Locality-Sensitive Deconvolution Networks with Gated Fusion for RGB-D Indoor Semantic Segmentation

Yanhua Cheng[1,2], *Rui Cai*[3], *Zhiwei Li*[3], *Xin Zhao*[1,2], *Kaiqi Huang*[1,2,4]

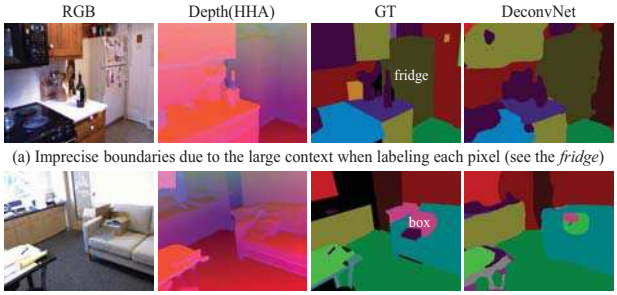
[1]CRIPAC&NLPR, CASIA [2]University of Chinese Academy of Sciences [3]Microsoft Research
[4]CAS Center for Excellence in Brain Science and Intelligence Technology

Abstract

This paper focuses on indoor semantic segmentation using RGB-D data. Although the commonly used deconvolution networks (DeconvNet) have achieved impressive results on this task, we find there is still room for improvements in two aspects. One is about the boundary segmentation. DeconvNet aggregates large context to predict the label of each pixel, inherently limiting the segmentation precision of object boundaries. The other is about RGB-D fusion. Recent state-of-the-art methods generally fuse RGB and depth networks with equal-weight score fusion, regardless of the varying contributions of the two modalities on delineating different categories in different scenes. To address the two problems, we first propose a locality-sensitive DeconvNet (LS-DeconvNet) to refine the boundary segmentation over each modality. LS-DeconvNet incorporates locally visual and geometric cues from the raw RGB-D data into each DeconvNet, which is able to learn to upsample the coarse convolutional maps with large context whilst recovering sharp object boundaries. Towards RGB-D fusion, we introduce a gated fusion layer to effectively combine the two LS-DeconvNets. This layer can learn to adjust the contributions of RGB and depth over each pixel for high-performance object recognition. Experiments on the large-scale SUN RGB-D dataset and the popular NYU-Depth v2 dataset show that our approach achieves new state-of-the-art results for RGB-D indoor semantic segmentation.

1. Introduction

Semantic segmentation of indoor scenes is a fundamental problem in computer vision, which can benefit many intelligent applications such as domestic robots, SLAM, content-based image retrieval, etc. However, it is a very tough task due to challenges from large variations of scene types, cluttered backgrounds, severe object occlusions and varying illuminations. Thanks to recent consumer depth cameras, e.g. Kinect, we are able to capture high-quality synchronized visual (RGB data) and geometrical (depth data) cues to depict one scene. It represents an opportunity to improve the performance of indoor scene segmentation by taking full advantage of the two complementary modalities.



(a) Imprecise boundaries due to the large context when labeling each pixel (see the *fridge*)

(b) Misclassified objects due to the improper fusion of RGB and depth (see the *box*)

Figure 1. Limitations of DeconvNet on indoor scene segmentation. Here a two-stream DeconvNet is used to represent RGB and depth, followed by score fusion with equal-weight sum just like the FCN model [19]. Note that the depth data in this paper is encoded to three-channel HHA image as the method [11]. See our results in Fig. 4 for comparison.

Extensive studies have been carried out on indoor semantic segmentation. Graphical models with handcrafted RGB features (e.g. SIFT, HOG, LBP, etc.) and depth features (e.g. SPIN images, depth kernels, surface normals, etc.) are used in many methods [23, 22, 10, 7, 15]. Instead of the handcrafted features, patch-wise CNN models [5] and R-CNN models [11] are proposed to learn RGB-D features of the superpixels or region proposals. Recently, fully convolutional networks (FCN) [19, 27] have significantly pushed forward the performance of semantic segmentation, including both indoor and outdoor scenes. FCN adapts the CNN model designed for classification into an end-to-end system for holistic scene segmentation. Through the repeated max-pooling and downsampling at multiple layers, FCN learns invariant features embedded with large context for robust prediction of each pixel, yet producing a coarse label map with low-resolution and imprecise boundaries.

Towards RGB-D fusion, a simple sum fusion with equal weights is adopted by [19] to combine the predictions of RGB and depth FCN models.

Remarkable efforts [3, 4, 21, 29, 17] have been invested to improve FCN for scene segmentation. Among these extensions, DeconvNet [21] is a very effective and efficient method to refine the coarse label map of FCN. The core idea of DeconvNet is to learn multi-layer deconvolution networks to upsample the low-resolution label map of FCN into full resolution with more details. We adapt DeconvNet to RGB-D indoor scene segmentation with the same fusion way of FCN, which achieves large performance gain compared to FCN in our experiments. Nonetheless, we find there is still room for improvements in two aspects. One is about the boundary segmentation. Though high-resolution label map can be generated, such convolution-deconvolution networks of DeconvNet aggregates large context for dense prediction, reducing its sensitiveness to object boundaries. As shown in Fig. 1(a), DeconvNet segments the *fridge* with inflated contours. The other one is about RGB-D fusion. RGB and depth can have varying contributions in recognition of different categories in different scenes. As shown in Fig. 1(b), both the visual and geometrical cues are beneficial to recognize *sofa*, while emphasizing the two modalities equally can confuse the recognition of *box* (misclassified as *pillow* due to the confused shape).

This paper aims to augment DeconvNet for indoor semantic segmentation with RGB-D data. **Our first contribution** is to address the problem of boundary segmentation. Inspired by recent CRF-RNN model [29], which leverages pixel-level cues such as intensity and location via conditional random fields (CRF) to refine label agreements of the large-context FCN maps, we try to benefit DeconvNet from pixel-level cues similarly but get rid of the complex training and inference of the CRF model. To this end, we propose a locality-sensitive DeconvNet for semantic segmentation. Specifically, an affinity matrix is constructed for each scene to describe pairwise relations between neighboring pixels (similar or not) based on low-level RGB-D features [10]. Then the affinity matrix is embedded into the DeconvNet to encourage the labeling consistency of local similar pixels (termed as "locality-sensitive") along with deconvolution operations for upsampling (See Fig. 2). Such a locality-sensitive DeconvNet can result in a high-resolution segmentation map with precise object boundaries. **Our second contribution** is to combine RGB and depth cues more effectively for semantic segmentation. Instead of the simple score fusion with equal weights for the two modalities like [19], we devise a gated fusion layer to automatically learn the varying contributions of each modality for classifying different categories in different scenes. The gated fusion layer is implemented by a series of standard layers with learnable parameters, which makes our whole system (RGB LS-DeconvNet + depth LS-DeconvNet + Gated Fusion, termed as "LSD-GF") can be trained end-to-end via efficient back propagation algorithms. Experimental results on the large-scale SUN RGB-D dataset [25] and the popular NYU-Depth v2 dataset [23] demonstrate that LSD-GF can significantly improve the semantic segmentation of RGB-D indoor scenes.

The rest of this paper is organized as follows. We first review related work in Section 2. Then the details of the proposed approach are introduced in Section 3. Extensive experimental results as well as analyses are reported in Section 4. Finally, we draw conclusions in Section 5.

2. Related Work

Refine Boundaries for Semantic Segmentation. Many studies have been made to refine object boundaries of the prediction map, since it highly affects the visualization and accuracy of semantic segmentation. Here we mainly focus on deep learning models, and divide previous work into two groups. One group utilizes post-processing method to ameliorate the resulted segmentation map. Couprie *et al.* [5, 9] apply the superpixels generated by graph cuts to smooth the predictions. Chen *et al.* [3, 4] adopt fully connected condition random fields (CRF) to optimize the holistic segmentation map. Another one focuses on designing particular deep learning models for dense prediction. CRF is incorporated into FCN by [29, 17] to encourage spatial and appearance consistency in the labelling outputs. Affinity CNNs [2, 20] embed additional pixel-wise similarity loss into FCN for dense prediction. Compared to these methods, DeconvNet [21] is a simple but effective and efficient method to refine the segmentation map by learning multi-layer deconvolution networks. However, the potentials of DeconvNet can be limited since the high-level prediction map aggregates large context for dense prediction. Similar to this paper, He *et al.* [12] also attempt to improve DeconvNet, while they only add one data driven pooling layer on top of DeconvNet to smooth the predictions in every superpixel. Different from them, this paper devises a locality-sensitivity DeconvNet to produce structured outputs with precise boundary segmentation. Experimental results show our model is superior to that of [12] on both the SUN RGB-D dataset and the NYU-Depth v2 dataset.

Combine RGB and Depth Data for Semantic Segmentation. An effective fusion of the two complementary modalities can improve the performance of semantic segmentation. Most methods [23, 22, 10] simply concatenate the handcrafted RGB and depth features to represent each pixel or superpixel. Some approaches [7, 15] incorporate both the RGB and depth cues into graphical models like MRFs or CRFs for semantic segmentation. Very recently,

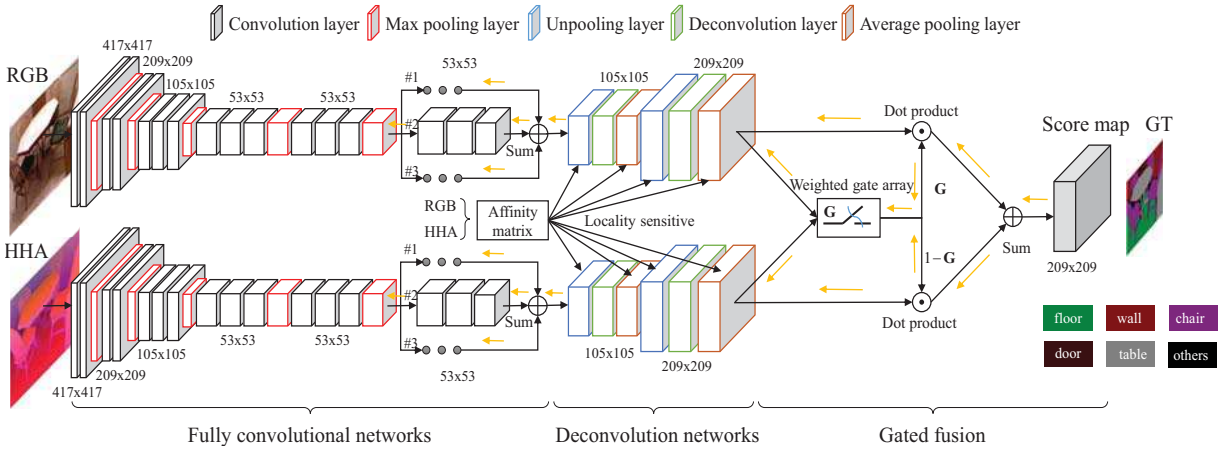


Figure 2. The overall architecture of our LSD-GF model. LSD-GF mainly consists of three parts: 1) the frontend fully convolutional networks (FCN). This paper adopts recent state-of-the-art FCN model [4], which leverages multi-scale atrous algorithm to alleviate the resolution loss and learn robust features; 2) the intermediate locality-sensitive deconvolution networks. An affinity matrix embedded with pairwise relations between neighboring RGB-D pixels is incorporated into the unpooling and average pooling operations to recover sharp boundaries of FCN maps. Due to the computational cost, only two-layer deconvolution networks are used; 3) the final gated fusion layer. We merge the RGB and depth score maps to learn a weighted gate array to weigh the contribution of each modality for object recognition in the scene. The overall networks can be trained efficiently as an end-to-end system (except for the affinity matrix). Best viewed in color.

recurrent networks [16] are explored for RGB-D fusion. Towards the popular convolutional neural networks (CNN), three levels of fusion are often used: Couprie *et al.* [5] concatenate the RGB and depth image as four-channel input for the CNN model (early fusion); Gupta *et al.* [11] leverage two CNN models to extract features from RGB and depth images independently, and then concatenate them to learn the final semantic classifier (middle fusion); Long *et al.* [19] also learn two independent CNN models but directly predict the score map of each modality, followed by score fusion with equal-weight sum (late fusion). Through comparison experiments, Long *et al.* find the late fusion can be more effective to benefit from the complementarities of the two modalities, compared to other fusion levels. This paper adopts the late fusion version, but embeds a gate fusion layer to further adapt our model to the varying contributions of the two modalities for recognition of different categories in different scenes. As shown in the experiments, the proposed fusion way can achieve performance gains for those confused categories.

3. Our Approach

3.1. Overall Architecture

Fig. 2 illustrates the overall architecture of the proposed LSD-GF model. LSD-GF is composed of three parts: the frontend fully convolutional networks (FCN), the intermediate locality-sensitive deconvolution networks (LS-DeconvNet), and the final gated fusion layer. FCN is to learn robust feature representation for each pixel by aggregating multi-scale contextual cues. The proposed LS-DeconvNet is used to restore high-resolution and precise scene details based on the coarse FCN map. Finally, a gated fusion layer is introduced to fuse the RGB and depth cues effectively for accurate scene semantic segmentation.

We adopt recent state-of-the-art fully convolutional version, termed as ASPP [4] as the frontend model. ASPP is derived from the VGG 16-layer net [24], but embedding atrous algorithm into the last convolution layers (i.e., conv5_1$\sim$conv5_3), whilst replacing all the fully connected layers (i.e., fc6$\sim$fc8) with multi-stream and multi-atrous convolution layers. LS-DeconvNet consists of a series of unpooling, deconvolution and average pooling layers. We employ the standard deconvolution operation as [21], but incorporate pixel-centric affinity matrix into both the unpooling and pooling operations to recover sharp boundaries along with upsampling. Towards the gated fusion layer, we concatenate the prediction maps of RGB and depth to learn a weighted gate array, which is able to weigh the contributions of each modality for accurate object recognition in the scene. More details of the proposed LSD-GF model are described in the following subsections.

3.2. Locality-Sensitive DeconvNet

We now discuss the details of the unpooling, deconvolution and average pooling operations in our LS-DeconvNet.

3.2.1 Locality-Sensitive Unpooling

The conventional unpooling [28, 21] performs reverse operation of max pooling to enlarge the activations of the responding map. For example, a max pooling layer in the convolutional networks employs a pooled window of 3×3 size, and the location of the maximum activation

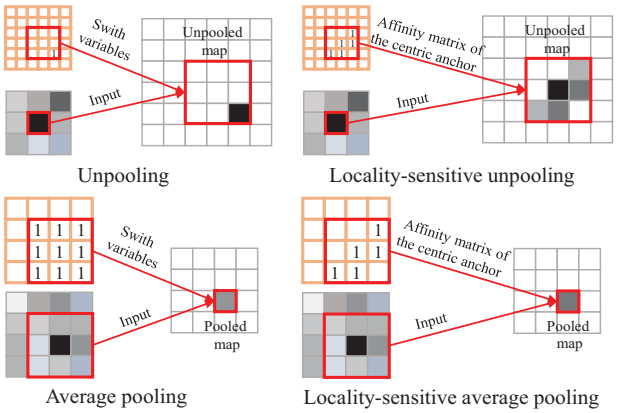


Figure 3. Illustration of the locality-sensitive unpooling as well as the locality-sensitive average pooling of LSD-GF. For clear comparisons, we only show the result of one filter window with 3×3 size (red rectangle) for both the conventional ones and ours.

places in the bottom right, which is recorded in the switch variables. For the corresponding unpooling in the deconvolution networks, it places each activation back to its original pooled location based on the switch variables, as illustrated in the top left of Fig. 3. Although those methods [28, 21] demonstrate unpooling is helpful to reconstruct detailed object boundaries, its capability can be limited a lot due to the excessive dependence on the input responding map with large context.

To address this issue, we incorporate locally visual and geometrical cues into unpooling for restoring precise object boundaries, and term it "locality-sensitive unpooling". Assuming $\mathbf{F}_{\mathrm{s}}^{\mathrm{un}} \in \mathbb{R}^{c\times h\times w}$ denotes the input responding map, where c is the number of feature channels, h is the height and w is the width. The output unpooled map is $\mathbf{F}_{\mathrm{t}}^{\mathrm{un}} \in \mathbb{R}^{c\times nh\times nw}$ with an amplification of n times. $\mathbf{A} \in \mathbb{R}^{hw\times hw}$ is the holistic affinity matrix denoting pairwise similarity between all pixels. For each feature vector $\mathbf{x} \in \mathbf{F}_{\mathbf{s}}^{\mathrm{un}}$ (regarded as an anchor), we generate a local pixel-centric affinity matrix $\mathbf{A}^{\mathbf{x}} = \{A^x_{i,j} | 1 \le i,j \le s\}$ with size $s \times s$ through cropping $\mathbf{A}$, where $A^x_{i,j} = 1$ indicates the neighboring pixel is similar to the centric anchor, and $A^x_{i,j} = 0$ indicates not. Let $\mathbf{Y} \subseteq \mathbf{F}_t^{\mathrm{un}}$ be the resulted $s \times s$ unpooled map corresponding to $\mathbf{x}$. We compute $\mathbf{Y}$ based $\mathbf{x}$ as

$$\begin{aligned} &\mathbf{Y_{i,j}} = \frac{(s-1-|i-o_i|)(s-1-|j-o_j|)}{(s-1)^2} A_{i,j}\mathbf{x}, \\ &\forall i,j \in [1,s], o_i = o_j = \tfrac{1+s}{2}. \end{aligned} \tag{1}$$

In the resulted $s \times s$ unpooled map, $\mathbf{Y_{i,j}}$ is the feature vector of the i-th row, and the j-th column. $o = (o_i, o_j)$ is the centric location mapping to the anchor $\mathbf{x}$. An example of $s = 3$ is shown in the top right of Fig. 3. It is noted that $\mathbf{Y_{i,j}}$ can also receive activations from other anchors, and we aggregate all these activations by linear addition to generate the final unpooled map. Indeed, the locality-sensitive unpooling performs like a bilinear interpolation but emphasize the influence of the neighboring similar pixels. Compared to the very sparse responding map produced by the conventional unpooling, the proposed one leads to much denser map whilst keeping sensitive to the local object boundaries.

3.2.2 Deconvolution

The output of our unpooling layer is an enlarged activation map, yet with many discontinuous boundary responses. We employ deconvolution to make up the missing details with multiple learned filters. Deconvolution performs like the reverse convolution operation. Instead of aggregating multiple input activations within a filter window to a single activation, it maps a single input activation to multiple outputs. Such an operation can effectively connect many discontinuous boundaries and reconstruct rich object structures for semantic segmentation. More details of deconvolution can be found in [21].

The resulted map of deconvolution is also enlarged, but more smoothing. We crop the map to keep it with the identical output size of the unpooling layer.

3.2.3 Locality-Sensitive Average Pooling

To further enhance the consistent representation of spatially neighboring pixels that have both similar appearance and geometry, we add the locality-sensitive average pooling layer (without downsampling) on top of the deconvolution layer. For better understanding of the proposed pooling strategy, we introduce the conventional version at first. As shown in the bottom left of Fig. 3, the conventional average pooling computes the mean value of the activations within a filter window for single output. Such an operation can achieve more robust feature representation against noise and clutter, while it is probable to blur object boundaries and result in imprecise semantic segmentation map. In order to keep the advantages of the conventional average pooling but get rid of its drawbacks, we leverage the aforementioned pixel-centric affinity matrix to force that only local similar pixels contribute to the average pooling for the corresponding outputs, as shown in the bottom right of Fig. 3.

Specifically, let $\mathbf{F}_{\mathrm{s}}^{\mathrm{avg}} \in \mathbb{R}^{c\times h\times w}$ and $\mathbf{F}_{\mathrm{t}}^{\mathrm{avg}} \in \mathbb{R}^{c\times h\times w}$ denote the input and output responding maps respectively for the locality-sensitive pooling layer. Given a feature set $\mathbf{X} \subseteq \mathbf{F}_{\mathrm{s}}^{\mathrm{avg}}$ within an $s \times s$ filter window, we compute the corresponding output feature vector $\mathbf{y} \in \mathbf{F}_{\mathrm{t}}^{\mathrm{avg}}$ by pooling $\mathbf{X}$ as

$$\mathbf{y} = \frac{1}{\sum_{i,j\in[1,s]} A^y_{i,j}} \sum\nolimits_{i,j\in[1,s]} A^y_{i,j}\mathbf{X}_{i,j}. \tag{2}$$

Similar to unpooling, $\mathbf{A}^{\mathbf{y}}$ is the local pixel-centric affinity matrix corresponding to the anchor $\mathbf{y}$. Through the locality-

sensitive average pooling, we can achieve consistent and robust feature representation for the consecutive object structures.

3.3. Gated Fusion

The gated fusion layer is proposed to effectively combine RGB and depth for semantic segmentation. Actually, it is composed of three layers, including a concatenation layer, a convolution layer and a sigmoid layer, which are not illustrated in Fig. 2 for brevity. Let $\mathbf{P}^{\mathrm{rgb}} \in \mathbb{R}^{c\times h\times w}$ and $\mathbf{P}^{\mathrm{depth}} \in \mathbb{R}^{c\times h\times w}$ denote the probability maps on RGB and depth, respectively. Here the number of feature channels c equals to the number of the categories. After concatenation, we obtain a fused probability map $\mathbf{P}^{\mathrm{fusion}} \in \mathbb{R}^{2c\times h\times w}$. Then we employ a convolution layer with weights $\mathbf{W} \in \mathbb{R}^{c\times 2c\times 1\times 1}$ (c filters with dimension of $2c \times 1 \times 1$ per filter) to learn the correlations of the two modalities and weigh their contributions for the prediction of each category. The output of the convolution layer is a coefficient matrix $\mathbf{G} \in \mathbb{R}^{c\times h\times w}$ with the value

$$\begin{aligned} &\mathbf{G}_{k,i,j} = \sum_{k'=1}^{2c} \mathbf{P}^{\mathrm{fusion}}_{k',i,j} \times \mathbf{W}_{k,k',i,j} \\ &\forall k \in [1,c],\ i \in [1,h],\ j \in [1,w]. \end{aligned} \tag{3}$$

The subsequent sigmoid layer is used to regularize $\mathbf{G}$ to keep $\mathbf{G}_{k,i,j} \in [0,1]$. We term $\mathbf{G}^{\mathrm{rgb}} = \mathbf{G}$ and $\mathbf{G}^{\mathrm{depth}} = 1 - \mathbf{G}$ as the weighted gates, where $\mathbf{G}^{\mathrm{rgb}}_{k,i,j}$ and $\mathbf{G}^{\mathrm{depth}}_{k,i,j}$ denote how confidently we can rely on RGB and depth respectively to predict the pixel (i,j) as category k. The two coefficient matrices are utilized to weigh the contributions of RGB and depth as follows:

$$\begin{aligned} \tilde{\mathbf{P}}^{\mathrm{rgb}} &= \mathbf{P}^{\mathrm{rgb}} \odot \mathbf{G}^{\mathrm{rgb}} \\ \tilde{\mathbf{P}}^{\mathrm{depth}} &= \mathbf{P}^{\mathrm{depth}} \odot \mathbf{G}^{\mathrm{depth}}, \end{aligned} \tag{4}$$

where $\odot$ denotes Hadamard product. Finally, we generate the gated fusion probability map as

$$\tilde{\mathbf{P}}^{\mathrm{fusion}} = \tilde{\mathbf{P}}^{\mathrm{rgb}} + \tilde{\mathbf{P}}^{\mathrm{depth}}. \tag{5}$$

We predict the label map by $\tilde{\mathbf{P}}^{\mathrm{fusion}}$ and leverage the ground truth label map to optimize the whole network via stochastic gradient descent.

3.4. Implementation Details

Preprocessing. Before starting to train the networks, we need to obtain the holistic affinity matric $\mathbf{A}$ for each RGB-D scene. Following the method [10], we extract low-level RGB-D features (gradients over visual and geometrical cues) for each pixel, and employ gPb-ucm [1] to generate over-segments. These over-segments can be used to calculate $\mathbf{A}$ by verifying that pairwise pixels belong to the same over-segment (similarity is 1) or not (similarity is 0). Note that we will scale $\mathbf{A}$ to match the resolution of the corresponding feature maps.

Optimization. We utilize the popular Caffe framework [13] to implement the proposed networks. The training process can be divided into two stages. In the first stage, we train two independent locality-sensitive DeconvNets on RGB and depth for semantic segmentation without the gated fusion layer. For each modality, we employ the "poly" learning rate policy (the learning rate is multiplied by $(1 - \frac{iter}{max\_iter})^{power}$) to optimize the networks, for which the base learning rate is set to 0.001, power is 0.9, weight decay is 0.0005 and max iteration is 20000. The frontend FCN model is initialized by VGG 16-layer net pretrained on imageNet [6]. The intermediate deconvolution layers is initialized with identity filters following [27], whilst a smaller layer learning rate $lr\_mult = 0.01$ is used instead of $lr\_mult = 1$ for other layers. We leverage 5×5 local pixel-centric affinity matrix for all the unpooling and average pooling layers, except for the last average pooling layer, which uses 11×11 size. We find these settings can be more effective to train the networks of each modality for semantic segmentation. In the second stage, we add the gated fusion layer, and then finetune the whole networks on the synchronized RGB and depth data. We use the same "poly" learning rate policy but with a smaller base learning rate (set to 10^{-6}). It is noted that the conventional DeconvNet [21] utilize additional region proposals and batch normalization to train their networks, while our networks are directly trained on the cropped images with 417×417 size very efficiently. During the testing phase, we utilize the trained LSD-GF model but enlarge the last average pooling size to 15×15 for more accurate segmentation.

4. Experiments

4.1. Experimental Setup

Datasets. We evaluate our approach for indoor scene segmentation on two benchmark RGB-D datasets, including the large-scale SUN RGB-D dataset [25] and the popular NYU-Depth v2 dataset [23]. The SUN RGB-D dataset consists of 10355 RGB-D images with pixel-wise labels, which are collected from five appealing datasets. Following the setting of [25], we divide the dataset into a training set with 5285 images and a test set with 5050 images. The NYU-Depth v2 dataset consists of 1449 RGB-D images from indoor scenes, which provides 795 images for training and the remaining 654 images for evaluation.

Metrics. Following recent methods [23, 10, 19, 25], this paper employs four metrics to evaluate the performance of semantic segmentation, such as pixel accuracy, mean accuracy, mean IOU and frequency weighted IOU (f.w. IOU). Let n_{ij} be the number of pixels of class i classified

Table 1. Comparison results of scene semantic segmentation on the SUN RGB-D dataset with class-wise accuracy as well as mean accuracy over all classes. Note that the pixels of the class "background" are ignored for performance evaluation.

	wall	floor	cabinet	bed	chair	sofa	table	door	window	bookshelf	picture	counter	blinds	desk	shelves	curtain	dresser	pillow	mirror
Song *et al.*[25]	37.8	45.0	17.4	21.8	16.9	12.8	18.5	6.1	9.6	9.4	4.6	2.2	2.4	7.3	1.0	4.3	2.2	2.3	6.9
Song *et al.*[25]	32.1	42.6	2.9	6.4	21.5	4.1	12.5	3.4	5.0	0.8	3.3	1.7	14.8	2.0	15.3	2.0	1.4	1.2	0.9
Song *et al.*[25]	36.4	45.8	15.4	23.3	19.9	11.6	19.3	6.0	7.9	12.8	3.6	5.2	2.2	7.0	1.7	4.4	5.4	3.1	5.6
Liu *et al.*[18]	38.9	47.2	18.8	21.5	17.2	13.4	20.4	6.8	11.0	9.6	6.1	2.6	3.6	7.3	1.2	6.9	2.4	2.6	6.2
Liu *et al.*[18]	33.3	43.8	3.0	6.3	22.3	3.9	12.9	3.8	5.6	0.9	3.8	2.2	32.6	2.0	10.1	3.6	1.8	1.1	1.0
Liu *et al.*[18]	37.8	48.3	17.2	23.6	20.8	12.1	20.9	6.8	9.0	13.1	4.4	6.2	2.4	6.8	1.0	7.8	4.8	3.2	6.4
Ren *et al.*[22]	43.2	78.6	26.2	42.5	33.2	40.6	34.3	33.2	43.6	23.1	57.2	31.8	42.3	12.1	18.4	59.1	31.4	49.5	24.8
Li *et al.*[16]	74.9	82.3	47.3	62.1	67.7	55.5	57.8	45.6	52.8	43.1	56.7	39.4	48.6	**37.3**	9.6	63.4	35.0	45.8	44.5
DeconvNet	90.4	92.7	57.7	75.9	83.0	61.2	64.2	43.0	64.7	42.3	59.8	42.5	48.3	29.5	17.5	64.9	54.0	61.7	51.3
Ours	**91.9**	**94.7**	**61.6**	**82.2**	**87.5**	**62.8**	**68.3**	**47.9**	**68.0**	**48.4**	**69.1**	**49.4**	**51.3**	35.0	**24.0**	**68.7**	**60.5**	**66.5**	**57.6**

	floormat	clothes	ceiling	books	fridge	tv	paper	towel	shower	box	board	person	nightstand	toilet	sink	lamp	bathhub	bag	Mean Acc.
Song *et al.*[25]	0.0	1.2	27.9	4.1	7.0	1.6	1.5	1.9	0.0	0.6	7.4	0.0	1.1	8.9	14.0	0.9	0.6	0.9	8.3
Song *et al.*[25]	0.0	0.3	9.7	0.6	0.0	0.9	0.0	0.1	0.0	1.0	2.7	0.3	2.6	2.3	1.1	0.7	0.0	0.4	5.3
Song *et al.*[25]	0.0	1.4	35.8	6.1	9.5	0.7	1.4	0.2	0.0	0.6	7.6	0.7	1.7	12.0	15.2	0.9	1.1	0.6	9.0
Liu *et al.*[18]	0.0	1.3	39.1	5.9	7.1	1.4	1.5	2.2	0.0	0.7	10.4	0.0	1.5	12.3	14.8	1.3	0.9	1.1	9.3
Liu *et al.*[18]	0.0	0.6	13.9	0.5	0.0	0.9	0.4	0.3	0.0	0.7	3.5	0.3	1.5	2.6	1.2	0.8	0.0	0.5	6.0
Liu *et al.*[18]	0.0	1.6	49.2	8.7	10.1	0.6	1.4	0.2	0.0	0.8	8.6	0.8	1.8	14.9	16.8	1.2	1.1	1.3	10.1
Ren *et al.*[22]	**5.6**	27.0	84.5	35.7	24.2	36.5	26.8	19.2	**9.0**	11.7	51.4	35.7	25.0	64.1	53.0	44.2	47.0	18.6	36.3
Li *et al.*[16]	0.0	28.4	68.0	47.9	**61.5**	52.1	36.4	36.7	0.0	38.1	48.1	**72.6**	36.4	68.8	67.9	58.0	65.6	23.6	48.1
DeconvNet	0.4	39.8	78.3	55.0	43.9	59.6	29.4	45.2	1.5	35.9	47.7	45.3	36.0	77.6	66.6	51.2	66.1	35.8	51.9
Ours	0.0	**44.7**	**88.8**	**61.5**	51.4	**71.7**	**37.3**	**51.4**	2.9	**46.0**	**54.2**	49.1	**44.6**	**82.2**	**74.2**	**64.7**	**77.0**	**47.6**	**58.0**

as class j. Assuming there are n_{cl} different classes, $t_i = \sum_j n_{ij}$ is the total number of pixels belonging to class i, and $t = \sum_i t_i$ record the number of all pixels. The four metrics are defined as follows:

- pixel accuracy: $\sum_i n_{ii}/t$;
- mean accuracy: $\frac{1}{n_{cl}} \sum_i n_{ii}/t_i$;
- mean IOU: $\frac{1}{n_{cl}} \sum_i n_{ii}/(t_i + \sum_j n_{ji} - n_{ii})$;
- f.w. IOU: $\frac{1}{t} \sum_i t_i n_{ii}/(t_i + \sum_j n_{ji} - n_{ii})$.

4.2. Overall Performance

Table 1 and Table 2 show performance comparisons of all recent methods on the two RGB-D scene benchmarks. In addition, we provide the result of DeconvNet [21] over each RGB-D dataset as a strong baseline. Note that the only differences between DeconvNet and the proposed approach are that we replace the conventional deconvolution networks with simple sum fusion by the locality-sensitive deconvolution networks with gated fusion.

SUN RGB-D. Following recent methods [25, 18, 22, 16], we also report the mean accuracy of our approach for labeling 37 classes on the SUN RGB-D dataset. As shown in Table 1, we achieve 58.0% mean accuracy with 9.9% improvement over the recent state-of-the-art method [16]. Specifically, we yield significant performance gains over 32 classes, which demonstrate the effectiveness of the proposed approach. To further verify the particular advantages of our locality-sensitive deconvolution networks with gated fusion, we compare the results of ours to that of DeconvNet. We can see that the improvements are remarkable. We owe the improvements to two factors: 1) the local visual and geometrical cues from raw data embedded into the deconvolution networks can effectively alleviate the imprecise boundary representation from the frontend FCN model with large context; 2) the gated fusion layer can effectively combine the two complementary modalities for accurate object recognition.

NYU-Depth v2. Following recent methods [19, 11, 7, 12][1], we evaluate the four aforementioned metrics of our approach for labeling 40 classes on the NYU-Depth v2 dataset. As illustrated in Table 2, we achieve the best results over all the four metrics. Compared to recent state-of-the-art method [12], our approach yields around 5.8% improvements on mean IOU. Since the metric of class-wise IOU is more sensitive to object boundary segmentation, the performance gain of our approach compared to DeconvNet further verifies that the proposed approach can boost the boundary precision and recognition accuracy effectively.

[1] Recent methods often augment the training set with synthetic data [11] or video frames [12]. Differently, we simply pretrain our model on SUN RGB-D dataset and then finetune it on NYU v2 dataset.

Table 2. Comparison results of scene semantic segmentation on the NYU-Depth v2 dataset with class-wise IOU as well as four mentioned metrics over all classes. Note that the pixels of the class "background" are ignored for performance evaluation.

	wall	floor	cabinet	bed	chair	sofa	table	door	window	bookshelf	picture	counter	blinds	desk	shelves	curtain	dresser	pillow	mirror	floormat	clothes	ceiling
Long *et al.*[19]	69.9	79.4	50.3	66.0	47.5	53.2	32.8	22.1	39.0	36.1	50.5	54.2	45.8	11.9	8.6	32.5	31.0	37.5	22.4	13.6	18.3	59.1
Gupta *et al.*[11]	68.0	81.3	44.9	65.0	47.9	47.9	29.9	20.3	32.6	18.1	40.3	51.3	42.0	11.3	3.5	29.1	34.8	34.4	16.4	28.0	4.7	60.5
Kendall *et al.*[14]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Eigen *et al.*[8]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Deng *et al.*[7]	65.6	79.2	51.9	66.7	41.0	55.7	36.5	20.3	33.2	32.6	44.6	53.6	49.1	10.8	9.1	47.6	27.6	42.5	30.2	32.7	12.6	56.7
He *et al.*[12]	72.7	85.7	55.4	**73.6**	58.5	60.1	42.7	30.2	42.1	41.9	52.9	59.7	46.7	13.5	9.4	40.7	**44.1**	42.0	34.5	**35.6**	22.2	55.9
Li *et al.*[16]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
DeconvNet	73.9	83.4	54.0	68.4	59.9	58.8	44.4	35.8	44.9	43.4	52.3	58.6	50.3	20.0	12.8	48.1	40.2	44.2	43.7	30.7	23.3	56.4
Ours	**78.5**	**87.1**	**56.6**	70.1	**65.2**	**63.9**	**46.9**	**35.9**	**47.1**	**48.9**	**54.3**	**66.3**	**51.7**	**20.6**	**13.7**	**49.8**	43.2	**50.4**	**48.5**	32.2	**24.7**	**62.0**

	books	fridge	tv	paper	towel	shower	box	board	person	nightstand	toilet	sink	lamp	bathhub	bag	ostuct	ofurn	oprops	Pixel Acc.	Mean Acc.	Mean IOU	f.w. IOU
Long *et al.*[19]	27.3	27.0	41.9	15.9	26.1	14.1	6.5	12.9	57.6	30.1	61.3	44.8	32.1	39.2	4.8	15.2	7.7	30.0	65.4	46.1	34.0	49.5
Gupta *et al.*[11]	6.4	14.5	31.0	14.3	16.3	4.2	2.1	14.2	0.2	27.2	55.1	37.5	34.8	38.2	0.2	7.1	6.1	23.1	60.3	-	28.6	47.0
Kendall *et al.*[14]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	68.0	45.8	32.4	-
Eigen *et al.*[8]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	65.6	45.1	34.1	51.4
Deng *et al.*[7]	8.9	21.6	19.2	**28.0**	28.6	22.9	1.6	1.0	9.6	30.6	48.4	41.8	28.1	27.6	0	9.8	7.6	24.5	63.8	-	31.5	48.5
He *et al.*[12]	29.8	41.7	52.5	21.1	34.4	15.5	7.8	29.2	**60.7**	42.2	62.7	47.4	38.6	28.5	7.3	**18.8**	15.1	**31.4**	70.1	53.8	40.1	55.7
Li *et al.*[16]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	49.4	-	-
DeconvNet	30.1	43.2	53.2	26.9	**42.9**	22.2	10.6	53.5	50.7	45.2	72.2	54.5	41.6	49.7	10.6	10.6	13.8	30.1	69.9	56.4	42.7	56.0
Ours	**34.2**	**45.3**	**53.4**	27.7	42.6	**23.9**	**11.2**	**58.8**	53.2	**54.1**	**80.4**	**59.2**	**45.5**	**52.6**	**15.9**	12.7	**16.4**	29.3	**71.9**	**60.7**	**45.9**	**59.3**

4.3. Ablation Study

To discover the importance of the proposed locality-sensitive DeconvNet and the gated fusion of LSD-GF, we conduct an ablation study via removing or replacing each component independently or both together for semantic segmentation on the NYU-Depth v2 dataset. Note that both the training and testing procedures of each ablation experiment are kept exactly the same for fair comparison. We report the results on RGB only, depth only and the both, as illustrated in Table 3. We can draw conclusions as follows: 1) Embedding local visual and geometrical cues (locality-sensitive) into deconvolution networks can boost the performance of semantic segmentation considerably (comparing a vs b, c vs d, e vs i, etc.). For each comparison pair, the only difference is with and without locality-sensitive module; 2) Gated fusion is superior to the sum fusion, as well as some other popular equal-weight score fusion like pixelwise production and Dempster-Shafer (DS) [26] (comparing $e \sim h$ and $i \sim l$). We owe the improvement to the accurate recognition of some hard objects in the scene by gated fusion, such as *box* on the sofa and *chair* in the weak lights. These objects need to effectively weigh the contributions of RGB an depth for recognition; 3) Cascading the locality-sensitive deconvolution networks and the gated fusion can achieve the best result, i.e., 45.9% mean IOU. Since each proposed component can benefit one aspect of semantic segmentation, combining the both is natural to achieve the state-of-the-art result.

Table 3. Ablation study of the proposed model on the NYU-Depth v2 dataset with mean IOU.

Model	Mean IOU
a. RGB + DeconvNet	37.4
b. RGB + LS-DeconvNet	40.5
c. HHA + DeconvNet	33.4
d. HHA + LS-DeconvNet	38.7
e. RGB-HHA + DeconvNet + Sum Fusion	42.7
f. RGB-HHA + DeconvNet + Product Fusion	40.6
g. RGB-HHA + DeconvNet + DS Fusion	42.8
h. RGB-HHA + DeconvNet + Gated Fusion	43.2
i. RGB-HHA + LS-DeconvNet + Sum Fusion	45.3
j. RGB-HHA + LS-DeconvNet + Product Fusion	44.9
k. RGB-HHA + LS-DeconvNet + DS Fusion	45.8
l. RGB-HHA + LS-DeconvNet + Gated Fusion (LSD-GF)	45.9

4.4. Visualized Comparisons

Fig. 4 illustrates the visualized comparisons of semantic segmentation on NYU-Depth v2 dataset, which involves cluttered objects from various indoor scenes. On the whole, our LSD-GF approach achieves very promising results for semantic segmentation. Specifically, rows (1)$\sim$(3) of the figure show some examples to witness the effectiveness of the proposed gated fusion, e.g., it helps to correctly recognize the *box* on the *sofa* (emphasize appearance), the faraway *fridge* against the *cabinet* (emphasize shape), and

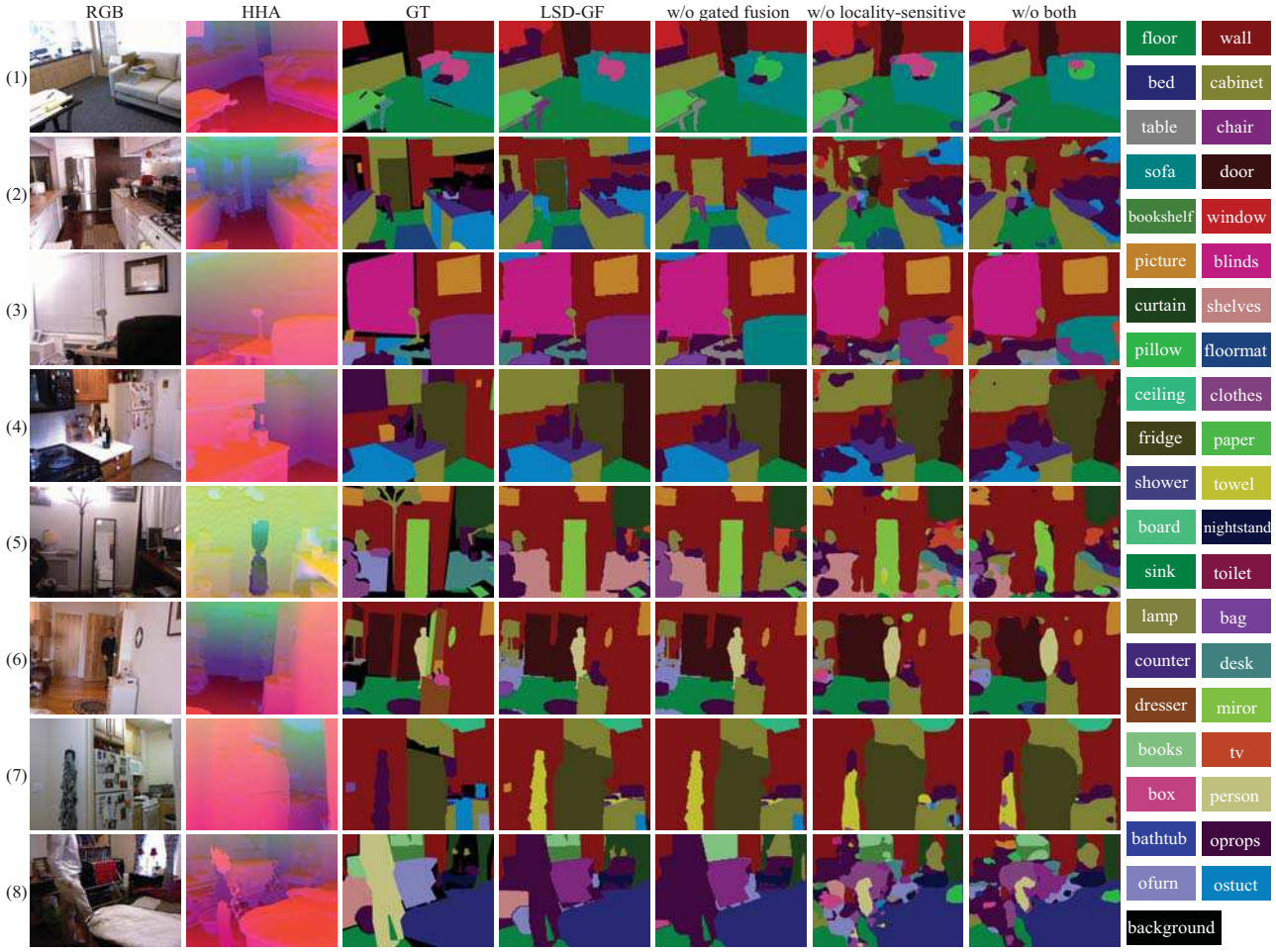


Figure 4. Visual comparison of scene semantic segmentation on the NYU-Depth v2 dataset. For the scene image in each row, we show: (column 1) the RGB image; (column 2) the HHA image; (column 3) the ground truth of semantic segmentation; (column 4) the result of our LSD-GF approach, i.e., l in Table 3; (column 5) the result of LSD-GF whose gated fusion is replaced by sum fusion, i.e., i in Table 3; (column 6) the result of LSD-GF whose locality-sensitive module is removed, i.e., h in Table 3; (column 7) the result of LSD-GF whose locality-sensitive is removed and the gated fusion is replaced by sum fusion, i.e., e in Table 3. See detailed analysis in the text. Best viewed in the magnified color image.

the *chair* with upper parts (emphasize both). Rows (4)$\sim$(6) demonstrate that the usage of locality-sensitive module can generate very precise boundary segmentation, such as the white *fridge* beside the white *door*, the *mirror* with various reflected objects, the *person* in front of the *door*. The networks without locality-sensitive module generally obtain inflated edges. Moreover, we show some failure examples in rows (7)$\sim$(8), our approach misclassify *oprops* (short for *other props*) as *towel* due to similar appearance, and mislabel *person* as *oprops* due to the occluded face.

5. Conclusion

In this paper, we propose a novel LSD-GF method for indoor semantic segmentation with RGB-D data. LSD-GF is composed of two main components: 1) the locality-sensitive deconvolution networks, which are designed for simultaneously upsamping the coarse fully convolutional maps and refining object boundaries; 2) gated fusion, which can adapt to the varying contributions of RGB and depth for better fusion of the two modalities for object recognition. Extensive experiments on recent RGB-D scene benchmarks demonstrate that LSD-GF can achieve significant performance gains compared to recent state-of-the-art methods.

Acknowledgments This work is funded by the National Key Research and Development Program of China (2016YFB1001005), the National Natural Science Foundation of China (Grant No.61673375, No.61602485), and the Projects of Chinese Academy of Science (Grant No. QYZDB-SSW-JSC006, Grant No.173211KYSB20160008). We thank the reviewers for their helpful comments to improve this paper.

References

[1] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *PAMI*, 33(5):898–916, 2011.

[2] G. Bertasius, L. Torresani, S. X. Yu, and J. Shi. Convolutional random walk networks for semantic image segmentation. *arXiv:1605.07681*, 2016.

[3] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv:1412.7062*, 2014.

[4] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv:1606.00915*, 2016.

[5] C. Couprie, C. Farabet, L. Najman, and Y. LeCun. Indoor semantic segmentation using depth information. In *ICLR*, 2013.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[7] Z. Deng, S. Todorovic, and L. Jan Latecki. Semantic segmentation of rgbd images with mutex constraints. In *ICCV*, pages 1733–1741, 2015.

[8] D. Eigen and R. Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *ICCV*, pages 2650–2658, 2015.

[9] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. *PAMI*, 35(8):1915–1929, 2013.

[10] S. Gupta, P. Arbelaez, and J. Malik. Perceptual organization and recognition of indoor scenes from rgb-d images. In *CVPR*, pages 564–571, 2013.

[11] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik. Learning rich features from rgb-d images for object detection and segmentation. In *ECCV*, pages 345–360, 2014.

[12] Y. He, W.-C. Chiu, M. Keuper, and M. Fritz. Rgbd semantic segmentation using spatio-temporal data-driven pooling. *arXiv:1604.02388*, 2016.

[13] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM MM*, pages 675–678, 2014.

[14] A. Kendall, V. Badrinarayanan, and R. Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv:1511.02680*, 2015.

[15] S. H. Khan, M. Bennamoun, F. Sohel, R. Togneri, and I. Naseem. Integrating geometrical context for semantic labeling of indoor scenes using rgbd images. *IJCV*, 117(1):1–20, 2016.

[16] Z. Li, Y. Gan, X. Liang, Y. Yu, H. Cheng, and L. Lin. Rgb-d scene labeling with long short-term memorized fusion model. *arXiv:1604.05000*, 2016.

[17] G. Lin, C. Shen, I. Reid, and A. van den Hengel. Efficient piecewise training of deep structured models for semantic segmentation. In *CVPR*, pages 345–360, 2016.

[18] C. Liu, J. Yuen, and A. Torralba. Sift flow: Dense correspondence across scenes and its applications. *PAMI*, 33(5):978–994, 2011.

[19] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015.

[20] M. Maire, T. Narihira, and S. X. Yu. Affinity cnn: Learning pixel-centric pairwise relations for figure/ground embedding. In *CVPR*, pages 174–182, 2016.

[21] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. In *ICCV*, pages 1520–1528, 2015.

[22] X. Ren, L. Bo, and D. Fox. Rgb-(d) scene labeling: Features and algorithms. In *CVPR*, pages 2759–2766, 2012.

[23] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, pages 746–760, 2012.

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

[25] S. Song, S. P. Lichtenberg, and J. Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *CVPR*, pages 567–576, 2015.

[26] H. Wu, M. Siegel, R. Stiefelhagen, and J. Yang. Sensor fusion using dempster-shafer theory [for context-aware hci]. In *Instrumentation and Measurement Technology Conference*, volume 1, pages 7–12. IEEE, 2002.

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

[28] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *ECCV*, pages 818–833, 2014.

[29] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr. Conditional random fields as recurrent neural networks. In *ICCV*, pages 1529–1537, 2015.