Object Recognition with and without Objects

Zhuotun Zhu, Lingxi Xie, Alan Yuille
{zhuotun, 198808xc, alan.l.yuille}@gmail.com

Johns Hopkins University

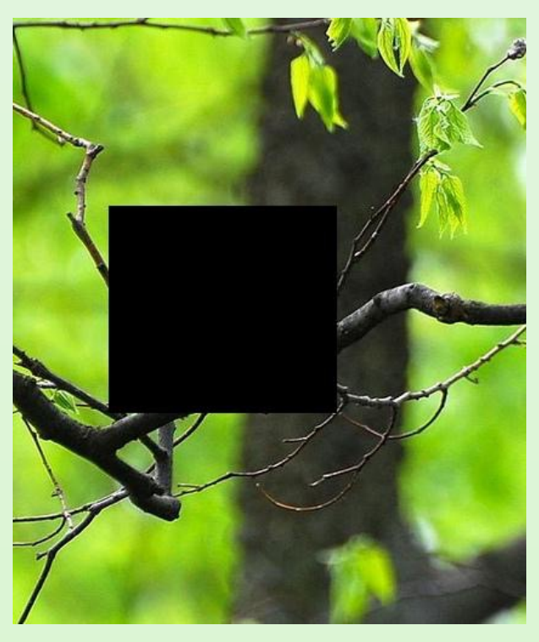
Introduction

While recent deep neural networks have achieved a promising performance on object recognition, they rely ***implicitly*** on the visual contents of the whole image. So we train deep neural networks on the foreground (object) and background (context) regions of images respectively.

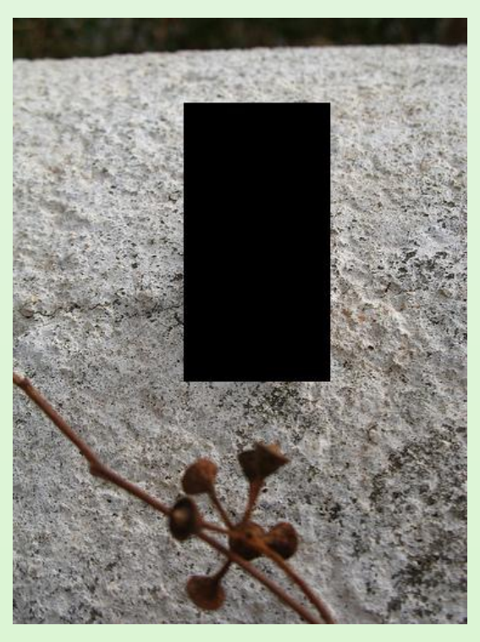
Contributions

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, which gives promising improvement over the baseline deep neural networks.

Examples

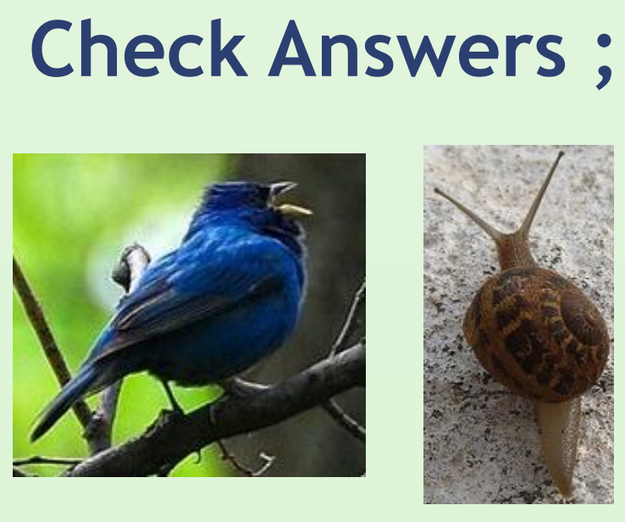


Bird?
Squirrel?
Monkey?
Bat?
…



Snake?
Snail?
Lizard?
Scorpion?
…

Check Answers ;)



Bird and Snail

Datasets

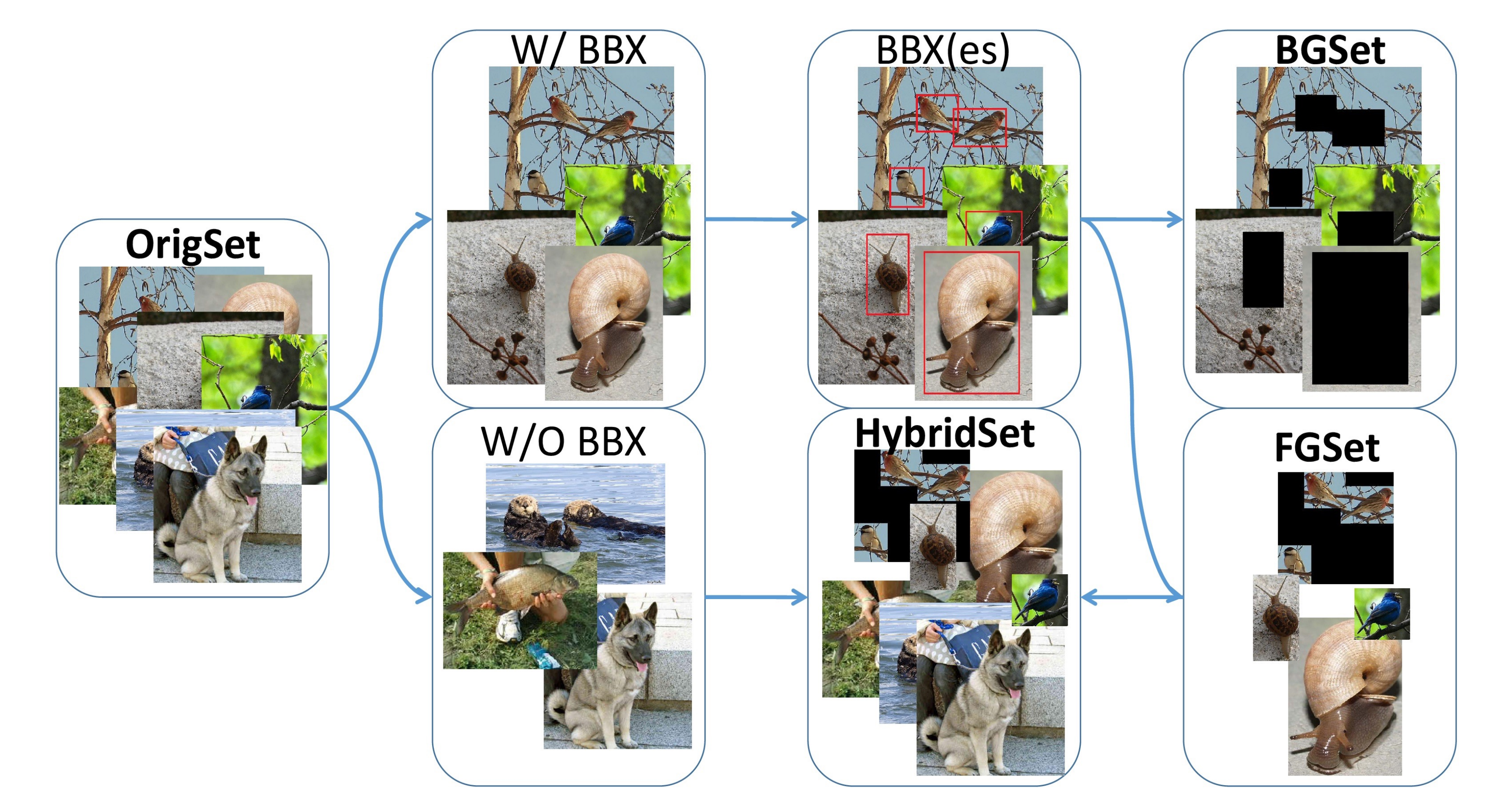


Figure 1: Procedures of dataset generation.

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.

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.

Acknowledgement

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..

Experiments

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.

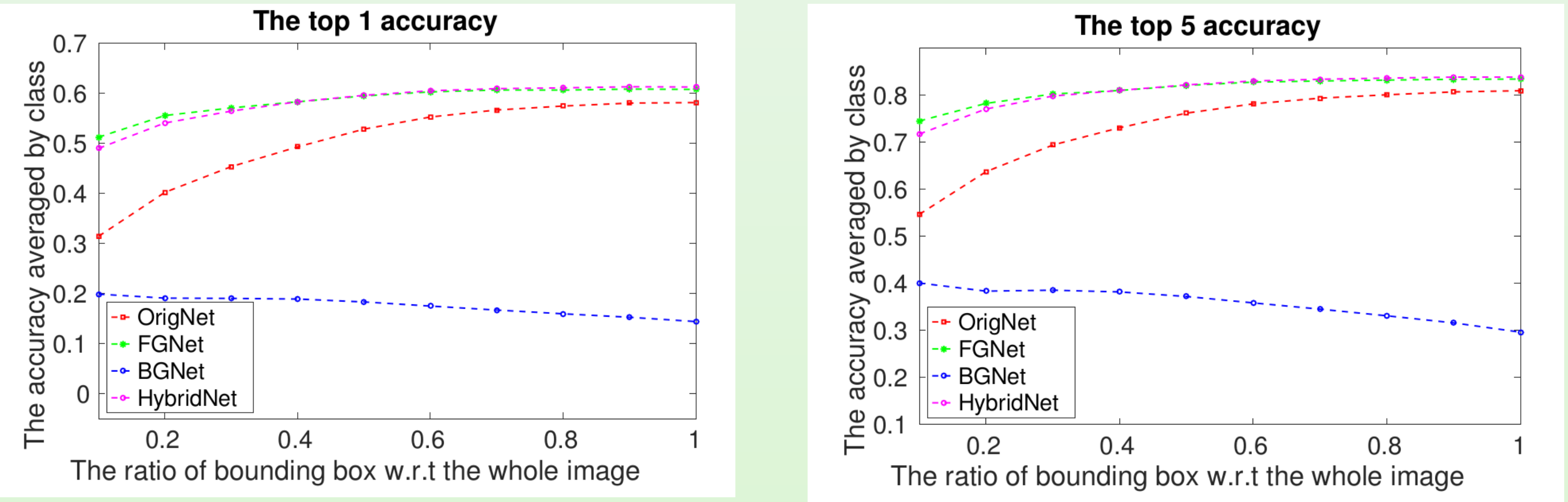


Figure 2: Classification accuracy with respect to the foreground ratio on testing images. The number at, say, 0.3, represents the testing accuracy on the set of all images with foreground ratio no greater than 30%. Best viewed in color.

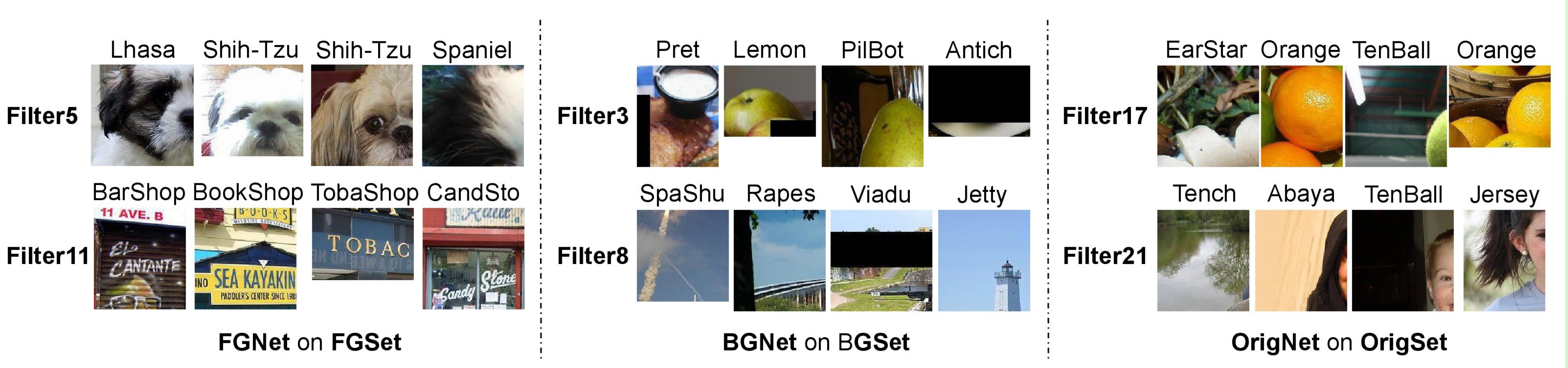


Figure 3: Patch visualization of FGNet on FGSet (left), BGNet on BGSet (middle) and OrigNet on OrigSet (right). Each row corresponds to one filter on the conv-5 layer, and each patch is selected from 13^2× 50000 ones, with the highest response on that kernel.

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.

Conclusions

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.

References

[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.
[Huh et al., 2016] M. Huh, P. Agrawal, and A.A. Efros. What Makes ImageNet Good for Transfer Learning? arXiv: 1608.08614, 2016.
[Krizhevsky et al., 2012] A. Krizhevsky, I. Sutskever, and G.E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. NIPS, 2012.
[Zitnick and Dollar, 2014] C.L. Zitnick and P. Dollar. Edge Boxes: Locating Object Proposals from Edges. ECCV, 2014.