5]. To match the resolution of the output, the first 2 stages reduce the resolution to 1/4 of the input. Next blocks follows encoder-decoder design and, following [10], we used dilated convolutions on lower resolutions to improve the receptive field. The last stage is responsible for reducing the number of channels to match the desired number of output classes and loss calculations.

4. Experiments

In this section we describe the datasets and the evaluation protocol used in this paper.

We train and validate our proposed approach on SUNCG [10] and NYUv2 [9] datasets. SUNCG dataset consists of about 45K synthetic scenes from which were extracted more than 130K 3D scenes with corresponding depth maps and ground truth divided in train and test datasets.

NYUv2 dataset includes depth and RGB images captured by the Kinect depth sensor divided in 795 depth images for training and 654 for test. Following the majority of works in semantic segmentation we used ground truth by voxelizing the 3D mesh annotations from [3] and and mapped object categories based on [4].

We follow exactly the same evaluation protocol as [10], with the same test datasets. For the semantic scene completion task, we report the IoU of each object class on both the observed and occluded voxels. For the scene completion task, all non-empty object classes are considered as one category, and we report Precision, Recall and IoU of the binary predictions on occluded voxels. Voxels outside the view or the room are not considered.

5. Results

In this section we report quantitative results of EdgeNet on NYUv2 and compare them to other end-to-end approaches.

Table 1 shows the results of EdgeNet on NYUv2 dataset and compare them to recent end-to-end semantic scene completion approaches, for models trained only on synthetic data, only on NYU and on both synthetic and NYU using fine tuning. We present results extracted from their original papers. Overall, EdgeNet achieved the best scores on each one of the training scenarios, improving the state-of-art on 3D SSC on NYU. Considering training on SUNCG and fine tuning on NYU the improvement was 3% on scene completion and 6.9% on semantic scene completion.

Qualitative results on NYU are shown in Figure 3. Models used to generate the inferences were trained on SUNCG and fine tuned on NYU. We compare results of SSCNet* to our model. It is visually perceptible that EdgeNet presents

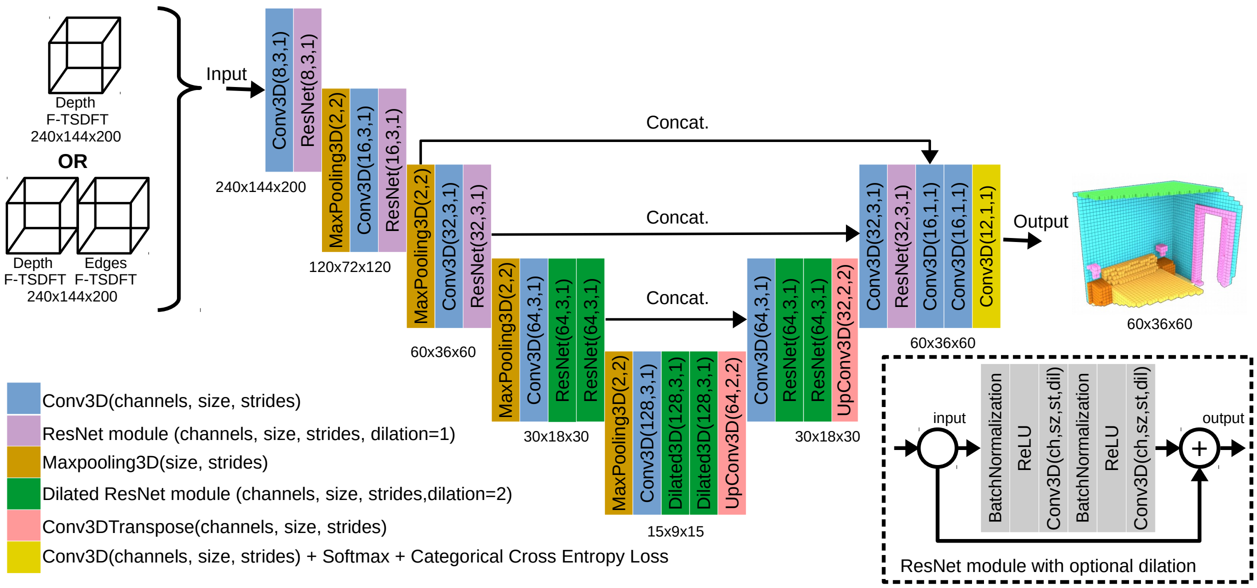


Figure 2: The proposed U-shaped architecture with two possible sets of input channels: the proposed EdgeNet, which uses depth and edges, and U-SSCNet, which has the same architecture but uses only depth as input (best viewed in colour).

train	input	model	scene completion			semantic scene completion (IoU, in percentages)											
			prec.	rec.	IoU	ceil.	floor	wall	win.	chair	bed	sofa	table	tvs	furn.	objs.	avg.
SUNCG	d	SSCNet[10]	55.6	**91.9**	53.2	**5.8**	81.8	**19.6**	**5.4**	12.9	34.4	26	**13.6**	**6.1**	9.4	7.4	20.2
	d+e	EdgeNet(Ours)	**59.4**	84.3	**53.5**	4.7	**88.1**	15.0	5.3	**13.8**	**42.5**	**28.5**	8.8	3.0	**12.5**	**9.7**	**21.1**
NYU	d	SSCNet[10]	57.0	**94.5**	55.1	15.1	94.7	24.4	0.0	**12.6**	32.1	35.0	**13.0**	**7.8**	27.1	10.1	24.7
		SGC[11]	71.9	71.9	**56.2**	17.5	75.4	25.8	**6.7**	15.3	**53.8**	**42.4**	11.2	0.0	33.4	11.8	26.7
	d+e	EdgeNet(Ours)	**78.4**	66.2	56.0	**19.7**	**94.9**	**28.1**	0.0	7.5	52.5	41.8	10.4	0.0	**34.7**	**12.8**	**27.5**
SUNCG +NYU	d	SSCNet[10]	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
	d+c	Guedes *et al.* [2]	-	-	56.6	-	-	-	-	-	-	-	-	-	-	-	30.5
	d+e	EdgeNet(Ours)	**76.3**	71.1	**58.3**	**23.6**	**95.0**	**28.6**	**12.6**	13.1	**57.7**	**51.1**	**16.4**	9.6	**37.5**	**13.4**	**32.6**

Table 1: **Semantic scene completion results of end-to-end approaches on NYU test set**. Column 'input' indicates the type of input: d=depth only; d+e=depth and edges. Column 'train' indicates dataset used for training the models. SUNCG + NYU means trained on SUNCG and fine tuned on NYU. EdgeNet presented the best results on all training scenarios.

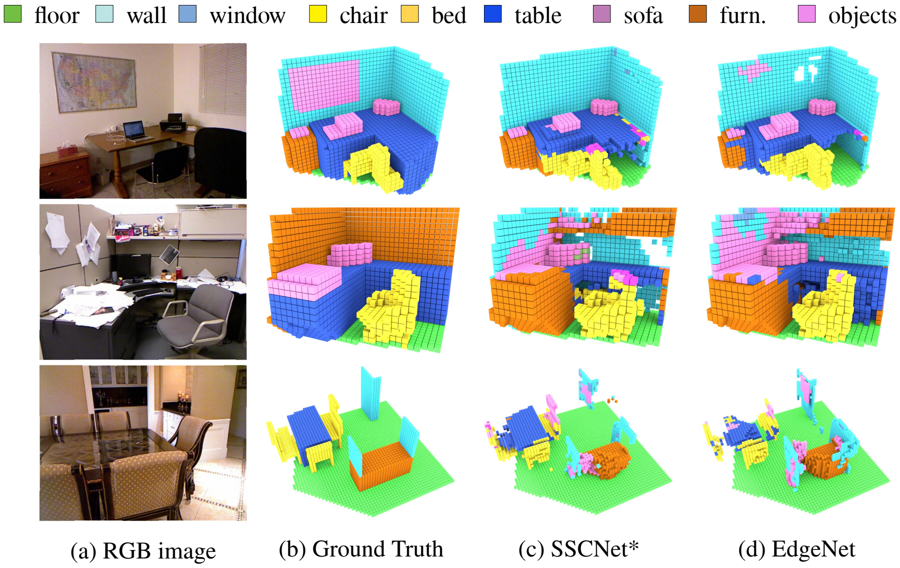
the best results.

In the first row of images of Figure 3, note how EdgeNet correctly captures the details of the laptop and other small objects on the table. The effect of the use of edges over flat objects is made clear on the forth row. While SSCNet and U-SSCNet are incapable of distinguishing the posters on the wall, all edge networks highlight the presence of those objects, being EdgeNet more precise.

The second row of Figure 3 depicts some problems related to Ground Truth annotations on NYU dataset. Note that neither the papers fixed on the wall nor the shelf appear in the Ground Truth. Most of the models captured the shelf, but only EdgetNet inferred the presence of objects fixed on the wall. When quantitative results are computed, ground truth annotation flaws like these unfairly benefit the less precise models and harm more precise models like EdgeNet.

6. Conclusion

This paper presented a new approach to fuse depth and colour into a CNN for semantic scene completion. We introduced the use of F-TSDF encoded 3D projected edges extracted from RGB images. We also presented a new end-to-end network architecture capable of properly aggregating edges and depth and extracting useful information from both sources, with no requirement for previous 2D seman-



(a) RGB image (b) Ground Truth (c) SSCNet* (d) EdgeNet

Figure 3: **Qualitative Results**. We compare our results (EdgeNet) with scene completion results from Song *et al*. [10] on SUNCG and NYU. Overall, EdgeNet gives more accurate voxel predictions (best viewed in colour).

tic segmentation training as previous depth plus colour approaches. Experiments with alternate models, showed that both aggregating edges and the new proposed architecture have positive impact on semantic scene completion, especially in hard to detect objects. Qualitative results show visually perceptible improvements in 3D label inferences and we have achieved improvement over the state-of-the-art result on the NYU depth v2 dataset, for end-to-end approaches.

We developed a lightweight training pipeline for the task, which reduced the memory footprint in comparison to the original implementation of SSCNet and reduced the training time on SUNCG from 7 to 4 days and on NYU from 30 to 6 hours.

References

[1] J. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679–698, Nov 1986. 2

[2] A. B. S. Guedes, T. E. de Campos, and A. Hilton. Semantic scene completion combining colour and depth: preliminary experiments. *CoRR*, abs/1802.04735, 2018. 2, 3

[3] R. Guo, C. Zou, and D. Hoiem. Predicting complete 3D models of indoor scenes. *CoRR*, abs/1504.02437, 2015. 2

[4] A. Handa, V. Patraucean, V. Badrinarayanan, S. Stent, and R. Cipolla. Scenenet: Understanding real world indoor scenes with synthetic data. *CoRR*, abs/1511.07041, 2015. 2

[5] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016. 2

[6] S. Liu, Y. HU, Y. Zeng, Q. Tang, B. Jin, Y. Han, and X. Li. See and think: Disentangling semantic scene completion. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Conference on Neural Information Processing Systems (NeurIPS)*, pages 263–274. Curran Associates, Inc., 2018. 2

[7] D. Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. MIT Press, 1982. 1

[8] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. 2

[9] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, editors, *Computer Vision – ECCV 2012*, pages 746–760, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. 2

[10] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser. Semantic Scene Completion from a Single Depth Image. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2, 3, 4

[11] J. Zhang, H. Zhao, A. Yao, Y. Chen, L. Zhang, and H. Liao. Efficient semantic scene completion network with spatial group convolution. In *The European Conference on Computer Vision (ECCV)*, September 2018. 2, 3