Semi-Supervised Multimodal Deep Learning for RGB-D Object Recognition

Yanhua Cheng[1*], **Xin Zhao**[1], **Rui Cai**[2], **Zhiwei Li**[2], **Kaiqi Huang**[1,3], **Yong Rui**[2]
[1]CRIPAC&NLPR, CASIA [2]Microsoft Research
[3]CAS Center for Excellence in Brain Science and Intelligence Technology
{yh.cheng, xzhao, kaiqi.huang}@nlpr.ia.ac.cn, {ruicai, zli, yongrui}@microsoft.com

Abstract

This paper studies the problem of RGB-D object recognition. Inspired by the great success of deep convolutional neural networks (DCNN) in AI, researchers have tried to apply it to improve the performance of RGB-D object recognition. However, DCNN always requires a large-scale annotated dataset to supervise its training. Manually labeling such a large RGB-D dataset is expensive and time consuming, which prevents DCNN from quickly promoting this research area. To address this problem, we propose a semi-supervised multimodal deep learning framework to train DCNN effectively based on very limited labeled data and massive unlabeled data. The core of our framework is a novel diversity preserving co-training algorithm, which can successfully guide DCNN to learn from the unlabeled RGB-D data by making full use of the complementary cues of the RGB and depth data in object representation. Experiments on the benchmark RGB-D dataset demonstrate that, with only 5% labeled training data, our approach achieves competitive performance for object recognition compared with those state-of-the-art results reported by fully-supervised methods.

1 Introduction

Recent years have witnessed RGB-D object recognition becoming a very active research area in computer vision and robotics with the rapid development of commodity depth cameras. Such off-the-shelf sensors, *e.g.*, Microsoft Kinect and Intel RealSense, are capable of providing high quality synchronized RGB and depth information, to depict multimodal characteristics of an object. Specifically, the RGB modality captures rich colors and textures, while the depth modality provides pure geometry and shape cues which are robust to lighting and color variations. It represents an opportunity to dramatically improve the performance of object recognition by combining the two complementary cues.

Remarkable efforts have been invested for RGB-D object recognition in the last few years. Most existing work falls into two kinds. One is about feature representation, including handcrafted features [Lai *et al.*, 2011a; Bo *et al.*, 2011a; R.C. *et al.*, 2012] and learning-based features [Blum *et al.*, 2012; Bo *et al.*, 2012; Socher *et al.*, 2012; Jhuo *et al.*, 2015]. The other one is about RGB-D fusion, like straightforward concatenation of RGB and depth features as well as learning-based fusion [Lai *et al.*, 2011b; Cheng *et al.*, 2015a]. Towards building a unified solution for feature learning and RGB-D fusion, a promising trend is to devise an end-to-end deep learning system via convolutional neural networks (DCNN) [Gupta *et al.*, 2014; Eitel *et al.*, 2015; Wang *et al.*, 2015], such as the one shown in Fig. 1 (a). Such ideas were inspired by the great success of deep learning for image classification (only RGB data). It should be noticed that DCNN models always require a large-scale dataset for supervised training, *e.g.*, ImageNet with millions of annotated images [Deng *et al.*, 2009]. However, labeling such a large dataset for the emerging RGB-D object recognition task is still expensive and time consuming. This prevents DCNN from quickly promoting this research area. Thus it is necessary to develop a new effective training framework for deep learning to benefit from the massive unlabeled RGB-D data, which is often cheap and easily available.

To handle the aforementioned problem, a natural idea is to incorporate the conventional semi-supervised learning methods into the deep learning framework. Although many successful semi-supervised methods exist in the literature [Zhu, 2005], we are particularly interested in the co-training algorithm due to its unique advantage over the multimodal data. Theoretical proofs have been given in [Blum and Mitchell, 1998; Balcan *et al.*, 2004] to guarantee the success of co-training in learning from the unlabeled data on condition that: 1) each example contains two views, either of which is able to depict the example well; and 2) the two views should not be highly correlated. RGB-D data matches the two conditions well by providing two complementary cues of objects (i.e., RGB and depth). Therefore, the goal of this paper is to develop a semi-supervised multimodal deep learning framework based on co-training, as shown in Fig. 1 (b).

The pipeline of the framework can be summarized as follows. First, the RGB- and depth-DCNN models are trained on the given labeled data of the respective views. Then each model is applied to predict the unlabeled pool and label the most confident examples for the other model, for which these

[*]The work was performed at Microsoft Research.

examples are random and informative to increase its capability through the next round training. The two steps are repeated until no confident examples can be chosen for each other. Finally, we add a fusion layer to combine the two stream networks for recognition and jointly train the whole model.

Although the proposed framework looks quite straightforward, it is not a trivial task to make it work. In fact, starting the framework directly doesn't show any inspiring results in our experiments. There are two obstacles during the training of the framework. One is about the initial phase. Such a limited labeled set is hard to provide a good deep learning model for either the RGB or the depth modality due to overfitting, even though each model can be pretrained based on other datasets like ImageNet and then finetuned on the RGB-D object recognition task. The other is about the co-training phase. Since each DCNN model selects those most confident examples from every predicted class, it is prone to result in a biased distribution over each category in the labeled pool along with co-training, e.g., almost all "apples" will be red but few are green, and the "cups" with handles will be dominant compared to those without handles, meaning that the intra-class diversity of each category is fading. As a result, the final DCNN models trained on the imbalanced labeled set have poor generalization ability for category-level object recognition on the unseen data.

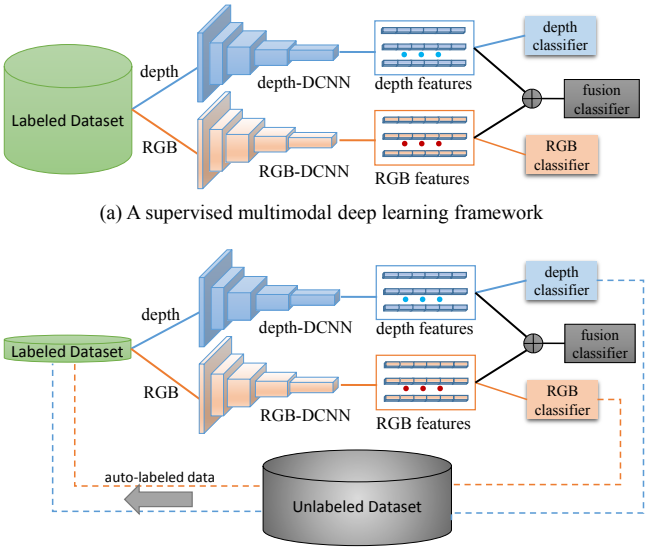
Two strategies are proposed in this paper to address the involved problems. First, we devise two reconstruction networks to better initialize the RGB- and depth-DCNN models for object recognition separately. The reconstruction networks make use of both the labeled and unlabeled data for unsupervised feature learning, which can help to relieve the over-fitting problem effectively. Second, we introduce a diversity preserving co-training algorithm to balance the added samples from the unlabeled pool. To this end, we adopt the convex clustering [Lashkari and Golland, 2007] to automatically discover various intra-class attributes over each category, and then keep the added samples to uniformly cover every attribute of every category during the iterations. Such informative and balanced samples can boost the RGB- and depth-DCNN models during every round training. We demonstrate the effectiveness of the two strategies in the experiments.

The rest of this paper is organized as follows. Section 2 briefly reviews related work. Section 3 introduces the proposed semi-supervised multimodal deep learning framework. Experimental results and detailed analysis are reported in Section 4. In Section 5, we finally draw our conclusions.

2 Related Work

RGB-D Object Recognition. Many successful methods have been proposed for RGB-D object recognition in recent years. Here we review those state-of-the-art supervised as well as semi-supervised approaches evaluated on the benchmark RGB-D datasets.

Supervised Methods. *Early work* can be divided into two groups. One was focused on feature extraction of the novel RGB-D data, including handcrafted features [Lai *et al.*, 2011a] (such as SIFT and spin images), a series of appearance and shape kernel descriptors [Bo *et al.*, 2011a], and automatically learning features via successful machine learning methods [Blum *et al.*, 2012; Bo *et al.*, 2012; Socher *et al.*, 2012; Jhuo *et al.*, 2015]. The other tried to explore a more effective way for RGB and depth fusion [Lai *et al.*, 2011b; Cheng *et al.*, 2015a] instead of a direct feature concatenation in the first group. Both the two groups of work were followed by SVM or random forest classifier, fully supervised by all the training data for better object recognition. *Very recently*, researches began to jointly learn the features, classifiers and RGB-D fusion using end-to-end deep learning [Gupta *et al.*, 2014; Eitel *et al.*, 2015; Wang *et al.*, 2015]. Due to the lack of a large scale annotated RGB-D object dataset, they spared no efforts to augment the data, e.g., synthesizing objects via CAD rendering, as well as generating new samples via geometrical transformations. Compared to the real data, the artificial data inevitably has a different distribution, and is hard to provide the same rich information to depict object categories. Thus the potential of deep learning to improve RGB-D object recognition is limited.



(a) A supervised multimodal deep learning framework

(b) Our semi-supervised multimodal deep learning framework

Figure 1: The structures of (a) supervised and (b) semi-supervised multimodal deep learning for RGB-D object recognition. Both (a) and (b) jointly learn the features and classifiers in an end-to-end fashion.

Semi-Supervised Methods. The most similar work to ours is that of [Cheng *et al.*, 2014; 2015c], who also employed co-training for RGB-D object recognition to reduce the dependence on large annotated training sets. However, they only adopted co-training to retrain the RGB and depth SVM classifiers based on the features extracted in advance. Different from them, this paper proposes a powerful semi-supervised deep learning method, which can jointly learn the features, classifiers and RGB-D fusion by making use of the unlabeled data. Experimental results show that our method achieves much better performance.

Semi-Supervised Deep Learning. [Weston *et al.*, 2012] pro-

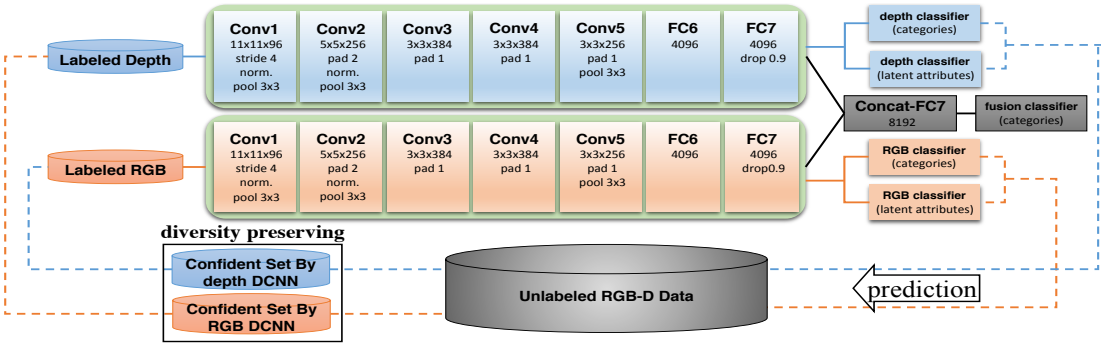


Figure 2: Overview of our semi-supervised multimodal deep learning framework for RGB-D object recognition. The training of the framework mainly contains two iterative steps: 1) Training the RGB- and depth-DCNN models over the respective data based on the labeled pool (indicated as solid lines); 2) Applying each model to predict the unlabeled pool and select the most confident samples over each category for the other, whilst keeping these newly labeled examples to preserve intra-class diversity (indicated as dashed lines. See details in the text.). After all iterations are completed, we add a fusion classification layer and optimize the entire model end-to-end. Note that all the classifiers are implemented with a hingeloss layer. Best viewed in color.

posed a semi-supervised deep learning method for the single-modal data by embedding a pairwise loss in the middle layer. However, they required extra information about whether a pair of unlabeled images belong to the same class, which were hard to obtain in reality. Differently, we focus on the multi-modal RGB-D object recognition, and propose a more explicit semi-supervised deep learning framework to learn from the unlabeled data via diversity preserving co-training.

3 Our Approach

3.1 Overview

We target on learning a powerful semi-supervised multimodal deep learning model for RGB-D object recognition based on limited labeled data and massive unlabeled data. To be specific, we have a small labeled pool $\mathcal{L} = \{(\mathcal{I}_1, \mathcal{D}_1, y_1), \cdots, (\mathcal{I}_M, \mathcal{D}_\mathrm{M}, y_M)\}$ with M pairwise RGB-D objects , where $\mathcal{I}_i$ and $\mathcal{D}_i$ denote the corresponding RGB and depth modalities of the i-th example with the category label $y_i \in \{1, \cdots, C\}$. Meanwhile, we have a large-scale unlabeled RGB-D dataset $\mathcal{U} = \{(\mathcal{I}_1, \mathcal{D}_1), \cdots, (\mathcal{I}_N, \mathcal{D}_\mathrm{N})\}$ with similar data distribution of the labeled pool. The proposed framework in this paper is shown in Fig. 2, for which a diversity preserving co-training algorithm is introduced to learn from the unlabeled RGB-D data. Now we detail three important phases of the framework.

Initialization. A well trained RGB- as well as depth-DCNN model before co-training is the first prerequisite of the whole system. This paper adopts the architecture of AlexNet [Krizhevsky *et al.*, 2012] to represent both the RGB and depth data. However, the small labeled set $\mathcal{L}$ is infeasible to supervise the training of such deep learning models for object recognition. To address this problem, we devise two reconstruction networks (Section 3.2) to initialize the convolution layer parameters (i.e., conv1, ...,conv5) of RGB- and depth-DCNN models, respectively. Each reconstruction network tries to encode and decode its inputs, taking advantage of all the labeled and unlabeled data to learn meaningful features. After pretrained by the corresponding reconstruction network, the RGB- and depth-DCNN models finetuned on $\mathcal{L}$ can generalize well for object recognition.

Training. The training of the semi-supervised deep learning framework mainly involves two iterative steps, including training each DCNN model and updating the labeled pool as illustrated in Fig. 2. For clarity, we denote the state of the system at the t-th iteration as $\mathcal{L}_t, \mathcal{U}_t$. To select an informative and balanced set $\mathcal{H}_t = \{\mathcal{H}_t^{RGB}, \mathcal{H}_t^{depth}\} = \{(\mathcal{I}_i, \mathcal{D}_i, \tilde{y}_i)\}$ ($\tilde{y}_i$ is the predicted category label) from $\mathcal{U}_t$ to update the next round train of the deep learning models effectively, a diversity preserving co-training algorithm (Section 3.3) is introduced.

The goal of the diversity preserving algorithm is to make sure that $\mathcal{H}_t$ captures as diverse intra-class attributes as possible for each category of each modality. To this end, the convex clustering [Lashkari and Golland, 2007] is utilized to discover latent attributes over each category of each modality based on $\mathcal{L}_t$, and then gives a RGB as well as a depth attribute tag for each object, i.e., the labeled pool can be recorded as $\mathcal{L}_t = \{(\mathcal{I}_i, z_i^{RGB}, \mathcal{D}_i, z_i^{depth}, y_i)\}$, where the attribute tag $z_i^{RGB} \in \{1, \cdots, |\mathcal{Z}^{RGB}|\}$, $z_i^{depth} \in \{1, \cdots, |\mathcal{Z}^{depth}|\}$. Note that $\mathcal{Z}^{RGB}$ (or $\mathcal{Z}^{depth}$) is an attribute set integrating all the generated attributes over all categories of the RGB (or depth) modality. Compared to those non-convex clustering methods like k-means, convex clustering is guaranteed to converge to the global minimum and automatically finds the optimal number of clusters given a temperature-like parameter. Such a characteristic is important for our method to search for the unknown representative attributes for each category. Now we train an extra attribute classifier for each modality, which can help the category classifier to select a diversity preserving confident set $\mathcal{H}_t$, consisting of uniform samples over each attribute of each category. It is noted that all the classifiers in Fig. 2 are implemented with a hingeloss layer.

When no confident examples can be selected from $\mathcal{U}_t$, the iteration stops. Finally, we add a fusion layer and train the whole network end-to-end based on the resulting model of the last iteration.

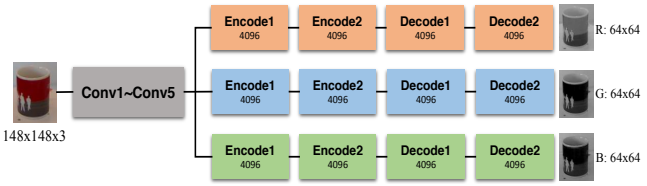


Figure 3: The reconstruction network for the RGB modality, which is the same for the depth modality (We compute 3-channel surface normals to represent the depth data).

Inference. Given an unseen RGB-D object, we utilize the final RGB model, depth model, and the fusion model to predict the category label, respectively. In the experiments, we compare the performance of our method with the state of the arts evaluated on each modality as well as the both.

3.2 Reconstruction Networks for Pretraining

The architecture of our reconstruction network for each modality is shown in Fig. 3, which consists of 5 convolutional layers (with the same structure of the convolutional layers in Fig. 2) and 12 fully connected layers to decode each channel of the inputs. It is noted that the depth data is represented as 3-channel surface normals in this paper, since researches [Bo *et al.*, 2012; Cheng *et al.*, 2015c] have demonstrated that surface normals can capture more robust geometry cues of object than the original depth data. For simplicity, we still use the term "depth" instead of the surface normals in this paper.

Both the labeled and unlabeled data are utilized to train the reconstruction network of each modality. Specifically, the input of the network is a rescaled RGB or depth image $x \in \mathbb{R}^{148\times148\times3}$. The corresponding output is a reconstructed map $R(x) \in \mathbb{R}^{64\times64\times3}$ with downsampled resolution due to memory and computational loss. We train the network by minimizing the mean square reconstruction error

$$Loss_R^v = \frac{1}{M+N} \sum_{x\in\{\mathcal{L}^v,\mathcal{U}^v\}} \sum_{ch=1}^{3} \left\| \tilde{x}^{ch} - R^{ch}(x) \right\|^2, \quad (1)$$

where M, N is the size of the labeled and unlabeled pool in the beginning, ch denotes the channel index of the input, and the modality v represents RGB or depth data. $\tilde{x} \in \mathbb{R}^{64\times64\times3}$ is the ground truth by resizing x via bilinear interpolation.

We use the standard back-propagation algorithm based on stochastic gradient descent (SGD) to optimize the reconstruction network. When the network of each modality achieves convergence, the parameters of the convolutional layers (conv1, ..., conv5 in Fig. 3) are utilized to initialize the corresponding convolutional layers of each modality in the proposed framework of Fig. 2. The experiments will show that such a pretraining strategy is able to largely improve the generalization ability of RGB- and depth-DCNN models for object recognition, even though very limited labeled samples in $\mathcal{L}$ are available to supervise the training.

3.3 Diversity Preserving Co-Training

The goal of the diversity preserving co-training algorithm is to select highly confident examples with predicted labels from the unlabeled pool, whilst keeping these examples uniformly cover each category, as well as each intra-class attribute for each category. Such newly labeled examples by one DCNN model can greatly boost the performance of the other one in the next round training. Now we introduce the three main components of the iterative algorithm. For clarity, sometimes we omit the iteration number t in equations below.

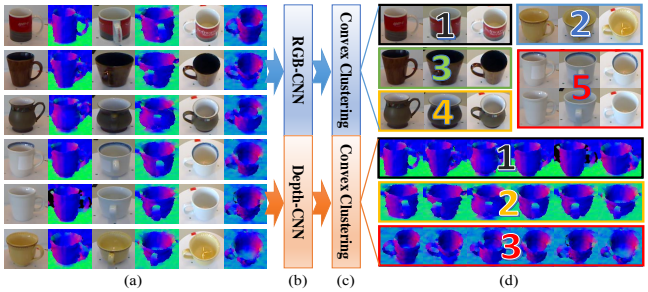


Figure 4: A sketch map illustrates the convex clustering over the category "coffee mug" in the labeled pool. For RGB modality, it finds 5 latent attributes, while 3 attributes for the depth modality (fc7 features are used for convex clustering).

Convex Clustering. We apply convex clustering [Lashkari and Golland, 2007] to discover the diverse intra-class attributes for each category of each modality independently based on the labeled pool $\mathcal{L}$. Convex clustering solve the clustering problem by maximizing the following log-likelihood function

$$\begin{aligned} & l(\{q_x\}; \mathcal{L}_c^v) = \tfrac{1}{|\mathcal{L}_c^v|} \sum_{x\in\mathcal{L}_c^v} \log[\sum_{x'\in\mathcal{L}_c^v} q_{x'} e^{-\beta d_\phi(x,x')}] \\ & s.t. \sum_{x\in\mathcal{L}_c^v} q_x = 1,\ q_x \geq 0 \end{aligned} \quad (2)$$

where the category $c \in \{1, \cdots, C\}$, the modality v denotes the RGB or depth data, and $d_\phi(x, x') = ||\phi(x) - \phi(x')||_2$ is the Euclidean distance between the DCNN features of two examples (fc7 is used). The scalar weight q_x denotes the representative degree of exemplar x, while β is a positive temperature-like parameter that controls the sparseness of $\{q_x\}$ ($q_x > 0$ indicates that x is a cluster center, and $q_x = 0$ denotes an exemplar). Given β, we follow [Lashkari and Golland, 2007] to optimize the likelihood function for each category of each modality individually. Due to space limits, please refer [Lashkari and Golland, 2007] for more details. As a result, we obtain two cluster sets

$$\begin{aligned} \mathcal{Z}^{RGB} &= \{\mathcal{Z}_1^{RGB}, \cdots, \mathcal{Z}_C^{RGB}\} \\ \mathcal{Z}^{depth} &= \{\mathcal{Z}_1^{depth}, \cdots, \mathcal{Z}_C^{depth}\}, \end{aligned} \quad (3)$$

where $\mathcal{Z}_c^v$ is a subset that contains all the clusters generated on category c of modality v. Fig. 4 illustrates how convex clustering is applied to discover latent attributes for the category "coffee mug" for example.

Multitask Learning. For each modality, we define each cluster as an attribute, and assign every exemplar to its closest cluster to tag the same attribute label. Now we train the DCNN model for both category and attribute recognition, as shown in Fig. 2. The loss function of the multitask learning

for each modality is

$$Loss^v_{MT} = \sum_{x \in \mathcal{L}^v} \max(0, 1 - y\varphi^v_{cat}(x)) + \lambda \sum_{x \in \mathcal{L}^v} \max(0, 1 - z\varphi^v_{attr}(x)), \quad (4)$$

where y, z denote the ground truth category label and attribute label for exemplar x of modality v, while φ^v_{cat}, φ^v_{attr} are the corresponding predicted probabilities of the DCNN model. We fix the coefficient $\lambda = 1$ in the experiments.

Co-Training. Finally, the two well-trained attribute DCNN models are utilized to predict the unlabeled pool $\mathcal{U}$ over the respective modalities. A highly confident set $\mathcal{H}^v$ for RGB or depth data can be selected as follows:

$$\mathcal{H}^v = \bigcup\nolimits_{z \in \mathcal{Z}^v} \{(x, z) | score^v_{attr}(z|x) > \tau, x \in \mathcal{U}^v\}, \quad (5)$$

where $score^v_{attr}(z|x) = f(\varphi^v_{attr}(z|x))$ is the predicted score for x carrying the attribute z via a softmax function f, and τ is a score threshold. To further keep the data balance and accuracy, we only reserve the top K examples with highest scores for each attribute in $\mathcal{H}^v$. Then for each remaining exemplar $x \in \mathcal{H}^v$, we are easy to obtain its category label y based on the predicted attribute label z according to Eq.(3). Now we attach both the RGB and depth data to x and update the labeled pool as

$$\mathcal{L}_{t+1} = \mathcal{L}_t \bigcup \mathcal{H}^{RGB} \bigcup \mathcal{H}^{depth}, \quad (6)$$

where every exemplar in $\mathcal{L}_{t+1}$ still contains pairwise RGB-D data with a category label (the attribute labels will be updated by the next round convex clustering). During the next round training, $\mathcal{H}^{RGB}$ can greatly improve the depth-DCNN model as they are new and informative labeled samples, which is the same to $\mathcal{H}^{depth}$ for RGB-DCNN.

4 Experiments

4.1 Experimental Setup

Dataset. We perform our experiments on the Washington RGB-D dataset [Lai *et al.*, 2011a] captured by Microsoft Kinect. The dataset consists of 300 household objects, grouped into 51 categories. Each object is imaged from 3 vertical angles as well as multiple horizontal angles, resulting roughly 600 images per object. We subsample every 5th frame from each instance and obtain around 41,877 images in total for category recognition.

To evaluate our semi-supervised learning, we first utilize one of the 10 random splits provided by [Lai *et al.*, 2011a] to divide the dataset into a training set and a testing set. For any split, there are around 35,000 examples for training and around 6,877 for testing. Then we randomly labeled 5% samples (around 1750) of the training set, and remain the rest unlabeled (around 33,250). Finally, we train our model based on both the labeled and unlabeled data in the training set, and evaluate its performance on the testing set.

Besides semi-supervised methods, we also compare our approach to those existing powerful supervised methods, for which all the objects in the training set are manually labeled to train their classifiers. All the experiments are repeated 10 times based on the given 10 splits, and the average accuracies are reported for comparison.

Table 1: Comparison of recent results on the Washington RGB-D object database for category recognition.

Supervised Methods	**Depth**	**RGB**	**Combine**
[Lai *et al.*, 2011a]$^{\text{linear svm}}$	53.1 ± 1.7	74.3 ± 3.3	81.9 ± 2.8
[Lai *et al.*, 2011a]$^{\text{kernel svm}}$	64.7 ± 2.2	74.5 ± 3.1	83.8 ± 3.5
[Lai *et al.*, 2011a]$^{\text{random forest}}$	66.8 ± 2.5	74.7 ± 3.6	79.6 ± 4.0
[Lai *et al.*, 2011b]$^{\text{IDL}}$	70.2 ± 2.0	78.6 ± 3.1	85.4 ± 3.2
[R.C. *et al.*, 2012]$^{\text{3D SPMK}}$	67.8	–	–
[Bo *et al.*, 2011a]$^{\text{KDES}}$	78.8 ± 2.7	77.7 ± 1.9	86.2 ± 2.1
[Blum *et al.*, 2012]$^{\text{CKM}}$	–	–	86.4 ± 2.3
[Bo *et al.*, 2011b]$^{\text{HMP}}$	70.3 ± 2.2	74.7 ± 2.5	82.1 ± 3.3
[Bo *et al.*, 2012]$^{\text{SP+HMP}}$	81.2 ± 2.3	82.4 ± 3.1	87.5 ± 2.9
[Socher *et al.*, 2012]$^{\text{CNNRNN}}$	78.9 ± 3.8	80.8 ± 4.2	86.8 ± 3.3
[Schwarz *et al.*, 2015]$^{\text{CNN}}$	–	83.1 ± 2.0	89.4 ± 1.3
[Jhuo *et al.*, 2015]$^{\text{R}^2\text{ICA}}$	83.9 ± 2.8	85.7 ± 2.7	89.6 ± 3.8
[Eitel *et al.*, 2015]$^{\text{FusCNN(HHA)}}$	83.0 ± 2.7	84.1 ± 2.7	91.0 ± 1.9
[Eitel *et al.*, 2015]$^{\text{FusCNN(jet)}}$	83.8 ± 2.7	84.1 ± 2.7	91.3 ± 1.4
[Cheng *et al.*, 2015a]$^{\text{warping}}$	–	–	**92.7** ± 1.0
[Wang *et al.*, 2015]$^{\text{NMSS}}$	75.6 ± 2.7	74.6 ± 2.9	88.5 ± 2.2
[Cheng *et al.*, 2015b]$^{\text{CFK}}$	**85.8** ± 2.3	**86.8** ± 2.2	91.2 ± 1.5
Semi-Supervised Methods	**Depth**	**RGB**	**Combine**
[Cheng *et al.*, 2014]$^{\text{CT+SVM}_1}$	71.8 ± 0.8	77.1 ± 2.3	81.6 ± 1.4
[Cheng *et al.*, 2015c]$^{\text{CT+SVM}_2}$	75.4 ± 2.4	78.7 ± 1.4	83.7 ± 1.3
our approach	82.6 ± 2.3	85.5 ± 2.0	89.2 ± 1.3

Parameter Setting of Our Approach. We fix $\tau = 0.5$, $K = 20$, $\beta = 1$ for our semi-supervised learning method, although dynamically finetuning each parameter could result in a better performance. For the reconstruction network of each modality, we use a mini-batch $b = 128$ of images and initial learning rate $\eta = 10^{-5}$, multiplying the learning rate by 0.1 at every $s = 4000$ iterations. Towards the training of the RGB- and depth-DCNN models for recognition during every iteration, we set $b = 128$, $\eta = 10^{-7}$, and $s = 3000$. It is noted that we only apply convex clustering for the first 5 iterations in consideration of efficiency, and then keep the attribute labels unchanged for the rest iterations (around 400 attributes for RGB, and 280 attributes for depth at last).

4.2 Overall Performance

Table 1 presents the recognition accuracies of all recent methods. We can find that: 1) with only 5% labeled data, our method can achieve very promising result on each modality (depth: 82.6%, RGB: 85.5%, both: 89.2%), demonstrating the effectiveness of the proposed semi-supervised multi-modal deep learning framework based on diversity preserving co-training algorithm; 2) Compared to other semi-supervised methods [Cheng *et al.*, 2014; 2015c], which adopted the co-training algorithm directly to retrain the RGB- and depth-SVM classifiers iteratively based on the extracted features in advance, our end-to-end deep learning system shows a large improvement (nearly 7% increase over each modality) for ob-

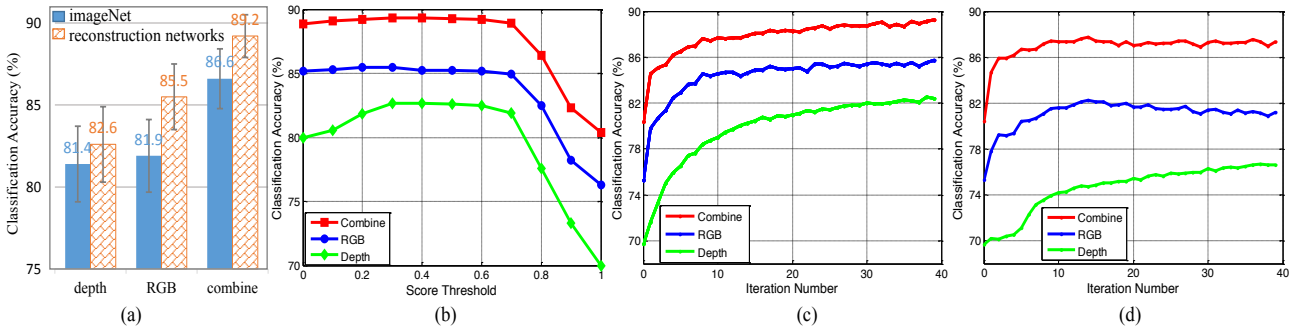


Figure 5: Performance analysis of the semi-supervised multimodal deep learning framework. See details in the text.

ject recognition; 3) Our method is comparable to all state-of-the-art supervised methods, except for the one of [Cheng *et al.*, 2015a], who employed dense matching to obtain a query adaptive similarity measure for RGB-D object recognition. Despite of the best performance, their algorithm is very time-consuming due to plentiful dense matching required for one query object, restricting their potentials in practical use. Our method is efficient in testing, since the recognition only needs one forward propagation of the network.

Furthermore, we also run our method on the fully supervised training data, and the average results are 84.0% accuracy for depth, 86.3% for RGB, and 91.3% for both, which is a little superior to our method based on only 5% labeled data. The results further demonstrate that our semi-supervised learning approach is able to make use of the unlabeled data very effectively. Readers may doubt why the deep learning model does not surpass the traditional methods like fisher kernel encoding [Cheng *et al.*, 2015b] and the dense matching methods [Cheng *et al.*, 2014]. We think the main reason is the scale of the Washington dataset, which is relatively very small compared to ImageNet [Deng *et al.*, 2009]. If given more labeled or unlabeled RGB-D data, we believe our method can achieve much higher performance for object recognition.

4.3 Detailed Analysis

In this section, we analyze the effects of the reconstruction networks, the score threshold τ, the diversity preserving co-training and the initial labeled size to the performance of our semi-supervised multimodal deep learning framework. Note that we evaluate each of them by keeping others the same to the experimental settings in Section 4.1.

The effect of the reconstruction networks. To demonstrate the effectiveness of our reconstruction networks for pretraining, we compare it with a popular pretraining skill, which utilizes the AlexNet model [Krizhevsky *et al.*, 2012] pretrained on imageNet to initialize the parameters of both the RGB- and depth-DCNN models. As shown in Fig. 5 (a), the reconstruction networks can better boost the performance of our semi-supervised learning. We explain that, compared to the knowledge learned from other domains like ImageNet, the reconstruction network trained on the RGB-D data is able to learn more proper cues for RGB-D object representation.

The effect of the score threshold τ. As shown in Fig. 5 (b), our semi-supervised learning is robust to τ when $0.3 < \tau < 0.7$. Such a characteristic is very important in practical usage since a wide range of τ can keep the algorithm successful. When τ is smaller than 0.3, it drops a little because some unconfident examples (probably with wrongly predicted labels) can be involved to distract the next round supervised training. When τ is larger than 0.7, the performance of our method begins to drop quickly. It is reasonable, since very few examples can be added to the labeled pool to benefit the network training. Note that we always constrain that the added confident examples of each attribute is no more than $K = 20$ for balance and accuracy, whatever the value of τ is set.

The effect of the diversity preserving co-training. As shown in Fig. 5 (c), our diversity preserving co-training algorithm can significantly increase the capability of each DCNN model along with iterations. When we employ the conventional co-training algorithm without diversity preserving constraint, the improvements are much inferior, as shown in Fig. 5 (d). We explain that the conventional co-training is prone to result in a biased labeled pool, which limits the potential of each DCNN model a lot.

The effect of the initial labeled size. When the initial labeled size of the training set is changed from 1% to 10%, the recognition accuracies of our method are increased from (68.3%, 79.3%, 84.4%) to (83.8%, 86.1%, 91.2%). Adding more labeled examples do not show obvious improvements, as 10% labeled data is already sufficient for our method to learn from the unlabeled data successfully.

5 Conclusion

This paper proposes a semi-supervised multimodal deep learning framework for RGB-D object recognition, which is capable of reducing the dependence of deep learning method on large-scale manually labeled RGB-D data. The key to the framework are two parts: 1) the reconstruction networks for good initialization and 2) the diversity preserving co-training algorithm for effective semi-supervised learning. Experimental results on the Washington RGB-D benchmark dataset demonstrate the effectiveness of our approach.

Acknowledgments

This work is funded by the National Basic Research Program of China (Grant No. 2012CB316302), National Natural Science Foundation of China (Grant No. 61322209), the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant XDA06040102).

References

[Balcan *et al.*, 2004] Maria-Florina Balcan, Avrim Blum, and Ke Yang. Co-training and expansion: Towards bridging theory and practice. In *NIPS*, 2004.

[Blum and Mitchell, 1998] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *COLT*, 1998.

[Blum *et al.*, 2012] Manuel Blum, Jost Tobias Springenberg, Jan Wulfing, and Martin Riedmiller. A learned feature descriptor for object recognition in rgb-d data. In *ICRA*, 2012.

[Bo *et al.*, 2011a] Liefeng Bo, Xiaofeng Ren, and Dieter Fox. Depth kernel descriptors for object recognition. In *IROS*, 2011.

[Bo *et al.*, 2011b] Liefeng Bo, Xiaofeng Ren, and Dieter Fox. Hierarchical matching pursuit for image classification: architecture and fast algorithms. In *NIPS*, 2011.

[Bo *et al.*, 2012] Liefeng Bo, Xiaofeng Ren, and Dieter Fox. Unsupervised feature learning for rgb-d based object recognition. *ISER, June*, 2012.

[Cheng *et al.*, 2014] Yanhua Cheng, Xin Zhao, Kaiqi Huang, and Tieniu Tan. Semi-supervised learning for rgb-d object recognition. In *ICPR*, 2014.

[Cheng *et al.*, 2015a] Yanhua Cheng, Rui Cai, Chi Zhang, Zhiwei Li, Xin Zhao, Kaiqi Huang, and Yong Rui. Query adaptive similarity measure for rgb-d object recognition. In *ICCV*, 2015.

[Cheng *et al.*, 2015b] Yanhua Cheng, Rui Cai, Xin Zhao, and Kaiqi Huang. Convolutional fisher kernels for rgb-d object recognition. In *3DV*, 2015.

[Cheng *et al.*, 2015c] Yanhua Cheng, Xin Zhao, Kaiqi Huang, and Tieniu Tan. Semi-supervised learning and feature evaluation for rgb-d object recognition. *Computer Vision and Image Understanding*, 139:149–160, 2015.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[Eitel *et al.*, 2015] Andreas Eitel, Jost Tobias Springenberg, Luciano Spinello, Martin Riedmiller, and Wolfram Burgard. Multimodal deep learning for robust rgb-d object recognition. *IROS*, 2015.

[Gupta *et al.*, 2014] Saurabh Gupta, Ross Girshick, Pablo Arbeláez, and Jitendra Malik. Learning rich features from rgb-d images for object detection and segmentation. In *ECCV*, 2014.

[Jhuo *et al.*, 2015] I-Hong Jhuo, Shenghua Gao, Liansheng Zhuang, DT Lee, and Yi Ma. Unsupervised feature learning for rgb-d image classification. In *ACCV*, 2015.

[Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[Lai *et al.*, 2011a] Kevin Lai, Liefeng Bo, Xiaofeng Ren, and Dieter Fox. A large-scale hierarchical multi-view rgb-d object dataset. In *ICRA*, 2011.

[Lai *et al.*, 2011b] Kevin Lai, Liefeng Bo, Xiaofeng Ren, and Dieter Fox. Sparse distance learning for object recognition combining rgb and depth information. In *ICRA*, 2011.

[Lashkari and Golland, 2007] Danial Lashkari and Polina Golland. Convex clustering with exemplar-based models. In *NIPS*, 2007.

[R.C. *et al.*, 2012] Carolina R.C., Roberto J. Lopez-Sastre, Javier Acevedo-Rodriguez, and Saturnino Maldonado-Bascon. Surfing the point clouds: selective 3d spatial pyramids for category-level object recognition. In *CVPR*, 2012.

[Schwarz *et al.*, 2015] Max Schwarz, Hannes Schulz, and Sven Behnke. Rgb-d object recognition and pose estimation based on pre-trained convolutional neural network features. In *ICRA*, 2015.

[Socher *et al.*, 2012] Richard Socher, Brody Huval, Bharath Bath, Christopher D Manning, and Andrew Ng. Convolutional-recursive deep learning for 3d object classification. In *NIPS*, 2012.

[Wang *et al.*, 2015] Anran Wang, Jianfei Cai, Jiwen Lu, and Tat-Jen Cham. Mmss: Multi-modal sharable and specific feature learning for rgb-d object recognition. In *ICCV*, 2015.

[Weston *et al.*, 2012] Jason Weston, Frédéric Ratle, Hossein Mobahi, and Ronan Collobert. Deep learning via semi-supervised embedding. In *Neural Networks: Tricks of the Trade*, pages 639–655. Springer, 2012.

[Zhu, 2005] Xiaojin Zhu. Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison, 2005.