Semantic Scene Completion Combining Colour and Depth:
preliminary experiments

André B. S. Guedes   Teófilo E. de Campos
Universidade de Brasília
Brasília-DF, 70910-900, Brazil
`t.decampos@st-annes.oxon.org`
`http://www.cic.unb.br/~teodecampos/`

Adrian Hilton
CVSSP, University of Surrey
Guildford, GU2 7XH, UK
`a.hilton@surrey.ac.uk`

Abstract

Semantic scene completion is the task of producing a complete 3D voxel representation of volumetric occupancy with semantic labels for a scene from a single-view observation. We built upon the recent work of Song et al. [13], who proposed SSCnet, a method that performs scene completion and semantic labelling in a single end-to-end 3D convolutional network. SSCnet uses only depth maps as input, even though depth maps are usually obtained from devices that also capture colour information, such as RGBD sensors and stereo cameras. In this work, we investigate the potential of the RGB colour channels to improve SSCnet.

1. Introduction

The task of reasoning about scenes in 3D is one of the seminal goals of Computer Vision [8]. If the 3D geometry of a scene is known, robots are able to plan trajectories, avoid collisions or clean surfaces. If the semantic labels of each surface or voxel is also known a robot can also figure interact with the environment and perform more complex tasks, such as moving objects from one location to another; opening/closing doors, drawers, windows; operating kitchen appliances etc. Three-dimensional maps with labelled voxels have several other applications, including surveillance, assistive computing, augmented reality and so on. One issue is that capturing the full geometry of a scene can be time consuming (if it is done using a scanning technique [9]) or expensive (if it is done using a rig of calibrated sensors).

It is well known that vision is a combination of so called bottom-up and top-down processes [8]. Bottom-up information can be obtained by matching local features for stere-

opsis and top-down is the use of prior knowledge from related scenes and objects. If both types of cues are combined, it is possible to estimate a complete scene geometry by using a single visual and depth map of a scene. This is well illustrated in Figure 2 of [13]. A visual sensor captures a single view of a scene which provides measurements (e.g. RGB and Depth) of the visible objects but it is not possible to measure the geometry of occluded regions. However, if the class of the objects is identified, it is possible to infer the complete scene geometry, enabling a full 3D representation to be proposed.

Solid computational demonstrations of this have started to be published recently. Notably, Song et al. [13] introduced the problem of Semantic Scene Completion (SSC), i.e., given an depth map, the goal is to generate a 3D image where each voxel is associated to one out of $N+1$ labels, where there are N known object labels plus an 'empty space' label.

In [13], this problem is approached using a Deep 3D Convolutional Neural Network coined SSCNet. That paper demonstrates impressive results on completing and labelling a full 3D scene generated from a single depth map. Using a combination of bottom-up dues (from the depth sensor) and top-down cues (learnt from the training set), their method is able to infer the geometry and labels of the whole scene, including heavily occluded regions, such as the regions under tables and behind sofas, as illustrated in Figure 1 of [13].

However, one of the main limitations of SSCNet is that it was not designed to use any colour information, only depth maps are used. This clearly impairs the method as indoor scenes generally include various sources of error in depth and geometry estimation. Highly reflective scenes with glass, mirrors or shiny surfaces usually induce false depth. If depth is captured using stereo cameras, textureless and non-Lambertian surfaces often result in errors in feature detection and matching. Colour information also

disambiguates between different objects that have similar shape or that are co-planar, like posters on the wall. Furthermore, it is clear that colour offers crucial information for semantic labelling that strongly complement depth information, as seen in papers that focus on semantic segmentation from RGBD images, e.g. [11, 12, 7, 4, 14, 1, 2].

In this paper, we propose to use colour in addition to depth for Semantic Scene Completion. For that, we propose modifications of the SSCnet architecture in order to fuse RGB and depth. A new input layer was proposed to encode colour in the visible frustum and we combined a feature extraction training technique for multiple view learning.

2. Colour SSCNet

Depth maps are acquired using an RGB-D sensor and using the sensor's intrinsic calibration parameters, a 3D point cloud is generated. The observed geometry is then encoded using flipped Truncated Signed Distance Function (fTSDF), proposed in [13]. This method associates a value to each point in the 3D space to a function of its distance to the nearest surface point. The sign of this value indicates if it is visible or occluded. Apart from the occlusion coding, this method is viewpoint independent.

The fTSDF encoding of voxels describe the geometry of the space, but it does not carry any information about the colour or grey level of the visible objects. We propose to encode the RGB values of the visible surfaces in another voxel representation of the scene. The three channels are normalised to range from 0 to 1. Empty spaces and occluded regions are coded with the -1 value for the three colour channels.

We apply these two encoding techniques to RGB and Depth signals and run them through a 3D CNN that learns to map from RGBD to a labelled 3D volume. Labelled volumes were obtained as described in [13], i.e., the binvox voxelisation technique [10] was applied to 3D models, which accounts for both surface and interior voxels using a space carving approach.

We built upon the 3D CNN architecture of SSCNet [13]. To combine RGB and Depth, we propose the two fusion schemes described below.

Early fusion. The first layer of SSCNet was adapted so that it takes as input a concatenation of fTSDF and the three colour channels encoded as described above. The remaining of the network is the same.

Mid-level fusion. This architecture is depicted in Figure 1, drawn using Caffe [6] (better viewed on a screen). The numbers in brackets are: kernel size, stride, pad and dilation factor, respectively. The branch on the left is essentially a copy of SSCNet. A colour 3D CNN was built following a similar architecture to SSCNet up to the concatenation layer. This layer originally aggregated the output of five scales gathered from previous layers and it is followed
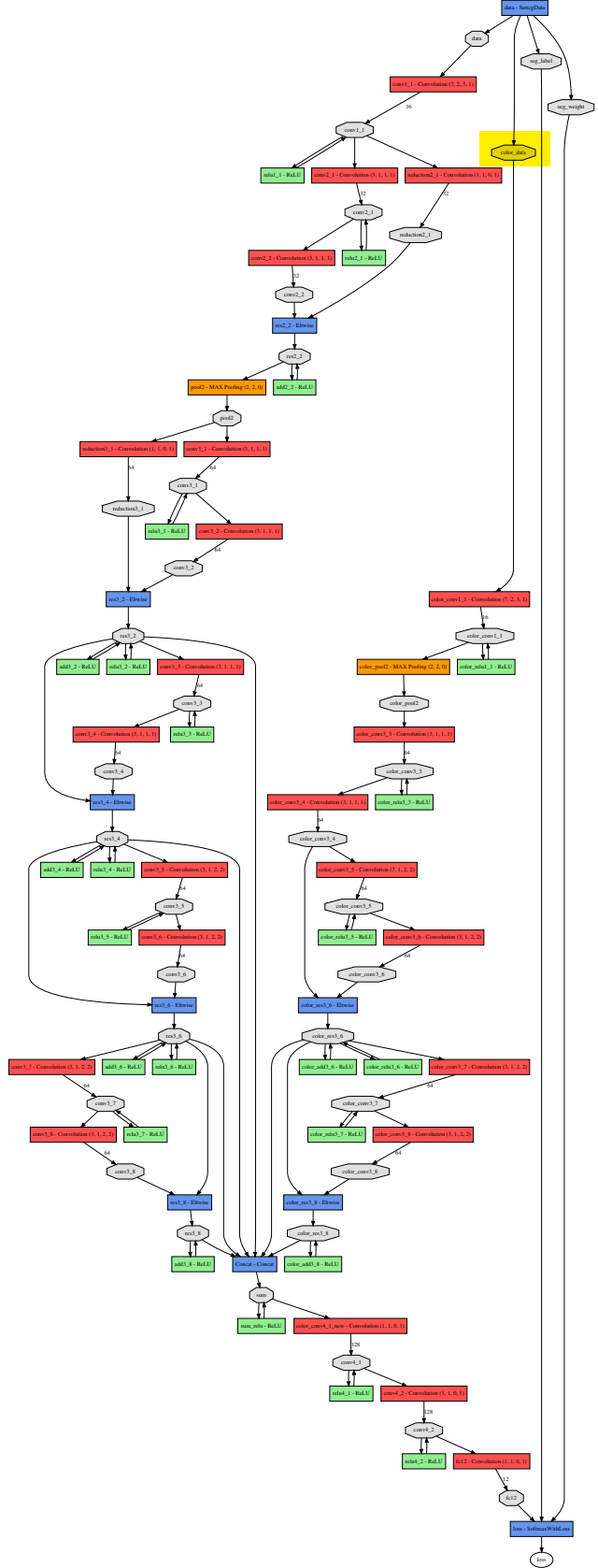

Figure 1. Network architecture for mid-level fusion.

by three 3D convolutional layers. Although this network does not have any fully connected (FC) layer, the last three layers perform the same class as FC layers in classification CNNs, as they are closer to the output, which produces labelled data. A typical mid-level fusion using CNNs is done at the input of the first FC layer. Therefore, we believe the most appropriated layer to fuse SSCNet-like sub-networks is at the concatenation layer.

However, the original SSCNet is very memory-intensive and it requires about 7GB of GPU memory to process a single depth image[1]. If we were to duplicate all layers of the network up to the concatenation layer, much more than 12GB would be required, whereas most GPU models available nowadays have up to 12GB of RAM. To add to the challenge, our RGB coding uses three channels per image, rather than one as in the fTSDF model, though this only affects the first convolutional layer. Therefore, some of the convolutional layers were removed from the colour branch of SSCNet, but we have preserved all the dilated convolution layers, as this is a significant feature of SSCNet which widely expands the receptive field of the network [15].

In addition, we also evaluated a **colour-only SSCNet**, which follows the same architecture as the colour branch of the mid-level RGB-D fusion network, but it is followed by the top three convolutional layers, without aggregating activations from the fTSDF branch.

3. Experiments and Training Strategies

Our evaluations focused on the NYU depth v2 dataset [12], using the standard split of 795 training samples and 654 test samples[2]. However, instead of the standard semantic segmentation labels, we used the labels devised for scene completion, where objects are grouped into 7 categories plus window, wall, floor, ceiling and another category that identifies free space. This set of labels originated from [5]. As explained in [13], ground truth volumes were obtained from 3D mesh annotations of [3]. Our implementation was developed using the Caffe framework [6].

We evaluated the two architectures proposed in Section 2: early and mid-level fusion and compared it against the original SSCNet and colour-only. For all methods that we proposed, training was done following these strategies:

- **Random initialisation**: all parameters were randomly initialised and the whole network was trained from scratch.

- **Feature learning**: we kept the original SSCNet parameters trained by Song et al. [13] for all the original layers and optimised only the colour layers, i.e., the original SSCNet parameters were frozen.

- **Fine tuning**: this is similar to the strategy above, except that instead of freezing the original layers, we also enabled their parameters to be optimised, but with the learning rate ratio of 0.2 times the ratio of the new layers.

- **Surgery**: was applied only for the early fusion approach. It is similar to fine tuning, except that the weights of the input layer which related to depth were set to the original parameters of the first layer of SSCNet and the other weights (linked to the colour channels) of the same convolutional kernel were initialised randomly.

Voxel labelling is done by applying soft-max to the scores of the last convolutional layer of the networks and optimisation is done using cross-entropy as a loss function, averaged out over all classes.

The results were evaluated using the Intersection over the Union (IoU) between predicted class labels and ground truth, averaging out over all voxels in the test set and all classes. We followed [13] and evaluated our results both in terms of completion (i.e., the ability to detect if an occluded voxel is occupied or free space) and in terms of semantic labelling of voxels of all classes.

4. Results and Discussion

Our results so far show that none of the proposed architectures and training strategies actually lead to results that are better than the original SSCNet based only on depth observations, i.e., through the training iterations, our results peaked at scene completion IoU of 56.6 and average semantic scene completion of 30.5, which are both results obtained by the original SSCNet on the test set of the NYU depth v2 dataset. In other words, our experiments in the NYU depth v2 dataset (with the 12 category labels [5]) show that the proposed method for coding colour information is not as discriminative as fTSDF, neither it complements depth information.

However, the performance of our colour-only network, initialised with randrom weights, followed a monotonic increase as the number of training iterations increased, though it did not converge with the same number of iterations as the architectures that use depth. Therefore, there is certainly relevant information in RGB, but it should probably be combined with fTSDF in a different way, perhaps using late fusion. Even if early or mid-level fusion are not the ideal strategies in this problem, further investigation is also needed to understand why RGB has not complemented

---

[1]The original implementation from the authors, obtained from `https://github.com/shurans/sscnet`, actually requires almost 12GB. We removed some redundancy from their code, freeing about 5GB of memory.

[2]We used the train+validation split for training and the test split for testing, following the sample indices available from `https://github.com/shelhamer/fcn.berkeleyvision.org/tree/master/data/nyud`.

Depth at all. It might be an artefact of the dataset and set annotated classes, as it is possible that geometry alone is already very discriminative. A suggestion is to verify this using more complex scenes with more occlusions or with finer object class labels.

Our results have also shown that unconstrained Fine Tuning leads to a higher decrease in the loss function than the constrained optimisation methods (Feature Learning and Surgery). However, after 1000 iterations, the test set performance (measured by IoU) starts to decrease due to over-fitting. Although the loss is lower for Fine Tuning, we did not observe a significant difference between the methods in terms of test set performance.

5. Conclusion

In this paper we reported ongoing work that considers the problem of Semantic Scene Completion in 3D from a single RGBD image. Starting from the 3D CNN architecture of SSCNet [13], which used only depth maps as input, we proposed to combine RGB and Depth information using early and mid-level fusion schemes.

Our preliminary results were not better than the original depth-only method. Therefore, further investigation is needed in order to verify if the dataset (NYU depth v2 with 12 labels obtained from [5]) prilidedges structural information such that depth is already very discriminative. A finer set of labels or a more complex dataset should be evaluated. Other directions of future work are to evaluate late fusion scheme and investigate other ways to encode RGB information.

6. Acknowledgements

References

[1] K. Chen, Y.-K. Lai, and S.-M. Hu. 3D indoor scene modeling from RGB-D data: a survey. *Computational Visual Media*, 1(4):267–278, 2015.

[2] D. Eigen and R. Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proc 15th Int Conf on Computer Vision, Santiago, Chile*, 2015.

[3] R. Guo, C. Zou, and D. Hoiem. Predicting complete 3D models of indoor scenes. *arXiv preprint arXiv:1504.02437*, 2015.

[4] S. Gupta, P. Arbelez, R. Girshick, and J. Malik. Indoor scene understanding with RGB-D images: Bottom-up segmenta-

tion, object detection and semantic segmentation. *Int Journal of Computer Vision*, pages 1–17, 2014.

[5] A. Handa, V. Patraucean, V. Badrinarayanan, S. Stent, and R. Cipolla. Understanding real world indoor scenes with synthetic data. In *Proc IEEE Conf on Computer Vision and Pattern Recognition, Las Vegas, NV, June 26 - July 1*, pages 4077–4085, 2016.

[6] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[7] O. Kahler and I. Reid. Efficient 3D scene labeling using fields of trees. In *Proc 14th Int Conf on Computer Vision, Sydney, Australia*, pages 3064–3071, 2013. DOI:10.1109/ICCV.2013.380.

[8] D. C. Marr. *Vision: a computational investigation into the human representation and processing of visual information*. MIT Press, 1982. ISBN-13: 9780262514620, DOI:10.7551/mitpress/9780262514620.001.0001.

[9] R. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. Davison, P. Kohli, J. Shotton, S. Hodges, and A. Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *Proc ISMAR*, 2011.

[10] F. S. Nooruddin and G. Turk. Simplification and repair of polygonal models using volumetric techniques. *IEEE Transactions on Visualization and Computer Graphics*, 9(2):191–205, 2003. Used in the implementation of binvox, available from Patrik Min's website: http://www.patrickmin.com/binvox/.

[11] N. Silberman and R. Fergus. Indoor scene segmentation using a structured light sensor. In *Proc ICCV Workshops*, 2011.

[12] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from RGBD images. In *Proc European Conf on Computer Vision*, pages 746–760, 2012. NYU depth v2 dataset available from http://cs.nyu.edu/~silberman/datasets/nyu_depth_v2.html.

[13] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser. Semantic scene completion from a single depth image. In *Proc 30th IEEE Conf on Computer Vision and Pattern Recognition, Honolulu, Hawaii, July 21-26*, 2017. Preprint available as technical report arXiv:1611.08974.

[14] A. Wang, J. Lu, J. Cai, G. Wang, and T.-J. Cham. Unsupervised joint feature learning and encoding for RGB-D scene labeling. *IEEE Trans Image Processing*, 2015. DOI:10.1109/TIP.2015.2465133.

[15] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. In *Proc Int Conf on Learning Representations ICLR*, Toulon, France, April 24-26 2016. Preprint available as arXiv technical report arXiv:1511.07122.

Semantic Scene Completion Combining Colour and Depth: preliminary experiments

André B. S. Guedes, Teo E. de Campos[1] and Adrian Hilton[2]

University of Brasília, Brazil   t.decampos@st-annes.oxon.org
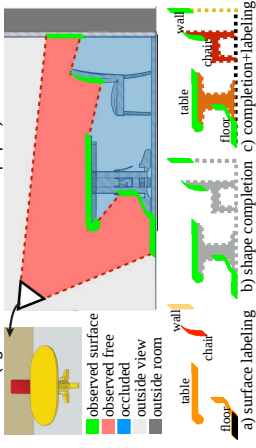CVSSP, University of Surrey, UK   a.hilton@surrey.ac.uk
http://www.cic.unb.br/~teodecampos

UNIVERSITY OF SURREY

UnB

Problem statement

Semantic Scene Completion

- **Introduced by Song et al. CVPR 2017**
  (figure extracted from their paper)
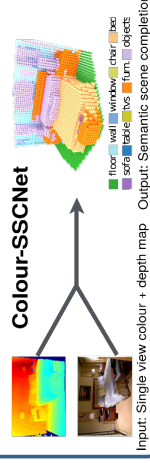

observed surface
observed free
occluded
outside view
outside room

wall
chair
table
floor

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

Previous method

SSC from depth only
by Song et al. CVPR 2017

Input: Single view depth map

SSCNet

floor | wall | window/ chair | bed
sofa | table | tvs | furn. | objects
Output: Semantic scene completion

This work

Colour-SSCNet

Input: Single view colour + depth map

floor | wall | window/ chair | bed
sofa | table | tvs | furn. | objects
Output: Semantic scene completion

Fusion schemes

Early fusion: concatenation at the first layer

conv (16,7,2,1)
conv (32,3,1,1)
conv (32,3,1,1)
add
conv (32,1,1,1)
conv (64,3,1,1)
conv (64,3,1,1)
add
conv (32,1,1,1)
conv (64,3,1,1)
conv (64,3,1,1)
add
dilated (64,3,1,2)
dilated (64,3,1,2)
add
dilated (64,3,1,2)
dilated (64,3,1,2)
add
concat
conv (128,1,1,1)
conv (128,1,1,1)
conv (12,1,1,1)

Mid-level fusion

conv (16,7,2,1)
conv (32,3,1,1)
conv (32,3,1,1)
add
conv (32,1,1,1)
conv (64,3,1,1)
conv (64,3,1,1)
add
dilated (64,3,1,2)
dilated (64,3,1,2)
add
dilated (64,3,1,2)
dilated (64,3,1,2)
add

pooling
conv (16,7,2,1)
conv (64,3,1,1)
conv (64,3,1,1)
add
dilated (64,3,1,2)
dilated (64,3,1,2)
add
dilated (64,3,1,2)
dilated (64,3,1,2)
add

concat
conv (128,1,1,1)
conv (128,1,1,1)
conv (12,1,1,1)

pooling

Encoding and learning

Depth and Colour encoding

- The input point cloud is converted to a 3D volume aligned with gravity and following Manhattan assumptions.
- **Depth** is encoded using the flipped Truncated Signed Distance Function (fTSDF).
- **Colour** is encoded for each voxel as an RGB triplet.
  - To compute the colour of a voxel, the inverse of the voxelation process is applied and colour is computed by averaging out the original RGB region.
  - Colour values are normalised to be between 0 and 1.
  - Occluded or empty voxels have their colour set to -1.

Learning strategies

- For the early fusion network, we tried these strategies:
  - **Feature learning**: where only the first layer was trained to learn colour parameters and the remaining parameters where locked.
  - **Fine tuning**: where a learning rate of 0.2 was used for the previously learnt layers and 1.0 for the new layer.
- For the mid-level fusion network, we also tried learning all parameters from scratch with random initialisation.

Results on the NYU dataset

Scene completion
- inferring geometry in 3D

	Learning Strategy	Precision (%)	Recall (%)	IoU (%)
Early Fusion	Fine tuning	59.4	92.4	56.3
	Feature learning	58.9	93.1	56.1
Mid-level fusion	Fine tuning	61.50	84.11	54.20
	Feature learning	62.46	82.26	54.31
	Learning from scratch	45.60	100.0	46.60
Depth only	Song et al CVPR'17	59.3	92.9	56.6

Semantic Scene Completion
- inferring labels in 3D

|	Learning Strategy	Precision (%)	Recall (%)	IoU (%)
Mid-level fusion	Fine tuning	36.92	34.23	17.19
	Feature learning	–	45.03	27.45
	Learning from scratch	–	18.30	05.92
Colour only	Fine tuning	21	46	16
Depth only	Song et al CVPR'17	–	–	30.50

Sample result

The results bellow illustrates one of the potential causes for the degeneration of our results
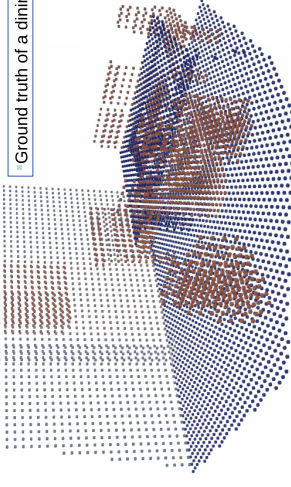
Ground truth of a dining room scene

Top view of prediction
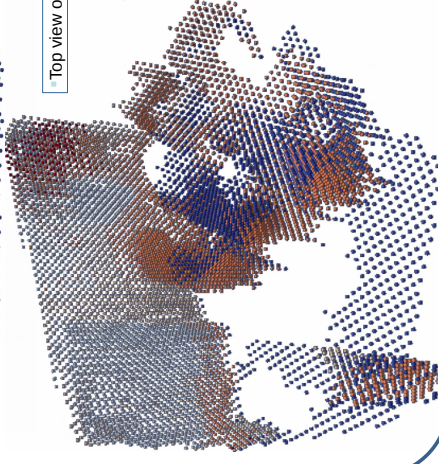
Observations

- Phantom cones are being predicted.
- It is likely that the data balancing strategy is not inducing a high penalty when an empty voxel is predicted as containing an object.
- Colour information will be more crucial on datasets with finer object labels (e.g., poster on the wall).