R-CNN
for
Object Detection

Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik

(UC Berkeley)

presented by

Ezgi Mercan

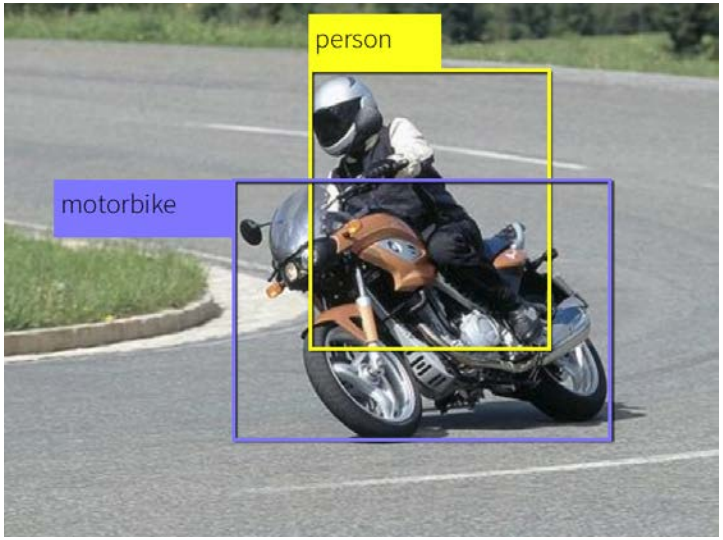
Outline

1. Problem Statement: Object Detection (and Segmentation)
2. Background: DPM, Selective Search, Regionlets
3. Method overview
4. Evaluation
5. Extensions to DPM and RGB-D
6. Discussion

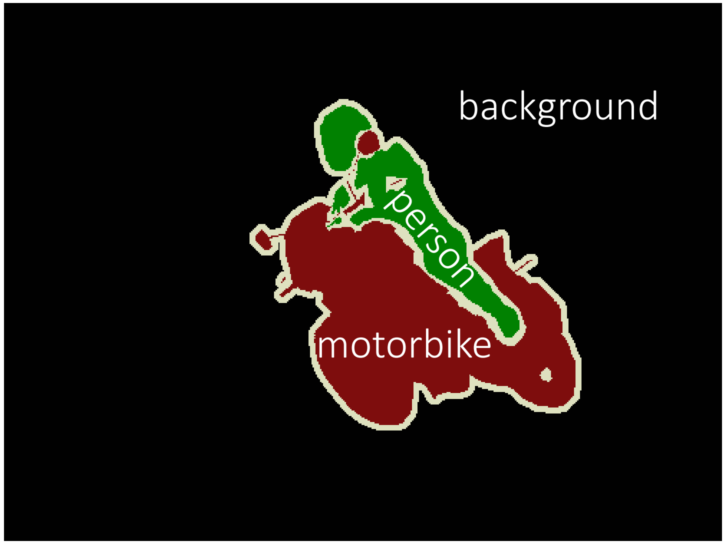
Detection and Segmentation

input image

object detection



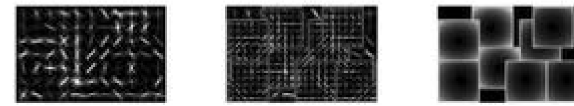
segmentation



Background: VOC

- PASCAL **V**isual **O**bject **C**lasses Challenge
- 20 classes, ~10K images, ~25K annotated objects
- Training, validation, test data sets.
- Evaluation:
  - **A**verage **P**recision (AP) per class
  - **m**ean **A**verage **P**recision

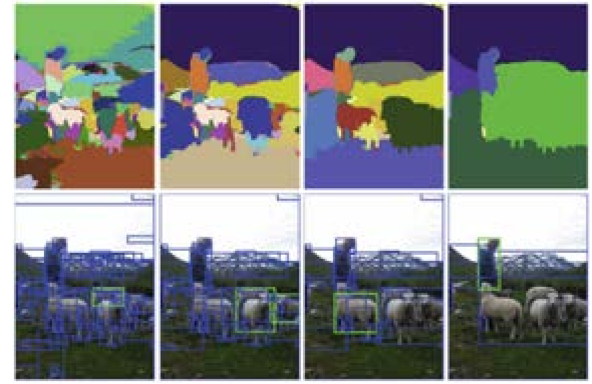
Background: Deformable Parts Model



- Strong low-level features based on histograms of oriented gradients (HOG)
- Efficient matching algorithms for deformable part-based models (pictorial structures)
- Discriminative learning with latent variables (latent SVM)
- mean Average Precision (mAP): 33.7% - 33.4%
- mAP with “context”: 35.4%
- mAP with “sketch tokens”: 29.1%
- mAP with “histograms of sparce codes”: 34.3%

P.F. Felzenszwalb et al., “Object Detection with Discriminatively Trained Part-Based Models”, PAMI 2010.
J.J. Lim et al., “Sketch Tokens: A Learned Mid-level Representation for Contour and Object Detection”, CVPR 2013.
X. Ren et al., “Histograms of Sparse Codes for Object Detection”, CVPR 2013.

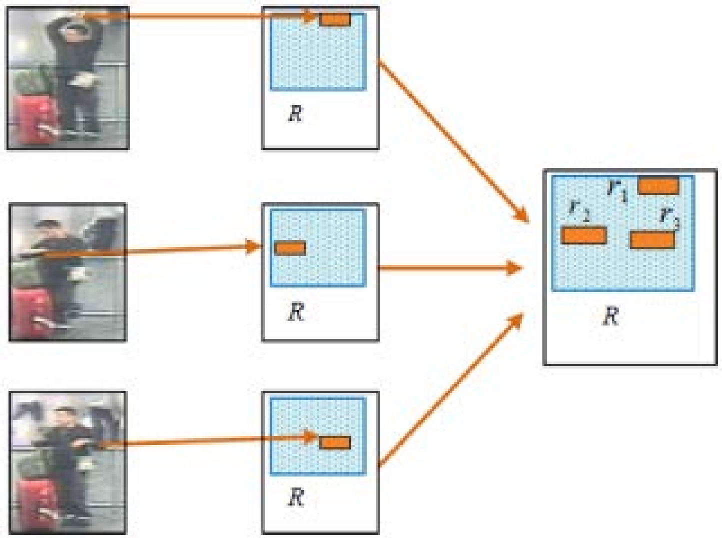
Background: Selective search



- Alternative to exhaustive search with sliding window.
- Starting with over-segmentation, merge *similar* regions and produce region proposals.
- Bag-of-Words Model with Dense SIFT, OpponentSIFT and RGB-SIFT, plus SVM.
- mAP: ? – 35.1%

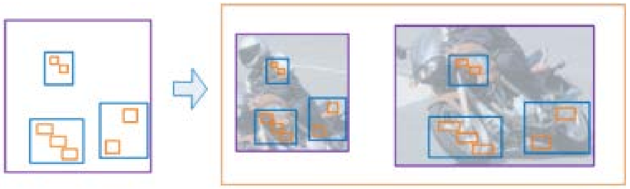
B.C. Russell et al., “Using Multiple Segmentations to Discover Objects and their Extent in Image Collections”, CVPR 2006.
C. Gu et al., “Recognition Using Regions”, CVPR 2009.
van de Sande et al., “Segmentation as Selective Search for Object Recognition”, ICCV 2011.

Background: Regionlets



- Start with *selective search*.
- Define sub-parts of regions whose position/resolution are relative and normalized to a detection window, as the basic units to extract appearance features.
- Features: HOG, LBP, Covarience.
- mAP: 41.7% - 39.7%

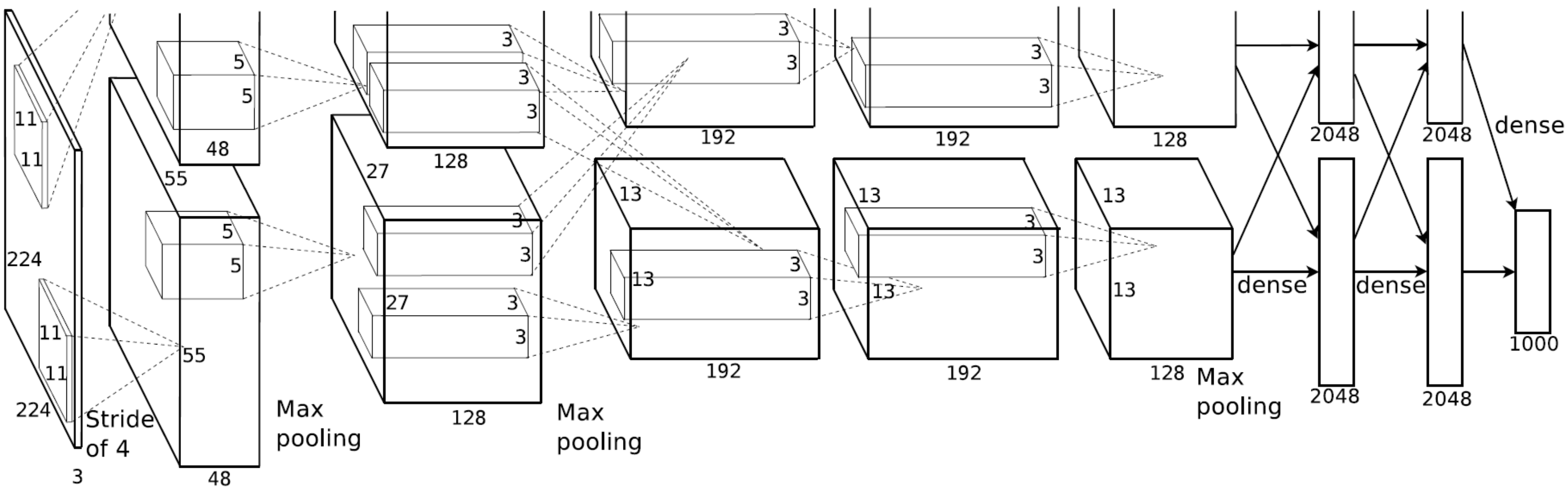
Wang et al., "Regionlets for Generic Object Detection", ICCV 2013.



Regionlet based model   Applied to candidate boxes

Deep Learning is back!

UToronto "SuperVision" CNN



Krizhevsky et al., "ImageNet Classification with Deep Convolutional Neural Networks", NIPS 2012.

ImageNet 2012

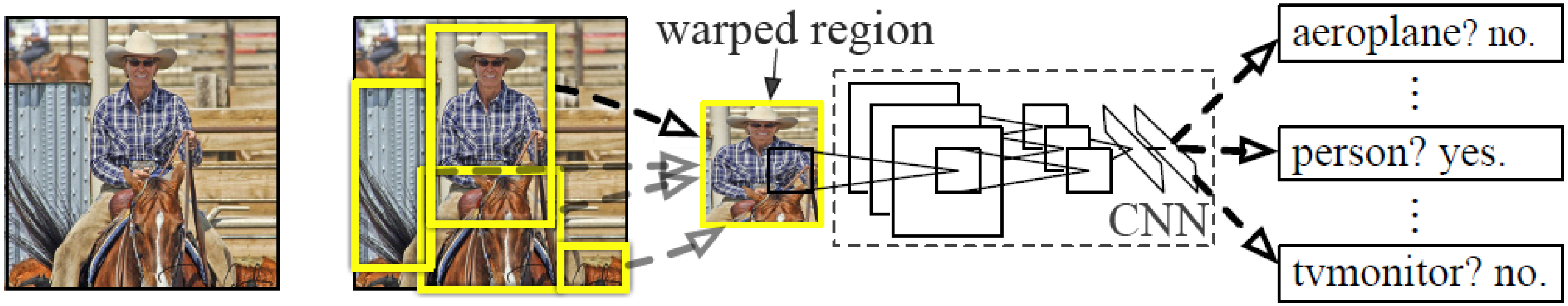
whole-image classification with 1000 categories

Model	Top-1 (val)	Top-5 (val)	Top-5 (test)
SIFT + Fisher Vectors	-	-	26.2%
1 CNN	40.7%	18.2%	-
5 CNNs	38.1%	16.4%	16.4%
1 CNN (pre-trained)	39.0%	16.6%	-
7 CNNs (pre-trained)	36.7%	15.4%	15.3%

- Can it be used in object recognition?
- Problems:
  - localization: Where is the object?
  - annotation: Labeled data is scarce.

Krizhevsky et al., “ImageNet Classification with Deep Convolutional Neural Networks”, NIPS 2012.

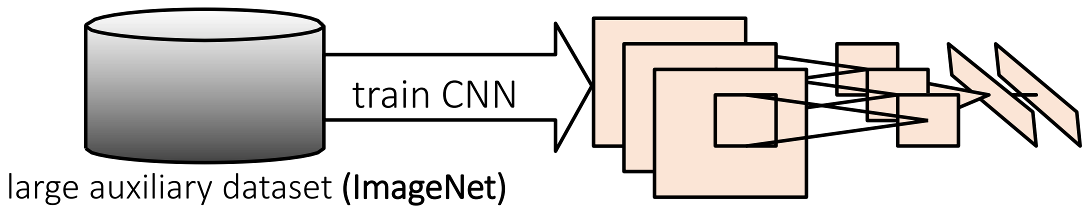
R-CNN: Region proposals + CNN



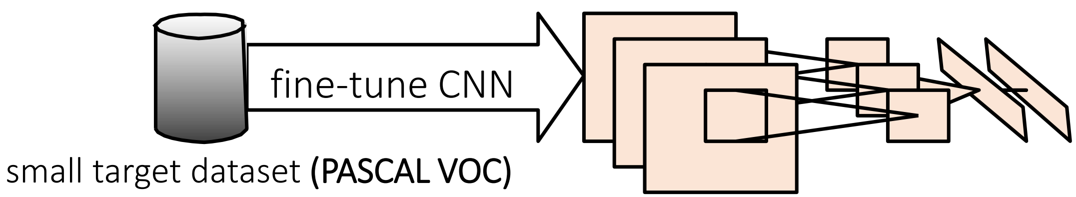
	localization	feature extraction	classification
this paper:	selective search	deep learning CNN	binary linear SVM
alternatives:	objectness, constrained parametric min-cuts, sliding window …	HOG, SIFT, LBP, BoW, DPM …	SVM, Neural networks, Logistic regression …

R-CNN: Training

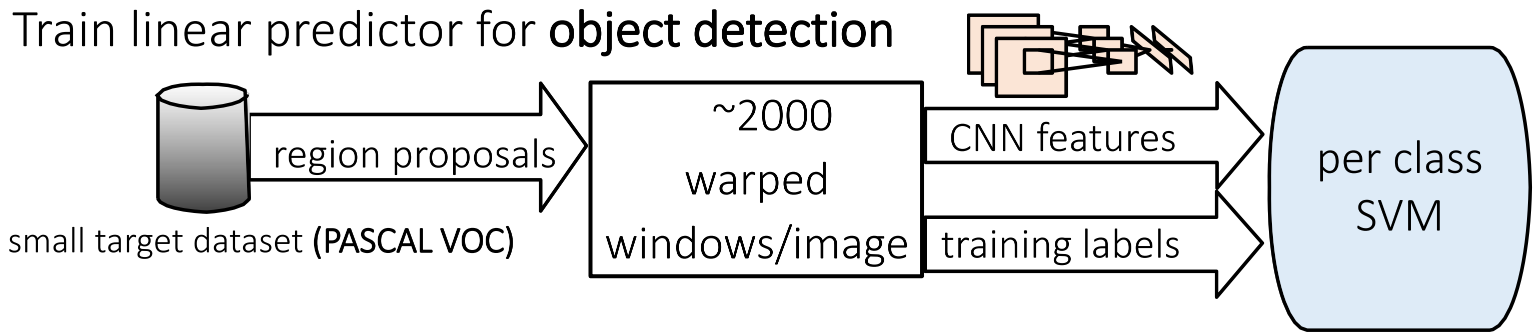
1. Pre-train CNN for **image classification**



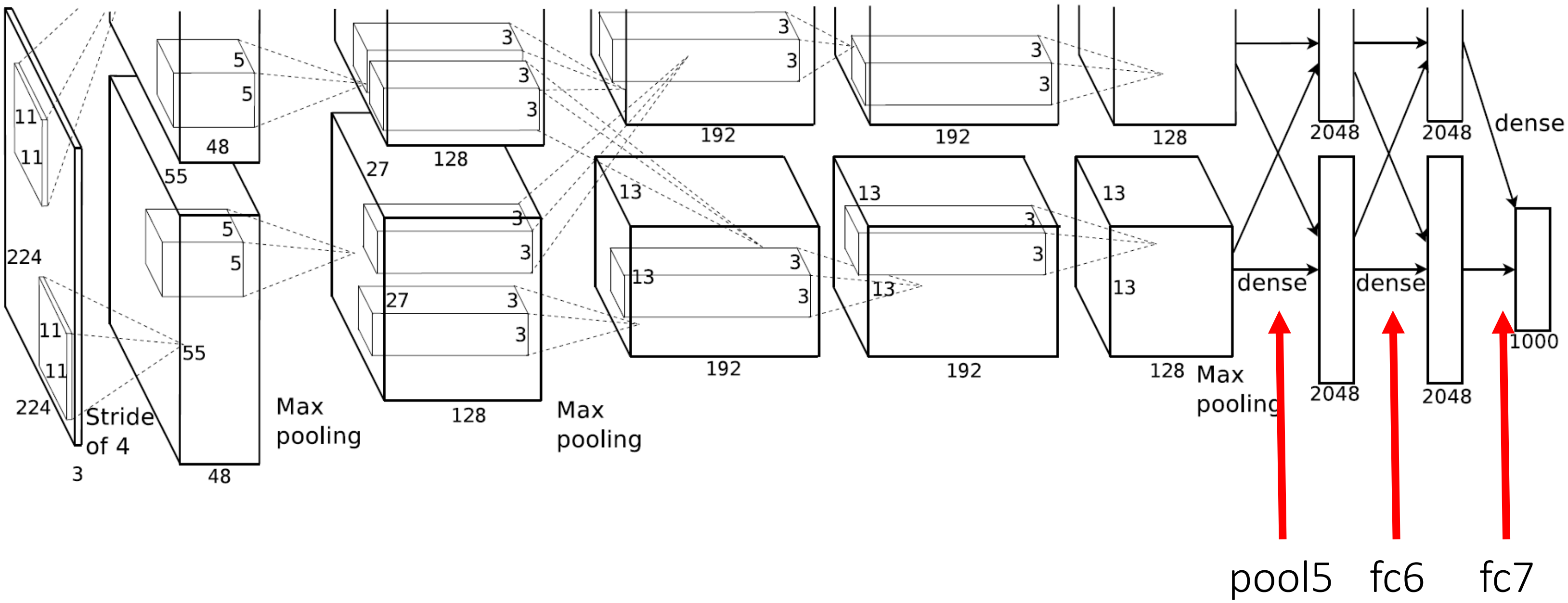
2. Fine-tune CNN for **object detection**



3. Train linear predictor for **object detection**



UToronto "SuperVision" CNN



Krizhevsky et al., "ImageNet Classification with Deep Convolutional Neural Networks", NIPS 2012.

Evaluation: mAP

		VOC 2007	VOC 2010
reference	DPM v5 (Girshick et al. 2011)	33.7%	29.6%
	UVA sel. search (Uijlings et al. 2012)		35.1%
	Regionlets (Wang et al. 2013)	41.7%	39.7%
pre-trained only	R-CNN $pool_5$	44.2%	
	R-CNN fc_6	46.2%	
	R-CNN fc_7	44.7%	
fine-tuned	R-CNN $pool_5$	47.3%	
	R-CNN fc_6	53.1%	
	R-CNN fc_7	54.2%	50.2%%
	R-CNN fc_7 (Bounding Box regression)	**58.5%**	53.7%

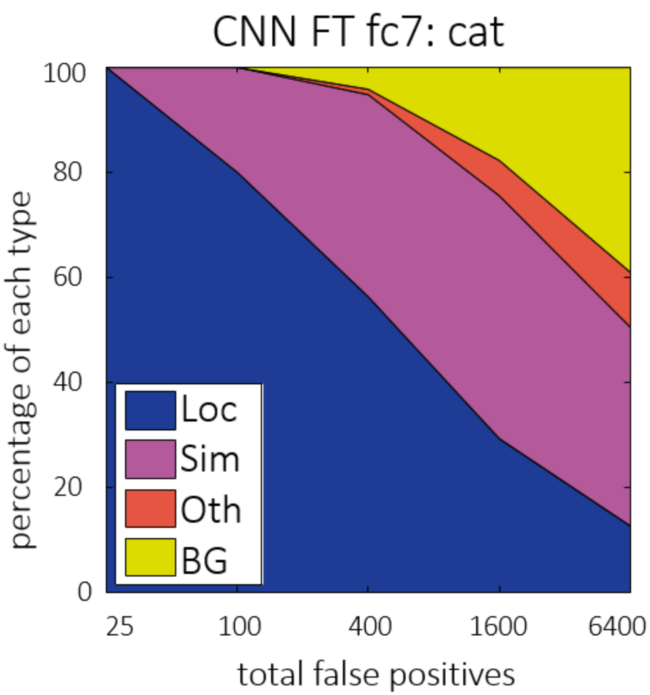
Evaluation: Top False Positives

Bicycle (AP 62.5%)

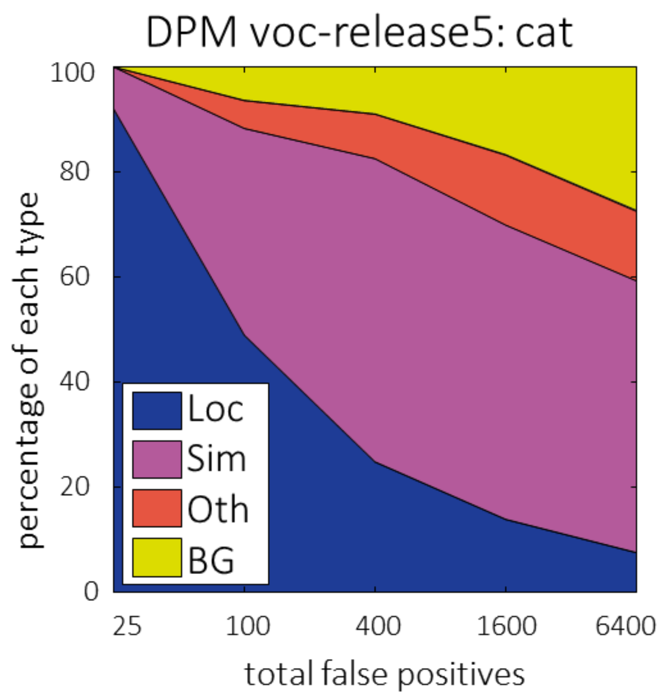
Evaluation: Top False Positives

Bird (AP 41.4%)

Evaluation: False positive types
Cat (AP 56.3%)



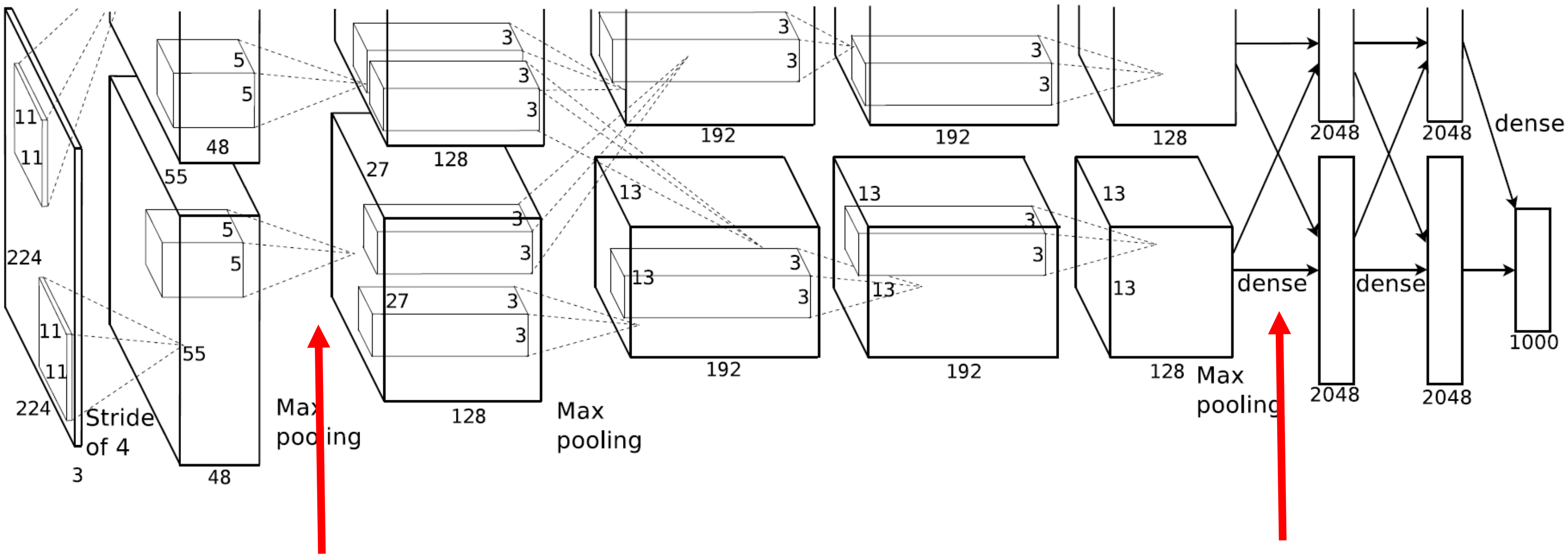
AP 56.3%



AP 23.0%

D. Hoiem et al., "Diagnosing Error in Object Detectors", ECCV 2012.

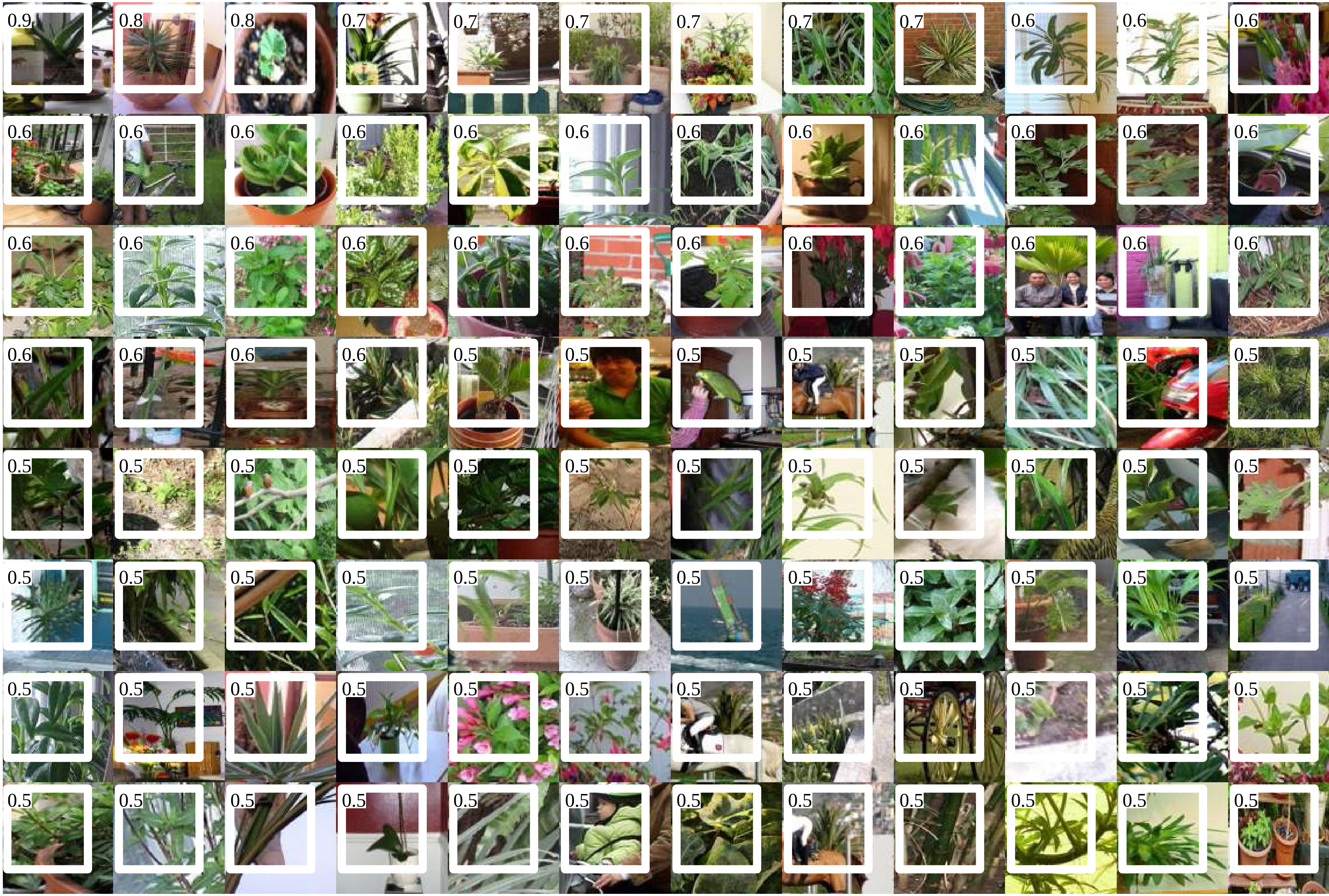
UToronto “SuperVision” CNN



pool5

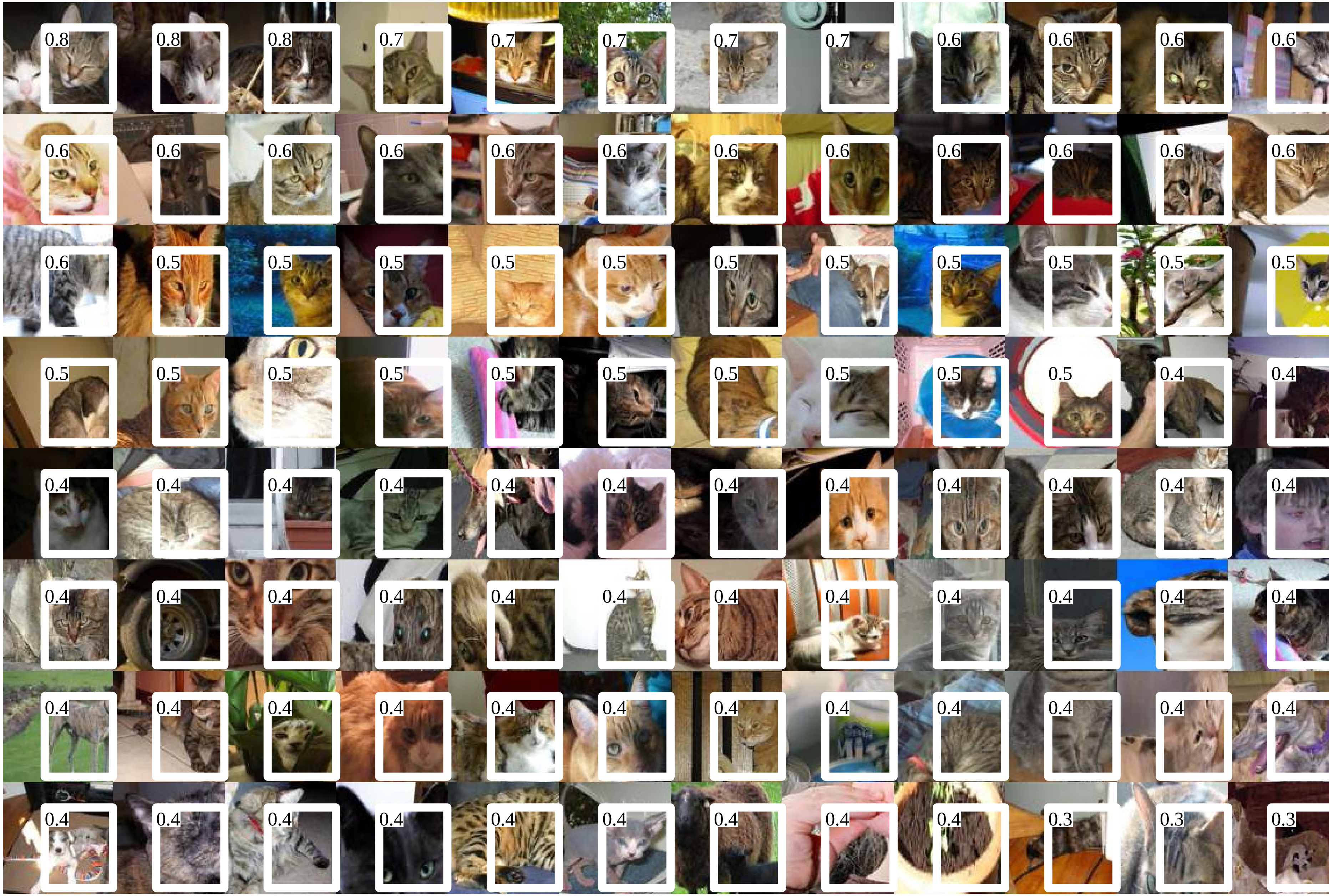
6x6x256 = 9216 dimensional

pool5 feature: (3,3,42) (top 1 − 96)

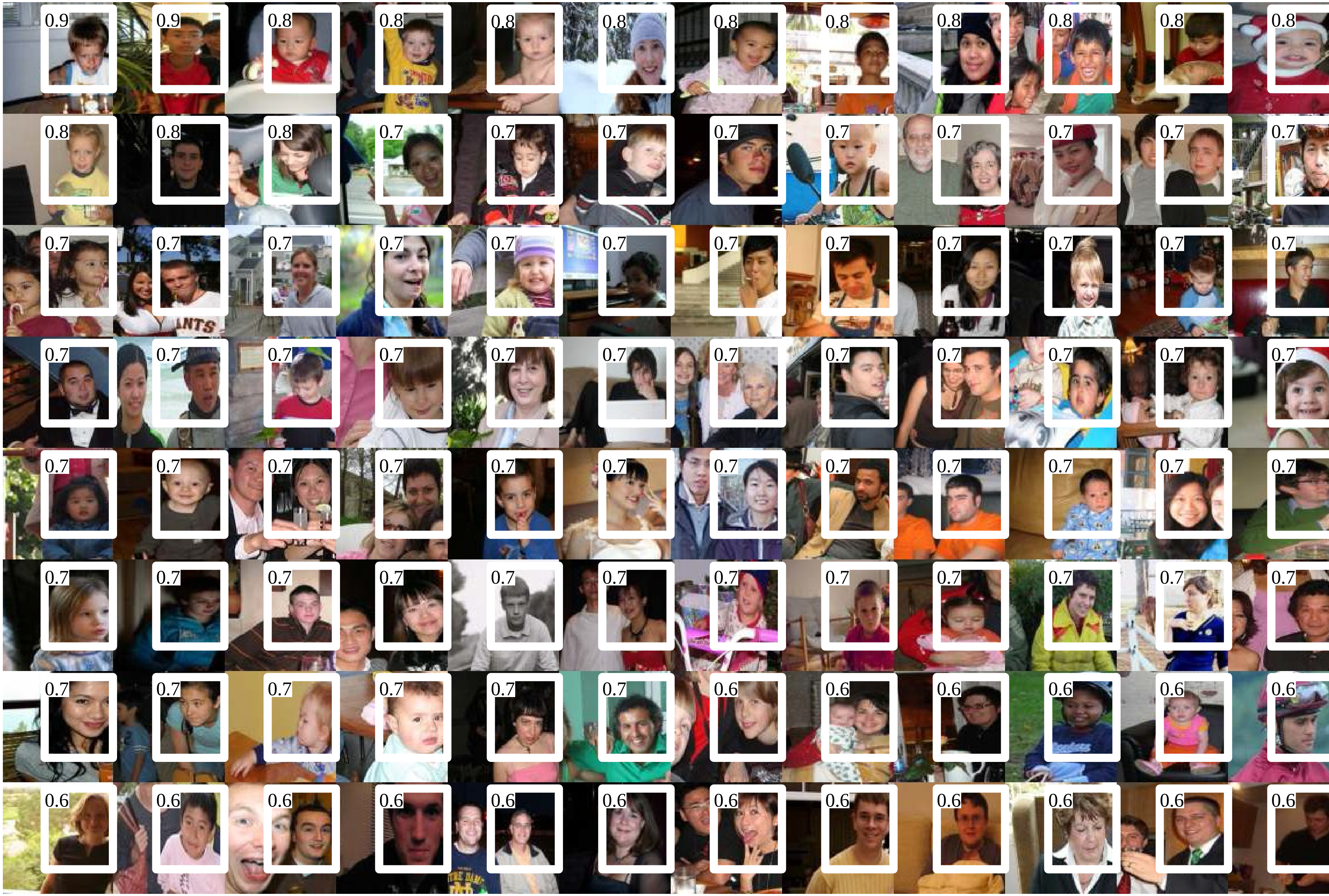


pool5 feature: (3,4,80) (top 1 − 96)

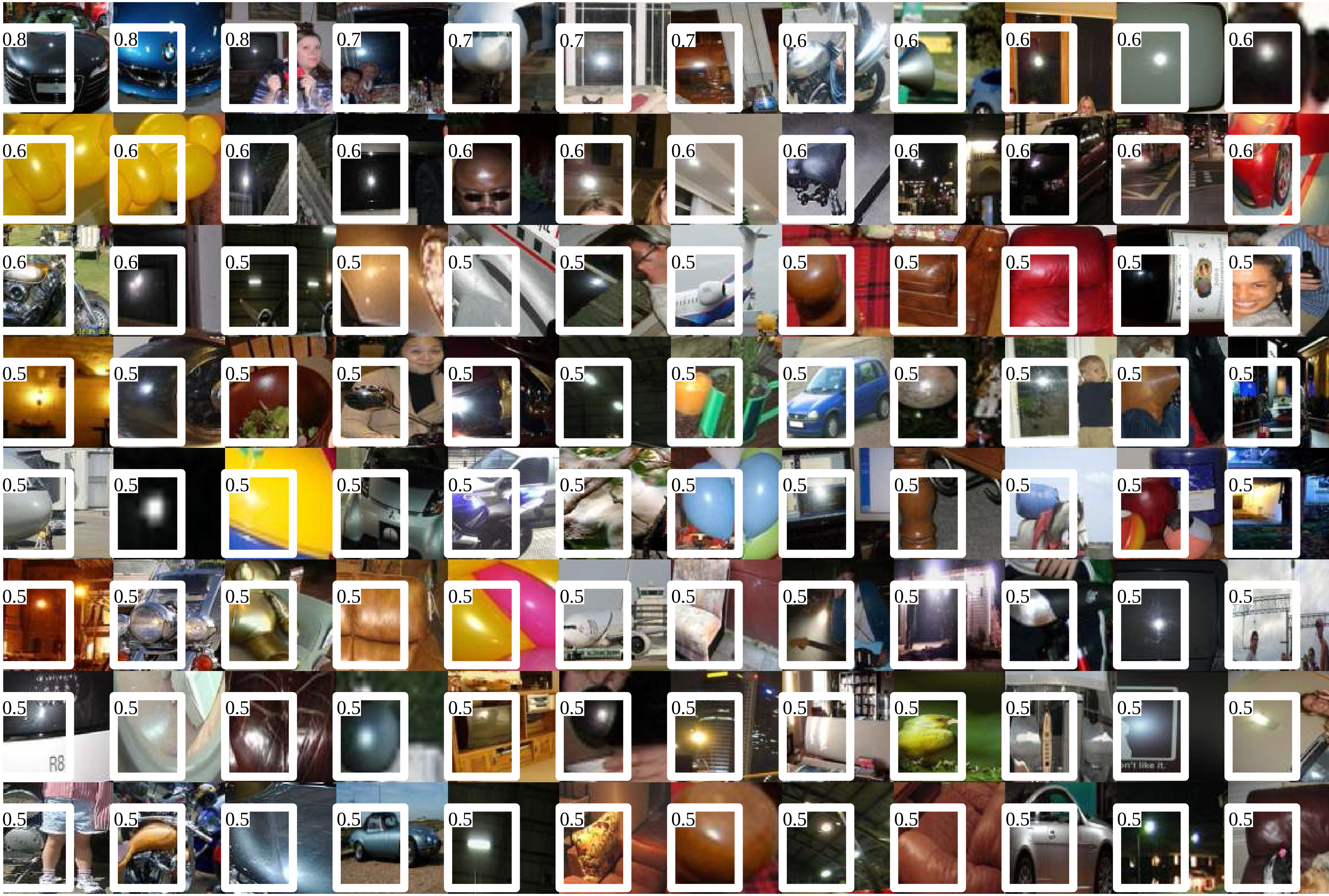
pool5 feature: (4,5,110) (top 1 − 96)



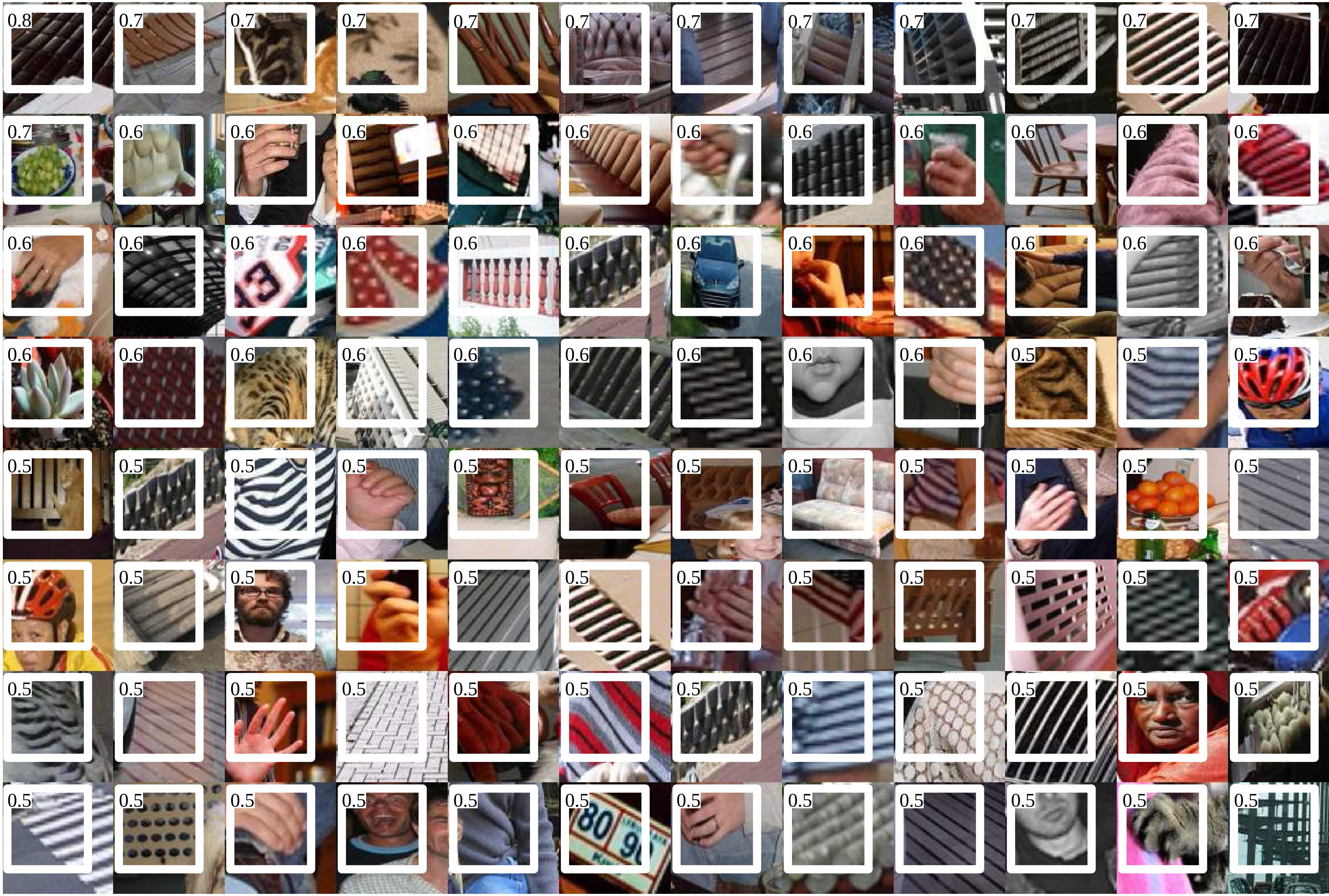
pool5 feature: (3,5,129) (top 1 − 96)



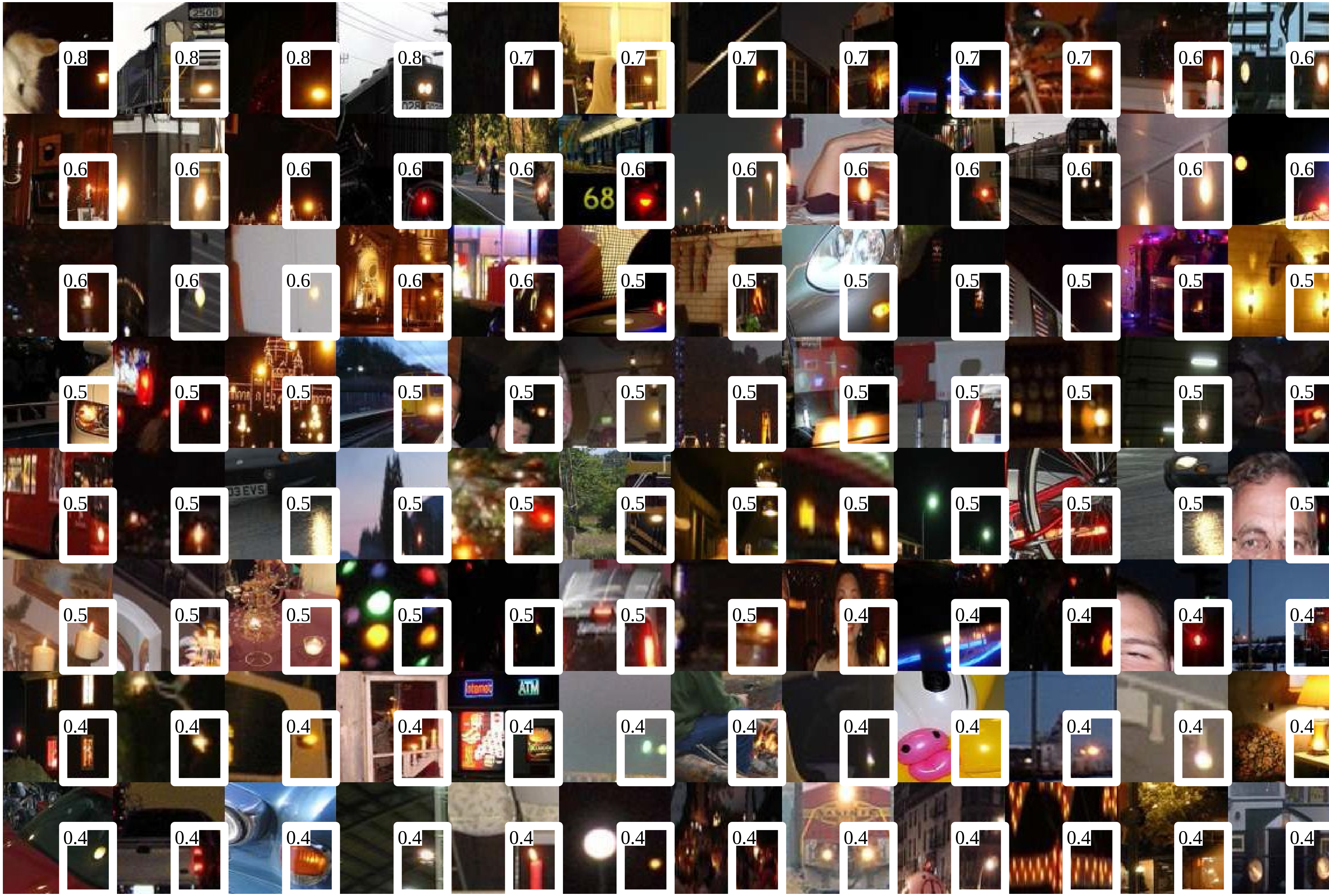
pool5 feature: (4,2,26) (top 1 − 96)



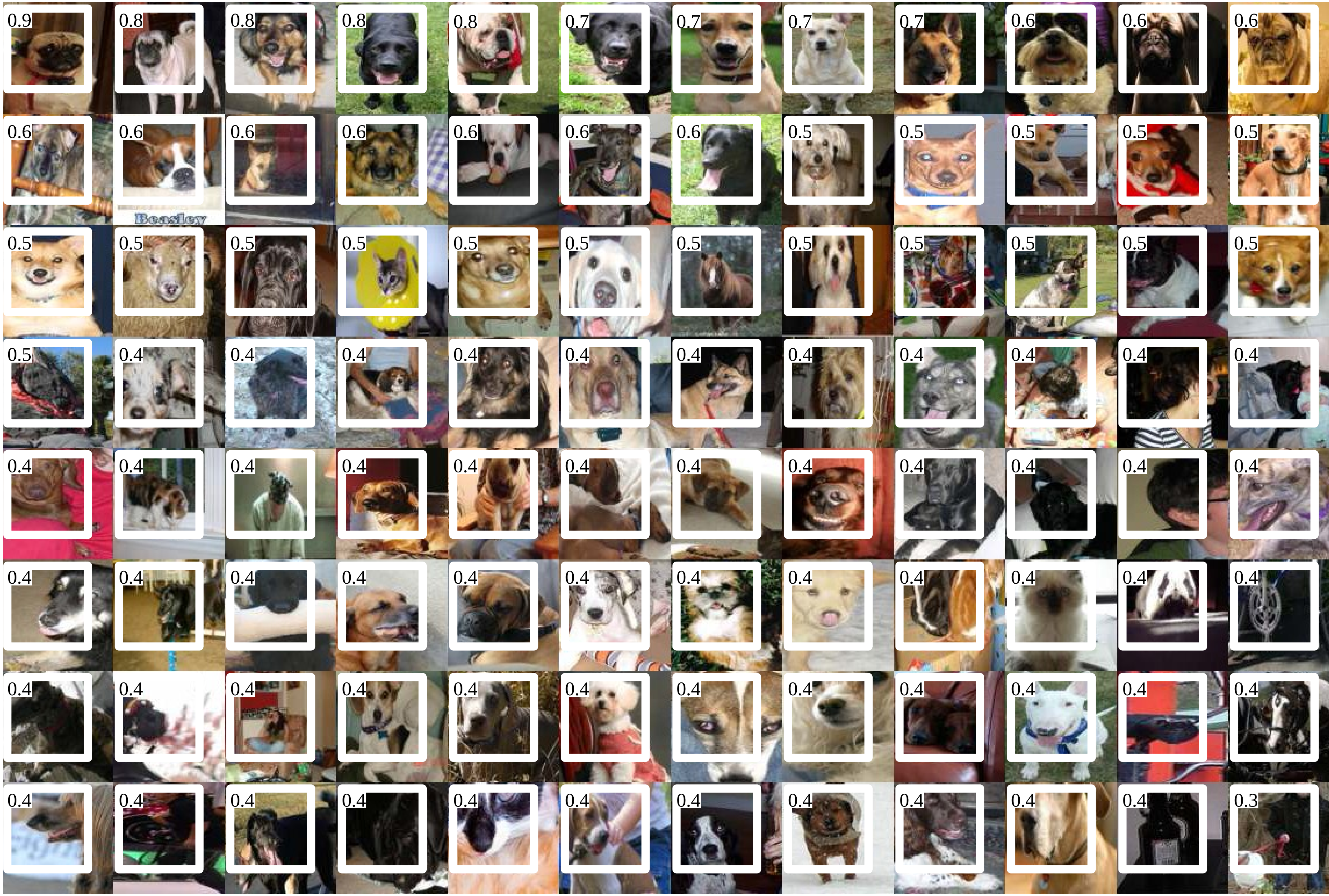
pool5 feature: (3,3,39) (top 1 − 96)



pool5 feature: (5,6,53) (top 1 − 96)

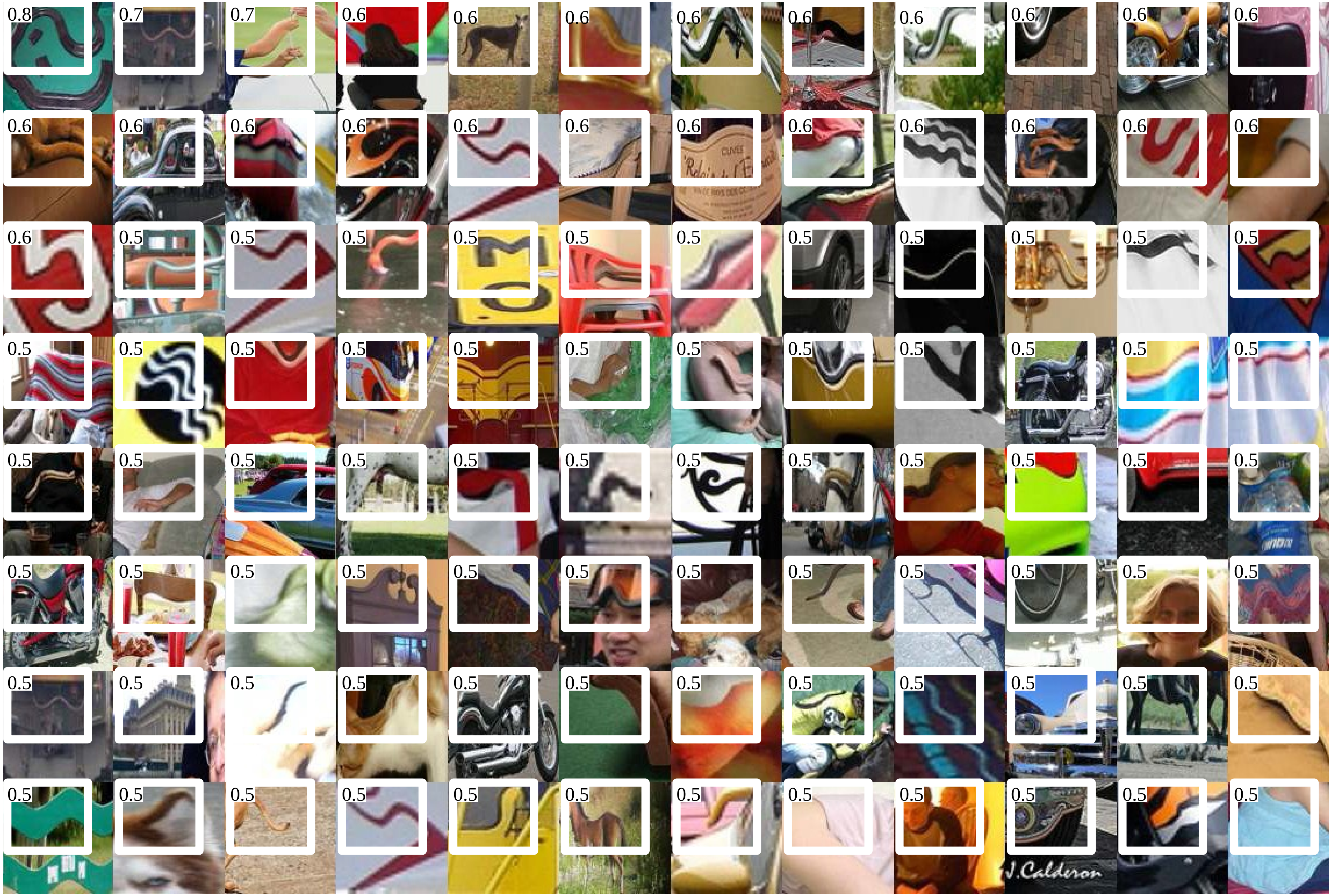


pool5 feature: (3,3,139) (top 1 − 96)



pool5 feature: (1,4,138) (top 1 − 96)

pool5 feature: (2,3,210) (top 1 − 96)



Discussion

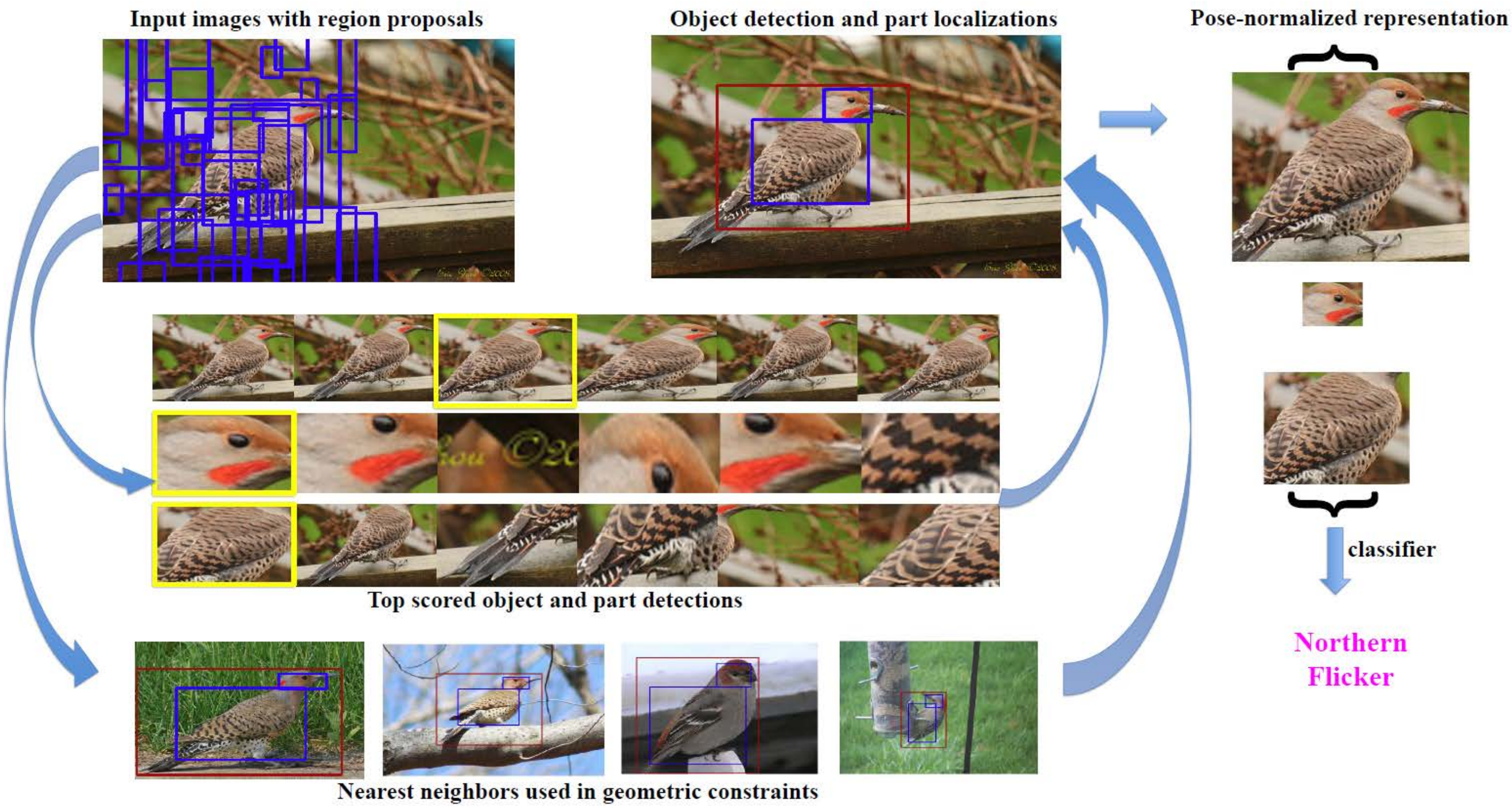
- Days of HOG, SIFT, LBP, and feature engineering are over?
- Machines can *design* better features than man?

Part-based R-CNNs for Fine-grained Category Detection

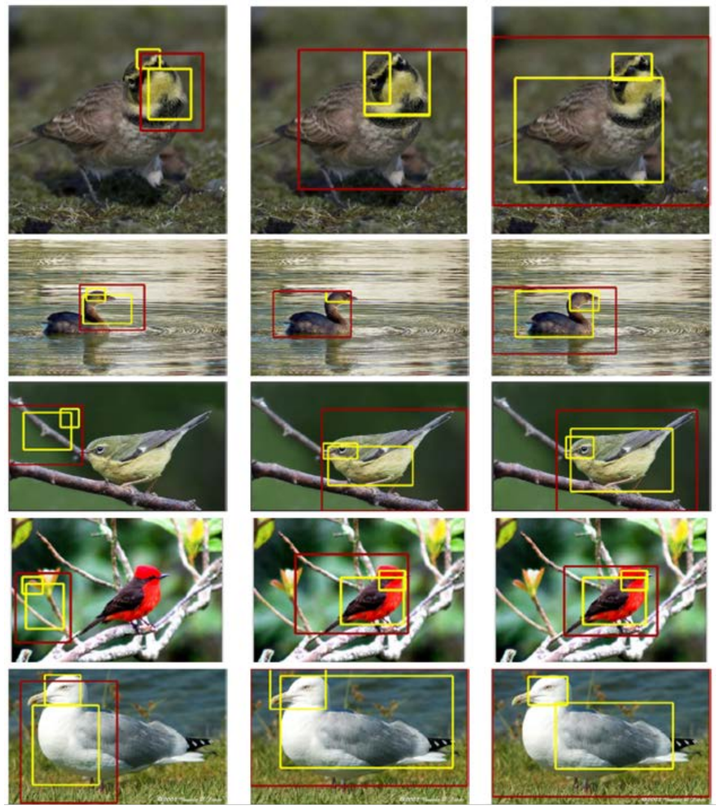
- Caltech-UCSD bird dataset (CUB200-2011) with ~12,000 images of 200 bird species.
- Strongly supervised setting in which ground truth bounding boxes of full objects (birds) and parts (head and body) are given.
- Each part + full object are treated as independent object categories to train SVMs in original R-CNN pipeline.
- Then geometric constraints (box + knn) are applied.

N. Zhang et al., “Part-based R-CNNs for Fine-grained Category Detection”, ECCV 2014.

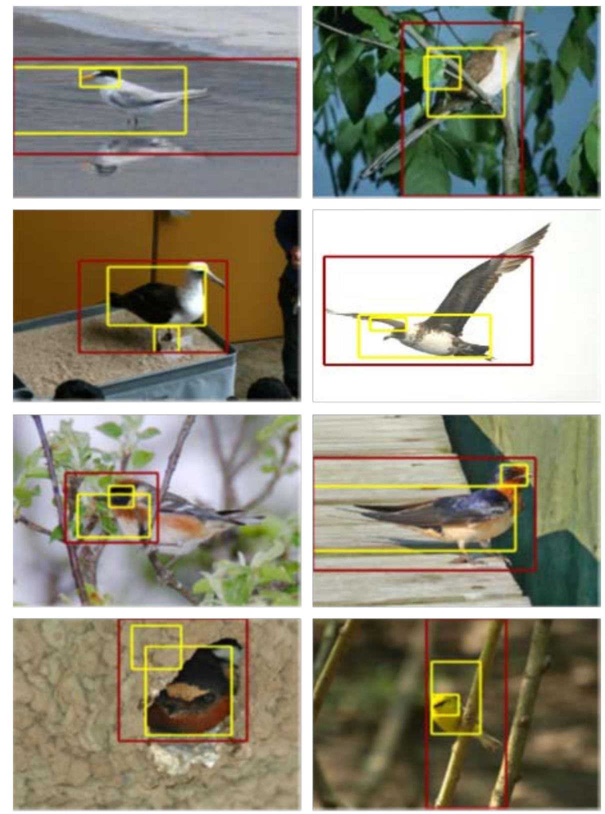
Part-based R-CNNs
for Fine-grained Category Detection



Part-based R-CNNs for Fine-grained Category Detection

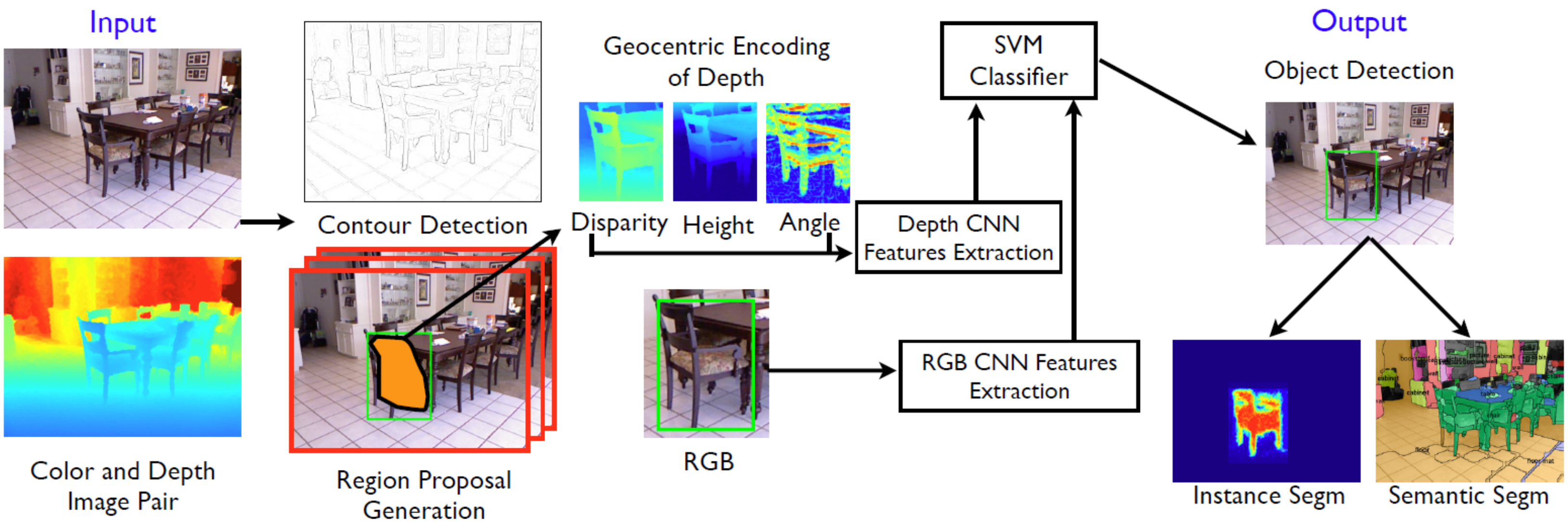


DPM | R-CNN w/box | R-CNN w/knn



some failures of R-CNN w/knn

R-CNNs on RGB-D
for Object Detection and Segmentation



Pre-trained on Image-Net using RGB images.
Fine-tuned on NYUD2 (400 images) and synthetic data.
SVM training on pool5, **fc6** and fc7.

S. Gupta et al., "Learning Rich Features from RGB-D Images for Object Detection and Segmentation", ECCV 2014.

R-CNNs on RGB-D
for Object Detection and Segmentation

Model	DPM	DPM	CNN	CNN	CNN	CNN	CNN	CNN	CNN	CNN	CNN	CNN
Fine-tuned			no	yes	no	yes	yes	yes	yes	yes	yes	yes
Input channels	RGB	RGBD	RGB	RGB	disp	disp	HHA	HHA	HHA	HHA	HHA	RGB+ HHA
synth data								2x	15x	2x	2x	2x
CNN layer			fc6	fc6	fc6	fc6	fc6	fc6	fc6	pool5	fc7	fc6
mAP	8.4	21.7	16.4	**19.7**	11.3	20.1	25.2	**26.1**	25.6	21.9	25.3	**32.5**

HHA: **H**orizontal disparity,
Height above ground,
Angle the pixel's local surface normal makes with the inferred gravity direction.

R-CNNs on RGB-D