RICH FEATURE HIERARCHIES FOR ACCURATE OBJECT DETECTION AND SEMANTIC SEGMENTATION

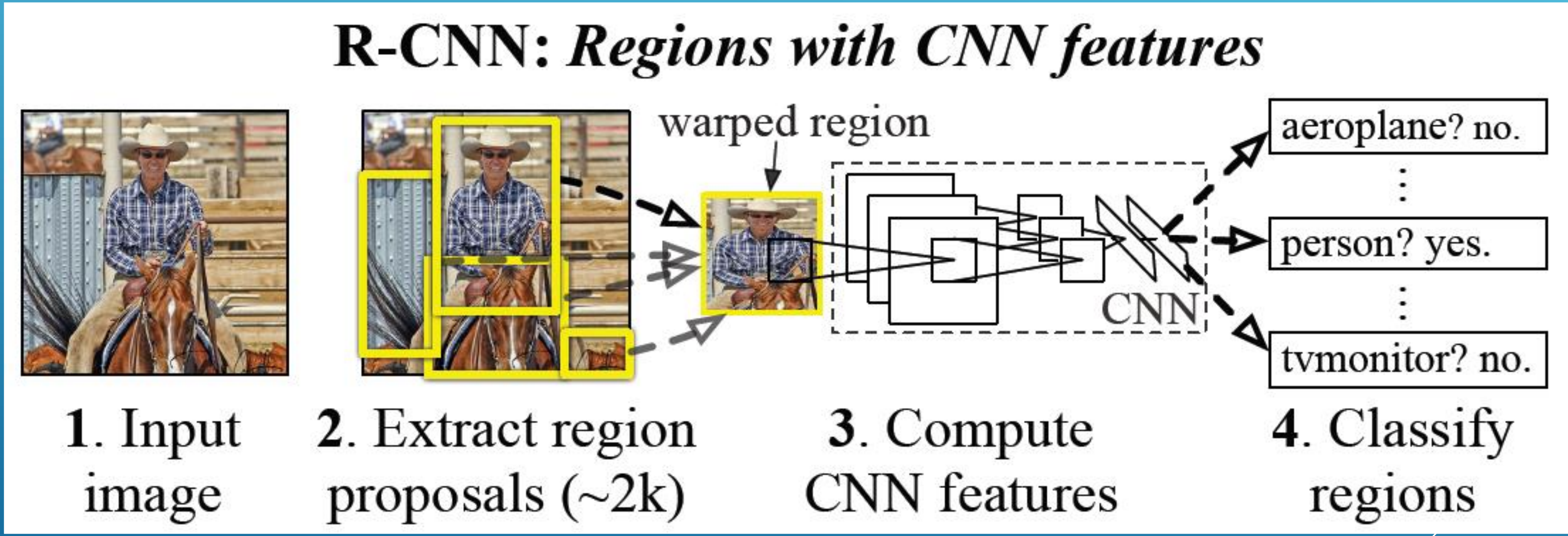
Ross Girshick, Jeff Donahue, Trevor Darrell, Jitandra Malik (UC Berkeley)

Presenter: Hossein Azizpour

ABSTRACT

- Can CNN improve s.o.a. object detection results?
  - Yes, it helps by learning rich representations which can then be combined with computer vision techniques.
- Can we understand what does a CNN learn?
  - Sort of!, we can check which positive (or negative) image regions stimulates a neuron the most
- It will evaluate different layers of the method
- Experiments on segmentation
- mAP on VOC 2007: ***48%*** !

APPROACH

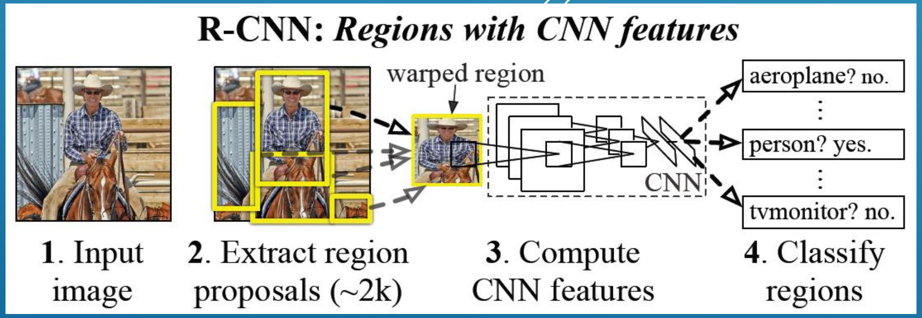

R-CNN: *Regions with CNN features*
warped region
aeroplane? no.
person? yes.
tvmonitor? no.
CNN
1. Input image
2. Extract region proposals (~2k)
3. Compute CNN features
4. Classify regions

REGION PROPOSALS

- over segmentation (initial regions)
- bottom-up grouping at multiple scales
- Diversifications (different region proposals, similarity for grouping,…)
- Enables computationally expensive methods
- Potentially reduce false positives

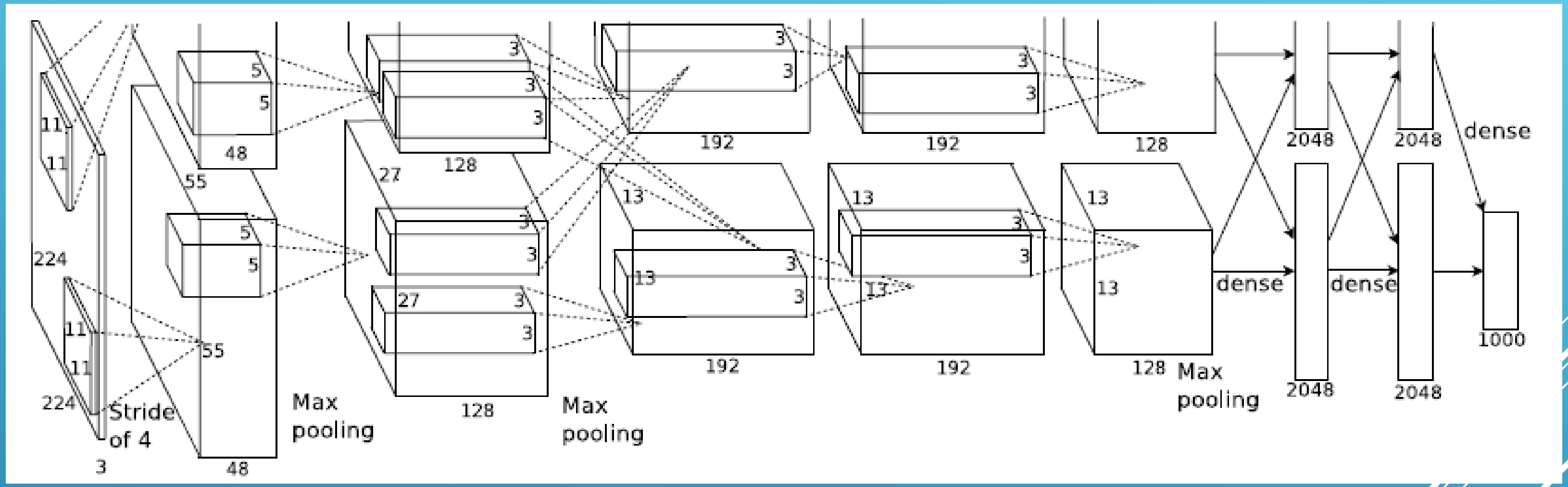
method	recall	MABO	# windows
Arbelaez *et al.* [3]	0.752	0.649 ± 0.193	418
Alexe *et al.* [2]	0.944	0.694 ± 0.111	1,853
Harzallah *et al.* [16]	0.830	-	200 per class
Carreira and Sminchisescu [4]	0.879	0.770 ± 0.084	517
Endres and Hoiem [9]	0.912	0.791 ± 0.082	790
Felzenszwalb *et al.* [12]	0.933	0.829 ± 0.052	100,352 per class
Vedaldi *et al.* [34]	0.940	-	10,000 per class
Single Strategy	0.840	0.690 ± 0.171	289
Selective search "Fast"	0.980	0.804 ± 0.046	2,134
Selective search "Quality"	0.991	0.879 ± 0.039	10,097

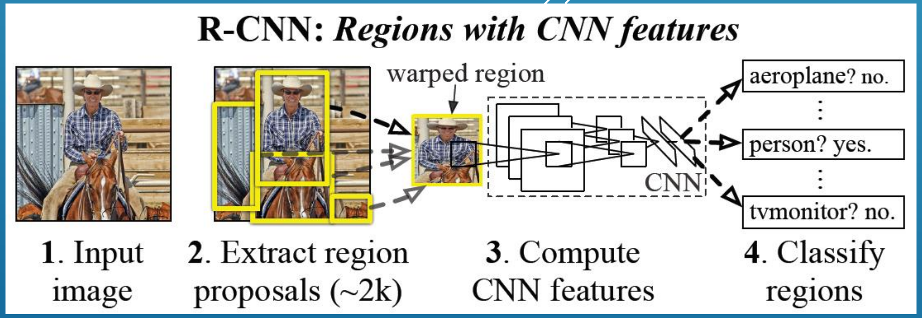
Version	Diversification Strategies	MABO	# win	# strategies	time (s)
Single Strategy	HSV C+T+S+F $k = 100$	0.693	362	1	0.71
Selective Search Fast	HSV, Lab C+T+S+F, T+S+F $k = 50, 100$	0.799	2147	8	3.79
Selective Search Quality	HSV, Lab, rgI, H, I C+T+S+F, T+S+F, F, S $k = 50, 100, 150, 300$	0.878	10,108	80	17.15



CNN PRE-TRAINING

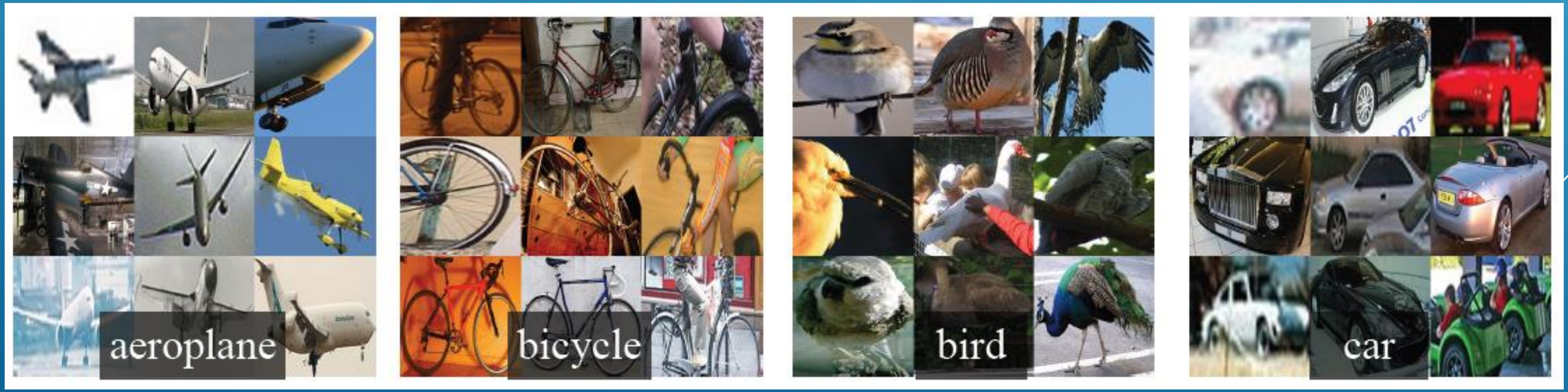
- Rectified non-linearity
- Local Response Normalization
- Overlapping max pooling
- 5 convolutional layers
- 2 fully connected layers
- Softmax
- Drop out
- 224x224x3 input
- ImageNet samples





CNN FINE-TUNING

- lower learning rate (1/100)
- only pascal image regions
- 128 patch per image
- Positives: overlap >= 0.5, Negative otherwise
-



LEARNING CLASSIFIER

- Positives: full patches
- Negatives: overlap < 0.3 (***very important!***)
- Linear SVM per each class
- Standard hard negative mining
- Pre-computed and saved features

TIMING

- Training SVM for all classes on a single core takes 1.5 hours
- Extracting feature for a window on GPU takes 5 ms
- Inference requires a matrix multiplication, for 100K classes it takes 10 secs
- Compared to Google Dean et al. paper (CVPR best paper): 16% mAP in 5 minutes. Here 48% in about 1 minute!

DETECTION RESULTS

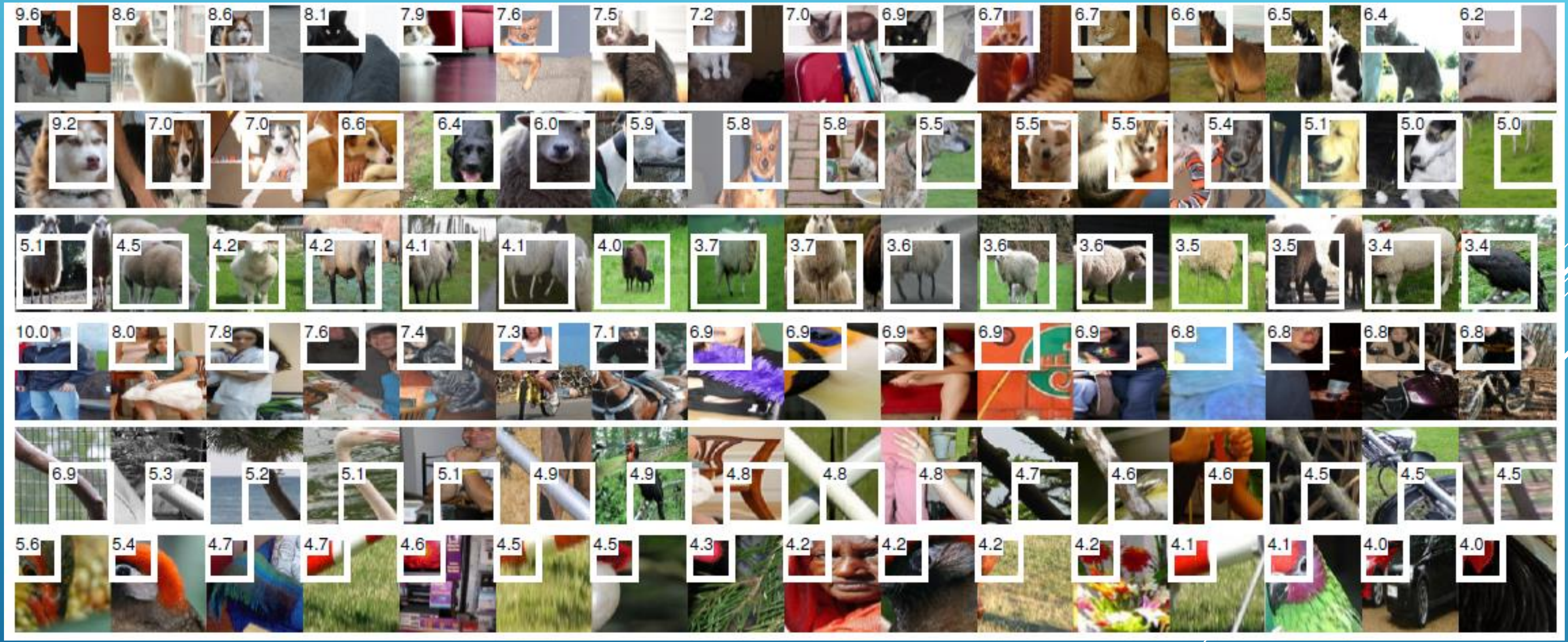
VOC 2010 test	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
DPM HOG [19]	45.6	49.0	11.0	11.6	27.2	50.5	43.1	23.6	17.2	23.2	10.7	20.5	42.5	44.5	41.3	8.7	29.0	18.7	40.0	34.5	29.6
SegDPM [18]	56.4	48.0	24.3	21.8	**31.3**	**51.3**	47.3	48.2	16.1	29.4	19.0	37.5	44.1	51.5	44.4	12.6	32.1	28.8	**48.9**	39.1	36.6
UVA [36]	56.2	42.4	15.3	12.6	21.8	49.3	36.8	46.1	12.9	32.1	30.0	36.5	43.5	52.9	32.9	15.3	41.1	**31.8**	47.0	44.8	35.1
ours (R-CNN FT fc_7)	**65.4**	**56.5**	**45.1**	**28.5**	24.0	50.1	**49.1**	**58.3**	**20.6**	**38.5**	**31.1**	**57.5**	**50.7**	**60.3**	**44.7**	**21.6**	**48.5**	24.9	48.0	**46.5**	**43.5**

- Pascal 2010
- UVA uses the same region proposals with large combined descriptors and HIK SVM

VISUALIZATION

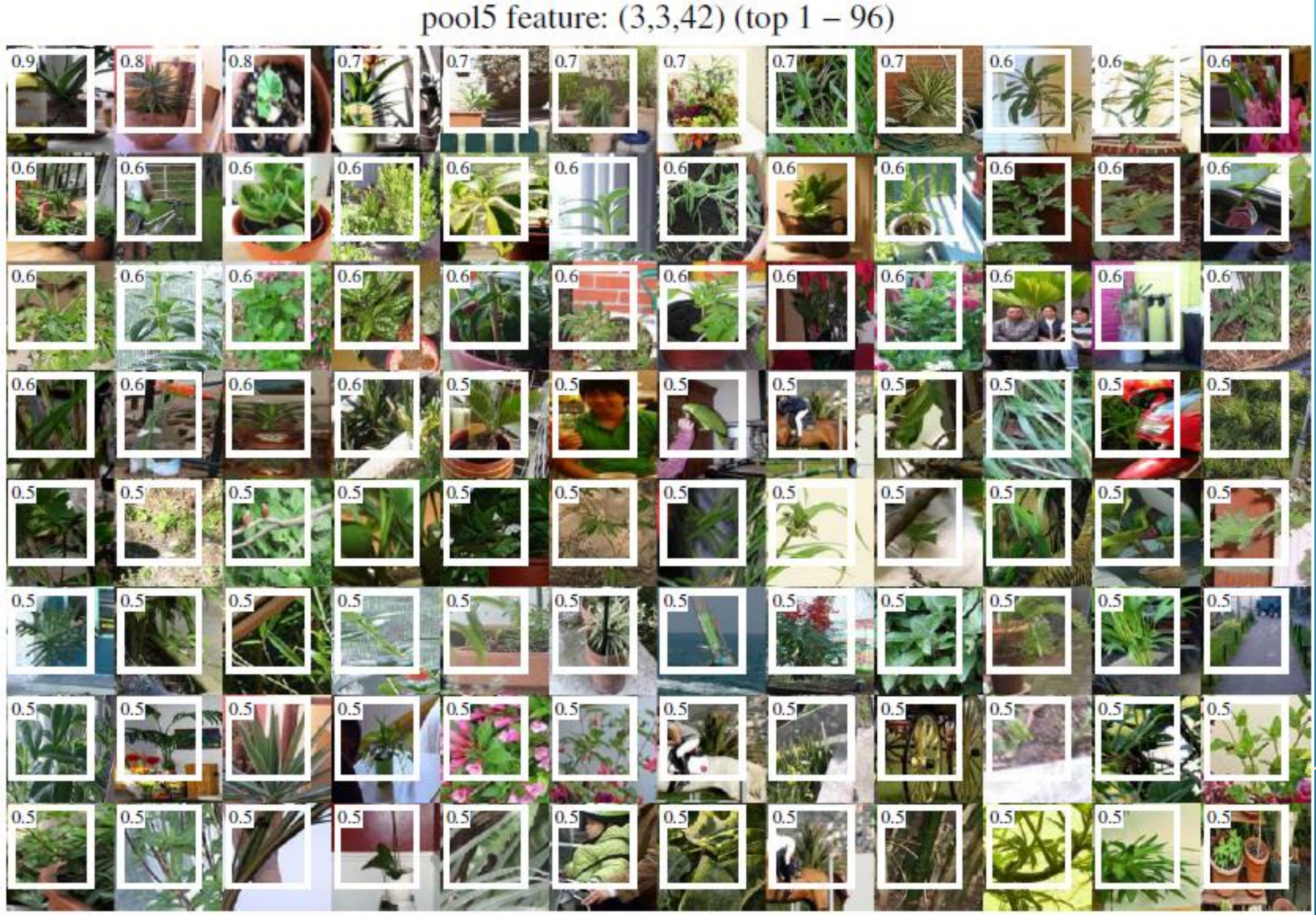
- 10 million held-out regions
- sort by the activation response
- potentially shows modes and invariances
- max pool layer #5 (6x6x256=9216D)

VISUALIZATION



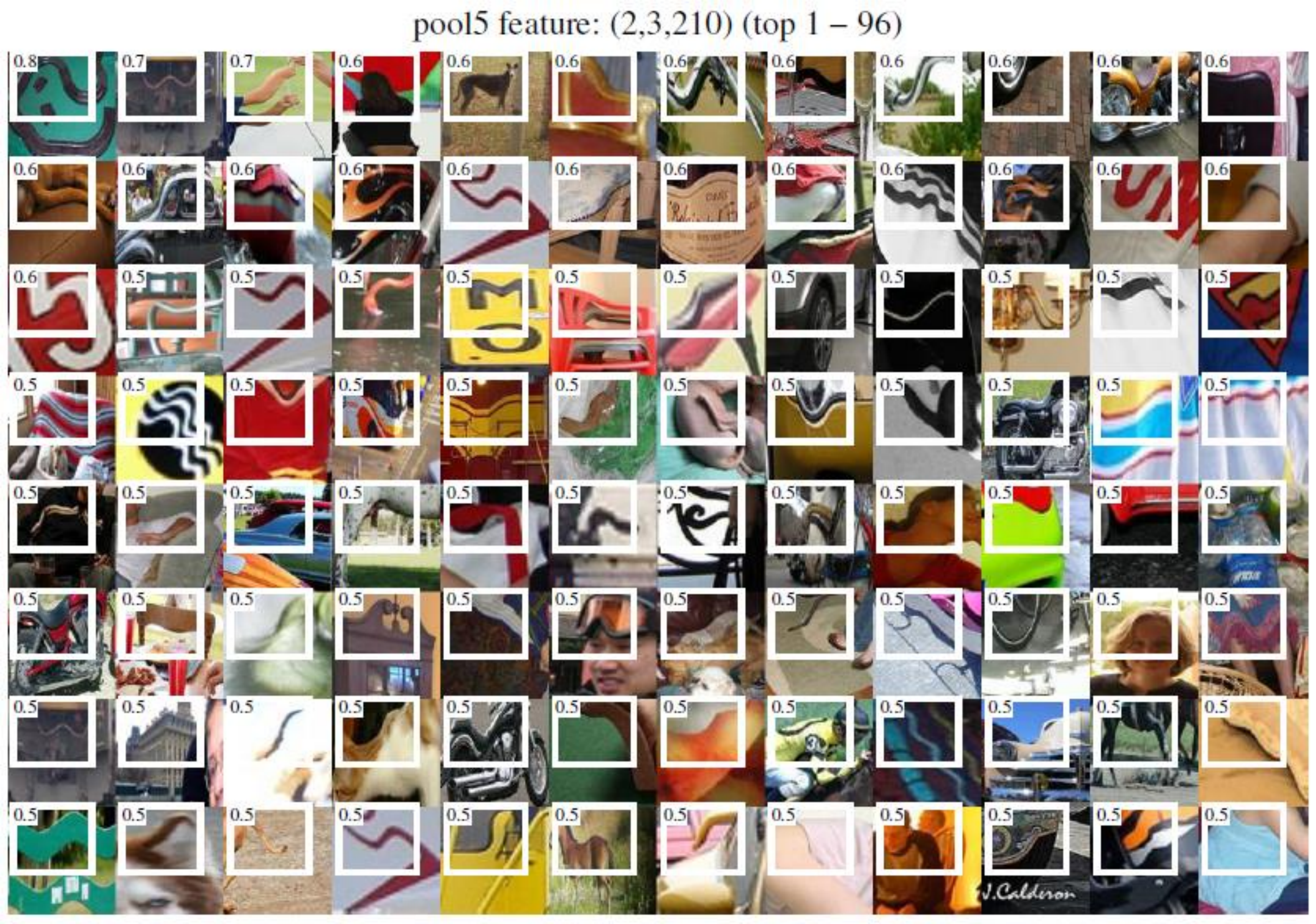
- 1- Cat (positive SVM weight) 2- Cat (negative SVM weight) 3- Sheep (Positive SVM Weight)
- 4- Person (positive SVM weight) 5,6- Some generic unit (diagonal bars, red blobs)

VISUALIZATION



VISUALIZATION

VISUALIZATION



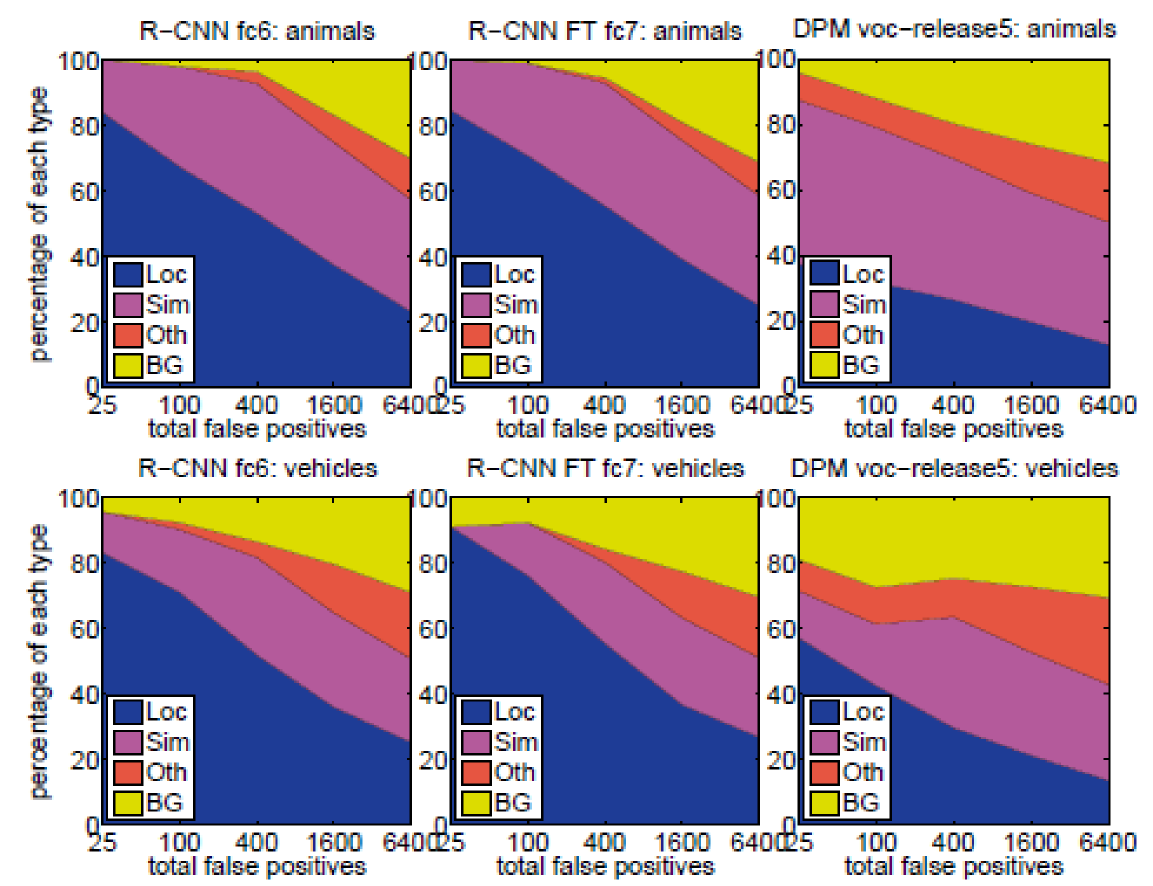
ABLATION STUDY

- With and without fine tuning on different layers
- Pool 5 (only **6%** of all parameters, out of ~60 million parmeters)
- No Color: (grayscale pascal input): 43.4% → 40.1% mAP

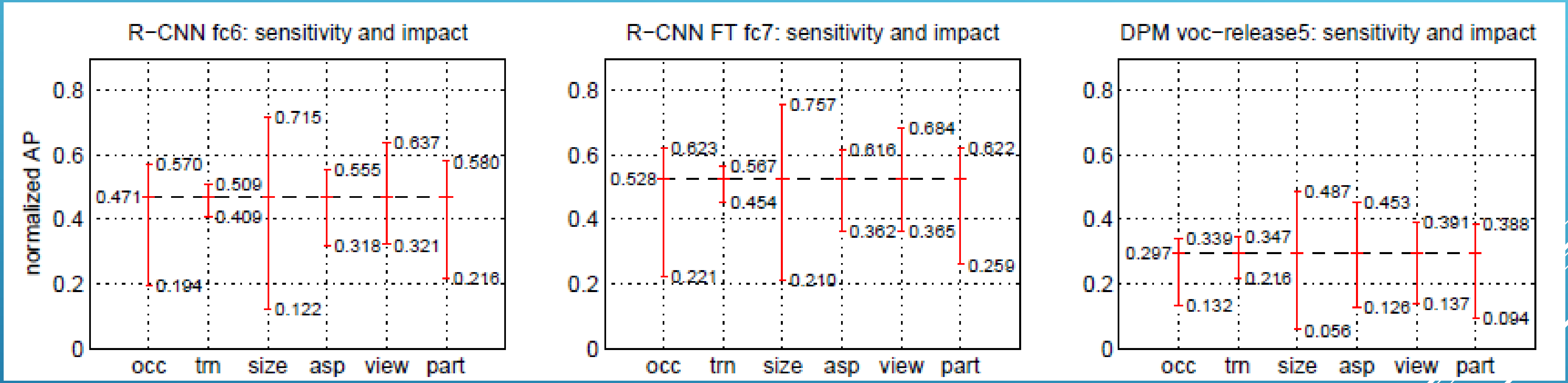
VOC 2007 test	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
R-CNN $pool_5$	49.3	58.0	29.7	22.2	20.6	47.7	56.8	43.6	16.0	39.7	37.7	39.6	49.6	55.6	37.5	20.6	40.5	37.4	47.8	51.3	40.1
R-CNN fc_6	56.1	58.8	34.4	29.6	22.6	50.4	58.0	52.5	18.3	40.1	41.3	46.8	49.5	53.5	39.7	23.0	46.4	36.4	50.8	59.0	43.4
R-CNN fc_7	53.1	58.9	35.4	29.6	22.3	50.0	57.7	52.4	19.1	43.5	40.8	43.6	47.6	54.0	39.1	23.0	42.3	33.6	51.4	55.2	42.6
R-CNN FT $pool_5$	55.6	57.5	31.5	23.1	23.2	46.3	59.0	49.2	16.5	43.1	37.8	39.7	51.5	55.4	40.4	23.9	46.3	37.9	49.7	54.1	42.1
R-CNN FT fc_6	61.8	62.0	38.8	35.7	29.4	52.5	**61.9**	53.9	22.6	49.7	40.5	**48.8**	49.9	**57.3**	44.5	**28.5**	50.4	**40.2**	54.3	61.2	47.2
R-CNN FT fc_7	**60.3**	**62.5**	**41.4**	**37.9**	29.0	**52.6**	61.6	**56.3**	**24.9**	**52.3**	**41.9**	48.1	54.3	57.0	**45.0**	26.9	**51.8**	38.1	**56.6**	**62.2**	**48.0**
DPM HOG [19]	33.2	60.3	10.2	16.1	27.3	54.3	58.2	23.0	20.0	24.1	26.7	12.7	**58.1**	48.2	43.2	12.0	21.1	36.1	46.0	43.5	33.7
DPM ST [29]	23.8	58.2	10.5	8.5	27.1	50.4	52.0	7.3	19.2	22.8	18.1	8.0	55.9	44.8	32.4	13.3	15.9	22.8	46.2	44.9	29.1
DPM HSC [32]	32.2	58.3	11.5	16.3	**30.6**	49.9	54.8	23.5	21.5	27.7	34.0	13.7	**58.1**	51.6	39.9	12.4	23.5	34.4	47.4	45.2	34.3

DETECTION ERROR ANALYSIS

- Compared to DPM, more of the FPs come from poor localization
- Animals: fine-tuning reduces the confusion with other animals
- Vehicles: fine-tuning reduces the confusion with other animals amongst the high scoring FPs



DETECTION ERROR ANALYSIS



- Sensitivity is the same, but we see improvements, in general, for all of the subsets

SEGMENTATION

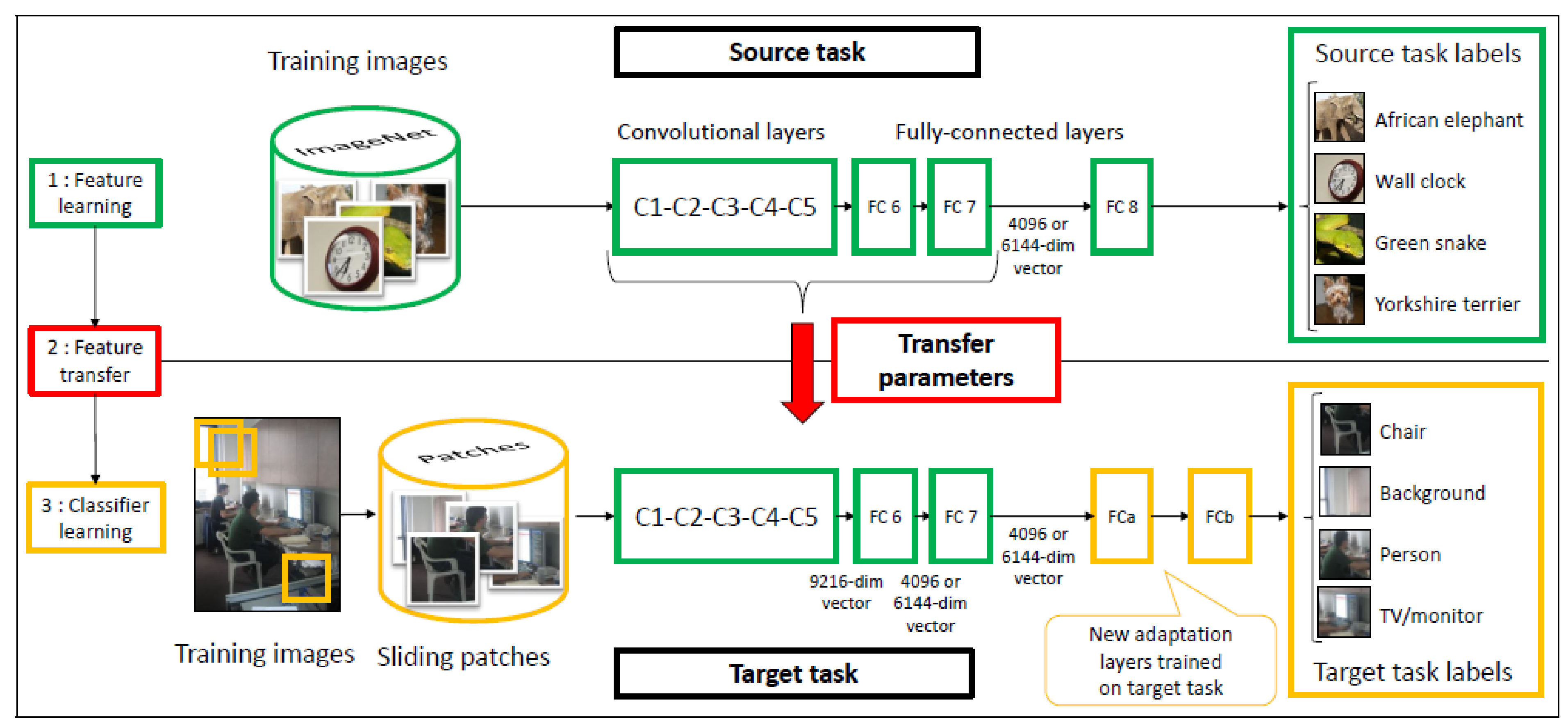
- CPMC region proposals
- SVR
- Compared to s.o.a. O2P
- VOC 2011
- 3 versions, full, foreground, full+foreground
- Fc6 better than fc7
- O2P takes 10 hours, CNN takes 1 hour

	full R-CNN		*fg* R-CNN		*full+fg* R-CNN	
O_2P [5]	fc_6	fc_7	fc_6	fc_7	fc_6	fc_7
46.4	43.0	42.5	43.7	42.1	**47.9**	45.8

VOC 2011 test	bg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mean
R&P [2]	83.4	46.8	18.9	36.6	31.2	42.7	57.3	47.4	44.1	8.1	39.4	**36.1**	36.3	49.5	48.3	50.7	26.3	47.2	22.1	42.0	43.2	40.8
O_2P [5]	**85.4**	**69.7**	22.3	45.2	**44.4**	46.9	66.7	57.8	56.2	**13.5**	**46.1**	32.3	41.2	**59.1**	55.3	51.0	**36.2**	50.4	**27.8**	46.9	**44.6**	47.6
ours (*full+fg* R-CNN fc_6)	84.2	66.9	**23.7**	**58.3**	37.4	**55.4**	**73.3**	**58.7**	**56.5**	9.7	45.5	29.5	**49.3**	40.1	**57.8**	**53.9**	33.8	**60.7**	22.7	**47.1**	41.3	**47.9**

LEARNING AND TRANSFERRING MID-LEVEL IMAGE REPRESENTATIONS USING CONVOLUTIONAL NEURAL NETWORKS

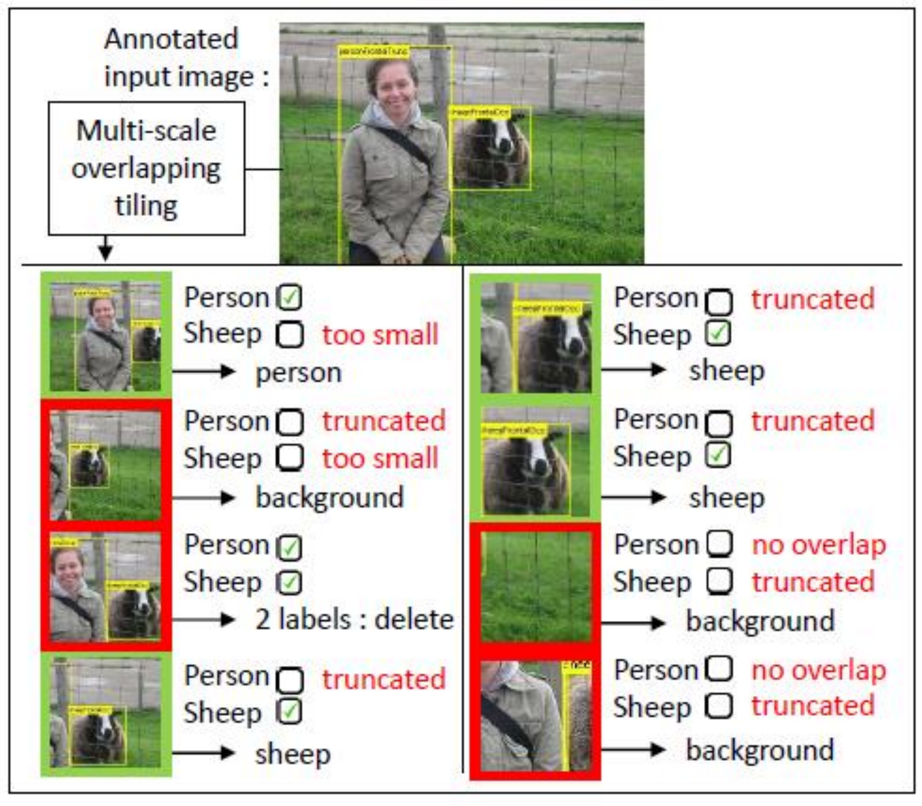
Maxime Oquab, Leon Bottou, Ivan Laptev, Josef Sivic (INRIA, WILLOW)



APPROACH

- Dense sampling of 500 patches per image instead of segmented regions
- Different positive/negative criteria
- Resampling positives to make the balance
- Classification

$$\text{score}(C_n) = \frac{1}{M}\sum_{i=1}^{M} y(C_n|P_i)^k,$$



FINAL RESULTS

	plane	bike	bird	boat	btl	bus	car	cat	chair	cow	table	dog	horse	moto	pers	plant	sheep	sofa	train	tv	mAP
INRIA [32]	77.5	63.6	56.1	71.9	33.1	60.6	78.0	58.8	53.5	42.6	54.9	45.8	77.5	64.0	85.9	36.3	44.7	50.6	79.2	53.2	59.4
NUS-PSL [44]	82.5	79.6	64.8	73.4	**54.2**	75.0	77.5	79.2	46.2	62.7	41.4	74.6	**85.0**	76.8	91.1	53.9	61.0	**67.5**	83.6	70.6	70.5
PRE-1000C	**88.5**	**81.5**	**87.9**	**82.0**	47.5	**75.5**	**90.1**	**87.2**	**61.6**	**75.7**	**67.3**	**85.5**	83.5	**80.0**	**95.6**	**60.8**	**76.8**	58.0	**90.4**	**77.9**	**77.7**

Table 1: Per-class results for object classification on the VOC2007 test set (average precision %).

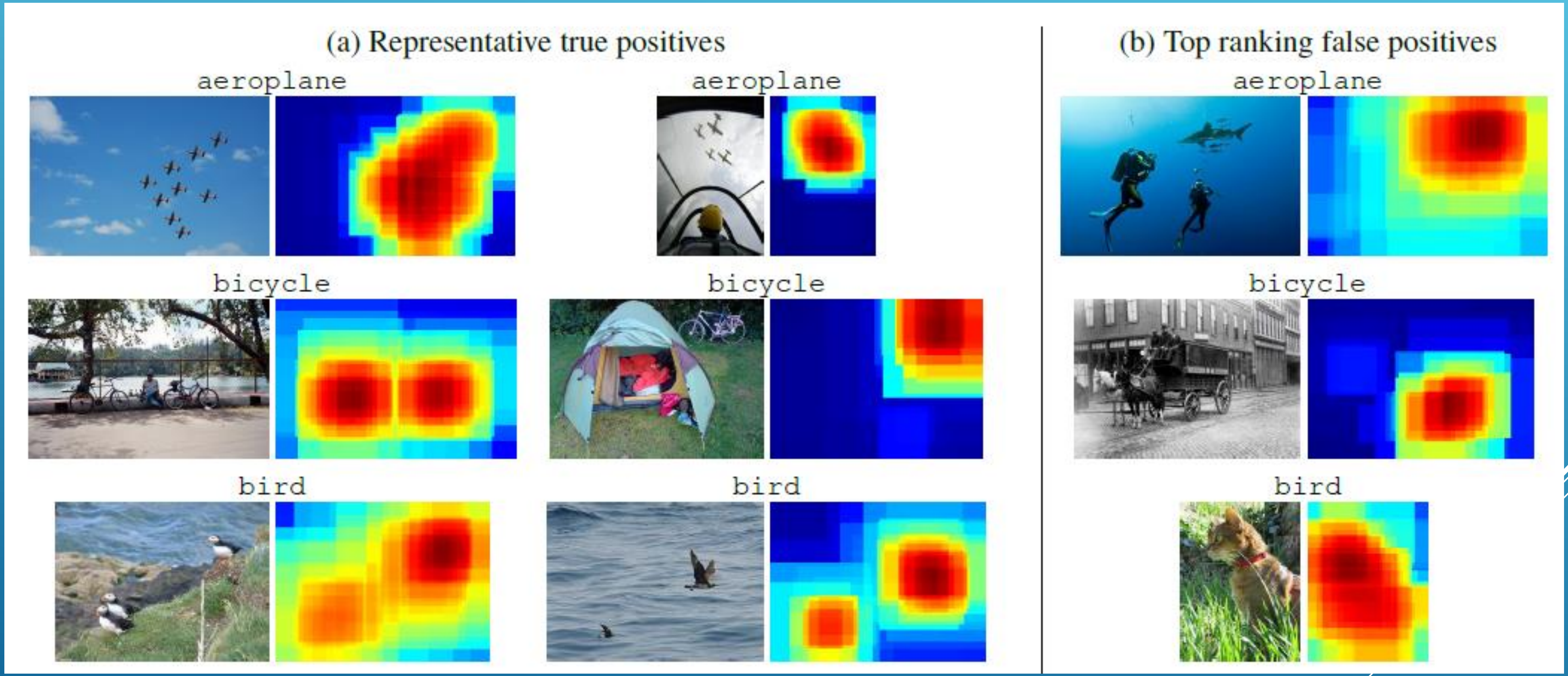
	plane	bike	bird	boat	btl	bus	car	cat	chair	cow	table	dog	horse	moto	pers	plant	sheep	sofa	train	tv	mAP
NUS-PSL [49]	**97.3**	**84.2**	80.8	**85.3**	**60.8**	**89.9**	**86.8**	89.3	**75.4**	77.8	**75.1**	83.0	87.5	**90.1**	95.0	57.8	79.2	**73.4**	**94.5**	**80.7**	82.2
NO PRETRAIN	85.2	75.0	69.4	66.2	48.8	82.1	79.5	79.8	62.4	61.9	49.8	75.9	71.4	82.7	93.1	59.1	69.7	49.3	80.0	76.7	70.9
PRE-1000C	93.5	78.4	*87.7*	80.9	57.3	85.0	81.6	*89.4*	66.9	73.8	62.0	*89.5*	83.2	87.6	*95.8*	*61.4*	79.0	54.3	88.0	78.3	78.7
PRE-1000R	93.2	77.9	83.8	80.0	55.8	82.7	79.0	84.3	66.2	71.7	59.5	83.4	81.4	84.8	95.2	59.8	74.9	52.9	83.8	75.7	76.3
PRE-1512	94.6	82.9	**88.2**	84.1	60.3	89.0	84.4	**90.7**	72.1	**86.8**	69.0	**92.1**	**93.4**	88.6	**96.1**	**64.3**	**86.6**	62.3	91.1	79.8	**82.8**

Table 2: Per-class results for object classification on the VOC2012 test set (average precision %).

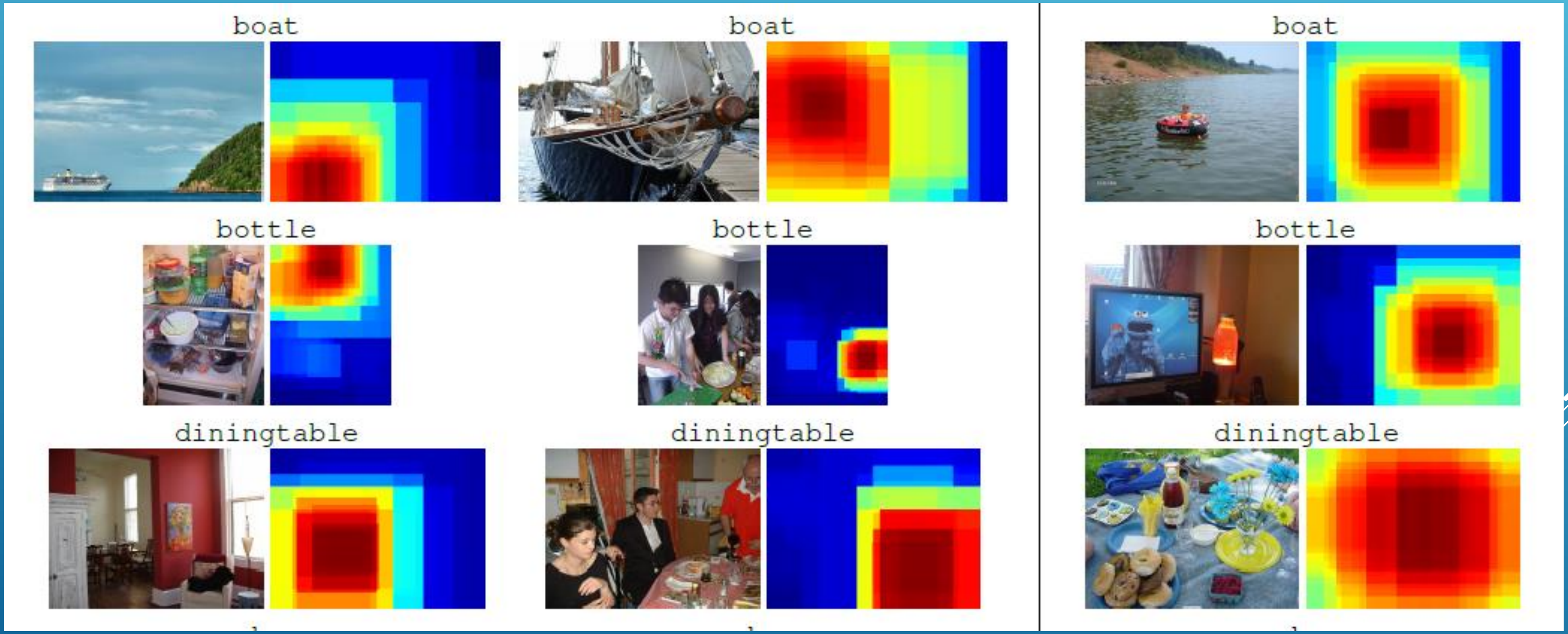
Action	jump	phon	instr	read	bike	horse	run	phot	comp	walk	mAP
STANFORD [1]	75.7	44.8	66.6	44.4	93.2	94.2	87.6	38.4	**70.6**	**75.6**	69.1
OXFORD [1]	**77.0**	**50.4**	65.3	39.5	**94.1**	**95.9**	**87.7**	42.7	68.6	74.5	69.6
NO PRETRAIN	43.2	30.6	50.2	25.0	76.8	80.7	75.2	22.2	37.9	55.6	49.7
PRE-1512	73.4	44.8	**74.8**	43.2	92.1	94.3	83.4	**45.7**	65.5	66.8	68.4
PRE-1512U	74.8	46.0	**75.6**	**45.3**	93.5	95.0	86.5	**49.3**	66.7	69.5	**70.2**

Table 3: Pascal VOC 2012 action classification results (AP %).

DETECTION POTENTIAL



DETECTION POTENTIAL



DETECTION POTENTIAL