SEMANTIC LABELING OF STRUCTURAL ELEMENTS IN BUILDINGS BY FUSING RGB AND DEPTH IMAGES IN AN ENCODER-DECODER CNN FRAMEWORK

D. Iwaszczuk[1,2]*, Z. Koppanyi[1], N. A. Gard[1], B. Zha[1], C. Toth[1], A. Yilmaz[1]

[1] Dept. of Civil, Environmental and Geodetic Engineering, The Ohio State University, Columbus, OH, 43212, USA
- (koppanyi.1, ajamgard.1, zha.44, toth.2, yilmaz.15)@osu.edu
[2] Dept. of Civil, Geo and Environmental Engineering, Technical University of Munich, 80333 Munich, Germany
- dorota.iwaszczuk@tum.de

Commission I, WG I/6

KEY WORDS: CNN, Sensor, Fusion, Semantic, Labeling

ABSTRACT:

In the last decade, we have observed an increasing demand for indoor scene modeling in various applications, such as mobility inside buildings, emergency and rescue operations, and maintenance. Automatically distinguishing between structural elements of buildings, such as walls, ceilings, floors, windows, doors etc., and typical objects in buildings, such as chairs, tables and shelves, is particularly important for many reasons, such as 3D building modeling or navigation. This information can be generally retrieved through semantic labeling. In the past few years, convolutional neural networks (CNN) have become the preferred method for semantic labeling. Furthermore, there is ongoing research on fusing RGB and depth images in CNN frameworks. For pixel-level labeling, encoder-decoder CNN frameworks have been shown to be the most effective. In this study, we adopt an encoder-decoder CNN architecture to label structural elements in buildings and investigate the influence of using depth information on the detection of typical objects in buildings. For this purpose, we have introduced an approach to combine depth map with RGB images by changing the color space of the original image to HSV and then substitute the V channel with the depth information (D) and use it utilize it in the CNN architecture. As further variation of this approach, we also transform back the HSD images to RGB color space and use them within the CNN. This approach allows for using a CNN, designed for three-channel image input, and directly comparing our results with RGB-based labeling within the same network. We perform our tests using the Stanford 2D-3D-Semantics Dataset (2D-3D-S), a widely used indoor dataset. Furthermore, we compare our approach with results when using four-channel input created by stacking RGB and depth (RGBD). Our investigation shows that fusing RGB and depth improves results on semantic labeling; particularly, on structural elements of buildings. On the 2D-3D-S dataset, we achieve up to 92.1% global accuracy, compared to 90.9% using RGB only and 93.6% using RGBD. Moreover, the scores of Intersection over Union metric have improved using depth, which shows that it gives better labeling results at the boundaries.

1. INTRODUCTION

Urbanization and population growth in the cities has increased interest in generating detailed hierarchical models of urban areas, including objects, such as buildings, roads, ponds, etc., described by their geometry and semantics. Among these objects, buildings have an important role as people perform most of their activities indoors. This requires systems that provide improved support for applications including but not limited to mobility inside buildings, emergency and rescue operations and maintenance tasks. The realization of such systems depends on the structural elements defined within the level of detail (LoD) in building models, e.g. doors, windows, walls, etc., provided in the Geographical Information Systems (GIS) and Building Information Model (BIM).

In literature, there are many ongoing research efforts on modeling building exterior (Becker, 2009, Tuttas and Stilla, 2011, Ripperda, 2008, Förstner and Korč, 2009), and interior (Del Pero et al., 2012, Becker et al., 2015, Chen et al., 2015, Liu et al., 2015, Armeni et al., 2016) using a number of remote sensing sensors, such as cameras, LiDAR and RADAR. These sensors provide high volume data which requires automation of interpretation. Particularly, automatically distinguishing between structural elements of buildings, such as walls, ceilings, floors, windows, doors etc., and movable objects, such as chairs, tables and shelves, is crucial for creating building models. This information can be retrieved through a semantic labeling process. In the past few years, convolutional neural networks (CNN) have become the state-of-the-art approach for semantic labeling of image data due their ability to learn and extract features at high performance. For pixel-wise labeling, which is of particular interest for 3D reconstruction purposes, the encoder-decoder architecture (Badrinarayanan et al., 2017) has been shown to be very effective. This CNN architecture has been widely investigated for RGB data. In order to incorporate depth in such networks, typically depth channel (D) is stacked with RGB generating a four-channel input, which leads to high number parameters to be estimated in the training. In this paper, we investigate an alternative approach to fuse depth with the RGB image through color space transformations and use this modified three-channel input for labeling.

The paper is organized as follows. Section 1.1 briefly presents the concept of convolutional neural networks for image segmentation and labeling as well as the most relevant studies on this topics. Then, in Section 2, we introduce our approach for fusion of imaging and depth sensors as a three-channel data and use it for labeling within an encoder-decoder CNN architecture. In Section 3, we describe the Stanford 2D-3D-Semantics Dataset (2D-3D-S) (Armeni et al., 2017) which we use in our experiments. Following this, we present our experimental setup and the results

*Corresponding author

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLII-1-225-2018

in Section 4. Finally, we discuss our results in Section 5 and conclude out work in Section 6.

1.1 Related Work

Image segmentation, interpretation and semantic understanding of images are typical problems in many computer vision and mapping applications. Strictly speaking, these problems typically have no exact solution, while the human solves them relatively efficiently and without much effort. Neural networks mimic this behavior by defining a layered network topology consisting of nodes, also called neurons, and activation functions attached to these neurons. In the last decade, deep neural networks, in particular, the convolutional network structures, have been successfully applied to solve various labeling problems on aerial, indoor and outdoor images. Generally, these problems can be categorized into three classes. The first attempts by the computer vision community were to recognize and label an entire image, claiming that, convolutional neural networks compete with humans in image recognition problems, see ILSVCR (Russakovsky et al., 2015, Karpathy, 2014). Algorithms in the second group extract regions from images representing various objects, and label these image regions; see, for instance, R-CNN (Girshick et al., 2013, Toth and Koppanyi, 2017). Finally, algorithms in the third class, which is the focus of this study, label all pixels on the images, also referred as pixel-wise or pixel-level segmentation.

Prior to deep networks, predicting the central pixel label in a patch of pixels (Shotton et al., 2008), and, later on, all the pixels in a patch using Random Forest (Kontschieder et al., 2011) were among the most successful methods for pixel-wise segmentation. In 2015, (Long et al., 2015) presented an end-to-end, pixel-to-pixel fully connected convolutional neural network (CNN) structure for semantic segmentation for RGB images. Later, (Badrinarayanan et al., 2017) introduced an encoder-decoder-based network structure called SegNet. The encoder-decoder structure utilizes a set of pooling layers that downsamples the original image to a smaller feature raster, and then it upsamples back to the original size, where each pixel represents a label category. Segmentation results were later improved by transferring max-pooling indices to the decoder. (Yu and Koltun, 2015) design a context module to enhance the performance of dense prediction architectures by means of dilated convolutions. Similar to dilated convolutions, Deeplab (Chen et al., 2018) uses upsampled filters that are referred to as atrous convolution, to control the field of view. They also exploit atrous spatial pyramid pooling (ASPP) to segment objects at multiple scales and then a conditional random fields (CRF) method is applied to localize the object boundaries. (Zhao et al., 2017) use the global context information as clues for scene parsing via a pyramid scene parsing network.

The use of depth information along with RGB data is not straightforward as it needs a special care for fusion of the depth to RGB images. (Gupta et al., 2014) propose to use Horizontal Height Angle (HHA) which encodes depth into three channels: horizontal disparity, height above ground, and the angle between local surface normal and the inferred gravity direction. (Zeng et al., 2017) state that HHA does not provide any notable performance improvement compared to that of VGG when trained on only RGB images. An alternative approach to use second branch for depth shows promising results (Hazirbas et al., 2016, Sherrah, 2016, Schneider et al., 2017), it generates, however, an architecture which is more difficult to train.

In this paper, we present an approach that uses RGB and depth data in a convolutional neural network (CNN) for labeling structural elements of buildings. In particular, we adopt the encoder-decoder network architecture introduced in (Badrinarayanan et al., 2017) and hypothesize that the depth information encoded in the color channels improve the semantic labeling result. In order to realize this, our algorithm converts the RGB color space to hue, saturation, value (HSV) color space, and replaces the intensity encoded in the V component with depth information. We show that two variations of the proposed approach outperform the RGB based encoder-decoder framework on the Stanford 2D-3D-Semantics Dataset (2D-3D-S) indoor dataset (Armeni et al., 2017). We compare this approach also with the a CNN implementation using a four-channel RGBD input.

2. SENSOR FUSION WITH CNN

Fusing RGB and depth sensor data within a CNN architecture can be performed in several ways. Typically, the three-channel RGB and one-channel depth inputs are stacked generating a four-channel RGBD input. This approach results in higher number of parameters to be estimated in the training and increases the time for training. At the same time, the RGB can be transformed to other color spaces, such as hue, saturation and value (HSV), where V, representing the brightness, is strongly illumination and view dependent. Therefore, we hypothesize that this channel of the HSV color space does not contribute to the image classification, and can be considered redundant and replaced by depth. This approach enables to keep the number of estimated parameters equal to using RGB only and to use a SegNet based encoder-decoder network.

2.1 Network Architecture

We build our method by adopting the SegNet (Badrinarayanan et al., 2017) architecture, which is an encoder-decoder type network design. The first 13 layers in the VGG16 network (Simonyan and Zisserman, 2014) comprise the encoder network in SegNet. Each layer is 3×3 convolution, which are stacked on each other. The encoder receives three channel image input to generate a low dimensional representation which is passed onto the decoder that classifies pixels in the image.

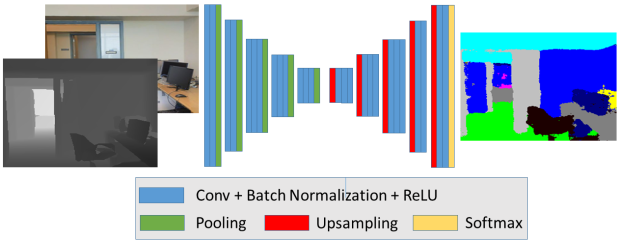

Figure 1. SegNet based encoder-decoder architecture for semantic labeling using RGB and depth images.

The decoder is the mirror of the encoder, such that for each layer in the decoder network there is a corresponding layer in the encoder network. Class labels are generated in the last layer of the decoder using a softmax classifier, which predicts pixel labels.

The breakdown of the *encoder* is as follows:

1. Generate feature maps from various filters
2. Batch normalize the generated feature maps
3. Apply rectified linear unit (ReLU)
4. Perform max-pooling with 2×2 window $+$ stride of size 2

5. Downsize the output by a factor of 2

The breakdown of the *decoder* is as follows:

1. Generate sparse feature maps using max-pooling indices from the corresponding encoder layer
2. Generate dense feature maps by convolving sparse feature maps with trainable filters
3. Batch normalize generated feature maps
4. Predicts pixel labels through a multi-class softmax function

It is noteworthy that, although excessive use of max-pooling may result in translation invariance and therefore better labeling performance, the network loses accuracy on pixels residing on object boundaries. Since, our goal is to label pixels within an image, it is crucial to have accurate boundary information after labeling.

2.2 Fusing Depth with RGB

It is a reasonable expectation that extending the color information with depth data of the scene increases the segmentation accuracy. RGB is a widely used color space for images, which contains red, blue, and green channels. As opposed to RGB, HSV is another representation of the color space that uses hue, saturation, and value. In the human visual system, colors are perceived when cone cells are excited, while luminance is perceived when rod cells are excited. Similar to how the rod and cone cells work for humans to perceive colors, separation of the luminance component of a pixel color from its chrominance components is adopted by the HSV model. Colors are represented by hue, and saturation gives a measure of the amount of gray in the color (Vadivel et al., 2005). Note that RGB representation allows for retrieving shapes, edges of object boundaries from one color channel. This indicates that RGB channels carry redundant information.

We fuse the RGB and depth information by combining the depth with the reduced color space. We perform this fusion in two different ways: first by transforming RGB image to HSV color space and replacing the value component with depth (fusion F1); and second, by transforming this HSD image back to RGB color space (fusion F2).

Fusion F1: Let r, g and b be the values of the RGB images normalized to [0,1], $c_{\max} = \max(r, g, b)$ the maximal value and $c_{\min} = \min(r, g, b)$ the minimum value of those three components. We generate images consisting of three channels HSD, where their two first components are calculated as

$$H = \begin{cases} 0, \text{ for } c_{\max} = 0 \\ 60^\circ \frac{g-b}{c_{\max}-c_{\min}} \text{mod} 6, \text{ for } c_{\max} = r \\ 60^\circ \frac{b-r}{c_{\max}-c_{\min}} + 2, \text{ for } c_{\max} = g \\ 60^\circ \frac{r-g}{c_{\max}-c_{\min}} + 4, \text{ for } c_{\max} = b, \end{cases} \tag{1}$$

$$S = \begin{cases} 0, \text{ for } c_{\max} = 0 \\ \frac{c_{\max}-c_{\min}}{c_{\max}}, \text{ otherwise}, \end{cases} \tag{2}$$

and the third component D is generated from depth values normalized to [0,1]. Practically, H is also normalized to values [0,1] to be consistent with S and D.

Fusion F2: In the second variation of the depth fusion, we use the representation from F1 and transform it back to RGB color space. Let c_1 be primary color defined as integer component of $H/60$. We perform colors space back transformation as follows

$$(R_d, G_d, B_d) = \begin{cases} (D, c, a), \text{ for } c_1 = 0 \\ (b, D, a), \text{ for } c_1 = 1 \\ (a, D, c), \text{ for } c_1 = 2 \\ (a, b, D), \text{ for } c_1 = 3 \\ (c, a, D), \text{ for } c_1 = 4 \\ (D, a, b), \text{ for } c_1 = 5 \end{cases} \tag{3}$$

where

$$a = \frac{D(c_{\max} - c_{\min})}{1 - c_{\max}}, \tag{4}$$

$$b = \frac{D(c_{\max} - c_{\min})(H/60 - c_1)}{1 - c_{\max}}, \tag{5}$$

$$c = \frac{D(c_{\max} - c_{\min})(H/60 - c_1)}{c_{\max} + 1}. \tag{6}$$

3. DATASET AND PREPROCESSING

To test the proposed approach, we used the Stanford 2D-3D-Semantics Dataset (2D-3D-S) (Armeni et al., 2017), which contains RGB images as well as the corresponding depth images for 11 types of indoor scenes. Most of the dataset are office rooms and hallways, but there is also a small part of other room types, such as lobby and auditorium. The data is collected using the Matterport Camera, which combines 3 structured-light sensors to capture RGB and 360-degree depth images. Each 360 degree sweep is performed in increments of 60 degrees. The sensor provides a reconstructed 3D textured mesh and the raw RGB-D images. The dataset consist of 6 indoor areas. It also contains semantic annotations. The annotations are pixel-wise, and correspond to 13 object classes, including ceiling, floor, wall, column, beam, window, door, table chair, bookcase, sofa, board, and clutter.

In order to use this data in our experiments as input for SegNet based network, preprocessing steps are required. These steps includes resizing, depth filtering, depth thresholding and normalization.

Resizing: Original image size in the dataset is 1080 × 1080. Large images require larger neural network architecture, and thus, the training is time consuming. In addition, larger image size does not necessarily mean more accurate results. Therefore, many authors (Hazirbas et al., 2016, Badrinarayanan et al., 2017) suggest resizing the images to 224 × 224 for indoor scales. Thus, the images are resized to this dimensions for network training and testing.

Depth filtering: Depth images in the dataset have missing pixels. These missing pixels can significantly influence the training and validation of the semantic labeling, therefore depth interpolation is performed using the *Inpainting* method (Garcia, 2010, Wang et al., 2012). In this approach, first rough initial guesses of the missing depth value are calculated by local statistical analysis. Then an iterative scheme continuously estimates missing depth values using discrete cosine transform (DCT) until convergence.

Depth thresholding and normalization: RGB values are normalized to 0 to 255 range, while HSV values range from 0 to 1. Depth depends on the scene and the range of the sensor. For indoor scene we can accept values up to few meters. Also, typical depth cameras operate in this range. Therefore, we perform a normalization of depth values based on these knowledge by setting 0 [m] to minimum and 10 [m] as maximum possible values. All values, which are larger than 10 [m] are set to 10 [m]. Then, the depth values are rescaled to 0 to 1 and fused with the HSV images.

Class weighting: In the preprocessing step, we also look at the dataset and the distribution of the classes. In order to measure this distribution, we calculate the class frequency, which is defined as

$$f = \frac{n_c}{N}, \tag{7}$$

where n_c = number of pixels of the class,
N = number of all pixels in the dataset.

This class frequency f is depicted in Fig. 2

As shown in Fig. 2, the class frequency in 2D-3D-Semantics Dataset is unbalanced. In ideal case, the number of samples used for training should be the same for each class. For the pixelwise segmentation tasks, however, this is usually not possible, as some objects always occupy a bigger region in an image. In order to overcome this problem, we use class weighting of the training data defined as:

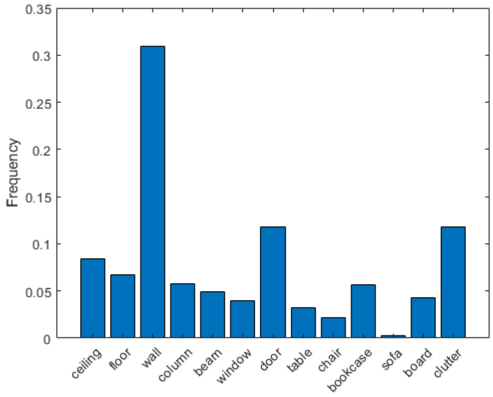

Figure 2. Class frequency in the training data.

$$w = \frac{\text{median}(\frac{n_c}{n})}{\frac{n_c}{n}}, \tag{8}$$

where n_c = number of pixels of the class,
n = total number of pixels in images,
that had an instance of the class.

This approach helps reduce the problem of unbalanced pixel number, but it does not handle extreme cases. In 2D-3D-S dataset sofa is defined as a separate class. In Area 1, however, the amount of pixels labeled as sofa is negligible compared to other classes. Therefore, we discard this class for our tests and treat it as clutter.

4. EXPERIMENTS

We test semantic labeling with the Stanford 2D-3D-Semantics Dataset described in Section 3. For our experiments, we select Area 1 that covers over 965 square meters and 2,850 cubic meters. This area contains 10,327 images and all the 13 object classes so that it is sufficient to represent a typical indoor scene.

After the data preprocessing, we fuse RGB and depth to create two new datasets: HSD images, and HSD images transformed back to R$_d$G$_d$B$_d$. Example for those images can be seen in Fig. 3.

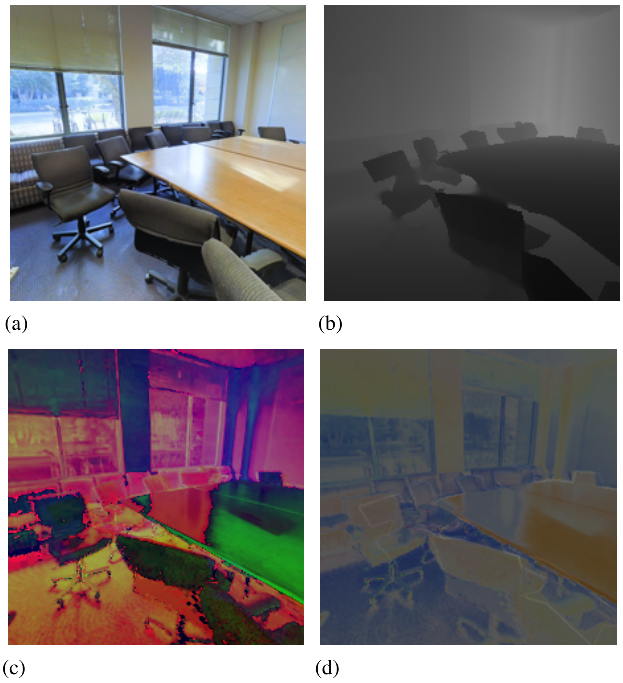

Figure 3. An exemplary image from the dataset: a) RGB; b) depth; c) HSD; d) R$_d$G$_d$B$_d$.

4.1 Experiment Setup

We conduct two experiments using the test dataset.

Test T1: First test shows the general performance of our approach. In this experiment we take 50% of the images for training (5164 images) and the other 50% for validation (5163 images).

Test T2: Then, we test the ability of our approach to label a scene based on the training in another scene. For this purpose, we select six rooms: three offices, two hallways and one conference room, which is about 10% of the data (1047 images). While selecting the training dataset, we payed attention that all 13 classes are included. In this test, the remaining 90% of the data is used a the test set.

Hyper-parameters and Training: The hyper-parameters of the CNN were the same for all tests. We used stochastic gradient decent solver with momentum, where the momentum was 0.9 and the weight decay was 0.0005. We used batch size of 10 and the input data was randomly shuffled after every epoch. The learning rate of 0.001 was set constant over the entire training, which ended after 100 epochs. Weights for our network are initialized based on VGG16 weights.

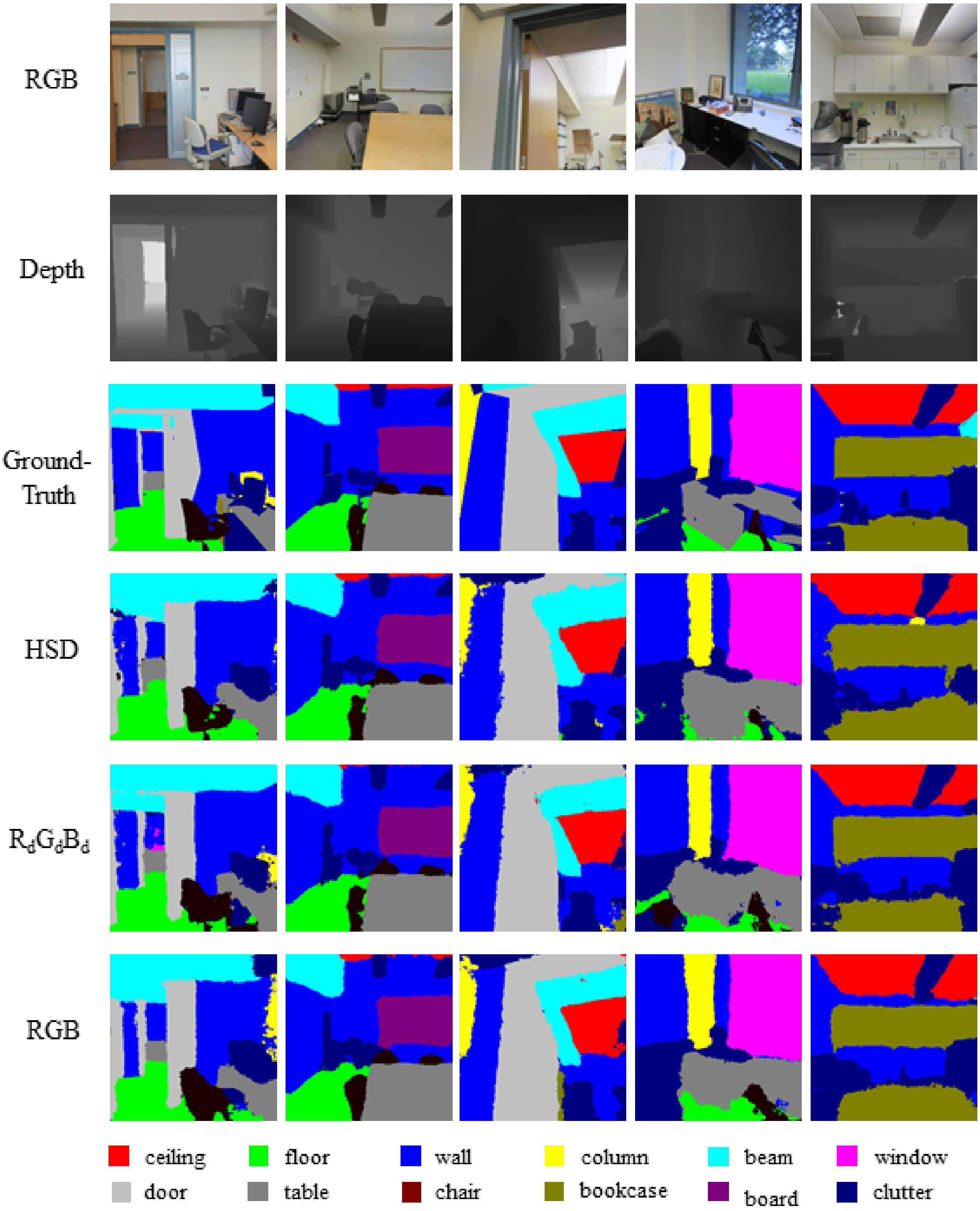

Figure 4. Results of the label prediction.

4.2 Evaluation

Results of the labeling on typical images are presented in Fig. 4. We evaluate our result of label prediction based on the confusion matrix elements, where TP denotes true positive, FP denotes false positive, FN denotes false negative samples and N denotes the total number of annotated pixels. For this evaluation, we use three metrics. First metrics, called global accuracy, represents the percentage of the correctly classified pixels and is defined as

$$\mathrm{GlobAcc} = \frac{1}{N}\sum \mathrm{TP}_c. \tag{9}$$

Second metrics is called mean accuracy and denotes the average accuracy over all classes. It is defined as

$$\mathrm{MeanAcc} = \frac{1}{K}\sum \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FP}_c}, \tag{10}$$

where $c \in \{1, 2, ..., K\}$ denotes the class index and K denotes the number of classes. Finally, we use for evaluation also Intersection over Union (IoU), which is the average value of the intersection of the prediction and ground truth over the union of them. It is defined as

$$\mathrm{IoU} = \frac{1}{K}\sum \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FP}_c + \mathrm{FN}_c}. \tag{11}$$

First we evaluate the results of test T1[1]. The evaluation metrics for the experiments are presented in Tab. 1. Here, it can be shown that our fusion F2 ($\mathrm{R}_d\mathrm{G}_d\mathrm{B}_d$) delivers the best results in terms of all three metrics: global accuracy, mean accuracy and IoU among three-channel inputs. In term of global accuracy and IoU, the four-channel RGBD input delivers better results.

Channels	GlobalAcc	MeanAcc	Mean IoU
RGB	90.9%	92.5%	81.2%
HSD	91.4%	92.5%	82.0%
$\mathrm{R}_d\mathrm{G}_d\mathrm{B}_d$	92.1%	**93.5%**	83.2%
RGBD	**93.6%**	92.8%	**86.3%**

Table 1. Results on semantic labeling in test T1 (using 50% of the dataset for training and 50% for testing)

We also investigate the metrics per class, which is presented in Tab. 2 and Tab. 3. Particularly, we investigate the accuracy of labeling structural elements of buildings: ceiling, floor, wall, column and beam.

We also calculate the dataset metrics for structural elements of buildings, more specifically of ceiling, floor, wall, column and beam (Tab. 4). Here, we calculate the MeanAcc and Mean IOU as average value for those classes (contained individually Tab. 2 and Tab. 3) and the GlobalAcc by treating other classes all as clutter.

Next, we evaluate the results of test T2, where selected rooms where used for training and other rooms for testing the approach. The metrics for the dataset are presented in Tab. 5

For test T2, we also present the metrics per class in Tab. 6 and Tab. 7 with special attention paid to structural elements. Also for test T2, we present the results on labeling of structural building elements which is presented in Tab. 8.

Class	RGB	HSD	$\mathrm{R}_d\mathrm{G}_d\mathrm{B}_d$	RGBD
ceiling	95.8%	96.8%	**97.2%**	97.0%
floor	92.1%	96.0%	**96.3%**	95.7%
wall	87.7%	89.0%	89.4%	**94.9%**
column	92.3%	89.4%	**93.4%**	83.9%
beam	94.9%	95.6%	**96.0%**	94.5%
window	97.1%	**97.6%**	**97.6%**	96.6%
door	92.9%	93.5%	93.1%	**95.5%**
table	**93.5%**	91.3%	92.7%	92.7%
chair	90.0%	89.7%	**91.6%**	84.9%
bookcase	**93.4%**	92.8%	92.7%	91.9%
board	**95.6%**	94.0%	95.5%	94.6%
clutter	85.1%	84.7%	86.2%	**87.8%**

Table 2. Accuracy per class in test T1 (50% of the dataset for training and 50% for testing).

Class	RGB	HSD	$\mathrm{R}_d\mathrm{G}_d\mathrm{B}_d$	RGBD
ceiling	90.2%	91.0%	92.5%	**93.6%**
floor	84.2%	87.2%	88.8%	**90.4%**
wall	84.5%	85.1%	86.1%	**88.9%**
column	76.0%	79.0%	**80.1%**	79.3%
beam	85.0%	85.6%	86.5%	**90.7%**
window	90.2%	90.1%	90.0%	**92.3%**
door	85.2%	86.6%	87.2%	**90.5%**
table	69.7%	70.4%	72.0%	**76.1%**
chair	65.0%	68.6%	71.0%	**75.8%**
bookcase	81.9%	81.2%	82.4%	**85.7%**
board	**90.0%**	87.9%	87.4%	**90.0%**
clutter	72.3%	71.8%	74.0%	**79.0%**

Table 3. IoU per class in test T1 (50% of the dataset for training and 50% for testing).

5. DISCUSSION AND OUTLOOK

Our investigations showed that incorporating depth improves slightly the labeling results in an indoor scene. For structural elements of buildings, this improvement is even more significant, which was shown in both tests T1 and T2. Here it is noticeable that $\mathrm{R}_d\mathrm{G}_d\mathrm{B}_d$ representation delivers better results than HSD representation. This can be related to the periodic nature of the definition of hue in the HSV representation. Another explanation can be the fact, that the utilized network was designed for RGB input and VGG16 weight come also from training on RGB images, which is closer to $\mathrm{R}_d\mathrm{G}_d\mathrm{B}_d$ rather than to HSD representation. An alternative would have been generation of a more complicated dual network architecture (Siamese networks) for the treatment of depth and color information on parallel networks that can be merged using a final fully connected layer to generate the labels through a decoder network. An added benefit is a simpler network architecture, resulting in faster learning and fewer hyperparameters used to setup the network.

Comparing the results across the RGB-based method and two variations of RGB and depth fusion, it can be observed that the boundaries of structural elements are better delineated in IoU measurements as tabulated in Tab. 7. This improvement in boundary detection is exemplified in Fig. 5. We have also observed reduction in performance when the training set is reduced to 10% of the data. We should also note that data used for our experiments contains biased training samples across classes. The class frequency shown in Fig. 2 suggests that a random sampling of {image-label} pair will create biased training set where *wall, ceiling and door* will have better representation than other labels,

[1] Tests with RGB, HSD and $\mathrm{R}_d\mathrm{G}_d\mathrm{B}_d$ were conducted using a SegNet-based MATLAB implementation, while RGBD test were performed using a Pytorch implementation using the same hyper parameters

Channels	GlobalAcc	MeanAcc	Mean IoU
RGB	92.2%	92.6%	84.0%
HSD	92.8%	93.4%	85.6%
$R_dG_dB_d$	93.4%	**94.5%**	86.8%
RGBD	**94.7%**	94.2%	**89.4%**

Table 4. Results on semantic labeling of structural elements of buildings in test T1 (50% of the dataset for training and 50% for testing).

Channels	GlobalAcc	MeanAcc	Mean IoU
RGB	65.0%	61.8%	45.2%
HSD	61.0%	55.7%	40.0%
$R_dG_dB_d$	65.7%	60.1%	45.4%
RGBD	**69.4%**	**64.0%**	**49.2%**

Table 5. Results on semantic labeling in test T2 (10% of the dataset for training and 90% for testing).

which can be considered a typical class distribution for most indoor scenes.

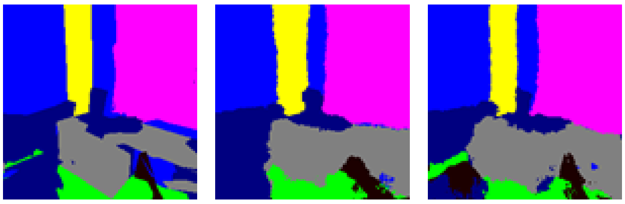

Figure 5. Improvement of the labeling at the boundaries using depth on example of class column (yellow). The images represent: ground truth (left), RGB based label prediction (middle), $R_dG_dB_d$ based label prediction

In our experiments, we compared our approach with utilization of stacked RGB and depth input (RGBD). It was observed, that using RGBD input up to 2% higher accuracy can be achieved. Analyzing the accuracy per class for structural elements, we can observe that significantly better accuracy was achieved only for wall, windows and door classes. In our dataset, wall was the most dominant class, therefore high accuracy for this class results in high global accuracy. Remarkable is also that using stacked RGBD input improves IoU for almost all classes compared to RGB and $R_dG_dB_d$ input.

6. CONCLUSION

In this study, we investigated the influence of using fused RGB with depth information to label indoor scenes within an encoder-decoder CNN framework and compared it with the performance of an RGB-based labeling. We use the same, three-channel-based, CNN architecture for RGB and fused RGB and depth. In order to do so, we proposed a fusion of RGB and depth information in two variations by taking advantage of the redundancy in information of the three color channels, used in the RGB representation. Our results showed that using depth shows better overall accuracy of the labeling. Also Intersection over Union metric can be improved by the RGB-depth fusion, which proves that this approach gives better labeling results at the boundaries.

In summary, we showed that depth particularly improves labeling on structural elements of buildings, such as ceilings, walls, floors, which was the focus of our study. These elements are the main components of Building Information Models (BIM), and thus identifying them in images is important for tasks such as reconstruction of building models and and coregistration of images with these models.

Class	RGB	HSD	$R_dG_dB_d$	RGBD
ceiling	86.3%	85.6%	**89.7%**	89.5%
floor	85.5%	88.4%	**92.1%**	90.2%
wall	70.7%	73.0%	76.7%	**86.2%**
column	12.8%	11.3%	**14.4%**	9.9%
beam	59.2%	38.8%	**66.1%**	54.4%
window	87.4%	74.7%	78.6%	**89.8%**
door	**80.9%**	70.2%	77.8%	74.6%
table	74.7%	60.0%	67.1%	**76.5%**
chair	52.1%	55.0%	**59.1%**	57.1%
bookcase	29.3%	28.0%	28.8%	**30.1%**
board	49.8%	39.1%	20.6%	**61.4%**
clutter	**52.7%**	44.5%	50.7%	48.0%

Table 6. Accuracy per class in test T2 (10% of the dataset for training and 90% for testing).

Class	RGB	HSD	$R_dG_dB_d$	RGBD
ceiling	64.5%	62.5%	**79.1%**	75.4%
floor	71.4%	72.3%	**80.5%**	78.5%
wall	55.6%	53.3%	56.4%	**60.4%**
column	10.0%	8.8%	**10.9%**	8.6%
beam	42.1%	31.4%	**48.0%**	41.1%
window	64.1%	63.8%	69.1%	**73.0%**
door	**61.4%**	56.5%	56.8%	61.3%
table	42.2%	35.3%	45.7%	**46.0%**
chair	32.9%	20.0%	25.0%	**38.3%**
bookcase	23.9%	24.6%	24.4%	**25.7%**
board	44.5%	26.8%	18.2%	**49.0%**
clutter	29.3%	24.7%	30.4%	**33.3%**

Table 7. IoU per class in test T2 (10% of the dataset for training and 90% for testing).

ACKNOWLEDGEMENTS

First author would like to thank German Academic Exchange Service (DAAD) for providing the scholarship which enabled conducting the research presented in this paper.

REFERENCES

Armeni, I., Sax, A., Zamir, A. R. and Savarese, S., 2017. Joint 2D-3D-Semantic Data for Indoor Scene Understanding. *ArXiv e-prints*.

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 the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1534–1543.

Badrinarayanan, V., Kendall, A. and Cipolla, R., 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence* 39(12), pp. 2481–2495.

Becker, S., 2009. Generation and application of rules for quality dependent facade reconstruction. *ISPRS Journal of Photogrammetry and Remote Sensing* 64(6), pp. 640–653.

https://doi.org/10.5194/isprs-archives-XLII-1-225-2018 |

Channels	GlobalAcc	MeanAcc	Mean IoU
RGB	71.8%	62.9%	48.7%
HSD	69.4%	59.4%	45.7%
$R_dG_dB_d$	72.8%	67.8%	55.0%
RGBD	**74.7%**	**70.7%**	**56.9%**

Table 8. Results on semantic labeling of structural elements of buildings in test T2 (10% of the dataset for training and 90% for testing).

Becker, S., Peter, M. and Fritsch, D., 2015. Grammar-supported 3d indoor reconstruction from point clouds for” as-built” bim. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 2(3), pp. 17.

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

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. and Yuille, A. L., 2018. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence* 40(4), pp. 834–848.

Del Pero, L., Bowdish, J., Fried, D., Kermgard, B., Hartley, E. and Barnard, K., 2012. Bayesian geometric modeling of indoor scenes. In: *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, IEEE, pp. 2719–2726.

Förstner, W. and Korč, F., 2009. etrims image database for interpreting images of man-made scenes. Technical report, Department of Photogrammetry, University of Bonn. (TR-IGG-P-2009-01).

Garcia, D., 2010. Robust smoothing of gridded data in one and higher dimensions with missing values. *Computational statistics & data analysis* 54(4), pp. 1167–1178.

Girshick, R. B., Donahue, J., Darrell, T. and Malik, J., 2013. Rich feature hierarchies for accurate object detection and semantic segmentation. *CoRR*.

Gupta, S., Girshick, R., Arbeláez, P. and Malik, J., 2014. Learning rich features from rgb-d images for object detection and segmentation. In: *European Conference on Computer Vision*, Springer, pp. 345–360.

Hazirbas, C., Ma, L., Domokos, C. and Cremers, D., 2016. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In: *Asian Conference on Computer Vision*, Springer, pp. 213–228.

Karpathy, A., 2014. What i learned from competing against a convnet on imagenet. Web: http://karpathy.github.io/2014/09/02/what-i-learned-from-competing-against-a-convnet-on-imagenet, Last Access: Apr. 1, 2018.

Kontschieder, P., Bulo, S. R., Bischof, H. and Pelillo, M., 2011. Structured class-labels in random forests for semantic image labelling. In: *Computer Vision (ICCV), 2011 IEEE International Conference on*, IEEE, pp. 2190–2197.

Liu, Z., Zhang, Y., Wu, W., Liu, K. and Sun, Z., 2015. Model-driven indoor scenes modeling from a single image. In: *Proceedings of the 41st Graphics Interface Conference*, Canadian Information Processing Society, pp. 25–32.

Long, J., Shelhamer, E. and Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440.

Ripperda, N., 2008. Grammar based façade reconstruction using rjmcmc. *PFG Photogrammetrie Fernerkundung Geoinformation* 2008(2), pp. 83–92.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C. and Fei-Fei, L., 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115(3), pp. 211–252.

Schneider, L., Jasch, M., Fröhlich, B., Weber, T., Franke, U., Pollefeys, M. and Rätsch, M., 2017. Multimodal neural networks: Rgb-d for semantic segmentation and object detection. In: P. Sharma and F. M. Bianchi (eds), *Image Analysis*, Springer International Publishing, Cham, pp. 98–109.

Sherrah, J., 2016. Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery.

Shotton, J., Johnson, M. and Cipolla, R., 2008. Semantic texton forests for image categorization and segmentation. In: *Computer vision and pattern recognition, 2008. CVPR 2008. IEEE Conference on*, IEEE, pp. 1–8.

Simonyan, K. and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Toth, C. K. and Koppanyi, Z., 2017. Localization using region-based convolution neural network: A comparison study. In: *Proceedings of the 10th International Conference on Mobile Mapping Technology*, Cairo, Eqypt.

Tuttas, S. and Stilla, U., 2011. Window detection in sparse point clouds using indoor points. In: *ISPRS International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vol. 38-3/W22, pp. 131–136. Proceedings of PIA11 - Photogrammetric Image Analysis.

Vadivel, A., Sural, S. and Majumdar, A., 2005. Human color perception in the hsv space and its application in histogram generation for image retrieval.

Wang, G., Garcia, D., Liu, Y., De Jeu, R. and Dolman, A. J., 2012. A three-dimensional gap filling method for large geophysical datasets: Application to global satellite soil moisture observations. *Environmental Modelling & Software* 30, pp. 139–142.

Yu, F. and Koltun, V., 2015. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*.

Zeng, A., Yu, K.-T., Song, S., Suo, D., Walker, E., Rodriguez, A. and Xiao, J., 2017. Multi-view self-supervised deep learning for 6d pose estimation in the amazon picking challenge. In: *Robotics and Automation (ICRA), 2017 IEEE International Conference on*, IEEE, pp. 1386–1383.

Zhao, H., Shi, J., Qi, X., Wang, X. and Jia, J., 2017. Pyramid scene parsing network. In: *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.