Object Recognition with and without Objects

Zhuotun Zhu, Lingxi Xie, Alan Yuille
Johns Hopkins University, Baltimore, MD, USA
{zhuotun, 198808xc, alan.l.yuille}@gmail.com

Abstract

While recent deep neural networks have achieved a promising performance on object recognition, they rely ***implicitly*** on the visual contents of the whole image. In this paper, we train deep neural networks on the foreground (object) and background (context) regions of images respectively. Considering human recognition in the same situations, networks trained on the pure background ***without*** objects achieves highly reasonable recognition performance that beats humans by a large margin if only given context. However, humans still outperform networks ***with*** pure object available, which indicates networks and human beings have different mechanisms in understanding an image. Furthermore, we straightforwardly combine multiple trained networks to explore different visual cues learned by different networks. Experiments show that useful visual hints can be ***explicitly*** learned separately and then combined to achieve higher performance, which verifies the advantages of the proposed framework.

1 Introduction

Object recognition is a long-lasting battle in computer vision, which aims to categorize an image according to the visual contents. In recent years, we have witnessed an evolution in this research field. Thanks to the availability of large-scale image datasets [Deng *et al.*, 2009] and powerful computational resources, it becomes possible to train a very deep convolutional neural network (CNN) [Krizhevsky *et al.*, 2012], which is much more efficient beyond the conventional Bag-of-Visual-Words (BoVW) model [Csurka *et al.*, 2004].

It is known that an image contains both foreground and background visual contents. However, most object recognition algorithms focus on recognizing the visual patterns only on the foreground region [Zeiler and Fergus, 2014]. Although it has been proven that background (context) information also helps recognition [Simonyan and Zisserman, 2014], it still remains unclear if a deep network can be trained individually to learn visual information only from the background region. In addition, we are interested in exploring different visual patterns by training neural networks on foreground and background separately for object recognition, which is less studied before.

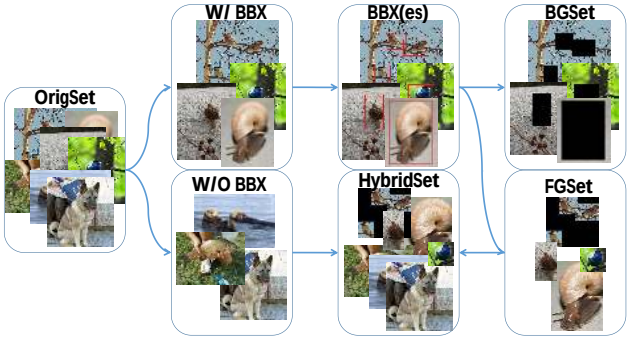

Figure 1: Procedures of dataset generation. First, we denote the original set as the **OrigSet**, divided into two sets, one with the ground-truth bounding box (W/ BBX) and the other one without (W/O BBX). Then the set with labelled bounding box(es) are further processed by setting regions inside all ground-truth to be 0's to compose the **BGSet** while cropping the regions out to produce the **FGSet**. In the end, add the images without bounding boxes with **FGSet** to construct the **HybridSet**. Please note that some images of the **FGSet** have regions to be black (0's) since these images are labelled with multiple objects belonging to the same class, which are cropped according to the smallest rectangle frame that includes all object bounding boxes in order to keep as less background information as possible on **FGSet**. Best viewed in color.

In this work, we investigate the above problems by explicitly training multiple networks for object recognition. We first construct datasets from **ILSVRC2012** [Russakovsky *et al.*, 2015], *i.e.*, one *foreground* set and one *background* set, by taking advantage of the ground-truth bounding box(es) provided in both training and testing cases. After dataset construction, we train deep networks individually to learn foreground (object) and background (context) information, respectively. We find that, even ***only*** trained on pure background contexts, the deep network can still converge and makes reasonable prediction ($\mathbf{14.4}\%$ top-1 and nearly $\mathbf{30}\%$ top-5 classification accuracy on the background validation set). To make a comparison, we are further interested in the human recognition performance on the constructed datasets. Deep neural networks outperform non-expert humans in fine-grained recognition, and humans sometimes make errors be-

cause they cannot memorize all categories of datasets [Russakovsky *et al.*, 2015]. In this case, to more reasonably compare the recognition ability of humans and deep networks, we follow [Huh *et al.*, 2016] to merge all the $1{,}000$ fine-grained categories of the original **ILSVRC2012**, resulting in a 127-class recognition problem meanwhile keeping the number of training/testing images unchanged. We find that human beings tend to pay more attention to the object while networks put more emphasis on context than humans for classification. By visualizing the patterns captured by the background net, we find that some visual patterns are not available in the foreground net. Therefore, we apply networks on the foreground and background regions respectively via the given ground-truth bounding box(es) or extracting object proposals without available ones. We find that the linear combination of multiple neural networks can give higher performance.

To summarize, our main contributions are three folds: 1) We demonstrate that learning foreground and background visual contents ***separately*** is beneficial for object recognition. Training a network based on pure background although being wired and challenging, is technically feasible and captures highly useful visual information. 2) We conduct ***human recognition*** experiments on either pure background or foreground regions to find that human beings outperform networks on pure foreground while are beaten by networks on pure background, which implies the different mechanisms of understanding an image between networks and humans. 3) We straightforwardly ***combine*** multiple neural networks to explore the effectiveness of different learned visual clues under two conditions ***with*** and ***without*** ground-truth bounding box(es), which gives promising improvement over the baseline deep neural networks.

2 Related Work

Object recognition is fundamental in computer vision field, which is aimed to understand the semantic meaning among an image via analyzing its visual contents. Recently, researchers have extended the traditional cases [Lazebnik *et al.*, 2006] to fine-grained [Wah *et al.*, 2011] [Nilsback and Zisserman, 2008] [Lin *et al.*, 2015], and large-scale [Xiao *et al.*, 2010] [Griffin *et al.*, 2007] tasks. Before the exploding development of deep learning, the dominant BoVW model [Csurka *et al.*, 2004] represents every single image with a high-dimensional vector. It is typically composed of three consecutive steps, *i.e.*, descriptor extraction [Lowe, 2004] [Dalal and Triggs, 2005], feature encoding [Wang *et al.*, 2010] [Perronnin *et al.*, 2010] and feature summarization [Lazebnik *et al.*, 2006].

The milestone Convolutional Neural Network (CNN) is treated as a hierarchical model for large-scale visual recognition. In past years, neural networks have already been proved to be effective for simple recognition tasks [LeCun *et al.*, 1990]. More recently, the availability of large-scale training data (*e.g.*, ImageNet [Deng *et al.*, 2009]) and powerful computation source like GPUs make it practical to train deep neural networks [Krizhevsky *et al.*, 2012] [Zhu *et al.*, 2016] [Fang *et al.*, 2015] [He *et al.*, 2016b] which significantly outperform the conventional models. Even deep features have been proved to be very successful on vision tasks like object discovery [Wang *et al.*, 2015b], object recognition [Xie *et al.*, 2017], etc. A CNN is composed of numerous stacked layers, in which responses from the previous layer are then convoluted and activated by a differentiable function, followed by a non-linear transformation [Nair and Hinton, 2010] to avoid over-fitting. Recently, several efficient methods were proposed to help CNNs converge faster and prevent over-fitting [Krizhevsky *et al.*, 2012]. It is believed that deeper networks produce better recognition results [Szegedy *et al.*, 2015][Simonyan and Zisserman, 2014], but also requires engineering tricks to be trained very well [Ioffe and Szegedy, 2015] [He *et al.*, 2016a].

Very few techniques on background modeling [Bewley and Upcroft, 2017] have been developed for object recognition, despite the huge success of deep learning methods on various vision tasks. [Shelhamer *et al.*, 2016] proposed the fully convolutional networks (FCN) for semantic segmentation, which are further trained on foreground and background defined by shape masks. They find it is not vital to learn a specifically designed background model. For face matching, [Sanderson and Lovell, 2009] developed methods only on the cropped out faces to alleviate the possible correlations between faces and their backgrounds. [Han *et al.*, 2015] modeled the background in order to detect the salient objects from the background. [Doersch *et al.*, 2014] showed using the object patch to predict its context as supervisory information can help discover object clusters, which is consistent with our motivation to utilize the pure context for visual recognition. To our best knowledge, we are the first to explicitly learn both the foreground and background models and then combine them together to be beneficial for the object recognition.

Recently, researchers pay more attention to human experiments on objects recognition. Zhou *et al.* [Zhou *et al.*, 2015] invited Amazon Mechanical Turk (AMT) to identify the concept for segmented images with objects. They found that the CNN trained for scene classification automatically discovers meaningful object patches. While in our experiments, we are particularly interested in the different emphasis between human beings and networks for recognition task.

Last but not the least, visualization of CNN activations is an effective method to understand the mechanism of CNNs. In [Zeiler and Fergus, 2014], a *de-convolutional* operation was proposed to capture visual patterns on different layers of a trained network. [Simonyan and Zisserman, 2014] and [Cao *et al.*, 2015] show that different sets of neurons are activated when a network is used for detecting different visual patterns. In this work, we will use a much simpler way of visualization which is inspired by [Wang *et al.*, 2015a].

3 Training Networks

Our goal is to explore the possibility and effectiveness of training networks on foreground and background regions, respectively. Here, foreground and background regions are defined by the annotated ground-truth bounding box(es) of each image. All the experiments are done on the datasets composed from the **ILSVRC2012**.

Dataset	Image Description	# Training Image	# Testing Image	Testing Accuracy
OrigSet	Original Image	1,281,167	50,000	58.19%, 80.96%
FGSet	Foreground Image	544,539	50,000	60.82%, 83.43%
BGSet	Background Image	289,031	50,000	14.41%, 29.62%
HybridSet	Original Image or Foreground Image	1,281,167	50,000	61.29%, 83.85%

Table 1: The configuration of different image datasets originated from the **ILSVRC2012**. The lass column denotes the testing performance of trained **AlexNet** in terms of top-1 and top-5 classification accuracy on corresponding datasets, *e.g.*, the **BGNet** gives 14.41% top-1 and 29.62% top-5 accuracy on the testing images of **BGSet**.

3.1 Data Preparation

The **ILSVRC2012** dataset [Russakovsky *et al.*, 2015] contains about 1.3M training and 50K validation images. Throughout this paper, we refer to the original dataset as **OrigSet** and the validation images are regarded as our testing set. Among **OrigSet**, 544,539 training images and all 50,000 testing images are labeled with at least one ground-truth bounding box. For each image, there is only one type of object annotated according to its ground-truth class label.

We construct three variants of training sets and two variants of testing sets from **OrigSet** by details below. An illustrative example of data construction is shown in Figure 1. The configuration of different image datasets are summarized in Table 1.

- The foreground dataset (**FGSet**) is composed of all images with at least one available ground-truth bounding box. For each image, we first compute the smallest rectangle frame which includes all object bounding boxes, then based on which the image inside the frame is cropped to be used as the training/testing data. Note that if an image has multiple object bounding boxes belonging to the same class, we set all the background regions inside the frame to be 0's to keep as little context as possible on **FGSet**. There are totally 544,539 training images and 50,000 testing images on **FGSet**. Since the annotation is on the bounding box level, images of the **FGSet** may contain some background information.
- The construction of the background dataset (**BGSet**) consists of two stages. First, for each image with at least one ground-truth bounding box available, regions inside every ground-truth bounding box are set to 0's. Chances are that almost all the pixels of one image are set to 0s if its object consists of nearly 100 percent of its whole region. Therefore during training, we discard those samples with less than 50% background pixels preserved, *i.e.*, the *foreground frame* is larger than half of the entire image, so that we can maximally prevent using those less meaningful background contents (see Figure 1). However in testing, we keep all the processed images, in the end, 289,031 training images and 50,000 testing images are preserved.
- To increase the amount of training data for foreground classification, we also construct a hybrid dataset, abbreviated as the **HybridSet**. The **HybridSet** is composed of all images of the original training set. If at least one ground-truth bounding box is available, we pre-process this image as described on **FGSet**, otherwise, we simply keep this image without doing anything. As bounding box annotation is available in each testing case, the **HybridSet** and the **FGSet** contain the same testing data. Training with the **HybridSet** can be understood as a semi-supervised learning process.

3.2 Training and Testing

We trained the milestone **AlexNet** [Krizhevsky *et al.*, 2012] using the CAFFE library [Jia *et al.*, 2014] on different training sets as mentioned in the Sec 3.1.

The base learning rate is set to 0.01, and reduced by $1/10$ for every 100,000 iterations. The moment is set to be 0.9 and the weight decay parameter is 0.0005. A total number of 450,000 iterations is conducted, which corresponds to around 90 training epochs on the original dataset. Note that both **FGSet** and **BGSet** contain less number of images than that of **OrigSet** and **HybridSet**, which leads to a larger number of training epochs, given the same training iterations. In these cases, we adjust the dropout ratio as 0.7 to avoid the overfitting issue. We refer to the network trained on the **OrigSet** as the **OrigNet**, and similar abbreviated names also apply to other cases, *i.e.*, the **FGNet**, **BGNet** and **HybridNet**.

During testing, we report the results by using the common data augmentation of averaging 10 patches from the 5 *crops* and 5 *flips*. After all forward passes are done, the average output on the final (*fc-8*) layer is used for prediction. We adopt the MatConvNet [Vedaldi and Lenc, 2015] platform for performance evaluation.

4 Experiments

The testing accuracy of **AlexNet** trained on corresponding dataset are given in the last column of Table 1. We can find that the **BGNet** produces reasonable classification results: 14.41% top-1 and 29.62% top-5 accuracy (while the random guess gets 0.1% and 0.5%, respectively), which is a bit surprising considering it makes classification decisions only on background contents ***without*** any foreground objects given. This demonstrates that deep neural networks are capable of learning pure contexts to infer objects even being fully occluded. Not surprisingly, the **HybridNet** gives better performance than the **FGNet** due to more training data available.

4.1 Human Recognition

As stated before, to alleviate the possibility of wrongly classifying images for humans beings due to high volume of classes up to 1,000 on the original **ILSVRC2012**, we follow [Huh *et al.*, 2016] by merging all the fine-grained categories, resulting

Dataset	AlexNet	Human
OrigSet	58.19%, 80.96%	−, 94.90%*
BGSet	14.41%, 29.62%	−, −
OrigSet-127	73.16%, 93.28%	−, −
FGSet-127	75.32%, 93.87%	81.25%, 95.83%
BGSet-127	41.65%, 73.79%	18.36%, 39.84%

Table 2: Classification accuracy (in terms of top-1, top-5) on five sets by deep neural networks and human, respectively.

Network	OrigSet	FGSet	BGSet
OrigNet	**58.19%**, **80.96%**	50.73%, 74.11%	3.83%, 9.11%
FGNet	33.42%, 53.72%	60.82%, 83.43%	1.44%, 4.53%
BGNet	4.26%, 10.73%	1.69%, 5.34%	**14.41%**, **29.62%**
HybridNet	52.89%, 76.61%	**61.29%**, **83.85%**	3.48%, 9.05%

Table 3: Cross evaluation accuracy (in terms of top-1, top-5) on four networks and three testing sets. Note that the testing set of **HybridSet** is identical to that of **FGSet**.

in a 127-class recognition problem meanwhile keeping the number of training/testing images unchanged. To distinguish the merged 127-class datasets with the previous datasets, we refer to them as the **OrigSet-127**, **FSet-127** and **BGSet-127**, respectively. Then we invite volunteers who are familiar with the merged 127 classes to perform the recognition task on **BGSet-127** and **FSet-127**. Humans are given 256 images covering all 127 classes and one image takes around two minutes to make the top-5 decisions. We do not evaluate humans on **OrigSet-127** since we believe humans can perform well on this set like on **OrigSet**. Human performance on **OrigSet** (labeled by $\star$) is reported by [Russakovsky *et al.*, 2015].

Table 2 gives the testing recognition performance of human beings and trained **AlexNet** on different datasets. It is well noted that humans are good at recognizing natural images [Russakovsky *et al.*, 2015], *e.g.*, on **OrigSet**, human labelers achieve much higher performance than **AlexNet**. We can find the human beings also surpass networks on the foreground (object-level) recognition by 5.93% and 1.96% in terms of top-1 and top-5 accuracy. Surprisingly, **AlexNet** beats human labelers to a large margin on the background dataset **BGSet-127** considering the 127% and 85% relative improvements from 18.36% to 41.65% and 39.84% to 73.79% for top-1 and top-5 accuracy, respectively. In this case, the networks are capable of exploring background hints for recognition much better than human beings. On the contrary, humans classify images mainly based on the visual contents of the foreground objects.

4.2 Cross Evaluation

To study the difference in visual patterns learned by different networks, we perform the cross evaluation, *i.e.*, applying each trained network to different testing sets. Results are summarized in Table 3.

We find that the transferring ability of each network is limited, since a model cannot obtain satisfying performance in the scenario of different distributions between training and testing data. For example, using **FGNet** to predict **OrigSet** leads to 27.40% absolute drop (45.05% relative) in top-1 accuracy, meanwhile using **OrigNet** to predict **FGSet** leads to 7.46% drop (12.82% relative) in top-1 accuracy. We conjecture that **FGNet** may store very little information on contexts, thus confused by the background context of **OrigSet**. On the other side, **OrigNet** has the ability of recognizing contexts but is wasted for the task on **FGSet**.

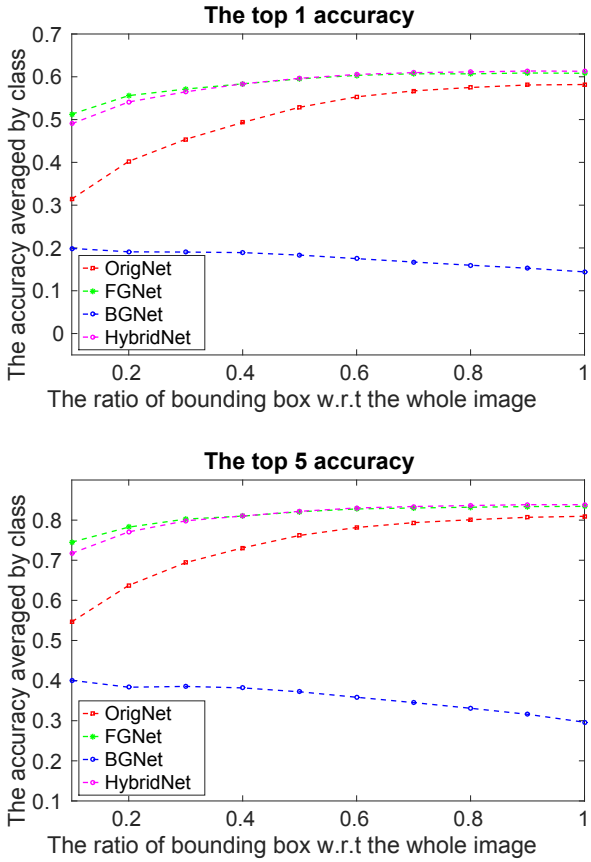

Figure 2: Classification accuracy with respect to the foreground ratio on testing images. The number at, say, 0.3, represents the testing accuracy on the set of all images with foreground ratio no greater than 30%. Best viewed in color.

4.3 Diagnosis

We conduct diagnostic experiments to study the property of different networks to fully understand the networks behaviors. Specifically, we report the classification accuracy of different networks with respect to keeping different foreground ratios of the testing image.

We split each testing dataset into 10 subsets, each of which contains all images with the foreground ratio no greater than a fixed value. Results are shown in Figure 2. **BGNet** gets higher classification accuracy on the images with a relatively smaller foreground ratio, while other three networks prefer a large object ratio since the foreground information is primarily learned for recognition in these cases. Furthermore when the foreground ratio goes larger, *e.g.*, greater than 80%,

Network	*Guided*	*Unguided*
OrigNet	58.19%, 80.96%	58.19%, 80.96%
BGNet	14.41%, 29.62%	8.30%, 20.60%
FGNet	60.82%, 83.43%	40.71%, 64.12%
HybridNet	61.29%, 83.85%	45.58%, 70.22%
FGNet+BGNet	61.75%, 83.88%	41.83%, 65.32%
HybridNet+BGNet	62.52%, 84.53%	48.08%, 72.69%
HybridNet+OrigNet	**65.63%**, **86.69%**	**60.36%**, **82.47%**

Table 4: Classification accuracy (in terms of top-1, top-5) comparison of different network combinations. It's worth noting that we feed the entire image into the **OrigNet** no matter whether the ground-truth bounding box(es) is given in order to keep the testing phase consistent with the training of **OrigNet**. Therefore, the reported results of **OrigNet** are same with each other under both *guided* and *unguided* conditions. To integrate the results from several networks, we weighted sum up the responses on the *fc-8* layer.

the performance gap among **OrigNet**, **FGNet** and **HybridNet** gets smaller.

4.4 Visualization

In this part, we visualize the networks to see how different networks learn different visual patterns. We adopt a very straightforward visualization method [Wang *et al.*, 2015a], which takes a trained network and reference images as input.

We visualize the most significant responses of the neurons on the *conv-5* layer. The *conv-5* layer is composed of 256 filter response maps, each of which has 13×13 different spatial positions. After all the 50,000 reference images are processed, we obtain $13^2 \times 50000$ responses for each of the 256 filters. We pick up those neurons with the highest response and trace back to obtain its receptive field on the input image. In this way, we can discover the visual patterns that best describe the concept this filter has learned. For diversity, we only choose at most one patch from a reference image with the highest response score.

Figure 3 shows visualization results using **FGNet** on **FGSet**, **BGNet** on **BGSet** and **OrigNet** on **OrigSet**, respectively. We can observe quite different visual patterns learned by these networks. The visual patterns learned by **FGNet** are often very specific to some object categories, such as the patch of a *dog face* (filter 5) or the *front side* of a *shop* (filter 11). These visual patterns correspond to some visual attributes, which are vital for recognition. However, each visual concept learned by **BGNet** tends to appear in many different object categories, for instance, the patch of *outdoor scene* (filter 8) shared by the *jetty, viaduct, space shuttle, etc.* These visual patterns are often found in the context, which plays an assistant role in object recognition. As for **OrigNet**, the learned patterns can be shared specific objects or scene.

To summarize, **FGNet** and **BGNet** learn different visual patterns that can be combined to assist visual recognition. In Sec 4.3 we quantitatively demonstrate the effectiveness of these networks via combining these information for better recognition performance.

5 Combination

We first show that the recognition accuracy can be significantly boosted using ground-truth bounding box(es) at the testing stage. Next, with the help of the EdgeBox algorithm [Zitnick and Dollar, 2014] to generate accurate object proposals, we improve the recognition performance without the requirement of ground-truth annotations. We name them as *guided* and *unguided* combination, respectively.

5.1 Guided vs. Unguided Combination

We start with describing guided and unguided manners of the model combination. For simplicity, we adopt the linear combination over different models, *i.e.*, forwarding several networks, and weighted summing up the responses on the *fc-8* layer.

If the ground-truth bounding box is provided (the *guided* condition), we use the ground-truth bounding box to divide the testing image into foreground and background regions. Then, we feed the foreground regions into **FGNet** or **HybridNet**, and background regions into **BGNet**, then fuse the neuron responses at the final stage.

Furthermore, we also explore the solution of combining multiple networks in an *unguided* manner. As we will see in Sec 5.2, a reliable bounding box helps a lot in object recognition. Motivated by which, we use an efficient and effective algorithm, EdgeBox, to generate a lot of potential bounding boxes proposals for each testing image, and then feed the foreground and background regions into neural networks as described before across top proposals.

To begin with, we demonstrate the EdgeBox proposals are good to capture the ground-truth object. After extracting top-k proposals with EdgeBox, we count the detected ground-truth if at least one of proposals has the IoU no less than 0.7 with the ground-truth. The cumulative distribution function (CDF) is plotted in Figure 4. Considering efficiency as well as accuracy, we choose the top-100 proposals to feed the foreground and background into trained networks, which give an around 81% recall. After obtaining 100 outputs for each network, we average responses of *fc-8* layer for classification.

5.2 Combination Results and Discussion

Results of different combinations are summarized in Table 4. Under either *guided* or *unguided* settings, combining multiple networks boosts recognition performance, which verifies the statement that different visual patterns from different networks can help with each other for the object recognition.

Take a closer look at the accuracy gain under the *unguided* condition. The combination of **HybridNet+BGNet** outperforms **HybridNet** by 2.50% and 2.47% in terms of top-1 and top-5 recognition accuracy, which are noticeable gains. As for the **FGNet+BGNet**, it improves 1.12% and 1.20% classification accuracy compared with the **FGNet**, which are promising. Surprisingly, the combination of **HybridNet** with **OrigNet** can still increase from the **OrigNet** by 2.17% and 1.51%. We hypothesize that the combination is capable of discovering the objects implicitly by the inference of where the objects are due to the visual patterns of **HybridNet** are learned from images with object spatial

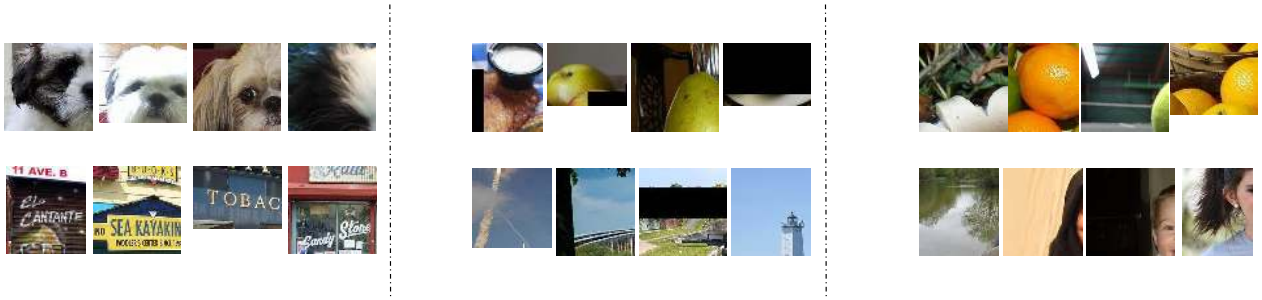

Figure 3: Patch visualization of **FGNet** on **FGSet** (left), **BGNet** on **BGSet** (middle) and **OrigNet** on **OrigSet** (right). Each row corresponds to one filter on the *conv-5* layer, and each patch is selected from $13^2 \times 50000$ ones, with the highest response on that kernel. Best viewed in color.

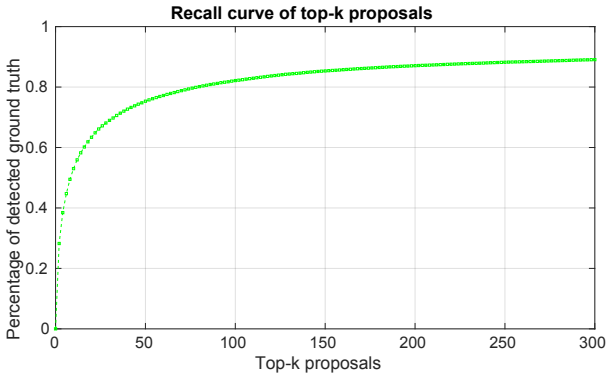

Figure 4: EdgeBox statistics on **ILSVRC2012** validation set, which denotes the curriculum distribution function of the detected ground-truth with respect to the top-k proposals. Here, we set the Intersection over Union (IoU) threshold to be 0.7 for EdgeBox algorithm.

information. One may conjecture that the performance improvement may come from the ensemble effect, which is not necessarily true considering: 1) object proposals are not accurate enough; 2) data augmentation (5 *crops* and 5 *flips*) is already done for the **OrigNet**, therefore the improvement is complementary to data augmentation. Moreover, we quantitatively verify that the improvements are not from simple data augmentation by giving the results of **OrigNet** averaged by 100 densely sampled patches (50 crops and corresponding 50 flips, $227 \times 227 \times 3$, referred to as **OrigNet100**) instead of the default (5 crops and 5 flips) setting. The top-1 and top-5 accuracy of **OrigNet100** are 58.08% and 81.05%, which are very similar to original 58.19% and 80.96%. This suggests that the effect of data augmentation by 100 patches is negligible. By contrast, **HybridNet+OrigNet100** reports 60.80% and 82.59%, significantly higher than **OrigNet100** alone, which reveals that **HybridNet** brings in some benefits that are not achieved via data augmentation. These improvements are super promising considering that the networks don't know where the accurate objects are under the *unguided* condition. Notice that the results under *unguided* condition cannot surpass those under *guided* condition, arguably because the top-100 proposals not good enough to capture the accurate ground-truth given that the **BGNet** cannot give high confidence on the predictions.

For the *guided* way of testing, by providing accurate separation of foreground from background, works better than the *unguided* way by a large margin, which makes sense. And the improvements can consistently be found after combinations with the **BGNet**. It is well worth noting that the combination of **HybridNet** with **OrigGNet** improves the baseline of **OrigGNet** to a significant margin by 7.44% and 5.73%. The huge gains are reasonable because of networks' ability to infer object locations trained on accurate bounding box(es).

6 Conclusions and Future Work

In this work, we first demonstrate the surprising finding that neural networks can predict object categories quite well even when the object is ***not*** present. This motivates us to study the human recognition performance on foreground ***with*** objects and background ***without*** objects. We show on the 127-classes **ILSVRC2012** that human beings beat neural networks for foreground object recognition, while perform much worse to predict the object category only on the background without objects. Then ***explicitly*** combining the visual patterns learned from different networks can help each other for the recognition task. We claim that more emphasis should be placed on the role of contexts for object detection and recognition.

In the future, we will investigate an end-to-end training approach for explicitly separating and then combining the foreground and background information, which explores the visual contents to the full extent. For instance, inspired by some joint learning strategy such as Faster R-CNN [Ren *et al.*, 2015], we can design a structure which predicts the object proposals in the intermediate stage, then learns the foreground and background regions derived from the proposals separately by two sub-networks and then takes foreground and background features into further consideration.

Acknowledgments

This work is supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center (DoI/IBC) contract number D16PC00007. We greatly thank the anonymous reviewers and JHU CCVL members who have given valuable and constructive suggestions which make this work better.

References

[Bewley and Upcroft, 2017] A. Bewley and B. Upcroft. Background appearance modeling with applications to visual object detection in an open-pit mine. *JFR*, 2017.

[Cao *et al.*, 2015] C. Cao, X. Liu, Y. Yang, Y. Yu, J. Wang, Z. Wang, Y. Huang, L. Wang, C. Huang, and W. Xu. Look and Think Twice: Capturing Top-Down Visual Attention with Feedback Convolutional Neural Networks. *ICCV*, 2015.

[Csurka *et al.*, 2004] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual Categorization with Bags of Keypoints. *Workshop on ECCV*, 2004.

[Dalal and Triggs, 2005] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. *CVPR*, 2005.

[Deng *et al.*, 2009] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. *CVPR*, 2009.

[Doersch *et al.*, 2014] C. Doersch, A. Gupta, and A.A. Efros. Context as supervisory signal: Discovering objects with predictable context. *ECCV*, 2014.

[Fang *et al.*, 2015] Y. Fang, J. Xie, G. Dai, M. Wang, F. Zhu, T. Xu, and E. Wong. 3d deep shape descriptor. In *CVPR*, 2015.

[Griffin *et al.*, 2007] G. Griffin, A. Holub, and P. Perona. Caltech-256 Object Category Dataset. *Technical Report: CNS-TR-2007-001, Caltech*, 2007.

[Han *et al.*, 2015] J. Han, D. Zhang, X. Hu, L. Guo, J. Ren, and F. Wu. Background prior-based salient object detection via deep reconstruction residual. *TCSVT*, 2015.

[He *et al.*, 2016a] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. *CVPR*, 2016.

[He *et al.*, 2016b] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In *ECCV*, 2016.

[Huh *et al.*, 2016] M. Huh, P. Agrawal, and A.A. Efros. What Makes ImageNet Good for Transfer Learning? *arXiv: 1608.08614*, 2016.

[Ioffe and Szegedy, 2015] S. Ioffe and C. Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *ICML*, 2015.

[Jia *et al.*, 2014] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. CAFFE: Convolutional Architecture for Fast Feature Embedding. *ACM MM*, 2014.

[Krizhevsky *et al.*, 2012] A. Krizhevsky, I. Sutskever, and G.E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *NIPS*, 2012.

[Lazebnik *et al.*, 2006] S. Lazebnik, C. Schmid, and J. Ponce. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. *CVPR*, 2006.

[LeCun *et al.*, 1990] Y. LeCun, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, and L.D. Jackel. Handwritten Digit Recognition with a Back-Propagation Network. *NIPS*, 1990.

[Lin *et al.*, 2015] T. Lin, A. RoyChowdhury, and S. Maji. Bilinear cnn models for fine-grained visual recognition. *ICCV*, 2015.

[Lowe, 2004] D.G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *IJCV*, 2004.

[Nair and Hinton, 2010] V. Nair and G.E. Hinton. Rectified linear units improve restricted boltzmann machines. *ICML*, 2010.

[Nilsback and Zisserman, 2008] M.E. Nilsback and A. Zisserman. Automated Flower Classification over a Large Number of Classes. *ICVGIP*, 2008.

[Perronnin *et al.*, 2010] F. Perronnin, J. Sanchez, and T. Mensink. Improving the Fisher Kernel for Large-scale Image Classification. *ECCV*, 2010.

[Ren *et al.*, 2015] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards Real-time Object Detection with Region Proposal Networks. *NIPS*, 2015.

[Russakovsky *et al.*, 2015] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, pages 1–42, 2015.

[Sanderson and Lovell, 2009] C. Sanderson and B.C. Lovell. Multi-region probabilistic histograms for robust and scalable identity inference. *ICB*, 2009.

[Shelhamer *et al.*, 2016] E. Shelhamer, J. Long, and T. Darrell. Fully convolutional networks for semantic segmentation. *TPAMI*, 2016.

[Simonyan and Zisserman, 2014] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *ICLR*, 2014.

[Szegedy *et al.*, 2015] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going Deeper with Convolutions. *CVPR*, 2015.

[Vedaldi and Lenc, 2015] A. Vedaldi and K. Lenc. MatConvNet – Convolutional Neural Networks for MATLAB. In *ACM MM*, 2015.

[Wah *et al.*, 2011] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. *Technical Report: CNS-TR-2011-001, Caltech*, 2011.

[Wang *et al.*, 2010] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-Constrained Linear Coding for Image Classification. *CVPR*, 2010.

[Wang *et al.*, 2015a] J. Wang, Z. Zhang, V. Premachandran, and A. Yuille. Discovering Internal Representations from Object-CNNs Using Population Encoding. *arXiv: 1511.06855*, 2015.

[Wang *et al.*, 2015b] X. Wang, Z. Zhu, C. Yao, and X. Bai. Relaxed multiple-instance svm with application to object discovery. *ICCV*, 2015.

[Xiao *et al.*, 2010] J. Xiao, J. Hays, K.A. Ehinger, A. Oliva, and A. Torralba. SUN Database: Large-Scale Scene Recognition from Abbey to Zoo. *CVPR*, 2010.

[Xie *et al.*, 2017] L. Xie, J. Wang, W. Lin, B. Zhang, and Q. Tian. Towards reversal-invariant image representation. *IJCV*, 2017.

[Zeiler and Fergus, 2014] M.D. Zeiler and R. Fergus. Visualizing and Understanding Convolutional Networks. *ECCV*, 2014.

[Zhou *et al.*, 2015] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Object Detectors Emerge in Deep Scene CNNs. *ICLR*, 2015.

[Zhu *et al.*, 2016] Z. Zhu, X. Wang, S. Bai, C. Yao, and X. Bai. Deep learning representation using autoencoder for 3d shape retrieval. *Neurocomputing*, 2016.

[Zitnick and Dollar, 2014] C.L. Zitnick and P. Dollar. Edge Boxes: Locating Object Proposals from Edges. *ECCV*, 2014.