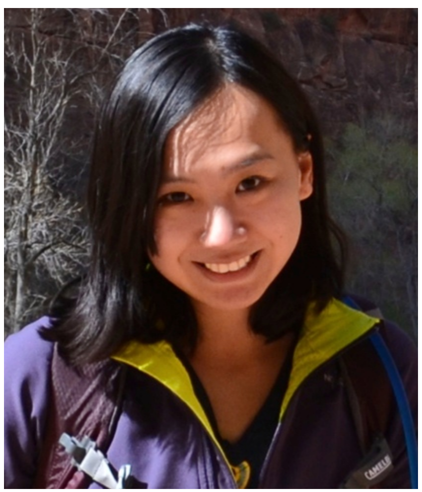
Part-based R-CNNs for Fine-grained Category Detection




Ning Zhang Jeff Donahue Ross Girshick Trevor Darrell

EECS, UC Berkeley

Challenges of Fine-grained Categorization

Black footed Albatross

Challenges of Fine-grained Categorization

Laysan Albatross

Finding correspondence

Blue headed vireo

???

White eyed vireo

Finding correspondence

Blue headed vireo

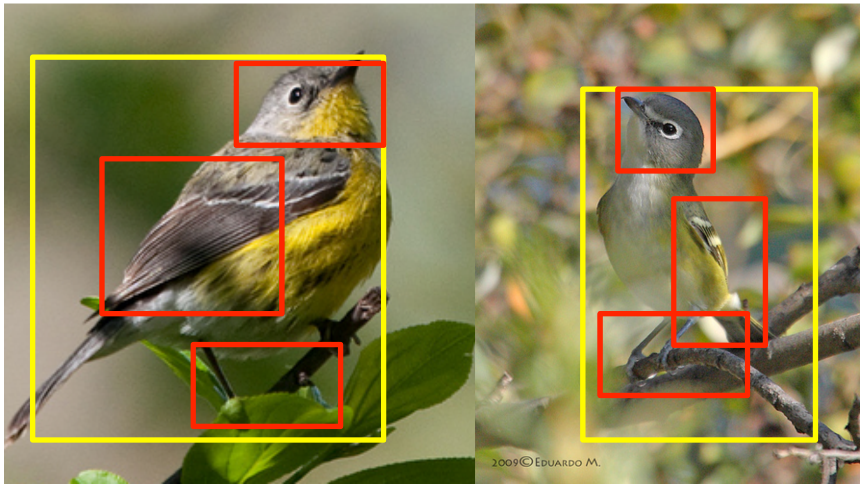
???

Blue headed vireo

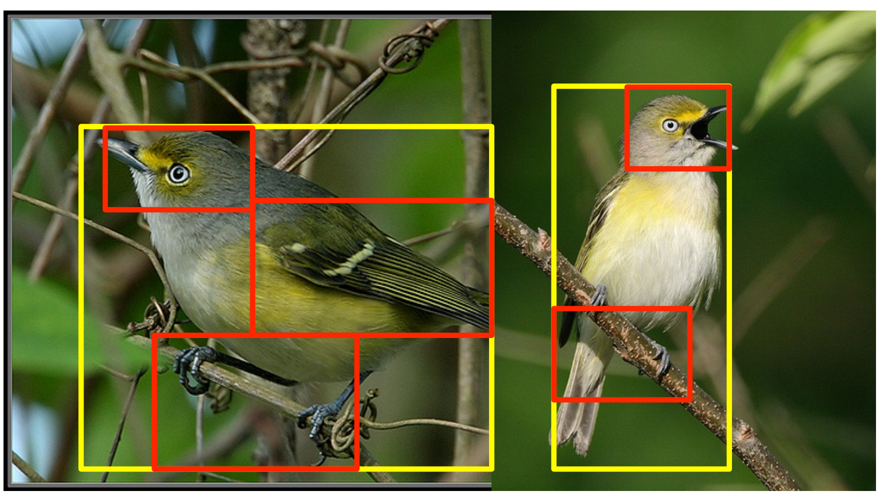
White eyed vireo

Pose-normalized correspondence

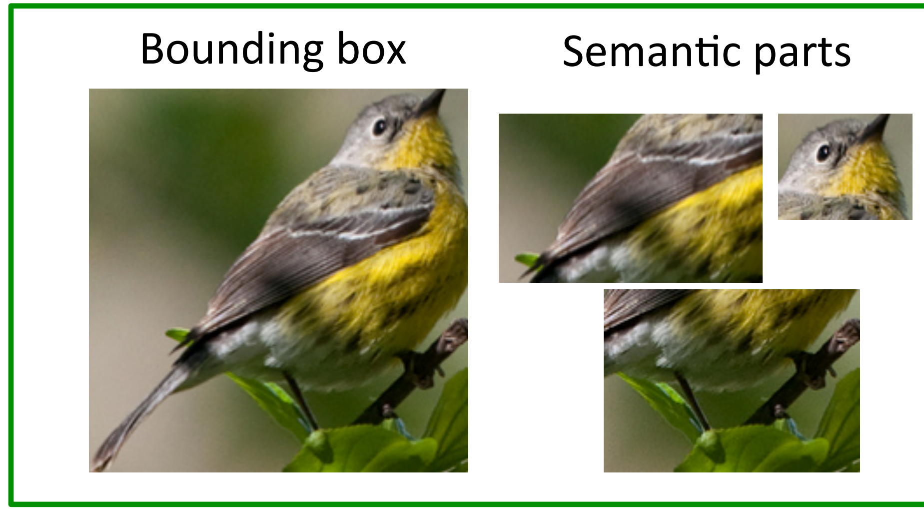
Blue headed vireo



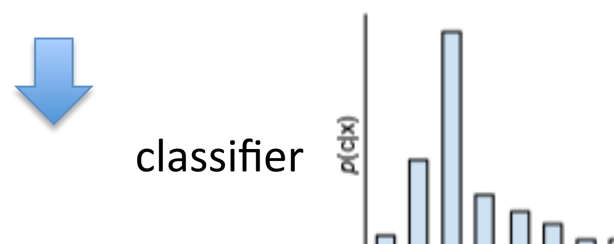
White eyed vireo



1) Correspondence

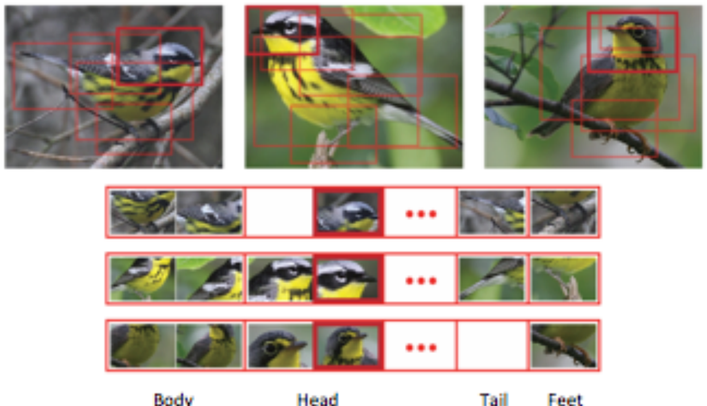


2) Feature representations



Prior work on fine-grained categorization

Correspondence

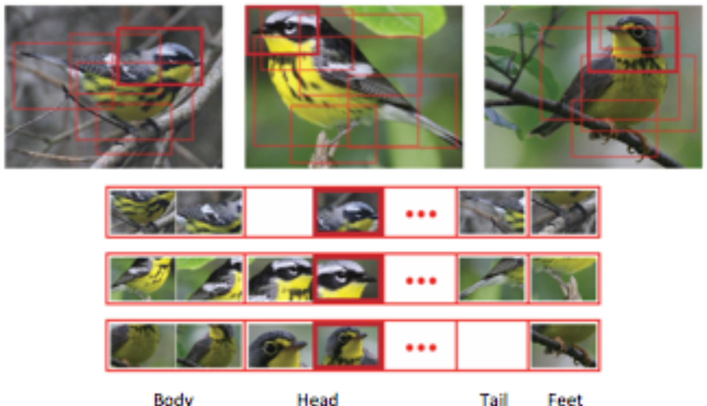


- [Farrell et.al. ICCV 2011]
- [Yao et.al. CVPR 2012]
- [Zhang et.al. CVPR 2012]
- [Liu et.al. ECCV 2012]
- [Yang et.al. NIPS 2012]
- [Berg et.al. CVPR 2013]
- [Chai et.al. ICCV 2013]
- [Gavves et.al. ICCV 2013]
- [Liu et.al. ICCV 2013]
- [Xie et.al. ICCV 2013]
- [Zhang et.al. ICCV 2013]
- [Göring et.al. CVPR 2014]

Bounding box assumed at test time

Prior work on fine-grained categorization

Correspondence



- [Farrell et.al. ICCV 2011]
- [Yao et.al. CVPR 2012]
- [Zhang et.al. CVPR 2012]
- [Liu et.al. ECCV 2012]
- [Yang et.al. NIPS 2012]
- [Berg et.al. CVPR 2013]
- [Chai et.al. ICCV 2013]
- [Gavves et.al. ICCV 2013]
- [Liu et.al. ICCV 2013]
- [Xie et.al. ICCV 2013]
- [Zhang et.al. ICCV 2013]
- [Göring et.al. CVPR 2014]

Feature representation

(color) SIFT:
- [Farrell et.al. ICCV 2011]
- [Zhang et.al. CVPR 2012]
- [Liu et.al. ECCV 2012]
- [Chai et.al. ECCV 2012]
- [Göring et.al. CVPR 2014]

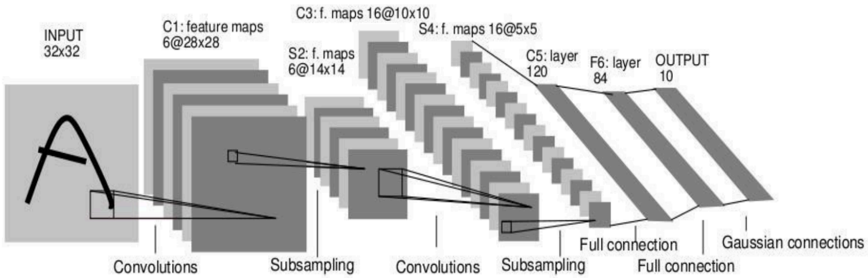
HOG:
- [Berg et al. CVPR 2013]
- [Liu et.al. ICCV 2013]

Fisher vector:
- [Chai et.al. ICCV 2013]
- [Gavves et.al. ICCV 2013]

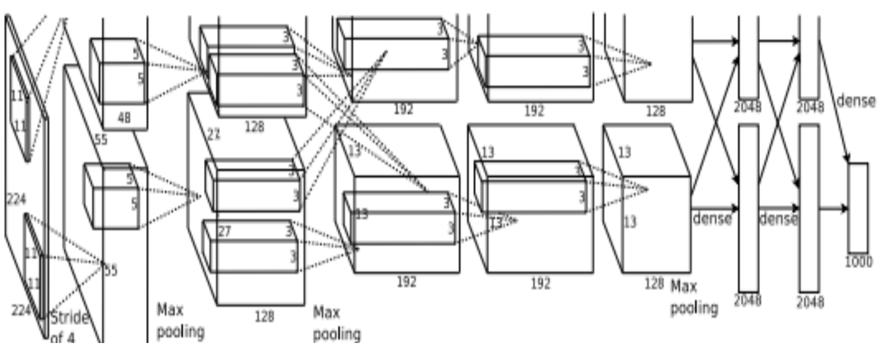
Kernel descriptors:
- [Yang et.al. NIPS 2012]
- [Zhang et.al. ICCV 2013]

Bounding box assumed at test time

Progress in deep learning



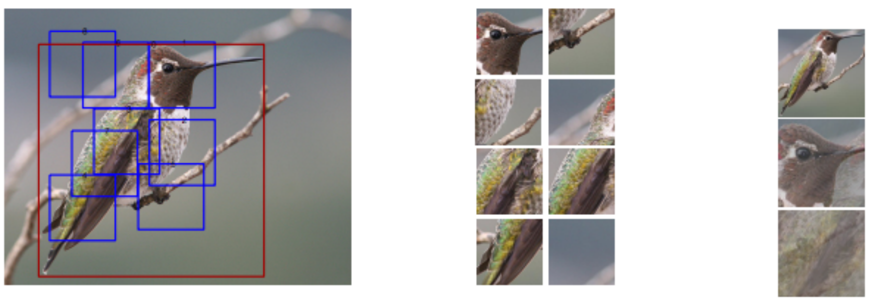
LeCun et.al. 1989-1998



[Krizhevsky et.al. NIPS 2012]

- **OCR** [Ciresan et.al. CVPR 2012] [Wen et.al. ICML 2013]
- **Pedestrian detection** [Sermanet et.al. CVPR 2013]
- **Scene parsing** [Farabet et.al. PAMI 2013]
- **Action recognition** [Karpathy et.al. CVPR 2014]
- **Face verification** [Taigman et.al. CVPR 2014]
- **Pose estimation** [Toshev et.al. CVPR 2014] [Jain et.al. ICLR 2014]
- **Object detection** [Girshick et.al. CVPR 2014] [Sermanet et.al. ICLR 2014]

Deep representations for fine-grained

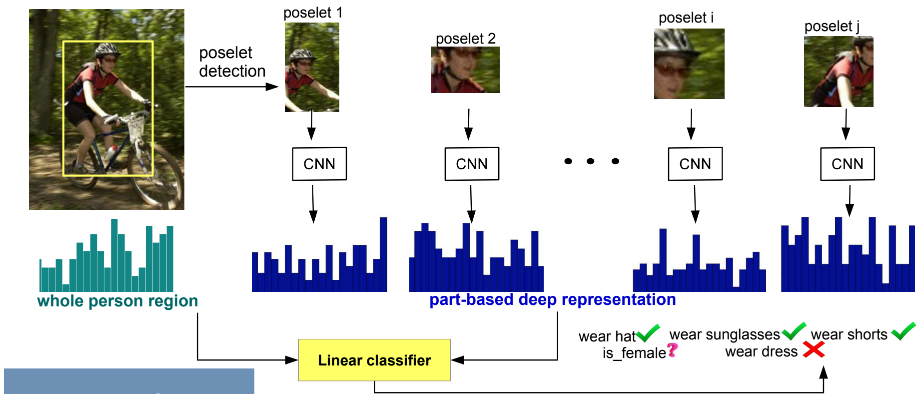


(a) DPM detections (b) Parts (c) DPD

Bounding box assumed

[Donahue et.al. ICML 2014]

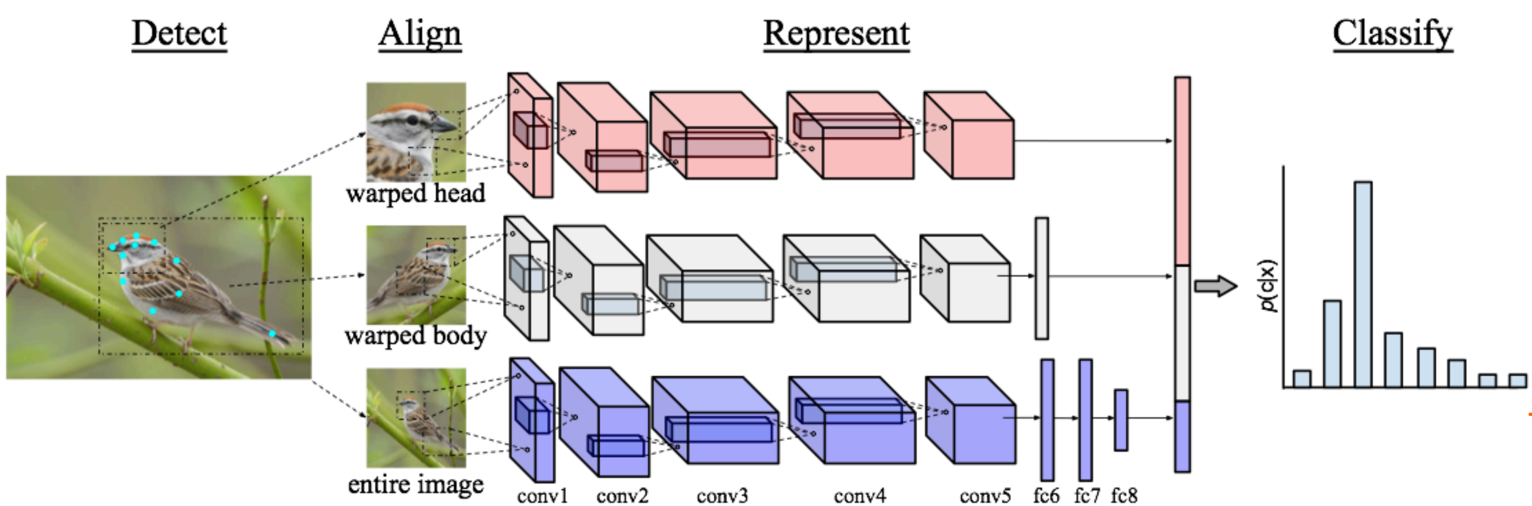
DPM detections + DeCAF feature



Bounding box assumed

[Zhang et.al. CVPR 2014]

poselet detections + deep network training from scratch



[Branson et.al. BMVC 2014.]

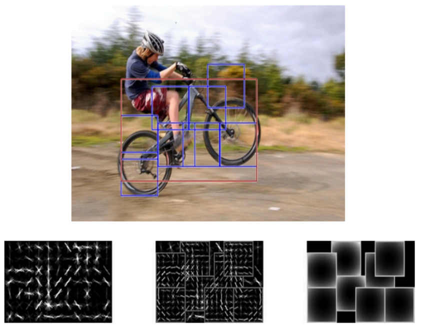
DPM keypoint detection

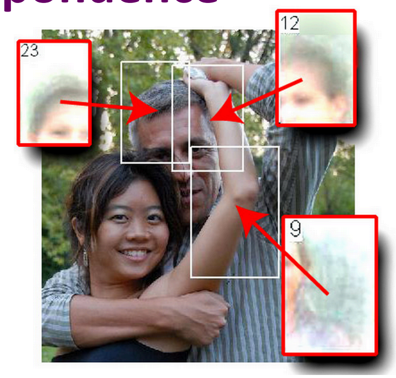
+

finetuned deep network

Limitations

To find correspondence





deformable part models poselets

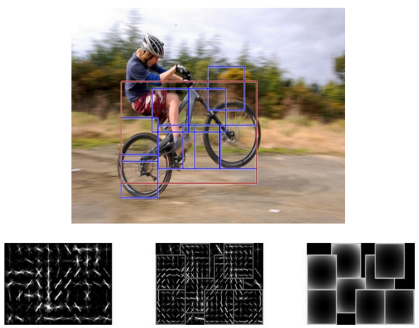
OR other part detectors

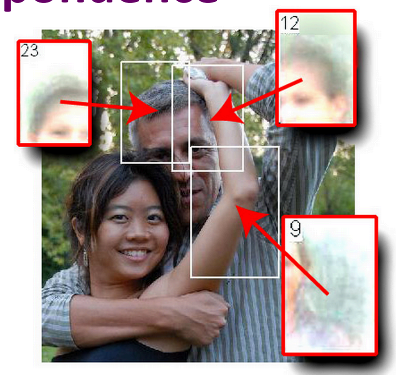
Hand-engineered feature(e.g. HOG)

Bounding box assumed at test time

Limitations

To find correspondence





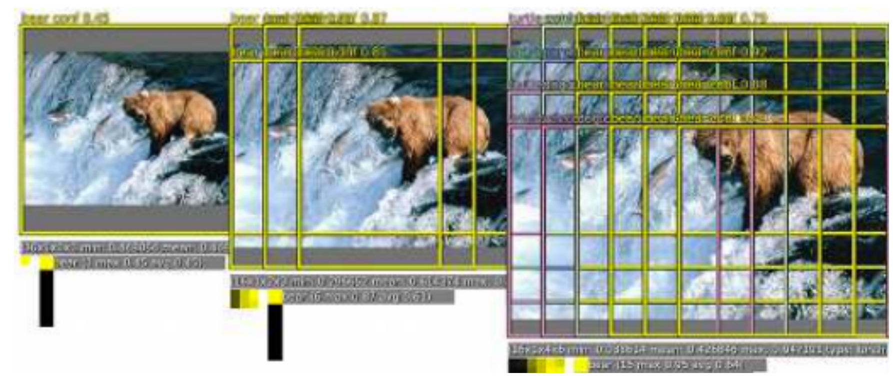
deformable part models poselets

OR other part detectors

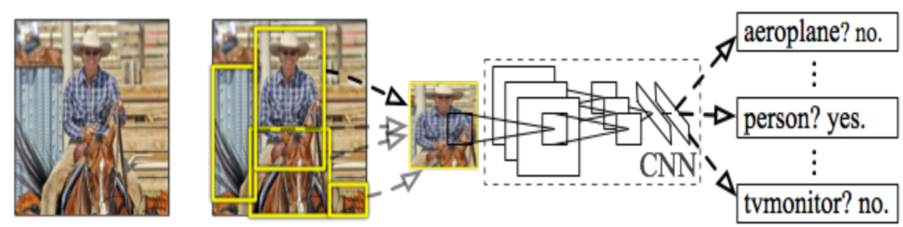
Hand-engineered feature(e.g. HOG)

Bounding box assumed at test time

Recent breakthrough for object detection



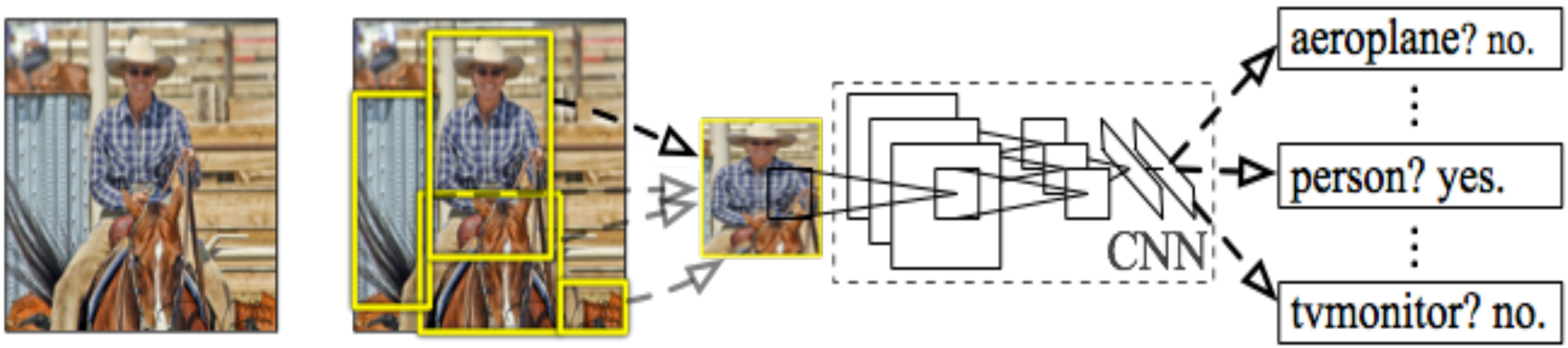
OverFeat [Sermanet et.al. ICLR 2014]



R-CNN [Girshick et.al. CVPR 2014]

Can we simultaneously detect objects and find part correspondences?

Extend RCNN to parts



Input image

Extract region proposals (~2k / image)

Compute CNN features

Classify regions (linear SVM)

Use part annotations. Treat object and parts as individual categories.

Try R-CNN https://github.com/rbgirshick/rcnn

Try CAFFE http://caffe.berkeleyvision.org

Girshick et.al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. CVPR, 2014

Unifying correspondence and feature learning

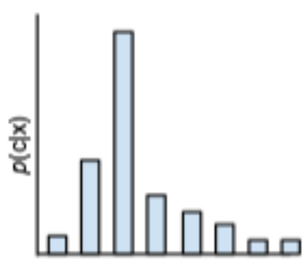
1) Correspondence

Bounding box

Semantic parts

2) Feature representations

classifier



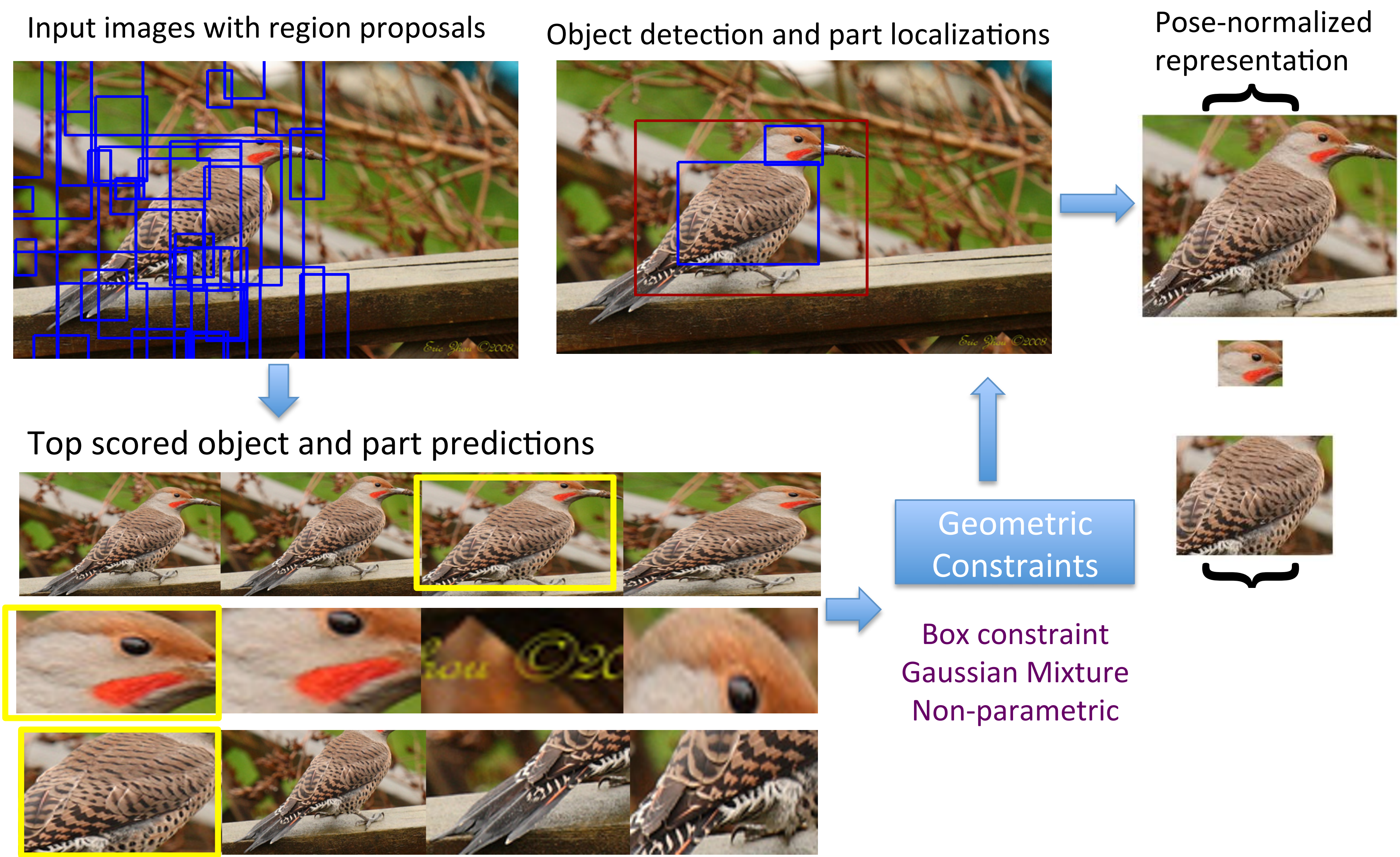
object detection
and part
localization

single deep
network

discriminative
feature learning

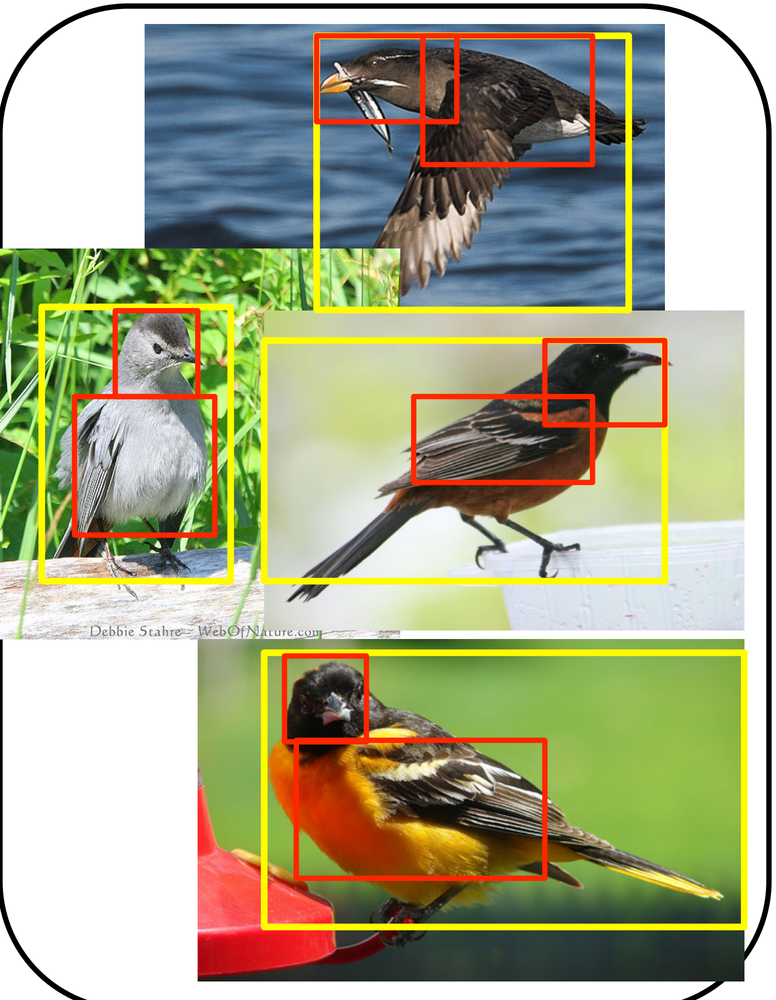
No more bounding box
assumption.

Overview of our approach

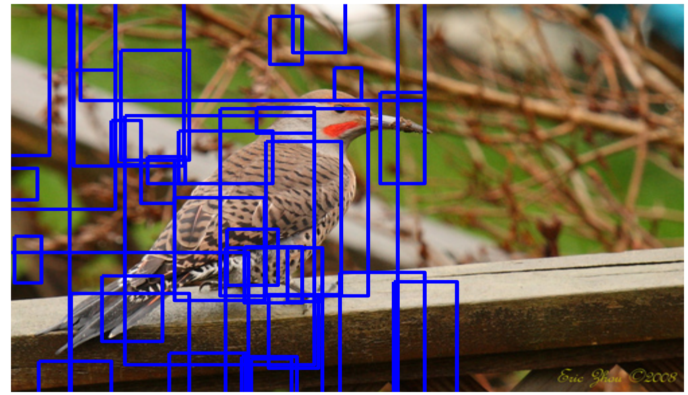


Object and Part detectors

Bounding box and part annotations

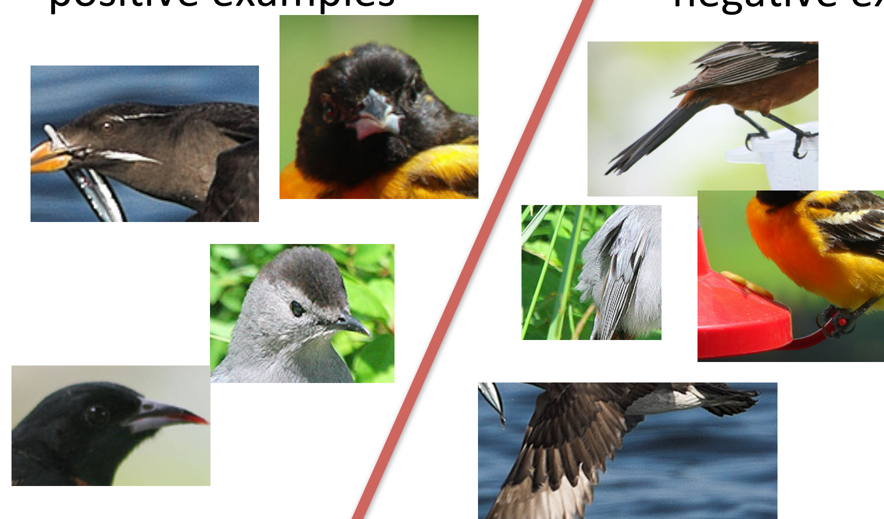


Region proposals using selective search



positive examples

negative examples



Object and Part detectors

R-CNN detection
for part i

$$d_i(x) = \sigma(w_i^\mathsf{T}\phi(x))$$

Learned detection weight

Deep convolution al feature

$\sigma(\cdot)$ is sigmoid function

Top scored object and part detections

d_0

d_1

d_2

Object and Part detectors

R-CNN detection
for part i

$$d_i(x) = \sigma(w_i^{\mathsf{T}} \phi(x))$$

Learned detection weight

Deep convolution al feature

$\sigma(\cdot)$ is sigmoid function

Top scored object and part detections

$$X^* = \arg\max_X \prod_{i=0}^{n} d_i(x_i)$$

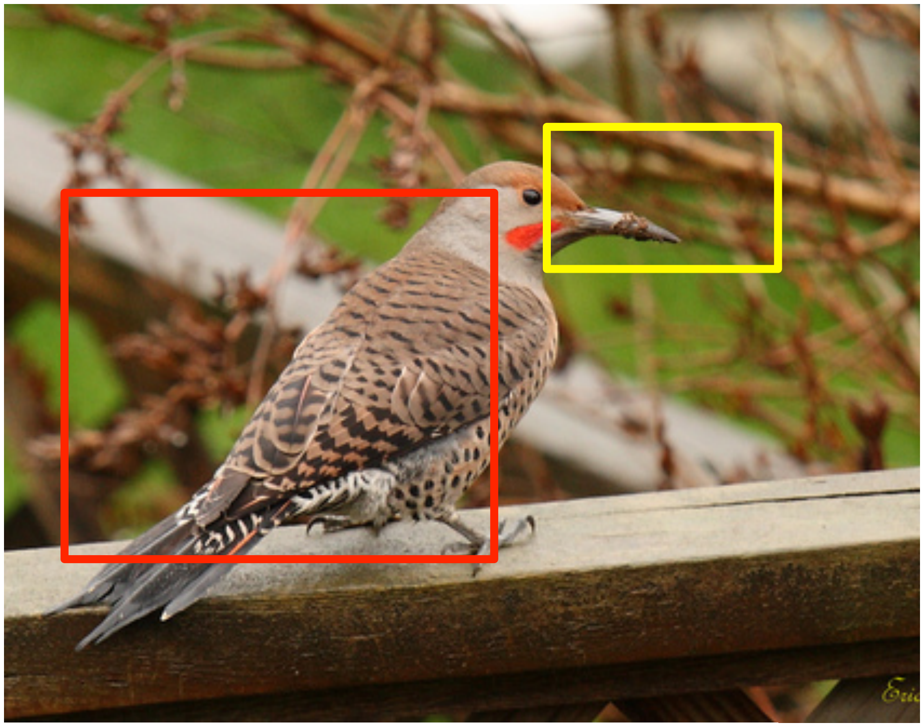
Geometric Constraints

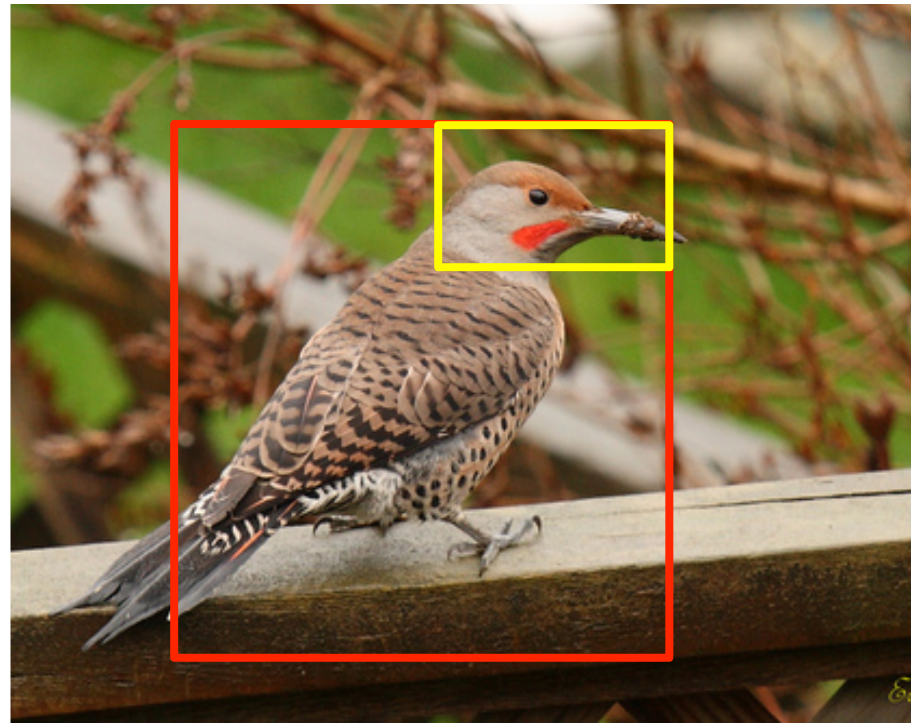
Box constraint
Gaussian Mixture
Non-parametric

$$X^* = \arg\max_X \boxed{\Delta(X)} \prod_{i=0}^{n} d_i(x_i)$$

Box constraint

head prediction bounding box prediction

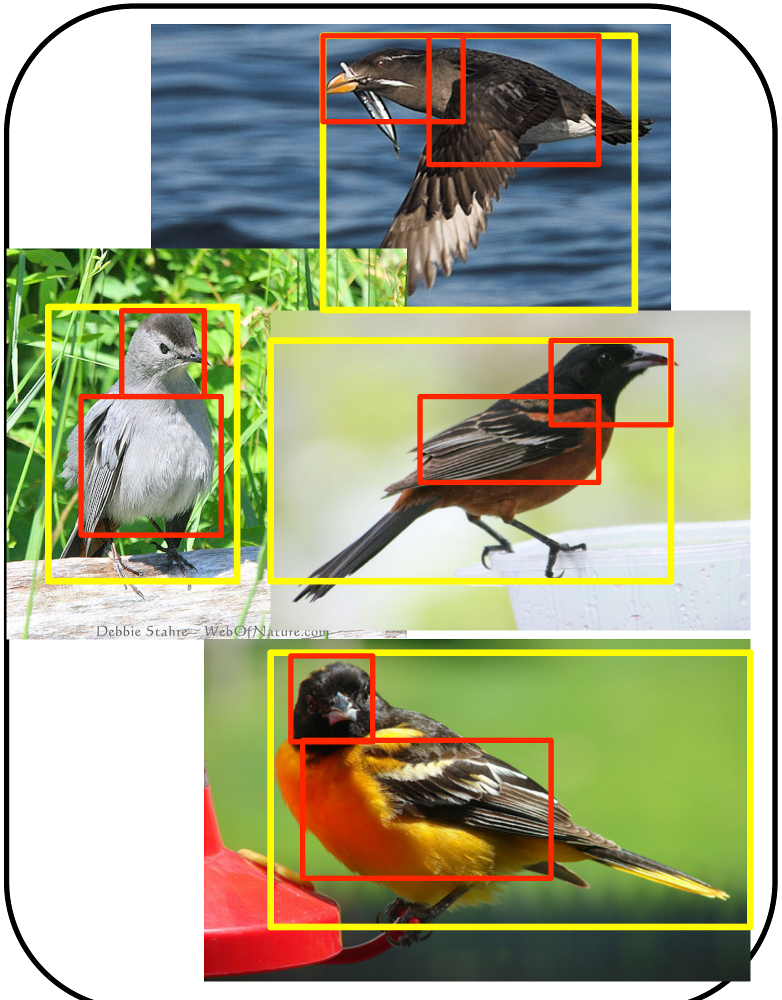




$$\Delta_{\text{box}}(X) = \prod_{i=1}^{n} c_{x_0}(x_i) \quad c_x(y) = \begin{cases} 1 \text{ if region } y \text{ falls outside region } x \\ 0 \text{ otherwise} \end{cases}$$

Geometric constraint: Gaussian Mixture

Bounding box and part annotations

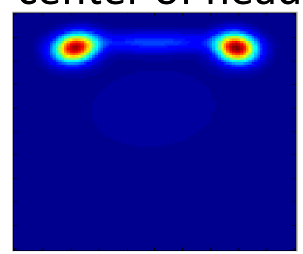


Normalize part box coordinates

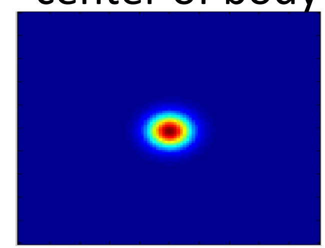
$$\begin{cases} x' = (x - x_b)/h_b \\ y' = (y - y_b)/w_b \end{cases}$$

Generate Gaussian mixture prior for each part

center of head



center of body

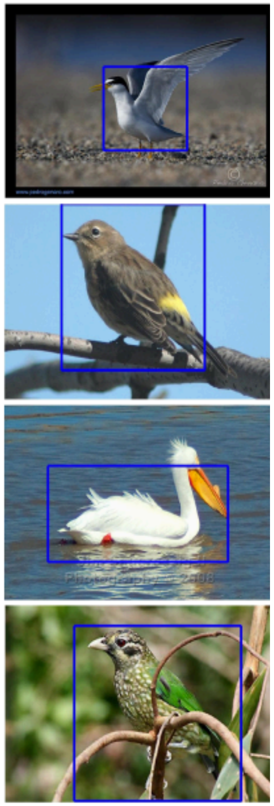


Incorporate prior into part detector scores

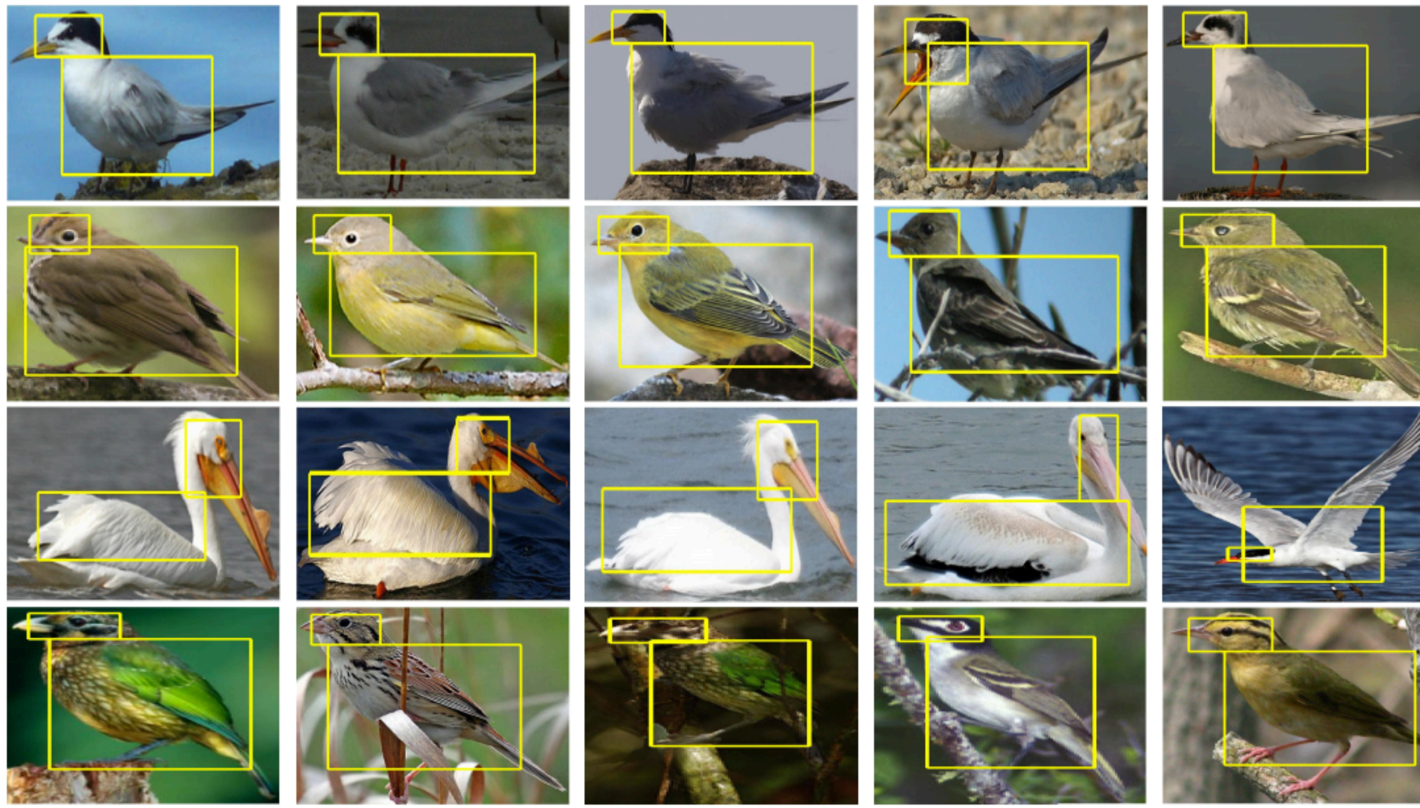
$$\Delta_{\text{geometric}}(X) = \Delta_{\text{box}}(X) \left(\prod_{i=1}^{n} \delta_i(x_i) \right)^{\alpha}$$

Geometric constraint: non-parametric

Predicted bounding box



Nearest neighbors using pool5 feature with cosine distance

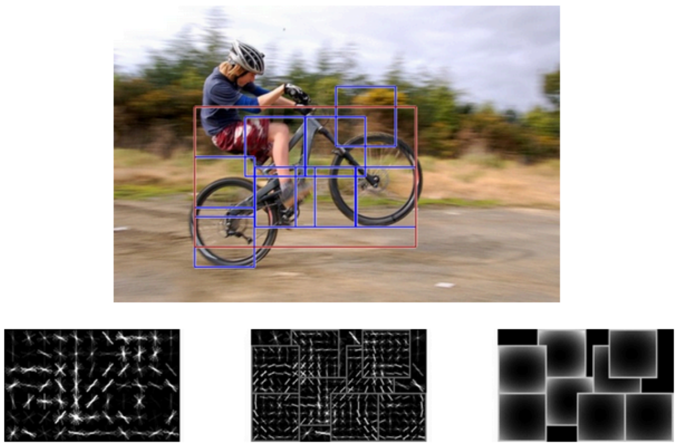


Fit one gaussian using top K neighbors

$$\Delta_{\text{geometric}}(X) = \Delta_{\text{box}}(X)\left(\prod_{i=1}^{n}\delta_i(x_i)\right)^{\alpha}$$

Comparison of constraints

Deformable part models



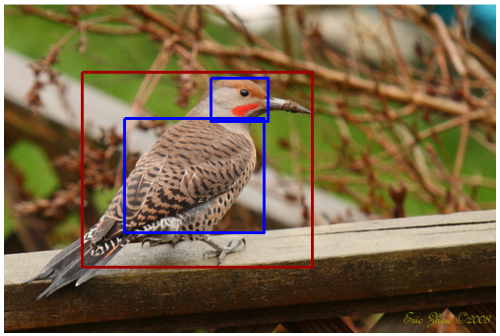
- Multiple components
- Deformation cost is a per-component Gaussian prior.
- R-CNN is a single-component model, motivating our MG and NP constraint.

Belhumeur et al. Localizing parts of faces using a consensus of exemplars. In CVPR 2011.

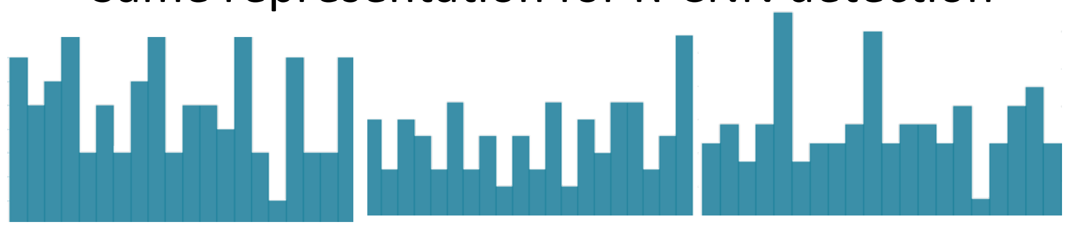
- Nonparametric prior on keypoint configuration space.
- Our non-parametric prior uses nearest neighbors on appearance space.

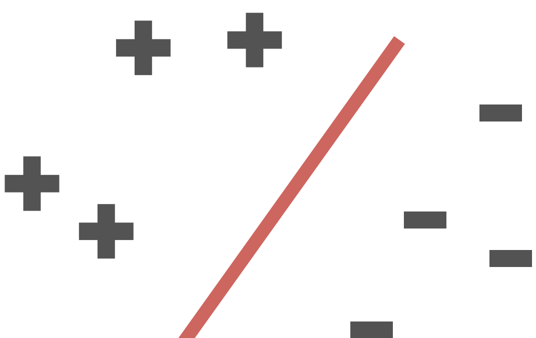
Fine-grained categorization

Bounding box and part predictions



Same representation for R-CNN detection





SVM classifier

Northern Flickr

RESULTS

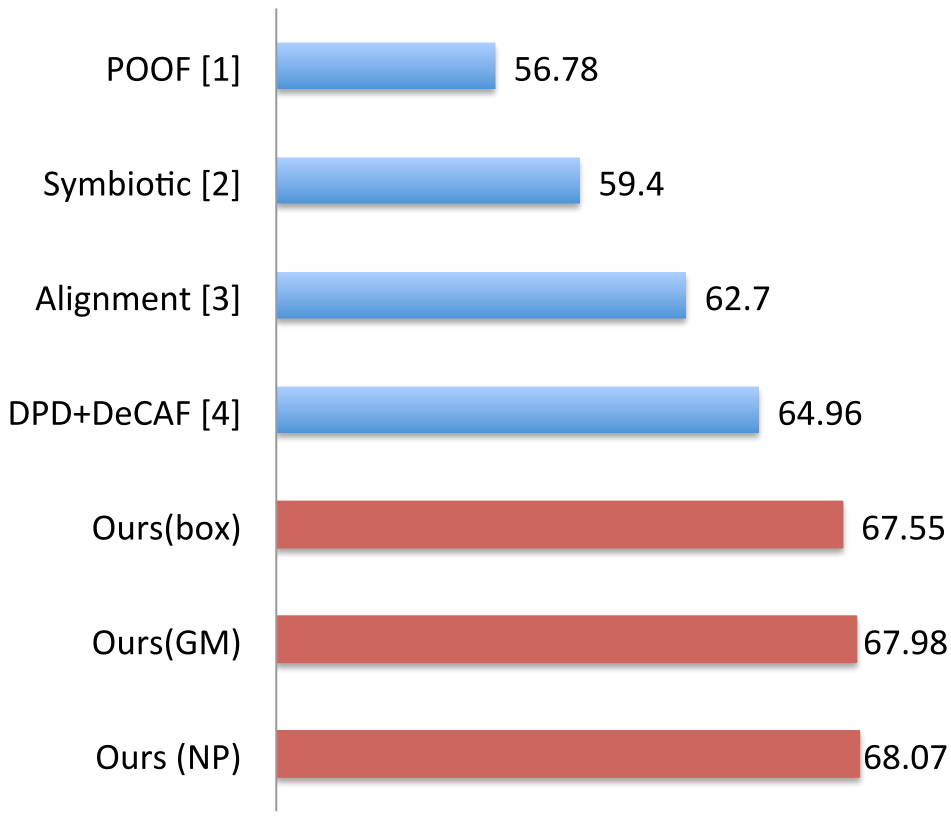
Dataset: CUB-200-2011

~12k images, 200 classes, 15 keypoints

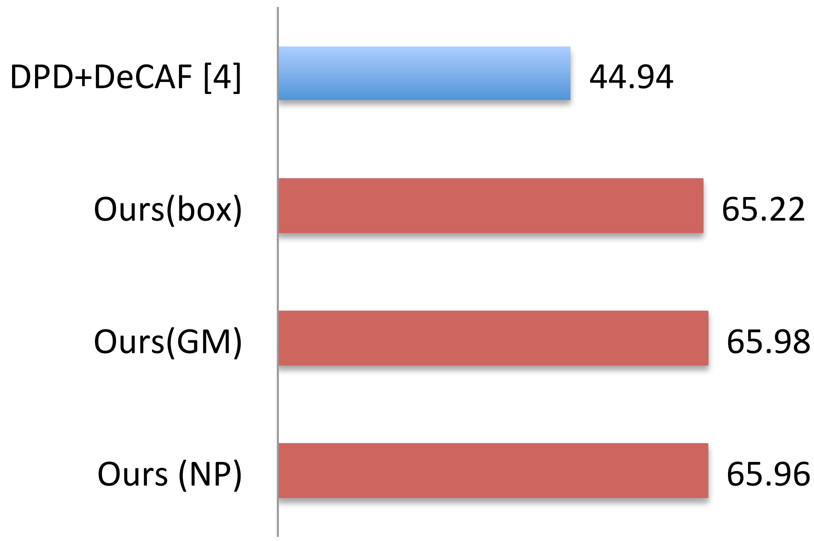
Fine-grained categorization results

Evaluation metric: classification accuracy (%)

Bounding box given

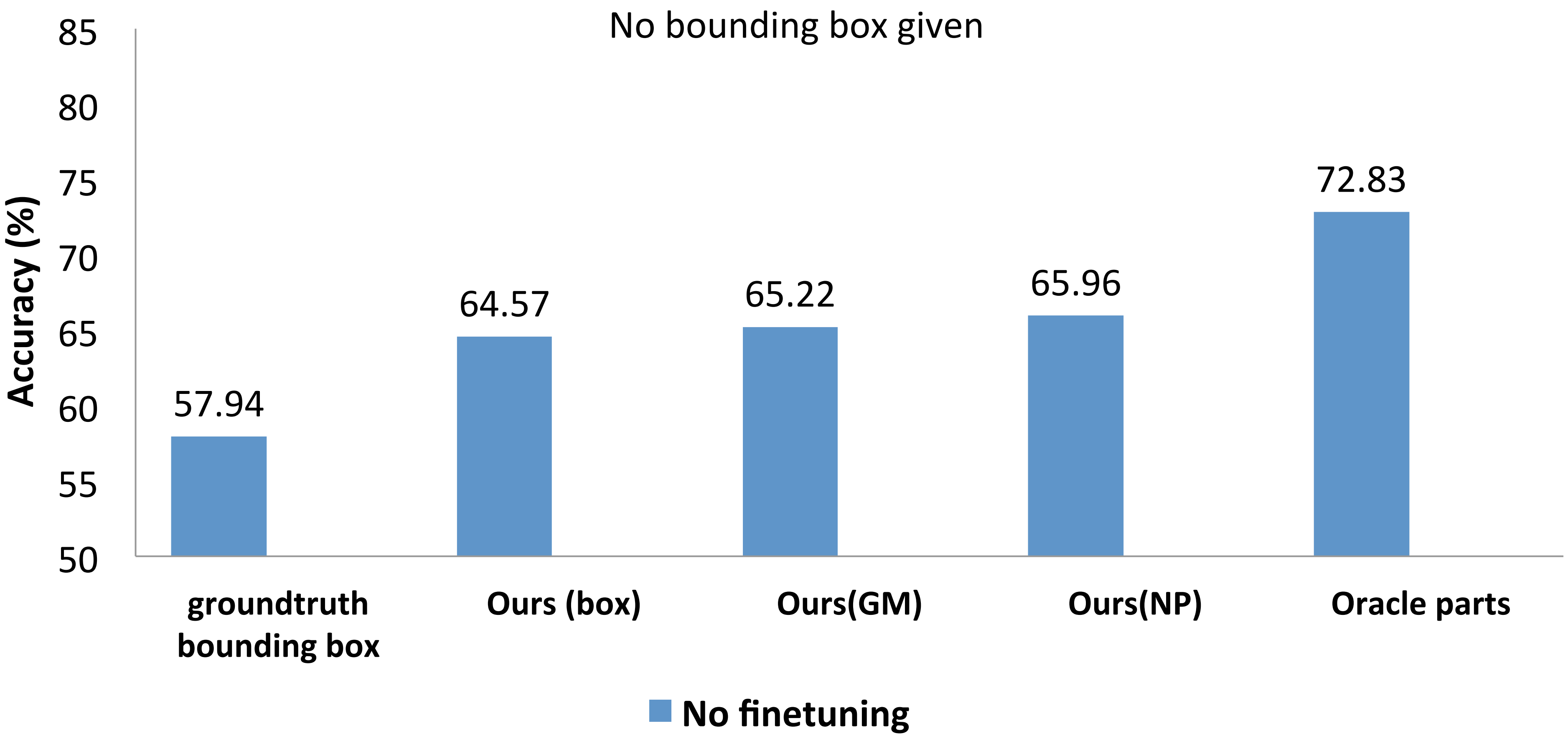


Bounding box not given

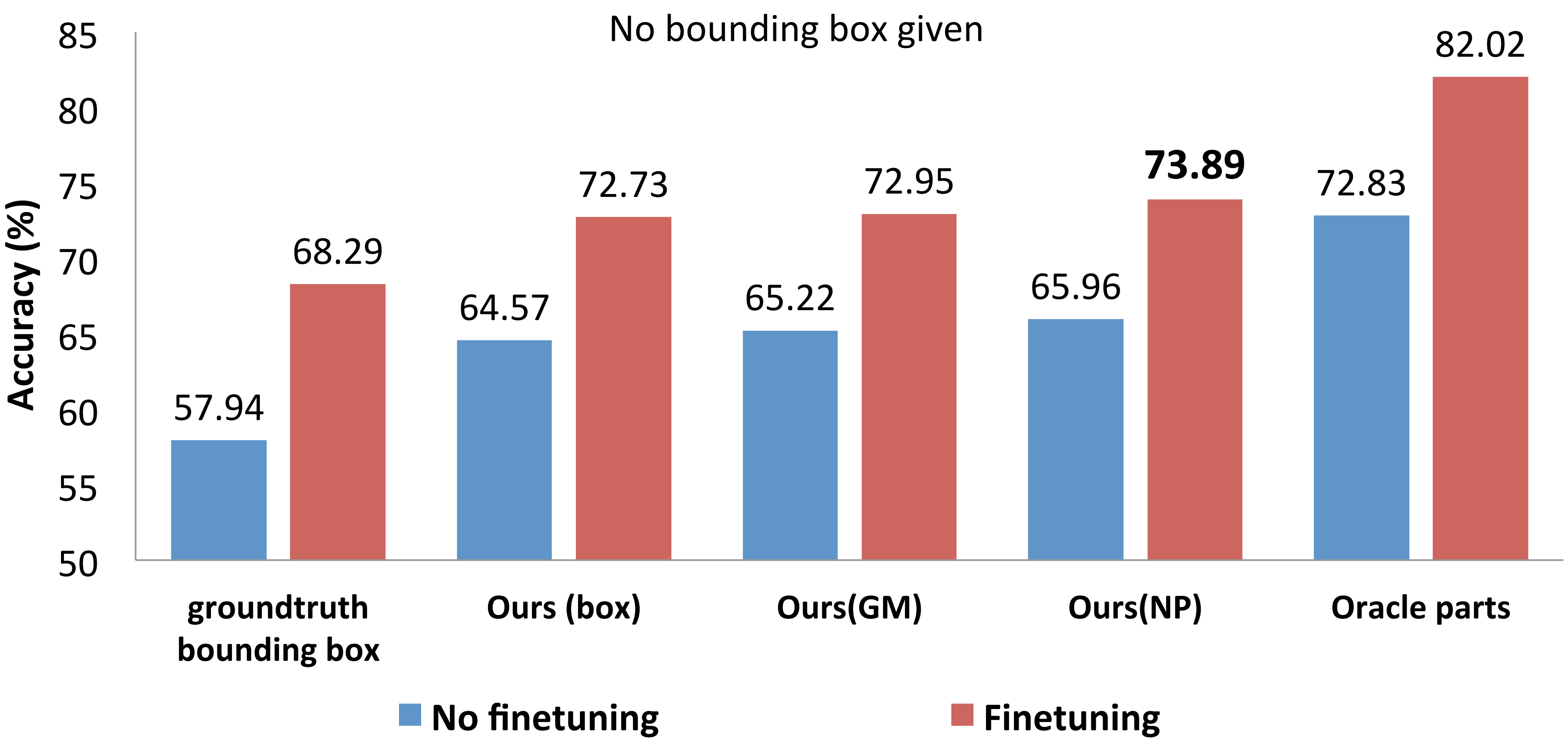


[1] Berg et.al. POOF: Part-based one-vs-one features for fine-grained categorization, face verification, and attribute estimation. In CVPR 2013.
[2] Chai et.al. Symbiotic segmentation and part localization for fine-grained categorization. In ICCV 2013.
[3] Gavves et.al. Fine-grained categorization by alignments. In ICCV 2013.
[4] Donahue et.al. DeCAF: A deep convolutional activation feature for generic visual recognition. In ICML 2014.

Does finetuning help?



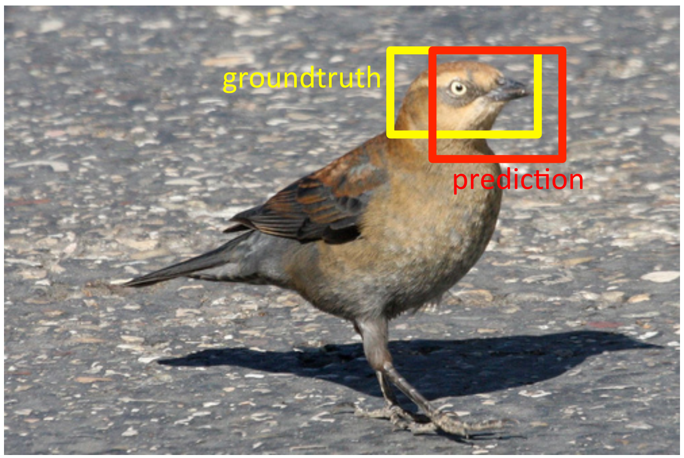
Does finetuning help?



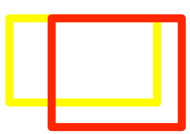
Part localization results

Evaluation metric:

Percentage of Correctly Localized Parts (PCP)



$$overlap(a,b) = \frac{a \cap b}{a \cup b}$$

if overlap of  > 0.5

part prediction is correct

Bounding Box Given		
	Head	Body
Strong DPM [1]	43.49%	75.15%
Ours (box)	61.40%	65.42%
Ours (GM)	66.03%	76.62%
Ours (NP)	**68.19%**	**79.82%**

Bounding Box Unknown		
	Head	Body
Strong DPM [1]	37.44%	47.08%
Ours (box)	60.56%	65.31%
Ours (GM)	**61.94%**	70.16%
Ours (NP)	61.42%	**70.68%**

[1] Azizipour et.al. Object detection using strongly-supervised deformable part models. In ECCV 2012.

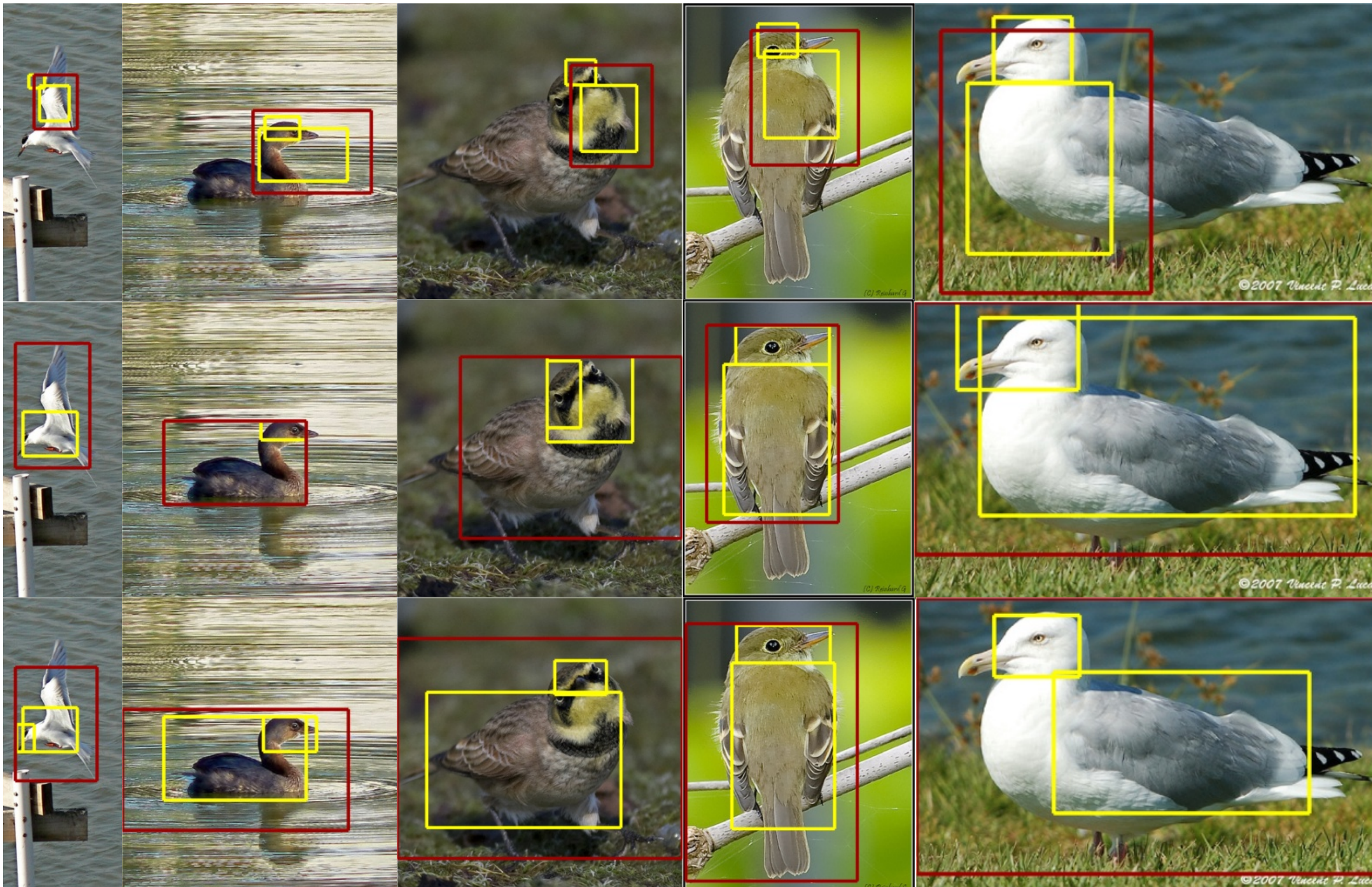
Part localization samples

part box prediction   bounding box prediction

Strong DPM

Ours (box)

Ours (NP)



Where doesn't it work?

- Limited performance of region proposal by selective search for small parts.
- Regional proposal is not designed to pick up parts.

Recall of selective search boxes on CUB200-2011 bird dataset

overlap	0.50	0.60	0.70
bounding box	96.70%	97.68%	89.50%
head	93.34%	73.87%	37.57%
body	96.70%	85.97%	54.68%

Where doesn't it work?

- Limited performance of region proposal by selective search for small parts.
- Regional proposal is not designed to pick up parts.

Recall of selective search boxes on CUB200-2011 bird dataset

overlap	0.50	0.60	0.70
bounding box	96.70%	97.68%	89.50%
head	93.34%	73.87%	37.57%
body	96.70%	85.97%	54.68%
belly	81.17%	51.82%	21.29%
leg	83.60%	51.48%	19.52%

Revisit sliding window for small parts…

Take away

- A unified deep network for both part-localization and fine-grained categorization.

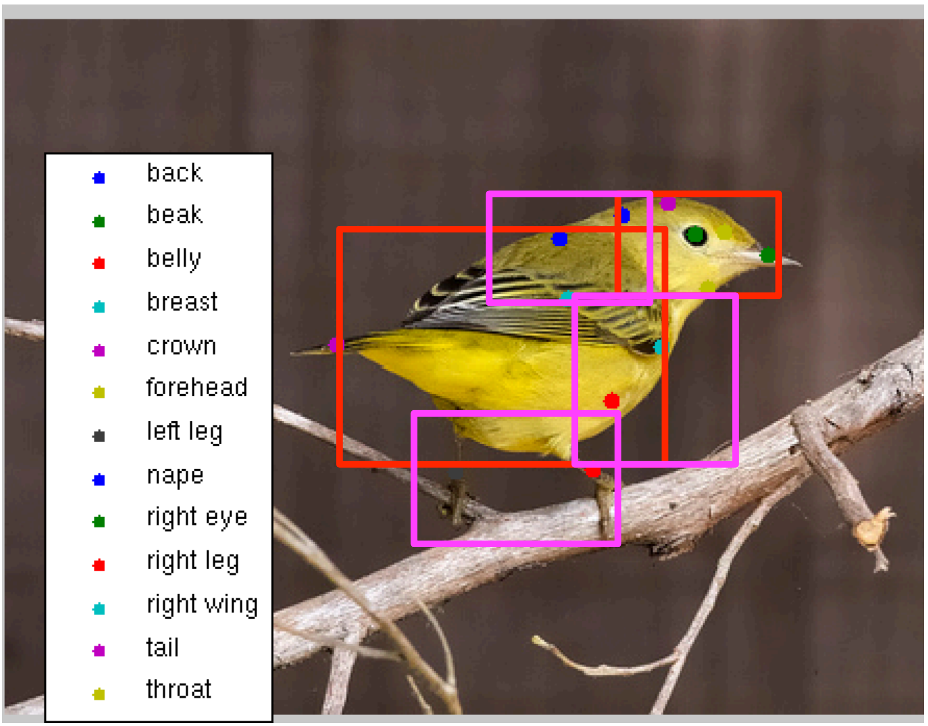
- Bounding box is not required at test time.

- Pose-normalized representation remains important for fine-grained categorization.

- R-CNN can also be used for part detections with geometric constraints.

Using more parts

Images with 5 parts annotation:
head, body, back, belly and leg

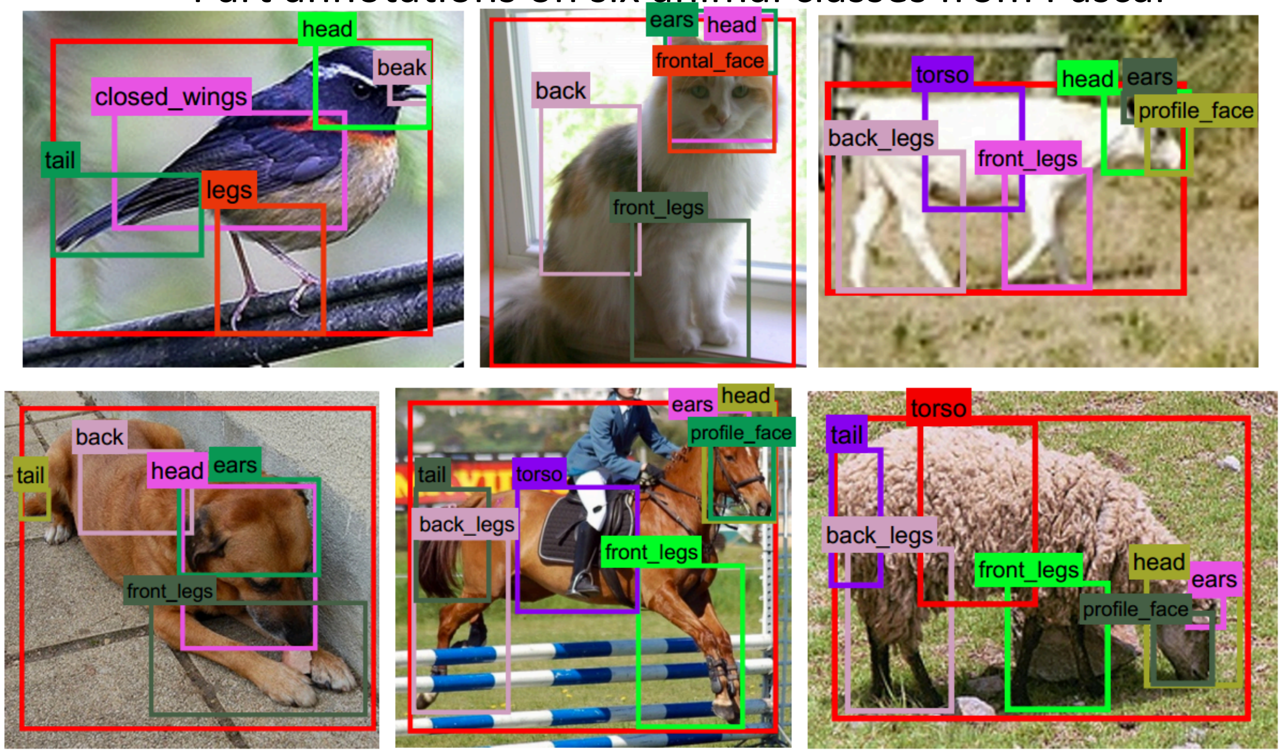


Bounding box not given at test time
without finetuning

	head+body	5 parts
Ours (box)	65.22%	62.75%
Ours(GM)	65.98%	65.43%
Ours(NP)	65.96%	65.72%

Region proposal on Pascal parts

Part annotations on six animal classes from Pascal



[Azizpour et.al. ECCV 2012]

<u>Recall on some parts from PASCAL</u>:

Cat head: 98.72 Cat back: 85.32

Dog frontal face: 95.65 Dog head: 98.98

Sheep tail: 31.25 Sheep torso: 38.24 Sheep ears: 42.54

Cow ears: 45.65 Cow head: 85.23

Bird beak: 48.41 Bird tail: 66.49

Results with no parts

Oracle (ground truth bounding box)	57.94%
Oracle-ft	68.29%
Strong DPM [3]	38.02%
R-CNN [21]	51.05%
Ours (Δ_{box})	50.17%
Ours ($\Delta_{\text{geometric}}$ with δ^{MG})	51.83%
Ours ($\Delta_{\text{geometric}}$ with δ^{NP})	52.38%
Ours-ft (Δ_{box})	62.13%
Ours-ft ($\Delta_{\text{geometric}}$ with δ^{MG})	62.06%
Ours-ft ($\Delta_{\text{geometric}}$ with δ^{NP})	**62.75%**