Improving Object Detection with Deep Convolutional Networks via Bayesian Optimization and Structured Prediction

Yuting Zhang[*,†], Kihyuk Sohn[†], Ruben Villegas[†], Gang Pan[*], Honglak Lee[†]

†

University of Michigan

*

ZHEJIANG UNIVERSITY 1897 浙江大学

Object detection using deep learning

- Object detection systems based on the deep convolutional neural network (CNN) have recently made ground-breaking advances.

  [LeCune et al. 1989; Sermanet et al. 2013; Girschick et al., 2014; Simoyan et al., 2014; Lin et al. 2014, and many others]

- State-of-the-art: “Regions with CNN features” (R-CNN)

  Girshick et al, “Region-based Convolutional Networks for Accurate Object Detection and Semantic Segmentation”, PAMI 2015 & CVPR 2014.

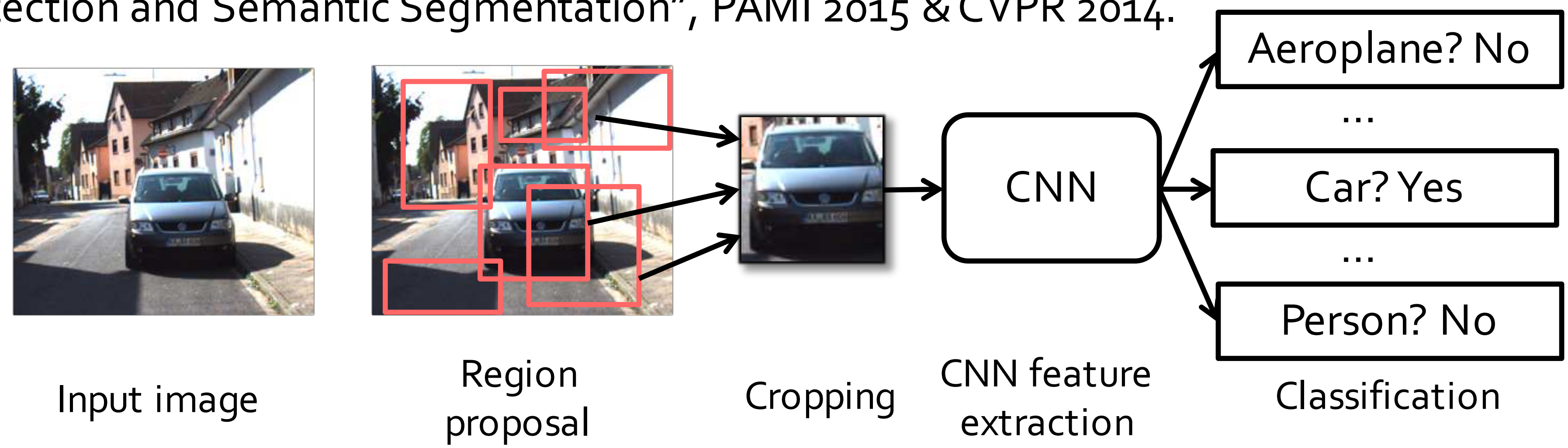
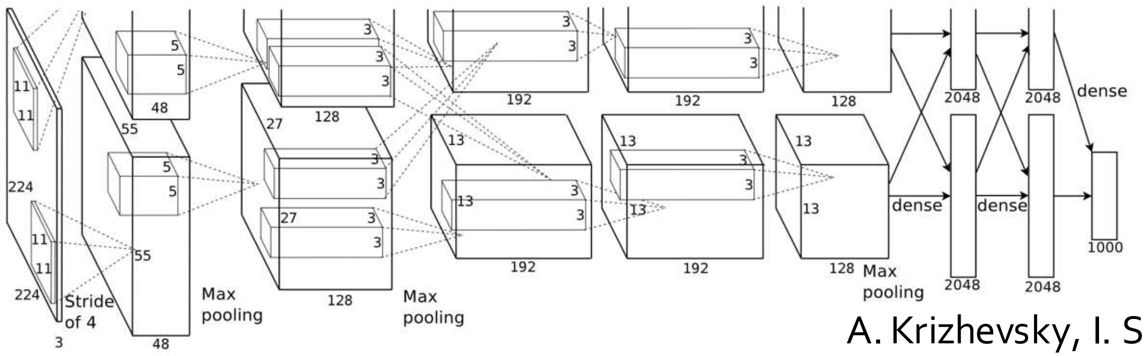


Image adapted from Girshick et al., 2014

R-CNN: Method

1) Convolutional neural network for classification



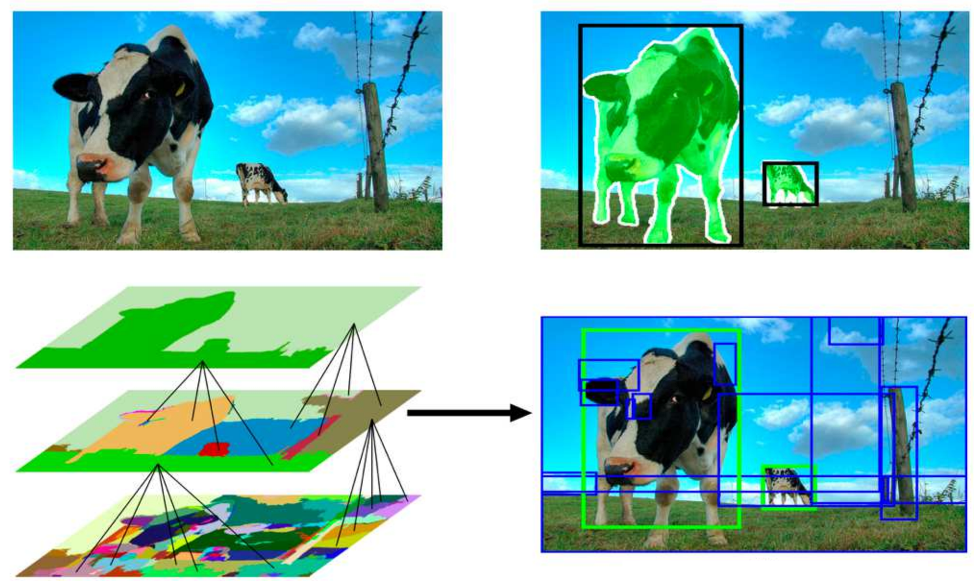
- Pretrained on ImageNet for 1000-category classification
- Finetuned on PASCAL VOC for 20 categories

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *NIPS*, 2012.

2) Selective search for region proposal:

- Hierarchical segmentation → bounding box

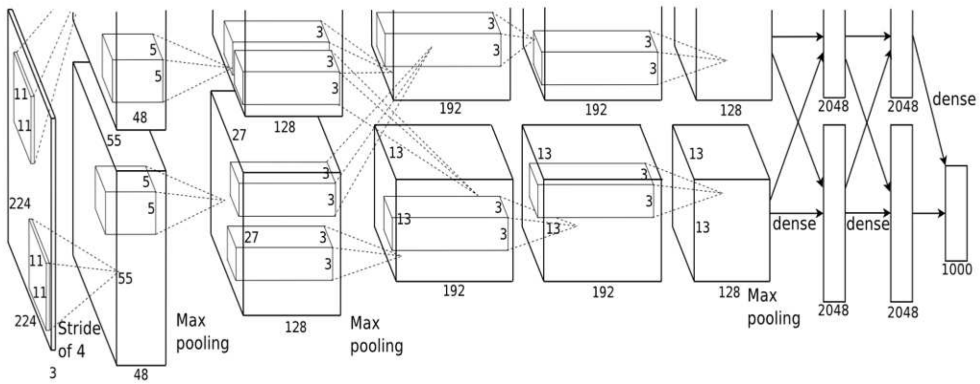
K. E. A. Sande, J. R. R. Uijlings, T. Gevers, and A. W. M. Smeulders. Segmentation as selective search for object recognition. *ICCV*, 2011.



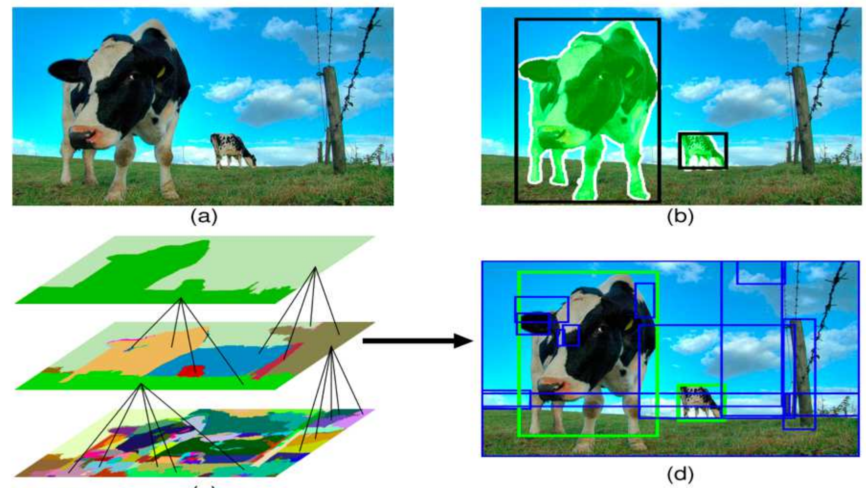
Images from Krizhevsky et al. 2012 & Sande et al. 2011

R-CNN: Detection

Classification confidence for sampled bounding boxes



A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.



K. E. A. Sande, J. R. R. Uijlings, T. Gevers, and A. W. M. Smeulders. Segmentation as selective search for object recognition. *ICCV*, 2011.

- Detection: locally solve

$$\mathrm{argmax}_y f(x, y)$$

where x is the image, and y is a bounding box, $f(x, y)$ is the classification confidence computed from CNN.

R-CNN: Pros and Cons

Pros:

- Surprisingly good performance (mean average precision, mAP), e.g., on PASCL VOC2007:
  - Deformable part model (old SOA): 33.4%
  - R-CNN: **53.7%**
- Strong discriminative ability from CNN
- Reasonable efficiency from region proposal

R-CNN: Pros and Cons

Pros:

- Surprisingly good performance (mean average precision, mAP), e.g., on PASCL VOC2007:
  - Deformable part model (old SOA): 33.4%
  - R-CNN: **53.7%**
- Strong discriminative ability from CNN
- Reasonable efficiency from region proposal

Cons:

- Poor localization (worse than DPM), due to
  - Ground truth bounding box (BBox) may be missing from (or have poor overlap with) region proposals
  - CNN is trained solely for classification, but not localization

Our solutions

1. Find better bounding boxes via Bayesian optimization
2. Improve localization sensitivity via structured objective

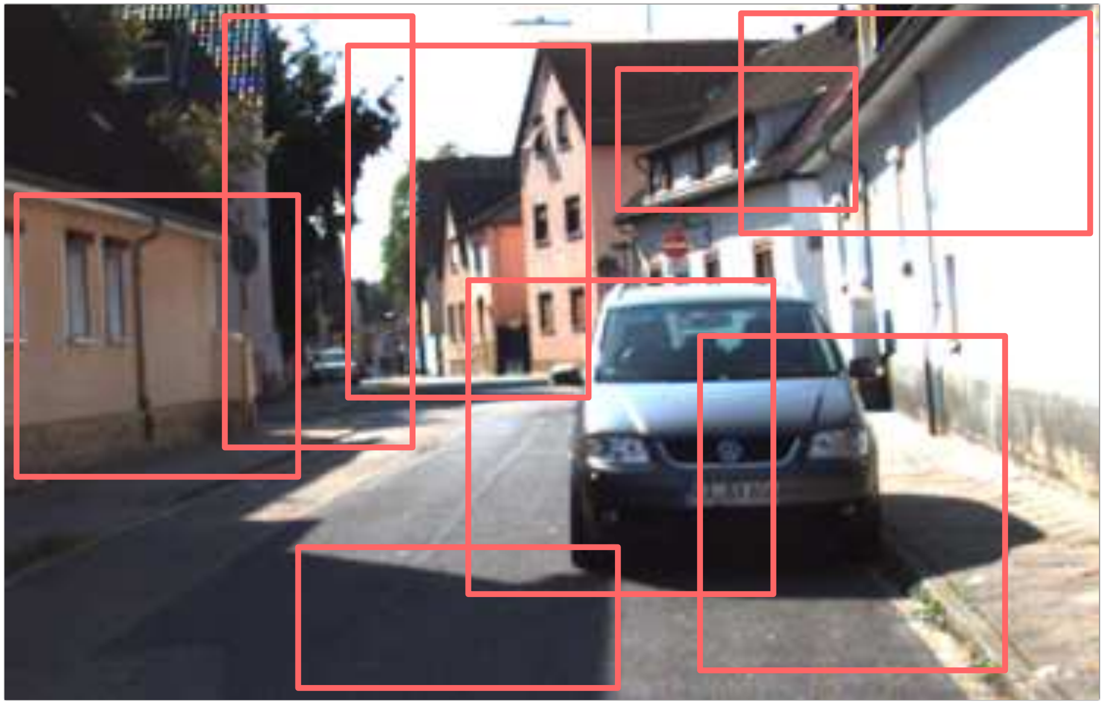
Thrust 1:
Find better bounding boxes via Bayesian optimization

Fine-grained search: Framework

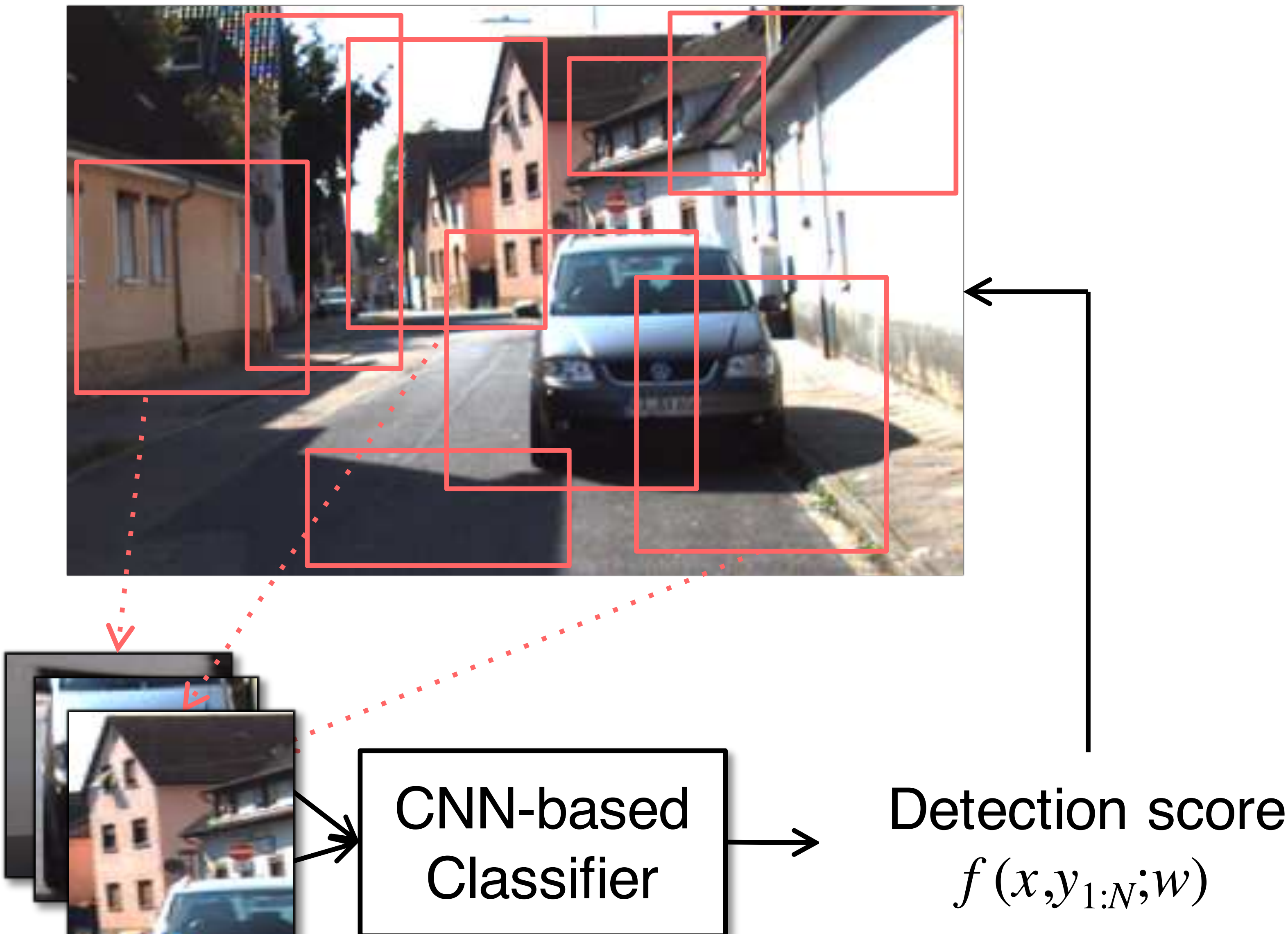
Given a test image

The image is from the KITTI dataset

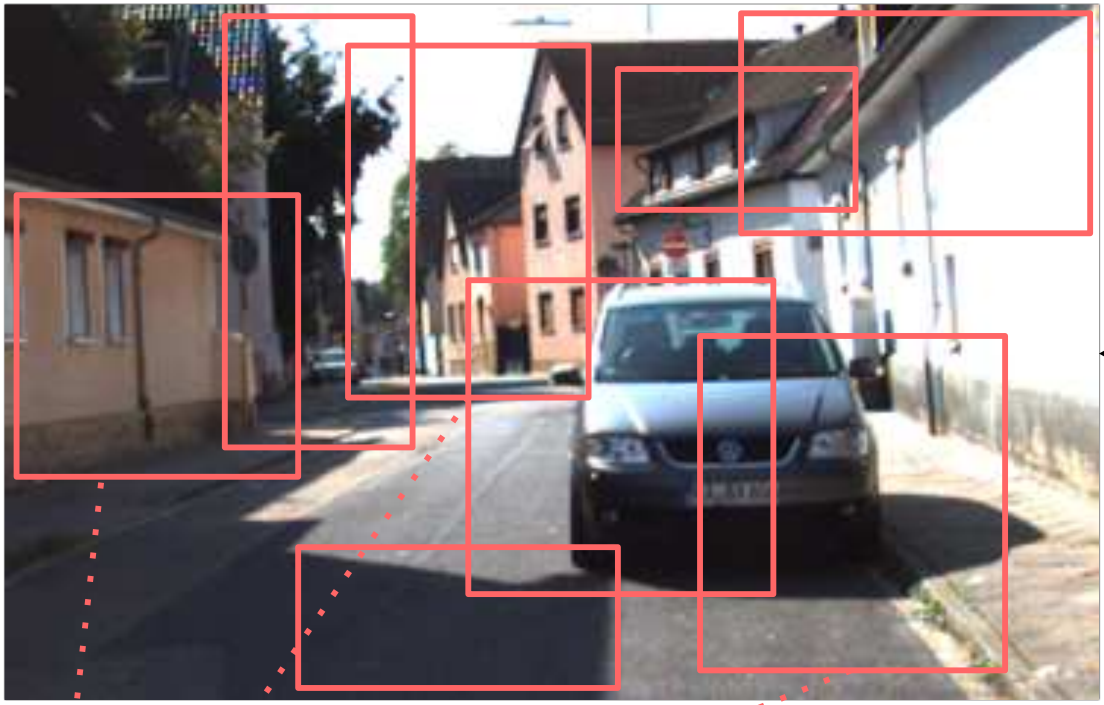
Propose initial regions via selective search



Compute classification scores

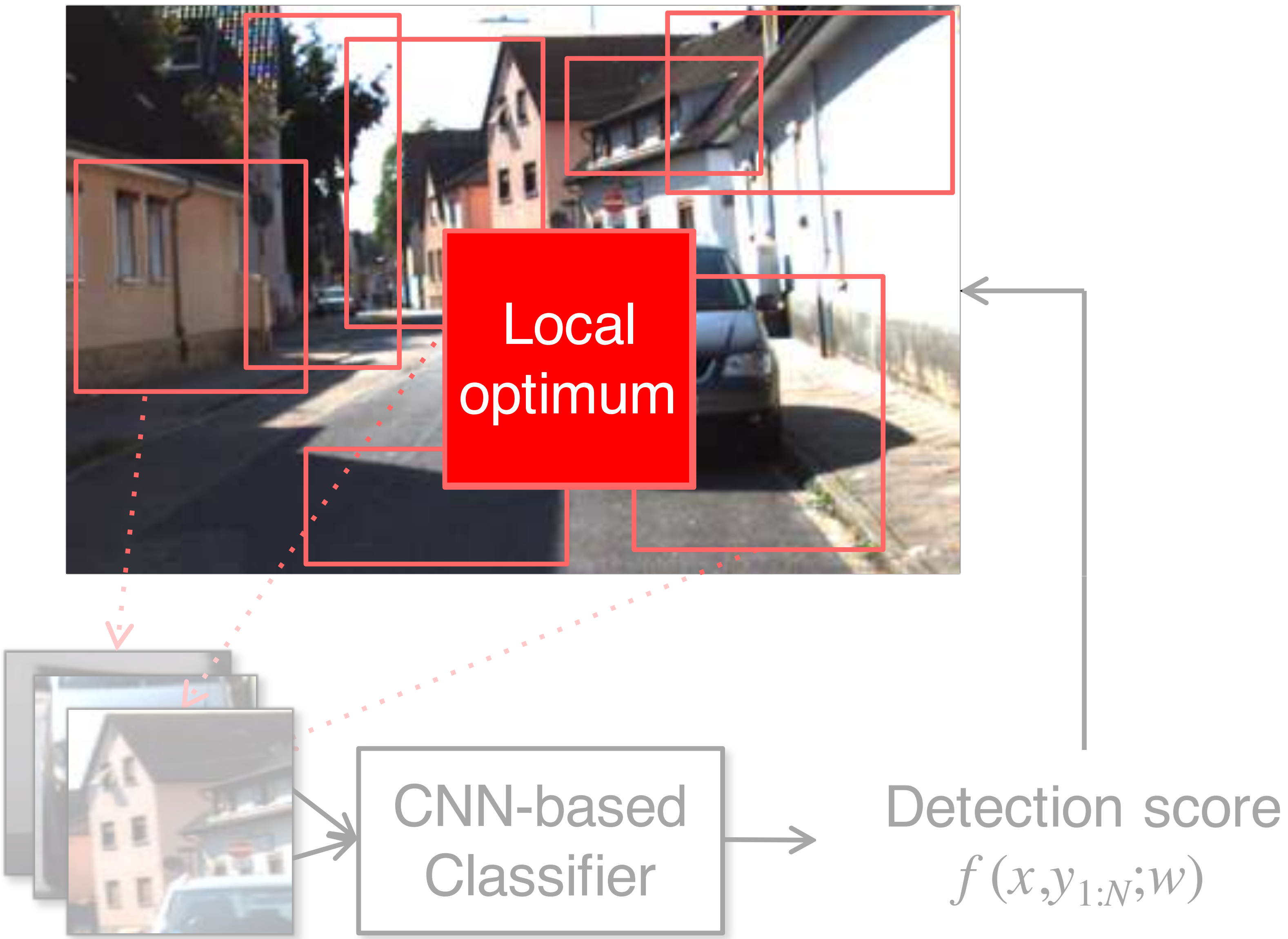


What if no existing bounding box is good enough?

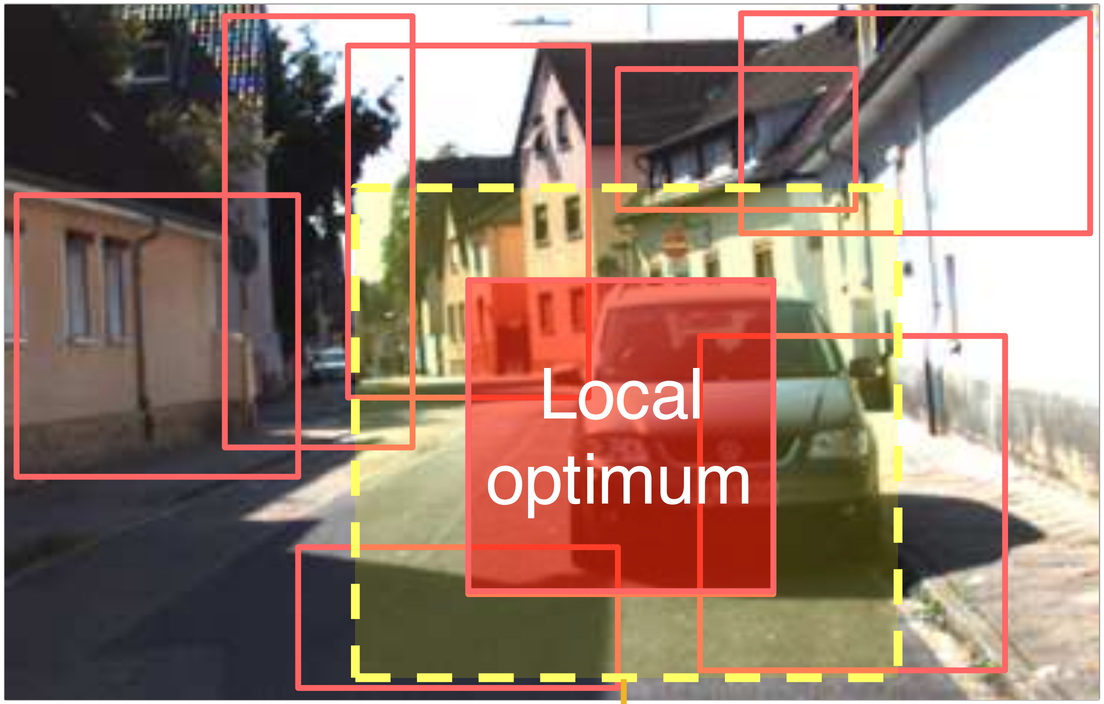


How to propose a better box?

Find a local optimal bounding box

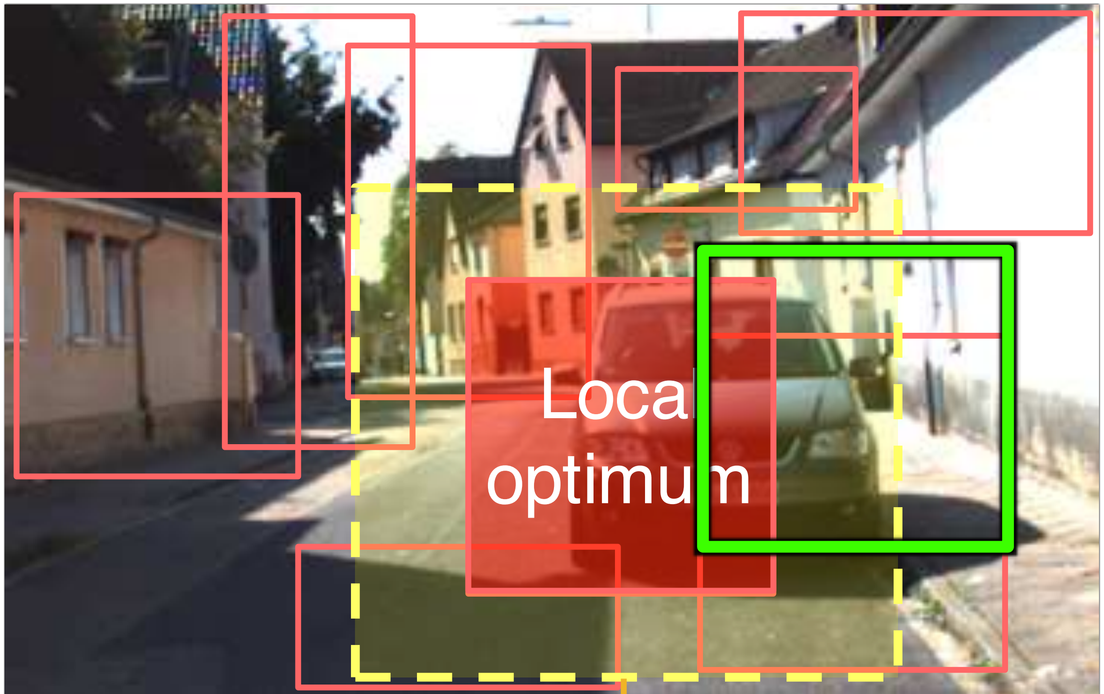


Determine a local search region



Search Region near
local optimum for
Bayesian optimization

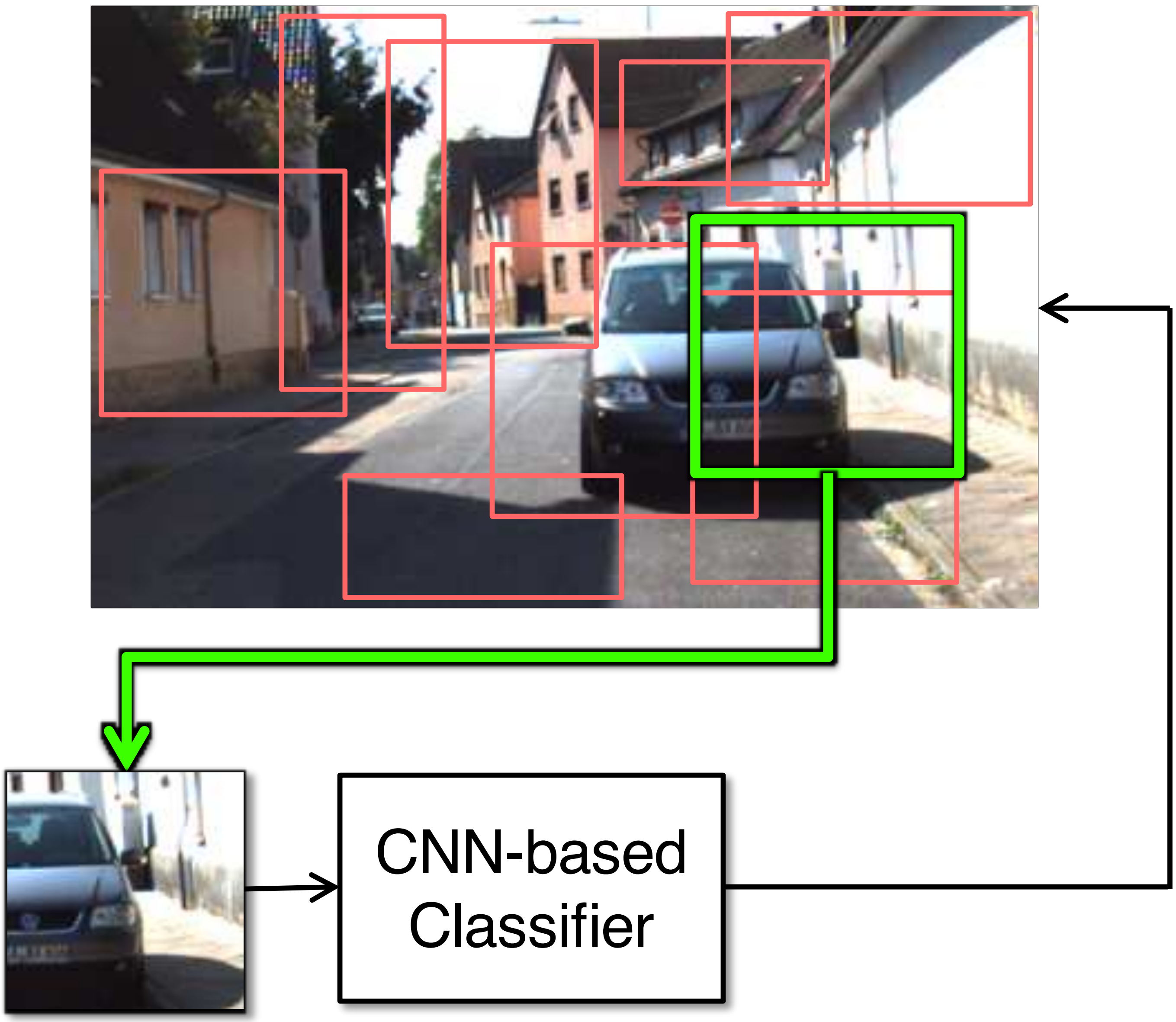
Propose a bounding box via Bayesian optimization



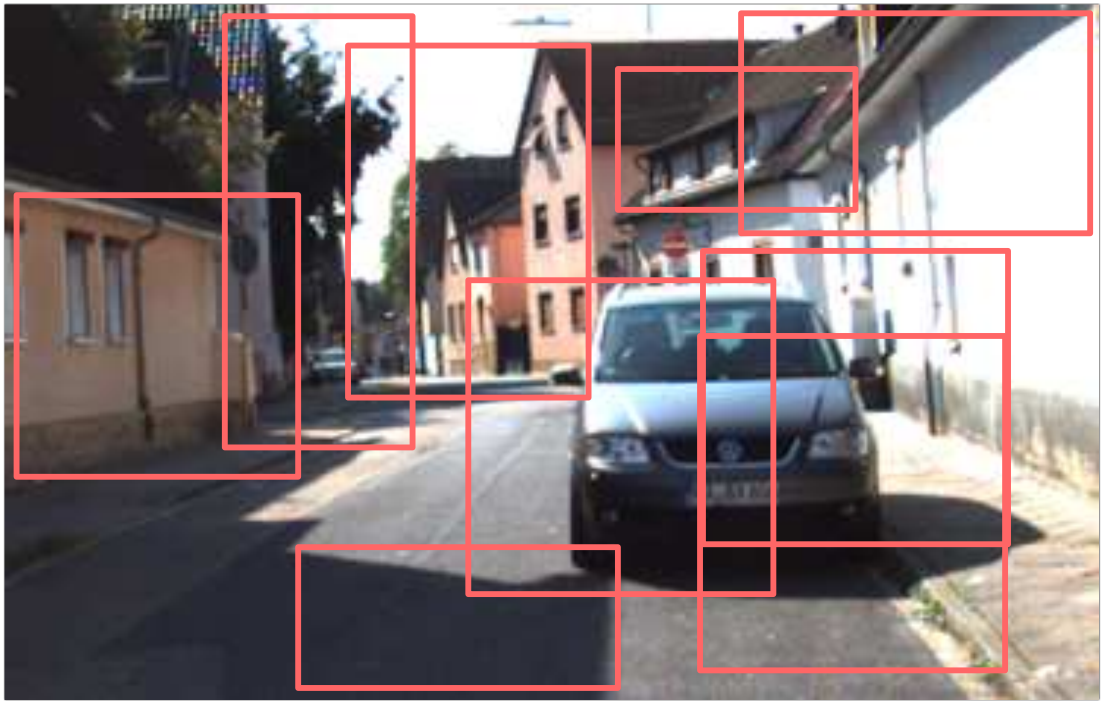
The new box Has a good chance to get better classification score

Search Region near local optimum for Bayesian optimization

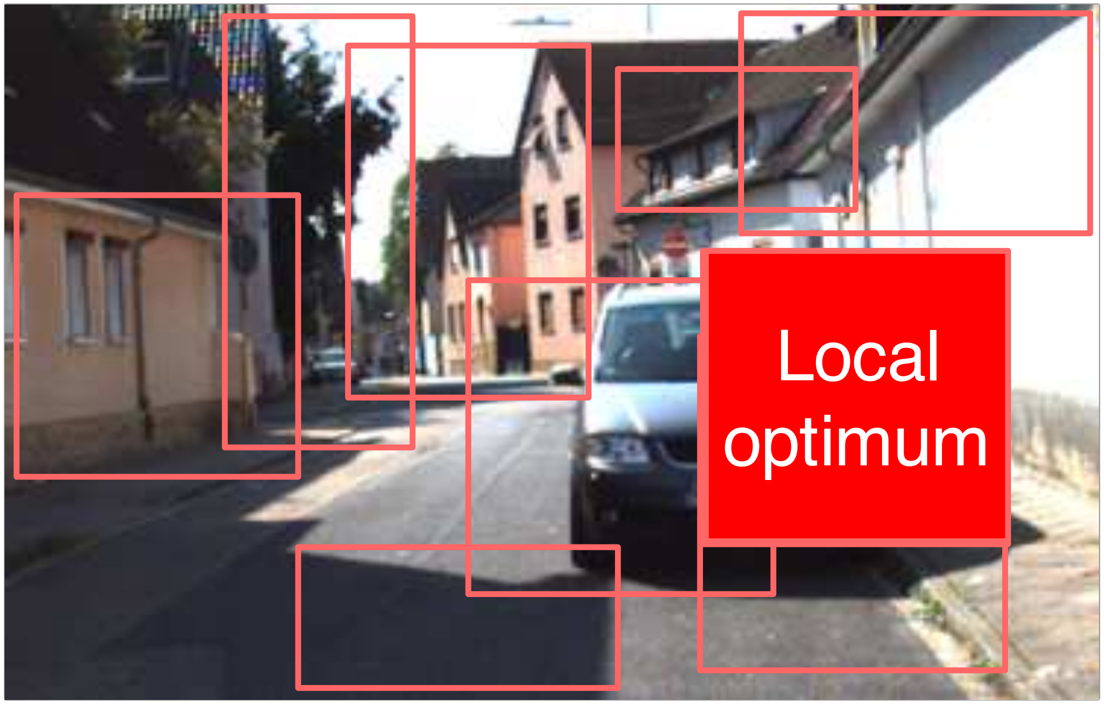
Compute the actual classification score



Iterative procedure : Iteration 2



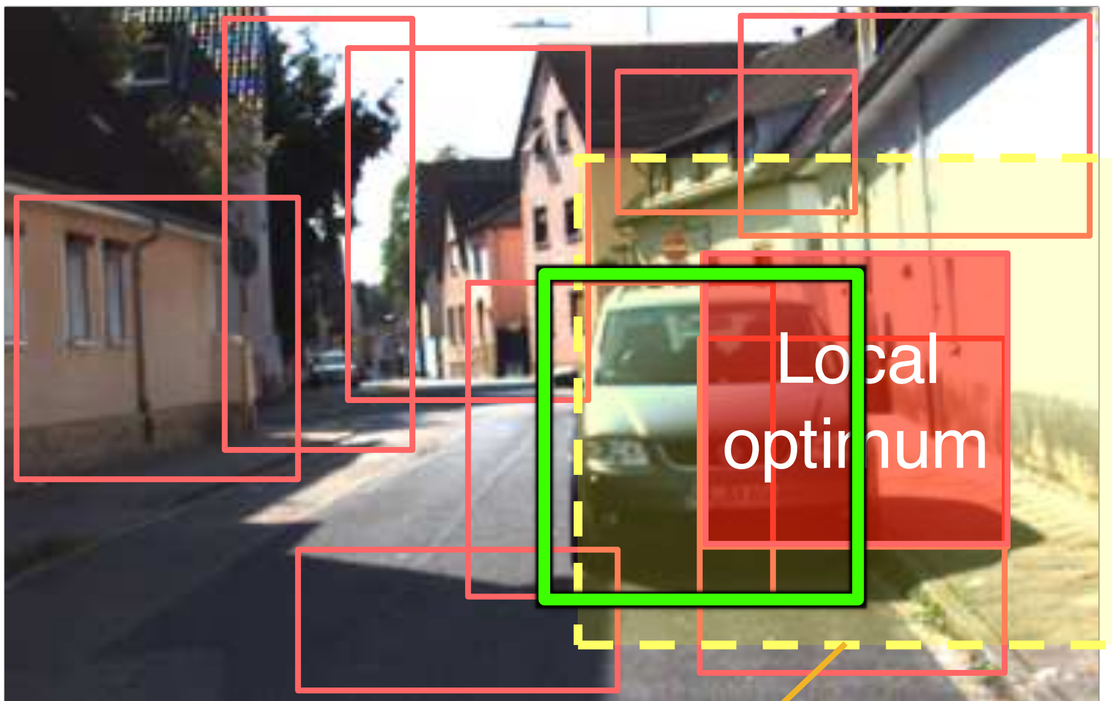
Iteration 2: Find a local optimum



Iteration 2: Determine a local search region

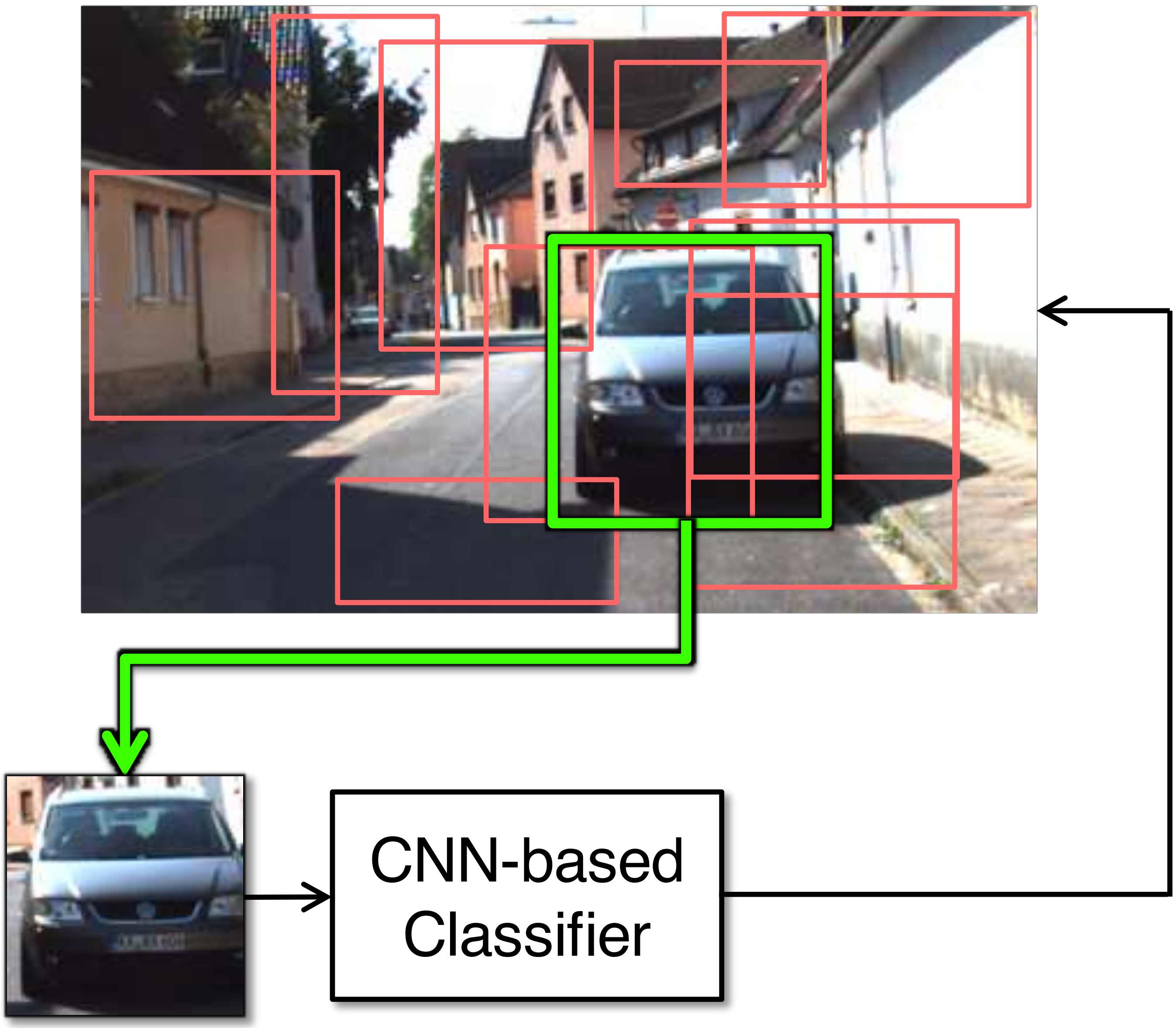
Search Region near local optimum for Bayesian optimization

Iteration 2: Propose a new box via Bayesian opt.

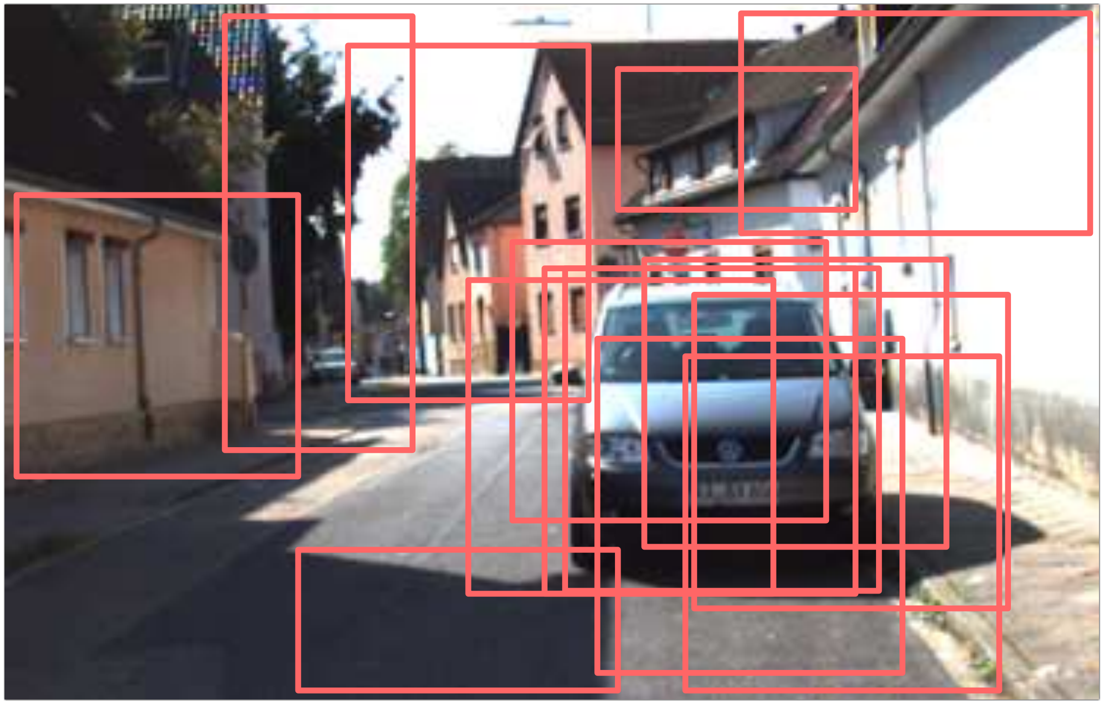


Search Region near
local optimum for
Bayesian optimization

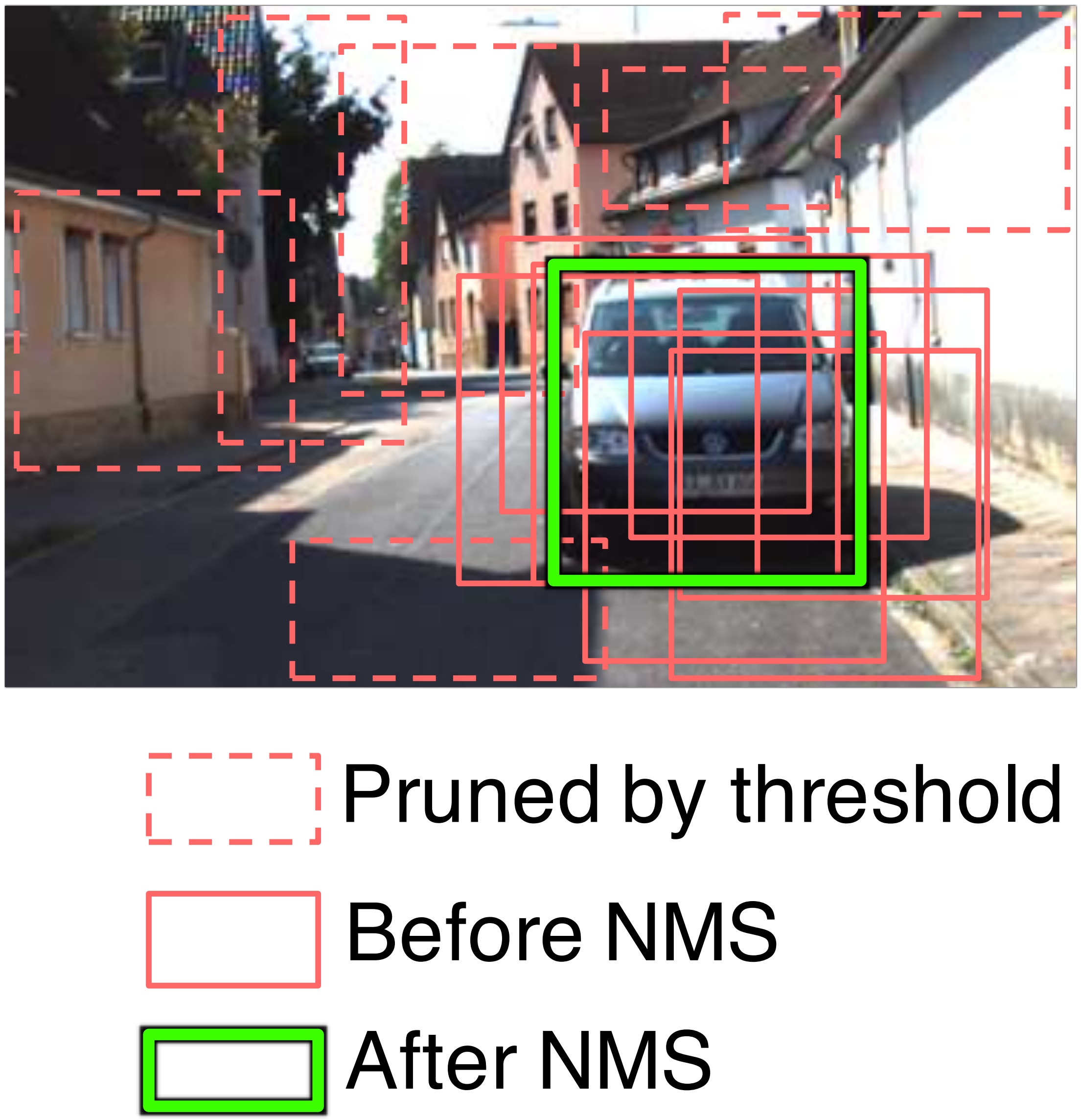
Iteration 2: compute the actual score



After a few iterations …



Final detection output



Bayesian optimization: General

e.g., CNN-based classifier or any score function of detection methods.

- Model the $\boxed{\text{complicated function } f(x, y)}$, whose evaluation cost is high, with a probabilistic distribution of function values.
- The distribution is defined with a relatively computationally efficient surrogate model.

Framework

- Let $\mathcal{D}_N = \{y_j, f_j\}_{j=1}^{N}$ and $f_j = f(x, y_j)$ be the known solutions. We want to model

$$p(f|\mathcal{D}_N) \propto p(\mathcal{D}_N|f)p(f)$$

- Try to find a new boxing box $y_{N+1} \neq y_j, \forall j \leq N$ with the highest chance s.t. $f_{N+1} > \max_{1 \leq j \leq N} f_j$

Bayesian optimization: Gaussian process

- Framework:

$$p(f|\mathcal{D}_N) \propto p(\mathcal{D}_N|f)p(f)$$

- Gaussian process is a general function prior, which used for $p(f)$.
- $p(f_{N+1}|y_{N+1},\mathcal{D}_N)$ can be expressed as a multivariate Gaussian, whose parameters can be obtained by **Gaussian process regression (GPR)** as a closed-form solution, when the square exponential covariance function is used.
- The chance of $f_{N+1} > \max_{1 \le j \le N} f_j = \hat{f}_N$ is measure by the **expected improvement**:

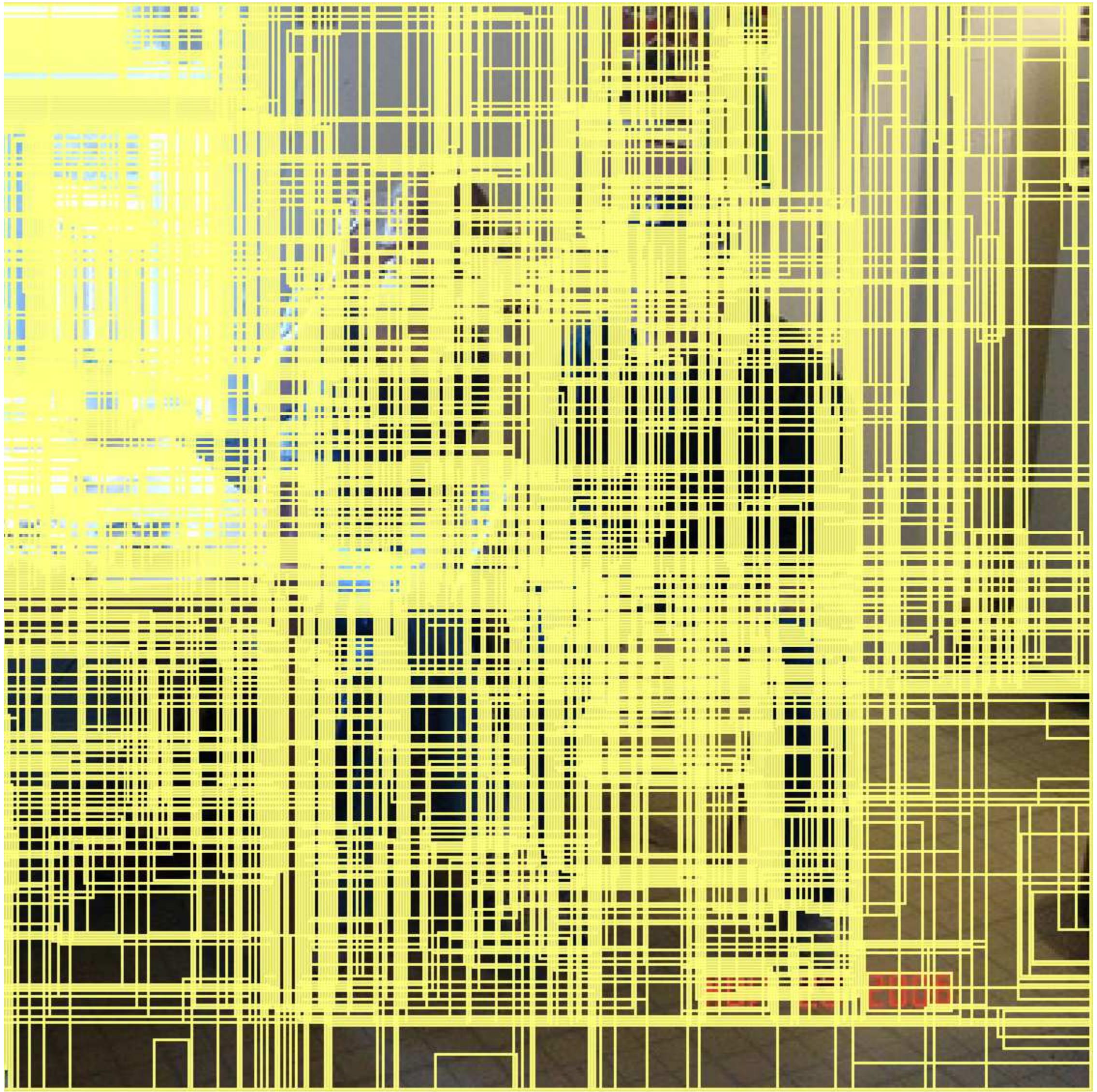
$$\int_{\hat{f}_N} (f - \hat{f}_N) \cdot p(f|y_{N+1},\mathcal{D}_N;\theta)df$$

FGS Procedure: a real example

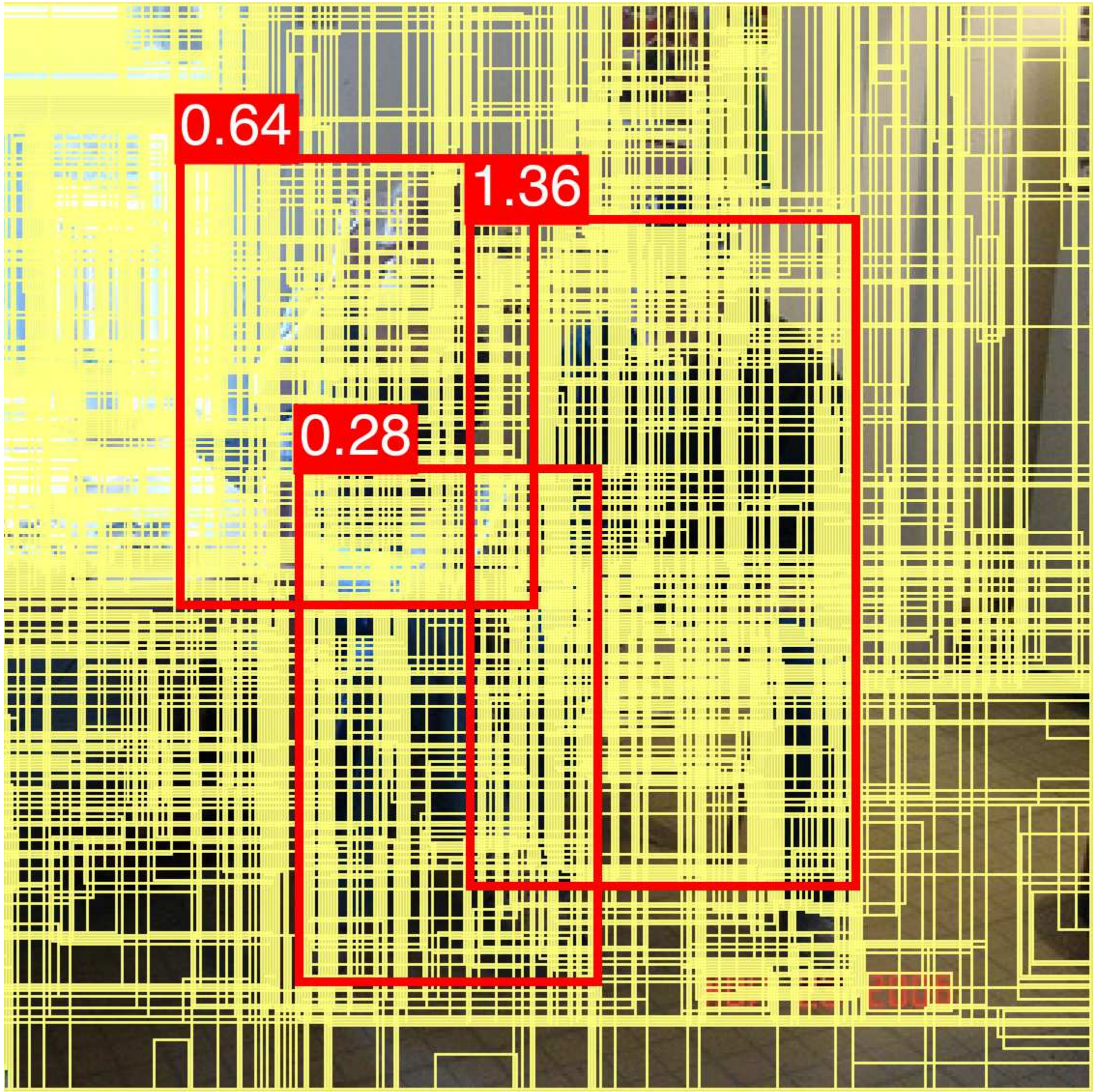
Original image

The image is from PASCAL VOC2007

Initial region proposals

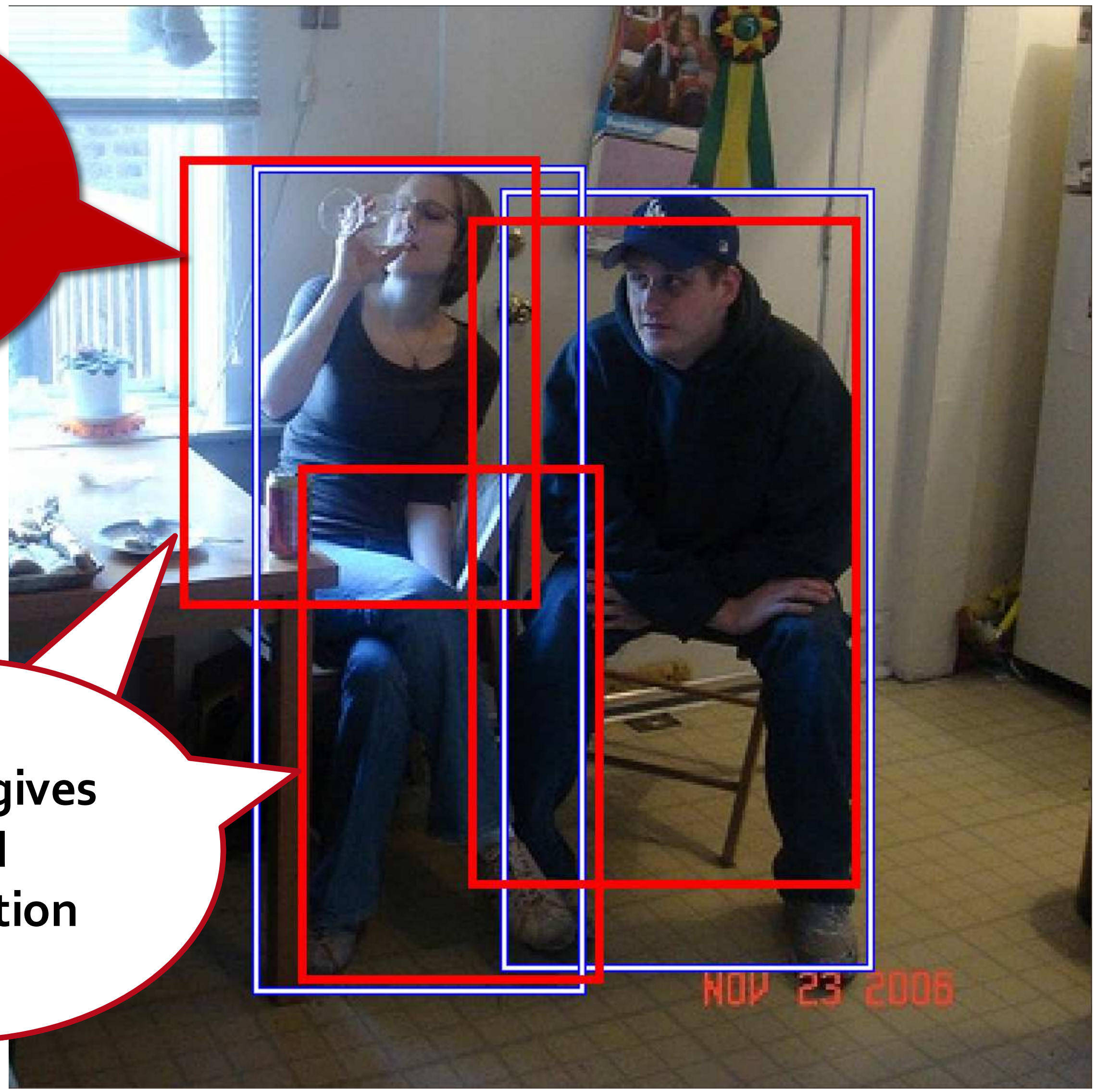


Initial detection (local optima)



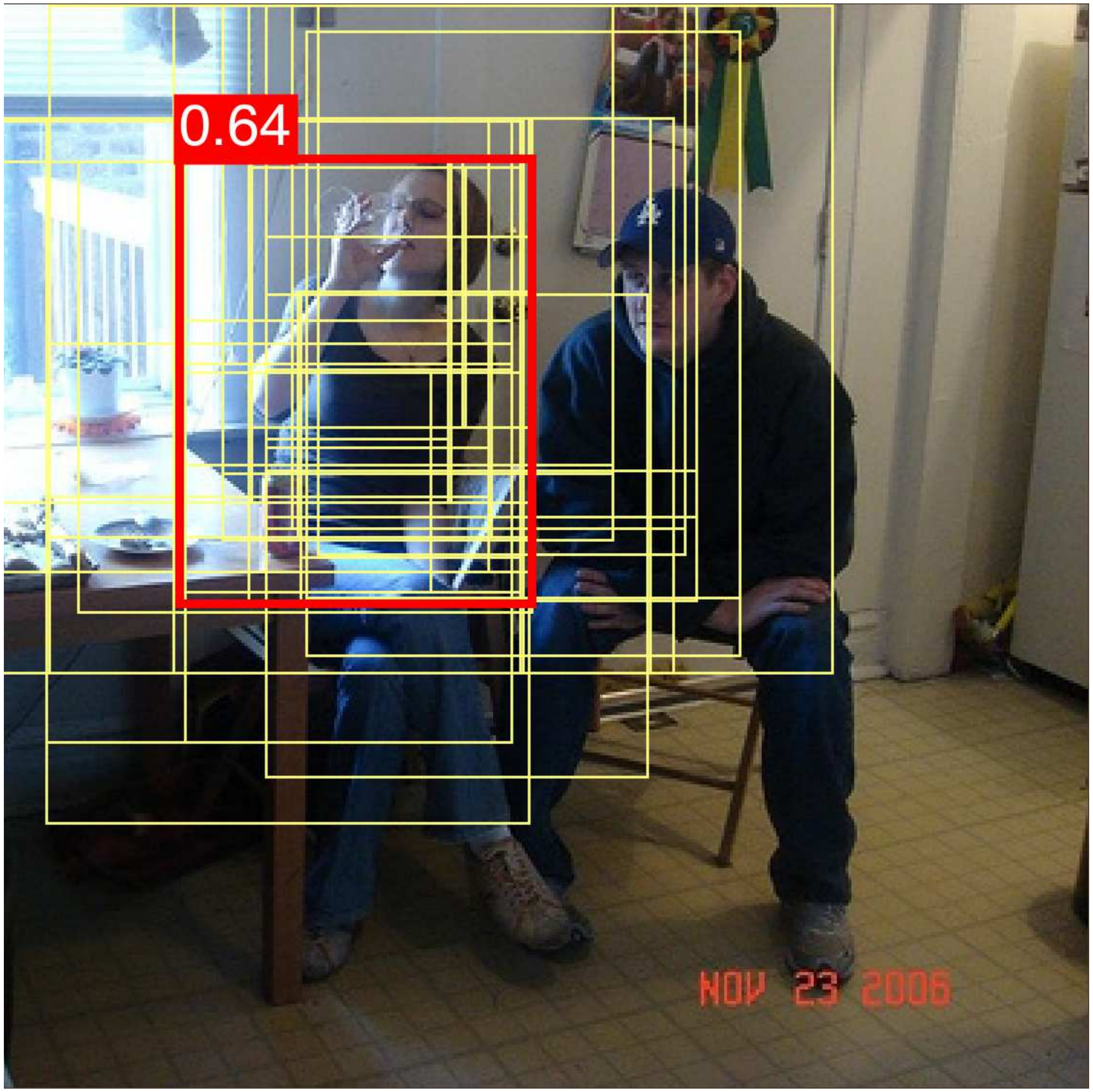
Initial detection & Ground truth

Take this as ONE starting point

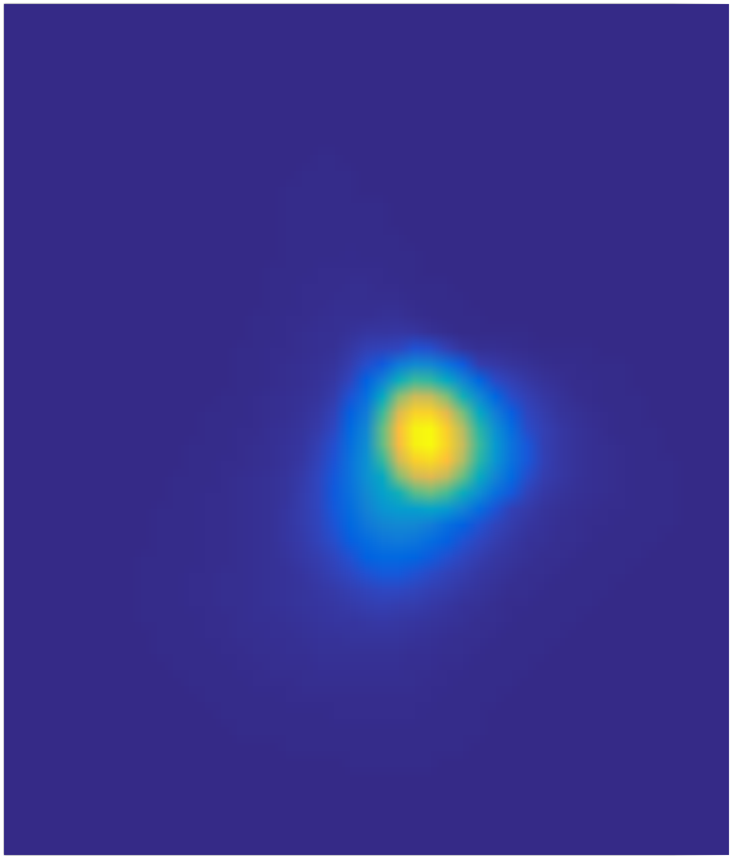


Neither gives good localization

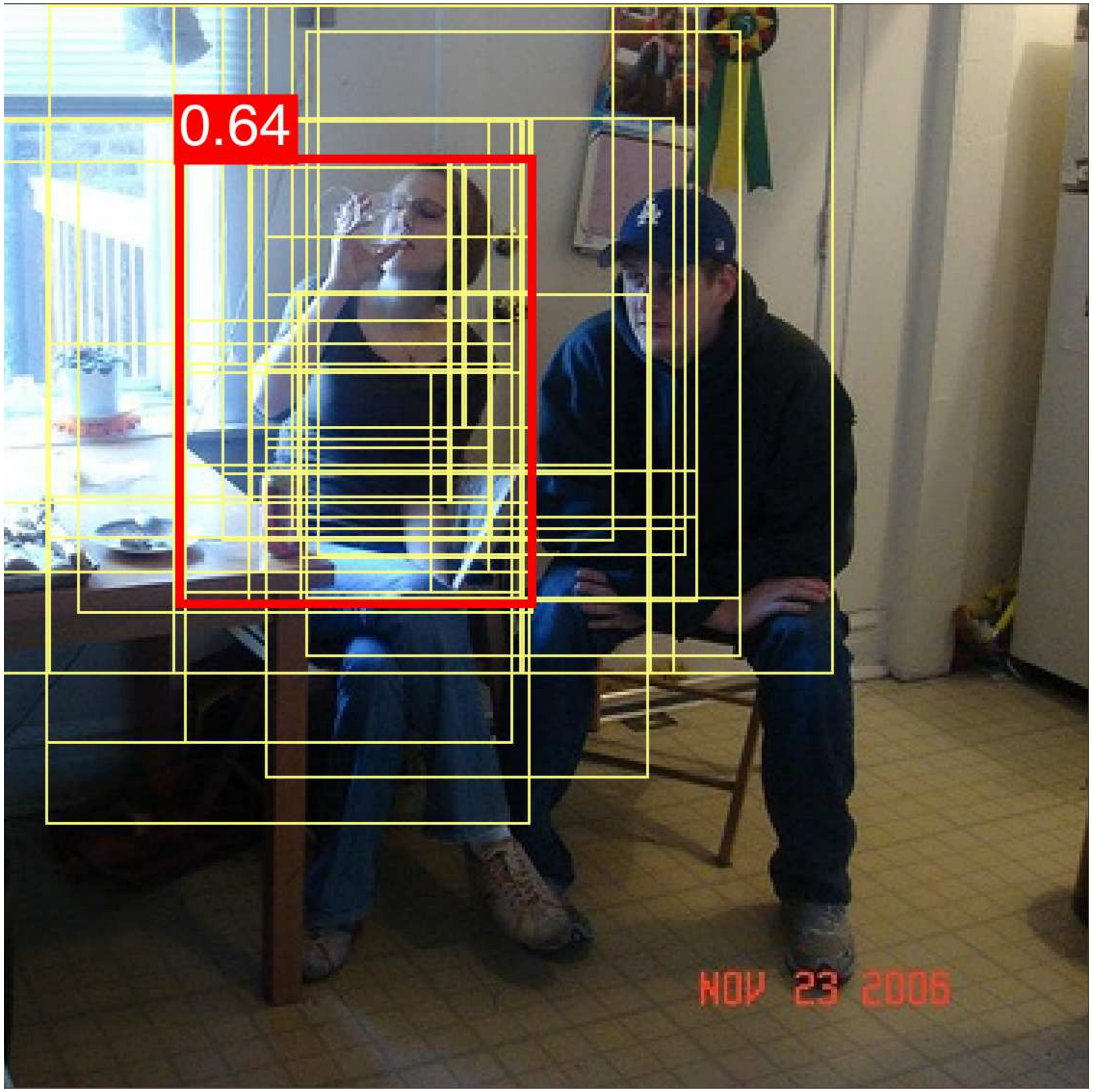
Iter 1: Boxes inside the local search region



Iter 1: Heat map of expected improvement (EI)

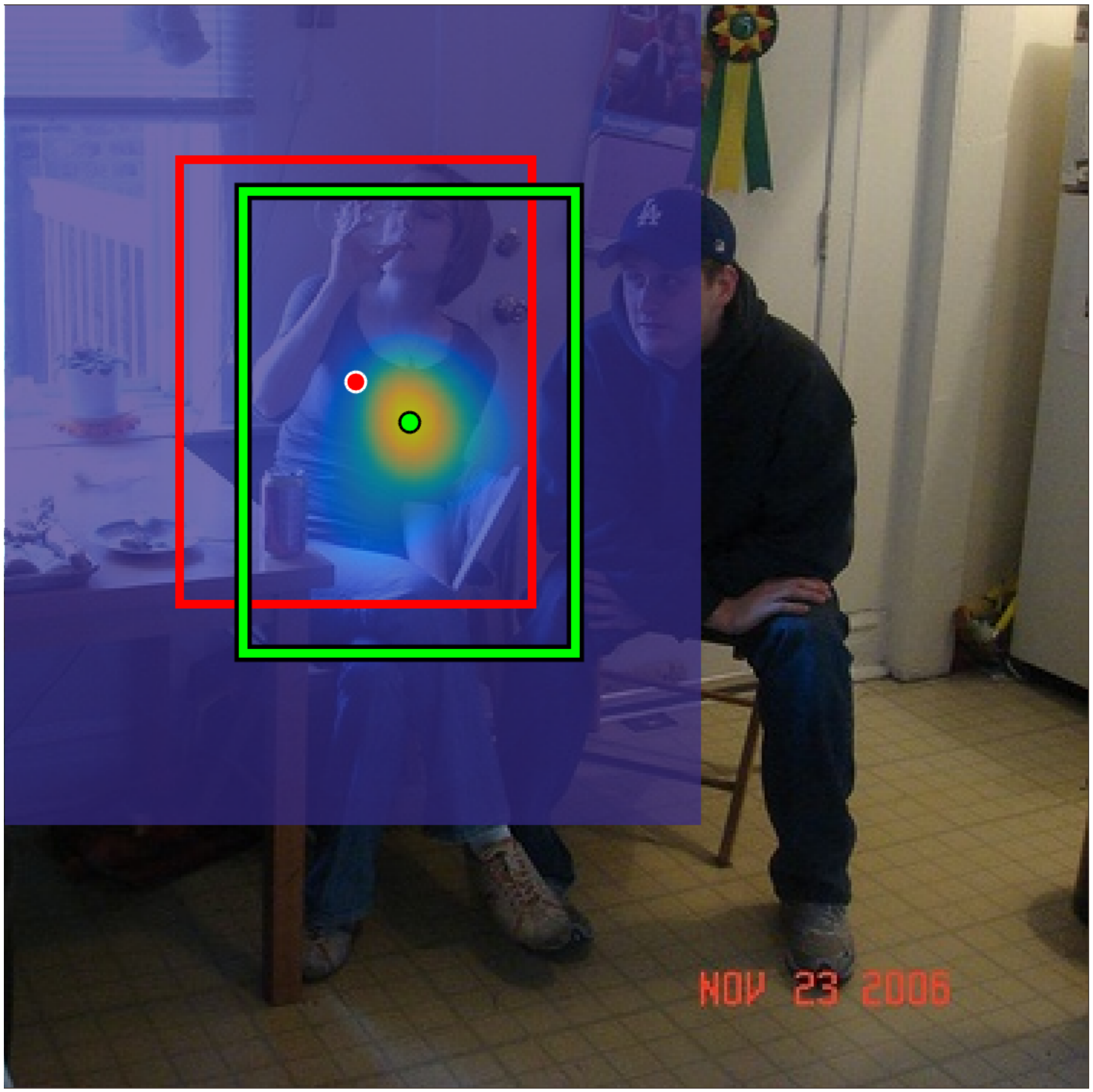


- A box has 4-coordinates: (centerX, centerY, height, width)
- The height and width are marginalized by max to visualize EI in 2D



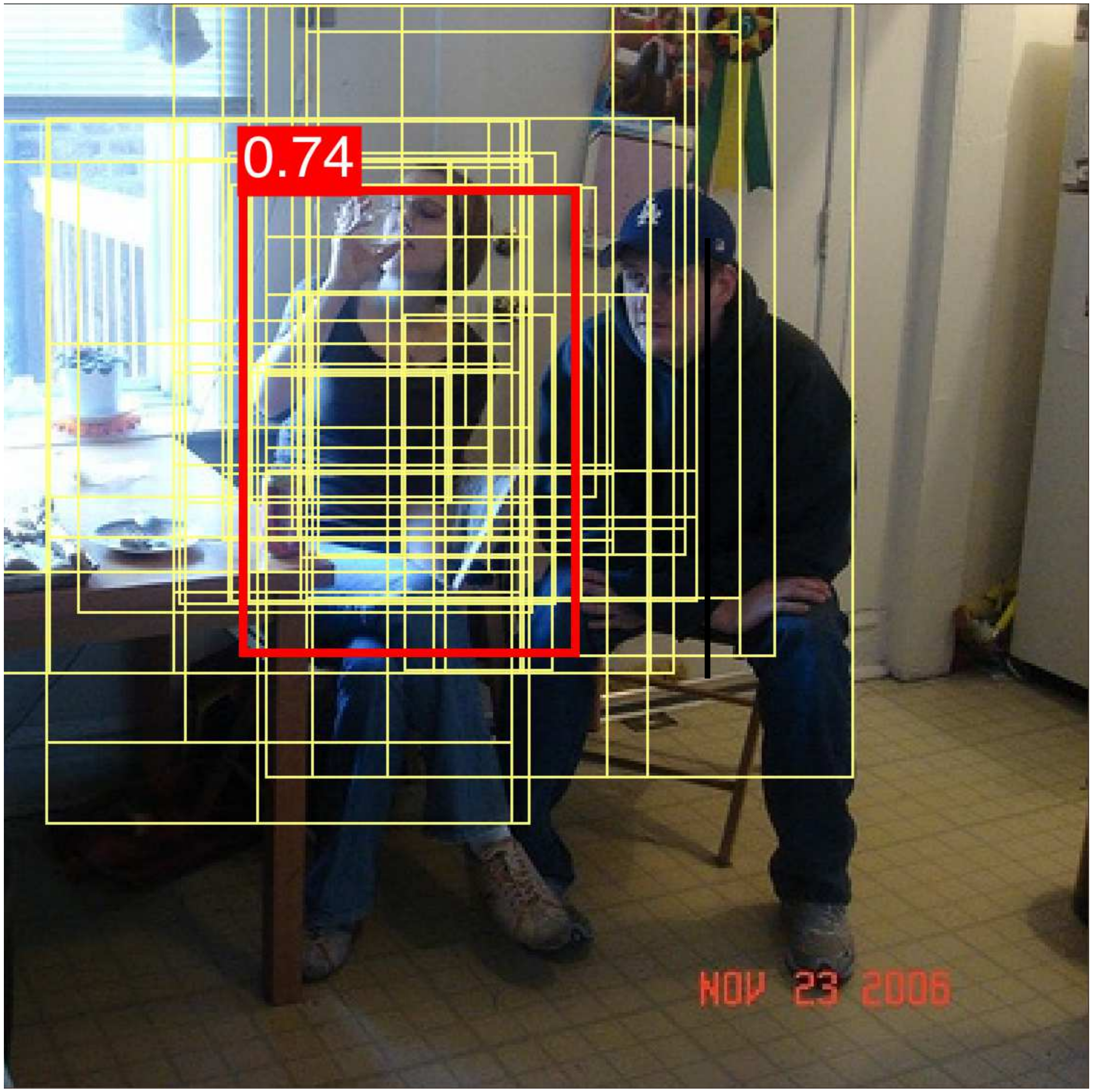
Iter 1: Heat map of expected improvement (EI)

Iter 1: Maximum of EI – the newly proposed box

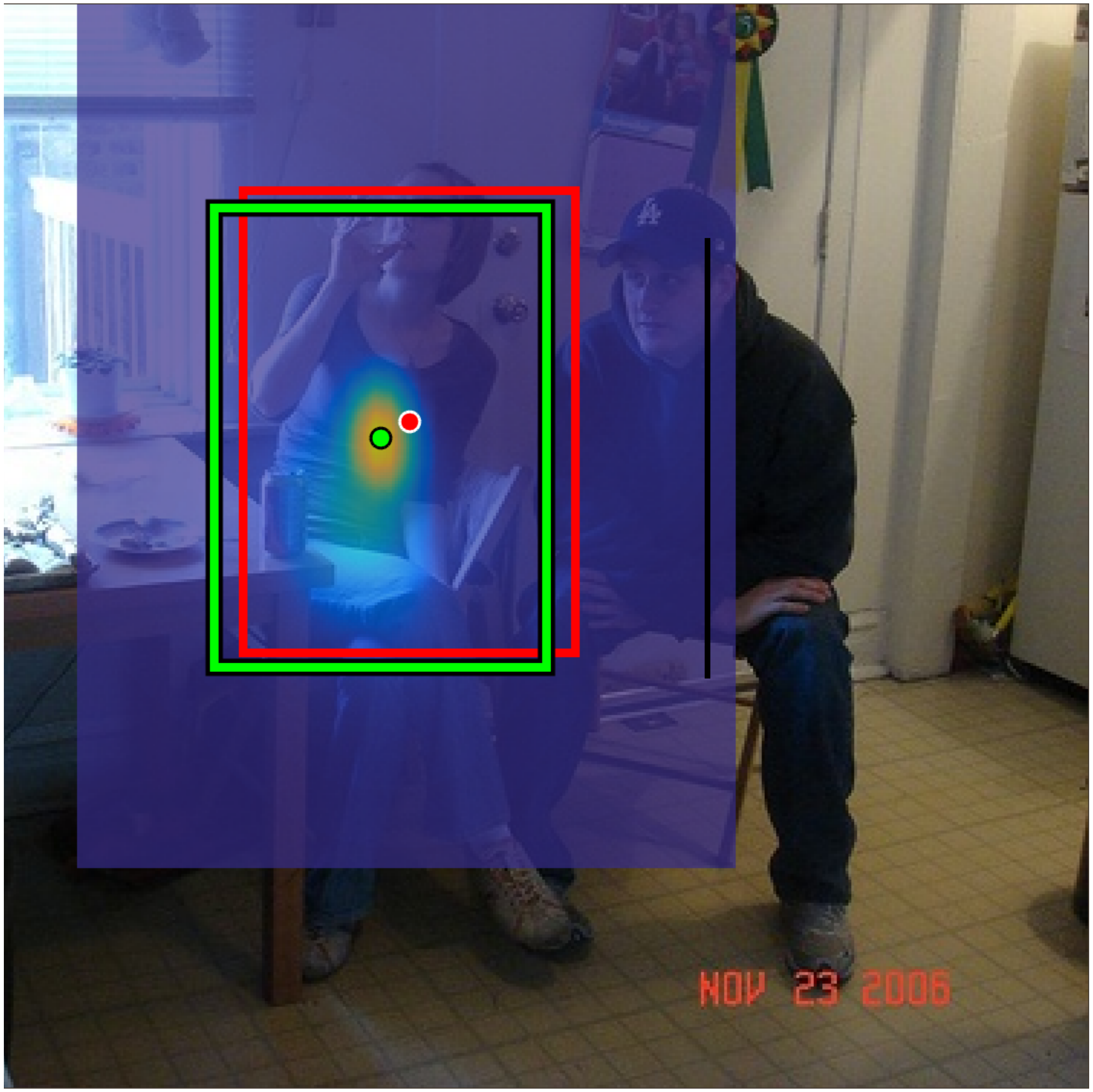


Iter 1: Complete

Iteration 2: local optimum & search region

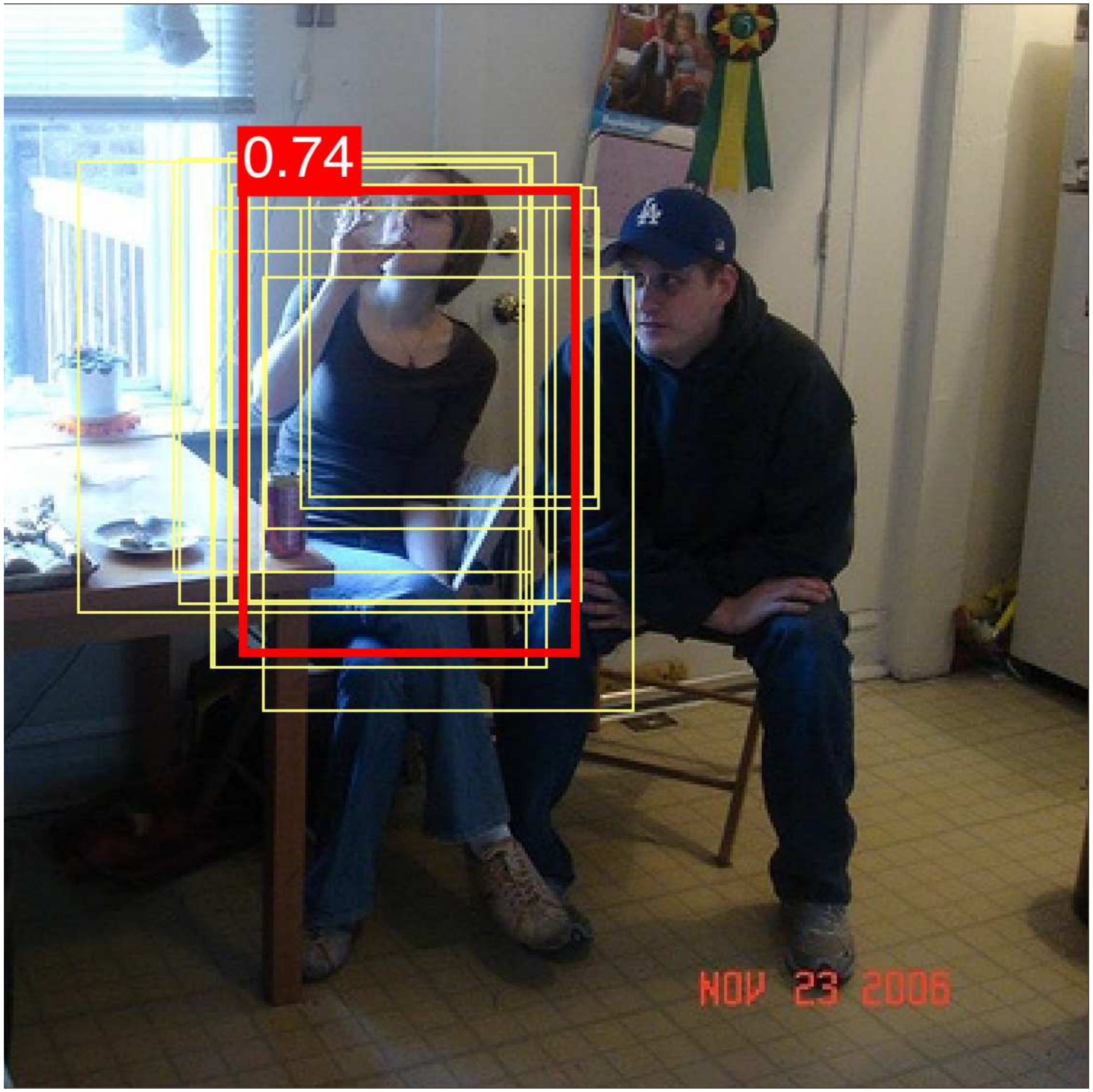


Iteration 2: EI heat map & new proposal



Iteration 2: Newly proposed box & its actual score

Iteration 3: local optimum & search region

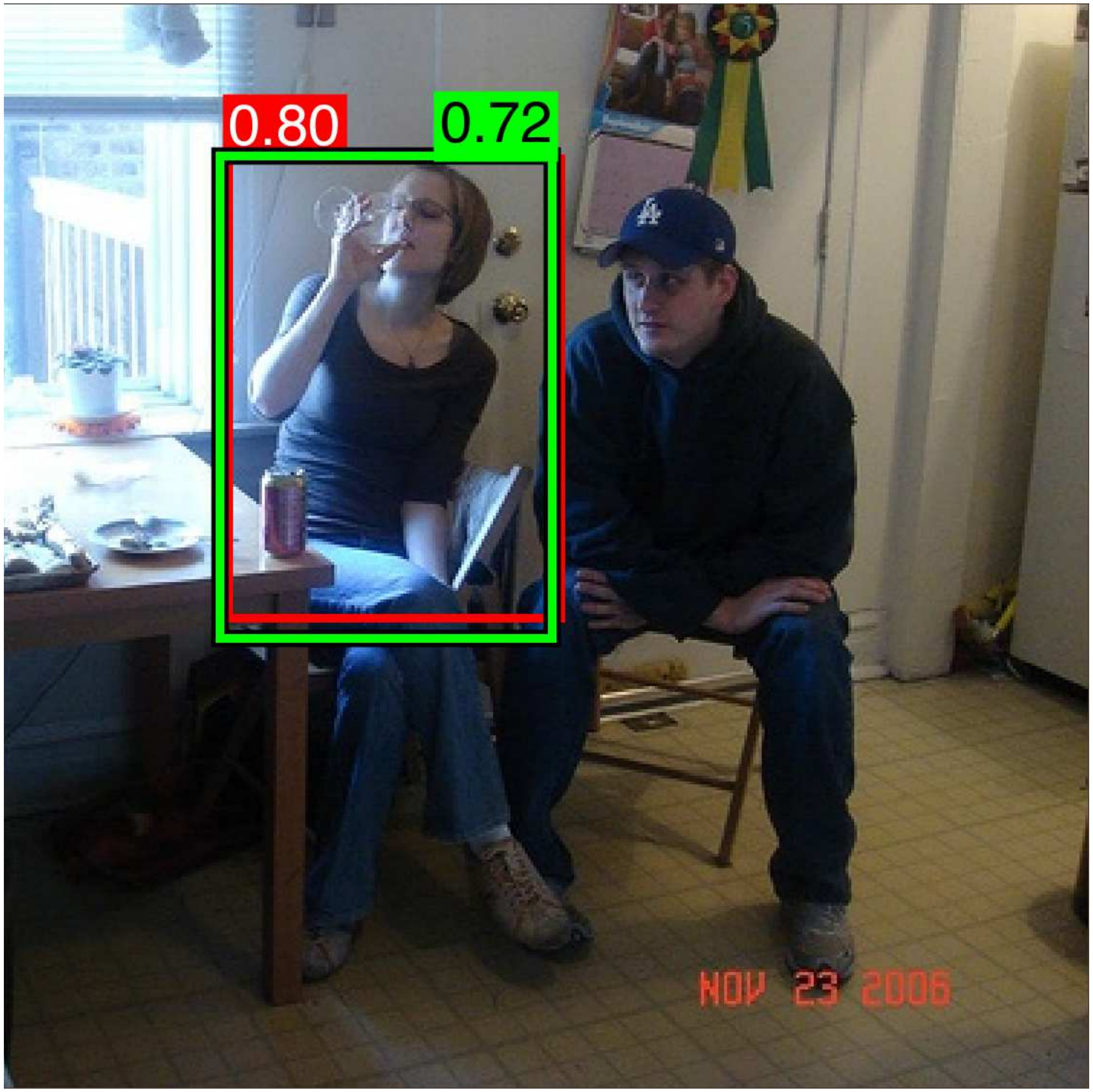


Iteration 3: EI heat map & new proposal

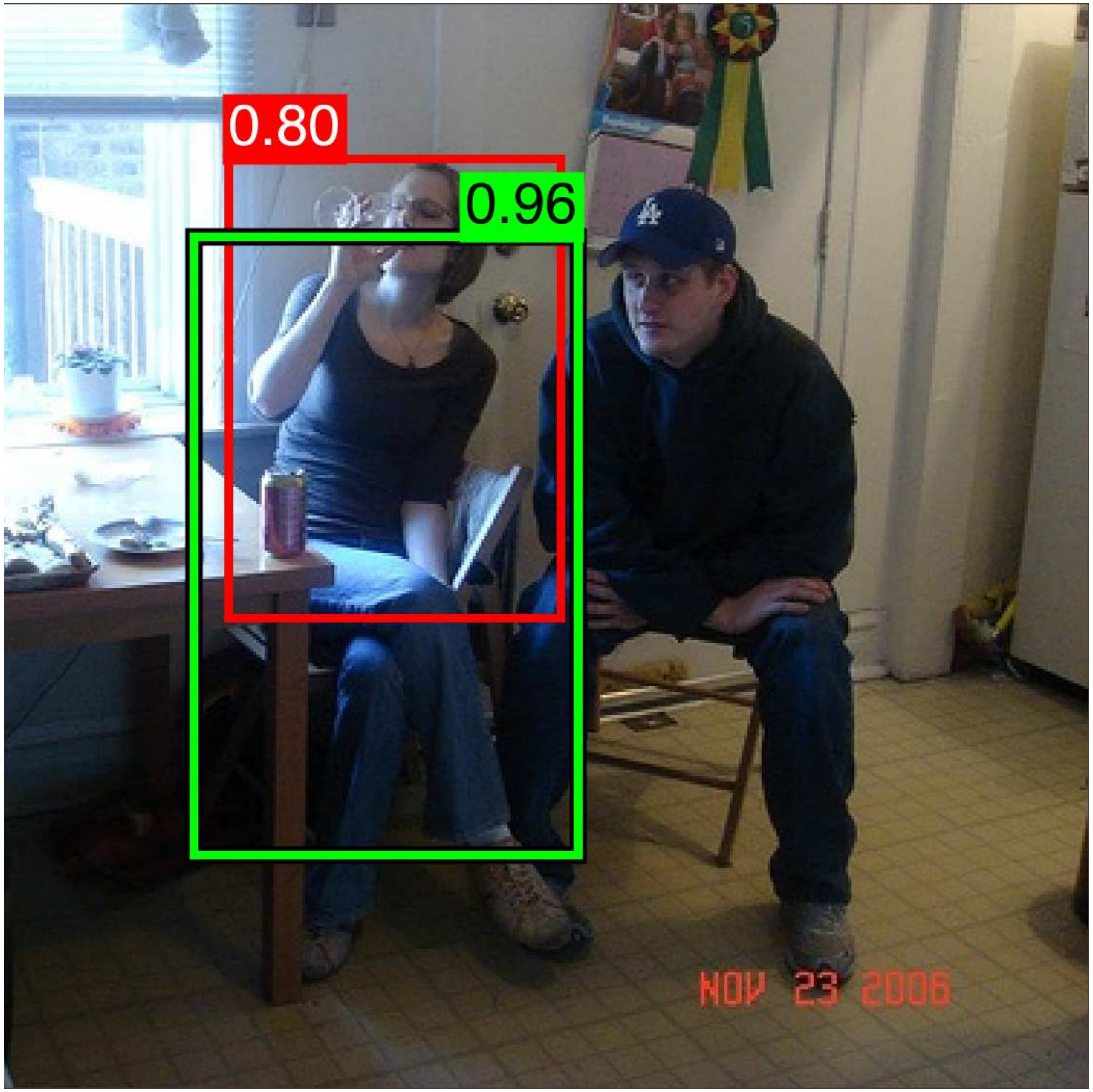
Iteration 3: Newly proposed box & its actual score

Iteration 4

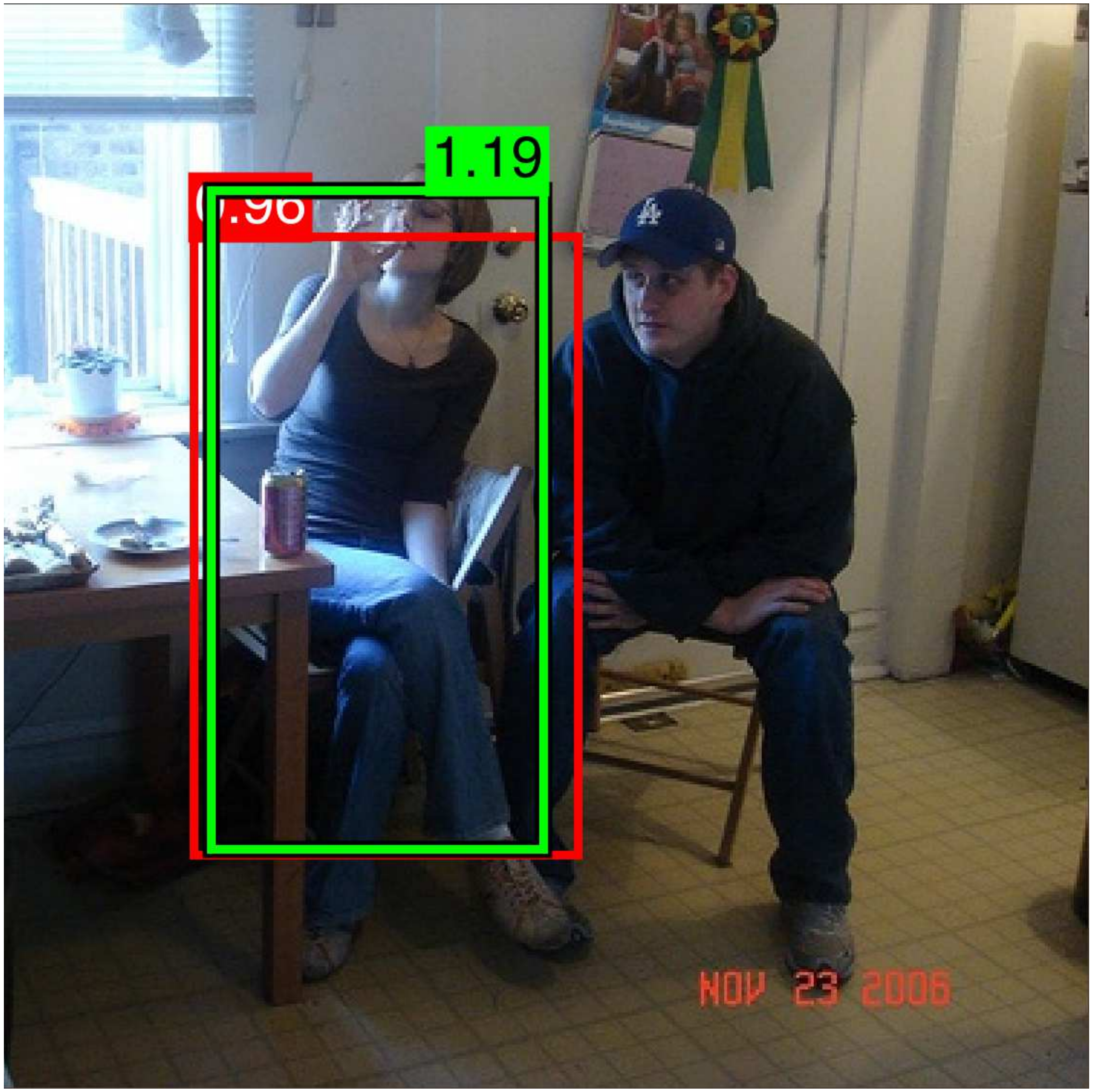
Iteration 5



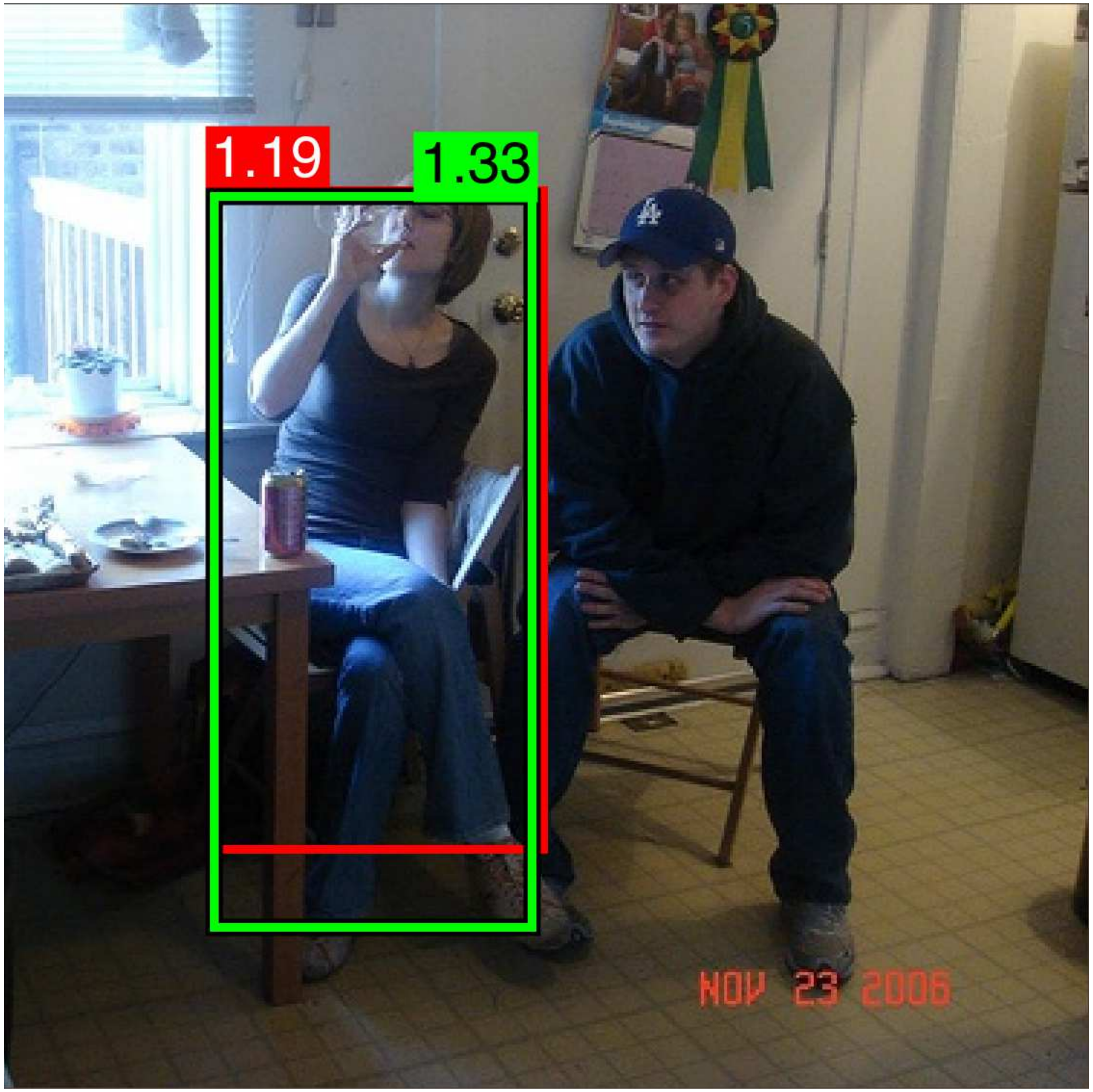
Iteration 6



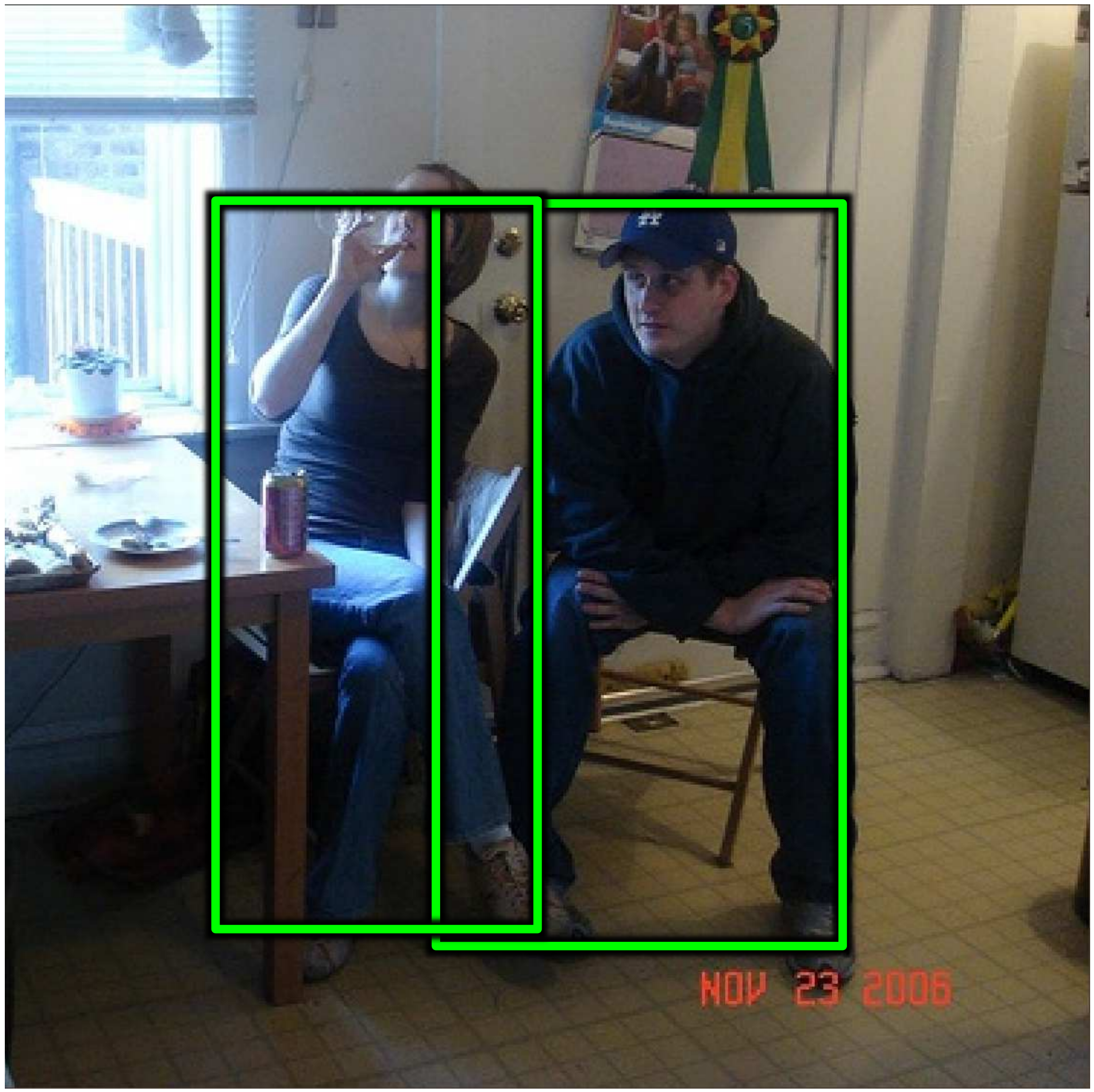
Iteration 7



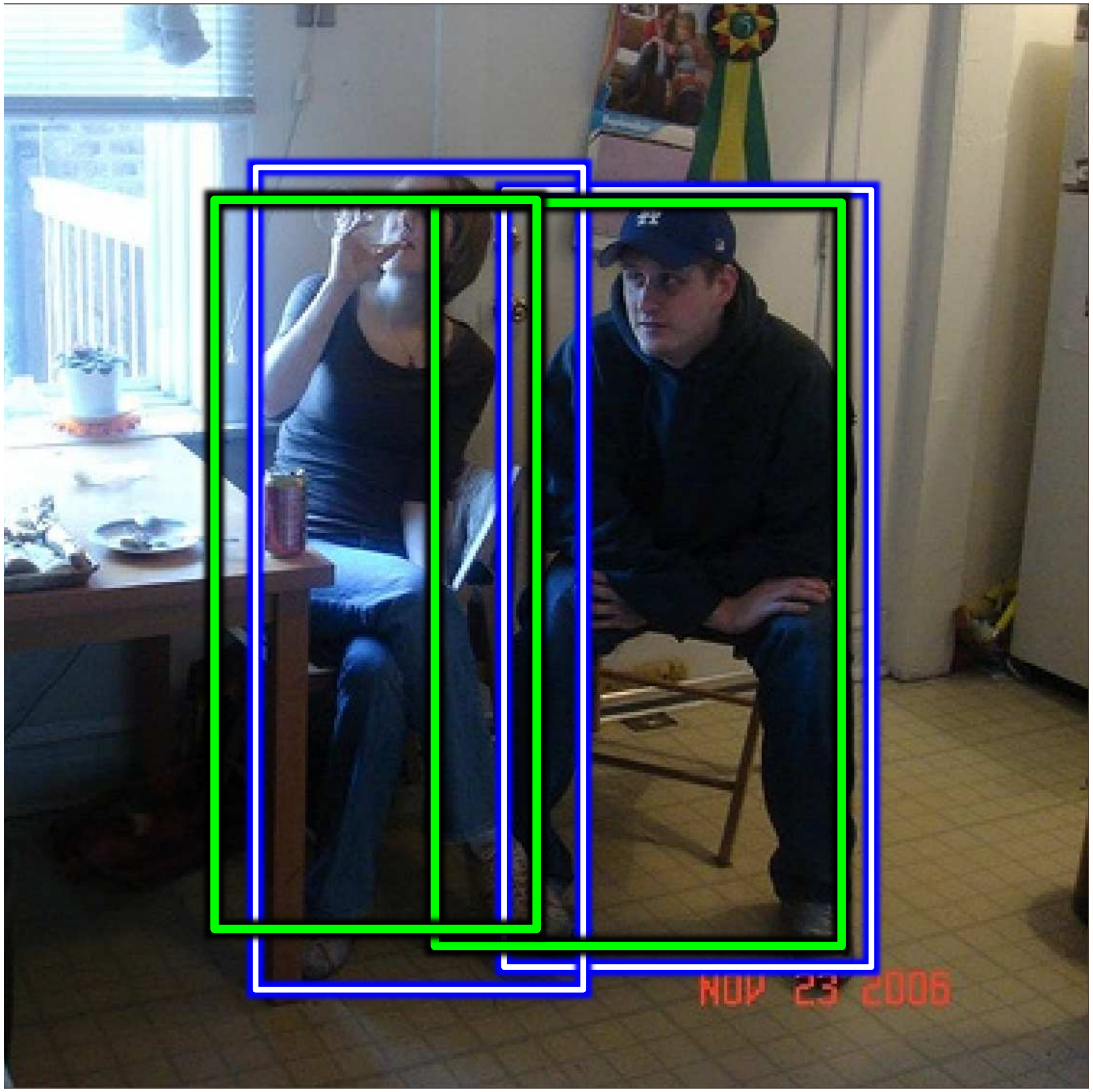
Iteration 8



Final results



Final results & Ground truth



Thrust 2:
Train CNN classifier with structured output regression

Structured loss for detection

- Linear classifier

$$g(x; w) = \mathrm{argmax}_{y \in \mathcal{Y}} f(x, y; w)$$

$$f(x, y; w) = w^\top \tilde{\phi}(x, y)$$

CNN features

$$\tilde{\phi}(x, y) = \begin{cases} \phi(x, y), & l = +1 \\ \mathbf{0}, & l = -1 \end{cases}$$

- Minimizing the structured loss (Blaschko and Lampert, 2008)*

$$\hat{w} = \mathrm{argmax}_w \sum_{i=1}^{M} \Delta(g(x_i; w), y_i)$$

$$\Delta(y, y_i) = \begin{cases} 1 - \mathrm{IoU}(y, y_i), & \text{if } l = l_i = 1 \\ 0, & \text{if } l = l_i = -1 \\ 1, & \text{if } l \neq l_i \end{cases}$$

* Blaschko and Lampert, "Learning to localize objects with structured output regression", ECCV, 2008.

Other related work: LeCun et al. 1989; Taskar et al. 2005; Joachims et al. 2005; Veldaldi et al. 2014; Thomson et al. 2014; and many others

Structured SVM for detection

- The objective is hard to solve. Replace it with an upper-bound surrogate using structured SVM framework

$$\min_{w} \quad \frac{1}{2} \parallel w \parallel^2 + \frac{C}{M} \sum_{i=1}^{M} \xi_i \qquad \text{, subject to}$$

$$w^\top \tilde{\phi}(x_i, y_i) \geq w^\top \tilde{\phi}(x_i, y) + \Delta(y, y_i) - \xi_i, \forall y \in \mathcal{Y}, \forall i$$

$$\xi_i \geq \; 0, \forall i$$

- The constraints can be re-written as:

$$\left.\begin{array}{ll} w^\top \phi(x_i, y_i) \geq \; 1 - \xi_i, & \forall i \in I_{\text{pos}}, \\ w^\top \phi(x_i, y) \leq \; -1 + \xi_i, & \forall y \in \mathcal{Y}, \forall i \in I_{neg}, \end{array}\right\} \text{Recognition}$$

$$\left.\begin{array}{l} w^\top \phi(x_i, y_i) \geq \; w^T \phi(x_i, y) + \Delta^{\text{loc}}(y, y_i) - \xi_i, \\ \forall y \in \mathcal{Y}, \forall i \in I_{\text{pos}}, \end{array}\right\} \text{Localization}$$

where $\Delta^{\text{loc}}(y, y_i) = 1 - \text{IoU}(y, y_i)$.

Solution for Structured SVM

- Approximate the structured output space $\mathcal{Y}$ with samples from selective search and random boxes near ground truths.

- Gradient-based method
  - Opt 1: LBFG-S for learning classification layer
  - Opt 2: SGD for fine-tuning the whole CNN

- Hard sample mining according to hinge loss
  - Not all the training samples can fit into memory
  - Significantly reduce the time consumption for searching the most violated sample

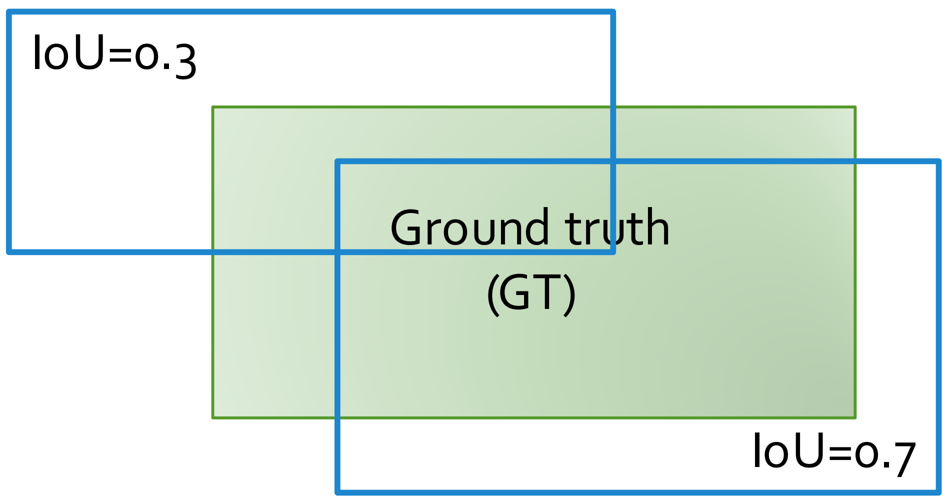
Experimental results

Control experiments with Oracle detector

- Oracle detector for image x_i, and ground truth box y_i

$$f_{ideal}(x_i, y) = \mathrm{IoU}(y, y_i)$$

where IoU is the **intersection over union**.



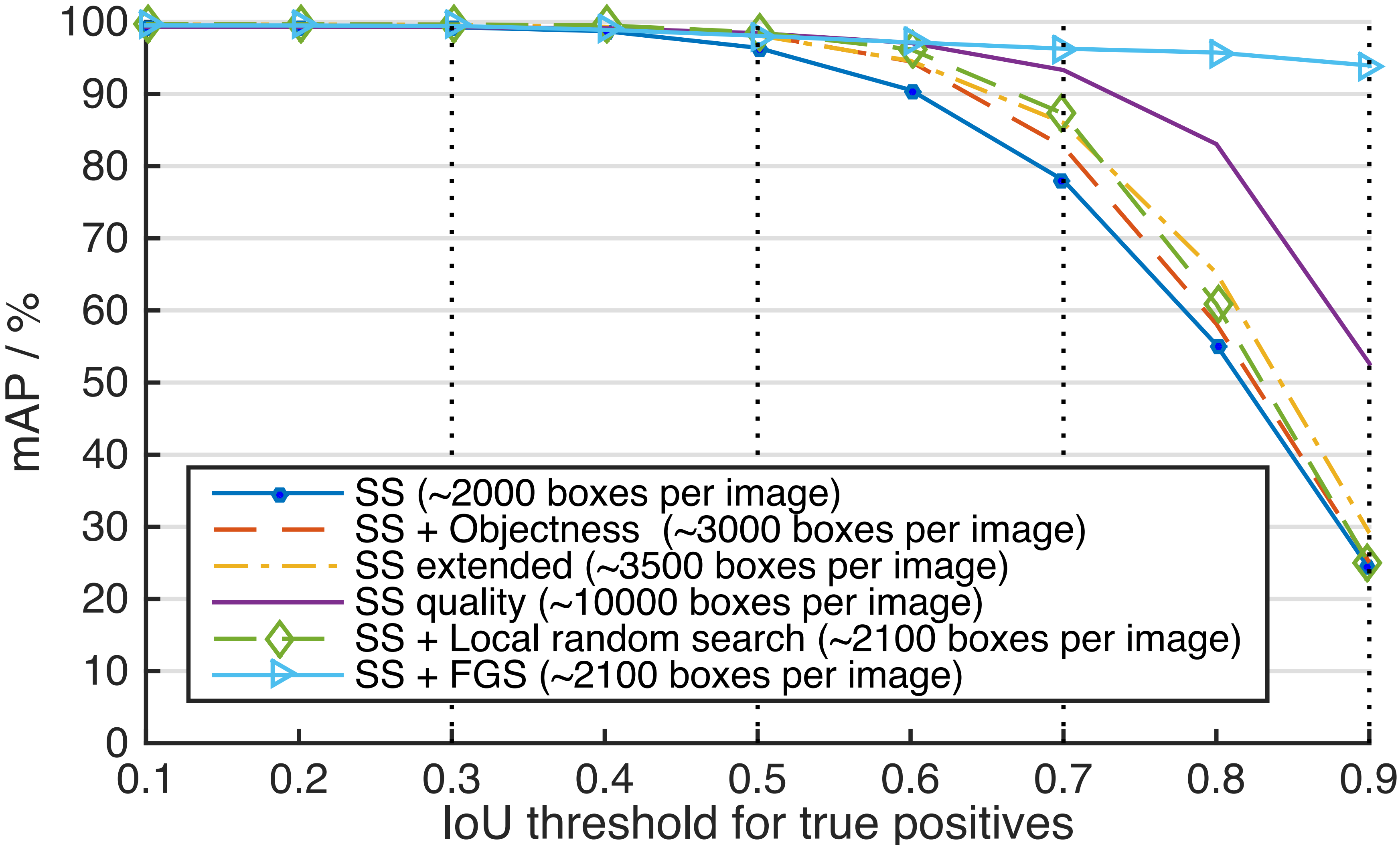
Controlled experiments with Oracle detector

More region proposal methods:

- SS: selective search
  fast (default) / extended / quality
- Objectness*
- Local random search:
  Random generate extra boxes without Bayesian optimization

* Alexe, B., Deselaers, T., & Ferrari, V. (2012). Measuring the objectness of image windows. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *34*(11), 2189-2202.

Controlled experiments with Oracle detector

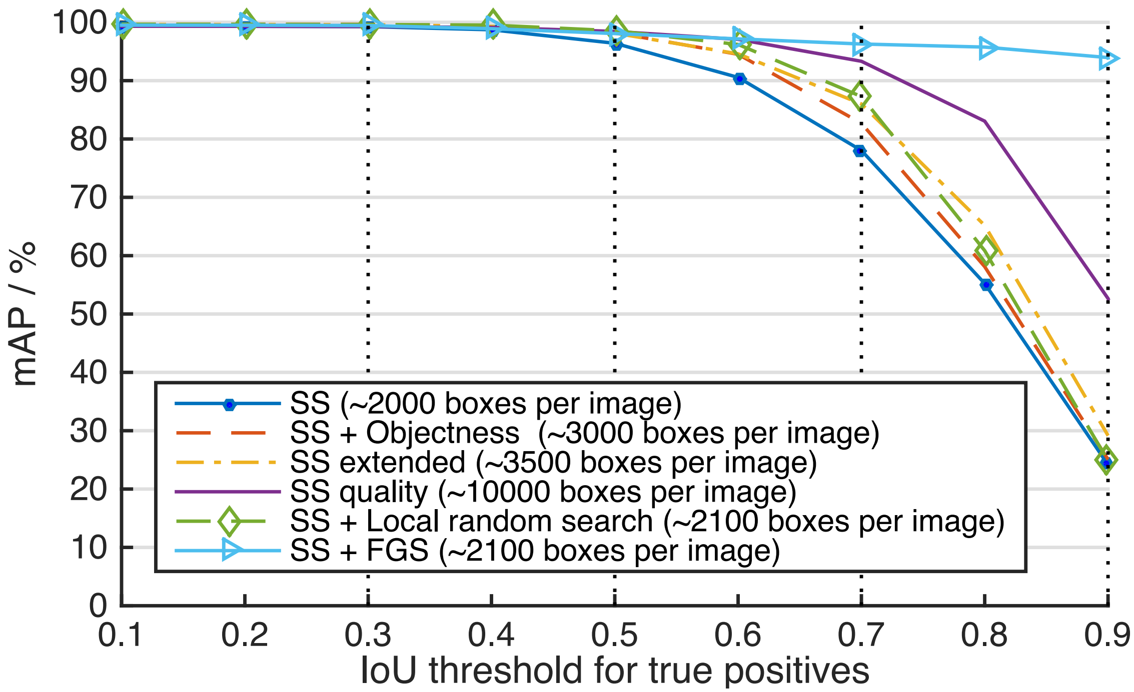


- x-axis: Different IoU thresholds for accepting a true positive
- y-axis: mean average precision (mAP)

Control experiments with Oracle detector

More region proposal methods:

- SS: selective search
  fast (default) / extended / quality
- Objectness
- Local random search:
  Random generate extra boxes without Bayesian optimization

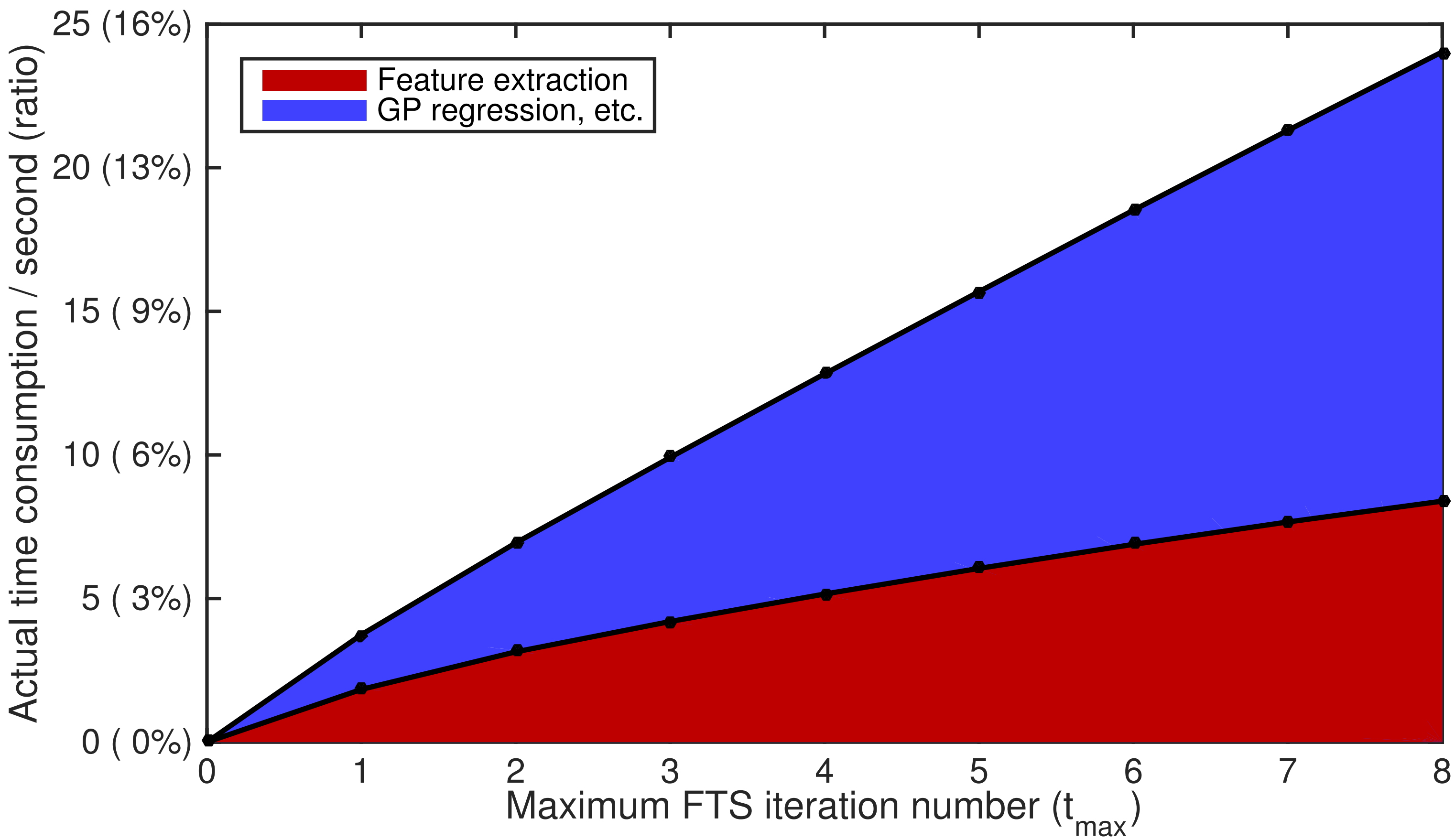


Results:

- x-axis:
  Different IoU thresholds for accepting a true positive
- y-axis:
  mean average precision (mAP)

FGS efficiency: time overhead

- Baseline time: Initial feature extraction time of R-CNN



mAP on VOC2007 test set

Mean Average Precision	Standard localization
R-CNN (AlexNet)	58.5
R-CNN (VGGNet)	65.4

Bounding box regression is always taken as a post-processing step.

mAP on VOC2007 test set

Mean Average Precision	Standard localization
R-CNN (AlexNet)	58.5
R-CNN (VGGNet)	65.4
+ StructObj	66.6
+ StructObj-FT	66.9

1.2%

mAP on VOC2007 test set

Mean Average Precision	Standard localization
R-CNN (AlexNet)	58.5
R-CNN (VGGNet)	65.4
+ StructObj	66.6
+ StructObj-FT	66.9
+ FGS	67.2

1.8%

mAP on VOC2007 test set

Mean Average Precision	Standard localization
R-CNN (AlexNet)	58.5
R-CNN (VGGNet)	65.4
+ StructObj	66.6
+ StructObj-FT	66.9
+ FGS	67.2
+ FGS + StructObj	**68.5**
+ FGS + StructObj-FT	68.4

3.1%

mAP on VOC2007 test set

	IoU>0.5	IoU>0.7
Mean Average Precision	**Standard localization**	**More accurate localization**
R-CNN (AlexNet)	58.5	?
R-CNN (VGGNet)	65.4	
+ StructObj	66.6	
+ StructObj-FT	66.9	
+ FGS	67.2	
+ FGS + StructObj	**68.5**	
+ FGS + StructObj-FT	68.4	

mAP on VOC2007 test set

	IoU>0.5	IoU>0.7
Mean Average Precision	**Standard localization**	**More accurate localization**
R-CNN (AlexNet)	58.5	35.2
R-CNN (VGGNet)	65.4	35.2
+ StructObj	66.6	40.5
+ StructObj-FT	66.9	41.8
+ FGS	67.2	42.7
+ FGS + StructObj	**68.5**	43.0
+ FGS + StructObj-FT	68.4	**43.7**

mAP on VOC2007 test set

Mean Average Precision	IoU>0.5 Standard localization	IoU>0.7 More accurate localization
R-CNN (AlexNet)	58.5	35.2
R-CNN (VGGNet)	65.4	35.2
+ StructObj	66.6	40.5
+ StructObj-FT	66.9	41.8
+ FGS	67.2	42.7
+ FGS + StructObj	**68.5**	43.0
+ FGS + StructObj-FT	68.4	**43.7**

7.8%

mAP on VOC2007 test set

	IoU>0.5	IoU>0.7
Mean Average Precision	**Standard localization**	**More accurate localization**
R-CNN (AlexNet)	58.5	35.2
R-CNN (VGGNet)	65.4	35.2
+ StructObj	66.6	40.5
+ StructObj-FT	66.9	41.8
+ FGS	67.2	42.7
+ FGS + StructObj	**68.5**	43.0
+ FGS + StructObj-FT	68.4	**43.7**

8.6%

mAP on VOC2012 test set

Mean Average Precision	IoU>0.5
R-CNN (AlexNet)	53.3
R-CNN (VGGNet)	63.0
+ StructObj	65.1
+ FGS	64.0
+ FGS + StructObj	**66.4**

3.4%

mAP on VOC2012 test set

Mean Average Precision	IoU>0.5
R-CNN (AlexNet)	53.3
R-CNN (VGGNet)	63.0
+ StructObj	65.1
+ FGS	64.0
+ FGS + StructObj	**66.4**
Network in Network*	63.8

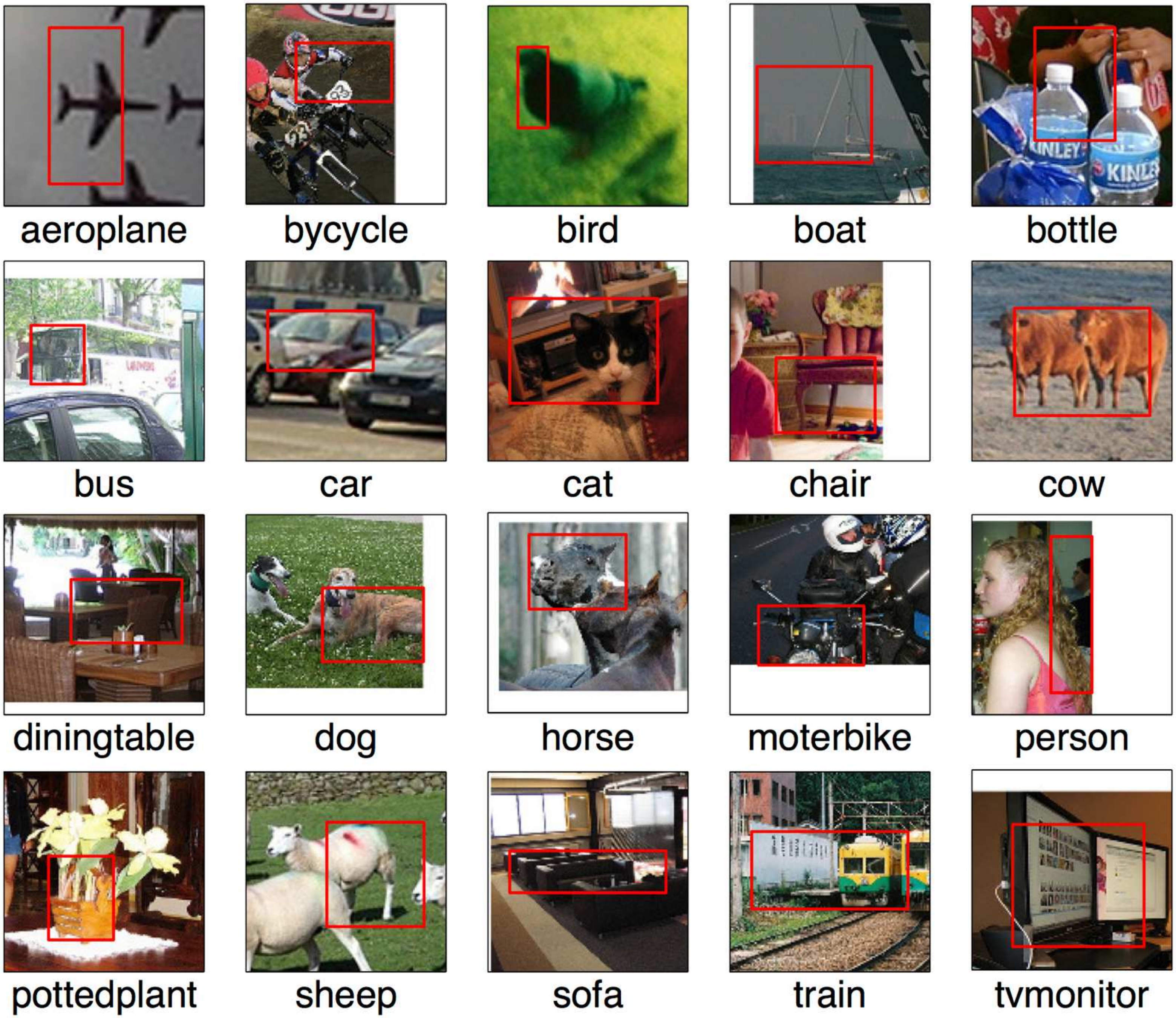
2.6%

*M. Lin, Q. Chen, S. Yan, Network In Network, ICLR 2014

Good examples on VOC 2007 (1)

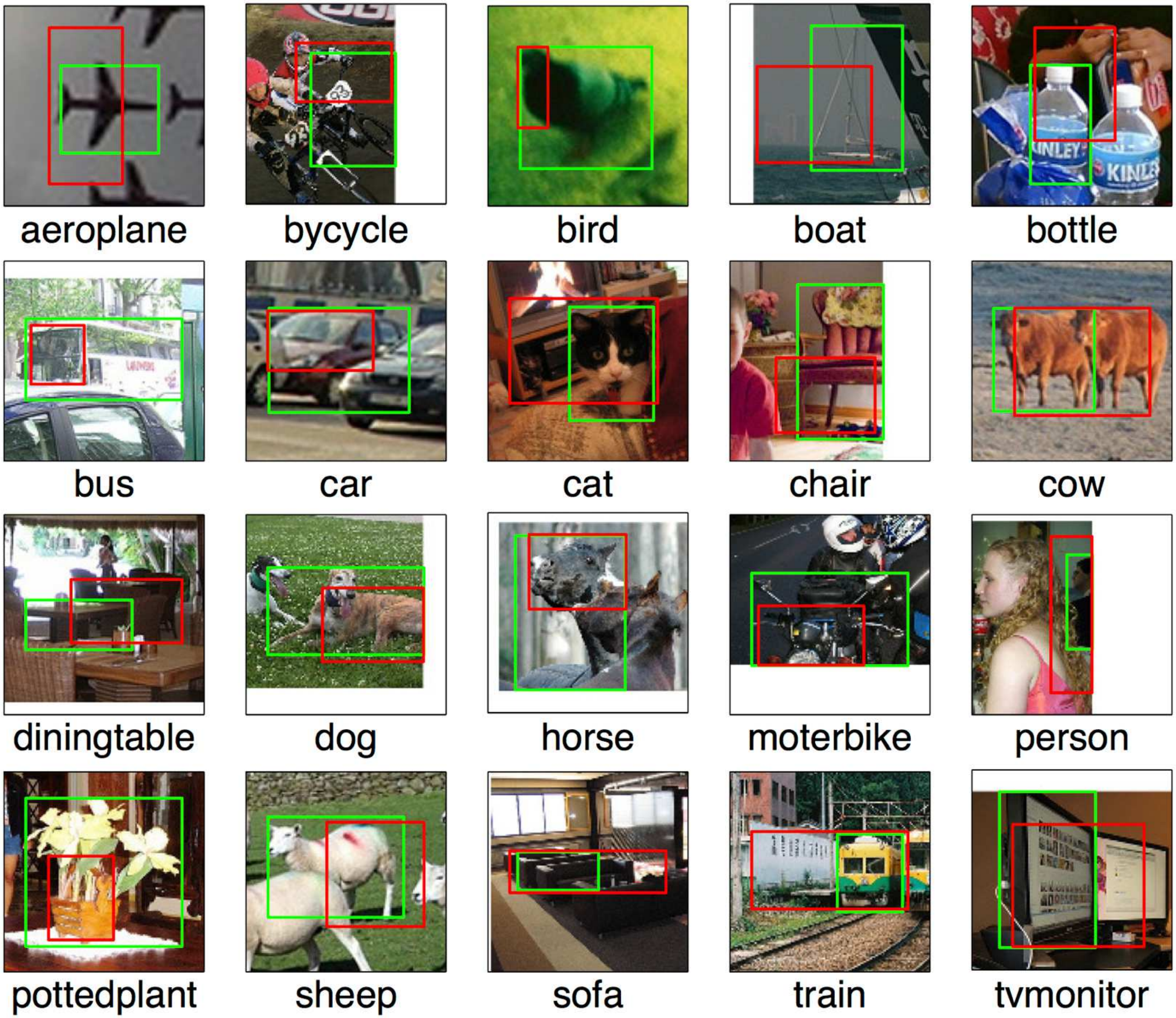
Original image

Good examples on VOC 2007 (1)



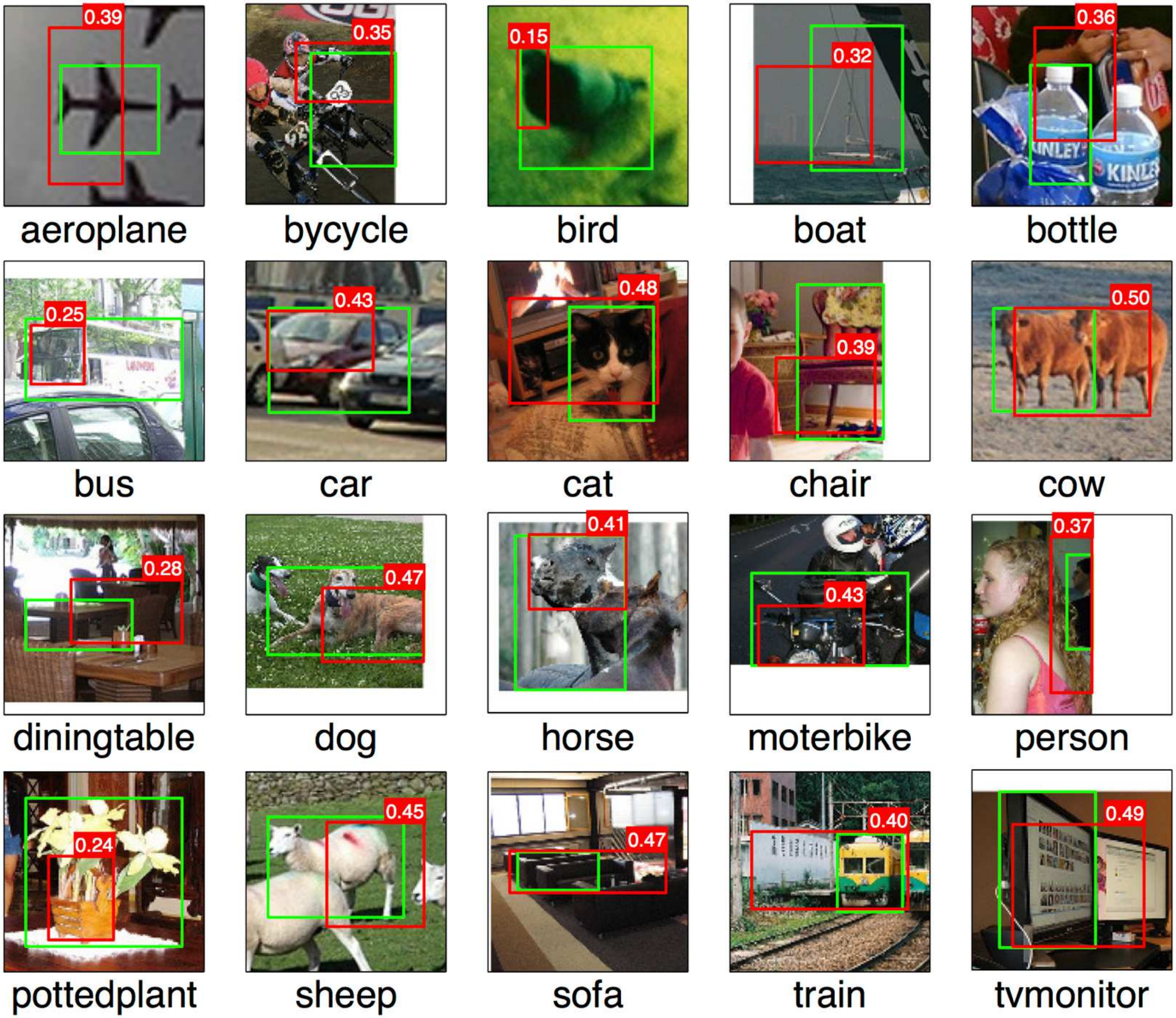
Red boxes:
R-CNN (VGGNet) baseline.

Good examples on VOC 2007 (1)



Red boxes: R-CNN (VGGNet) baseline.

Green boxes: Ground truth(GT)

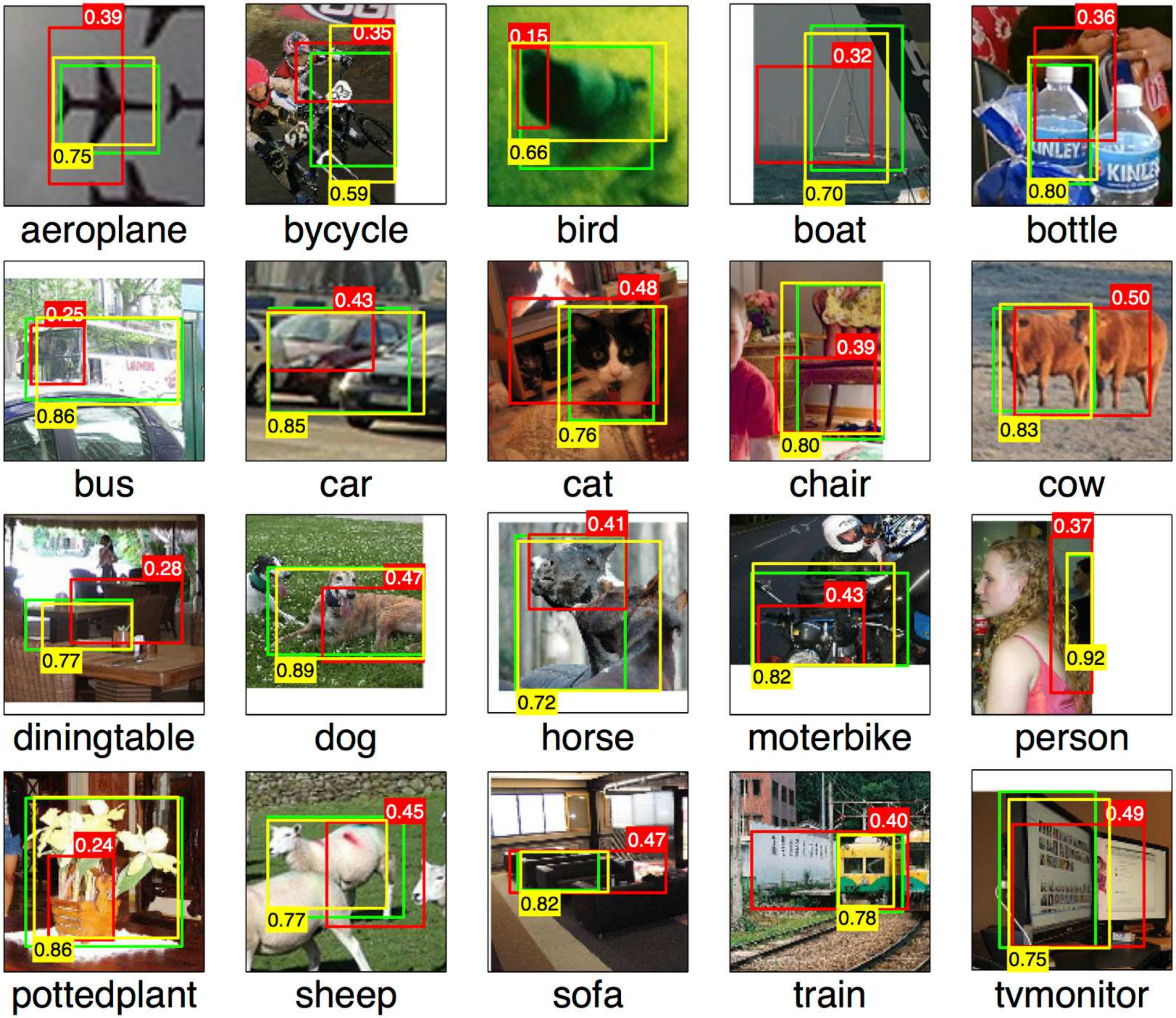


Good examples on VOC 2007 (1)

Numbers:
Overlap (IoU) with GT

Red boxes:
R-CNN (VGGNet) baseline.

Green boxes:
Ground truth(GT)



Good examples on VOC 2007 (1)

Numbers:
Overlap (IoU) with GT

Red boxes:
R-CNN (VGGNet) baseline.

Green boxes:
Ground truth(GT)

Yellow boxes:
Ours (+ StructObj + FGS)

Good examples on VOC 2007 (2)

Original image


aeroplane


bycycle


bird


boat


bottle

bus


car


cat


chair


cow


diningtable


dog


horse


moterbike


person


pottedplant

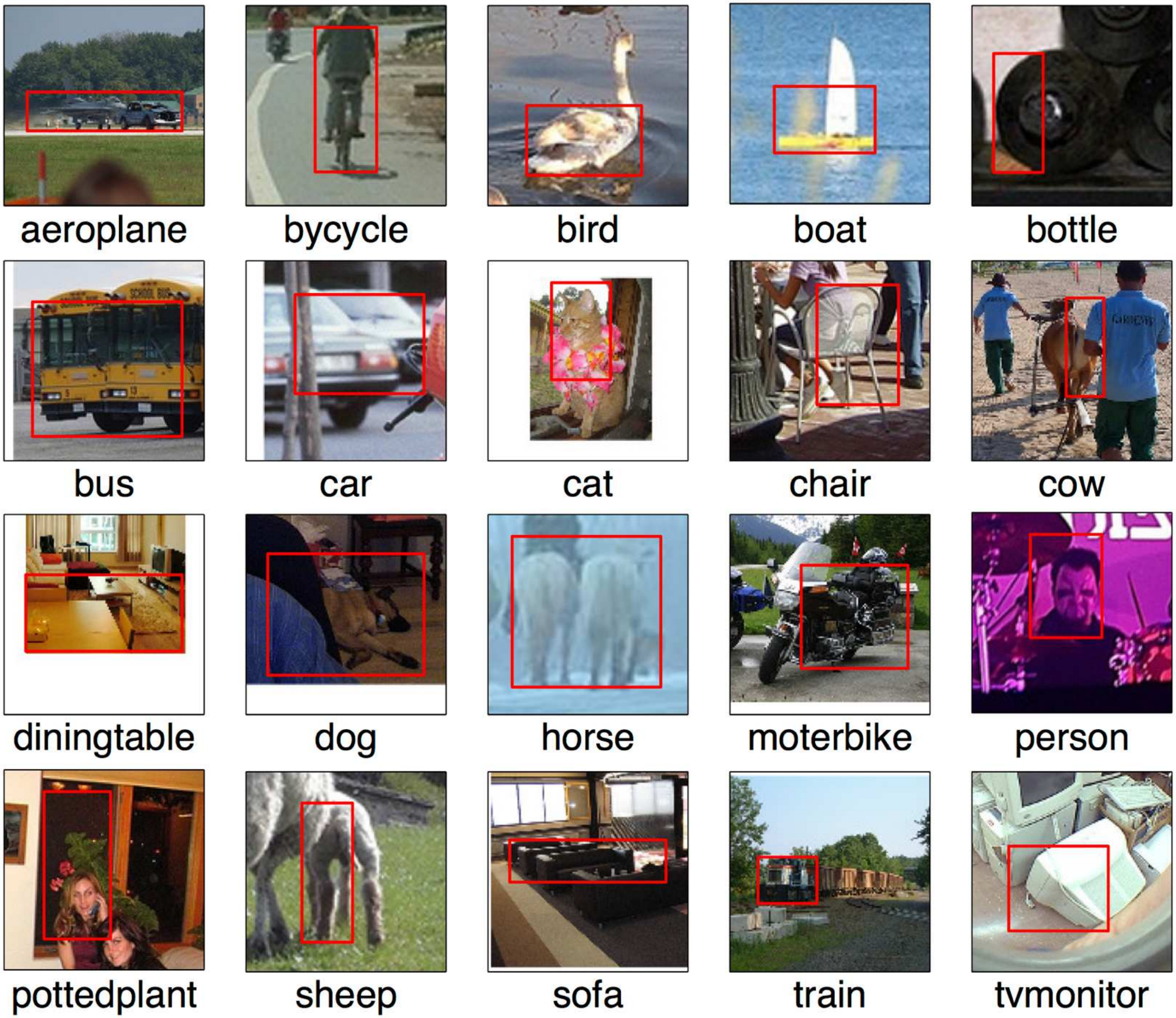

sheep


sofa


train

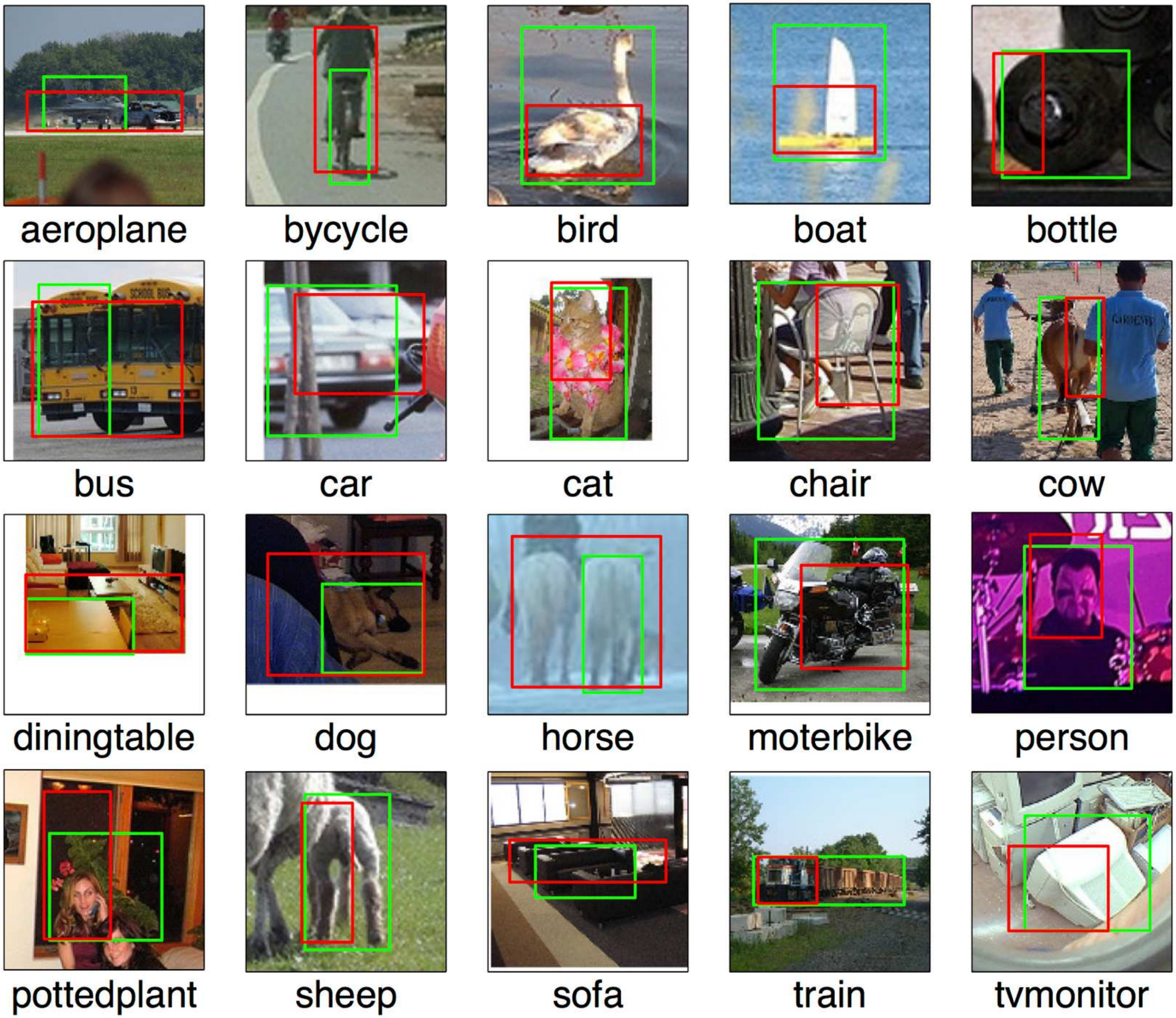

tvmonitor

Good examples on VOC 2007 (2)



Red boxes:
R-CNN (VGGNet) baseline.

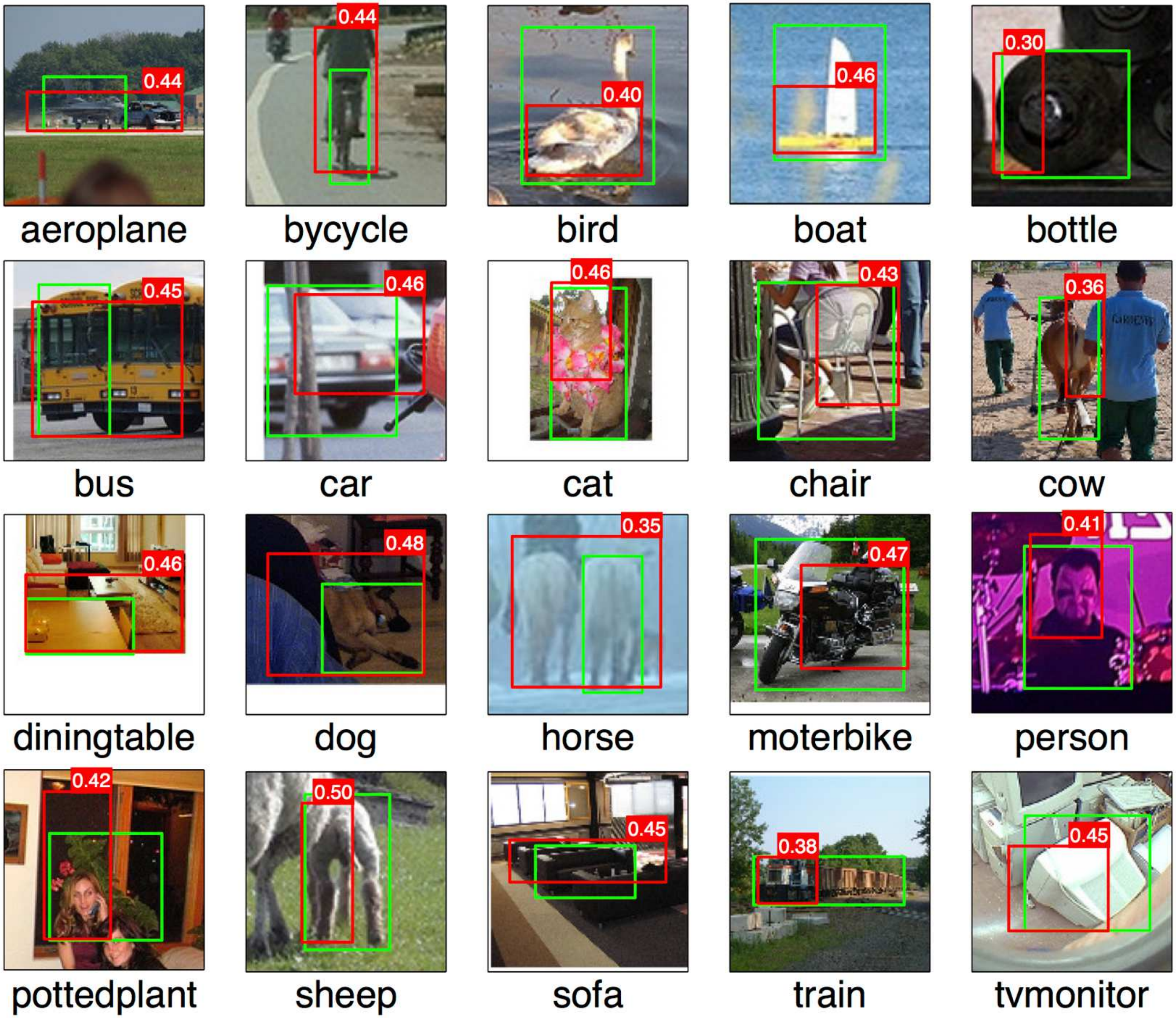
Good examples on VOC 2007 (2)



Red boxes: R-CNN (VGGNet) baseline.

Green boxes: Ground truth(GT)

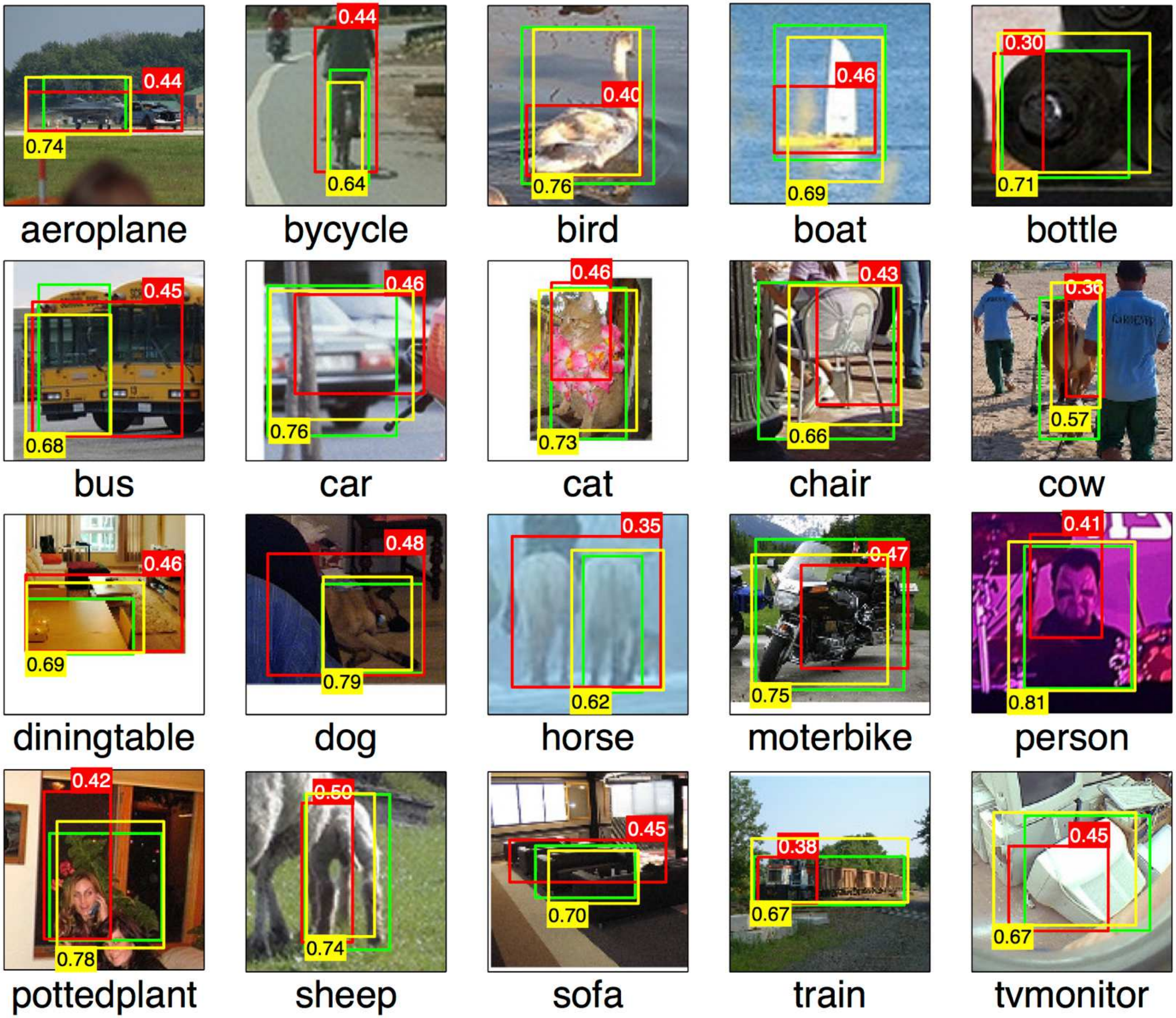
Good examples on VOC 2007 (2)



Numbers:
Overlap (IoU) with GT

Red boxes:
R-CNN (VGGNet) baseline.

Green boxes:
Ground truth(GT)



Good examples on VOC 2007 (2)

Numbers:
Overlap (IoU) with GT

Red boxes:
R-CNN (VGGNet) baseline.

Green boxes:
Ground truth(GT)

Yellow boxes:
Ours (+ StructObj + FGS)

Conclusion

- We proposed two complementary methods for improving object detection
  1. Find better bounding boxes via Bayesian optimization
  2. Improve localization sensitivity via structured objective
- If the object classifier is accurate, our fine-grained search algorithm is almost as good as doing exhaustive search.
  - compatible with most detection methods.
- We significantly improve over the previous state-of-the-art in object detection both for VOC 2007 and 2012 benchmarks.

Code available at :

bit.ly/fgs-obj

Q & A

Thank you!

References

- B. Alexe, T. Deselaers, and V. Ferrari. Measuring the objectness of image windows. IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(11):2189–2202, Nov 2012. 6
- Y. Bengio, P. Lamblin, D. Popovici, H. Larochelle, et al. Greedy layer-wise training of deep networks. In NIPS, 2007.
- Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. IEEE Transac- tions on Pattern Analysis and Machine Intelligence, 35(8): 1798–1828, Aug 2013.
- M. B. Blaschko and C. H. Lampert. Learning to localize objects with structured output regression. In ECCV, 2008.
- Y.-l. Boureau, Y. L. Cun, et al. Sparse feature learning for deep belief networks. In NIPS, pages 1185–1192, 2008.
- J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In CVPR, 2009.
- J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. DeCAF: A deep convolutional activation feature for generic visual recognition. CoRR, abs/1310.1531, 2013.
- D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov. Scalable object detection using deep neural networks. In CVPR, 2014.
- M. Everingham, L. VanGool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Chal- lenge 2007 (VOC2007) Results, 2007.
- M.Everingham,L.VanGool,C.K.I.Williams,J.Winn,and A. Zisserman. The PASCAL Visual Object Classes Chal- lenge 2012 (VOC2012) Results, 2012.
- P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part- based models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(9):1627–1645, 2010.

- R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In CVPR, 2014.
- R. Girshick, J. Donahue, T. Darrell, J. Malik. Region-based Convolutional Networks for Accurate Object Detection and Semantic Segmentation. In IEEE PAMI, 2015.
- C. Gu, J. J. Lim, P. Arbelaez, and J. Malik. Recognition using regions. In CVPR, 2009.
- D. Hoiem, Y. Chodpathumwan, and Q. Dai. Diagnosingerror in object detectors. In ECCV, 2012.
- Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. B. Girshick, S. Guadarrama, and T. Darrell. Caffe: Con- volutional architecture for fast feature embedding. CoRR, abs/1408.5093, 2014.
- D. R. Jones. A taxonomy of global optimization methods based on response surfaces. Journal of Global Optimization, 21(4):345–383, 2001.
- A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In NIPS, 2012.
- Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. Neural Computation, 1(4):541–551, 1989.
- H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng. Unsuper- vised learning of hierarchical representations with convolutional deep belief networks. Communications of the ACM, 54(10):95–103, 2011.
- M. Lin, Q. Chen, and S. Yan. Network in network. CoRR, abs/1312.4400, 2013.
- J. Mockus, V. Tiesis, and A. Zilinskas. The application of bayesian methods for seeking the extremum. Towards Global Optimization, 2(117-129):2, 1978.
- C. Rasmussenand and C. Williams. Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning). The MIT Press, 2006.
- 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, 2014.
- J. Schmidhuber. Deep learning in neural networks: An overview. Neural Networks, 61:85–117, 2015.

- S. Schulter, C. Leistner, P. Wohlhart, P. M. Roth, and H. Bischof. Accurate object detection with joint classification-regression random forests. In CVPR, 2014.
- P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. OverFeat: Integrated recognition, localization and detection using convolutional networks. In ICLR, 2014.
- K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In ICLR, 2015.
- J. Snoek, H. Larochelle, and R.P.Adams. Practical bayesian optimization of machine learning algorithms. In NIPS, 2012.
- C. Szegedy, A. Toshev, and D. Erhan. Deep neural networks for object detection. In NIPS, 2013.
- C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. arXiv preprint arXiv:1409.4842, 2014. 1
- I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. Journal of Machine Learning Research, (6): 1453–1484, 2005.
- J. R. R. Uijlings, K. E. A. Sande, T. Gevers, and A. W. M. Smeulders. Selective search for object recognition. International Journal of Computer Vision, 104(2):154–171, 2013