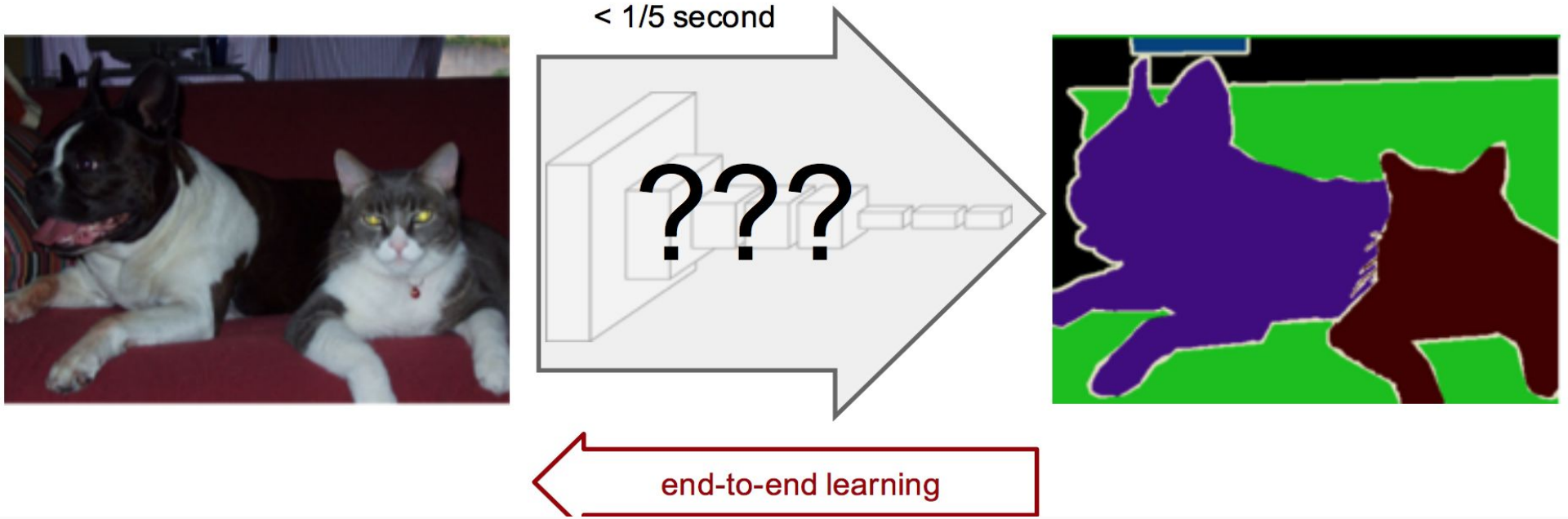
Variations on Semantic Segmentation Supervision

COS 598B, SPRING 2018 - YANNIS KARAKOZIS

Lecture Focus

- Semantic Segmentation = Instance Segmentation + Instance Classification
- Goal: Reduce training time annotation cost while maintaining high test time accuracy
- BoxSup: Bounding Boxes for CNN Supervision
- Point-level Supervision and Objectness Potential

How can we modify CNNs for semantic segmentation?



Slide Credit: Jonathan Long ft. Rohan Doshi

BoxSup: Exploiting Bounding Boxes to Supervise Convolutional Networks for Semantic Segmentation

Dai, He, Sun; ICCV’15

BoxSup at a Glance

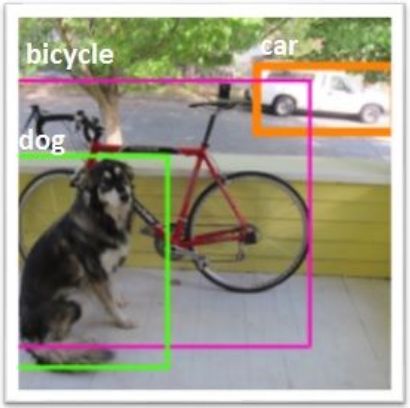
- Underlying Goal: Increase Training Set Size by Minimizing Human Input
- Intuition: Exploit Bounding Box Annotation to attain Large Scale Datasets for Instance Segmentation
- Iterative approach alternating between Region Proposal Generation and Deep CNN Training
- Datasets Used: PASCAL-VOC , Microsoft COCO and PASCAL-CONTEXT

Pixel Annotations vs Bounding Boxes

- Refined Masks
- Commercially Expensive
- Specially trained staff needed
- Not explicitly harnessed

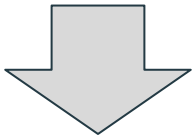
(c) Semantic segmentation

- Coarse Masks
- 15 times lighter workload
- Highly available datasets



Why Do We Care?
State of the Art Deep CNN Training

- ImageNet → largest classification dataset with quality labels
  - Special purposes datasets significant smaller



Use Transfer Learning using ImageNet for Pre-Training

Capture Generic Features

BUT: Task Specific Dataset Size still Matters

ADD MORE LAYERS

(just kidding)

The BoxSup Approach:

Region Proposal Generation

Deep CNN Training

Baseline Architecture: Mask-Supervised FCN

- Focus: Choice of appropriate objective function.

Fully Convolutional Net with CRF post-processing

- Loss function incorporates pixel-based ground-truth label

$$\mathcal{E}(\theta) = \sum_{p} e(X_\theta(p), l(p)),$$

Step 1: Initial Unsupervised Segmentation

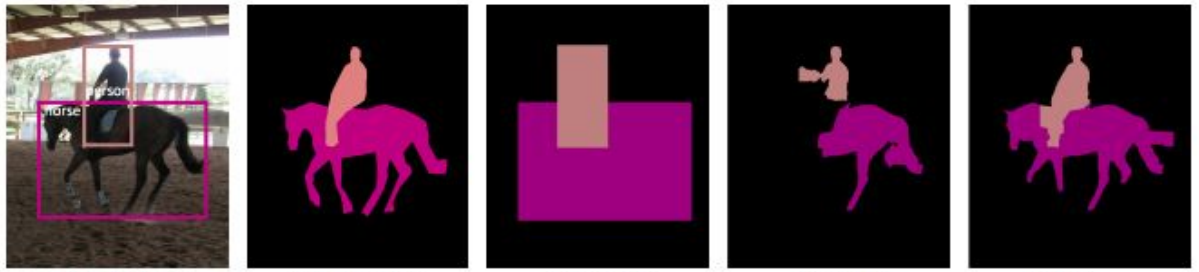
Goal: Estimate Initial Segmentation Masks from Ground Truth Bounding Boxes

Method: Unsupervised Region Proposal Methods (think GrabCut)

Produce Multiple Masks per Object

High Recall Rates of having a good candidate in the Proposal Pool

High Variance Candidates → Data Augmentation

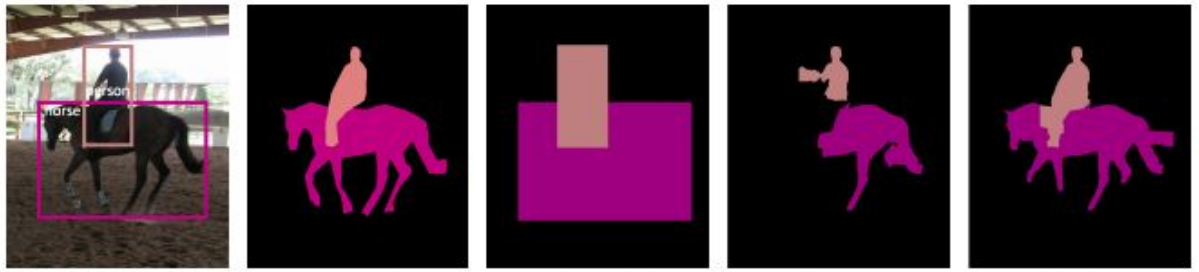


(a) training image (b) ground-truth (c) rectangles (d) GrabCut (e) ours

Step 1: Initial Unsupervised Segmentation

Proposed candidate masks pool fixed throughout training.
Only labels assigned to each candidate mask are updated.
Better masks picked by the algorithm with greater number of epochs.

Labels: Pre-learnt Semantic Category or Background



(a) training image (b) ground-truth (c) rectangles (d) GrabCut (e) ours

Step 2a: Overlapping Objective Function

Primary Goal: Label every pixel p correctly.

Goal: Pick candidate segment overlapping with the box as much as possible.

$$\mathcal{E}_o = \frac{1}{N} \sum_S (1 - \text{IoU}(B, S)) \delta(l_B, l_S).$$

Minimizing entails finding the candidate segment S that has the largest IoU with B, for each bounding box B.

Step 2b: Regression Objective Function

Primary Goal: Label every pixel p correctly.

Goal: Update Network Parameters.

$$\mathcal{E}_r = \sum_{p} e(X_\theta(p), l_S(p)).$$

Regression Target: Candidate Mask.

Minimizing entails finding optimal pixel labeling for current image.

Equivalent to FCN objective function.

Step 2c: Overarching Objective Function

Primary Goal: Label every pixel p correctly.

How: Optimize network parameters θ and candidate segment labels.

$$\min_{\theta,\{l_S\}} \sum_i (\mathcal{E}_o + \lambda \mathcal{E}_r)$$

Step 3 - Strawman: Training Algorithm

1. Generate and label initial candidate masks
2. For each semantic label, pick candidate mask minimizing objective function
3. Assign all other pixels to background
4. Update Network Parameters using one training epoch - all images are visited once
5. Update Mask labeling on all images using the updated network

$$\min_{\theta,\{l_S\}} \sum_i (\mathcal{E}_o + \lambda \mathcal{E}_r)$$

Step 3: Training Algorithm

1. Generate and label initial candidate masks
2. For each semantic label, **randomly sample candidate mask from the first k** minimizing objective function
3. Assign all other pixels to background
4. Update Network Parameters using one training epoch - all images are visited once
5. Update mask labeling on all images using the updated network

$$\min_{\theta,\{l_s\}} \sum_i (\mathcal{E}_o + \lambda \mathcal{E}_r)$$

BoxSup: Putting it all Together

1. Transfer Learning using ImageNet for Model Initialization
2. Generate and label initial candidate masks using Unsupervised Region Proposal
3. Perform Training Algorithm for a number of Epochs

Outcome: Segmentation Network ready to be applied directly on images

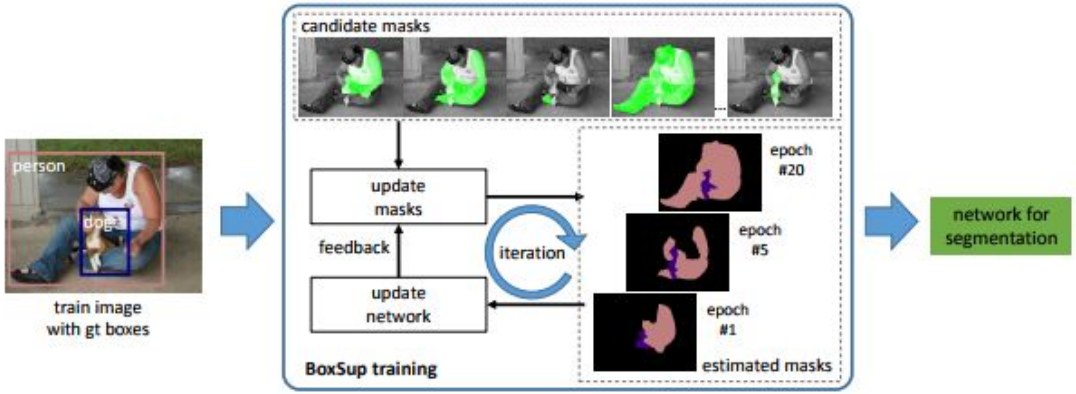


Figure 1: Overview of our training approach supervised by bounding boxes.

Performance Metrics

Comparing Supervision Strategies

data	VOC train			VOC train + COCO	
total #	10,582			133,869	
supervision	mask	box	semi	mask	semi
mask #	10,582	-	1,464	133,869	10,582
box #	-	10,582	9,118	-	123,287
mean IoU	63.8	62.0	63.5	68.1	68.2

Table 1: Comparisons of supervision in PASCAL VOC 2012 validation.

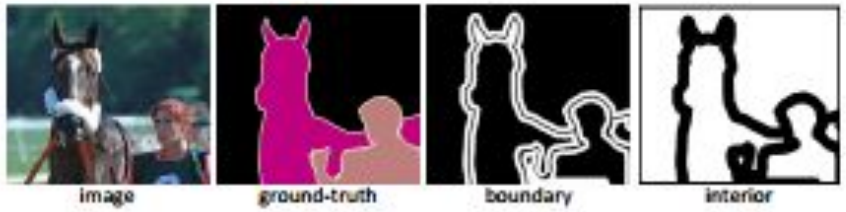
Error Analysis

Semantic Segmentation Error Types:

- Recognition Errors: Confusion in recognizing object
- Boundary Errors: Misalignment of pixel-level labels at object boundary

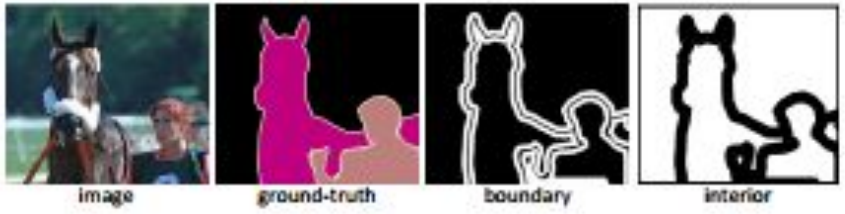
Bounding box annotations → Extra Instances for Recognizing Objects
Expectation: Recognition Error Reduction

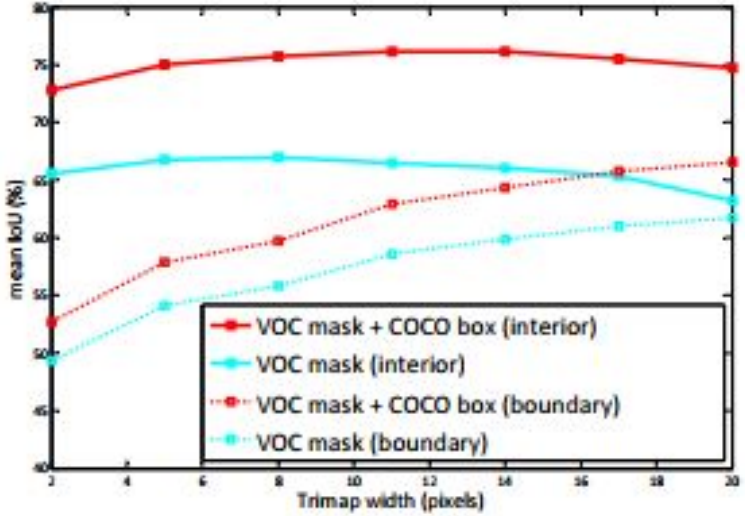
Evaluation Technique: Separate Boundary from Interior Regions of the object
Create bands of various pixel lengths around the ground truth boundaries.



Evaluation Technique: Separate Boundary from Interior Regions of the object

Create bands of various pixel lengths around the ground truth boundaries.



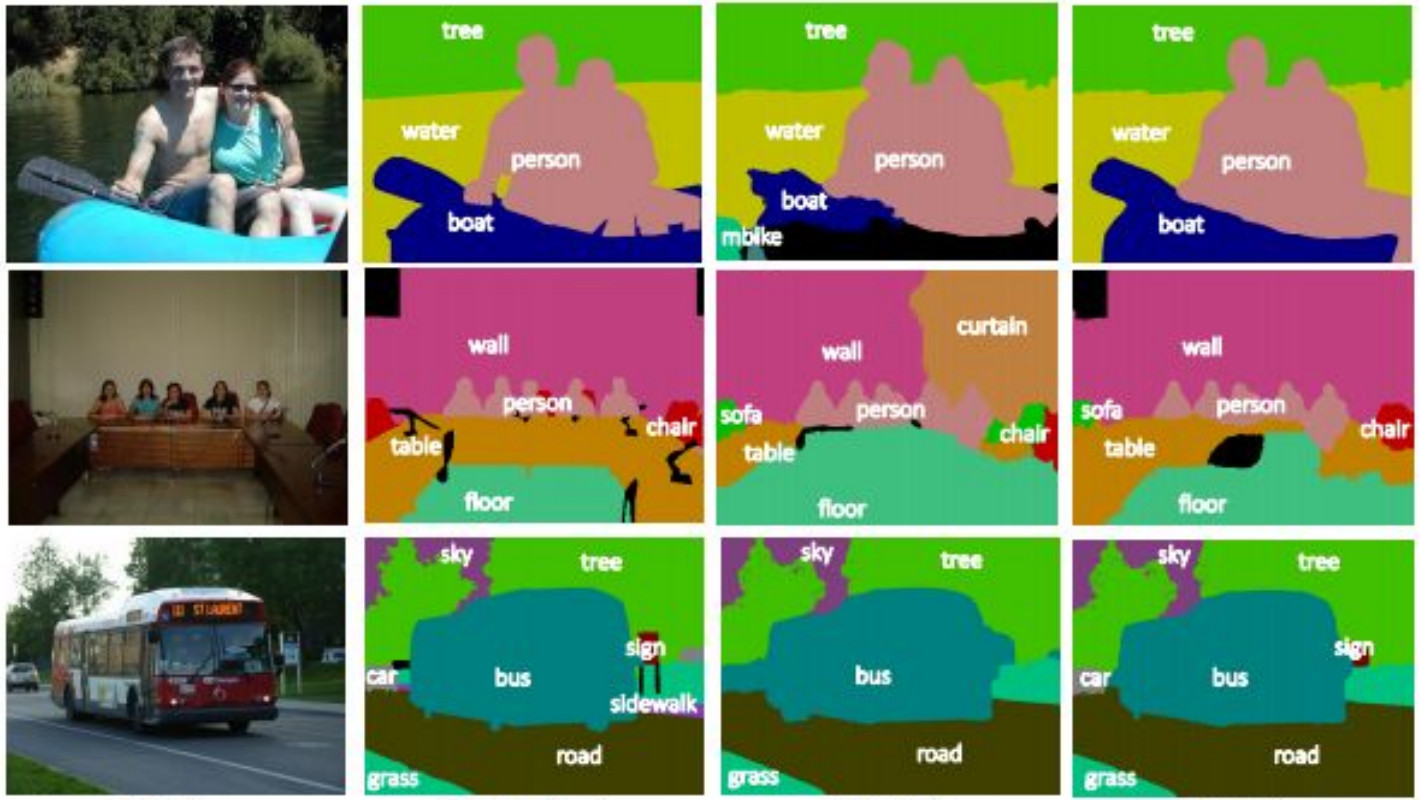


Improvement in recognition accuracy in interior regions.

Improvement in boundary regions is secondary (due to CRF post-processing).

More boxes/instances → Better Recognition

Reducing Recognition Error



(a) image (b) ground-truth (c) baseline (d) BoxSup

masks	mean IoU
rectangles	52.3
GrabCut	55.2
WSSL [25]	58.5
ours w/o sampling	59.7
ours	62.0

Estimated Masks for Supervision

Iteratively updating the masks by the networks yield higher performance.

Random sampling strategy for data augmentation and higher sample variance.

Comparison with other FCN-augmented models

method	sup.	mask #	box #	mIoU
FCN [22]	mask	V 10k	-	62.2
DeepLabCRF [5]	mask	V 10k	-	66.4
WSSL [25]	box	-	V 10k	60.4
BoxSup	box	-	V 10k	64.6
BoxSup	semi	V 1.4k	V 9k	66.2
WSSL [25]	mask	V+C 133k	-	70.4
BoxSup	semi	V 10k	C 123k	71.0
BoxSup	semi	V 10k	V_{07}+C 133k	73.1
BoxSup+	semi	V 10k	V_{07}+C 133k	**75.2**

Quick Note on BoxSup+

Test-time Scale Augmentation

- Compute pixel-wise prediction scores at three image scales: 80%, 100%, 120%
- Bilinearly rescale the score maps to the original size and average scores per pixel.

BoxSup Conclusions

Failure of recognizing objects (i.e. Recognition Error) is the main obstacle for semantic segmentation right now.

Large-scale data experimentally proven to help in this area.

Bounding box annotation-based models enable bigger data approaches to semantic segmentation, leading to comparable performance to state-of-the art models.

Novel test-time augmentation: Boxsup+

What's the Point: Semantic Segmentation with Point Supervision

Bearman, Russakovsky, Ferrari, Li;
ECCV'16

Starting Point: Weak Supervision

- Image-level labels (**presence or absence of class**)
- Cheap to obtain
- Labeling procedure:
  - For each image, label class as *present* (at least one pixel has this class) or *absent*

person
horse
background

Slide Credit: Deepak Pathak ft. Rohan Doshi


√ Person
√ Ray

Point Supervision at a Glance

- Goal: Reduce Training-time Cost without deteriorating Test-time Accuracy.
- Annotators point to an object of a particular class if one exists.
- Introducing Objectness Potential into the training loss-function → Objectness Prior to guide the training of the CNN to help separate objects from background.
- How can one make the most out of a fixed annotation budget?
- Benchmark: PASCAL-VOC 2012

Motivation: Pixel Level Annotations are Costly

Scarce

Commercially Expensive

Tens of minutes to annotate an image

Not explicitly harnessed

Weak Supervision Alternatives: Image-level Labels, Bounding Boxes, Squiggles, etc.

Unexplored: Training-time CNN Point Supervision

Point Supervision:

Image-Level Labels

+

Point Supervision per Object

+

Objectness Prior

Core Trade-off:

Training-time Cost
Vs
Test-time Accuracy

Key Implementation Points

- Supervised points only provided at training-time → no human input at test-time.
- Baseline = Typical semantic segmentation CNN network (FCN in this case):

  W x H image → W x H x N score map → W x H per-pixel predictions

- N = number of classes the CNN has been trained to recognize
- Focus: Choice of appropriate objective function.

Full Supervision Loss Function

Goal: Optimize sum of per-pixel cross-entropy errors estimated using the softmax loss. Per-pixel ground truth labels G_i known.

$$\mathcal{L}_{pix}(S, G) = -\sum_{i \in \mathcal{I}} \log(S_{iG_i})$$

Recap: Multi-class MIL Loss

$$(x_l, y_l) = \arg\max_{\forall(x,y)} \hat{p}_l(x, y) \qquad \forall l \in \mathcal{L}_I$$

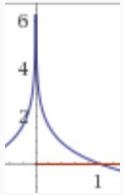
$$\implies \text{MIL LOSS} = \frac{-1}{|\mathcal{L}_I|} \sum_{l \in \mathcal{L}_I} \log \hat{p}_l(x_l, y_l)$$

- Maximize classification score based on each pixel-instance
- Takes advantage of inter-class competition to narrow down instance hypotheses

L_I: Label set of present classes

(x_l, y_l): max scoring pixel in coarse heat-maps of a class I

$\hat{p}(xl, yl)$: output heat-map for the l^{th} label at location (x, y)



y = -log(x)

Slide Credit: Rohan Doshi

Image-Level Supervision Loss Function

Goal 1: Encourage each **class present** in the image to have **high probability on at least one pixel in the image**.

Goal 2: Encourage each **class not present** in the image to have **low probability on all pixels in the image**.

Info available: which classes are present (set L) and which are not (set L')

$$\mathcal{L}_{img}(S, L, L') = -\frac{1}{|L|}\sum_{c \in L} \log(S_{t_c c}) - \frac{1}{|L'|}\sum_{c \in L'} \log(1 - S_{t_c c})$$

$$\text{with } t_c = \arg\max_{i \in \mathcal{I}} S_{ic}$$

Point-Level Supervision Loss Function

Goal 0: Optimize sum of per-pixel cross-entropy errors on supervised pixels.

Goal 1: Encourage each **class present** in the image to have **high probability on at least one pixel in the image**.

Goal 2: Encourage each **class not present** in the image to have **low probability on all pixels in the image**.

Info available: Supervised pixel classes + sets L and L’

$$\mathcal{L}_{point}(S, G, L, L') = \mathcal{L}_{img}(S, L, L') - \sum_{i \in \mathcal{I}_s} \alpha_i \log(S_{iG_i})$$

Setting a_i: Annotation Methods

1. At most 1 Point Annotation per Object Class → a_i is uniform for every point.
2. Multiple annotators performing (1) → a_i represents the confidence of the accuracy of the annotator providing that point.
3. 1 Point per Object Instance→ a_i corresponds to the order of point annotation.

$$\mathcal{L}_{point}(S, G, L, L') = \mathcal{L}_{img}(S, L, L') - \sum_{i \in \mathcal{I}_n} \alpha_i \log(S_{iG_i})$$

Introduction to Objectness Prior

- Measures likelihood an image window is part of an object.
- Computed from 35 images depicting a broad range of classes using mix of low level features (e.g.edges, corners, loops etc)
- Pixel score: Average of scores of all windows it containing it.
- Precomputed $\longrightarrow$ 0.28 seconds of extra annotation per second
- Function: Better defines spatial extent of recognized objects

Objectness Prior Loss Function

Goal: Infer the spatial extent of the object.

Objectness P_i: Probability pixel i belongs to a foreground object class.

$$\mathcal{L}_{obj}(S,P) = -\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} P_i \log\left(\sum_{c\in\mathcal{O}} S_{ic}\right) + (1-P_i)\log\left(1-\sum_{c\in\mathcal{O}} S_{ic}\right)$$

Pixels with high Pi values → Place probability mass on object classes.

Pixels with low Pi values → Place probability mass on background class.

L_obj requires no human supervision → Can be combined with any loss.

Objective Function

Goal: Avoid local minima that over/under-define spatial extent of class instances

$$\mathcal{L}_{obj}(S,P) = -\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} P_i \log\left(\sum_{c\in\mathcal{O}} S_{ic}\right) + (1-P_i)\log\left(1-\sum_{c\in\mathcal{O}} S_{ic}\right)$$

combined with

$$\mathcal{L}_{point}(S,G,L,L') = \mathcal{L}_{img}(S,L,L') - \sum_{i\in\mathcal{I}_x} \alpha_i \log(S_{iG_i})$$

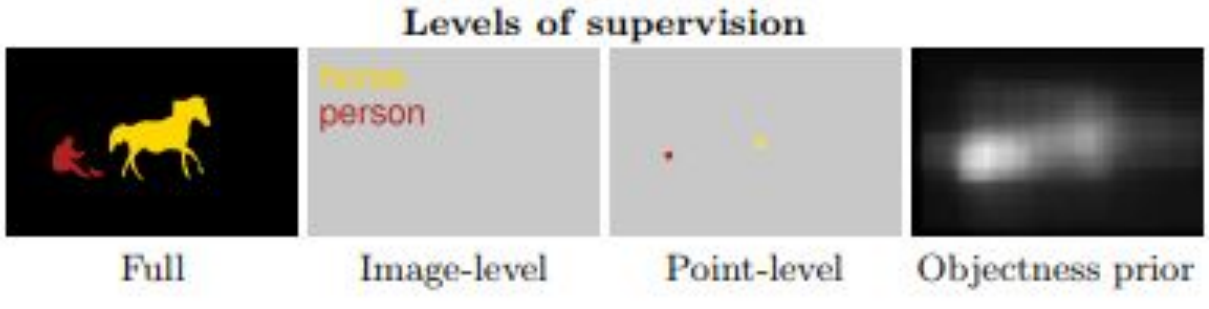
Objective Function Minimization Achieves

Goal 0: Optimize sum of per-pixel cross-entropy errors on supervised pixels.

Goal 1: Encourage each **class present** in the image to have **high probability on at least one pixel in the image**.

Goal 2: Encourage each **class not present** in the image to have **low probability on all pixels in the image**.

Goal 3: Avoid local minima that over/under-define spatial extent of class instances



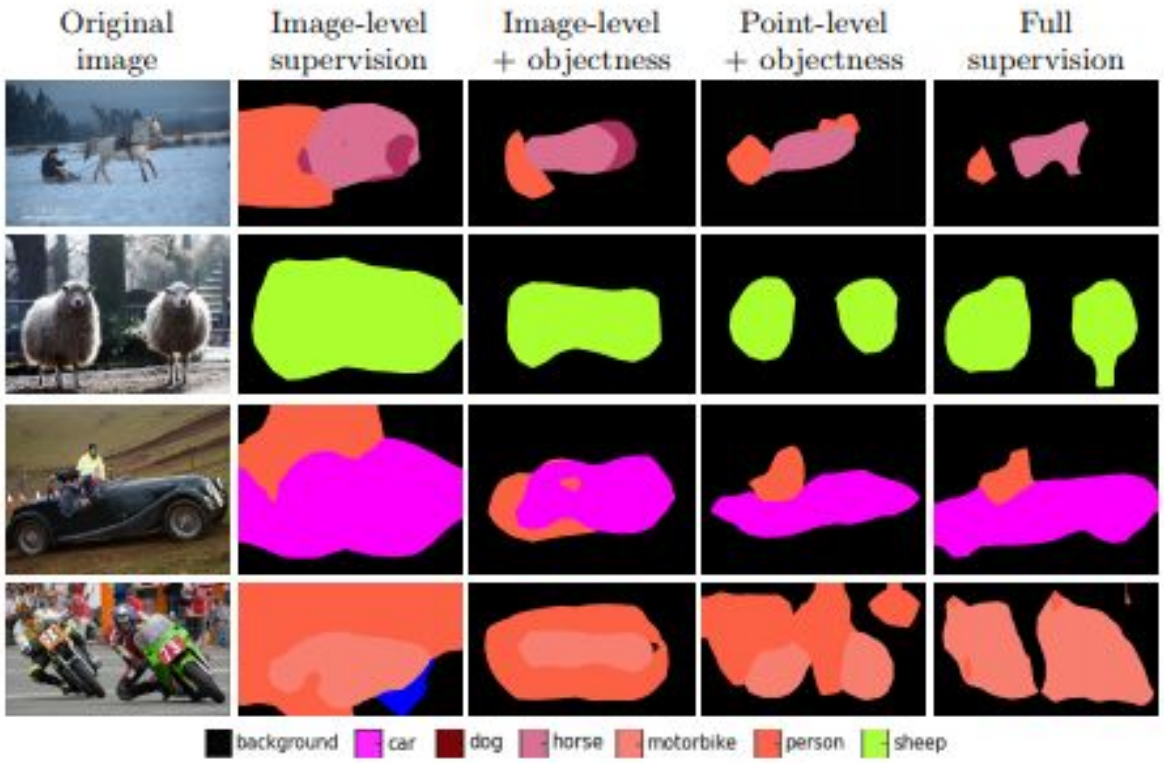
Crowdsourced Annotation

- Crowdsourced annotations via Amazon Mechanical Turk
- Dataset: PASCAL VOC 2012 (20 object classes)
- 1Point vs AllPoint annotations; Squiggle-Level Annotations reproduced
- Quality Control via evaluation images incorporated into the dataset.
- Much smaller annotation errors compared to bounding box annotations

Supervision Category		Average Annotation Time per Image
	Image-Level	20 seconds
	Full (Pixel-Level)	239.7 seconds
Point-Level	1 point per object class	22.1 seconds
	1 point per object instance	23.3 seconds
	Squiggle-Level	34.9 seconds

Baseline: FCN with Image-Level Supervision

Qualitative Evaluation



Quantitative Evaluation (1)

Key Metric: Mean IOU averaged over the 21 PASCAL VOC classes.

Baseline image-level supervision with no additional info: 25.1%

Supervision	Time (s)	Model	mIOU (%)
Image-level labels	20.0	*Img*	29.8
Image-level labels	20.3	*Img + Obj*	32.2
1*Point*	22.1	*Img*	35.1
1*Point*	22.4	*Img + Obj*	42.7
AllPoints	23.6	*Img + Obj*	42.7
AllPoints (weighted)	23.5	*Img + Obj*	43.4
1*Point* (3 annotators)	29.6	*Img + Obj*	43.8
1*Point* (random annotators)	22.4	*Img + Obj*	42.8 - 43.8
1*Point* (random points)	240	*Img + Obj*	46.1
Full supervision	239.7	*Img*	58.3
Hybrid approach	24.5	*Img + Obj*	53.1
1 squiggle per class	35.2	*Img + Obj*	49.1

Quantitative Evaluation (2)

Point-Level Supervision Variations

Supervision	Time (s)	Model	mIOU (%)
Image-level labels	20.0	*Img*	29.8
Image-level labels	20.3	*Img* + *Obj*	32.2
1*Point*	22.1	*Img*	35.1
1*Point*	22.4	*Img* + *Obj*	42.7
AllPoints	23.6	*Img* + *Obj*	42.7
AllPoints (weighted)	23.5	*Img* + *Obj*	43.4
1*Point* (3 annotators)	29.6	*Img* + *Obj*	43.8
1*Point* (random annotators)	22.4	*Img* + *Obj*	42.8 - 43.8
1*Point* (random points)	240	*Img* + *Obj*	46.1
Full supervision	239.7	*Img*	58.3
Hybrid approach	24.5	*Img* + *Obj*	53.1
1 squiggle per class	35.2	*Img* + *Obj*	49.1

Quantitative Evaluation (3)

Training-time Cost vs Test-time Accuracy Trade-off

Decreasing returns of full supervision with respect to cost

Supervision	Time (s)	Model	mIOU (%)
Image-level labels	20.0	*Img*	29.8
Image-level labels	20.3	*Img + Obj*	32.2
1*Point*	22.1	*Img*	35.1
1*Point*	22.4	*Img + Obj*	42.7
AllPoints	23.6	*Img + Obj*	42.7
AllPoints (weighted)	23.5	*Img + Obj*	43.4
1*Point* (3 annotators)	29.6	*Img + Obj*	43.8
1*Point* (random annotators)	22.4	*Img + Obj*	42.8 - 43.8
1*Point* (random points)	240	*Img + Obj*	46.1
Full supervision	239.7	*Img*	58.3
Hybrid approach	24.5	*Img + Obj*	53.1
1 squiggle per class	35.2	*Img + Obj*	49.1

Quantitative Evaluation (4)

Training-time Cost vs Test-time Accuracy Trade-off

Supervision	mIOU (%)
Full (883 imgs)	22.1
Image-level (10,582 imgs)	29.8
Squiggle-level (6,064 imgs)	40.2
Point-level (9,576 imgs)	**42.9**

Table 3. Accuracy of models on the PASCAL VOC 2012 validation set given a fixed budget (and number of images annotated within that budget). Point-level supervision provides the best tradeoff between annotation time and accuracy. Details in Section 5.5.

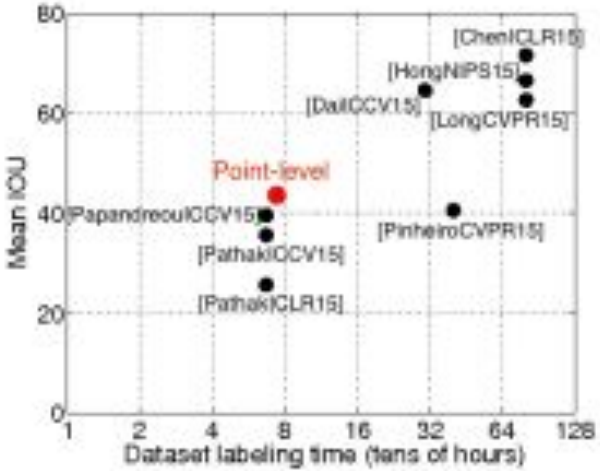


Fig. 5. Results without resource constraints on the PASCAL VOC 2012 *test* set. The x-axis is log-scale.

Point-Level Supervision Summary

- Point-Level Supervision to reinforce Image-Level Supervision.
- Point-Level Supervision directly incorporated in objective function for CNN training.
- Objectness Prior helps infer about spatial extent of the object.
- 1Point+Obj+Img achieves best performance under fixed annotation time budget.
- Hybrid approach achieves best accuracy-cost tradeoff.

Presenter's Note

- Experimental rather than rigorous mathematical proofs
- It is all about identifying the optimal objective function or the one that makes one's hypothesis work
- Why BoxSup+ perform better? → Intuition driven field based on past experience
- Decreasing marginal returns of extra sample (is higher efficiency really necessary once you have a huge dataset of pixel-based annotated examples to train on)
- Segmentation accuracies are still very low