LSTM-CF: Unifying Context Modeling and Fusion with LSTMs for RGB-D Scene Labeling

Zhen Li[1], Yukang Gan[2], Xiaodan Liang[2], Yizhou Yu[1], Hui Cheng[2], and Liang Lin[2(✉)]

[1] Department of Computer Science, The University of Hong Kong, Hong Kong, China
lizhen36@hku.hk, yizhouy@acm.org

[2] School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China
ganyk@mail2.sysu.edu.cn, xdliang328@gmail.com, chengh9@mail.sysu.edu.cn, linliang@ieee.org

Abstract. Semantic labeling of RGB-D scenes is crucial to many intelligent applications including perceptual robotics. It generates pixelwise and fine-grained label maps from simultaneously sensed photometric (RGB) and depth channels. This paper addresses this problem by (i) developing a novel Long Short-Term Memorized Context Fusion (LSTM-CF) Model that captures and fuses contextual information from multiple channels of photometric and depth data, and (ii) incorporating this model into deep convolutional neural networks (CNNs) for end-to-end training. Specifically, contexts in photometric and depth channels are, respectively, captured by stacking several convolutional layers and a long short-term memory layer; the memory layer encodes both short-range and long-range spatial dependencies in an image along the vertical direction. Another long short-term memorized fusion layer is set up to integrate the contexts along the vertical direction from different channels, and perform bi-directional propagation of the fused vertical contexts along the horizontal direction to obtain true 2D global contexts. At last, the fused contextual representation is concatenated with the convolutional features extracted from the photometric channels in order to improve the accuracy of fine-scale semantic labeling. Our proposed model has set a new state of the art, i.e., **48**.**1**% and **49**.**4**% average class accuracy over 37 categories (**2**.**2**% and **5**.**4**% improvement) on the large-scale SUNRGBD dataset and the NYUDv2 dataset, respectively.

Keywords: RGB-D scene labeling · Image context modeling · Long short-term memory · Depth and photometric data fusion

This work was support by Projects on Faculty/Student Exchange and Collaboration Scheme between the Higher Education in Hong Kong and the Mainland, Guangzhou Science and Technology Program under grant 1563000439, and Fundamental Research Funds for the Central Universities.

B. Leibe et al. (Eds.): ECCV 2016, Part II, LNCS 9906, pp. 541–557, 2016.
DOI: 10.1007/978-3-319-46475-6_34

1 Introduction

Scene labeling, also known as semantic scene segmentation, is one of the most fundamental problems in computer vision. It refers to associating every pixel in an image with a semantic label, such as table, road and wall, as illustrated in Fig. 1. High-quality scene labeling can be beneficial to many intelligent tasks, including robot task planning [1], pose estimation [2], plane segmentation [3], context-based image retrieval [4], and automatic photo adjustment [5].

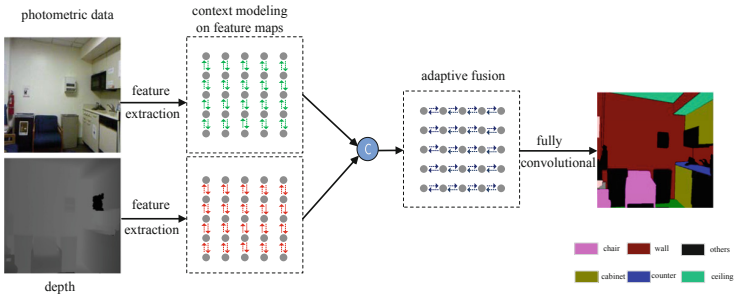


Fig. 1. An illustration of global context modeling and fusion for RGB-D images. Our LSTM-CF model first captures vertical contexts through a memory network layer encoding short- and long-range spatial dependencies along the vertical direction. After a concatenation operation (denoted by "C") over photometric and depth channels, our model utilizes another memory network layer to fuse vertical contexts from all channels in a data-driven way and performs bi-directional propagation along the horizontal direction to obtain true 2D global contexts. Best viewed in color. (Color figure online)

Previous work on scene labeling can be divided into two categories according to their target scenes: indoor and outdoor scenes. Compared with outdoor scene labeling [6–8], indoor scene labeling is more challenging due to a larger set of semantic labels, more severe object occlusions, and more diverse object appearances [9]. For example, indoor object classes, such as beds covered with different sheets and various appearances of curtains, are much harder to characterize than outdoor classes, e.g., roads, buildings, and sky, through photometric channels only. Recently, utilizing depth sensors to augment RGB data have effectively improved the performance of indoor scene labeling because the depth channel complements photometric channels with structural information. Nonetheless, two key issues remain open in the literature of RGB-D scene labeling.

(I) **How to effectively represent and fuse the coexisting depth and photometric (RGB) data.** For data representation, a batch of sophisticated hand-crafted features have been developed in previous methods. Such hand-crafted features are somewhat ad hoc and less discriminative than those RGB-D representations learned using convolutional neural networks (CNNs) [10–14]. However, in these CNN-related works, the fusion

of depth and photometric data has often been oversimplified. For instance, in [13,14], two independent CNNs are leveraged to extract features from depth and photometric data separately, and such features are simply concatenated before used for final classification. Overlooking the strong correlation between depth and photometric channels could inevitably harm semantic labeling.

(II) **How to capture global scene contexts during feature learning.** Current CNN-based scene labeling approaches can only capture local contextual information for every pixel due to their restricted receptive fields, resulting in suboptimal labeling results. In particular, long-range dependencies sometimes play a key role in distinguishing among different objects having similar appearances, e.g., labeling "ceiling" and "floor" in Fig. 1, according to the global scene layout. To overcome this issue, graphical models, such as a conditional random field [9,11] or a mean-field approximation [15], have been applied to improve prediction results in a post-processing step. These methods, however, separate context modeling from convolutional feature learning, which may give rise to suboptimal results on complex scenes due to less discriminative feature representation [16]. An alternative class of methods adopts cascaded recurrent neural networks (RNNs) with gate structures, e.g., long short-term memory (LSTM) networks, to explicitly strengthen context modeling [16–18]. In these methods, the long- and short-range dependencies can be well memorized by sequentially running the network over individual pixels.

To address the aforementioned challenges, this paper proposes a novel Long Short-Term Memorized Context Fusion (LSTM-CF) model and demonstrates its superiority in RGB-D scene labeling. Figure 1 illustrates the brief idea of using memory networks for context modeling and fusion of different channels. Our LSTM-CF model captures 2D dependencies within an image by exploiting the cascaded bi-directional vertical and horizontal RNN models as introduced in [19].

Our method constructs HHA images [13] for the depth channel through geometric encoding, and uses several convolutional layers for extracting features. Inspired by [19], these convolutional layers are followed by a memorized context layer to model both short-range and long-range spatial dependencies along the vertical direction. For photometric channels, we generate convolutional features using the Deeplab network [12], which is also followed by a memorized context layer for context modeling along the vertical direction. Afterwards, a memorized fusion layer is set up to integrate the contexts along the vertical direction from both photometric and depth channels, and perform bi-directional propagation of the fused vertical contexts along the horizontal direction to obtain true 2D global contexts. Considering the features differences, e.g., signal frequency and other characteristics (color/geometry) [20], our fusion layer facilitates deep integration of contextual information from multiple channels in a data-driven manner rather than simply concatenating different feature vectors. Since photometric channels usually contain finer details in comparison to the depth channel [20], we further

enhance the network with cross-layer connections that append convolutional features of the photometric channels to the fused global contexts before the final fully convolutional layer, which predicts pixel-wise semantic labels. Various layers in our LSTM-CF model are tightly integrated, and the entire network is amenable to end-to-end training and testing.

In summary, this paper has the following contributions to the literature of RGB-D scene labeling.

- It proposes a novel Long Short-Term Memorized Context Fusion (LSTM-CF) Model, which is capable of capturing image contexts from a global perspective and deeply fusing contextual information from multiple sources (i.e., depth and photometric channels).
- It proposes to jointly optimize LSTM layers and convolutional layers for achieving better performance in semantic scene labeling. Context modeling and fusion are incorporated into the deep network architecture to enhance the discriminative power of feature representation. This architecture can also be extended to other similar tasks such as object/part parsing.
- It is demonstrated on the large-scale SUNRGBD benchmark (including 10355 images) and canonical NYUDv2 benchmark that our method outperforms existing state-of-the-art methods. In addition, it is found that our scene labeling results can be leveraged to improve the groundtruth annotations of newly captured 3943 RGB-D images in SUNRGBD dataset.

2 Related Work

Scene Labeling: Scene labeling has caught researchers' attention frequently [6,11,12,16–18,21] in recent years. Instead of extracting features from over-segmented images, recent methods usually utilize powerful CNN layers as the feature extractor, taking advantage of fully convolutional networks (FCNs) [10] and its variants [22] to obtain pixel-wise dense features. Another main challenge for scene labeling is the fusion of local and global contexts, i.e., taking advantage of global contexts to refine local decisions. For instance, [6] exploits families of segmentations or trees to generate segment candidates. [23] utilizes an inference method based on graph cut to achieve image labeling. A pixel-wise conditional random forest is used in [11,12] to directly optimize a deep CNN-driven cost function. Most of the above models improve accuracy through carefully designed processing on the predicted confidence map instead of proposing more powerful discriminative features, which usually results in suboptimal prediction results [16]. The topological structure of recurrent neural networks (RNNs) is used to model short- and long-range dependencies in [16,18]. In [17], a multi-directional RNN is leveraged to extract local and global contexts without using a CNN, which is well suited for low-resolution and relatively simple scene labeling problems. In contrast, our model can jointly optimize LSTM layers and convolutional layers to explicitly improve discriminative feature learning for local and global context modeling and fusion.

Scene Labeling in RGB-D Images: With more and more convenient access to affordable depth sensors, scene labeling in RGB-D images [9,13,14,24–26] enables a rapid progress of scene understanding. Various sophisticated hand-crafted features are utilized in previous state-of-the-art methods. Specifically, kernel descriptions based on traditional multi-channel features, such as color, depth gradient, and surface normal, are used as photometric and depth features [24]. A rich feature set containing various traditional features, e.g., SIFT, HOG, LBP and plane orientation, are used as local appearance features and plane appearance features in [9]. HOG features of RGB images and HOG+HH (histogram of height) features of depth images are extracted as representations in [25] for training successive classifiers. In [27], proposed distance-from-wall features are exploited to improve scene labeling performance. In addition, an unsupervised joint feature learning and encoding model is proposed for scene labeling in [26]. However, due to the limited number of RGB-D images, deep learning for scene labeling in RGB-D images was not as appealing as that for RGB images. The release of the SUNRGBD dataset, which includes most of the previously popular datasets, may have changed this situation [13,14].

Another main challenge imposed by scene labeling in RGB-D images is the fusion of contextual representations of different sources (i.e., depth and photometric data). For instance, in [13,14], two independent CNNs are leveraged to extract features from the depth and photometric data separately, which are then simply concatenated for class prediction. Ignoring the strong correlation between depth and photometric channels usually negatively affects semantic labeling. In contrast, instead of simply concatenating features from multiple sources, the memorized fusion layer in our model facilitates the integration of contextual information from different sources in a data-driven manner,

RNN for Image Processing: Recurrent neural networks (RNNs) represent a type of neural networks with loop connections [28]. They are designed to capture dependencies across a distance larger than the extent of local neighborhoods. In previous work, RNN models have not been widely used partially due to the difficulty to train such models, especially for sequential data with long-range dependencies [29]. Fortunately, RNNs with gate and memory structures, e.g., long short-term memory (LSTM) [30], can artificially learn to remember and forget information by using specific gates to control the information flow. Although RNNs have an outstanding capability to capture short-range and long-range dependencies, there exist problems for applying RNNs to image processing due to the fact that, unlike data in natural language processing (NLP) tasks, images do not have a natural sequential structure. Thus, different strategies have been proposed to overcome this problem. Specifically, in [19], cascaded bi-directional vertical and horizonal RNN layers are designed for modeling 2D dependencies in images. A multi-dimensional RNN with LSTM unit has been applied to handwriting [31]. A parallel multi-dimensional LSTM for image segmentation has been proposed in [32]. In this paper, we propose an LSTM-CF model consisting of memorized context layers and a memorized fusion layer to capture image contexts from a global perspective and fuse contextual representations from different sources.

3 LSTM-CF Model

As illustrated in Fig. 2, our end-to-end LSTM-CF model for RGB-D scene labeling consists of four components, layers for vertical depth context extraction, layers for vertical photometric context extraction, a memorized fusion layer for incorporating vertical photometric and depth contexts as true 2D global contexts, and a final layer for pixel-wise scene labeling given concatenated convolutional features and global contexts. The inputs to our model include both photometric and depth images. The path for extracting global contexts from the photometric image consists of multiple convolutional layers and an extra memorized context layer. On the other hand, the depth image is first encoded as an HHA image, which is fed into three convolutional layers [14] and an extra memorized context layer for global depth context extraction. The other component, a memorized fusion layer, is responsible for fusing previously extracted global RGB and depth contexts in a data-driven manner. On top of the memorized fusion layer, the final convolutional feature of photometric channels and the fused global context are concatenated together and fed into the final fully convolutional layer, which performs pixel-wise scene labeling with the softmax activation function.

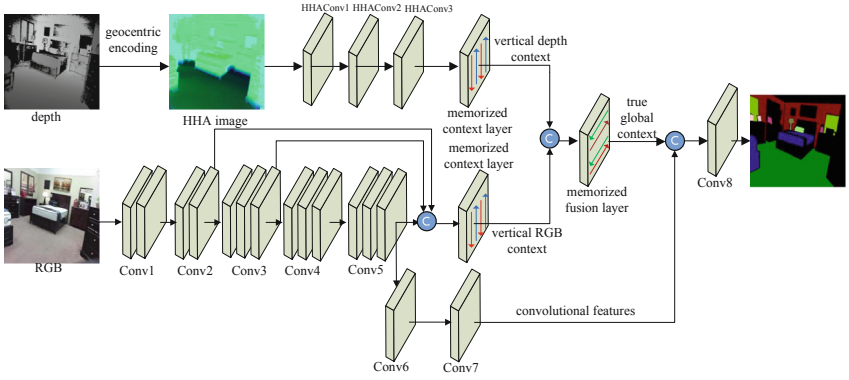


Fig. 2. Our LSTM-CF model for RGB-D scene labeling. The input consists of both photometric and depth channels. Vertical contexts in photometric and depth channels are computed in parallel using cascaded convolutional layers and a memorized context layer. Vertical photometric (color) and depth contexts are fused and bi-directionally propagated along the horizontal direction via another memorized fusion layer to obtain true 2D global contexts. The fused global contexts and the final convolutional features of photometric channels are then concatenated together and fed into the final convolutional layer for pixel-wise scene labeling. "C" stands for the concatenation operation. (Color figure online)

3.1 Memorized Vertical Depth Context

Given a depth image, we use the HHA representation proposed in [13] to encode geometric properties of the depth image in three channels, i.e., disparity, surface

normal and height. Different from [13], the encoded HHA image in our pipeline is fed into three randomly initialized convolutional layers (to obtain a feature map with the same resolution as that in the RGB path) instead of layers taken from the model pre-trained on the ILSVRC2012 dataset. This is because the color distribution of HHA images is different from that of natural images (see Fig. 2) according to [20]. One top of the third convolutional layer (i.e., HHAConv3), there is an extra memorized context layer from Renet [19], which performs bi-directional propagation of local contextual features from the convolutional layers along the vertical direction. For better understanding, we denote the feature map HHAConv3 as $F = \{f_{i,j}\}$, where $F \in \mathbb{R}^{w \times h \times c}$ with w, h and c representing the width, height and the number of channels. Since we perform pixel-wise scene labeling, every patch in this Renet layer only contains a single pixel. Thus, vertical memorized context layer (here we choose LSTM as recurrent unit) can be formulated as

$$h^f_{i,j} = \text{LSTM}(h^f_{i,j-1}, f_{i,j}), \quad \text{for} \quad j = 1, \ldots, h \tag{1}$$

$$h^b_{i,j} = \text{LSTM}(h^b_{i,j+1}, f_{i,j}), \quad \text{for} \quad j = h, \ldots, 1, \tag{2}$$

where h^f and h^b stand for the hidden states of the forward and backward LSTM. In the forward LSTM, the unit at pixel (i, j) takes $h^f_{i,j-1} \in \mathbb{R}^d$ and $f_{i,j} \in \mathbb{R}^c$ as input, and its output is calculated as follows according to [30]. The operations in the backward LSTM can be defined similarly.

$$\begin{aligned}
\text{gate}_i &= \delta(W_{if} f_{i,j} + W_{ih} h^f_{i,j-1} + b_i) \\
\text{gate}_f &= \delta(W_{ff} f_{i,j} + W_{fh} h^f_{i,j-1} + b_f) \\
\text{gate}_o &= \delta(W_{of} f_{i,j} + W_{oh} h^f_{i,j-1} + b_o) \\
\text{gate}_c &= \tanh(W_{cf} f_{i,j} + W_{ch} h^f_{i,j-1} + b_c) \\
c_{i,j} &= \text{gate}_f \odot c_{i,j-1} + \text{gate}_i \odot \text{gate}_c \\
h^f_{i,j} &= \tanh(\text{gate}_o \odot c_{i,j})
\end{aligned} \tag{3}$$

Finally, pixel-wise vertical depth contexts are collectively represented as a map, $C_{\text{depth}} \in \mathbb{R}^{w \times h \times 2d}$, where $2d$ is the total number of output channels from the vertical memorized context layer.

3.2 Memorized Vertical Photometric Context

In the component for extracting global RGB contexts, we adapt the Deeplab architecture proposed in [12]. Different from existing Deeplab variants, we concatenate features at three different scales to enrich the feature representation. This is inspired by the network architecture in [33]. Specifically, since there exists hole operations in Deeplab convolutional layers, feature maps from Conv2_2, Conv3_3 and Conv5_3 have sufficient initial resolutions. They can be further elevated to the same resolution using interpolation. Corresponding pixel-wise features from these three elevated feature maps are then concatenated together

before being fed into the subsequent memorized fusion layer, which again performs bi-directional propagation to produce vertical photometric contexts. Here pixel-wise vertical photometric contexts can also be represented as a map, $C_{\text{RGB}} \in \mathbb{R}^{w \times h \times 2d}$, which has the same dimensionalities as the map for vertical depth contexts.

3.3 Memorized Context Fusion

So far vertical depth and photometric contexts are computed independently in parallel. Instead of simply concatenating these two types of contexts, the memorized fusion layer, which performs horizontal bi-directional propagation from Renet, is exploited for adaptively fusing vertical depth and RGB contexts in a data-driven manner, and the output from this layer can be regarded as the fused representation of both types of contexts. Such fusion can generate more discriminative features through end-to-end training. The input and output dimensions of the fusion layer are set to $\mathbb{R}^{w \times h \times 4d}$ and $\mathbb{R}^{w \times h \times 2d}$, respectively.

Note that there are two separate memorized context layers in the photometric and depth paths of our architecture. Since the memorized context layer and the memorized fusion layer are two symmetric components of the original Renet [19], a more natural and symmetric alternative would have a single memorized context layer preceding the memorized fusion layer in our model (i.e., whole structure of Renet including cascaded bi-directional vertical and horizonal memorized layer) and let the memorized fusion layer incorporate the features from the RGB and depth paths. Nonetheless, in our experiments, this alternative network architecture gave rise to slightly worse performance.

3.4 Scene Labeling

Between photometric and depth images, photometric images contain more details and semantic information that can help scene labeling in comparison with sparse and discontinuous depth images [14]. Nonetheless, depth images can provide auxiliary geometric information for improving scene labeling performance. Thus, we design a cross-layer combination that integrates pixel-wise convolutional features (i.e., Conv7 in Fig. 2) from the photometric image with fused global contexts from the memorized fusion layer as the final pixel-wise features, which are fed into the last fully convolutional layer with softmax activation to perform scene labeling at every pixel location.

4 Experimental Results

4.1 Experimental Setting

Datasets: We evaluate our proposed model for RGB-D scene labeling on three public benchmarks, SUNRGBD, NYUDv2 and SUN3D. SUNRGBD [20] is the largest dataset currently available, consisting of 10355 RGB-D images captured

from four different depth sensors. It includes most previous datasets, such as NYUDv2 depth [34], Berkeley B3DO [35], and SUN3D [36], as well as 3943 newly captured RGB-D images [20]. 5285 of these images are predefined for training and the remaining 5050 images constitute the testing set [14].

Implementation Details: In our experiments, a slightly modified Deeplab pipeline [12] is adopted as the basic network in our RGB path for extracting convolutional feature maps because of its high performance. It is initialized with the publicly available VGG-16 model pre-trained on ImageNet. For the purpose of pixel-wise scene labeling, this architecture transforms the last two fully connected layers in the standard VGG-16 to convolutional layers with 1×1 kernels. For the parallel depth path, three randomly initialized CNN layers with max pooling are leveraged for depth feature extraction. In each path, on top of the aforementioned convolutional network, a vertically bi-directional LSTM layer implements the memorized context layer, and models both short-range and long-range spatial dependencies. Then, another horizontally bi-directional LSTM layer implements the memorized fusion layer, and is used to adaptively integrate the global contexts from the two paths. In addition, there is a cross-layer combination of final convolutional features (i.e., Conv7) and the integrated global representation from the horizontal LSTM layer.

Since the SUNRGBD dataset was collected by four different depth sensors, each input image is cropped to 426×426 (the smallest resolution of these four sensors) [14]. During fine-tuning, the learning rate for newly added layers, including HHAConv1, HHAConv2, HHAConv3, the memorized context layers, the memorized fusion layer and Conv8, is initialized to 10^{-2}, and the learning rate for those pre-trained layers of VGG-16 is initialized to 10^{-4}. All weights in the newly added convolutional layers are initialized using a Gaussian distribution with a standard deviation equal to 0.01, and the weights in the LSTM layers are randomly initialized with a uniform distribution over $[-0.01, 0.01]$. The number of hidden memory cells in a memorized context layer or a memorized fusion layer is set to 100, and the size of feature maps is 54×54. We train all the layers in our deep network simultaneously using SGD with a momentum 0.9, the batch size is set to one (due to limited GPU memory) and the weight decay is 0.0005. The entire deep network is implemented on the publicly available platform Caffe [37] and is trained on a single NVIDIA GeForce GTX TITAN X GPU with 12GB memory [1]. It takes about 1 day to train our deep network. In the testing stage, an RGB-D image takes 0.15s on average, which is significantly faster than pervious methods, i.e., the testing time in [9,24] is around 1.5 s.

4.2 Results and Comparisons

According to [14,22], performance is evaluated by comparing class-wise Jaccard Index, i.e., n_{ii}/t_i, and average Jaccard Index, i.e., $(1/n_{cl}) \sum_i n_{ii}/t_i$, where n_{ij} is the number of pixels annotated as class i and predicted to be class j, n_{cl} is

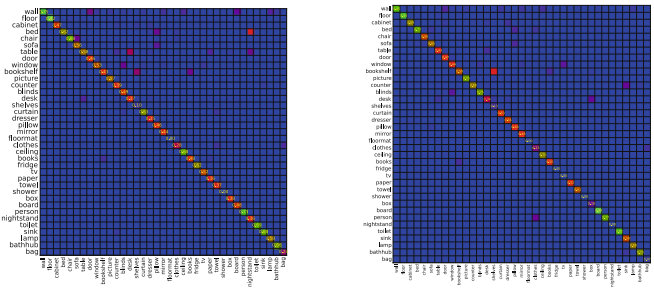
[1] LSTM-CF model is publicly available at: https://github.com/icemansina/LSTM-CF.

the number of different classes, and $t_i = \sum_j n_{ij}$ is the total number of pixels annotated as class i [10].

SUNRGBD Dataset [20]**:** The performance and comparison results on SUNRGBD are shown in Table 1. Our proposed architecture can outperform existing techniques: 2.2 % higher than the performance reported in [22], 11.8 % higher than that in [24], 38 % higher than that in [38] and 39.1 % higher than that in [20] in terms of 37-class average Jaccard Index. Improvements can be observed in 15 class-wise Jaccard Indices. For a better understanding, we also show the confusion matrix for this dataset in Fig. 3(a).

Table 1. Comparison of scene labeling results on SUNRGBD using class-wise and average Jaccard Index. We compare our model with results reported in [20,24,38] and previous state-of-the-art result in [22]. Boldface numbers mean best performance.

	Wall	floor	cabinet	bed	chair	sofa	table	door	window	bookshelf	picture	counter	blinds	desk	shelves	curtain	dresser	pillow	mirror
[20]	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
[20]	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
[20]	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
[38]	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
[38]	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
[38]	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
[24]	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
[22]	**80.2**	**90.9**	**58.9**	**64.8**	**76.0**	**58.6**	**62.6**	**47.7**	**66.4**	31.2	**63.6**	33.8	46.7	19.7	16.2	**67.0**	**42.3**	**57.1**	39.1
Ours	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**
	floormat	clothes	ceiling	books	fridge	tv	paper	towel	shower	box	board	person	nightstand	toilet	sink	lamp	bathhub	bag	mean
[20]	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
[20]	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
[20]	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
[38]	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
[38]	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
[38]	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
[24]	**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
[22]	0.1	24.4	84.0	**48.7**	21.3	49.5	30.6	18.8	0.1	24.1	**56.8**	17.9	**42.9**	**73.0**	66.2	48.8	45.1	**24.1**	45.9
Ours	0.0	**28.4**	68.0	47.9	**61.5**	**52.1**	**36.4**	**36.7**	0	**38.1**	48.1	**72.6**	36.4	68.8	**67.9**	**58.0**	**65.6**	23.6	**48.1**



(a) SUNRGBD (b) NYUDv2

Fig. 3. Confusion matrix for SUNRGBD and NYUDv2. Class-wise Jaccard Index is shown on the diagonal. Best viewed in color. (Color figure online)

It is worth mentioning that our proposed architecture and most previous methods achieve zero accuracy on two categories, i.e., floormat and shower, which mainly results from an imbalanced data distribution instead of the capacity of our model.

NYUDv2 Dataset: To further verify the effectiveness of our architecture and have more comparisons with existing state-of-the-art methods, we also conduct experiments on the NYUDv2 dataset. The results are presented in Table 2, where the 13-class average Jaccard Index of our model is 20.3 % higher than that in [39]. Class frequencies and the confusion matrix are also shown in Table 2 and Fig. 3(b) respectively. According to the reported results, our proposed architecture gains 5.6 % and 5.5 % improvement in average Jaccard Index over [9] and FCN-32s [10] respectively. Considering the listed class frequencies, our proposed model significantly outperforms existing methods on high frequency categories and most low frequency categories, which primarily owes to the convolutional features of the RGB image and the fused global contexts of the complete RGB-D image. In terms of labeling categories with small and complex regions, e.g., pillows and chairs, our method also achieves a large improvement, which can be verified in the following visual comparisons.

Table 2. Comparison of scene labeling on NYUDv2. We compare our proposed model with existing state-of-the-art methods, i.e., [9,24–26,34]. Class-wise Jaccard Index and average Jaccard Index of 37 classes are presented. 'Freq' stands for class frequency. Boldface numbers mean best performance.

	Wall	floor	cabinet	bed	chair	sofa	table	door	window	bookshelf	picture	counter	blinds	desk	shelves	curtain	dresser	pillow	mirror
Freq	21.4	9.1	6.2	3.8	3.3	2.7	2.1	2.2	2.1	1.9	2.1	1.4	1.7	1.1	1.0	1.1	0.9	0.8	1.0
[34]	60.7	77.8	33.0	40.3	32.4	25.3	21.0	5.9	29.7	22.7	35.7	33.1	40.6	4.7	3.3	27.4	13.3	18.9	4.4
[24]	60.0	74.4	37.1	42.3	32.5	28.2	16.6	12.9	27.7	17.3	32.4	38.6	26.5	10.1	6.1	27.6	7.0	19.7	17.9
[25]	67.4	80.5	41.4	56.4	40.4	44.8	30.0	12.1	34.1	20.5	38.7	50.7	44.7	10.1	1.6	26.3	21.6	31.3	14.6
[26]	61.4	66.4	38.2	43.9	34.4	33.8	22.6	8.3	27.6	17.6	27.7	30.2	33.6	5.1	2.7	18.9	16.8	12.5	10.7
[9]	65.7	62.5	40.1	32.1	44.5	50.8	43.5	**51.6**	**49.2**	36.3	41.4	39.2	55.8	**48.0**	**45.2**	53.1	55.3	**50.5**	46.1
Ours	**79.6**	**83.5**	**69.3**	**77.0**	**58.3**	**64.9**	**42.6**	47.0	43.6	**59.5**	**74.5**	**68.2**	**74.6**	33.6	13.1	**53.2**	**56.5**	48.0	**47.7**
	floormat	clothes	ceiling	books	fridge	tv	paper	towel	shower	box	board	person	nightstand	toilet	sink	lamp	bathhub	bag	mean
Freq	0.7	0.7	1.4	0.6	0.6	0.5	0.4	0.4	0.4	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.2	
[34]	7.1	6.5	73.2	5.5	1.4	5.7	12.7	0.1	3.6	0.1	0.0	6.6	6.3	26.7	25.1	15.9	0.0	0.0	17.5
[24]	20.1	9.5	53.9	14.8	1.9	18.6	11.7	12.6	5.4	3.3	0.2	13.6	9.2	35.2	28.9	14.2	7.8	1.2	20.2
[25]	28.2	8.0	61.8	5.8	14.5	14.4	14.1	19.8	6.0	1.1	12.9	1.5	15.7	52.5	47.9	31.2	29.4	0.2	30.0
[26]	13.8	2.7	46.1	3.6	2.9	3.2	2.6	6.2	6.1	0.8	28.2	5	6.9	32	20.9	5.4	16.2	0.2	29.2
[9]	**54.1**	**35.4**	50.6	39.1	**53.6**	**50.1**	35.4	39.9	**41.8**	**36.3**	60.6	35.6	**32.5**	31.8	22.5	26.3	38.5	**37.3**	43.9
Ours	0.0	22.7	**70.2**	**49.7**	0.0	0.0	**52.1**	**60.6**	0	17.6	**93.9**	**77.0**	0	**81.8**	**58.4**	**67.6**	**72.6**	7.5	**49.4**

SUN3D Dataset: Table 3 gives comparison results on the 1539 test images in the SUN3D dataset. For fair comparison, the 12-class average Jaccard Index is used in the comparison with the state-of-the-art results recently reported in [9]. Note that the 12-class accuracy of our network is calculated through the model previously trained for 37 classes. Our model substantially outperforms the one from [9] on large planar regions such as those labeled as floors and ceilings. This also results from the incorporated convolutional features and the fused global contexts.

Table 3. Comparison of class-wise Jaccard Index and 12-class average Jaccard Index on SUN3D.

	Wall	Floor	Bed	Chair	Table	Counter	Curtain	Ceiling	Tv	Toilet	Bathtub	Bag	Mean
[9]	**73**	35	**71**	35	30	**52**	68	27	56	23	49	**29**	45.7
Ours	**73**	**86**	32	**65**	**57**	22	**76**	**69**	**75**	**62**	**62**	23	**58.5**

These comparison results further confirm the power and generalization capability of our LSTM-based model.

4.3 Ablation Study

To discover the vital elements in our proposed model, we conduct an ablation study to remove or replace individual components in our deep network when training and testing on the SUNRGBD dataset. Specifically, we have tested the performance of our model without the RGB path, the depth path, multi-scale RGB feature concatenation, the memorized context layers or the memorized fusion layer. In addition, we also conduct an experiment with a model that does not combine the final convolutional features of photometric channels (i.e., Conv7 in Fig. 2) with the global contexts of the complete RGB-D image to figure out the importance of different components. The results are presented in Table 4. From the given results, we find that the final convolutional features of the photometric channels is the most vital information, i.e., the cross-layer combination is the most effective component as the performance drops to 15.2 % without it, which is consistent with previously mentioned properties of depth and photometric data. In addition, multi-scale RGB feature concatenation before the memorized context layer also plays a vital role as it directly affects the vertical contexts in the photometric channels and the performance drops to 42.1 % without it. It is obvious that performance would be inevitably harmed without the depth path. Among the memorized layers, the memorized fusion layer is more important than the memorized context layers in our pipeline as it accomplishes the fusion of contexts in photometric and depth channels.

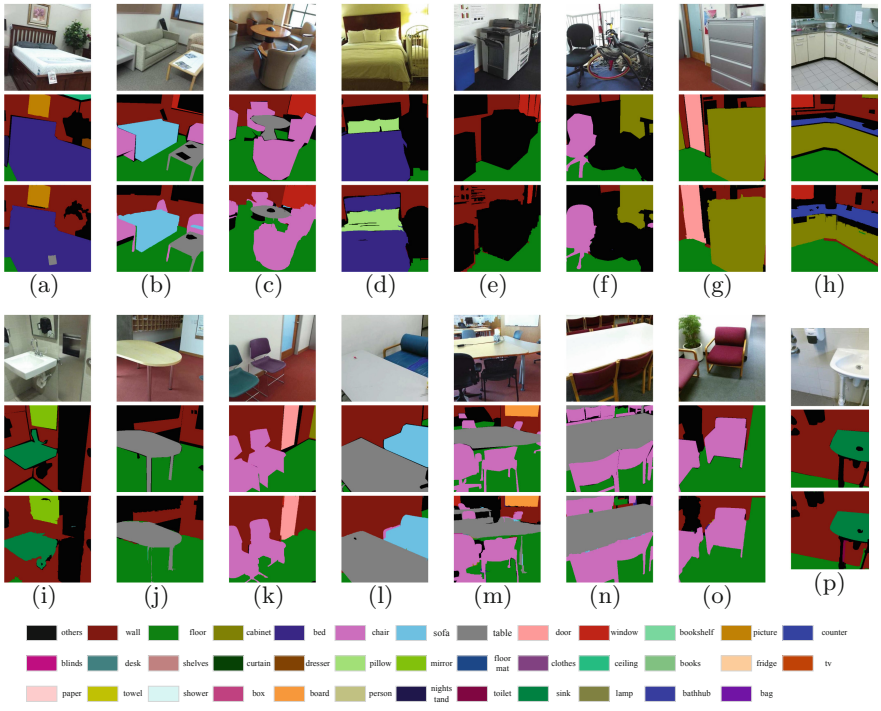
Table 4. Ablation study

Model	Mean accuracy
Without RGB path, using Deeplab+Renet for depth path	15.8 %
Without depth path	43.7 %
Without multi-scale RGB feature concatenation	42.1 %
Without cross-layer integration of RGB convolutional features	15.2 %
Without memorized fusion layer	44.7 %
Without memorized context layers	45.7 %
Without any memorized (context or fusion) layers	45.0 %

4.4 Visual Comparisons

SUNRGBD Dataset: We present visual results of RGB-D scene labeling in Fig. 4. Here, we leverage super-pixel based averaging to smooth visual labeling results as being done in [9]. The algorithm in [40] is used for performing super-pixel segmentation. As can be observed in Fig. 4, our proposed deep network

produces accurate and semantically meaningful labeling results, especially for large regions and high frequency labels. For instance, our model takes advantage of global contexts when labeling 'bed' in Fig. 4(a), 'wall' in Fig. 4(e) and 'mirror' in Fig. 4(i). Our proposed model can precisely label almost all 'chairs' (a high frequency label) by exploiting integrated photometric and depth information, regardless of occlusions.



(q) legend of semantic labels

Fig. 4. Examples of semantic labeling results on the SUNRGBD dataset. The top row shows the input RGB images, the bottom row shows scene labeling obtained with our model and the middle row has the ground truth. Semantic labels and their corresponding colors are shown at the bottom.

NYUDv2 Dataset: We also perform visual comparisons on the NYUDv2 benchmark, which has complicated indoor scenes and well-labeled ground truth. We compare our scene labeling results with those publicly released labeling results from [25]. It is obvious that our results are clearly better than those from [25] both visually and numerically (under the metric of average Jaccard Index) even though scene labeling in [25] is based on sophisticated segmentation.

Label Refinement: Surprisingly, our model can intelligently refine certain region annotations, which might have inaccuracies due to under-segmentation,

especially in the newly captured 3943 RGB-D images, as shown in Fig. 6. Specifically, the cabinets in Fig. 6(a) were annotated as 'background', the pillows in Fig. 6(g) as 'bed', and the tables in Fig. 6(n) as 'wall' by mistake. Our model can effectively deal with these difficult regions. For example, the annotation of

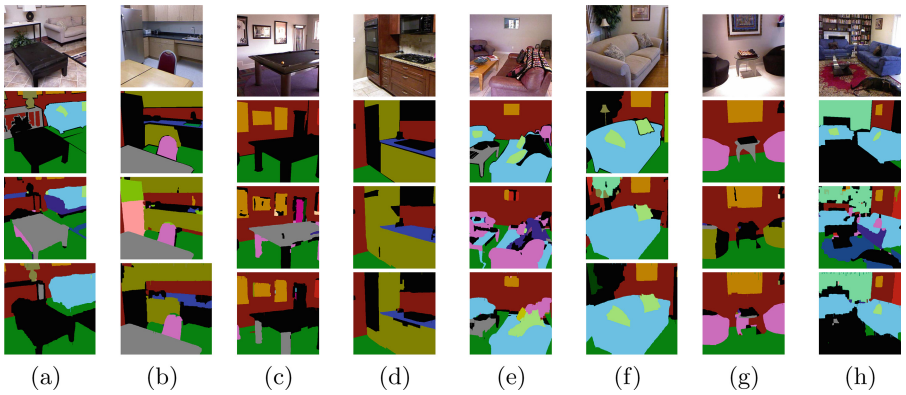


Fig. 5. Visual comparison of scene labeling results on the NYUDv2 dataset. The first and second rows show the input RGB images and their corresponding ground truth labeling. The third row shows the results from [25] and the last row shows the results from our model.

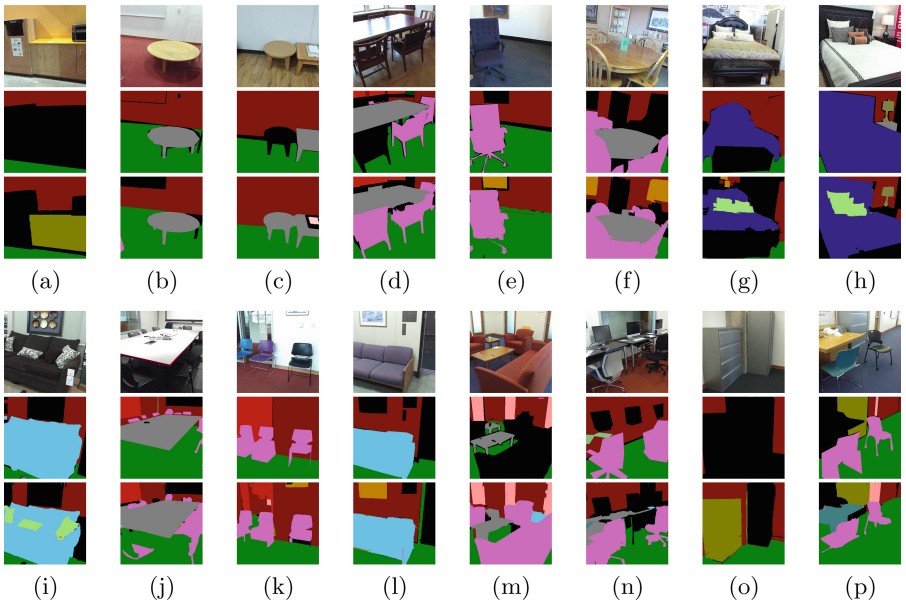


Fig. 6. Annotation refinement on the SUNRGBD dataset. The top row shows the input RGB images, the middle row shows the original annotations, and the bottom row shows scene labeling results from our model.

the picture in Fig. 6(e) and that of the pillows in Fig. 6(g) have been corrected. Thus, our model can be exploited to refine certain annotations in the SUNRGBD dataset, which is another contribution of our model.

5 Conclusions

In this paper, we have developed a novel Long Short-Term Memorized Context Fusion (LSTM-CF) model that captures image contexts from a global perspective and deeply fuses contextual representations from multiple sources (i.e., depth and photometric data) for semantic scene labeling. In future, we will explore how to extend the memorized layers with an attention mechanism, and refine the performance of our model in boundary labeling.

References

1. Wu, C., Lenz, I., Saxena, A.: Hierarchical semantic labeling for task-relevant RGB-D perception. In: Robotics: Science and Systems (RSS) (2014)
2. Hinterstoisser, S., Lepetit, V., Ilic, S., Holzer, S., Bradski, G., Konolige, K., Navab, N.: Model based training, detection and pose estimation of texture-less 3D objects in heavily cluttered scenes. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) ACCV 2012, Part I. LNCS, vol. 7724, pp. 548–562. Springer, Heidelberg (2013)
3. Holz, D., Holzer, S., Rusu, R.B., Behnke, S.: Real-time plane segmentation using RGB-D cameras. In: Röfer, T., Mayer, N.M., Savage, J., Saranlı, U. (eds.) RoboCup 2011. LNCS, vol. 7416, pp. 306–317. Springer, Heidelberg (2012)
4. Schuster, S., Krishna, R., Chang, A., Fei-Fei, L., Manning, C.D.: Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In: Proceedings of the Fourth Workshop on Vision and Language, pp. 70–80 (2015)
5. Yan, Z., Zhang, H., Wang, B., Paris, S., Yu, Y.: Automatic photo adjustment using deep neural networks. ACM Trans. Graph. **35**(2), 11 (2016)
6. Farabet, C., Couprie, C., Najman, L., LeCun, Y.: Learning hierarchical features for scene labeling. IEEE Trans. Pattern Anal. Mach. Intell. **35**(8), 1915–1929 (2013)
7. Gould, S., Fulton, R., Koller, D.: Decomposing a scene into geometric and semantically consistent regions. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 1–8. IEEE (2009)
8. Tighe, J., Lazebnik, S.: SuperParsing: scalable nonparametric image parsing with superpixels. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part V. LNCS, vol. 6315, pp. 352–365. Springer, Heidelberg (2010)
9. Khan, S.H., Bennamoun, M., Sohel, F., Togneri, R., Naseem, I.: Integrating geometrical context for semantic labeling of indoor scenes using RGBD images. Int. J. Comput. Vis. **117**, 1–20 (2015)
10. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
11. Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., Torr, P.H.: Conditional random fields as recurrent neural networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1529–1537 (2015)

12. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected CRFs. arXiv preprint arXiv:1412.7062 (2014)
13. Gupta, S., Girshick, R., Arbeláez, P., Malik, J.: Learning rich features from RGB-D images for object detection and segmentation. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part VII. LNCS, vol. 8695, pp. 345–360. Springer, Heidelberg (2014)
14. Song, S., Xiao, J.: Deep sliding shapes for amodal 3D object detection in RGB-D images. arXiv preprint arXiv:1511.02300 (2015)
15. Liu, Z., Li, X., Luo, P., Loy, C.C., Tang, X.: Semantic image segmentation via deep parsing network. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1377–1385 (2015)
16. Liang, X., Shen, X., Xiang, D., Feng, J., Lin, L., Yan, S.: Semantic object parsing with local-global long short-term memory. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)
17. Byeon, W., Breuel, T.M., Raue, F., Liwicki, M.: Scene labeling with LSTM recurrent neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3547–3555 (2015)
18. Pinheiro, P., Collobert, R.: Recurrent convolutional neural networks for scene labeling. In: Proceedings of the 31st International Conference on Machine Learning (ICML 2014), pp. 82–90 (2014)
19. Visin, F., Kastner, K., Cho, K., Matteucci, M., Courville, A., Bengio, Y.: Renet: a recurrent neural network based alternative to convolutional networks. arXiv preprint arXiv:1505.00393 (2015)
20. Song, S., Lichtenberg, S.P., Xiao, J.: Sun RGB-D: a RGB-D scene understanding benchmark suite. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 567–576 (2015)
21. Kumar, M.P., Koller, D.: Efficiently selecting regions for scene understanding. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3217–3224. IEEE (2010)
22. Kendall, A., Badrinarayanan, V., Cipolla, R.: Bayesian segnet: model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. arXiv preprint arXiv:1511.02680 (2015)
23. Lempitsky, V., Vedaldi, A., Zisserman, A.: Pylon model for semantic segmentation. In: Advances in Neural Information Processing Systems, pp. 1485–1493 (2011)
24. Ren, X., Bo, L., Fox, D.: RGB-(D) scene labeling: features and algorithms. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2759–2766. IEEE (2012)
25. Gupta, S., Arbeláez, P., Girshick, R., Malik, J.: Indoor scene understanding with RGB-D images: bottom-up segmentation, object detection and semantic segmentation. Int. J. Comput. Vis. **112**(2), 133–149 (2015)
26. Wang, A., Lu, J., Cai, J., Wang, G., Cham, T.J.: Unsupervised joint feature learning and encoding for RGB-D scene labeling. IEEE Trans. Image Process. **24**(11), 4459–4473 (2015)
27. Husain, F., Schulz, H., Dellen, B., Torras, C., Behnke, S.: Combining semantic and geometric features for object class segmentation of indoor scenes. IEEE Rob. Autom. Lett. **2**(1), 49–55 (2017)
28. Schmidhuber, J.: A local learning algorithm for dynamic feedforward and recurrent networks. Connection Sci. **1**(4), 403–412 (1989)
29. Bengio, Y., Simard, P., Frasconi, P.: Learning long-term dependencies with gradient descent is difficult. IEEE Trans. Neural Netw. **5**(2), 157–166 (1994)

30. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
31. Graves, A., Schmidhuber, J.: Offline handwriting recognition with multidimensional recurrent neural networks. In: Advances in Neural Information Processing Systems, pp. 545–552 (2009)
32. Stollenga, M.F., Byeon, W., Liwicki, M., Schmidhuber, J.: Parallel multi-dimensional LSTM, with application to fast biomedical volumetric image segmentation. In: Advances in Neural Information Processing Systems, pp. 2980–2988 (2015)
33. Li, G., Yu, Y.: Deep contrast learning for salient object detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
34. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from RGBD images. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part V. LNCS, vol. 7576, pp. 746–760. Springer, Heidelberg (2012)
35. Janoch, A., Karayev, S., Jia, Y., Barron, J.T., Fritz, M., Saenko, K., Darrell, T.: A category-level 3D object dataset: putting the kinect to work. In: Fossati, A., Gall, J., Grabner, H., Ren, X., Konolige, K. (eds.) Consumer Depth Cameras for Computer Vision, pp. 141–165. Springer, Heidelberg (2013)
36. Xiao, J., Owens, A., Torralba, A.: SUN3D: a database of big spaces reconstructed using SfM and object labels. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1625–1632 (2013)
37. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093 (2014)
38. Liu, C., Yuen, J., Torralba, A.: Sift flow: dense correspondence across scenes and its applications. IEEE Trans. Pattern Anal. Mach. Intell. **33**(5), 978–994 (2011)
39. Couprie, C., Farabet, C., Najman, L., LeCun, Y.: Toward real-time indoor semantic segmentation using depth information. J. Mach. Learn. Res. (2014)
40. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. Int. J. Comput. Vis. **59**(2), 167–181 (2004)