Semantic Labeling of Structural Elements in Buildings by Fusing RGB and Depth Images in an Encoder-Decoder CNN

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

[a] Dept. of Civil, Geo and Environmental Engineering, Technical University of Munich, 80333 Munich, Germany

[b] Dept. of Civil, Environmental and Geodetic Engineering, The Ohio State University, Columbus, OH, 43212, USA

1 Motivation

- Building modelling including geometry and semantics important for Geographical Information Systems (GIS) and Building Information Model (BIM)
- Focus on indoor mapping
- Deep Learning as state-of-the-art approach for semantic labelling
- Using 3D data together with image data is expected to improve segmentation results

2 Sensor fusion with CNN

- SegNet-based architecture (Badrinarayanan et al., 2017)
- 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 3x3 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
- Pixel-wise classification using Softmax classifier

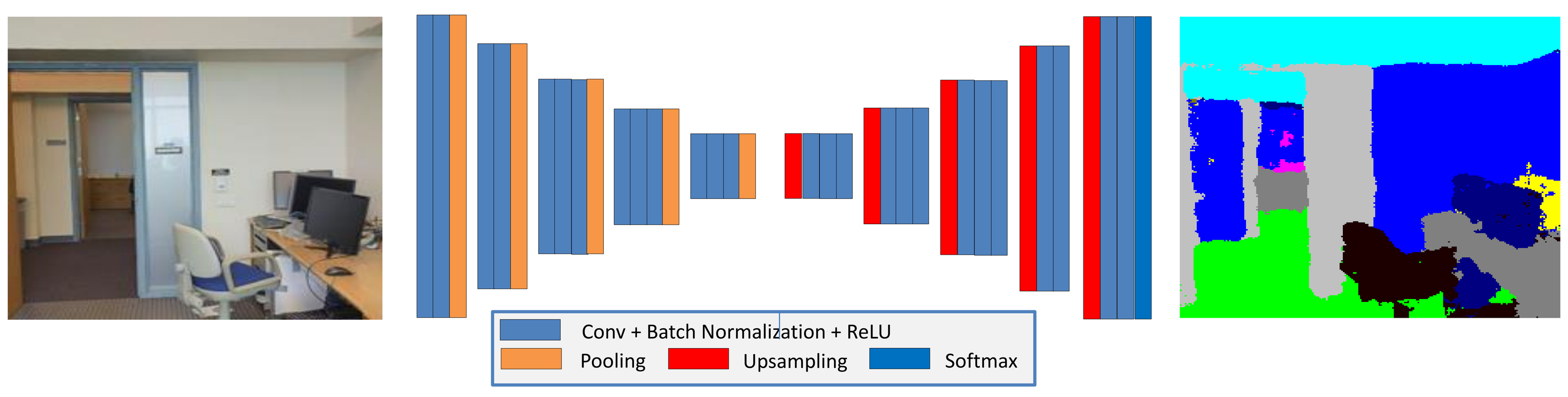

Fig. 1: SegNet-based encoder-decoder architecture for semantic labeling using RGB and depth images

We fuse the RGB and depth information by combining the depth with the reduced color space. We perform this fusion in two different ways:

- **Fusion F1**: transforming RGB image to HSV color space and replacing the value component with depth
- 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

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

- The third component D is generated from depth values normalized to [0,1]

- **Fusion F2**: transforming this HSD image 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

$$(R_d, G_d, B_d) = \begin{cases} (D, c, a), \text{ for } c_1 = 0 \\ (b, D, a), \text{ for } c_1 = 1 \\ (a, D, c), \text{ for } c_1 = 2 \\ (a, b, D), \text{ for } c_1 = 3 \\ (c, a, D), \text{ for } c_1 = 4 \\ (D, a, b), \text{ for } c_1 = 5 \end{cases} \quad \text{where} \quad \begin{aligned} a &= \frac{D(c_{\max} - c_{\min})}{1 - c_{\max}}, \\ b &= \frac{D(c_{\max} - c_{\min})(H/60 - c_1)}{1 - c_{\max}}, \\ c &= \frac{D(c_{\max} - c_{\min})(H/60 - c_1)}{c_{\max} + 1}. \end{aligned}$$

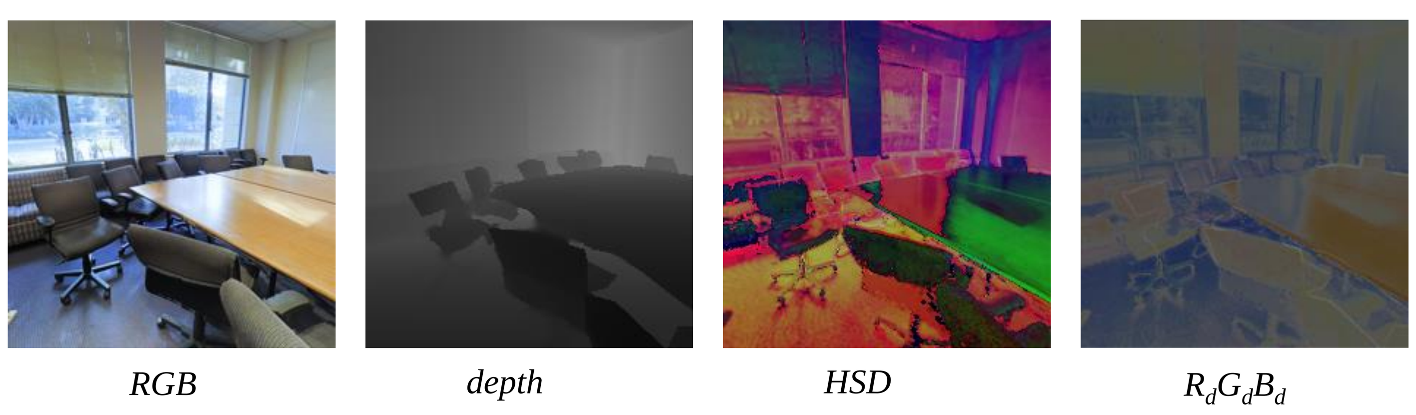

Fig. 2: An exemplary image from the dataset

* corresponding author: dorota.iwaszczuk@tum.de

3 Dataset

- Stanford 2D-3D Semantics Dataset (2D-3D-S) (Armeni et al., 2017).
- Collected using the Matterport Camera, which combines 3 structured-light sensors to capture RGB and 360° depth images.
- Consist of 6 indoor areas including 3D textured mesh, RGB-D images and semantic pixel-wise annotations.
- 13 object classes, including ceiling, floor, wall, column, beam, window, door, table chair, bookcase, sofa, board, and clutter. Sofa class is, however, underrepresented, therefore this class was merged with class clutter.
- Preprocessing
  - Resizing: 224x224
  - Depth filtering: Inpainting

4 Results

- We use Area 1 of 2D-3D-S dataset for our experiments
- Test T1: 50% of the images for training (5164 images) and the other 50% for validation (5163 images)
- Test T2: 10% of the data (1047 images) for training (six selected rooms: three offices, two hallways and one conference room) and 90% for testing
- Focus on <u>structural elements in buildings</u>
- Evaluation:

$$\text{GlobAcc} = \frac{1}{N}\sum \text{TP}_c$$

$$\text{MeanAcc} = \frac{1}{K}\sum \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c}$$

$$\text{IoU} = \frac{1}{K}\sum \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c + \text{FN}_c}$$

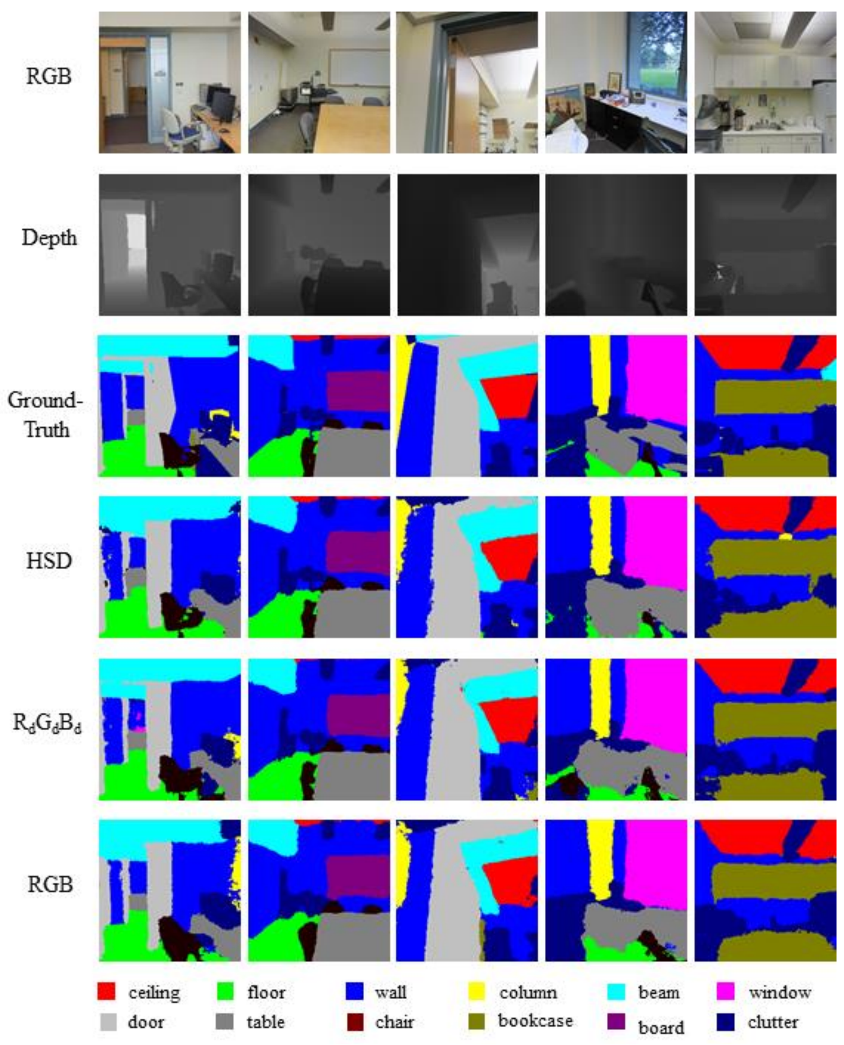

Fig. 3: Results of the label prediction.

Tab. 1: Results on semantic labeling in test T1

Channels	GlobalAcc	MeanAcc	Mean IoU
RGB	90.9%	92.5%	81.2%
HSD	91.4%	92.5%	82.0%
$R_dG_dB_d$	92.1%	**93.5%**	83.2%
RGBD	**93.6%**	92.8%	**86.3%**

Tab. 2: Results on semantic labeling in test T2

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%**

Tab. 3: Results on semantic labeling of structural elements of buildings in test T1

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%**

Tab. 4: Results on semantic labeling of structural elements of buildings in test T2

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%**

5 Discussion & Outlook

- Incorporating depth improves slightly the labeling results in an indoor scene
- For structural elements of buildings, this improvement is even more significant
- $R_dG_dB_d$ representation delivers better results than HSD representation
- Using RGBD input up to 2% higher accuracy can be achieved
- RGBD input improves IoU for almost all classes compared to RGB and $R_dG_dB_d$ input

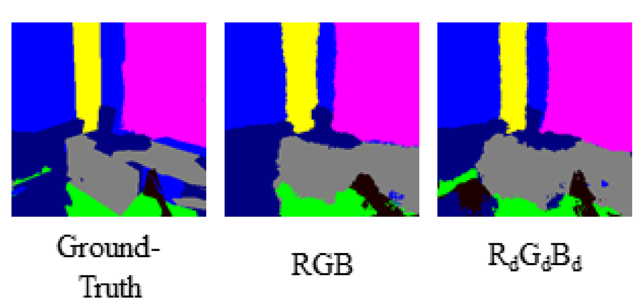

Fig. 4: Improvement of the labeling at the boundaries using depth on example of class column (yellow).

References:

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