Multi-modal Deep Feature Learning for RGB-D Object Detection

Xiangyang Xu[a,*], Yuncheng Li[b], Gangshan Wu[a], Jiebo Luo[b]

[a] *State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China.*
[b] *Department of Computer Science, University of Rochester, Rochester, NY 14627 USA.*

Abstract

We present a novel multi-modal deep feature learning architecture for RGB-D object detection. The current paradigm for object detection typically consists of two stages: objectness estimation and region-wise object recognition. Most existing RGB-D object detection approaches treat the two stages separately by extracting RGB and depth features individually, thus ignore the correlated relationship between these two modalities. In contrast, our proposed method is designed to take full advantages of both depth and color cues by exploiting both modality-correlated and modality-specific features and jointly performing RGB-D objectness estimation and region-wise object recognition. Specifically, shared weights strategy and a parameter-free correlation layer are exploited to carry out RGB-D-correlated objectness estimation and region-wise recognition in conjunction with RGB-specific and depth-specific procedures. The parameters of these three networks are simultaneously optimized via end-to-end multi-task learning. The multi-modal RGB-D objectness estimation results and RGB-D object recognition results are both boosted by late-fusion ensemble. To validate the effectiveness of the proposed approach, we conduct extensive experiments on two challenging RGB-D benchmark datasets, NYU Depth v2 and SUN RGB-D. The experimental results show that by introducing the modality-

*Corresponding author. E-mail addresses: xiangyang.xu@smail.nju.edu.cn (Xiangyang Xu), yli@cs.rochester.edu (Yuncheng Li), gswu@nju.edu.cn (Gangshan Wu) and jluo@cs.rochester.edu (Jiebo Luo).

correlated feature representation, the proposed multi-modal RGB-D object detection approach is substantially superior to the state-of-the-art competitors. Moreover, compared to the expensive deep architecture (VGG16) that the state-of-the-art methods preferred, our approach, which is built upon more lightweight deep architecture (AlexNet), performs slightly better.

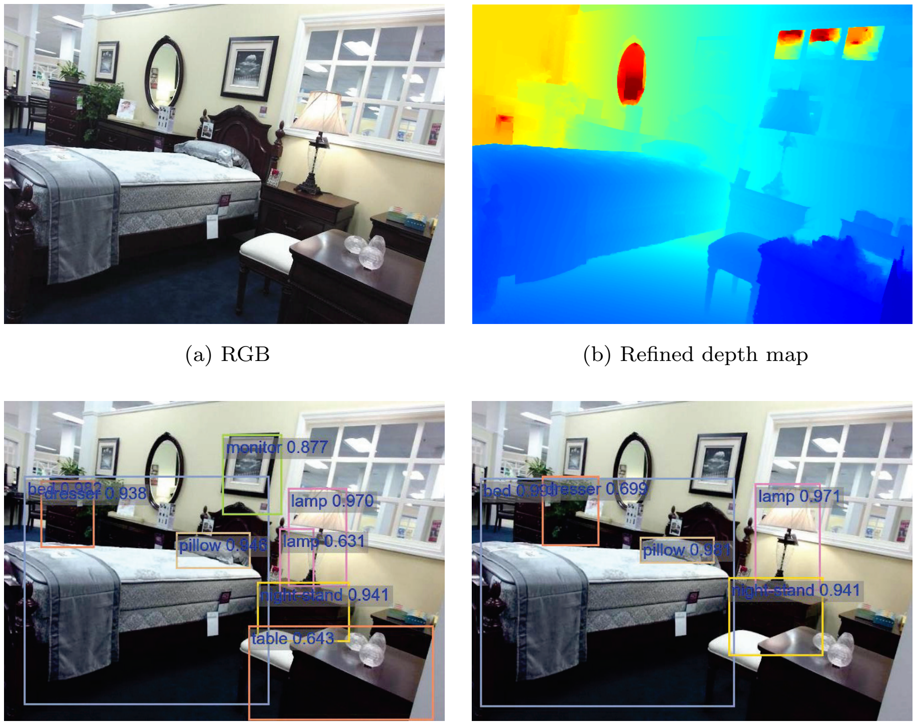
Keywords: RGB-D objectness estimation, RGB-D object detection, multi-modal learning, convolutional neural networks

---

1. Introduction

Object detection, which aims to determine what objects are present in the scene and where they are located, is one of the most challenging problems in computer vision [1, 2, 3]. It has been successfully addressed in many applications, including content analysis [4], image retrieval [5], image relevance prediction [6] and object-level editing [7]. With the recent advent of large-scale labeled image corpora [8, 9] and region-based convolutional neural networks [10, 11], the research on object detection has made remarkable achievements in recent years.

Nevertheless, many challenges remain when seeking to effectively detect objects in practice. For instance, in cluttered scenes, it is still quite difficult to discriminate objects due to the variance of object's appearance, position, pose, lighting, and background. As shown in Figure 1, the light and shadow is erroneously detected as a lamp and the paint on the wall appears to be a television (see Figure 1c). Fortunately, with the development of consumer-grade depth cameras, such as Microsoft Kinect, Intel RealSense, and Asus Xtion, increasing amounts of depth data offer us additional cues to revisit these problems. Since geometrical and structural properties of the scene are mostly invariant to visual changes, depth information shows powerful benefits in many vision tasks, including salient object detection [12, 13], image segmentation [14, 15] and activity classification [16, 17]. Specifically, in object detection task, as shown in Figures 1b and 1d, we could effortlessly infer several objects (*e.g.*, lamp and bed) from the depth map. The whole object body can be well estimated in this



(a) RGB (b) Refined depth map

(c) RGB object detection results (d) Depth object detection results

Figure 1: An exemplar of RGB image and its corresponding refined depth map (color indicates depths: red is far, blue is near). (c) and (d) highlight the detected objects from modality RGB and depth, respectively. For each bounding-box, one kind of color indicates one object category.

scene even regardless of its RGB map. This is mainly owing to the obvious object boundaries, layered structures and elegant object bodies in the depth map. Meanwhile, the erroneously detected objects (*e.g.*, lamp and television) could also be corrected by the depth map, as shown in Figure 1d. Therefore, we consider introducing the depth information into object detection.

On the other hand, it should be noted that depth is not perfect for general object description. First, the discriminative power of depth decays rapidly when the object depth increases. For example, the depth difference between the upper left dresser and the background is hard to discern in Figure 1b as it is too far from the viewer. Second, depth boundaries only describe the structural properties of objects, which are inadequate to detect objects due to the lack of appearance discrimination. An example is given in Figure 1d that the plant is detected as a dresser because of the similar shapes. Moreover, it is still nontrivial to obtain accurate depth map with the current techniques. The inaccuracy of depth map will inevitably bring in noises in object description. A simple solution is to straightforwardly fuse RGB and depth results. It is, however, not difficult to find such solution is sub-optimal. RGB and depth maps encode different aspects of scenes or objects, and the straightforward fusion is vulnerable to disagreements between RGB and depth results. Motivated by these observations, apart from the complementarity of these two modalities (*i.e.*, the specific constituents), the consistency between RGB and depth modalities (*i.e.*, the correlated ingredients) should be jointly exploited as well. Toward this end, we aim to take full advantages of both the depth and color cues for RGB-D object detection in this study.

The currently dominant object detection paradigm includes two key components: objectness estimation [18, 19] and region-wise object recognition [20, 10]. Objectness estimation generates a sparse set of category-agnostic object proposals[1] in the form of region candidates, which could substantially improve the efficiency and accuracy of the subsequent object classifiers. Moreover, the region-wise recognition accuracy can be further improved by enabling more sophisticated and discriminative classifiers due to the sparse search space. To the best of our knowledge, little attention has been paid to end-to-end object detection that leverages multi-modal information, especially incorporating multi-modal objectness estimation in an end-to-end manner. With object detection as the final goal, objectness estimation procedure could be boosted in return to generate more high-quality and recognition-favorable proposals. For this purpose, we come up with an end-to-end multi-modal multi-task deep learning approach to

[1]In this paper, we use objectness estimation, object proposals, and region proposals interchangeably.

tackle RGB-D object detection. More specifically, we develop modality-correlated and modality-specific deep convolutional neural networks to learn discriminative RGB-D-correlated, RGB-specific, and depth-specific representations for RGB-D object detection. It could simultaneously generate RGB-D region proposals and perform region-wise object recognition. The learning pipeline of the proposed approach is illustrated in Figure 2. We first adopt three-way deep convolutional neural networks (CNNs) to learn features from RGB and depth modalities correlatively and specifically. In particular, the *shared weights* strategy and a new parameter-free *correlation* layer are proposed to learn the modality-correlated representations. At the last convolution layer, the Region Proposal Networks (RPNs) [19] are utilized to predict object proposals. We then feed the learned feature maps with the late-fusion ensemble proposals generated from multi-modal RPNs to the subsequent RGB-D object recognition task. The recognition task has two sibling outputs per proposal: softmax probabilities and per-class bounding-box regression offsets. Finally, we assemble the correlated and specific outputs via late fusion to boost the RGB-D object detection performance. More importantly, by introducing the proposed modality-correlated model, disagreements between modality-specific results could be alleviated.

To evaluate the performance of the proposed approach, we conduct extensive experiments on two RGB-D benchmark datasets: NYU Depth v2 [14] and SUN RGB-D [21]. On these two challenging datasets, we compare the proposed approach to the state-of-the-art RGB-D objectness estimation methods and RGB-D object detection methods. The experimental results show that our proposed approach is superior to the state-of-the-art competing candidates. In summary, the main technical contributions of this study are three-fold:

- We develop a multi-modal deep feature learning approach for RGB-D object detection, which exploits both modality-correlated and modality-specific relationships between RGB and depth images. Notably, disagreements between modality-specific results can be alleviated with the proposed modality-correlated representation learning component.

- We adopt the *shared weights* strategy in the correlated detection network and introduce a parameter-free *correlation* layer to extract the modality-correlated representations. Together with modality-specific representations, the proposed approach provides consistent and significant performance boosts on RGB-D objectness estimation and object detection in terms of recall and mean average precision (mAP), respectively.
- We expand the state-of-the-art object proposal generator to perform multi-modal object detection. In particular, the modality-correlated and modality-specific detection networks are optimized via end-to-end multi-task learning, which can simultaneously generate RGB-D region proposals and perform region-wise RGB-D object recognition.

The remainder of this paper is organized as follows. After reviewing related works on the corresponding fields in Section 2, we describe our multi-modal deep feature learning approach for RGB-D object detection in Section 3. Section 4 presents the experimental results and analyses. The last section concludes this paper with remarks on the future work.

2. Related Work

The goal of this work is to incorporate depth information to multi-modal object detection, which consists of two key components: objectness estimation and object recognition. In this section, we first discuss the representative objectness estimation works briefly, which are mainly performed on traditional RGB images. After that, we will go through RGB-D object recognition and object detection works.

The objectness estimation task aims to generate a moderate number of generic-over-classes object proposals and is expected to cover all objects in an image [18, 22, 23]. According to the object distinctive characteristics, Alexe *et al.* [18] explored five window cues for measuring the objectness, including multi-scale saliency, color contrast, edge density, superpixels straddling, and window location and size. These cues are formulated in a Bayesian framework and each region proposal is assigned an objectness score, which reflects how likely the region covers an object of any category. But this framework takes much time to train and predict. Cheng *et al*. [22] and Zitnick *et al*. [23] tried to assess each potential window with carefully defined objectness scores in near real-time. It is worth noting that, they all share a common idea that the borders or edges of the objects play a much more important role in objectness estimation and should be incorporated into this task. We argue that the depth map provides much more salient object boundaries, layered scene structures and apparent object bodies. To the best of our knowledge, little attention has been paid to adopt the depth information into objectness estimation. Xu *et al*. [24] tried to adaptively integrate RGB and depth information into this task. However, their method is built upon Bing [22], which is optimized for intersection-over-union (IoU) of 0.5 and not well suited for object detectors. In contrast, we leverage not only RGB images but also depth maps to carry out the objectness estimation in this work, which is based on the recent region proposal networks (RPN) [19] and improves the region proposal quality and the overall object detection accuracy in return.

With the powerful deep convolutional neural networks (CNN) [20], recent works on RGB-D object recognition have considered neural networks for learning representations from RGB and depth images [25, 26]. Socher *et al*. [25] and Bo *et al*. [27] focused on recognizing small prop-like RGB-D objects imaged in controlled lab settings. Instead of using the depth image directly, in [28], the authors proposed a geocentric embedding for depth images and tackled RGB-D object detection in cluttered scenes. In the RGB scenario, object detection has witnessed great improvements starting from generic features that are learned on a large-scale RGB image corpus, ImageNet [8]. However, for depth modality, there are no such large amounts of labeled data as ImageNet. Gupta *et al*. [29] utilized learned representations from a large labeled RGB dataset as a supervisory signal for training representations for unlabeled paired depth modality, which provides better parameter initialization for the depth network. However, they either simply treat the depth map as an additional channel of

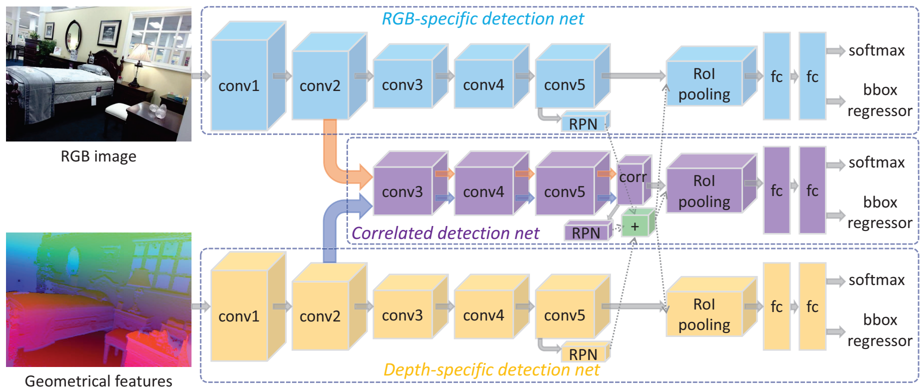


Figure 2: The proposed modality-correlated and modality-specific deep feature learning architecture for RGB-D object detection. In correlated detection net, `conv2` feature maps from the RGB branch and the depth branch will separately go through the same convolutional network (first three layers in the correlated net). After the last convolutional layer, the activations are integrated as RGB-D correlated features using `corr` operation. Each detection network has two outputs per proposal: softmax probabilities and per-class bounding-box regression offsets. This multi-modal object detection approach is trained with end-to-end multi-task learning. For clarity, the *ReLU*, *pooling* and *local response normalization* layers are omitted. The "+" operator denotes that the RGB-D region proposals are boosted from ensemble of modality-correlated and modality-specific RPN results, and the output proposals are fed into `RoI pooling` layers. `conv` stands for convolutional layer, `corr` operator is short for the *correlation* layer and `fc` means fully connected layer.

corresponding RGB image or separately learn representations from RGB and depth modalities. In [30], Wang *et al*. embedded the RGB and depth deep features into a transformed space to learn the shared and specific representations for RGB-D object recognition. In contrast, we aim to take full advantages of both depth and color cues by directly exploiting the modality-correlated and modality-specific deep feature representations for RGB-D object detection in uncontrolled, cluttered environments as in the datasets NYU Depth v2 [14] and SUN RGB-D [21].

3. Proposed Approach

In this section, we describe the proposed multi-modal deep feature learning approach for RGB-D object detection.

Currently, the dominant paradigm for mono-modal object detection contains two key components: objectness estimation (*e.g.*, selective search [31], edge box [23]) and deep feature based region-wise object recognition (*e.g.*, R-CNN [10], Fast R-CNN [11]). Similarly, the RGB-D object detection is broken down into these two sub-tasks. Most of the existing RGB-D object detection methods either simply treat depth map as an additional channel of corresponding RGB image in an undifferentiated way as in [32], or separately learn features from RGB and depth modalities as in [26]. However, neither the intrinsic characteristic of depth information nor the relationship between different modalities can be adequately exploited in such ways. As a result, sub-optimal results are produced. Instead, we employ geocentric encoding of depth map, HHA (Horizontal disparity, Height above ground and Angle with gravity) embedding[2] [28], to capture the scene geometrical features, which emphasize the complementary discontinuities in the image (*i.e.*, depth, surface normal and height) and are proven to be useful in several works [21, 33, 29]. Moreover, motivated by the intuition that different modalities should contain not only modality-specific information but also modality-correlated information [34, 30], we propose to learn correlated features that are shared between RGB and depth modalities as well as specific features that are only captured at each single modality for RGB-D object detection, and the learned modality-correlated features and modality-specific features are complementary to each other. By incorporating the proposed correlated features, disagreements between modality-specific results can be rectified. The pipeline of our proposed approach is depicted in Figure 2.

[2]We use the term depth and HHA interchangeably.

3.1. Multi-modal deep feature learning

In [30], Wang *et al*. employed multi-modal feature learning carried out in conjunction with convolutional neural network feature learning in RGB-D object recognition. They argued that in the transformed feature space, RGB and depth modality should have common parts and individual parts. In contrast, we employ convolutional neural networks to learn discriminative modality-correlated and modality-specific features in an end-to-end manner.

First, we develop a three-way fully convolutional neural network to learn multi-modal deep features as shown in Figure 2, which is explicitly designed to learn RGB-D-correlated, RGB-specific, and depth-specific feature representations. In [35], Li *et al*. found that network-specific features can be learned in multiple networks even with the same modality. More notably, Gupta *et al*. [29] demonstrated that even though the depth network is supervised by the RGB network, the learned features on the depth images are still complementary to the features on the RGB images. Therefore, it is reasonable to assume that the RGB network and depth network (shown in Figure 2) are able to learn modality-specific features with our configuration.

It is well known that the shared weights strategy has been demonstrated very effective in convolutional neural networks. On one hand, it can substantially lower the complexity of the model. Another important aspect is that the shared weights policy is dedicated to detecting the consistent or common patterns at all possible locations [36], which can increase the invariance of the learned features. Inspired by the latter aspect, we make efforts to learn the modality-correlated features through the shared weights policy across RGB and depth modalities. More formally, the shared weights in CNNs correspond to different filters or templates Ws, and for a specific W with inputs x_m ($m \in \{RGB, Depth\}$), the activations (*i.e.*, feature maps) h_m ($m \in \{RGB, Depth\}$) are obtained as follows:

$$h_m = \sigma(W * x_m + b)\,, \tag{1}$$

where $\sigma(\cdot)$ stands for the activation function (*e.g.*, ReLU [20], hyperbolic tangent or sigmoid function), operator $*$ denotes the convolution and b is bias term. When the inputs x_m comprise the similar pattern to W, h_m could be maximized. That is why the shared weights filters are dedicated to detecting different kinds of common patterns. In consequence, we may reasonably interpret that strong feature activations in the correlated network are responded from *similar patches*, which are all similar to W and shared by RGB and depth images or their feature maps. However, due to the hierarchical nature of deep convolutional neural networks, the layered feature maps or vectors reveal progressive properties. Low-level features are shown to be local and activated by edge-like patterns. In contrast, mid-level semantic representations can tell the context information (*e.g.*, texture and shape) and respond to parts of objects. In consideration of this nature, the mid-level semantic representations, instead of raw RGB images and depth maps, are utilized to learn the modality-correlated features via the shared weights manner.

It is easy to see that the learned similar activations h_m ($m \in \{RGB, Depth\}$) in Eq. (1) are not exactly the same because the input x_m are not equivalent. To encourage the network to learn the integrated and correlated representations, we introduce a parameter-free *correlation* layer, which performs *multiplicative comparisons* between similar feature maps of two modalities. Given two feature maps h_{RGB} and h_{Depth}, the correlated feature maps h_{corr} are defined as:

$$h_{corr} = \sqrt{h_{RGB} \circ h_{Depth}}\,, \tag{2}$$

in which $\circ$ denotes the Hadamard product. The multiplicative comparisons only keep the activations occurred both in RGB and depth feature maps, which guarantee that modality-correlated network dedicates to learning consistent and common representations between RGB and depth modalities.

The shared weights strategy comes with several advantages in multi-modal setting. First, as mentioned earlier, the intra-modalities and inter-modalities common patterns could be learned. Moreover, the shared weights in the modality-correlated network enable a favorable alignment between the learned RGB and depth feature maps, which makes the Hadamard product in Eq. (2) reasonable.

In practice, `conv2` feature maps from the RGB branch and the depth branch will separately go through the same convolutional network (first three layers of "correlated" branch in Figure 2). After the last convolutional layer, the activations are integrated as RGB-D correlated features using Eq. (2).

However, to learn modality-correlated features, there is a straightforward approach, *i.e.*, simply treating RGB and depth images or feature maps indistinguishably and concatenating their channels. Its detection performance is significantly worse than ours (46.4% *vs.* 49.5% on the NYU Depth v2 test set). We suspect that straightforward concatenation largely explores the "linear" combination of RGB and depth modalities, while failing to learn discriminative correlated relationship between the two modalities and producing suboptimal results.

3.2. RGB-D objectness estimation

In order to generate multi-modal object proposals, three Region Proposal Networks (RPNs) [19] are slid over the last `conv` feature maps (as shown in Figure 2). One is for modality-correlated objectness estimation and the other two are for modality-specific objectness estimation. Each RPN is performed as a multi-task learning module, which ends up with two sibling 1×1 convolutional layers for binary classification (object or not) and bounding-box regression [1, 11]. Specifically, the binary classification is carried out by a two-class softmax layer, and its sibling layer outputs bounding-box regression deviations. Given an anchor box with (x_a, y_a, w_a, h_a), bounding-box regression is developed to predict deviations $t^* = (t_x^*, t_y^*, t_w^*, t_h^*)$ following [1, 10]:

$$\begin{aligned} t_x^* &= (x^* - x_a)/w_a & t_y^* &= (y^* - y_a)/h_a \\ t_w^* &= \log(w^*/w_a) & t_h^* &= \log(h^*/h_a)\,, \end{aligned} \tag{3}$$

where x, y, w and h denote the bounding box's center coordinates and its width and height. Variables x^* and x_a are for the ground-truth box and anchor box respectively (likewise for y, w, h). The smoothed ℓ_1 loss [11] is adopted as the bounding-box regression loss function.

$$s_{\ell_1}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise}\,. \end{cases} \tag{4}$$

With these definitions, the objectness estimation multi-task loss L is defined as:

$$L(p, p^*, t, t^*) = \lambda L_{cls}(p, p^*) + \sum_{i \in \{x,y,w,h\}} p^* s_{\ell_1}(t_i - t_i^*)\,, \tag{5}$$

where the mini-batch size is ignored. p and p^* are the predicted objectness probability of an anchor and ground-truth label (1 if the anchor is positive, and 0 if the anchor is negative), respectively. Two types of anchors are treated as positive: the anchors with the highest IoU overlap with a ground-truth box, and the ones that have an IoU overlap higher than 0.7 with any ground-truth box [19]. An anchor is considered as negative example if its IoU ratio is lower than 0.3 for all ground-truth boxes. $L_{cls}(p, p^*) = -\log pp^*$ is the standard cross-entropy loss. The modality-correlated RPN and modality-specific RPNs are trained simultaneously with the same supervision. At last, the ensemble object proposal scores and bounding-box deviations are computed from the average of three RPNs predictions.

3.3. Region-wise RGB-D object recognition

With the *recognition using region proposals* framework (*e.g.*, R-CNN [10]), the objects detection capability has been greatly improved. For the recognition networks, we build upon the more recent Fast R-CNN [11]. Similar to the RGB-D objectness estimation, the recognition networks consist of three independent parts: one is modality-correlated and the other two are modality-specific, which are trained separately with the same supervision. Each recognition network simultaneously optimizes two tasks: K-class softmax classification and bounding-box regression. The multi-task loss for object recognition is similar to Eq. (5), except for the number of classes changed from 2 to K, and the bounding-box regression in this stage uses the similar parameterization as Eq. (3). The bounding-box regression in previous RGB-D objectness estimation stage could be considered as differentiating coarse-grained class-agnostic object candidates from chaos, and the latter one in this stage aims to refine the coarse object proposals. Moreover, with object detection as the final goal, previous objectness estimation procedure could be further boosted in return to generate more high-quality and recognition-favorable region proposals.

Likewise in RGB-D objectness estimation, the ensemble detection performance is based on the simple arithmetic average of class probabilities and bounding-box deviations predicted by these three constituent detection networks.

3.4. Training

The proposed multi-modal object detection networks can be trained end-to-end with back-propagation and stochastic gradient descent (SGD) [37]. For RPN networks, each mini-batch arises from a single image that contains many positive and negative example anchors.

During region-wise recognition training, RPNs generates region proposals which are treated as being fixed, *i.e.*, the derivatives with regard to the proposal boxes' coordinates are ignored during back-propagation. Some proposals generated from RPNs highly overlap with each other. To reduce redundancy, non-maximum suppression (NMS) is performed over the proposals according to their ensemble objectness scores with an IoU threshold of 0.7, which leaves about 2,000 proposal regions per image. In each SGD iteration, we uniformly sample 128 positives (≥ 0.5 IoU overlap with a ground-truth box over all classes) and 128 negatives (a maximum IoU with any ground-truth boxes in the interval $[0.1, 0.5)$, following [38]) from the rest of proposals to construct a mini-batch of size 256, which are treated as inputs to the following recognition networks.

4. Experiments and Results

4.1. Dataset

We comprehensively evaluate our algorithm on the NYU Depth v2 [14] and SUN RGB-D [21] benchmark datasets. NYU Depth v2 is comprised of 1,449 densely labeled pairs of aligned RGB and depth images, which are captured by Microsoft Kinect v1. Similarly, SUN RGB-D is comprised of 3,784 Microsoft Kinect v2 images, 3,389 Asus Xtion images, 2,003 Microsoft Kinect v1 images and 1,159 Intel RealSense images. NYU Depth v2 is a subset of SUN RGB-D. Since sensor bias does exist [21], we use these two datasets for evaluation. Due to measurement noises, diffuse or specular reflections, and occlusion boundaries, *etc.*, the depth maps in SUN RGB-D come with missing a significant amount of points. We first fill the missing values with colorization algorithm [39]. Following [28, 29, 21], we only work with 19 major furniture categories available in the datasets: *bathtub, bed, bookshelf, box, chair, counter, desk, door, dresser, garbage bin, lamp, monitor, night stand, pillow, sink, sofa, table, television*, and *toilet.*

4.2. Implementation details

In [19], the authors integrated the RPNs with Fast R-CNN [11], called Faster R-CNN, which is built upon the popular deep learning framework Caffe [40]. The proposed correlation layer can be easily implemented in two steps: element-wise square root followed by element-wise product. Faster R-CNN shares the computation for convolutional layers. Therefore, the cost for object proposal prediction is marginal (*e.g.*, 10ms per image typically). Moreover, the generated object proposals are somewhat adaptive to the subsequent recognition networks.

In addition, due to the GPU memory consumption, we only conduct the experiment on the AlexNet architecture [20] with an NVIDIA GeForce GTX TITAN Black. We fine-tune the proposed multi-modal object detection networks for 70,000 iterations with a base learning rate of 0.001 and reduce it by a factor of 10 after every 40,000 iterations from pre-trained models. The RGB-specific detection network is initialized with ImageNet [8] RGB classification model.[3] To better leverage the depth information, the modality-correlated and depth-specific networks are initialized from a supervision transfer model [29]. All new layers are initialized by drawing weights from a Gaussian distribution $\mathcal{N}(0, 0.01^2)$. A

[3] `https://github.com/BVLC/caffe/wiki/Model-Zoo`

momentum term with a weight of 0.9 and weight decay factor of 0.0005 are used in all experiments. For simplicity, we choose to weight category loss and bounding-box regression loss equally, *i.e.*, the balancing parameter λ in Eq. (5) is set to 1. We follow the default setup for Faster R-CNN that the input images are re-scaled such that their shorter side is $s = 600$ pixels. During training and testing, only the single re-scaled images ($s = 600$) are passed through both region proposal and object recognition networks. For RPN anchors, we use 3 scales with box areas of 128^2, 256^2, and 512^2 pixels, and 3 aspect ratios of $1:1$, $1:2$, and $2:1$ following [19]. During testing, object detection is carried out on the top 2,000 proposals.

4.3. Evaluation metrics

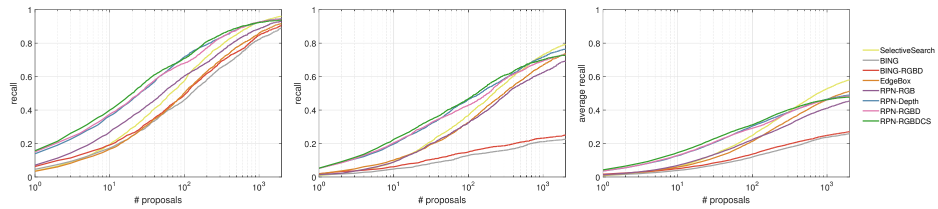
Evaluating class-agnostic object proposals is quite different from the traditional object detection task [41]. It is not practical to evaluate the object proposals' class confusion and background confusion and so forth. Instead, we report *recall* at a particular IoU threshold with a given number of proposals (#PRPSL):

$$recall(\epsilon, \#\text{PRPSL}) = \frac{\#(\text{IoU} \geq \epsilon)@\#\text{PRPSL}}{\#\text{GT}}\,, \tag{6}$$

where IoU is the de facto criterion to determine whether a proposal covers an object. $\epsilon\,(\epsilon \in [0.5, 1])$ is IoU threshold and GT means object ground-truth bounding-boxes. In addition, we also report the average recall (AR) [42] with IoU between 0.5 to 1:

$$\begin{aligned} AR(\#\text{PRPSL}) &= 2\int_{0.5}^{1} recall(\epsilon, \#\text{PRPSL})d\epsilon \\ &= \frac{2}{n}\sum_{i=1}^{n} f(gt_i, \#\text{PRPSL})\,, \end{aligned} \tag{7}$$

where $f(gt_i, \#\text{PRPSL})$ denotes the IoU between the ground-truth annotation gt_i and the best detection proposal with different #PRPSL. When the IoU between the ground-truth annotation gt_i and the best detection proposal is less than 0.5, $f(gt_i, \#\text{PRPSL})$ is set to 0. It has been demonstrated that the average recall correlates surprisingly well with almost all object detectors' performance [42].



(a) Recall at 0.5 IoU (b) Recall at 0.7 IoU (c) AR with IoU $\in [0.5, 1]$

Figure 3: Controlled experiments on the NYU Depth v2 test set. (a) and (b) demonstrate recall versus the number of proposals at different IoU threshold. (c) shows average recall (AR) versus the number of proposals between $[0.5, 1]$ IoU.

As to RGB-D object detection, the commonly used average precision (AP) is adopted to assess the detection performance.

4.4. Experiments on NYU Depth v2

We use the standard splits of 795 training and validation images for training and remaining 654 images for testing. These splits are all carefully selected by making sure images from the same scene do not spread across both sets.

4.4.1. Object proposal evaluation

In our initial experiments, we fine-tune two modality-specific Faster R-CNNs as baselines, RPN-RGB and RPN-Depth. Apart from the modality-correlated networks in Figure 2, the straightforward integration of modality-specific ones, RPN-RGBD, is also treated as a baseline, which leverages the RGB and Depth information in a preliminary way. Furthermore, by taking into account the efficiency of objectness estimation, we compare our approach with the state-of-the-art methods, SS [31], BING [22], EdgeBox [23] and BING-RGBD [24], which all perform reasonably in terms of proposal quality and speed [42]. In all experiments, we adopt the authors' open-source codes with the suggested parameters in their papers.

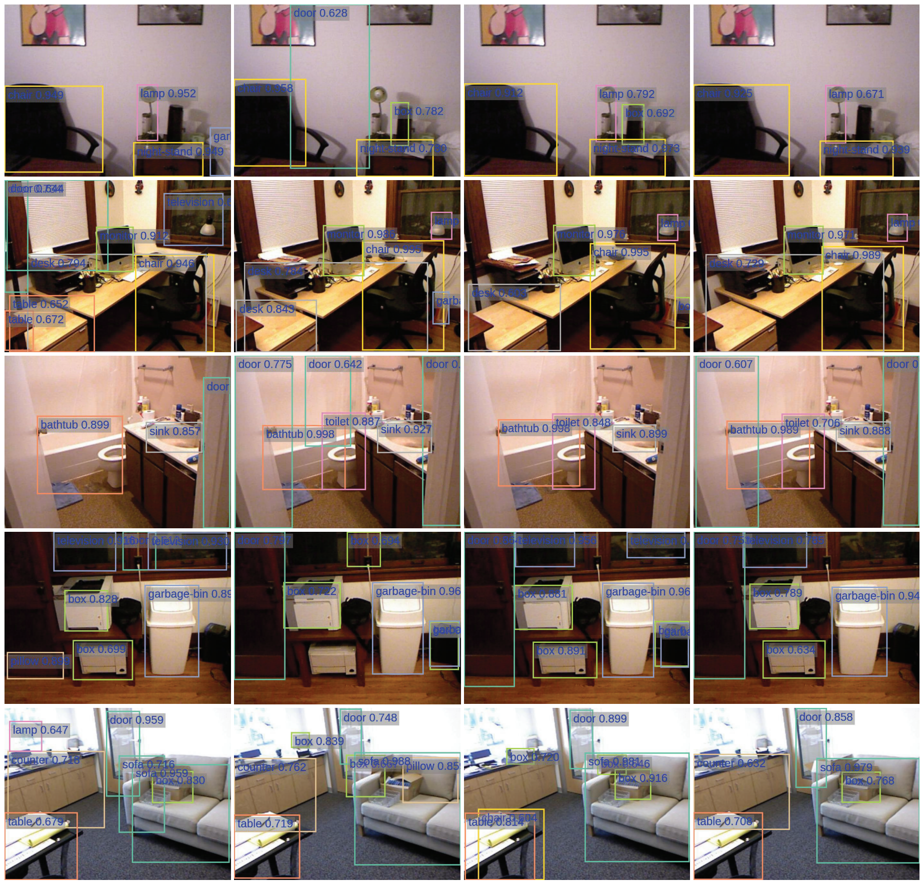
Since Eq. (6) indicates a proposal method's effectiveness, we first evaluate the recall with respect to various numbers of candidate proposals. Figure

3a illustrates the $recall(0.5, \#\text{PRPSL})$ with different truncated numbers of proposals. The proposed modality-correlated and modality-specific approach, RPN-RGBDCS, outperforms both the baselines and the competitors. However, the IoU score above 0.5 is quite loose for objectness estimation, and the detection algorithm may not benefit much from this setup. Therefore, we further report the detection rate at IoU above 0.7, as shown in Figure 3b. Due to the bounding-box regression, RPN-RGBDCS produces much tighter proposals compared with state-of-the-art methods. It is noteworthy that the *recall* metric is more appropriate to diagnose the proposal method and loosely related to the downstream detection accuracy [42, 19]. Therefore, in addition to reporting the recall with different truncated number of proposals, we also highlight the novel metric, average recall (AR), between IoU 0.5 to 1 for a varying number of proposals in Figure 3c. AR summarizes proposal performance across different IoU thresholds, which has proven to be an excellent indicator for downstream object detection performance [42]. As can be seen in Figure 3c, RPN-RGBDCS performs well across the entire range of number of proposals.

Overall, we have shown in this subsection that the proposed RPN-RGBDCS outperforms the existing objectness estimation methods.

4.4.2. Object detection evaluation

In this subsection, we report the performance of the proposed multi-modal object detection results on the NYU Depth v2 test set in Table 1 and compare our performance against the state-of-the-art methods. The proposal methods are utilized to denote the baseline detection methods. RGB Arch., Depth Arch., and RGBD Arch. refer to the CNN architecture used by the modality-specific detectors and modality-correlated detectors, respectively. We can see when using only the depth information, the detection rate is well above that of only using RGB images. We attribute this to the robust characteristics of depth information, which is largely invariant to visual changes. By investigating the consistent and common ingredients between RGB and depth cues, the modality-correlated detector (RPN-corr), predicts more accurate objects. The object detection



(a) RGB (b) Depth (c) RGB-D Correlation (d) RGBDCS

Figure 4: Detection results of examples from the NYU Depth v2 test set, comparing on different detection networks. (a) RGB-specific detection results. (b) Depth-specific detection results. (c) RGB-D correlated detection results. (d) Ensemble detection results from the proposed modality-specific and modality-correlated detection networks. Each detected box is associated with a category label and a softmax score in $[0, 1]$. A score threshold with 0.6 is used to display these images. For each bounding-box, one kind of color indicates one object category.

performance can be significantly boosted from the late fusion of modality-specific detectors (from 41.5% to 47.3%). This also holds true for the state-of-the-art competitor [29]. Moreover, the detection results can be further rectified by incorporating the additional modality-correlated recognition network. Figure 4 illustrates some examples evaluated on each detection network. Regions with similar appearance are easily misclassified (*e.g.*, televisions and lamp in Figure 4 (a)). In contrast, depth maps generally predict more precise object locations. The detection results can be improved by exploring modality-correlated features (as shown in Figure 4 (c) and (d)).

The proposed multi-modal RGB-D object detection approach is substantially superior to the state-of-the-art competitor, supervision transfer [29], in terms of mAP with the normal deep architecture (AlexNet [20]). Surprisingly, compared to Gupta *et al.*'s supervision transfer VGG model [29], our approach, which is built upon AlexNet, performs slightly better. Moreover, it is noteworthy that the proposed modality-correlated and modality-specific object detection approach is built upon an almost cost-free objectness estimation. In comparison, supervision transfer [29] employed a prohibitively time-consuming object proposal method, RGBD MCG [28], which typically takes about 30s for a 500×400 image.

4.4.3. Ensemble strategy

In practice, we find training three independent detection models all the way and then assembling the correlated and specific outputs (class scores and bounding-box deviations) via late fusion perform better than early fusion, which is also verified in [29]. Moreover, to investigate the relative importance amongst the three branches, we conduct an experiment to perform weighted averaging instead of simple averaging among the three branches on the NYU Depth v2 validation set as follows:

$$G(x) = \alpha g_{RGB}(x) + \beta g_{Depth}(x) + (1 - \alpha - \beta) g_{corr}(x)\,. \tag{8}$$

where $g(\cdot)$ is the output of detection network, $\alpha, \beta (\alpha \geq 0, \beta \geq 0, \alpha + \beta \leq 1)$ are the ensemble weights for RGB and depth branch, respectively. α, β vary in $[0, 1]$ with a step size of 0.05. The ensemble results are shown in Figure 5. $\alpha = 0.3, \beta = 0.35$ give the best result (36.7%). Furthermore, we also experiment stacking strategy to learn a meta-learner based on three branches' outputs, which is not better than simple averaging either. We suspect that the weighted averaging and meta-learner are prone to overfitting and are not always superior

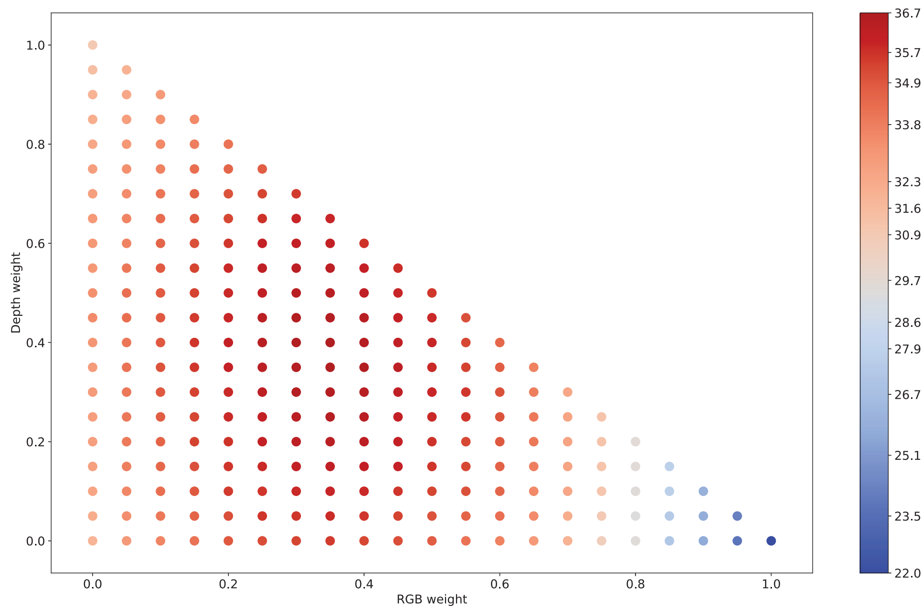


Figure 5: Weighted averaging results of the modality-correlated, RGB-specific and depth-specific networks on the NYU Depth v2 validation set. Warmer colors correspond to larger values, cooler colors are small values. $\alpha = 0.3, \beta = 0.35$ give the best result (36.7%).

to simple averaging [43, 44, 45]. Therefore, we use simple averaging in the following experiments. In addition, these experiments also imply that apart from the modality-specific constituents, the correlated ingredients are another complementary view of multi-modal data and should be jointly exploited as well.

4.4.4. Control experiments on ensemble of multiple detection networks

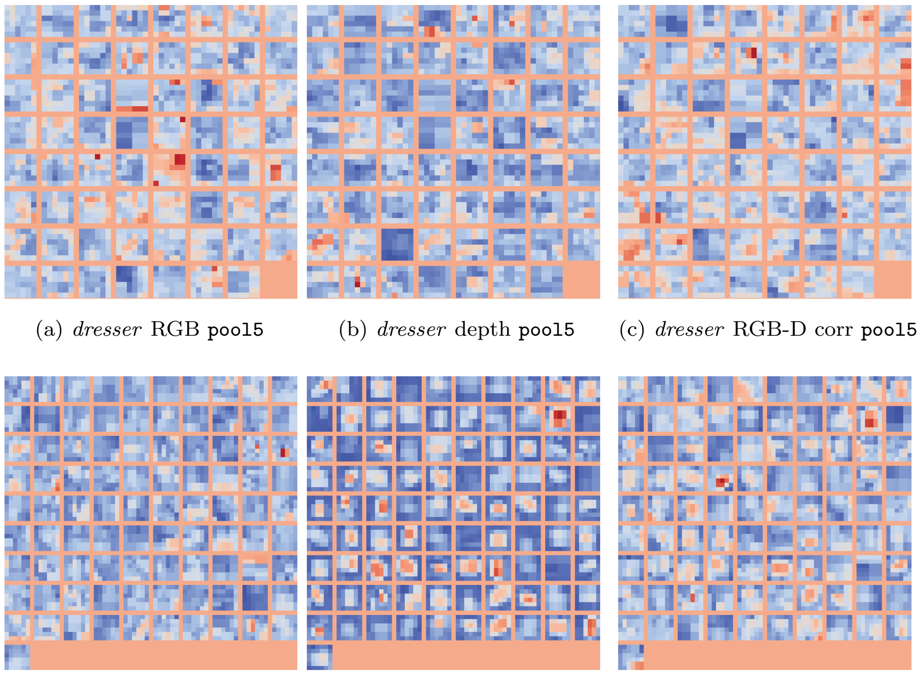
There exists a suspicion that most of the detection performance gain comes from the ensemble of multiple detection models rather than from the learned modality-correlated representation. To better understand the effects of adding modality-correlated detection network, we perform control experiments on ensemble of multiple detection models. With same experimental settings as in Section 4.2, we fine-tune modality-specific detection networks twice resulting two color and two depth detection networks. The ensemble detection rates are 46.2% and 48.2% for two color + one depth detection networks and one color + two depth detection networks respectively, which are much lower than the proposed detection approach. Even with two color + two depth detection networks (*i.e.*, four detection models), the achieved object detection performance is 48.7%. It is 0.8% worse than ours, which only relies on three detection models. We attribute that disagreements between RGB-specific and depth-specific object detection results can be rectified with the additional proposed modality-correlated model. In consequence, these control experiments imply that the proposed detection approach may take full advantages from the developed modality-correlated and modality-specific feature representations, and performs more effectively and powerfully than the straightforward combination with just modality-specific features dose, which is vulnerable to disagreements between modality-specific object detection results.

4.4.5. Features visualization

Figure 6 shows the `pool5` feature maps from RGB-specific, depth-specific, and modality-correlated networks for each *dresser* and *sink* in the NYU Depth v2 trainval set. [4] We adopt the "aggregation map" [46] to visualize the resulted $6 \times 6 \times 256$ `pool5` feature maps, which is proven effective in fine-grained image retrieval. More specifically, `pool5` features are aggregated via global average pooling across the channels to produce 6×6 object pool5 descriptors. It is worth noting that although the weights in the correlated detection network and depth-specific detection network were initialized from the same source, the learned features are diverse after fine-tuning (as illustrated in Figures 6b, 6c and 6e, 6f). Furthermore, different regions are activated in modality-correlated and modality-specific feature maps. In other words, the correlated, RGB-specific, and depth-specific detection nets are dedicated to covering different aspects of an object.

In addition, to give an overview visualization of the learned multi-modal

[4] More object pool5 features of the NYU Depth v2 trainval set can be visualized from `http://mcg.nju.edu.cn/dataset/pool5/`.



(a) *dresser* RGB `pool5` (b) *dresser* depth `pool5` (c) *dresser* RGB-D corr `pool5`

(d) *sink* RGB `pool5` (e) *sink* depth `pool5` (f) *sink* RGB-D corr `pool5`

Figure 6: `pool5` feature maps from RGB-specific, depth-specific, and modality-correlated networks for each *dresser* and *sink* in the NYU Depth v2 trainval set (best viewed in color). Each object's $6 \times 6 \times 256$ `pool5` maps are aggregated via the "aggregation map" [46] to produce 6×6 pool5 descriptors. The features are normalized to $[0, 1]$. Warmer colors correspond to larger values, cooler colors are small values.

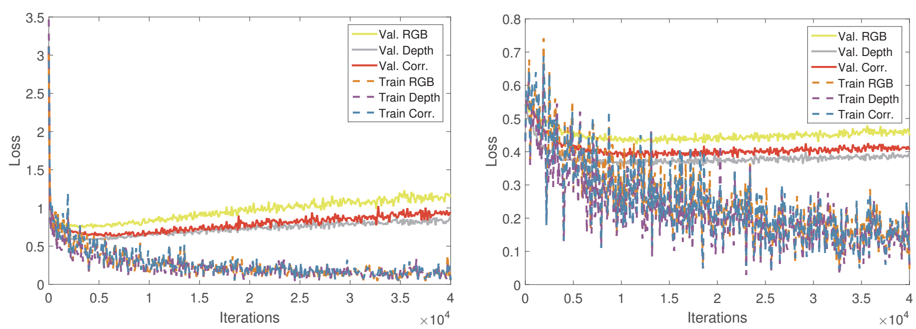
features for each object, we employ a high-dimensional data visualization technique, t-distributed stochastic neighbor embedding (t-SNE) [47], to map the learned high-dimensional features to two-dimensional locations. We can obtain a rough idea about the feature space's topology through t-SNE, because it is capable of retaining the local structure [47]. Figure 7 depicts `pool5` object features from different networks. These features are extracted from the NYU Depth v2 trainval set. We believe that the object feature distributions in RGB-specific, depth-specific, and modality-correlated feature spaces are essentially different by investigating the differences of object feature distributions. In particular,

(a) RGB `pool5` features (b) Depth `pool5` features (c) RGB-D corr `pool5` features

Figure 7: t-SNE [47] embedded `pool5` features from RGB-specific, depth-specific, and modality-correlated networks (best viewed in color). One kind of color indicates one object category.

differences of inter-object feature distributions between object categories and intra-object feature distributions in each category demonstrate that the feature's implicit structure varies in these three feature spaces. We can draw a conclusion that the developed components in Figure 2 are dedicated to exploring different aspects of the RGB-D data from Figure 6 and Figure 7. Consequently, the complementarity among features can considerably benefit the proposed multi-modal RGB-D object detection approach.

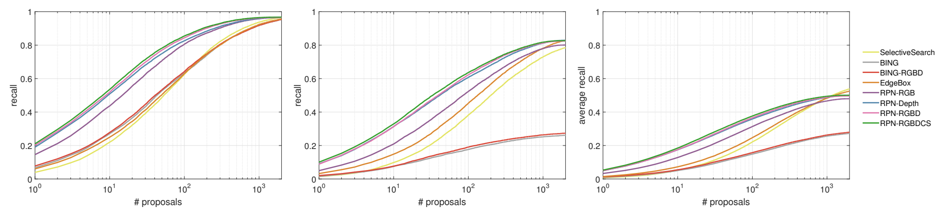
4.4.6. Convergent rate for different modality networks



(a) Softmax loss (b) Bounding-box regression loss

Figure 8: The training loss and validation error for modality-correlated and modality-specific RGB-D object detection networks, which are evaluated on the NYU Depth v2 training and validation set.

The proposed modality-correlated and modality-specific RGB-D object



(a) Recall at 0.5 IoU (b) Recall at 0.7 IoU (c) AR with IoU $\in [0.5, 1]$

Figure 9: Controlled experiments on the SUN RGB-D test set. (a) and (b) demonstrate recall versus the number of proposals at different IoU threshold. (c) shows average recall (AR) versus the number of proposals between $[0.5, 1]$ IoU.

detection networks are simultaneously optimized. However, the convergent rate for these three networks may vary. In this subsection, we examine the convergent speed for modality-correlated and modality-specific networks on the NYU Depth v2 training and validation set. The training softmax, bounding-box regression loss and validating error are shown in Figures 8a and 8b, respectively. Note that the tendency of convergence is very close for optimizing modality-correlated network and modality-specific networks. Therefore, when training the proposed multi-modal object detection approach, we choose the same learning rate and loss weights for different networks.

4.5. Experiments on SUN RGB-D

SUN RGB-D [21] is a very recent PASCAL VOC [48] scale RGB-D dataset, which is a superset of NYU Depth v2. This data set consists of RGB-D image pairs captured by various RGB-D sensors. Song *et al.* pointed out that sensor bias does exist due to the diverse capabilities for different devices [21]. It is crucial that an algorithm can generalize to different types of RGB-D sensors, because real data usually come from different sensors. For this reason, we also present extensive experimental results on this much more challenging dataset. However, SUN RGB-D consists of RGB-D image pairs captured by Intel RealSense, whose effective range for reliable depth is very short. Besides, we found that its depth map quality is too low for use in the accurate object detection task. Therefore, we leave out the RGB-D images captured by Intel RealSense and adopt the remaining standard splits in following experiments: 4,698 for training and 4,478 for testing. These splits are also carefully selected as suggested in [21]. There are a few minor changes of our system made for this dataset. First, SUN RGB-D consists of RGB-D captured by several devices, thus the modality-correlated and modality-specific networks are all fine-tuned from pre-trained ImageNet RGB classification models. Second, SUN RGB-D is a much larger dataset, thus it is trained for 100,000 iterations with a step size of 50,000.

4.5.1. Object proposal evaluation

Under the same protocol as in Section 4.4, we first evaluate the object proposal performance with the same experimental setup to the NYU Depth v2 dataset, as shown in Figure 9. The modality-correlated and modality-specific objectness estimation method, RPN-RGBDCS, consistently performs better than the baselines and the state-of-the-art competitors, which indicates that the proposed RPN-RGBDCS can be well generalized to different types of RGB-D devices. Better proposals do matter for better object detection performance [11]. In the following, we will see that the high-quality and recognition-favorable proposals generated from the modality-correlated and modality-specific objectness estimation models can benefit downstream object detection task.

4.5.2. Object detection evaluation

Next, we evaluate the detection performance on the SUN RGB-D dataset. Compared to the NYU Depth v2 dataset, RGB detection performance is greatly improved with more training object examples in SUN RGB-D. The detection performance gap between RGB and depth models is not as significant as on NYU Depth v2. We conjecture that this is because the scenes in SUN RGB-D are much more diverse. Object poses and relative object positions vary much more. Consequently, it is much more difficult to detect the objects with only depth maps. The modality-correlated detection network (RPN-corr) and late fusion of modality-specific detection networks (RPN-RGBD), which both take advantages of RGB and depth modalities, perform much better than mono-modality detection networks. Likewise, the detection performance can be further improved by the proposed modality-correlated and modality-specific detection networks (from 51.8% to 52.9%). The detailed numbers are reported in Table 2. Figure 10 shows some detection results on the SUN RGB-D test set returned from the proposed multi-modal object detection approach.

With the convolutional features shared for proposal generation and region-wise recognition, the proposed modality-correlated and modality-specific RGB-D object detection approach takes a total of $\sim$ **0.290s** for a RGB-D image pair, which is much more efficient than the supervision transfer [29].

4.6. From SUN RGB-D to NYU Depth v2

A large-scale labeled dataset is of crucial importance for improving the performance of object detection. In this subsection, we investigate how the SUN RGB-D dataset can help improving the detection performance on the NYU Depth V2 dataset.

As the original training and testing splits from NYU Depth V2 are kept in SUN RGB-D, we first directly evaluate the trained SUN RGB-D detection models on the NYU Depth V2 test set without fine-tuning. The mAP under this setting is 47.5%, which is lower than the performance fine-tuned from the ImageNet and supervision transfer models (49.5%), as shown in Table 1. We attribute this to the scene and device biases. We then fine-tune the SUN RGB-D detection models on the NYU Depth v2 trainval set. In this experiment, the trained SUN RGB-D models are in place of the ImageNet model and supervision transfer model, and are used to initialize the weights in multi-modal detection networks. The networks are fine-tuned for 40,000 iterations with a step size of 15,000. Doing so leads to 53.0% mAP on the NYU Depth v2 test set. The extra data from the SUN RGB-D set increase the mAP by 3.5%. Details are reported in the last row of Table 1.

Figure 10: SUN RGB-D test detection results returned from the proposed modality-correlated and modality-specific RGB-D object detection networks. Each detected box is associated with a category label and a softmax score in $[0, 1]$. A score threshold with 0.6 is used to display these images. For each bounding-box, one kind of color indicates one object category.

5. Conclusions and Future Work

We have presented a multi-modal deep feature learning approach for RGB-D object detection. More specifically, our method allows learning modality-correlated and modality-specific feature representations. The shared weights strategy and a parameter-free correlation layer are employed to learn modality-correlated features. In order to demonstrate the effectiveness of the learned modality-correlated and modality-specific feature representations, we have conducted extensive experimental analyses in RGB-D objectness estimation and RGB-D object detection tasks. Experimental results on two challenging standard datasets, NYU Depth v2 and SUN RGB-D, show that the proposed approach outperforms the state-of-the-art competitors, and confirms the benefits for joint consideration of modality-correlated and modality-specific components in RGB-D object detection.

Our experimental results show consistent improvements in overall detection accuracy (mAP). However, for some categories, the improvements are not as noticeable. It will be an interesting future direction to study the specific impact of depth information on various object classes.

Acknowledgment

This work is primarily supported by the National Science Foundation of China under Grant No.61321491 and No.61202320, Research Project of Excellent State Key Laboratory under Grant No.61223003 and the Collaborative Innovation Center of Novel Software Technology and Industrialization, and partially by the New York State through the Goergen Institute for Data Science at the University of Rochester.

References

[1] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, TPAMI 32 (9) (2010) 1627–1645.

[2] Y. Li, S. Wang, Q. Tian, X. Ding, Feature representation for statistical-learning-based object detection: A review, PR 48 (11) (2015) 3542–3559.

[3] E. Ohn-Bar, M. M. Trivedi, Multi-scale volumes for deep object detection and localization, PR 61 (2016) 557–572.

[4] T. Chen, F. X. Yu, J. Chen, Y. Cui, Y.-Y. Chen, S.-F. Chang, Object-based visual sentiment concept analysis and application, in: MM, ACM, 2014, pp. 367–376.

[5] Y. Jiang, J. Meng, J. Yuan, J. Luo, Randomized spatial context for object search, TIP 24 (6) (2015) 1748–1762.

[6] Y. Yang, L. Yang, G. Wu, S. Li, Image relevance prediction using query-context bag-of-object retrieval model, TMM 16 (6) (2014) 1700–1712.

[7] J. Xue, L. Wang, N. Zheng, G. Hua, Automatic salient object extraction with contextual cue and its applications to recognition and alpha matting, PR 46 (11) (2013) 2874–2889.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: CVPR, IEEE, 2009, pp. 248–255.

[9] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: ECCV, Springer, 2014, pp. 740–755.

[10] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: CVPR, IEEE, 2014, pp. 580–587.

[11] R. Girshick, Fast r-cnn, in: ICCV, IEEE, 2015, pp. 1440–1448.

[12] H. Peng, B. Li, W. Xiong, W. Hu, R. Ji, Rgbd salient object detection: A benchmark and algorithms, in: ECCV, Springer, 2014, pp. 92–109.

[13] R. Ju, Y. Liu, T. Ren, L. Ge, G. Wu, Depth-aware salient object detection using anisotropic center-surround difference, Signal Processing: Image Communication 38 (2015) 115–126.

[14] N. Silberman, D. Hoiem, P. Kohli, R. Fergus, Indoor segmentation and support inference from rgbd images, in: ECCV, Springer, 2012, pp. 746–760.

[15] R. Ju, X. Xu, Y. Yang, G. Wu, Stereo grabcut: Interactive and consistent object extraction for stereo images, in: PCM, Springer, 2013, pp. 418–429.

[16] O. Oreifej, Z. Liu, Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences, in: CVPR, IEEE, 2013, pp. 716–723.

[17] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, A. Blake, Real-time human pose recognition in parts from single depth images, in: CVPR, IEEE, 2011, pp. 1297–1304.

[18] B. Alexe, T. Deselaers, V. Ferrari, Measuring the objectness of image windows, TPAMI 34 (11) (2012) 2189–2202.

[19] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in: NIPS, MIT Press, 2015, pp. 91–99.

[20] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: NIPS, MIT Press, 2012, pp. 1097–1105.

[21] S. Song, S. P. Lichtenberg, J. Xiao, Sun rgb-d: A rgb-d scene understanding benchmark suite, in: CVPR, IEEE, 2015, pp. 567–576.

[22] M.-M. Cheng, Z. Zhang, W.-Y. Lin, P. H. S. Torr, BING: Binarized normed gradients for objectness estimation at 300fps, in: CVPR, IEEE, 2014.

[23] C. L. Zitnick, P. Dollár, Edge boxes: Locating object proposals from edges, in: ECCV, Springer, 2014, pp. 391–405.

[24] X. Xu, L. Ge, T. Ren, G. Wu, Adaptive integration of depth and color for objectness estimation, in: ICME, IEEE, 2015, pp. 1–6.

[25] R. Socher, B. Huval, B. Bath, C. D. Manning, A. Y. Ng, Convolutional-recursive deep learning for 3d object classification, in: NIPS, MIT Press, 2012, pp. 665–673.

[26] A. Wang, J. Lu, J. Cai, T.-J. Cham, G. Wang, Large-margin multi-modal deep learning for rgb-d object recognition, TMM 17 (11) (2015) 1887–1898.

[27] L. Bo, X. Ren, D. Fox, Unsupervised feature learning for rgb-d based object recognition, in: ISER, Springer, 2013, pp. 387–402.

[28] S. Gupta, R. Girshick, P. Arbeláez, J. Malik, Learning rich features from rgb-d images for object detection and segmentation, in: ECCV, Springer, 2014, pp. 345–360.

[29] S. Gupta, J. Hoffman, J. Malik, Cross modal distillation for supervision transfer, in: CVPR, IEEE, 2016, pp. 2827–2836.

[30] A. Wang, J. Cai, J. Lu, T.-J. Cham, Mmss: Multi-modal sharable and specific feature learning for rgb-d object recognition, in: ICCV, IEEE, 2015, pp. 1125–1133.

[31] J. R. Uijlings, K. E. van de Sande, T. Gevers, A. W. Smeulders, Selective search for object recognition, IJCV 104 (2) (2013) 154–171.

[32] C. Couprie, C. Farabet, L. Najman, Y. Lecun, Convolutional nets and watershed cuts for real-time semantic labeling of rgbd videos, JMLR 15 (1) (2014) 3489–3511.

[33] S. Song, J. Xiao, Deep sliding shapes for amodal 3d object detection in RGB-D images, in: CVPR, IEEE, 2016, pp. 808–816.

[34] J. Liu, Y. Jiang, Z. Li, Z.-H. Zhou, H. Lu, Partially shared latent factor learning with multiview data, TNNLS 26 (6) (2015) 1233–1246.

[35] Y. Li, J. Yosinski, J. Clune, H. Lipson, J. Hopcroft, Convergent learning: Do different neural networks learn the same representations?, in: ICLR, 2015.

[36] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE 86 (11) (1998) 2278–2324.

[37] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, L. D. Jackel, Backpropagation applied to handwritten zip code recognition, Neural computation 1 (4) (1989) 541–551.

[38] K. He, X. Zhang, S. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, in: ECCV, Springer, 2014, pp. 346–361.

[39] A. Levin, D. Lischinski, Y. Weiss, Colorization using optimization, TOG 23 (3) (2004) 689–694.

[40] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: Convolutional architecture for fast feature embedding, in: arXiv preprint arXiv:1408.5093, 2014.

[41] D. Hoiem, Y. Chodpathumwan, Q. Dai, Diagnosing error in object detectors, in: ECCV, Springer, 2012, pp. 340–353.

[42] J. Hosang, R. Benenson, P. Dollár, B. Schiele, What makes for effective detection proposals?, TPAMI 38 (4) (2015) 814–830.

[43] T. K. Ho, J. J. Hull, S. N. Srihari, Decision combination in multiple classifier systems, TPAMI 16 (1) (1994) 66–75.

[44] J. Kittler, M. Hatef, R. P. Duin, J. Matas, On combining classifiers, TPAMI 20 (3) (1998) 226–239.

[45] Z.-H. Zhou, Ensemble methods: foundations and algorithms, CRC press, 2012.

[46] X.-S. Wei, J.-H. Luo, J. Wu, Z.-H. Zhou, Selective convolutional descriptor aggregation for fine-grained image retrieval, TIP 26 (6) (2017) 2868–2881.

[47] L. v. d. Maaten, G. Hinton, Visualizing data using t-sne, JMLR 9 (Nov) (2008) 2579–2605.

[48] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, IJCV 88 (2) (2010) 303–338.

Table 1: Object detection average precision (%) on the NYU Depth v2 test set.

method	RGB Arch.	Depth Arch.	RGBD Arch.	bathtub	bed	bookshelf	box	chair	counter	desk	door
RPN-RGB	AlexNet	-	-	20.0	57.2	33.8	2.0	33.2	34.3	12.6	16.4
RPN-Depth	-	AlexNet	-	34.9	80.3	41.8	2.3	53.0	49.3	18.1	17.2
RPN-Corr	-	-	AlexNet	41.7	82.7	43.5	3.2	53.5	52.0	18.3	17.2
RPN-RGBD	AlexNet	AlexNet	-	46.5	79.8	49.1	3.1	53.9	55.0	23.8	23.5
Super. Trans. [29]	AlexNet	AlexNet	-	45.6	78.7	48.5	4.3	50.5	57.8	21.4	29.6
Super. Trans. [29]	**VGG16**	AlexNet	-	50.6	81.0	52.6	5.4	53.0	56.1	20.9	34.6
Ours	AlexNet	AlexNet	AlexNet	49.1	82.7	50.4	4.4	57.5	56.8	22.3	25.0
SUN w/o fine-tune	AlexNet	AlexNet	AlexNet	61.3	79.2	46.2	5.6	53.6	12.7	22.0	23.6
SUN w/ fine-tune	AlexNet	AlexNet	AlexNet	61.9	85.7	50.5	6.2	60.5	59.1	27.3	26.0

method	dresser	garbage-bin	lamp	monitor	night-stand	pillow	sink	sofa	table	television	toilet	mAP
RPN-RGB	24.6	21.6	23.1	33.0	20.8	22.0	30.1	31.6	18.0	32.1	56.6	27.5
RPN-Depth	49.2	29.4	38.5	42.6	49.1	40.2	47.9	63.1	34.0	34.6	62.7	41.5
RPN-Corr	54.5	29.4	39.7	39.1	50.7	37.4	51.1	63.9	30.9	45.1	63.9	43.0
RPN-RGBD	56.1	44.3	44.1	62.2	49.3	43.7	46.3	64.8	33.6	54.2	65.5	47.3
Super. Trans. [29]	54.0	41.6	45.4	61.2	57.9	47.3	48.9	63.2	29.5	50.0	60.1	47.1
Super. Trans. [29]	57.9	46.2	42.5	62.9	54.7	49.1	50.0	65.9	31.9	50.1	68.0	49.1
Ours	57.9	42.4	45.2	59.8	60.7	44.1	55.6	65.0	33.4	55.6	71.9	49.5
SUN w/o fine-tune	44.8	43.2	45.0	68.7	61.7	47.7	51.8	72.0	34.1	58.3	70.4	47.5
SUN w/ fine-tune	54.2	51.1	48.4	74.7	60.2	48.8	62.2	68.4	38.2	57.3	66.3	53.0

Table 2: Object detection average precision (%) on the SUN RGB-D test set.

method	RGB Arch.	Depth Arch.	RGBD Arch.	bathtub	bed	bookshelf	box	chair	counter	desk	door
RPN-RGB	AlexNet	-	-	36.4	67.0	34.8	6.2	43.3	45.3	17.5	22.0
RPN-Depth	-	AlexNet	-	55.8	77.7	26.7	5.0	49.9	47.6	21.8	6.9
RPN-Corr	-	-	AlexNet	50.7	78.9	36.1	8.7	53.1	47.9	24.2	21.8
RPN-RGBD	AlexNet	AlexNet	-	70.3	79.2	35.1	11.0	53.5	50.6	25.1	18.8
RGBD RCNN [28]	AlexNet	AlexNet	-	49.6	76.0	35.0	5.8	41.2	8.1	16.6	4.2
Ours	AlexNet	AlexNet	AlexNet	70.5	80.8	38.4	11.4	54.3	51.5	26.5	22.2

method	dresser	garbage-bin	lamp	monitor	night-stand	pillow	sink	sofa	table	television	toilet	mAP
RPN-RGB	38.3	57.8	42.0	51.6	43.6	31.6	46.2	57.2	37.1	20.1	74.7	40.7
RPN-Depth	31.3	54.0	46.4	50.6	45.0	47.0	50.5	67.8	43.3	12.9	74.2	42.9
RPN-Corr	46.6	65.5	54.4	60.6	51.2	47.6	61.9	69.3	45.7	34.9	78.5	49.3
RPN-RGBD	48.0	67.1	57.5	62.0	56.5	49.4	62.3	71.9	47.4	34.7	83.4	51.8
RGBD RCNN [28]	31.4	46.8	22.0	10.8	37.2	16.5	41.9	42.2	43.0	32.9	69.8	33.2
Ours	49.2	68.1	58.0	63.4	56.6	50.8	63.1	72.1	48.4	38.5	81.6	52.9