Semantic Scene Completion from a Single 360-Degree Image and Depth Map

Aloisio Dourado[1][a], Hansung Kim[2][b], Teofilo E. de Campos[1][c] and Adrian Hilton[2][d]
[1]*University of Brasilia, Brasilia, Brazil*
[2]*CVSSP, University of Surrey, Surrey, U.K.*

Keywords: Semantic Scene Completion, 360-Degree Scene Reconstruction, Scene Understanding, 360-Degree Stereo Images.

Abstract: We present a method for Semantic Scene Completion (SSC) of complete indoor scenes from a single 360° RGB image and corresponding depth map using a Deep Convolution Neural Network that takes advantage of existing datasets of synthetic and real RGB-D images for training. Recent works on SSC only perform occupancy prediction of small regions of the room covered by the field-of-view of the sensor in use, which implies the need of multiple images to cover the whole scene, being an inappropriate method for dynamic scenes. Our approach uses only a single 360° image with its corresponding depth map to infer the occupancy and semantic labels of the whole room. Using one single image is important to allow predictions with no previous knowledge of the scene and enable extension to dynamic scene applications. We evaluated our method on two 360° image datasets: a high-quality 360° RGB-D dataset gathered with a Matterport sensor and low-quality 360° RGB-D images generated with a pair of commercial 360° cameras and stereo matching. The experiments showed that the proposed pipeline performs SSC not only with Matterport cameras but also with more affordable 360° cameras, which adds a great number of potential applications, including immersive spatial audio reproduction, augmented reality, assistive computing and robotics.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

1 INTRODUCTION

Automatic understanding of the complete 3D geometry of a indoor scene and the semantics of each occupied 3D voxel is one of essential problems for many applications, such as robotics, surveillance, assistive computing, augmented reality, immersive spatial audio reproduction and others. After years as an active research field, this still remains a formidable challenge in computer vision. Great advances in scene understanding have been observed in the past few years due to the large scale production of inexpensive depth sensors, such as Microsoft Kinect. Public RGB-D datasets have been created and widely used for many 3D tasks, including prediction of unobserved voxels (Firman et al., 2016), segmentation of visible surface (Silberman and Fergus, 2011; Ren et al., 2012; Qi et al., 2017b; Gupta et al., 2013), object detection (Shrivastava and Mulam, 2013) and single object completion (Nguyen et al., 2016).

In 2017, a new line of work was introduced, focusing on the complete understanding of the scene: Semantic Scene Completion (SSC) (Song et al., 2017). SSC is the joint prediction of occupation and semantic labels of visible and occluded regions of the scene. The works in this area are mostly based on the use of Convolution Neural Networks (CNNs) trained on both synthetic and real RGB-D data (Garbade et al., 2018; Guedes et al., 2017; Zhang et al., 2018a; Zhang et al., 2018b; Liu et al., 2018). However, due to the limited field-of-view (FOV) of RGB-D sensors, those methods only predict semantic labels for a small part of the room and at least four images are required to understand the whole scene.

This scenario recently started to change with the use of more advanced technology for large-scale 3D scanning, such as Light Detection and Ranging (LIDAR) sensor and Matterport cameras. LIDAR is one of the most accurate depth ranging devices using a light pulse signal but it acquires only a point cloud set without colour or connectivity. Some recent LIDAR devices provide coloured 3D structure by map-

[a] https://orcid.org/0000-0002-5037-7178
[b] https://orcid.org/0000-0003-4907-0491
[c] https://orcid.org/0000-0001-6172-0229
[d] https://orcid.org/0000-0003-4223-238X

Dourado, A., Kim, H., E. de Campos, T. and Hilton, A.
Semantic Scene Completion from a Single 360-Degree Image and Depth Map.
DOI: 10.5220/0008877700360046
In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2020)*, pages 36-46
ISBN: 978-989-758-402-2
Copyright © 2020 by SCITEPRESS – Science and Technology Publications, Lda. All rights reserved

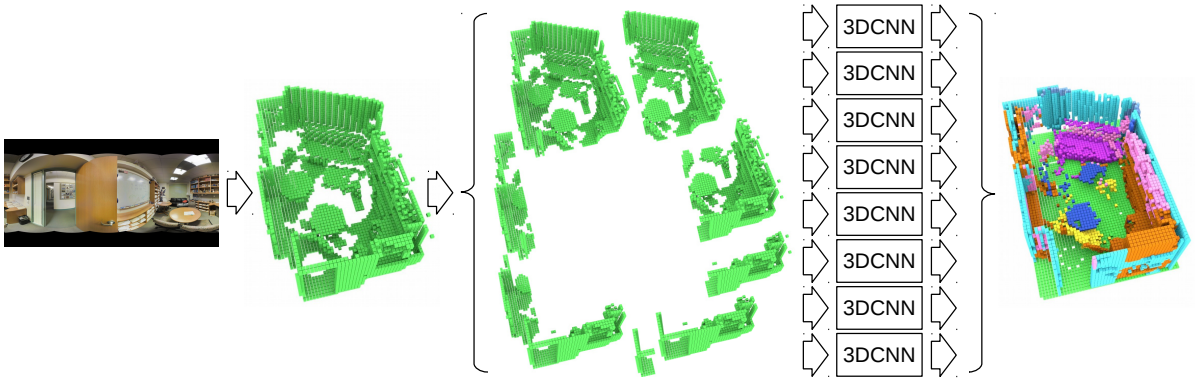
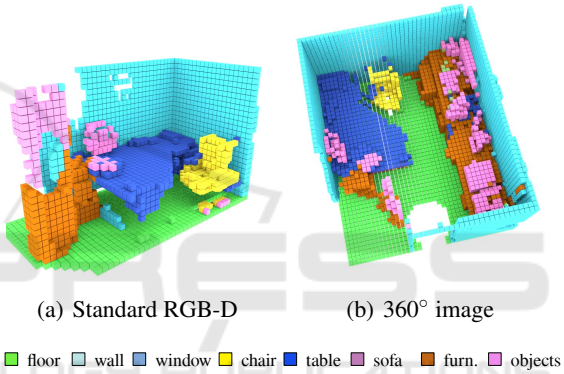


Figure 1: **Overview of our proposed approach**. The incomplete voxel grid generated from input panoramic depth map is automatically partitioned in 8 overlapping views that are individually submitted to our 3D CNN. The resulting prediction is generated from an automatic ensemble of the 8 individual predictions. The result is a complete 3D voxel volume with corresponding semantic labels for occluded surfaces and objects interior.

ping photos taken during the scan[1], but it does not provide full texture maps. The Matterport camera[2] using structured light sensors allows 3D datasets that comprise high-quality panoramic RGB images and its corresponding depth maps of indoor scenes (Armeni et al., 2017; Chang et al., 2017) for a whole room. Figure 2 depicts the difference on SSC results from normal RGB-D and 360° RGB-D image.

Alongside the advanced sensors like Matterport, there are currently many low cost consumer-level spherical cameras available with stereoscopic support, allowing high-resolution 360° RGB image capture, that made widely possible the generation of 360° images and corresponding depth maps through stereo matching. A system created to perform SSC for high-quality 360° images should be also able to work on images generated from low cost cameras, widening the possibilities of applications.

Despite the interesting features of the new large scale 3D datasets, the lack of variety in the type of the scenes is an important drawback. For instance, while NYU v2 regular RGB-D dataset (Silberman et al., 2012) comprises a wide range of commercial and residential buildings in three different cities across 26 scene classes, Stanford 2D-3D-Semantics large-scale dataset (Armeni et al., 2016) only comprises 6 academic buildings and Matterport 3D (Chang et al., 2017) dataset covers only 90 private homes. As most of the SSC solutions are data-driven and CNN-based, a dataset containing a large variety of scene types and object compositions is important to train generalized models. Another limitation of recent scene completion or segmentation methods that use large scans is



(a) Standard RGB-D (b) 360° image

Figure 2: SSC prediction from a regular RGB-D image in (a) covers only a small part of the Scene, while the result from panoramic RGB-D images in (b) covers the whole scene.

that they usually take, as input, a point cloud generated from multiple points of view, implying preprocessing and some level of prior knowledge of the scene.

To overcome these limitations of both previous approaches, we propose a SSC method for a single 360° RGB image and its corresponding depth map image that uses 3D CNN trained on standard synthetic RGB-D data and fine tuned on real RGB-D scenes. The overview of our proposed approach is presented in Figure 1. The proposed method decomposes a single 360° scene into several overlapping partitions so that each one simulates a single view of a regular RGB-D sensor, and submits to our pre-trained network. The final result is obtained aligning and ensembling the partial inferences.

We evaluated our method on the Stanford 2D-3D-Semantics Dataset (2D-3D-S) (Armeni et al., 2017) gathered with the Matterport sensor and stereo 360°

[1]FARO LiDAR, https://www.faro.com/products/construction-bim-cim/faro-focus/

[2]Matterport, https://matterport.com/pro2-3d-camera/

images captured by a pair of low cost 360° cameras. For the experiments with low-cost cameras, we propose a pre-processing method to enhance noisy 360° depth maps before submitting the images to the network for prediction. Our qualitative analysis show that the proposed method achieves reliable results with the low-cost 360° cameras.

Here are our main contributions:

- We are the first to extend the SSC task to complete scene understanding using 360° imaging sensors or stereoscopic spherical cameras;
- We propose a novel approach to perform SSC for 360° images taking advantage of existing standard RGB-D datasets for network training;
- We propose a pre-processing method to enhance depth maps estimated from a stereo pair of low-cost 360° cameras.

2 RELATED WORK

This paper relates to three fields of computer vision, discussed below.

2.1 RGB-D Semantic Scene Completion

3D SSC from standard RGB-D images is a problem that was established quite recently (Song et al., 2017), and consist of, given a single RGB-D image, classifying the semantic labels of all the voxels within the voxel space of the field-of-view, including occluded and non surface regions. The authors used a large synthetic dataset (SUNCG) to generate approximately 140 thousand depth maps that were used to train a typical contracting fully convolutional neural network with 3D dilated convolutions. They showed that jointly training for segmentation and completion leads to better results, as both tasks are inherently interlinked. To deal with data sparsity after projecting depth maps from 2D to 3D, the authors used a variation of Truncated Signed Distance Function (TSDF) that they called Flipped TSDF (F-TSDF). After training on SUNCG, the network was fine-tuned on the NYU depth v2 dataset (Silberman et al., 2012), which was acquired using the first version of MS Kinect.

After this initial work, some authors explored different approaches and architectures achieving good improvements on SSC results, still using the F-TSDF encoded depth map projected to 3D (Guo and Tong, 2018; Zhang et al., 2018a). Another line of work tried to aggregate information from the RGB channels to the SSC network (Guedes et al., 2017), however, better results were observed using a two step training protocol, where a 2D semantic segmentation CNN is first trained and then it is used to generate input to a 3D semantic scene completion CNN (Garbade et al., 2018; Liu et al., 2018).

In order to avoid the need for a two-step training process, we follow a line of work to combine grey-level information with depth maps(Dourado et al., 2019). This is done by using edges detected on the RGB images to highlight flat objects on flat surfaces. Such regions have no depth discontinuity and therefore cannot be captured as boundaries between objects on depth maps. In this approach, the detected edges are projected to 3D and then F-TSDF is applied. The authors observed better results without the need of two training processes.

The main advantage of the regular RGB-D approaches is the abundance and variety of available datasets with densely annotated ground truth which favours the training of deep CNNs. On the other hand, their main drawback is the limited FOV of the sensor, as depicted in Figure 2. Our proposed approach benefits from existing RGB-D datasets for training and presents a way to overcome the limited FOV drawback using 360° images to achieve complete scene coverage.

2.2 Scene Understanding from Large Scale Scans

The Scene Understanding research field observed a boost after the public availability of high quality datasets like Stanford 2D-3D-Semantics Dataset (Armeni et al., 2017) and Matterport3D (Chang et al., 2017), acquired with the Matterport camera, which comprises point cloud ground truth of the whole buildings, 360° RGB panoramas and corresponding depth maps and other features. The scanning process uses a tripod-mounted sensor that comprises three color and three depth cameras pointing slightly up, horizontal, and slightly down. It rotates and acquires RGB photos and depth data. Resulting 360° RGB-D panoramas are software-generated from this data (Chang et al., 2017). These datasets allowed the development of several scene understanding works (Charles et al., 2017; Liu et al., 2017; Qi et al., 2017a). Most of these works focus only on the visible surfaces, rather than on the full understanding of the scene including occluded regions and inner parts of the objects.

In a different line of work, Im2Pano3D (Song et al., 2018) uses data from large scale scans to train a CNN that generates a dense prediction of a full 360° view of an indoor scene from a given partial view of the scene corresponding to a regular RGB-D image.

The work that is most related to our proposal is ScanComplete (Dai et al., 2018). Using data from synthetic or real large scale datasets and a generative 3D CNN, it tries to complete the scene and classify all surface points. However, unlike our proposal, it takes inputs from multiple viewpoints.

Although large-scale scans provide a workaround to surpass the FOV limitations of popular RGB-D sensors, they have the significant drawback that multiple captures of the scene are required to cover a complete scene layout. In addition, each acquisition is a slow scanning process that can only work if the scene remains static for the duration of all captures. Therefore it may be unfeasible to apply them for dynamic scene understanding.

2.3 Scene Understanding using 360° Stereo Images

Spherical imaging provides a solution to overcome the drawbacks inherent to large scale scans. Schoenbein et al. proposed a high-quality omnidirectional 3D reconstruction from catadioptric stereo video cameras (Schoenbein and Geiger, 2014). However, these catadioptric omnidirectional cameras have a large number of systematic parameters including the camera and mirror calibration. In order to get high resolution spherical images with accurate calibration and matching, Spheron developed a line-scan camera, Spheron VR [3], with a fish-eye lens to capture the full environment as an accurate high resolution / high dynamic range image. Li (Li, 2006) has proposed a spherical image acquisition method using two video cameras with fish-eye lenses pointing in opposite directions. Various inexpensive off-the-shelf 360° cameras with two fish-eye lenses have recently become popular[4,5,6]. However, 360° RGB-D cameras which automatically generate depth maps are not yet available. Kim and Hilton proposed depth estimation and scene reconstruction methods using a pair of 360° images from various types of 360° cameras (Kim and Hilton, 2013; Kim et al., 2019). We applied this stereo based method to acquire depth maps for image captured with 360° cameras in the experiments.

[3]Spheron, https://www.spheron.com/products.html

[4]Insta360, https://www.insta360.com

[5]GoPro Fusion, https://shop.gopro.com/EMEA/cameras/fusion/CHDHZ-103-master.html

[6]Ricoh Theta, https://theta360.com/en/

3 DATASETS

We take advantage of existing diverse RGB-D training datasets to train our networks for general semantic scene completion. After training, we evaluate the performance of our model on datasets never seen before by the networks. This section describes the datasets used for training and evaluation.

3.1 Training Datasets

We train our 3D CNN on RGB-D depth maps from the training set of SUNCG (Song et al., 2017) and fine-tuned the networks on train set of NYUv2 dataset(Silberman et al., 2012). SUNCG dataset consists of about 45K synthetic scenes from which were extracted more than 130K 3D snapshots with corresponding depth maps and ground truth divided in train and test datasets. As the provided training data did not include RGB images, we generated images as specified in (Dourado et al., 2019).

NYU v2 dataset includes depth and RGB images captured by the Kinect depth sensor gathered from commercial and residential buildings, comprising 464 different indoor scenes. We generated ground truth by voxelizing the 3D mesh annotations from (Guo et al., 2015) and mapped object categories based on (Handa et al., 2015) to label occupied voxels with semantic classes.

3.2 Evaluation Datasets

Two distinct datasets are used for evaluation: Stanford 2D-3D-Semantics (Armeni et al., 2017) and a dataset created by off-the-shelf 360° cameras.

Stanford 2D-3D-Semantics is large-scale scan dataset gathered with a Materpport camera in academic indoor spaces. The dataset covers over 6,000 m^2 from 7 distinct buildings areas. For each room of the building areas, two or more 360° scans containing several RGB-D images are taken. The images from the scans are aligned, combined, and post-processed to generate one large scale point cloud file for each building area. The point cloud is then annotated with 13 class labels, to be used as ground truth. Each point of the point cloud is also annotated with the room which it belongs to. The dataset also provides a complete RGB 360° panorama, with corresponding depths for each room scan, camera rotation/translation information, and other features useful for 3D understanding tasks. Depth maps are provided as 16 bits png images, with a sensibility of 1/512 m. The value $2^{16} - 1$ is used for pixels without a valid depth measurement.

In order to show general applications of the proposed pipeline, we also used three general 360° image sets captured by various 360° cameras: Meeting Room, Usability Lab and Kitchen. The Meeting Room is similar to a normal living room environment in our daily lives including various objects such as sofas, tables, bookcases, etc. The Usability Lab is similar to the Meeting Room in its size but includes more challenging objects for scene understanding such as large windows and a big mirror on the walls. The Kitchen is a small and narrow room with various kitchen utensils. The scenes are captured as a vertical stereo image pair and dense stereo matching with spherical stereo geometry (Kim and Hilton, 2015) is used to recover depth information.

4 PROPOSED APPROACH

Our proposed approach, illustrated in Figure 1, is described in details in the next subsections. All source code and pretrained models required to reproduce our experiments is publicly available in https://gitlab.com/UnBVision/edgenet360.

4.1 Input Partitioning

From the 360° panoramic depth map, we generate a voxel grid of all the visible surfaces from the camera position, resulting in an incomplete and sparse 3D volume ($480 \times 144 \times 480$ voxels). The preferred voxel size throughout this work is 0.02m which gives an coverage of $9.6 \times 2.8 \times 9.6$ m, but this value can be set to a higher value to reach larger areas, with little impact in prediction accuracy. The resulting volume is then automatically partitioned into 8 views using a 45° step, each of them emulating the field of view of one standard RGB-D sensor. The emulated sensor is positioned 1.7m back from the original position of the 360° sensor, in order to get a better overlapped coverage, especially when the camera is placed near a wall, as is the case of scene from Figure 1 (in that scene, the camera is placed in the bottom left corner of the room). The reason for taking overlapping partitions is to improve the final prediction in the borders of the emulated sensors FOV, by ensembling multiple SSC estimates. Voxels behind the original sensor position are not included in the partition. Each partition size is $240 \times 144 \times 240$ voxels.

4.2 Semantic Scene Completion Network

The resulting partitions are individually submitted to the SSC network for prediction. In our experiments, we used EdgeNet (Dourado et al., 2019), which is a 3D CNN inspired by the U-Net design (Ronneberger et al., 2015) that, uses a surface volume together with a volume generated from the edges present in the RGB image in order to enhance the predictions of objects that are hard to see in depth maps. Its architecture is presented in Figure 3. Both input volumes are encoded using F-TSDF (Song et al., 2018) and the network can be optionally trained to work without the edges volume, as depicted in Figure 1. In our case, the edge volume was generated from the edges present in the RGB panorama projected to 3D using the depth information of corresponding pixels. The partition scheme for the edge volume is the same as that used for the surface volume. For edge detection we used the standard Canny edge detector (Canny, 1986). The final activation function of EdgeNet is a Softmax, each voxel of the output volume contains the predicted probabilities of the 12 classes used for training. The output resolution for each partition is $60 \times 36 \times 60$ voxels.

We trained EdgeNet on standard RGB-D images extracted from the SUNCG training set and fine-tuned on NYU v2 (these datasets are described in Section 3). For the training phase, we used the One Cycle Learning policy (Smith, 2018), which is a combination of Curriculum Learning (Bengio et al., 2009) and Simulated Annealing (Aarts and Korst, 1989), using the same hyperparameters as (Dourado et al., 2019). For fine tuning, we initialized the network with parameters trained on SUNCG and used standard training with SGD with a fixed learning rate of 0.01 and 0.0005 of weight decay. Using the training pipeline described in (Dourado et al., 2019), with offline F-TSDF pre-processing, our training time was 4 days on SUNCG and 6 hours on NYU, using a Nvidia GTX 1080 Ti GPU.

4.3 Prediction Ensemble

Each partition of the input data is processed by our CNN, generating 8 predicted 3D volumes. There are significant overlaps between the FOV of each CNN (some voxels are even captured from 3 different viewpoints), and their predictions need to be combined. We use a simple yet effective strategy of summing the *a posteriori* probability for each class over all classifier outputs, i.e., we apply the "sum rule", demonstrated by Kittler *et al*. (Kittler et al., 1998). Firstly,

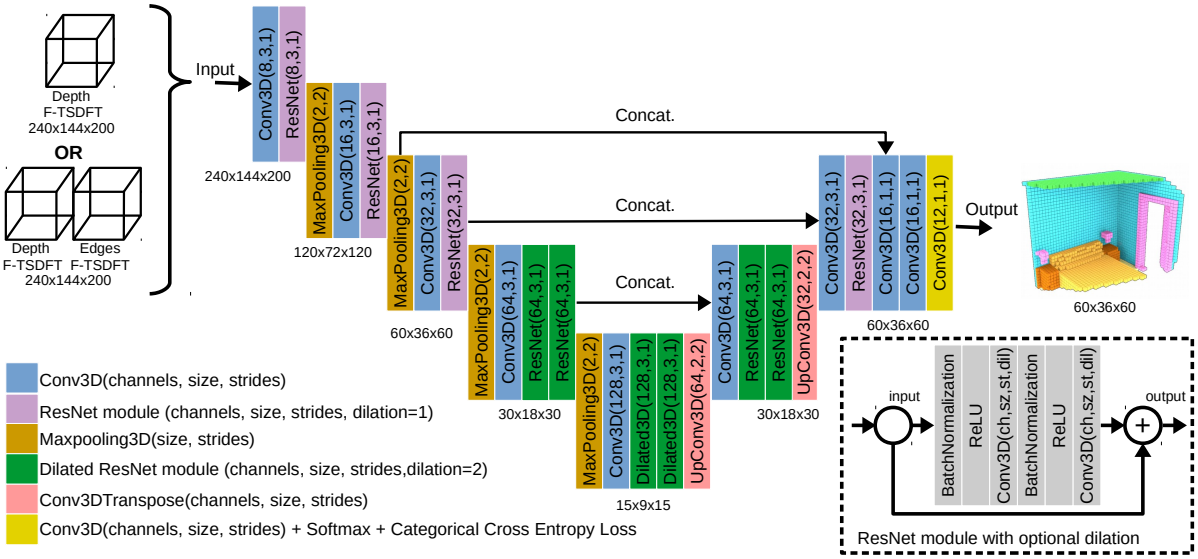


Figure 3: The U-shaped architecture of EdgeNet, with two possible sets of input channels: depth only or depth plus edges (best viewed in colour).

the prediction of each partition is aligned according its position in the final voxel volume. If a given voxel is not covered by a given partition, then the corresponding classifier *a posteriori* probabilities for all classes for that voxel and that partition will be 0, i.e., the softmax result is overruled in voxels outside the field of view of a given partition. Otherwise, the sum of the *a posteriori* probabilities for all classes for that voxel and that classifier will be 1. Given that, for any arbitrary voxel, being n the number of partitions and P_{ij} the *a posteriori* probability of the class i predicted by the classifier j, then, the sum of the probabilities for class i over all classifiers is given by

$$S_i = \sum_{j=1}^{n} P_{ij} \tag{1}$$

and the winning class C for that voxel is

$$C = \arg\max_i (S_i) \ . \tag{2}$$

4.4 Depth Map Enhancement

Stereo capture using commercial 360° cameras is one of realistic approaches as a 360° RGB-D system is not available in the market. However, depth estimation from stereoscopic images is subject to errors due to occlusions between two camera views and correspondence matching errors. These depth errors would lead to noisy and incomplete predictions in SSC. We propose a pre-processing step to enhance this erroneous depth map by taking into account two characteristics of most of the indoor scenes: 1) the majority of indoor scenes can be aligned to the Cartesian axis, following the Manhattan principle (Gupta et al., 2010); 2) the edges present in the RGB images are usually distinguishing features for stereo matching, providing good depth estimates on their neighbourhood.

The Canny Edge detector (Canny, 1986), with low and high thresholds of 30 and 70, is applied to the image and the edges are dilated to 3 pixels width. We observed that those parameters works well for a wide range of RGB images. Using the dilated edges as a mask, we extract the most reliable depth estimations from the original depth map. Vertical edges are removed from the mask as they do not contribute to the stereo matching procedure in the given vertical stereo camera set up. Using the thin edges as a border delimitation, coherent regions with similar colours are searched by a flood fill approach in the RGB image. With this procedure, we expect to get featureless planar surfaces like single colored walls and table tops whose depth surfaces are hard to be estimated by stereo matching. RANSAC (Fischler and Bolles, 1981) is used to fit a plane over those regions eliminating outliers from false stereo matching. If the normal vector of the fitted plane is close to one of the principal axes, we replace the original depth information of the region with the back-projected depths estimated from the plane. Discarding non-orthogonal planes is important to avoid planes estimated from non-planar regions, like wall corners, where the contrast in not enough to produce an edge between two walls. We keep the original depth estimations from the regions where we were not able to fit good planes. We also re-estimate the depths of the south and north poles of the image, as they usually have bad depth es-

timations as proved in (Kim and Hilton, 2013). Good depth estimations from the outer neighborhood of the poles are used as a source for the RANSAC plane fitting.

5 EVALUATION

We quantitatively evaluated our approach on the Stanford 2D-3D-Semantics dataset. We also provide a qualitative evaluation on that dataset and on our stereoscopic images. In this section we describe the experiments and discuss the results.

5.1 Evaluation Metric

As previous works on SSC, we evaluate our proposed approach using Intersection Over Union (IOU) for each class, over visible occupied and occluded voxels inside the room. However, unlike RGB-D works that only evaluate voxels inside the field of view of the sensor, we evaluate over the whole scene. Unfortunately, Stanford 2D-3D does not provide ground truth for the interior of the objects nor for areas that are not visible from at least one of the scanning points, so, we limit our quantitative evaluation to the areas to which ground is provided. We kept the predictions not covered by the ground truth for qualitative evaluation purposes.

5.2 Experiments on Stanford 2D-3D-Semantics Dataset

In order to feed our SSC network with aligned volumes, we rotated the provided 360° RGB panoramas and depth maps using the camera rotation matrix before generating a corresponding input point cloud. Using the room dimensions provided by the dataset, we discarded depth estimations outside room and generated the voxel volume placing camera in the center of the X and Z axis and keeping the capture height so that the floor level is at the voxel plane y=0.

For quantitative evaluation, we extracted only the points belonging to the room from the provided ground-truth (GT) point cloud and translated them to the camera position. In order to align the GT to our input volumes, we voxelized the point cloud using the same voxel size as our input volumes.

Stanford 2D-3D-Semantics dataset classifies each point in 13 classes, while the ground truth extracted from the datasets used to train our network (SUNCG and NYU) classifies the voxels in 12 classes. We mapped the classes *board* and *bookcase* from Stanford 2D-3D-Semantics dataset to classes $objects$ and $furniture$; and both classes *beam* and *door* to $wall$. Predictions of the classes *bed* and tv from SUNCG that have no correspondence in Stanford 2D-3D-Semantics dataset were remapped to *table* and $objects$, respectively. We evaluated all the panoramas from all rooms of types office, conference room, pantry, copy room, and storage. We discarded room types open space, lounge, hallway and WC. We evaluated 669 pairs of 360° RGB images and depth maps from Stanford 2D-3D-Semantics dataset.

Quantitative results for the Stanford 2D-3D-Semantics dataset are provided in table 1. As a baseline, we compare our results to previous works on SSC evaluated on the NYU v2 dataset. It is worth mentioning that, as those results are from different datasets and tasks(our work is the only one that covers the whole scene), they cannot be taken as a direct comparison of models performance.

Our 360° EdgeNet-based ensemble achieved very good overall results and a high level of semantic segmentation accuracy was observed on structural elements floor and wall. Good results were also observed on common scene objects like chairs, sofas, tables and furniture, as well. On the other hand, the same level of performance was not observed on the ceiling, due to domain shift (Csurka, 2017) between training and evaluation datasets. Ceiling in the Stanford dataset is in average higher than that in the NYU dataset where the network was trained. Even so, given that our model had no previous knowledge of the dataset being evaluated, results shows that the proposed model generalizes the scenes well.

Qualitative results presented in Figure 4 depicts the high level of completion achieved by our approach, as seen by comparing the input volume (green) to the prediction. The level of completion is even higher than the the ground truth models, which was manually composed and labelled by the authors using the surface gathered from multiple viewpoints. Note that the missing and occluded regions in the ground truth of scenes were completed in their correspondent predicted volumes. For instance, observe that the floor and wall surrounding the chair in the second scene that are missing in GT was completed by our solution. Semantic labelling results also show high accuracy. In the first scene, the majority of the objects are correctly labelled, even when partially occluded. Hard to detect objects where also correctly labeled. The window behind the sofa, for instance, which is invisible on the depth map, is correctly identified by the proposed approach.

Table 1: **Quantitative results.** We compare our 360° semantic scene completion results on Stanford 2D-3D dataset to partial view state-of-the-art approaches in a normal RGB-D dataset (NYU). Our network had no previous knowledge of the evaluation dataset and predicts result for the whole scene. Previous approaches where fine tuned on the target dataset and only gives a partial prediction. Even so, our proposed solution achieved better overall results.

evaluation dataset	model	scene coverage	semantic scene completion (IoU, in percentages)											
			ceil.	floor	wall	win.	chair	bed	sofa	table	tvs	furn.	objs.	avg.
NYU v2 RGB-D	SSCNet	partial	15.1	94.6	24.7	10.8	17.3	53.2	45.9	15.9	**13.9**	31.1	12.6	30.5
	SGC		17.5	75.4	25.8	**6.7**	15.3	53.8	42.4	11.2	0.0	33.4	11.8	26.7
	EdgeNet		**23.6**	**95.0**	28.6	**12.6**	13.1	**57.7**	**51.1**	16.4	9.6	**37.5**	13.4	32.6
Stanford 2D-3D-S	**Ours**	full (360°)	15.6	92.8	**50.6**	6.6	**26.7**	-	35.4	**33.6**	-	32.2	**15.4**	**34.3**

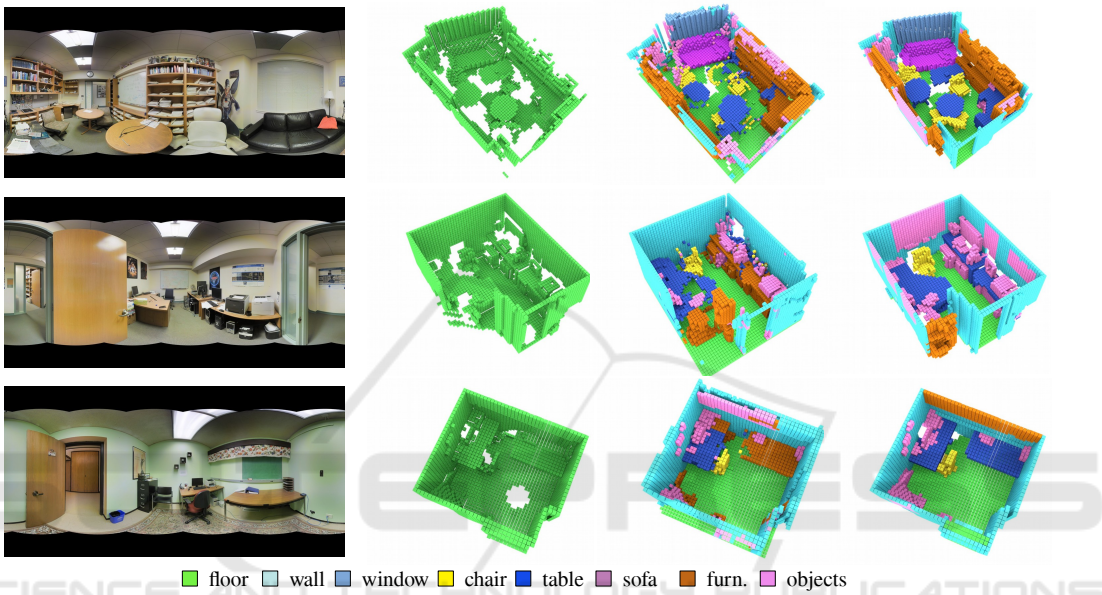


Figure 4: **Stanford 2D-3D-Semantics qualitative results**. From left to right: RGB image; incomplete input volume; semantically completed prediction output; ground truth (best viewed in colour).

5.3 Experiments on Spherical Stereo Images

For spherical stereo images, we first rotated them to align to the Cartesian axis, and applied the enhancement procedure described in Section 4.4. From the resulting images we generated a point cloud and voxelized the surface and edges with a voxel size of 0.02m, before encoding the volumes with F-TSDF and submitting them to the networks. Room dimensions are inferred from the point clouds.

The qualitative results are shown in Figure 5. Most of the stereo matching errors of the estimated depth maps are fixed by our enhancement approach. The cabinet in the extreme left part of the Meeting Room (first scene) originally had several depth estimation errors due to the vertical stripped patterns, but most errors were eliminated by the enhancement step, though some errors still remained in dark regions where borders are not clear. The lower border of the leftmost sofa in the second scene (Usability Lab) was not detected, and some part of its original depth was replaced by the depth of the floor. However, the proposed depth enhancement step improved the erroneous depth maps estimated by stereo matching over the entire regions.

The SSC results with the enhanced depth maps were also satisfactory. As in the large-scale dataset, the levels of scene completion and semantic labelling were high. Although the input images still carries some depth errors, the final predictions were generally clear enough. Comparing the final predictions from the stereo 360° dataset to the ones from Stanford 2D-3D-Semantics dataset, the results of spherical stereo ones are noisier than those of the scanned counter parts, but they are still accurate. Results demonstrate that the use of a pair of 360° images gives an inexpensive alternative to perform 360° SSC for dynamic scenes, where large-scale depth scans are not applicable.

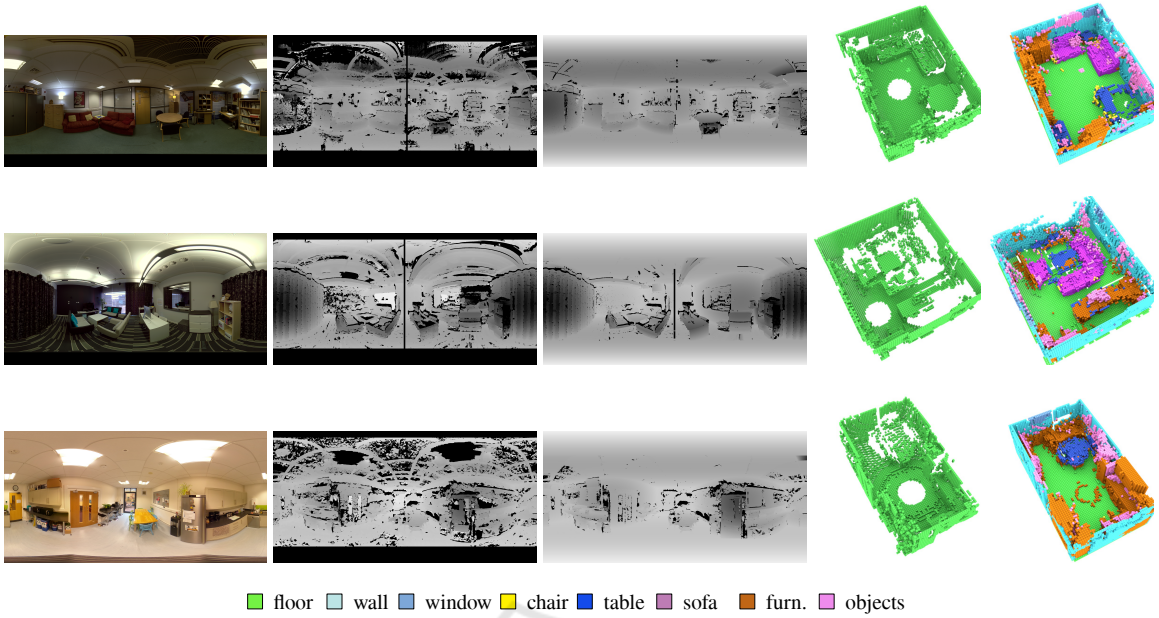


Figure 5: **360°stereoscopic qualitative results**. From left to right: RGB image; estimated depth map; enhanced depth map; incomplete input volume; semantically completed output. From top to bottom: Meeting Room; Usability Lab; Kitchen. Black regions in the estimated disparity maps are unknown regions due to ambiguous matching or stereo occlusion. Most of the failed stereo matching are fixed after enhancement. Predicted volumes present a high level of accuracy (best viewed in colour).

6 CONCLUSION

This paper introduced the task of Semantic Scene Completion from a pair of 360° image and depth map. Our proposed method to predict 3D voxel occupancy and its semantic labels for a whole scene from a single point of view can be applied to various range of images acquired from high-end sensors like Matterport to off-the-shelf 360° cameras. The proposed method is based on a CNN which relies on existing diverse RGB-D datasets for training. For images from spherical cameras, we also presented an effective method to enhance stereoscopic 360° depth maps to be used prior to submit the images to the SSC network.

Our method was evaluated on two distinct datasets: the publicly available Stanford 2D-3D-Semantics high quality large-scale scan dataset and a collection of 360° stereo images gathered with off-the-shelf spherical cameras. Our SSC network requires no previous knowledge of the datasets to perform the evaluation. Even so, when we compare our results to previous approaches using RGB-D images that only give results for part of the scene and were trained on the target datasets, the proposed method achieved better overall results with full coverage. Qualitative analysis shows high levels of completion of occluded regions on both Matteport and spherical images. On the large-scale scan dataset, completion levels achieved from a single point of view were superior to the ones of the ground truth obtained from multiple points of view.

The results show that our approach can be extended to applications that requires a complete understanding of dynamic scenes from images gathered from off-the-shelf stereo cameras.

ACKNOWLEDGEMENTS

The authors would like to thank FAPDF (fap.df.gov.br), CNPq grant PQ 314154/2018-3 (cnpq.br) and EPSRC Audio-Visual Media Platform Grant EP/P022529/1 for the financial support to this work. Mr. Dourado also would like to thank to TCU (tcu.gov.br) for supporting his PhD studies.

REFERENCES

Aarts, E. and Korst, J. (1989). *Simulated Annealing and Boltzmann Machines: A Stochastic Approach to Combinatorial Optimization and Neural Computing*. John Wiley & Sons, Inc., New York, NY, USA.

Armeni, I., Sax, S., Zamir, A. R., and Savarese, S. (2017). Joint2D-3D-semantic data for indoor scene understanding. Technical Report arXiv:1702.01105,

Cornell University Library. http://arxiv.org/abs/1702.01105.

Armeni, I., Sener, O., Zamir, A. R., Jiang, H., Brilakis, I., Fischer, M., and Savarese, S. (2016). 3D semantic parsing of large-scale indoor spaces. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, June 26 - July 1*, pages 1534–1543, Piscataway, NJ. IEEE.

Bengio, Y., Louradour, J., Collobert, R., and Weston, J. (2009). Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML, pages 41–48, New York, NY, USA. ACM.

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

Chang, A. X., Dai, A., Funkhouser, T. A., Halber, M., Nießner, M., Savva, M., Song, S., Zeng, A., and Zhang, Y. (2017). Matterport3D: Learning from RGB-D data in indoor environments. Technical Report arXiv:1709.06158, Cornell University Library. http://arxiv.org/abs/1709.06158.

Charles, R. Q., Su, H., Kaichun, M., and Guibas, L. J. (2017). PointNet: Deep learning on point sets for 3D classification and segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, Hawaii, July 21-26*, pages 77–85, Piscataway, NJ. IEEE.

Csurka, G. (2017). A comprehensive survey on domain adaptation for visual applications. In Csurka, G., editor, *Domain Adaptation in Computer Vision Applications*, pages 1–35. Springer International Publishing, Cham.

Dai, A., Ritchie, D., Bokeloh, M., Reed, S., Sturm, J., and Niessner, M. (2018). ScanComplete: Large-Scale Scene Completion and Semantic Segmentation for 3D Scans. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, June 18-22*, pages 4578–4587, Piscataway, NJ. IEEE.

Dourado, A., de Campos, T. E., Kim, H., and Hilton, A. (2019). EdgeNet: Semantic scene completion from RGB-D images. Technical Report arXiv:1908.02893, Cornell University Library. http://arxiv.org/abs/1908.02893.

Firman, M., Aodha, O. M., Julier, S., and Brostow, G. J. (2016). Structured prediction of unobserved voxels from a single depth image. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, June 26 - July 1*, pages 5431–5440, Piscataway, NJ. IEEE.

Fischler, M. A. and Bolles, R. C. (1981). Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395.

Garbade, M., Sawatzky, J., Richard, A., and Gall, J. (2018). Two stream 3D semantic scene completion. Technical Report arXiv:1804.03550, Cornell University Library. http://arxiv.org/abs/1804.03550.

Guedes, A. B. S., de Campos, T. E., and Hilton, A. (2017). Semantic scene completion combining colour and depth: preliminary experiments. In *ICCV workshop on 3D Reconstruction Meets Semantics (3DRMS)*, Venice, Italy. Event webpage: http://trimbot2020.webhosting.rug.nl/events/events-2017/3drms/. Also published at arXiv:1802.04735.

Guo, R., Zou, C., and Hoiem, D. (2015). Predicting complete 3D models of indoor scenes. Technical Report arXiv:1504.02437, Cornell University Library. http://arxiv.org/abs/1504.02437.

Guo, Y. and Tong, X. (2018). View-Volume Network for Semantic Scene Completion from a Single Depth Image. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 726–732, Stockholm, Sweden. International Joint Conferences on Artificial Intelligence Organization.

Gupta, A., Efros, A. A., and Hebert, M. (2010). Blocks world revisited: Image understanding using qualitative geometry and mechanics. In *Proceedings of 11th European Conference on Computer Vision (ECCV), Crete, Greece, September 5-11*, pages 482–496, Berlin, Heidelberg. Springer Berlin Heidelberg.

Gupta, S., Arbeláez, P., and Malik, J. (2013). Perceptual organization and recognition of indoor scenes from rgb-d images. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, OR, June 23-28*, pages 564–571, Piscataway, NJ. IEEE.

Handa, A., Patraucean, V., Badrinarayanan, V., Stent, S., and Cipolla, R. (2015). SceneNet: Understanding real world indoor scenes with synthetic data. Technical Report arXiv:1511.07041, Cornell University Library. http://arxiv.org/abs/1511.07041.

Kim, H. and Hilton, A. (2013). 3D scene reconstruction from multiple spherical stereo pairs. *Int Journal of Computer Vision (IJCV)*, 104(1):94–116.

Kim, H. and Hilton, A. (2015). Block world reconstruction from spherical stereo image pairs. *Computer Vision and Image Understanding (CVIU)*, 139(C):104–121.

Kim, H., Remaggi, L., Jackson, P. J., and Hilton, A. (2019). Immersive spatial audio reproduction for VR/AR using room acoustic modelling from 360 images. In *Proceedings of 26th IEEE Conference on Virtual Reality and 3D User Interfaces, Osaka Japan*, Piscataway, NJ. IEEE.

Kittler, J., Hatef, M., Duin, R. P. W., and Matas, J. (1998). On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 20(3):226–239.

Li, S. (2006). Real-time spherical stereo. In *Proceedings of International Conference on Pattern Recognition (ICPR)*, pages 1046–1049, Piscataway, NJ. IEEE.

Liu, F., Li, S., Zhang, L., Zhou, C., Ye, R., Wang, Y., and Lu, J. (2017). 3DCNN-DQN-RNN: A deep reinforcement learning framework for semantic parsing of large-scale 3D point clouds. In *Proceedings of 16th International Conference on Computer Vision (ICCV), Venice, Italy*, pages 5679–5688, Piscataway, NJ. IEEE.

Liu, S., HU, Y., Zeng, Y., Tang, Q., Jin, B., Han, Y., and Li, X. (2018). See and think: Disentangling semantic scene completion. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N.,

and Garnett, R., editors, *Procedings of Conference on Neural Information Processing Systems 31 (NIPS)*, pages 263–274, Reed Hook, NY. Curran Associates, Inc.

Nguyen, D. T., Hua, B., Tran, M., Pham, Q., and Yeung, S. (2016). A field model for repairing 3D shapes. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, June 26 - July 1*, pages 5676–5684, Piscataway, NJ. IEEE.

Qi, C. R., Yi, L., Su, H., and Guibas, L. J. (2017a). PointNet++: Deep hierarchical feature learning on point sets in a metric space. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Procedings of Conference on Neural Information Processing Systems 30 (NIPS)*, pages 5099–5108. Curran Associates, Inc., Reed Hook, NY.

Qi, X., Liao, R., Jia, J., Fidler, S., and Urtasun, R. (2017b). 3D graph neural networks for RGBD semantic segmentation. In *Proceedings of 16th International Conference on Computer Vision (ICCV), Venice, Italy*, pages 5209–5218, Piscataway, NJ. IEEE.

Ren, X., Bo, L., and Fox, D. (2012). RGB-(D) scene labeling: Features and algorithms. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, June 16-21*, pages 2759–2766, Piscataway, NJ. IEEE.

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

Schoenbein, M. and Geiger, A. (2014). Omnidirectional 3D reconstruction in augmented manhattan worlds. In *Proceedings of IEEE/RSJ Conference on Intelligent Robots and Systems IROS*, pages 716 – 723, Piscataway, NJ. IEEE.

Shrivastava, A. and Mulam, H. (2013). Building part-based object detectors via 3D geometry. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, OR, June 23-28*, pages 1745–1752, Piscataway, NJ. IEEE.

Silberman, N. and Fergus, R. (2011). Indoor scene segmentation using a structured light sensor. In *IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 601–608, Piscataway, NJ. IEEE.

Silberman, N., Hoiem, D., Kohli, P., and Fergus, R. (2012). Indoor segmentation and support inference from RGBD images. In Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., and Schmid, C., editors, *Proceedings of 12th European Conference on Computer Vision (ECCV), Florence, Italy, October 7-13*, pages 746–760, Berlin, Heidelberg. Springer Berlin Heidelberg.

Smith, L. N. (2018). A disciplined approach to neural network hyper-parameters: Part 1 - learning rate, batch size, momentum, and weight decay. Technical Report arXiv:1803.09820, Cornell University Library. http://arxiv.org/abs/1803.09820.

Song, S., Yu, F., Zeng, A., Chang, A. X., Savva, M., and Funkhouser, T. (2017). Semantic Scene Completion from a Single Depth Image. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, Hawaii, July 21-26*, pages 190–198, Piscataway, NJ. IEEE.

Song, S., Zeng, A., Chang, A. X., Savva, M., Savarese, S., and Funkhouser, T. (2018). Im2Pano3D: Extrapolating 360° structure and semantics beyond the field of view. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, June 18-22*, pages 3847–3856, Piscataway, NJ. IEEE.

Zhang, J., Zhao, H., Yao, A., Chen, Y., Zhang, L., and Liao, H. (2018a). Efficient semantic scene completion network with spatial group convolution. In *Proceedings of 15th European Conference on Computer Vision (ECCV), Munich, Germany, September 8-14*, pages 749–765, Cham. Springer International Publishing.

Zhang, L., Wang, L., Zhang, X., Shen, P., Bennamoun, M., Zhu, G., Shah, S. A. A., and Song, J. (2018b). Semantic scene completion with dense CRF from a single depth image. *Neurocomputing*, 318:182–195.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS