Object Recognition with and without Objects

Zhuotun Zhu, Lingxi Xie, Alan Yuille
Johns Hopkins University

Object Recognition

- A fundamental vision problem
  - This task traditionally means each image has exactly one label that can take a single value among a finite number of choices. The assumption is that each image contains exactly one recognisable object (or perhaps none, in which case it takes the "background" label).

Object Recognition

- Before deep learning

SIFT
HOG
SURF
etc…

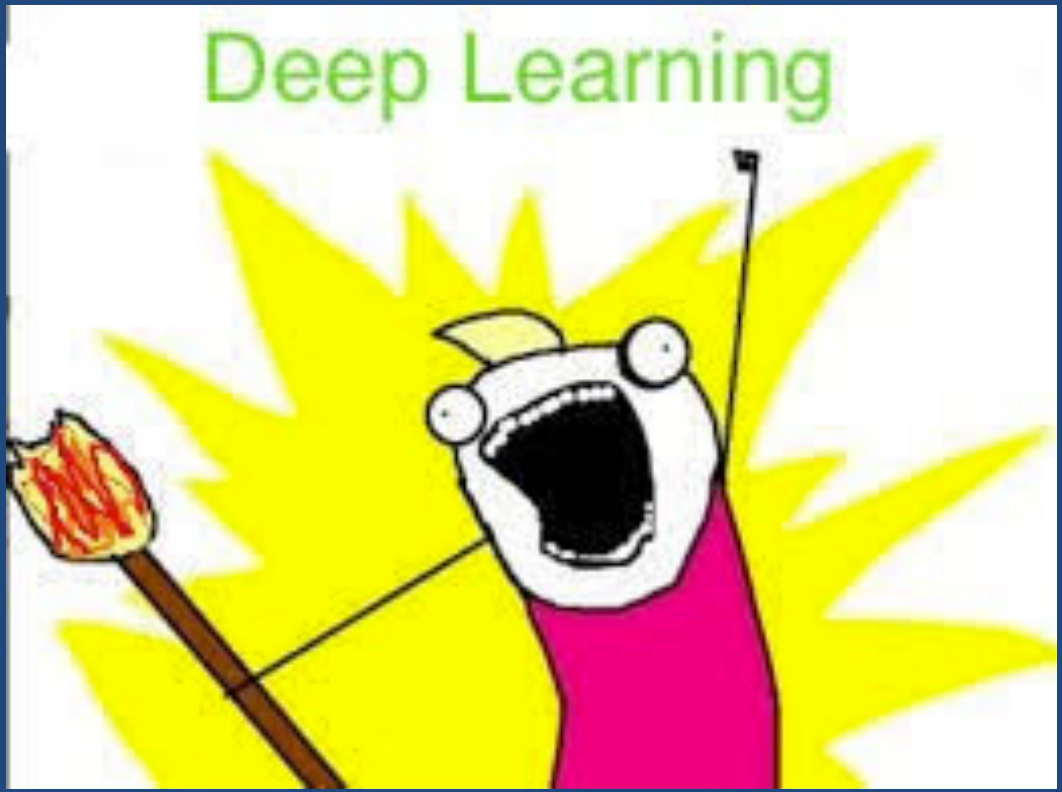
BoW
LLC
VLAD
etc…

SVM
KNN
etc…

Cat?

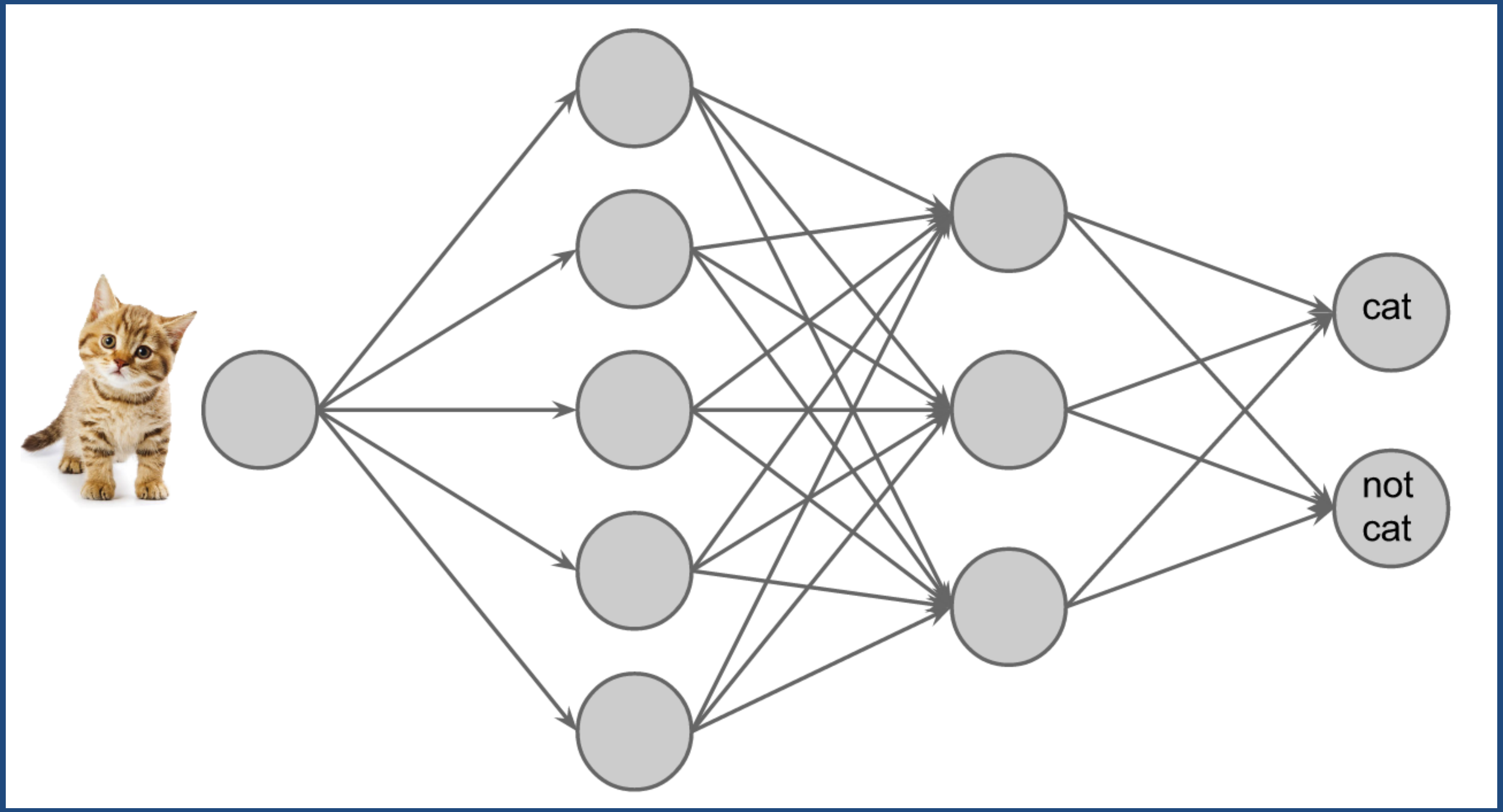
Object Recognition

- Deep learning
  - Computational resources, *e.g.*, GPU
  - Large Dataset, *e.g.*, ImageNet



Object Recognition

- Deep learning
  - ✦ Computational resources: GPU
  - ✦ Large Dataset: ImageNet



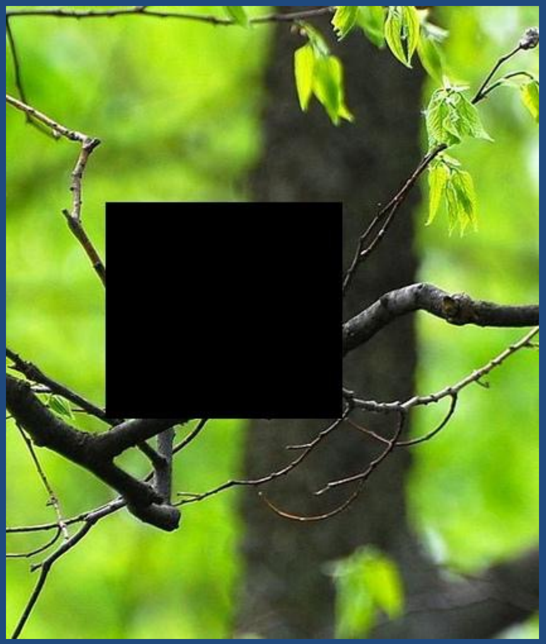
Object Recognition

- Multiple layers of learned feature detectors :)
- Local feature detectors are replicated across space :)
- Detectors get bigger in higher layers in space :)
- Foreground and background are learnt together *implicitly* :(

First three claims are borrowed from G.E. Hinton's recent talk, "What is wrong with convolutional neural nets".

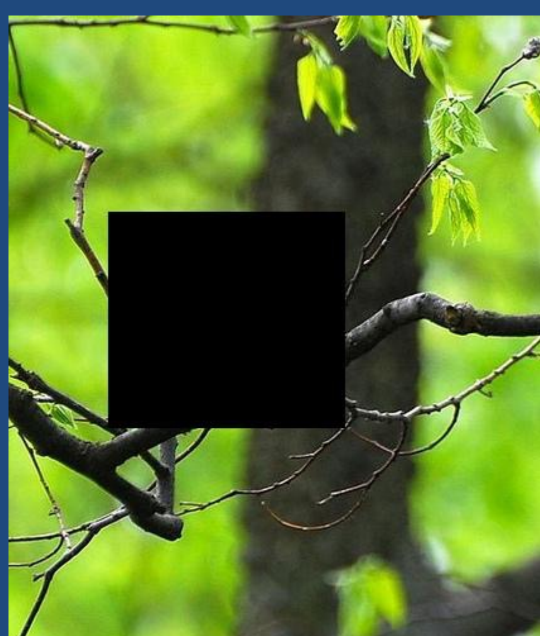
Intuitions

- Two examples

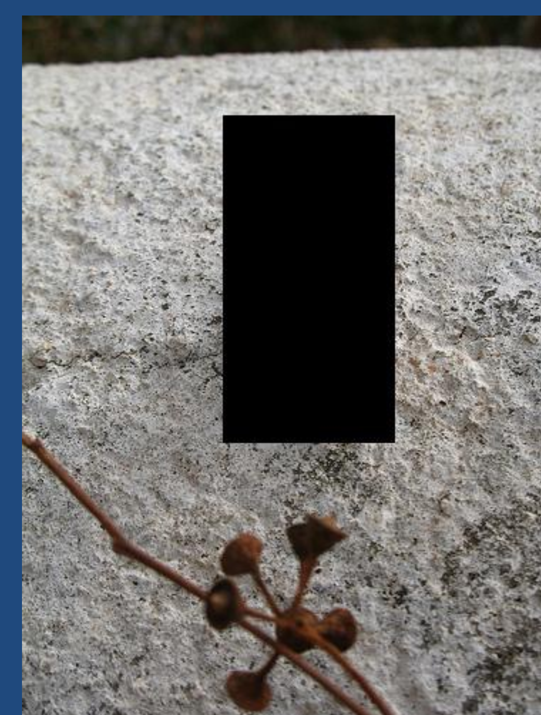


Intuitions

- Two examples



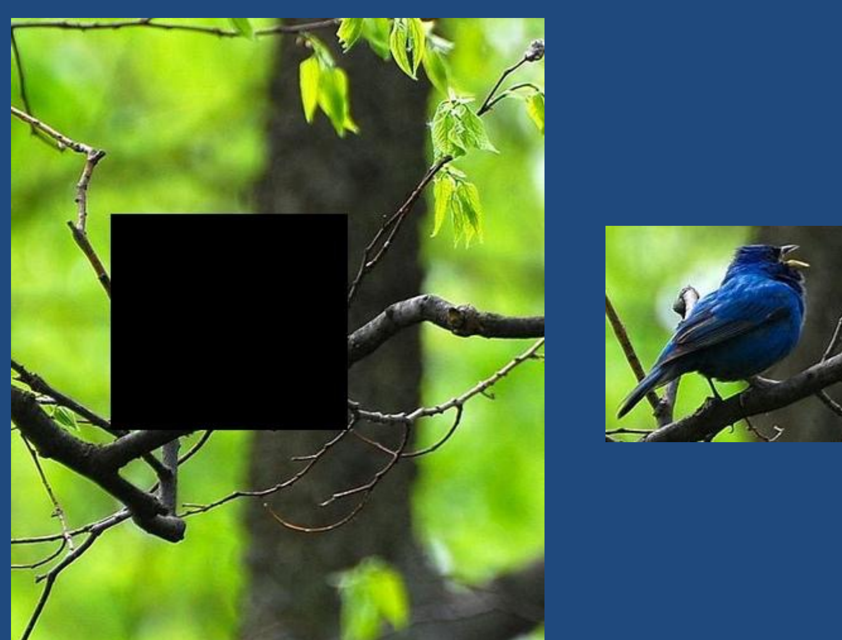
Bird?
Squirrel?
Monkey?
Bat?
...

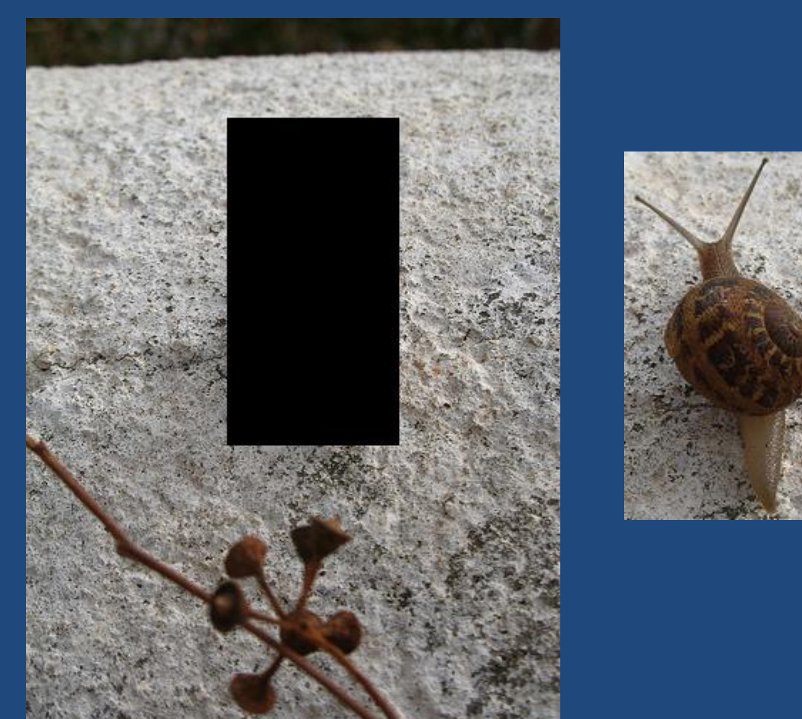


Snake?
Snail?
Lizard?
Scorpion?
...

Intuitions

- Two examples



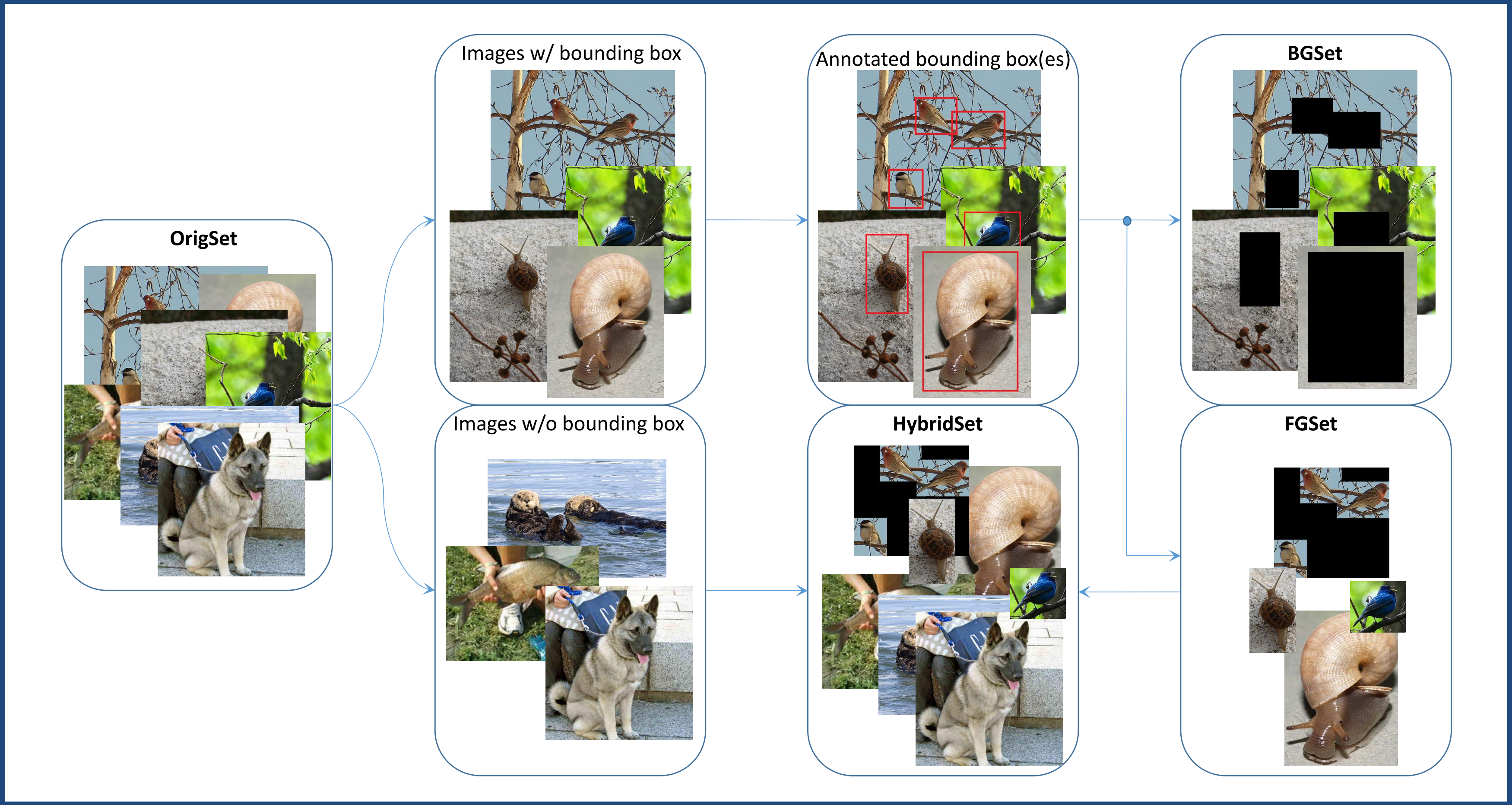


Key Questions

- How well can deep neural networks learn on the pure foreground (object) and background (context)?
- Could there be any difference between human and networks for understanding image (especially the foreground and background)?
- What can the networks do by learning the foreground and background models separately?

Datasets

- ILSVRC2012[2]: 1K classes, 1.28M training, 50K testing



[2] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, pages 1–42, 2015.

Datasets

- Summary of the datasets

Dataset	Image Description	# Training Image	# Testing Image
OrigSet	Original Image	1,281,167	50,000
FGSet	Foreground Image	544,539	50,000
BGSet	Background Image	289,031	50,000
HybridSet	Original Image or Foreground Image	1,281,167	50,000

Experiments

- AlexNet[3] v.s. Human

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%

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

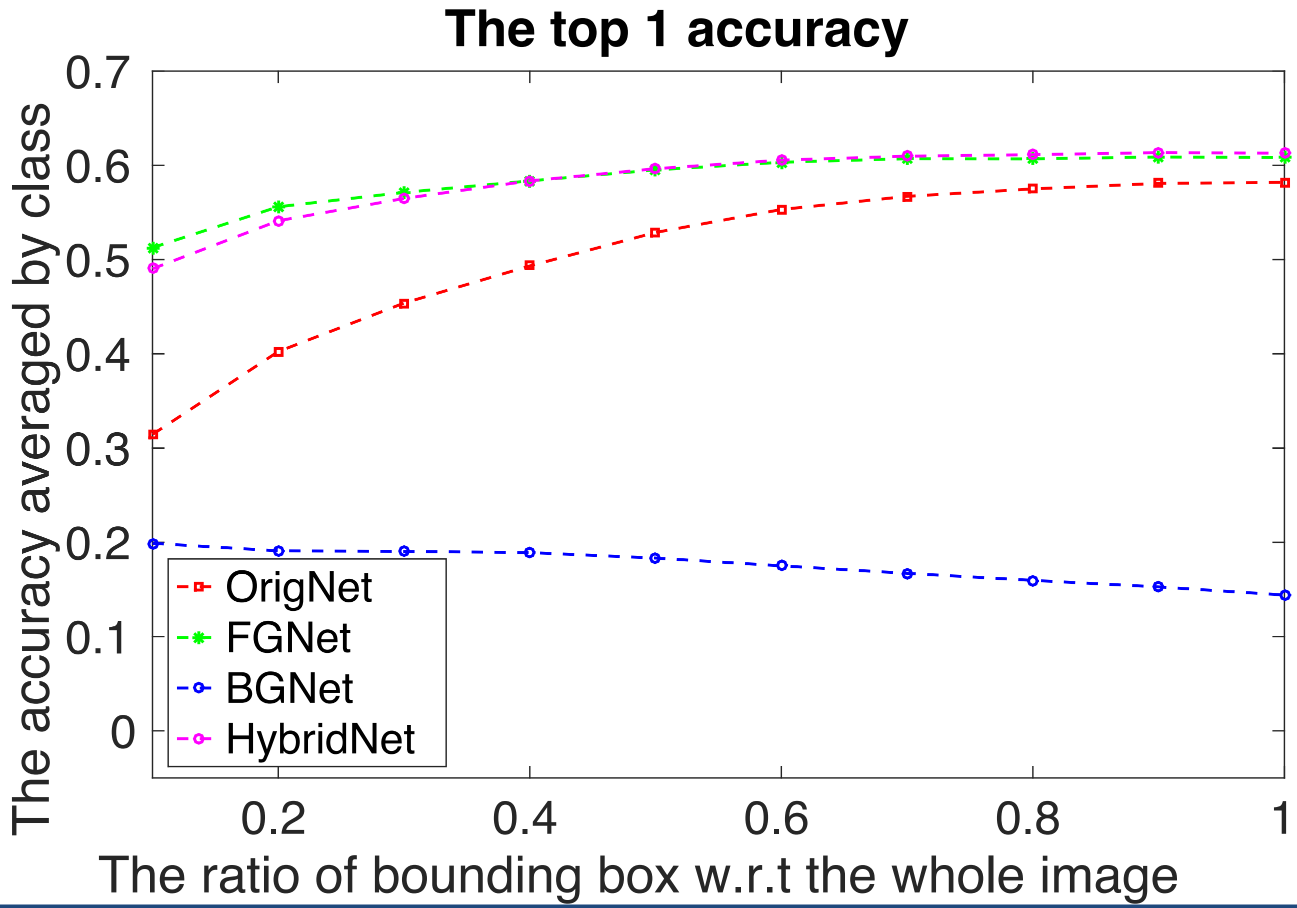
Experiments

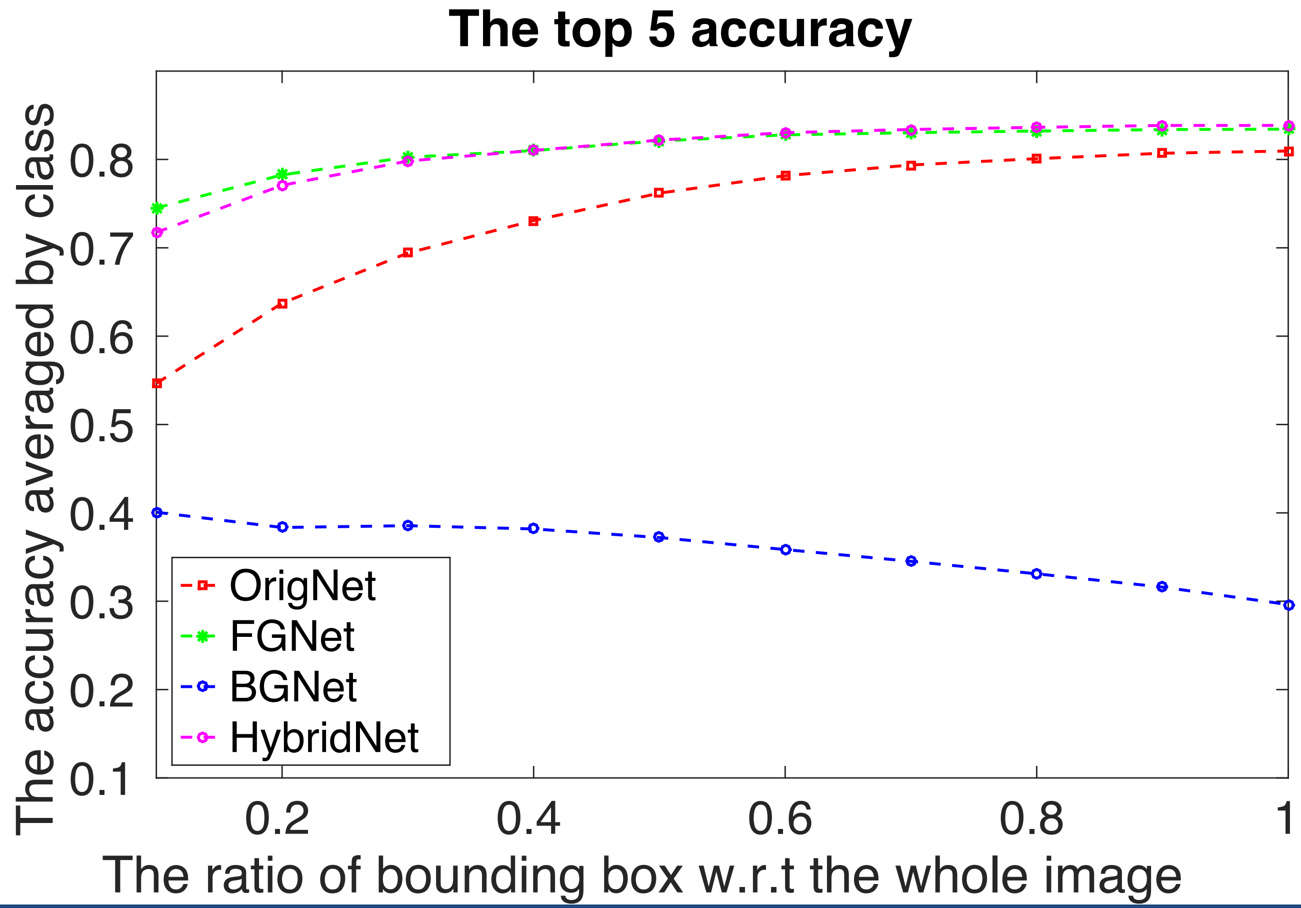
- Cross Validation

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%

Experiments

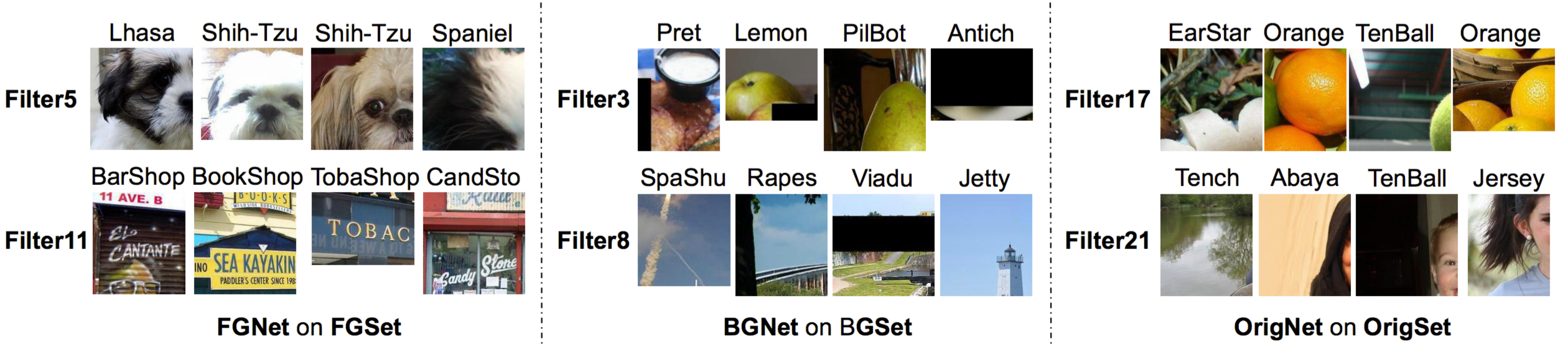
- Ratio of bounding box





Experiments

- Patches Visualization[4]



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

Experiments

- Recognition w. & w/o. objects

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%**	**60.84%, 82.56%**

Conclusions

- AlexNet can learn ***reasonable*** models to explore the correlation between the foreground object and background context

- AlexNet tend to perform better than human on background ***without*** objects but is beaten on foreground ***with*** object

- Combining the learnt networks can be ***beneficial*** for object recognition

Future Works

- An end-to-end training framework for explicitly separating and then combining the foreground and background information