Cross-Modal Attentional Context Learning for RGB-D Object Detection

Guanbin Li, *Member, IEEE*, Yukang Gan, Hejun Wu, *Member, IEEE*, Nong Xiao, *Member, IEEE*, and Liang Lin, *Senior Member, IEEE*

***Abstract*— Recognizing objects from simultaneously sensed photometric (RGB) and depth channels is a fundamental yet practical problem in many machine vision applications, such as robot grasping and autonomous driving. In this paper, we address this problem by developing a cross-modal attentional context (CMAC) learning framework, which enables the full exploitation of the context information from both RGB and depth data. Compared to existing RGB-D object detection frameworks, our approach has several appealing properties. First, it consists of an attention-based global context model for exploiting adaptive contextual information and incorporating this information into a region-based CNN (e.g., fast RCNN) framework to achieve improved object detection performance. Second, our CMAC framework further contains a fine-grained object part attention module to harness multiple discriminative object parts inside each possible object region for superior local feature representation. While greatly improving the accuracy of RGB-D object detection, the effective cross-modal information fusion as well as attentional context modeling in our proposed model provide an interpretable visualization scheme. Experimental results demonstrate that the proposed method significantly improves upon the state of the art on all public benchmarks.**

***Index Terms*— RGB-D object detection, attentional context modeling, cross modal feature, convolutional neural network.**

I. Introduction

RGB-D object detection attempts to localize and classify objects within an image with depth information. It is one of the core technologies in the field of robotics application and can be beneficial to many intelligent tasks, including pose estimation [1], [2], content-based image retrieval [3] and robot task planning [4]. In recent years, the successful application of deep convolutional neural networks has pushed this research into a new phase and achieved very good results.

Manuscript received August 11, 2017; revised March 5, 2018 and August 18, 2018; accepted October 22, 2018. Date of publication October 31, 2018; date of current version November 28, 2018. This work was supported in part by the State Key Development Program under Grant 2018YFC0830103, in part by the National Natural Science Foundation of China under Grant 61702565 and Grant 61672552, in part by the Science and Technology Planning Project of Guangdong Province under Grant 2017B010116001, in part by the Fundamental Research Funds for the Central Universities under Grant 18lgpy63, in part by GD-NSF under Grant 2017A030312006, and in part by the SenseTime Research Fund. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Shuicheng Yan. *(Guanbin Li and Yukang Gan contributed equally to this work.) (Corresponding author: Liang Lin.)*

The authors are with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510006, China (e-mail: linliang@ieee.org).

Digital Object Identifier 10.1109/TIP.2018.2878956

Most CNN-based RGB-D object detection frameworks are extended from RCNN-based object detectors [5]–[7] for RGB images. R-CNN-Depth [8] is the first deep learning framework for RGB-D object detection that extends the R-CNN system [5] to take advantage of depth information by incorporating two parallel network streams for both RGB and depth modalities. This two-stream pipeline later became the basis for many visual perception tasks in RGB-D images [9]–[12]. In this framework, the features from the RGB and depth modalities are computed independently and concatenated after applying fully connected layers for final proposal classification. However, this pipeline has its own limitations: (1) Independent feature computation and simple feature concatenation ignore the correlation between the two modalities. (2) Only information inside the object proposal is used for object classification, which neglects the auxiliary role of context information outside the bounding box in object classification.

In this paper, we propose a Cross Modal Attentional Context (CMAC) learning framework for RGB-D object detection that incorporates the consistency and complementary information between two diverse modalities (RGB and depth), as well as an attentional model for global context mining and discriminative object part discovery. To exploit the correlation between RGB and depth modalities, the CMAC model employs a cross-modal feature fusion component to fuse the features extracted from the output feature maps of the two fully convolutional networks (with different input sources). Instead of directly applying fused features to classification and object location refinement, our proposed CMAC model further learns attentional context and explores discriminative object parts based on the fused features. We believe that both the attentional global context and the discriminative parts attended inside each possible object region (object proposal) are crucial for accurate RGB-D object detection.

To capture the global context, our model employs a recurrent attention model that consists of multiple stacked Long Short-Term Memory (LSTM) units. The recurrent neural network is optimized to infer relevant regions for each given region proposal. As shown in Figure 1, the regions that are considered helpful for classification of the object proposal are highlighted. As can be seen, our proposed CMAC model can identify an adaptive global context for different object proposals (*i.e.,* the regions of the keyboard, parts of the table around the target monitor as well as the other monitor are highlighted when the input region proposal contains a monitor.

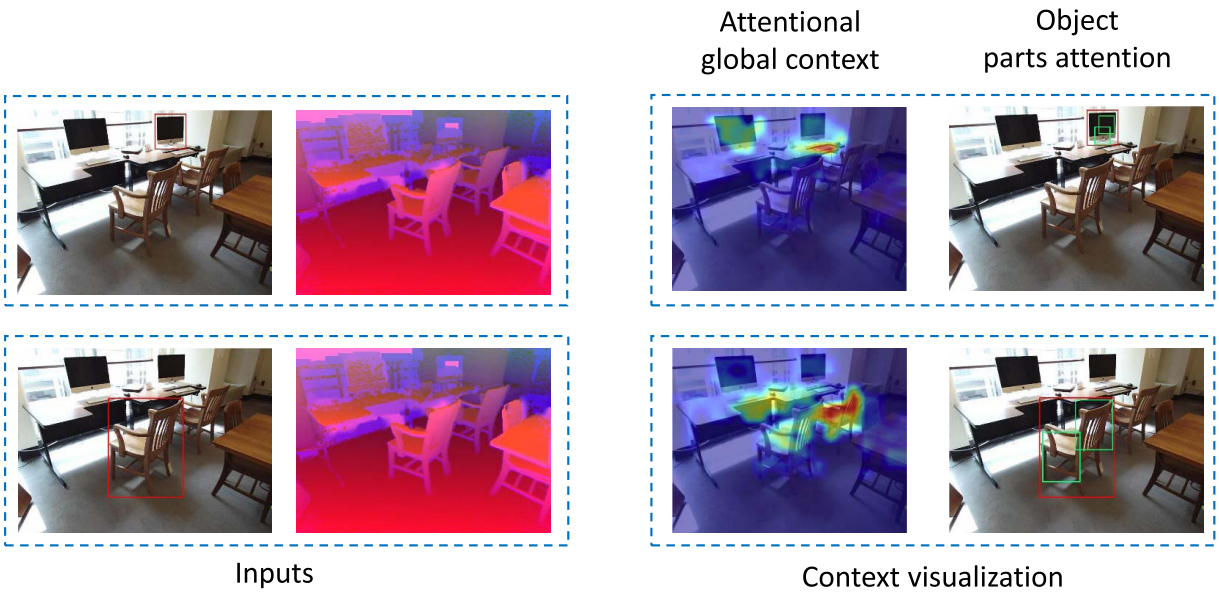


Fig. 1. Example visualization results for global context and object part attention generated by our proposed CMAC model. For global context, information from relevant regions (the highlighted regions) of the object proposals is obtained through a recurrent attentional model. For local context, multiple parallel spatial transformers are utilized to exploit information from the discriminative parts (green rectangles) of the object proposals. Red rectangles indicate the object proposals.

When the input region proposal contains a chair, the regions including parts of the table and other chairs are assigned higher weights in the final classification.). Moreover, inspired by the fact that humans tend to quickly capture distinguishable parts for more accurate object classification judgment when observing objects with occluded regions, we propose to further incorporate a fine-grained object part attention module in our network framework. Considering the flexible attention mechanism and the excellent spatial manipulation ability of Spatial Transform Networks (STNs), we adopt multiple STNs in parallel to examine the discriminative parts located inside a specific object proposal for capturing local context. As illustrated in Figure 1, the CMAC model is able to successfully locate the most discriminative location that can differentiate an object's category (*i.e.,* the main screen and the base of the monitors, as well as the back and legs of the chairs). Acquiring such fine-grained object parts provides enhanced feature representations for region proposals.

In summary, the main contributions of the proposed CMAC model can be listed as follows:

- We propose a novel Cross Modal Attentional Context (CMAC) deep learning framework that effectively incorporates the correlated information between different modalities and successfully identifies useful contextual information both locally and globally for RGB-D object detection.
- An attention-based global context module, based on an LSTM network, is utilized to recurrently generate contextual information from a global view for each object proposal.
- Multiple spatial transform networks are adopted in parallel to localize discriminative object parts for accurate object recognition.
- Extensive experiments on the SUNRGBD and NYUv2 datasets well demonstrate the effectiveness of the proposed CMAC model, which outperforms the state-of-the-art method [10] by 3.7% and 3.2%, respectively, in terms of mAP.

II. Related Work

A. Object Detection in RGB-D Images

Object detection in RGB-D images has attracted increased attention because of the rapid development of affordable depth sensors and their diverse application scenarios. Many successful algorithms have been proposed to effectively exploit information from RGB-D data. References [13] and [14] took advantage of hand-designed features such as SIFT and multiple shape features in the depth channel for RGB-D object recognition. Schwarz *et al.* [11] utilized two-stream CNNs pre-trained on ImageNet to extract features from RGB-D images. While most work mainly focuses on the RGB modality, some recent work has been dedicated to improving the object detection performance by taking depth information into consideration. Gupta *et al.* [8] proposed a geocentric embedding to convert each single-channel depth map into a three-channel depth image (HHA image), in which they encoded each pixel with three channels of information, *i.e.,* the height above the ground, the horizontal disparity and the angle with respect to gravity. They also introduced a generalized method for the R-CNN detector that can be applied to RGB-D images; they used large CNNs pre-trained on RGB images to extract features from HHA data. To learn rich representations for the depth modality, [10] transferred supervisions from labeled RGB images to unlabeled depth images. In this paper, we follow [8] and encode depth information into HHA images

for improved feature learning and take the model in [10] as our compared baseline model.

Another core issue of RGB-D object detection is how to merge the features from different sources. Existing fusion strategies can be divided into two streams: (1) Early fusion [13], [15], [16], in which the depth channel is being treated as an extra channel to RGB images and is concatenated with the RGB channels for feature extraction. (2) Late fusion [8]–[10], [17], where features are separately learned for each modality and are concatenated at later stages for object classification. Our model is similar to the late fusion approach, but instead of directly concatenating features for classification, we apply the attention model to the fused features to learn a better global context and discriminative object parts to achieve more accurate object recognition.

B. Context Information in Object Detection

Context information has been applied in many methods to enhance the performance of object detection [18]–[24]. For instance, [24] exploited context from information about the entire scene for object detection and localization. Reference [21] explored contextual relationships between regions in an unsupervised manner, where objects are detected using a discriminative approach. Spatial support and geographic information are used as context clues in [20]. Context models have also been applied to deep-learning-based object detectors. Reference [25] proposed a group recursive learning approach to refine object proposals by incorporating semantic and spatial layout correlations of surrounding proposals. Chu and Cai [26] formulated a fully connected conditional random field (CRF) to incorporate the local appearance and the contextual information in terms of relationships among objects and the global scene based on contextual features generated by a convolutional neural network. Inside-Outside net (ION) [27] introduced spatial recurrent networks (RNNs) to integrate the contextual information outside the region of interest while utilizing skip pooling to extract fine-grained information from multiple low-level convolutional layers. Although our proposed model also explores global contextual information through recurrent networks, it explicitly learn to attend the most relevant regions of the object proposal by generating a weight map for each proposal. The weight map can well reveal the contextual region that corresponds to the final classification result. One the other hand, instead of directly extracting local features from the whole object bounding box, our model can achieve better object feature representations by recurrently discovering the most discriminative object parts inside the object proposal and performing part-level feature fusion.

C. Recurrent Attention Models

Recurrent attentional models have been widely incorporated in deep-learning-based computer vision tasks [28]–[31] to achieve better performance. Bahdanau *et al.* [28] introduced recurrent attention to neural machine translation, which allows the model to adaptively attend to the most relevant part of a sentence. Reference [30] adopted visual attention to dynamically select a sequence of regions and only processed the selected regions for efficient computation. A recent work in [31] used an LSTM-based attention model to learn a description of static images. More recently, an attention mechanism has also been applied to vision tasks for videos. For instance, [32] extended an attention model for video description and employed a temporal attention mechanism to model the dynamic temporal structure of videos. Reference [33] optimized the attention model to attend to the relevant parts within a single frame and attached higher importance to them while performing action recognition.

The work that is most relevant to our proposed method is the attentive context proposed in [29], which also incorporated a recurrent attention model to exploit global contextual information. However, the attention model used in [29] generated a static attentive location map for all object proposals. Instead of utilizing a fixed attentive context, our model generates an attentional context feature adaptive to the input region proposals. Furthermore, we employ a fine-grained object part attention module to harness multiple discriminative object parts inside each object proposal for achieving a superior local feature representation. Experimental analysis in Sec. IV-C demonstrates that our method is more robust to background and inter-class noise.

III. Framework

An overview of our framework is illustrated in Fig. 2. Our RGB-D object detection system, which is based on cross modal attentional context learning, is composed of four components, including fully convolutional networks based feature extraction, cross-modal feature fusion, attention-based global context modeling and fine-grained object part attention. We term this network Cross-Modal Attentional Context (CMAC) network. Specifically, given an RGB-D image, we first employ Multiscale Combinatorial Grouping (MCG) [34] to generate a number of object proposals from RGB information and encode the original depth value to the three-channel HHA representation, as proposed in [8]. Following the benchmark object detection framework of Fast R-CNN [6], our CMAC model takes as input an RGB image, an HHA image and corresponding object proposals to generate class labels as well as a refined bounding box for each object proposal.

As shown in Fig. 2, the feature extraction module is built on two separate fully convolutional sub-networks, including the VGG16 model [35] for RGB modality and AlexNet model [36] for depth modality. The output of the last convolutional layer is being treated as our initial feature for object detection, therein including D convolutional maps. The two fully convolutional sub-networks take as input the RGB image and the HHA image to generate the corresponding feature cube. Region-of-Interest (RoI) pooling operations are performed on the two feature cubes to obtain both global (whole image) and local features (object proposal) of the two modalities before being fed to a cross-modal feature fusion module. Moreover, both the fused global feature and the fused local feature are fed to a global context modeling module to obtain an attentional global context feature for the corresponding object proposal, while

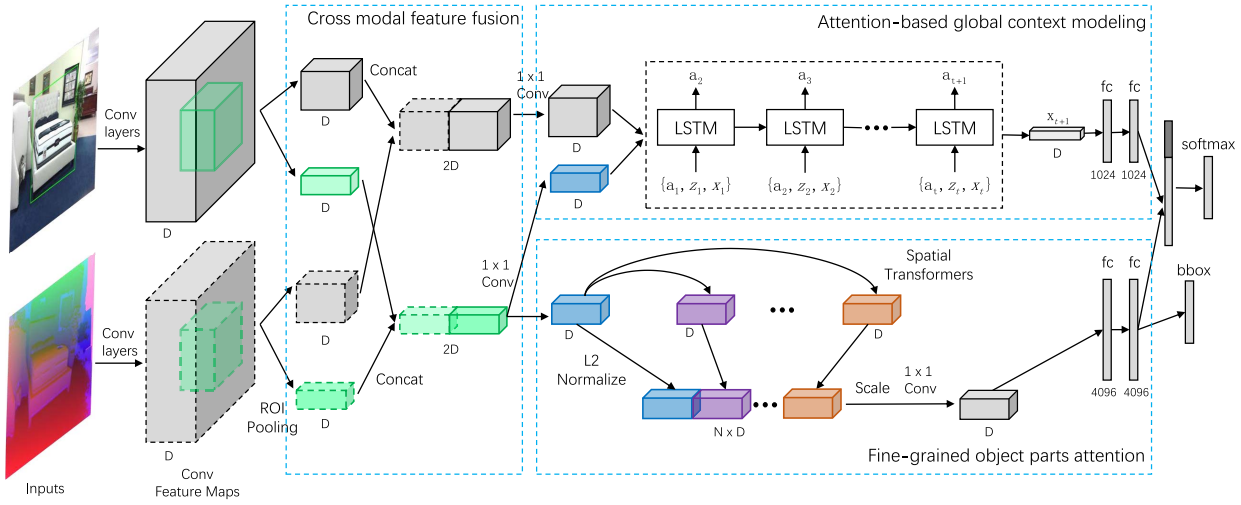


Fig. 2. The network architecture of our proposed cross-modal attentional context (CMAC) learning framework. The input consists of one RGB image and one HHA image (geocentric encoding of the depth image). Our network framework is composed of four components: convolutional feature extraction, cross-modal feature fusion, attention-based global context modeling and fine-grained object part attention.

the fused local feature itself is also treated as an input for the fine-grained object part attention, which generates an embedded local feature. Finally, the concatenation of the global context feature and the embedded local feature are employed for final object detection, while local feature embedding is applied for further bounding box regression.

A. Cross-Modal Feature Fusion

It has been widely verified that the RGB modality and depth modality are complementary, the combination of which can help to boost the RGB-D object detection performance [8], [10]. In this paper, we exploit the features extracted from the two modalities for both global context modeling and local proposal feature description. Specifically, we design a simple yet effective sub-network to fuse features extracted from both modalities. For each object proposal, we extract a fixed-size feature representation using ROI pooling [6] in both modalities, denoted as $\boldsymbol{F}_{l\_rgb}$ and $\boldsymbol{F}_{l\_depth}$. We also apply a pooling operation to the output feature map of the last convolution layer of the two fully convolutional networks to generate fixed-size feature cubes, denoted as $\boldsymbol{F}_{g\_rgb}$ and $\boldsymbol{F}_{g\_depth}$, respectively. The feature fusion between RGB and depth modality can be represented by

$$\boldsymbol{F}_{l\_fused} = concat(\boldsymbol{F}_{l\_rgb}, \boldsymbol{F}_{l\_depth}) \quad (1)$$

$$\boldsymbol{F}_{g\_fused} = concat(\boldsymbol{F}_{g\_rgb}, \boldsymbol{F}_{g\_depth}) \quad (2)$$

where $\boldsymbol{F}_{l\_fused}$ and $\boldsymbol{F}_{g\_fused}$ are the global context feature and local object proposal feature after fusion, respectively, and $concat(\cdot)$ indicates the concatenation operation of feature representations along the channel axis.

In contrast to [8], [10], and [37], which apply two independent CNNs to separately extract features from both modalities and directly perform simple concatenation for final classification, our cross-modal feature fusion operation is treated as a feature generation step for further global context modeling and local feature embedding before final classification. In the experiment section, we verify that our proposed cross-modal feature representation can help to produce more effective local and global context information, greatly improving the performance of the final classification.

B. Attention-Based Global Context Modeling

It is well known that contextual representation is crucial for accurate visual recognition [27], [29], [38]–[41]. Instead of directly obtaining fixed context information to assist in object detection [29], [39], we focus on exploiting adaptive context information for each object proposal. Specifically, we design a soft attention model based on multi-layered RNNs with LSTM units to spatially weight the features and generate an adaptive global context feature for each object proposal. Average pooling and max pooling operations over the feature map of the whole image can be considered as special cases of our method.

The attentional context model takes as input the concatenation of the global feature cube and that of the local feature cube before being fed to a 1×1 convolutional layer for feature embedding. The dimensions of the embedded global and local feature are denoted as $K \times K \times D$ ($20 \times 20 \times 512$ in our experiments) and $S \times S \times D$ ($7 \times 7 \times 512$ in our experiments), respectively. Based on these embedded feature cubes, the RNN model learns an attentional map of size $K \times K$ to determine the effectiveness of the contextual region that may be beneficial to the object detection.

Inspired by the LSTM-based soft attention model proposed in [33], we apply an LSTM network to generate a contextual attention map at every time step conditioned on the previous hidden state, the globally embedded feature vector as well as the local feature. Specifically, at each time-step t, we extract K^2 D-dimensional global feature vectors as well as S^2 local object proposal feature vectors. As in [33], we refer to these feature vectors as global feature slices and local feature slices,

TABLE I

DETECTION RESULTS FROM DIFFERENT METHODS ON SUNRGBD AND NYUV2. AC-CNN* INDICATES OUR IMPLEMENTATION OF THE RGB-D VERSION OF AC-CNN [29]. G AND L DENOTE OUR PROPOSED MODEL INCORPORATED WITH A SINGLE LSTM MODULE (G) OR STN MODULE (L), RESPECTIVELY. (W/O FUSION) AND (W/ FUSION) DENOTE WITHOUT AND WITH MULTI-MODAL CONTEXT FUSION, RESPECTIVELY

Method	G	L	mAP	
			SUNRGBD	NYUv2
ST(baseline) [10]			43.8	49.1
AC-CNN* [29]	√	√	45.4	50.2
Ours (w/o fusion)		√	46.3	50.9
	√		46.2	51.3
	√	√	46.9	51.9
Ours (w/ fusion)	√	√	**47.5**	**52.3**

TABLE II

COMPARISON OF EXPLOITING GLOBAL CONTEXT USING DIFFERENT METHODS ON SUNRGBD AND NYUV2

Method	mAP	
	SUNRGBD	NYUv2
Average Pooling	44.3	49.4
Fixed Attentive Context [29]	44.8	49.7
Adaptive Attentive Context (Ours)	46.2	51.3

respectively, denoted as

$$\begin{cases} G_t = \left[G_{t,1}, \ldots, G_{t,K^2}\right] & G_{t,i} \in \mathbb{R}^D \\ L_t = \left[L_{t,1}, \ldots, L_{t,S^2}\right] & L_{t,i} \in \mathbb{R}^D \end{cases} \tag{3}$$

Each vertical column of G_t and L_t denotes the feature representation (receptive field) in the input image. We follow the implementation of the LSTM network in [42], which is formulated as

$$\begin{pmatrix} \boldsymbol{i}_t \\ \boldsymbol{f}_t \\ \boldsymbol{o}_t \\ \boldsymbol{g}_t \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ tanh \end{pmatrix} \boldsymbol{T} \begin{pmatrix} \boldsymbol{h}_{t-1} \\ \boldsymbol{x}_t \\ \boldsymbol{z}_t \end{pmatrix}, \tag{4}$$

$$\boldsymbol{c}_t = \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + \boldsymbol{i} \odot \boldsymbol{g}_t \tag{5}$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t \odot tanh(\boldsymbol{c}_t) \tag{6}$$

where $\boldsymbol{i}_t$, $\boldsymbol{f}_t$, $\boldsymbol{c}_t$, $\boldsymbol{o}_t$, and $\boldsymbol{h}_t$ are the input gate, forget gate, cell state, output gate and hidden state of the LSTM, respectively; $\boldsymbol{x}_t$ is the global context feature vector input to the LSTM at time step t; the vector $\boldsymbol{z} \in \mathbb{R}^D$ is the local feature embedding of the object proposal with the global average pooling operation; $\boldsymbol{T} \in \mathbb{R}^{(2D+d)\times 4d}$ denotes a simple affine transformation with trainable parameters, where d is the dimensionality of i_t, f_t, c_t and h_t; and σ and $\odot$ denote the logistic sigmoid activation and element-wise multiplication, respectively.

At each time step t, our LSTM model learns to predict a weight map α_{t+1} of size $K \times K$, where its value corresponds to the spatial attention that should be paid when performing proposal classification. The weight map α_i is computed by a multilayer perception ϕ conditioned on the previous hidden state h_{t-1}. The spatial weight of α_t at location i can thus be computed as follows:

$$e_{ti} = \phi(\boldsymbol{h}_{t-1}) \tag{7}$$

$$\alpha_{ti} = \frac{exp(e_{ti})}{\sum_{k=1}^{K\times K} exp(e_{tk})} \tag{8}$$

Based on the weight map, the global context feature vector x at time step t is computed as an average of the feature slices weighted according to α_t, formulated as

$$\boldsymbol{x}_t = \sum_{i=1}^{K^2} \alpha_{t,i} \boldsymbol{F}_{g\_fused,i} \tag{9}$$

where $F_{g\_fused,i}$ is the i^{th} global feature slice. Because the relevant regions are given higher weights, the global feature $\boldsymbol{x}_t$ will be dominated by features from these regions and hence provide more useful contextual information for more accurate object detection.

During the initialization stage, we follow the same strategy proposed in [43] for faster convergence. Specifically, we initialize the cell state $\boldsymbol{c}_t$ and the hidden state $\boldsymbol{h}_t$ of the LSTM network as

$$\boldsymbol{c}_0 = f_{init,\,c}\left(\frac{1}{K^2}\sum_{i=1}^{K^2} \boldsymbol{F}_{g\_fused,i}\right) \tag{10}$$

$$\boldsymbol{h}_0 = f_{init,\,h}\left(\frac{1}{K^2}\sum_{i=1}^{K^2} \boldsymbol{F}_{g\_fused,i}\right) \tag{11}$$

where $f_{init,\,c}$ and $f_{init,\,h}$ are two multi-layer perceptions. The two initial values are applied to infer the initial weights α_1 for the initialization of the global context feature $\boldsymbol{x}_1$.

As shown in Fig. 2, the output of our LSTM model is a D-dimensional global context feature, which is further fed to two fully connected layers to produce the final feature representation, denoted as F_G.

C. Fine-Grained Object Part Attention

Because the local salient parts inside a specific object proposal play an important guiding role in judging the classification of an object (especially for partially occluded objects), we further propose to employ multiple STNs [44] in parallel to infer discriminative object parts for each object proposal. The spatial transformer is a differential module that learns to spatially transform the input feature maps U to the output feature maps V. A spatial transformer is applied in the following three steps. First, a localization network is employed to predict the affine transformation matrix A_θ to be applied to the input feature map. Second, A_θ is being applied to create a sampling grid in U by the grid generator. Finally, a sampler is adopted to produce the output maps sampled from the regions of input maps at the sampling grid. As shown in Figure 3, we train each transformer to automatically attend to discriminative object parts inside an object proposal. During training, we fix the scaling factor to 0.5 and only accept scaling

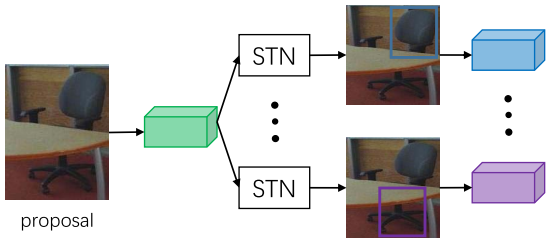


Fig. 3. Illustration of the STN module. The STN module takes the feature of the object proposal as input and attends to the most discriminative parts. The feature from these parts will subsequently serve as an enhanced local feature in object classification and bounding box regression.

and translating in each spatial transformer. Thus, A_θ is given by

$$A_\theta = \begin{bmatrix} 0.5 & 0 & t_x \\ 0 & 0.5 & t_y \end{bmatrix} \tag{12}$$

where $\theta = [t_x, t_y]$ are the translation parameters that are predicted based on the localization network.

Taking the local context feature map $\boldsymbol{F}_{l\_fused} \in \mathbb{R}^{D\times S\times S}$ as input, each transformer in our object part attention module transforms and samples the input map to the output map $\boldsymbol{q} \in \mathbb{R}^{D\times S\times S}$. After normalization, the outputs of each transformer are concatenated with the local context feature to form a mid-level feature representation for an object proposal, defined as

$$\boldsymbol{F}_{mid} = concat(\boldsymbol{F}_{l\_fused},\ \boldsymbol{q}_i,\ \ldots,\ \boldsymbol{q}_N). \tag{13}$$

where $\boldsymbol{q}_i$ is the output of the i_{th} transformer and N is the number of spatial transformers.

As shown in Figure 2, we use a 1×1 convolution layer after re-scaling to reduce the dimensions of $\boldsymbol{F}_{mid}$ from $S \times S \times (N \times D)$ to $S \times S \times D$, which is then fed to two fully connected layers to infer the final feature representation for the object proposal, denoted as $\boldsymbol{F}_L$.

D. Training Objective

Denote $p = (p_0, \ldots, p_L)$ as the predicted discrete probability distribution (per ROI) over $C+1$ categories and t^* as the predicted bounding-box regression offsets. Given the obtained local and global context features $\boldsymbol{F}_L$ and $\boldsymbol{F}_G$, p and t^* can be computed as follows:

$$p = Softmax\left(f_{cls}\left(concat(\boldsymbol{F}_L, \boldsymbol{F}_G)\right)\right) \tag{14}$$

$$t^* = f_{loc}\left(\boldsymbol{F}_L\right) \tag{15}$$

where $Softmax(\cdot)$ indicates the softmax operation and f_{cls} and f_{loc} are two fully connected layers with $C+1$ units and $4 \times C$ units, respectively.

Note that we only incorporate local contextual information for bounding-box regression. Finally, we minimize an objective function following the multi-task loss given in Fast-RCNN [6], which is defined as

$$L(p, u, t^u, v) = L_{cls}(p, u) + [u \geq 1]L_{loc}(t^u, v) \tag{16}$$

where u is the ground-truth label, v is the regression target, L_{cls} is the log loss for ground-truth class u, and L_{loc} is the smooth L_1 loss proposed in [6]. $[u \geq 1]$ evaluates to 1 when $u \geq 1$ and 0 otherwise. By convention, the background class is labeled as $u = 0$.

IV. Experimental Results

A. Experimental Settings

1) Datasets and Evaluation Metrics: We evaluate our model on two RGB-D datasets: SUNRGBD [45] and NYUv2 [46]. The SUNRGBD and NYUv2 datasets contain 10335 and 1449 RGB-D images, respectively, and are divided into *train* and *test* subsets. We adopt *Average Precision* (AP) and *mean of Average Precision* (mAP) following the PASCAL challenge protocols as our evaluation metrics.

2) Implementation Details: In our experiments, we implement our model based on Fast R-CNN [6], an open-source framework for traditional RGB object detection built on the Caffe platform [47]. We utilize the network architecture from Gupta *et al.* [10] as our basic CNN network structure for convolutional feature map extraction. All the newly added fully connected and convolutional layers are randomly initialized with a zero-mean Gaussian distribution with standard deviations of 0.01 and 0.001. The recurrent attention model consists of 4 stacked LSTM units with shared parameters. All the parameters of the LSTM units are initialized based on the xavier algorithm [48].

We apply Stochastic Gradient Decent (SGD) to fine tune our model. Each SGD mini-batch is composed of 128 randomly sampled object proposals from 2 randomly chosen images. In each mini-batch, we select 25% of the ROIs as foreground from object proposals that have intersection over union (IoU) overlap with a ground-truth bounding box of at least 0.5. The remaining ROIs are sampled from object proposals that have a maximum IoU with ground truth in the interval $[0.1, 0.5)$ and act as background with ground truth label $u = 0$. During training, images are horizontally flipped with a probability of 0.5 for data augmentation, and no other augmentation is used. We run SGD for approximately 10 epochs on the training set to fine tune the network parameters. The momentum is set to 0.9, and the learning rate is initialized to 0.001 and decreased by 10 every 4 epochs. It takes approximately 1.5 days to train our model on a single NVIDIA GeForce GTX TITAN X GPU with 12 GB of memory.

It costs approximately 10 GB of GPU memory to train our model. The average training time for each iteration is approximately 1.23 seconds. However, the testing process is particularly efficient and takes approximately 0.58 seconds (excluding object proposal extraction) to process one image.

B. Performance Comparisons

1) RGB-D Datasets: We compare our proposed method against recent state-of-the-art RGB-D object detection methods, including rich image and depth feature-based RGB-D object detection [8] and the supervision-transfer-based model [10]. Moreover, to better validate the superiority of the attention-based global context and fine-grained object part attention on RGB-D datasets, we also implement an RGB-D version (denoted as AC-CNN*) of the AC-CNN model

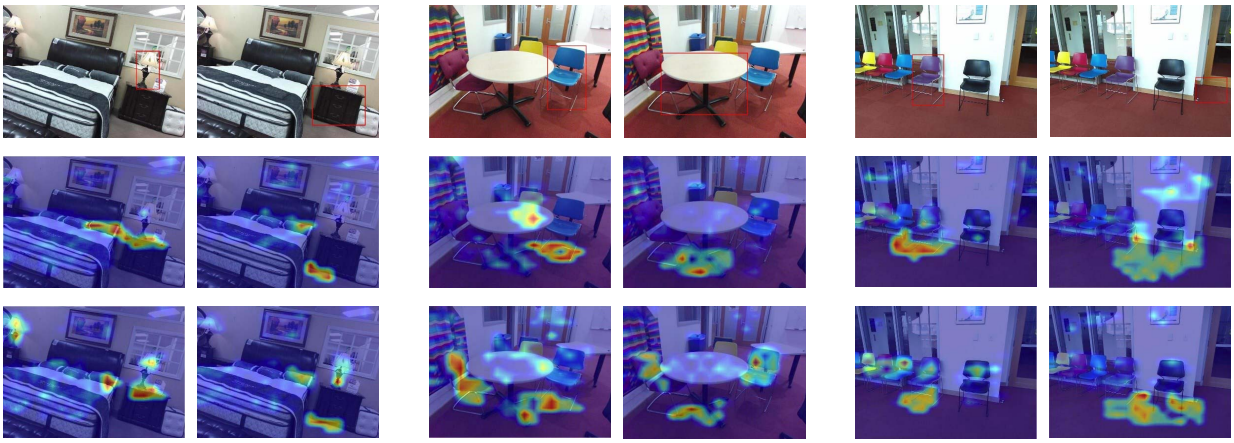


Fig. 4. Illustration of the attentional weight maps generated by the attention-based global context modeling module. The top rows are the input images and region proposals. The middle and bottom rows are the attentional weight maps generated by our model without context fusion and those with context fusion, respectively. The bottom two rows show that our model can perceive the most relevant regions to the given object proposal and that more useful regions can be acquired through context fusion. A detailed discussion can be found in section IV-D.

TABLE III

DETECTION RESULTS ON SUNRGBD. AC-CNN* INDICATES OUR IMPLEMENTATION OF THE RGB-D VERSION OF AC-CNN [29]. (W/O FUSION) AND (W/ FUSION) DENOTE WITHOUT AND WITH MULTI-MODAL CONTEXT FUSION, RESPECTIVELY

Method	mAP	bathtub	bed	bookshelf	box	chair	counter	desk	door	dresser	garbage_bin	lamp	monitor	night_stand	pillow	sink	sofa	table	tv	toilet
RGB-D RCNN [8]	35.2	49.6	76.0	35.0	5.8	41.2	8.1	16.6	4.2	31.4	46.8	22.0	10.8	37.2	16.5	41.9	42.2	43.0	32.9	69.8
ST(baseline) [10]	43.8	65.3	83.0	54.4	14.4	46.9	14.6	23.9	15.3	41.3	51.0	32.1	36.8	46.6	23.4	43.9	61.3	48.7	50.5	79.4
AC-CNN*	45.4	65.8	83.3	56.2	16.4	47.5	16.0	24.9	16.6	42.7	53.4	33.8	39.5	47.1	25.2	45.3	61.9	49.0	54.1	84.2
Ours (w/o fusion)	46.9	68.2	85.7	56.0	17.3	49.8	17.1	25.2	16.9	43.5	54.2	35.5	40.7	49.4	26.1	**46.6**	66.3	52.0	56.5	84.3
Ours (w/ fusion)	**47.5**	**69.0**	**86.1**	**57.9**	**18.2**	**50.3**	**17.4**	**26.8**	**17.3**	**44.4**	**54.4**	**35.6**	**40.5**	**49.8**	**26.7**	46.6	**67.2**	**52.9**	**56.7**	**84.9**

TABLE IV

DETECTION RESULTS ON NYUV2. AC-CNN* INDICATES OUR IMPLEMENTATION OF THE RGB-D VERSION OF AC-CNN [29]. (W/O FUSION) AND (W/ FUSION) DENOTE WITHOUT AND WITH MULTI-MODAL CONTEXT FUSION, RESPECTIVELY

Method	mAP	bathtub	bed	bookshelf	box	chair	counter	desk	door	dresser	garbage_bin	lamp	monitor	night_stand	pillow	sink	sofa	table	tv	toilet
RGB-D RCNN [8]	32.5	22.9	66.5	21.8	3.0	40.8	37.6	10.2	20.5	26.2	37.6	29.3	43.4	39.5	37.4	24.2	42.8	24.3	37.2	53.0
ST(baseline) [10]	49.1	50.6	81.0	52.6	5.4	53.0	56.1	21.0	34.6	57.9	46.2	42.5	62.9	54.7	49.1	50.0	65.9	31.9	50.1	68.0
AC-CNN*	50.2	52.2	82.4	52.5	8.6	54.8	57.3	22.7	34.1	58.1	46.5	42.9	63.6	55.2	49.7	51.4	66.8	33.5	51.8	70.4
Ours (w/o fusion)	51.9	55.2	83.4	**54.2**	9.4	55.1	58.5	24.0	35.9	58.3	46.6	44.8	65.7	57.0	**52.7**	53.6	68.4	**35.3**	54.8	73.9
Ours (w/ fusion)	**52.3**	**55.6**	**83.9**	54.0	**9.8**	**55.4**	**59.2**	**24.1**	**36.3**	**58.5**	**47.2**	**45.0**	**65.8**	**57.6**	52.7	**53.8**	**69.1**	35.0	**56.9**	**74.7**

proposed in [29] for comparison. AC-CNN follows a similar idea to our proposed method but incorporates fixed global and local attentive contexts to assist in improving the object detection performance. In the implementation, we apply the Fast RCNN [6] framework based on AlexNet [36] to the depth modality for proposal classification and bounding-box position regression. The final results are obtained by averaging the results from the RGB modality and depth modality. For fair comparison, we also apply the same depth modality processing as in AC-CNN* to our model; we call this custom model RGB-D detection without cross-modal fusion (denoted as w/o fusion).

Table III and Table IV illustrate the object detection results of our model, AC-CNN*, and the other two state-of-the-art RGB-D object detection models on the SUNRGBD and NYUv2 datasets. As shown in the table, our proposed method obtains state-of-the-art mAP scores of 47.5% and 52.3% on SUNRGBD and NYUv2, which outperforms the ST model [10] by 3.7% and 3.2%, respectively. The improvements validate the effectiveness of our model in RGB-D object detection by incorporating the proposed attention-based global context and fine-grained attentional object parts learned from the fused cross-modal context. Furthermore, our model (Ours (w/o fusion)) gains 1.5% and 1.7% improvements in mAP scores

TABLE V

COMPARISON OF DIFFERENT LSTM SETTINGS UTILIZED IN THE ATTENTION-BASED GLOBAL CONTEXT SUB-MODULE. THE EXPERIMENTS ARE CONDUCTED ON SUNRGBD. (2 × LSTM) DENOTES THAT THERE ARE 2 STACKED LSTM UNITS IN THE GLOBAL CONTEXTUALIZED SUB-NETWORK

LSTM Settings	mAP
Ours (2 × LSTM)	45.4
Ours (3 × LSTM)	46.0
Ours (4 × LSTM)	**46.2**
Ours (5 × LSTM)	**46.2**

TABLE VI

COMPARISON OF DIFFERENT STN SETTINGS UTILIZED IN FINE-GRAINED OBJECT PART ATTENTION SUB-MODULE. THE EXPERIMENTS ARE CONDUCTED ON SUNRGBD. (2 × STN) INDICATES THAT THERE ARE 2 PARALLEL SPATIAL TRANSFORMERS IN THE LOCAL CONTEXTUALIZE SUB-NETWORK

STN Settings	mAP
Ours (1 × STN)	45.7
Ours (2 × STN)	**46.3**
Ours (3 × STN)	46.0

over AC-CNN* on the SUNRGBD and NYUv2 datasets, respectively, and achieve better detection results on most of the categories.

2) RGB Dataset: To compare our model with the AC-CNN model [29] in a more equitable way, we remove the depth modality from our model and perform an extra evaluation on PASCAL VOC 2007, which contains 9963 RGB images. Specifically, we implement a variant of our model (denoted as Ours*) that performs global context modeling and object part attention only on the RGB modality without incorporating information from the depth modality. As shown in Table VII. Our model outperforms the baseline FRCN [6] and AC-CNN [29] by 3.6% and 1.2% in terms of mAP scores, respectively. The improvement on the RGB dataset as well as the favorable results achieved for RGB-D object detection well demonstrate the superiority of the proposed attention-based global context and fine-grained object part attention over the fixed global context and multi-scale local context proposed in [29]. Table VIII provides the comparisons of the proposed method with several state-of-the-art methods [27], [49]–[52] on PASCAL VOC 2012. It can be observed that our model obtains an mAP score of 76.7%, which outperforms the baseline model by 2.9%. Our model also achieves competitive results compared with the state-of-the-art methods, which validates the effectiveness of the proposed method.

C. Ablation Studies

In this subsection, we show the effectiveness and necessity of each component in our proposed model and also demonstrate the effectiveness of the network design.

1) Contribution of Each Component in CMAC Model: As described in Section III, our proposed CMAC model consists of three newly added sub-networks on the top of deep feature representation, including cross-modal feature fusion, attention-based global context modeling and fine-grained object part attention, which are employed to incorporate the strong correlation between different modalities and capture the global and local contextual information, respectively. We investigate the contributions of each component by gradually applying each sub-network to the object detection. Table I shows that 2.5% and 1.8% improvements in mAP scores over the baseline model are obtained using only fine-grained object part attention. Similar improvements of 2.4% and 2.2% on SUNRGBD and NYUv2 can be observed when only incorporating attention-based global context modeling. The better performance achieved by exploiting both global context features and discriminative object parts evidences the complementarity of the two sub-networks. Furthermore, incorporating cross-modal feature fusion into our detection framework brings an extra performance increase of 0.6% and 0.4% on SUNRGBD and NYUv2, respectively. The above experimental results and analysis well demonstrate the effectiveness of each component in our proposed CMAC framework.

2) Comparison of Diverse Global Context Modeling: To validate the effectiveness of our attention-based global context, which is generated based on a recurrent model, we compare our model with two variants: the global average pooling method in which the global contextual information is produced by applying the average pooling operation to the extracted feature map, and AC-CNN, which utilizes an attention-based recurrent model to generate the fixed global context. We conduct experiments on the SUNRGBD dataset, and the results are listed in Table II. No local context is used during these experiments. It can be observed that our model outperforms the global averaging pooling method and AC-CNN by 1.9% and 1.5%, respectively. Simply averaging the features of all regions may introduce both background and inter-class noise, which may deteriorate the object detection performance. Although background noise can be overcome by AC-CNN, which generates a fixed attention map for global context feature extraction and benefits the proposal classification, AC-CNN still suffers from a decreased performance caused by inter-class noise (e.g., regions that are beneficial for desk classification might provide noisy information to garbage_bin classification). Note that our attention map for global context weighting is generated according to the diverse contents of each ROI feature and can be optimized to attend to the most effective regions related to the input content. The results shown in Table II verify that our model performs better in mitigating both background and inter-class noise by incorporating global context and thus greatly enhances the accuracy of object detection.

3) Effectiveness of LSTM Settings: In our proposed CMAC model, we have employed a recurrent model to exploit the attentional global context, in which multiple stacked LSTM units are utilized to generate the attentional weight map in an iterative manner. To investigate the effectiveness of different LSTM settings, we implement several variants, whereby the recurrent model is constructed with different numbers (2 to 5) of LSTM units. The experimental results are listed in Table V. As shown in the table, the mAP metric increases

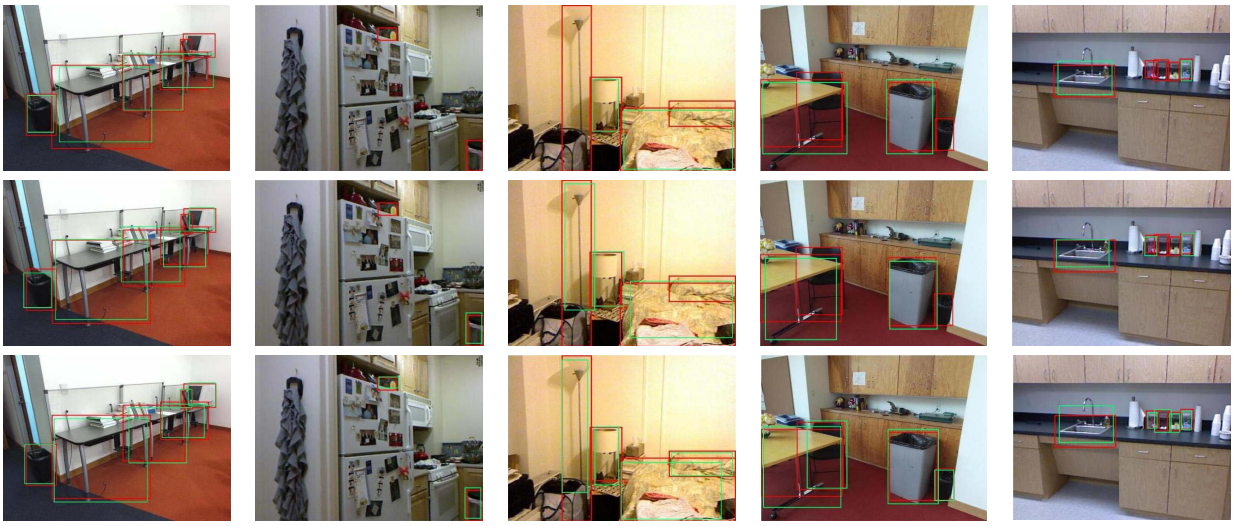


Fig. 5. Comparison of detection results produced by ST [10] (top row), AC-CNN [29] (middle row) and our model (bottom row). The red and green rectangles indicate the ground-truth bounding box and the predicted results, respectively.

TABLE VII

DETECTION RESULTS ON VOC 2007. OURS* DENOTES A VARIANT OF OUR MODEL IN WHICH WE INCORPORATE ONLY RGB INFORMATION FOR OBJECT DETECTION

Method	mAP	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv
FRCN(Baseline) [6]	70.0	77.0	78.1	69.3	59.4	38.3	81.6	78.6	86.7	42.8	78.8	68.9	84.7	82.0	76.6	69.9	31.8	70.1	74.8	80.4	70.4
AC-CNN [29]	72.4	79.1	79.2	71.9	**61.0**	43.2	83.0	81.4	87.7	50.0	82.1	73.6	83.4	84.2	77.5	72.0	35.8	71.9	74.7	**85.8**	71.0
Ours*	**73.6**	**81.0**	**80.2**	**72.4**	60.5	**45.3**	**84.1**	**82.8**	**88.0**	**51.6**	**82.5**	**74.8**	**85.7**	**84.9**	**79.6**	**72.2**	**36.9**	**72.1**	**76.8**	85.5	**74.3**

by 0.6% and 0.8% when the number of stacked LSTM units is increased from 2 to 3 and 4, respectively. When this number reaches or exceeds 5, no significant performance boosts are achieved, indicating that our model can obtain better context information through recurrent iterations and will converge quickly. We believe that good performance can be obtained in complicated images through more recurrent iterations.

4) Effectiveness of STN Settings: In the proposed method, we adopt several parallel multiple transform networks (STNs) to attend to discriminative object parts inside an object proposal. To investigate the most effective STN setting, we implement several variants whereby the fine-grained object parts are inferred from different numbers (2 to 4) of spatial transformers. As shown in Table VI, the detection performance increases from 43.8% (baseline) to 45.7% and 46.3% with 1 and 2 spatial transformers, respectively, which indicates that STNs are able to mine discriminative object parts to enhance the local feature representation. However, increasing the number of spatial transformers does not always bring about a better performance. We observe a 3% decrease in mAP when increasing the number of spatial transformers from 2 to 3, indicating that the STNs may start to enroll confusing object parts after most of the discriminative parts have been detected.

D. Visualization

In this subsection, we present some visual comparisons of the RGB-D object detection results as well as some visual effects of the attentional weight maps generated by our global context modeling component. Figure 5 shows some detection results of the ST [10] model, the AC-CNN [29] model and our model. It can be observed that our model performs best in detecting small and occluded objects (e.g., monitor, box, garbage_bin and the occluded chair). Furthermore, as shown in the third column, our proposed method is also more robust to appearance-similar instances because of the fusion of the geometry context (e.g., the pillow with similar texture to the bed). Figure 4 demonstrates the attentional weight maps generated by our model without (middle row) and with (bottom row) context fusion. Obviously, our attentional model is able to perceive regions most relevant to the specific object

TABLE VIII

PASCAL VOC 2012 TEST DETECTION RESULTS. **07+12+S:** 07 TRAINVAL + 12TRAINVAL + SEGMENTATION LABELS, **07++12:** 07 TRAINVAL + 07 TEST + 12 TRAINVAL

Method	data	mAP
Faster-Rcnn [7] (Baseline)	07++12	73.8
Multi-stage [25]	07++12	74.9
SSD300 [50]	07++12	75.8
DOSD [51]	07++12	76.3
ION [27]	07+12+S	76.4
R-FCN multi-scale [52]	07++12	**77.6**
ours	07++12	76.7

proposal, *i.e.,* a lamp is likely to be placed on top of a night stand near a bed, and a night stand is also likely to be placed on the floor near a bed and often co-occurs with a lamp. Moreover, our model obtains more accurate attentional weight maps by fusing information from both RGB and depth modalities since the depth image can provide geometric information. For example, our model is capable of attending to the chairs near the target chair, as they share similar geometric structures. The last column in Fig. 4 shows that our model will attend to the background regions when the proposal does not contain objects, which helps in making correct classifications.

V. Conclusion

In this paper, we have introduced an approach to effectively learn the cross-modal attentive context for RGBD object detection. In our model, the contextual representations from different sources (*i.e.,* RGB and depth modalities) are fused in the cross-modal feature fusion module. Based on the fused local and global feature, a recurrent attention model including several stacked LSTM units is employed to capture a global context that is closely related to the object proposal. Furthermore, our model adopts several parallel spatial transformers, which learn to attend to discriminative parts inside each object proposal, to generate the enhanced local context information. Extensive experiments and state-of-the-art detection results on SUNRGBD and NYUv2 well demonstrate the effectiveness of our model in exploiting contextual information.

References

[1] S. Hinterstoisser *et al.*, "Model based training, detection and pose estimation of texture-less 3D objects in heavily cluttered scenes," in *Proc. Asian Conf. Comput. Vis.* Springer, 2012, pp. 548–562.

[2] K. Wang, S. Zhai, H. Cheng, X. Liang, and L. Lin, "Human pose estimation from depth images via inference embedded multi-task learning," in *Proc. ACM Multimedia Conf.*, Oct. 2016, pp. 1227–1236.

[3] C. Wu, I. Lenz, and A. Saxena, "Hierarchical semantic labeling for task-relevant RGB-D perception," in *Proc. Robot. Sci. Syst.*, Jul. 2014.

[4] S. Schuster, R. Krishna, A. Chang, L. Fei-Fei, and C. D. Manning, "Generating semantically precise scene graphs from textual descriptions for improved image retrieval," in *Proc. 4th Workshop Vis. Lang.*, 2015, pp. 70–80.

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Oct. 2014, pp. 580–587.

[6] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1440–1448.

[7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Jun. 2015, pp. 91–99.

[8] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from RGB-D images for object detection and segmentation," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2014, pp. 345–360.

[9] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard, "Multimodal deep learning for robust RGB-D object recognition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst. (IROS)*, Sep. 2015, pp. 681–687.

[10] S. Gupta, J. Hoffman, and J. Malik, "Cross modal distillation for supervision transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2827–2836.

[11] M. Schwarz, H. Schulz, and S. Behnke, "RGB-D object recognition and pose estimation based on pre-trained convolutional neural network features," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2015, pp. 1329–1335.

[12] Z. Li, Y. Gan, X. Liang, Y. Yu, H. Cheng, and L. Lin, "LSTM-CF: Unifying context modeling and fusion with lstms for RGB-D scene labeling," in *Proc. Eur. Conf. Comput. Vis.* Springer, Sep. 2016, pp. 541–557.

[13] L. Bo, X. Ren, and D. Fox, "Depth kernel descriptors for object recognition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst. (IROS)*, Sep. 2011, pp. 821–826.

[14] K. Lai, L. Bo, X. Ren, and D. Fox, "A large-scale hierarchical multi-view RGB-D object dataset," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2011, pp. 1817–1824.

[15] M. Blum, J. T. Springenberg, J. Wülfing, and M. Riedmiller, "A learned feature descriptor for object recognition in RGB-D data," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2012, pp. 1298–1303.

[16] L. Bo, K. Lai, X. Ren, and D. Fox, "Object recognition with hierarchical kernel descriptors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 1729–1736.

[17] L. Spinello and K. O. Arras, "Leveraging RGB-D Data: Adaptive fusion and domain adaptation for object detection," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2012, pp. 4469–4474.

[18] P. Carbonetto, N. De Freitas, and K. Barnard, "A statistical model for general contextual object recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2004, pp. 350–362.

[19] G. Li and Y. Yu, "Visual saliency detection based on multiscale deep CNN features," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5012–5024, Nov. 2016.

[20] S. K. Divvala, D. Hoiem, J. H. Hays, A. A. Efros, and M. Hebert, "An empirical study of context in object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1271–1278.

[21] G. Heitz and D. Koller, "Learning spatial context: Using stuff to find things," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2008, pp. 30–43.

[22] D. Hoiem, A. A. Efros, and M. Hebert, "Geometric context from a single image," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 1, Oct. 2005, pp. 654–661.

[23] G. Li and Y. Yu, "Contrast-oriented deep neural networks for salient object detection," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published.

[24] A. Torralba, K. P. Murphy, and W. T. Freeman, "Using the forest to see the trees: Exploiting context for visual object detection and localization," *Commun. ACM*, vol. 53, no. 3, pp. 107–114, Mar. 2010.

[25] J. Li *et al.*, "Multistage object detection with group recursive learning," *IEEE Trans. Multimedia*, vol. 20, no. 7, pp. 1645–1655, Jul. 2017.

[26] W. Chu and D. Cai. (Apr. 2016). "Deep feature based contextual model for object detection." [Online]. Available: https://arxiv.org/abs/1604.04048

[27] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick, "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2874–2883.

[28] D. Bahdanau, K. Cho, and Y. Bengio. (May 2014). "Neural machine translation by jointly learning to align and translate." [Online]. Available: https://arxiv.org/abs/1409.0473

[29] J. Li *et al.*, "Attentive contexts for object detection," *IEEE Trans. Multimedia*, vol. 19, no. 5, pp. 944–954, May 2016.

[30] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2204–2212.

[31] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. ICML*, vol. 14, 2015, pp. 4507–4515.

[32] L. Yao *et al.*, "Describing videos by exploiting temporal structure," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 4507–4515.

[33] S. Sharma, R. Kiros, and R. Salakhutdinov. (Feb. 2015) "Action recognition using visual attention." [Online]. Available: https://arxiv.org/abs/1511.04119

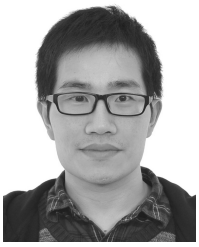
[34] P. Arbelaez, J. Pont-Tuset, J. T. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 328–335.

[35] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556

[36] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[37] S. Song and J. Xiao, "Deep sliding shapes for amodal 3D object detection in RGB-D images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 808–816.

[38] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 5455–5463.
[39] R. Mottaghi *et al.*, "The role of context for object detection and semantic segmentation in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 891–898.
[40] H. Li, G. Li, L. Lin, H. Yu, and Y. Yu, "Context-aware semantic inpainting," *IEEE Trans. Cybern.*, to be published.
[41] G. Li, Y. Xie, L. Lin, and Y. Yu, "Instance-level salient object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 247–256.
[42] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[43] K. Xu *et al.* (2015). "Show, attend and tell: Neural image caption generation with visual attention." [Online]. Available: https://arxiv.org/abs/1502.0304
[44] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2017–2025.
[45] S. Song, S. P. Lichtenberg, and J. Xiao, "Sun RGB-D: A RGB-D scene understanding benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 567–576.
[46] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2012, pp. 746–760.
[47] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, pp. 675–678.
[48] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," *J. Mach. Learn. Res.*, vol. 9, pp. 249–256, May 2010.
[49] S. Ravishankar, A. Jain, and A. Mittal, "Multi-stage contour based detection of deformable objects," in *Proc. Eur. Conf. Comput. Vis.* Springer, Oct. 2008, pp. 483–496.
[50] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* Springer, Oct. 2016, pp. 21–37.
[51] Z. Shen, Z. Liu, J. Li, Y.-G. Jiang, Y. Chen, and X. Xue, "DSOD: Learning deeply supervised object detectors from scratch," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 3, no. 6, 2017, p. 7.
[52] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 379–387.



Guanbin Li (M'18) received the Ph.D. degree from the University of Hong Kong in 2016. He is currently a Research Associate Professor with the School of Data and Computer Science, Sun Yat-sen University. His current research interests include computer vision, image processing, and deep learning. He has authored and co-authored over 20 papers in top-tier academic journals and conferences. He was a recipient of the Hong Kong Postgraduate Fellowship. He serves as an Area Chair for the conference of VISAPP. He has been serving as a reviewer for numerous academic journals and conferences, such as TPAMI, TIP, TMM, TC, CVPR2018, and IJCAI2018.

Yukang Gan received the B.E. degree from the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China, in 2015, where he is currently pursuing the M.Sc. degree in computer science. His research interests include computer vision and machine learning.

Hejun Wu (M'07) received the Ph.D. degree in computer science and engineering from the Hong Kong University of Science and Technology in 2008. He is currently an Associate Professor with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China. His main research interests include distributed computing and machine learning. He is a member of the ACM.

Nong Xiao (M'97) received the B.S. and Ph.D. degrees in computer science from the College of Computer, National University of Defense Technology (NUDT), China, in 1990 and 1996, respectively. He is currently a Professor with the State Key Laboratory of High Performance Computing, NUDT. He has more than 130 publications to his credit in journals and international conferences, including IEEE TSC, IEEE TMM, JPDC, JCST, HPCA, ICCAD, MIDDLEWARE, MSST, IPDPS, CLUSTER, SYSTOR, and MASCOTS. His current research interests include large-scale storage system, network computing, and computer architecture. He is a member of the ACM.

Liang Lin (M'09–SM'15) is the Executive R&D Director of the SenseTime Group Ltd. and a Full Professor with Sun Yat-sen University. He is the Excellent Young Scientist of the National Natural Science Foundation of China. From 2008 to 2010, he was a Post-Doctoral Fellow at the University of California at Los Angeles, Los Angeles, CA, USA. From 2014 to 2015, as a Senior Visiting Scholar, he was with The Hong Kong Polytechnic University and The Chinese University of Hong Kong. He currently leads the SenseTime R&D teams to develop cutting-edges and deliverable solutions on computer vision, data analysis and mining, and intelligent robotic systems. He has authored and co-authored more than 100 papers in top-tier academic journals and conferences. He is a fellow of IET. He served as Area/Session Chairs for numerous conferences, such as ICME, ACCV, and ICMR. He was a recipient of the Best Paper Runners-Up Award in ACM NPAR 2010, the Google Faculty Award in 2012, the Best Paper Diamond Award in IEEE ICME 2017, and the Hong Kong Scholars Award in 2014. He has been serving as an Associate Editor for the IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS, *The Visual Computer*, and *Neurocomputing*.