Knowledge transfer and human-machine collaboration for object detection and segmentation

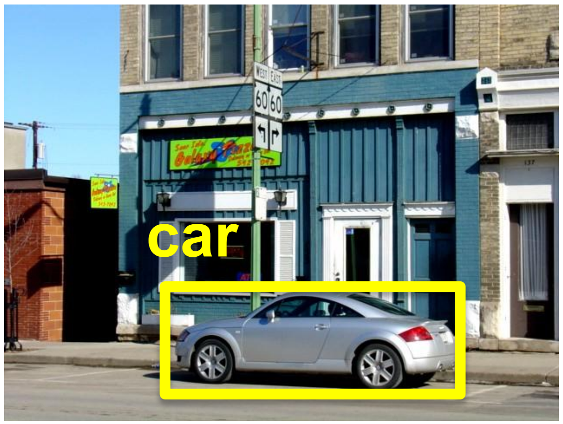
June 2018

Vittorio Ferrari
Google Research

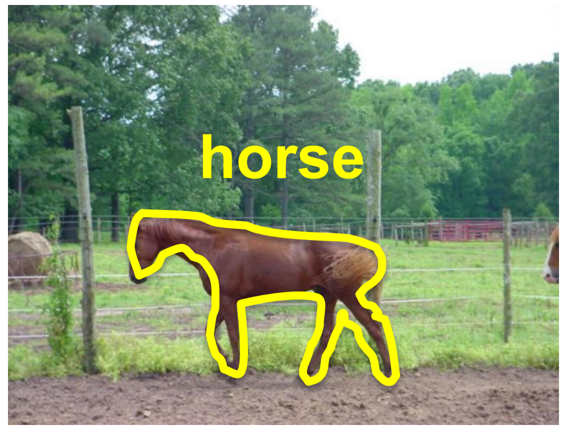
CVPR 2018 WebVision workshop

Manual annotation is expensive

Training modern models requires:



26s
(ImageNet)



80s
(COCO)

1000s
(COCO-Stuff)

Su et al., Crowdsourcing annotations for visual object detection, AAAI 2012

Deng et al., Scalable multi-label annotation, CHI 2014

Lin et al., Microsoft COCO: common objects in context, ECCV 2014

Caesar et al., COCO-Stuff: Things and Stuff classes in context, CVPR 2018

Papadopoulos et al., Extreme Clicking for efficient object annotation, ICCV 2017: *box in 7s!*

Fully supervised learning

motorbike

...

Object detection model

...

Panoptic segmentation model

Annotation to the ***same*** *degree as outputs on test images*

Weakly supervised learning

…

Object detection model

motorbike

…

Panoptic segmentation model

person, book, wall, light, …

sky, horse, train, …

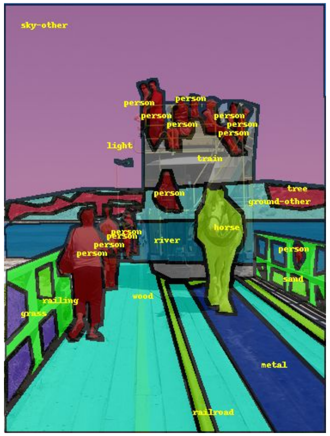
*Annotation to a **lower** degree than outputs on test images*

Weakly supervised learning

...

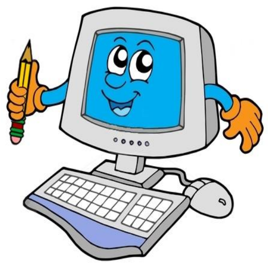
Object detection model

motorbike



...

Panoptic segmentation model



Core issue: need for auto-annotation

Human-machine collaboration

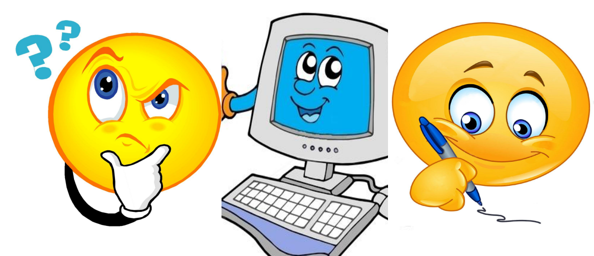
...

Object detection model

motorbike

...

Panoptic segmentation model



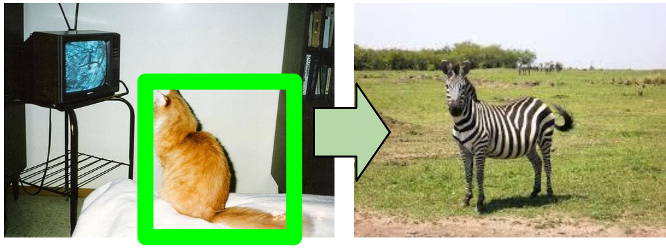
Human intervenes during machine process

Weak Supervision for bounding-boxes

- basic: image-level labels
  [Nguyen ICCV 09, Deselaers ECCV 10, Siva ICCV 11, Song ICML 14, Cinbis CVPR 14, Wang TIP 15, Bilen CVPR 16, Kantorov ECCV 16, Dong ACMMM 17]

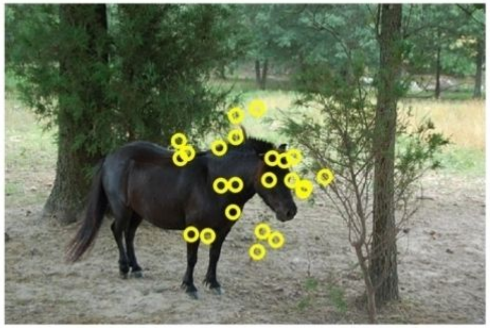
\+ point click on object
  [Mettes ECCV 16, Papadopoulos CVPR 17]

\+ video
  [Prest CVPR 12, Tang CVPR 13, Joulin ECCV 14, Kuznetsova CVPR 15, Liang ICCV 15, Liang ICCV 15, Kalogeiton PAMI 15]



\+ knowledge transfer from classes with boxes
  [Salakhutdinov CVPR 11, Aytar ICCV 11, Guillaumin CVPR 12, Vezhnevets CVPR 14, Hoffman NIPS 14, Rochan CVPR 15, Tang CVPR 16, Redmon CVPR 17]

\+ eye-tracks
  [Papadopulous ECCV 14, Mathe arXiv 14, Karthikeyan CVPR 15]

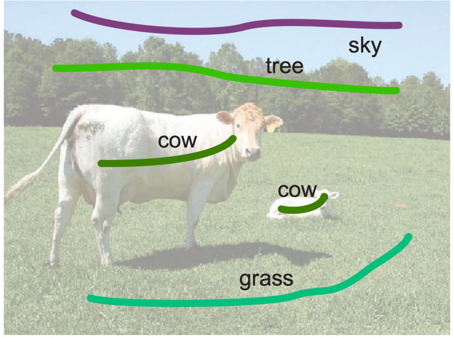


Weak Supervision for object segments

- basic: image-level labels
  [Verbeek CVPR 07, Vezhnevets ICCV 11, Xu CVPR 14, Pinheiro CVPR15, Pathak ICLR 2015, Papandreou ICCV 15, Kolesnikov ECCV 16, Wei CVPR 2017, Singh ICCV 2017]

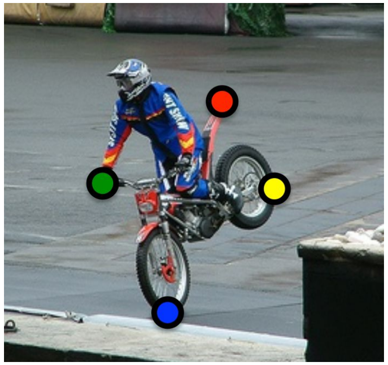
\+ point click on object
  [Wang CVIU 14, Bell CVPR 15, Bearman ECCV 16, Jain AAAI 16]

\+ scribbles
  [Xu CVPR 15, Lin CVPR 16]



\+ object bounding-boxes
  [Dai ICCV 15, Khoreva CVPR 17]

\+ extreme points on object
  [Papadopoulos ICCV 17, Maninis CVPR 18]



\+ transfer from other pre-segmented classes
  [Kuettel ECCV 12, Rubinstein ECCV 12]

\+ video
  [Tokmakov ECCV 16]

Human-Machine collaboration

- classic active learning (ask label of samples)

  for box: [Vijayanarasimhan IJCV14, Yao CVPR 12]
  for segmentations [Vezhnevets CVPR 12, Jain CVPR 16]

- box verification series
  [Papadopoulos CVPR 16]

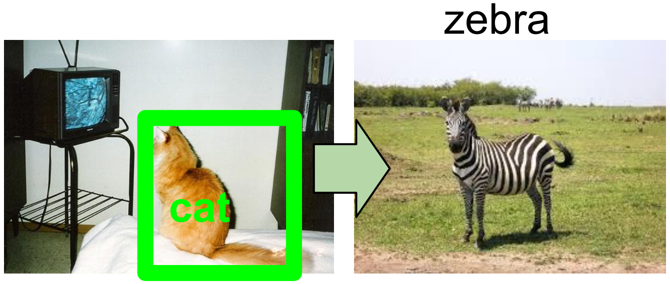
- select which annotation micro-task to ask for
  [Vijayanarasimhan CVPR 09, Jain ICCV 13,
  Russakovsky CVPR 15]

- interactive object segmentation
  [Boykov ICCV 01, Rother SIGGRAPH 04, Wang ICCV 05,
  Liew ICCV 17, Xu BMVC 17, Castrejon CVPR 17, Li CVPR 18]

- fine-grained classification by asking attributes
  [Branson ECCV 10, Biswas CVPR 13, Wah CVPR 14]

This talk

- **Revisiting knowledge transfer for training object class detectors**
  Uijlings, Popov, Ferrari
  CVPR 2018



- Learning intelligent dialogs for bounding box annotation
  Konyushkova, Uijlings, Lampert, Ferrari
  CVPR 2018

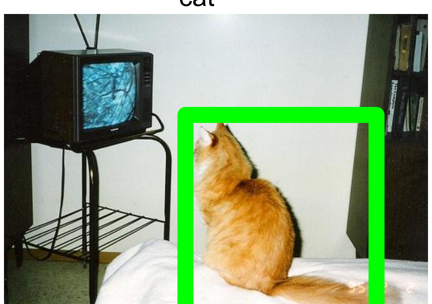
- Fluid Annotation: human-machine collaboration for full image annotation
  Andriluka, Uijlings, Ferrari,
  arXiv June 2018

Proble Problem setting: knowledge transfer on

Source train set (bounding boxes)

bicycle

cat

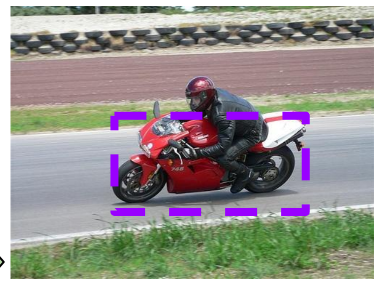


knowledge transfer

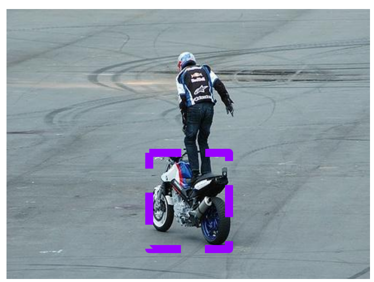
Target train set (image-level labels)

Intermediate goal: Localize target class in train set

motorbike



motorbike



no motorbike

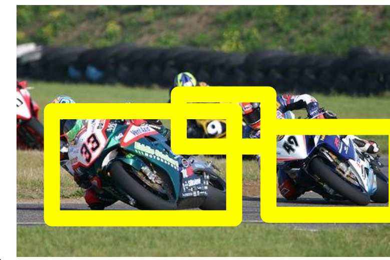
Train

Detection model

Apply

Target test set (no labels)

Goal: detect target class in test set



[Guillaumin CVPR 12, Hoffman NIPS 14, Rochan CVPR 15, Tang CVPR 16, Redmon CVPR 17]

Why relevant?

- Image-level labels are cheaper to obtain (2s)

- But weakly supervised methods lead to lower quality detectors (~60% the mAP of full supervision)

- There are large datasets with bounding boxes

IMAGENET

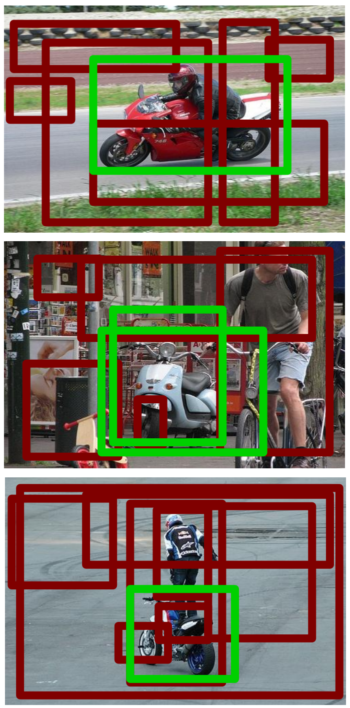
COCO
Common Objects in Context

OPEN IMAGES

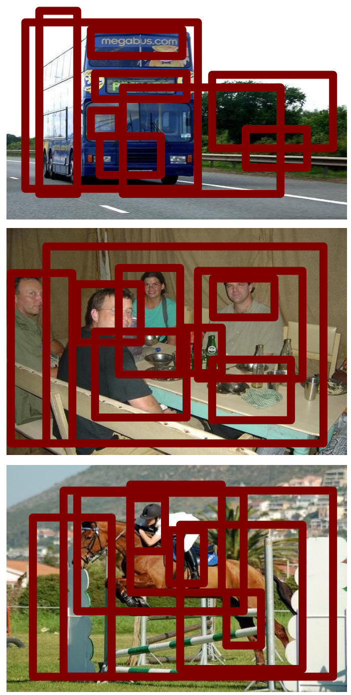
[Deselaers ECCV 10, Bilen CVPR 15, Blaschko NIPS 10, Cinbis CVPR 14, Nguyen ICCV 09, Pandey ICCV 11, Russakovsky ECCV 12, Shi PAMI15, Shi BMVC 12, Siva CVPR 13, Siva ICCV 11, Siva ECCV 12, Song NIPS 14, Song ICML 14, Wang TIP 15, Bilen CVPR 16, Dong ACMMM 17]

A typical WSOL framework

Positive images



Negative images



Multiple instance learning (MIL)

images = bags
windows = instances

Object proposals

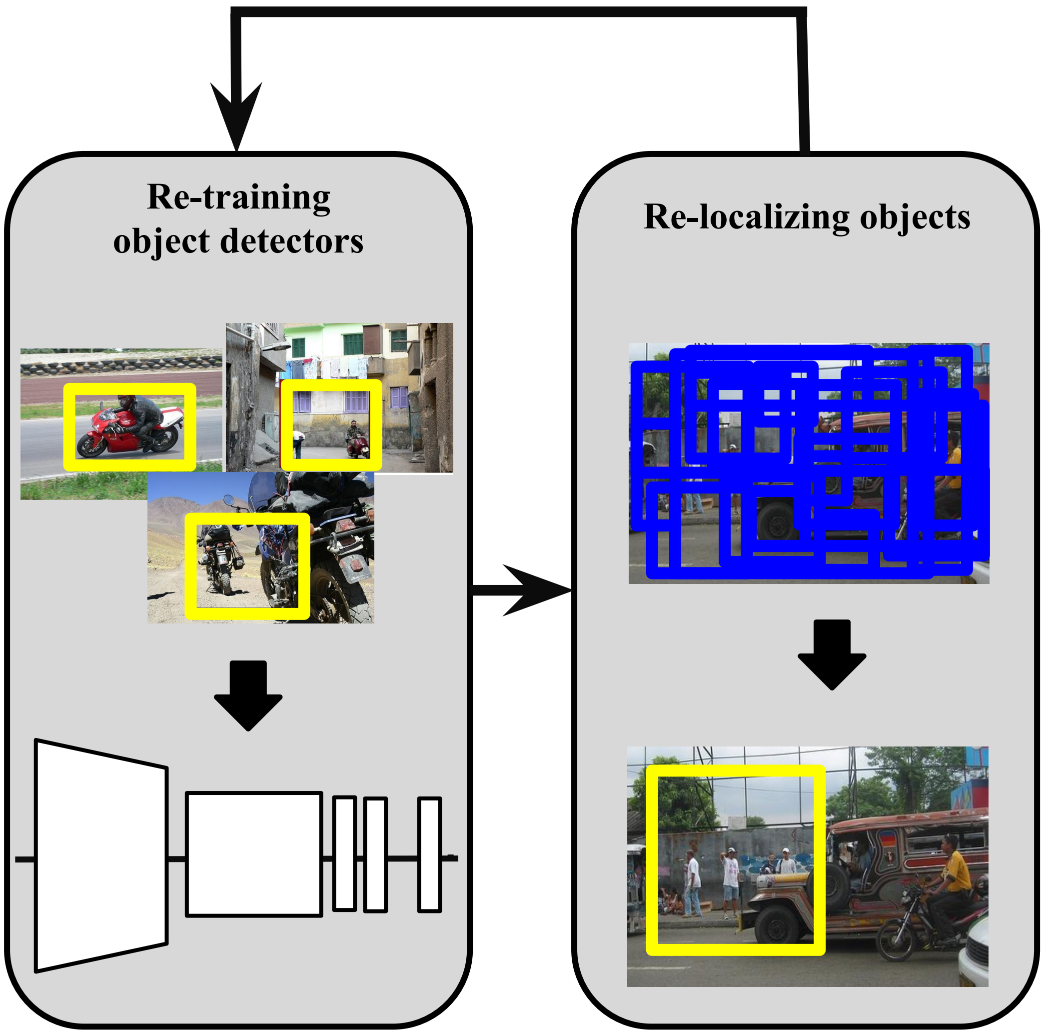
[Alexe CVPR 10, Dollar ECCV 14, van de Sande ICCV 11, …]

Goals:

- find true positive instances
- train window classifier

[Blaschko NIPS 10, Cinbis PAMI 16, Deselaers ECCV 10, Nguyen ICCV 09, Russakovsky ECCV 12, Shi PAMI 15, Shi BMVC 12, Siva ICCV 11 & ECCV 12, Song NIPS 14 & ICML 14, Bilen BMVC 14, Sangineto PAMI 18, Li BMVC 17]

MIL-style WSOL



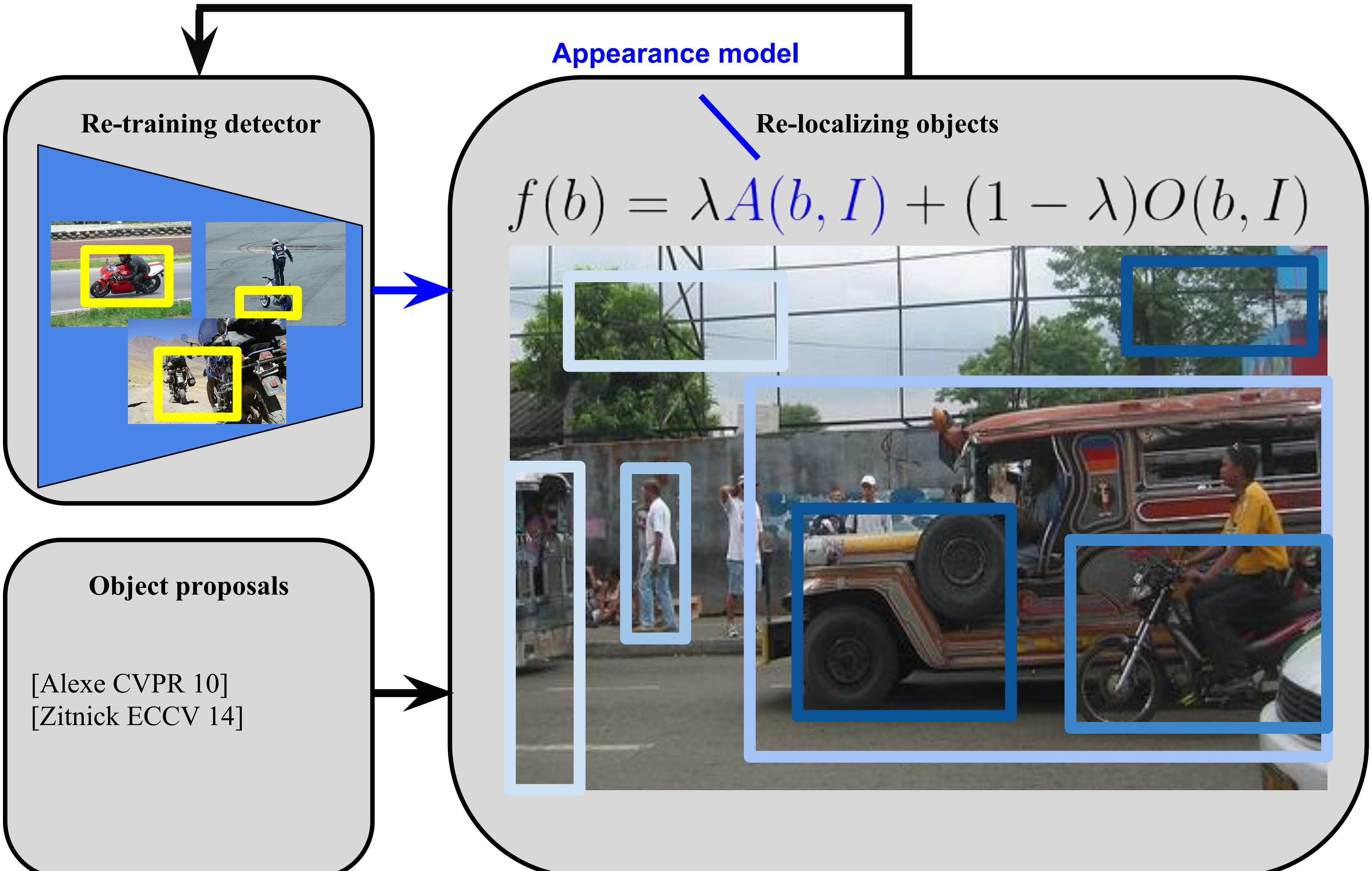
Initialization: full images
[Cinbis CVPR 14, Nguyen ICCV 09, Russakovsky ECCV 12, Pandey ICCV 11]

Objectness and multi-folding
[Deselaers ECCV 10] [Cinbis CVPR 14]

Re-localization:
pick proposal with
highest appearance score

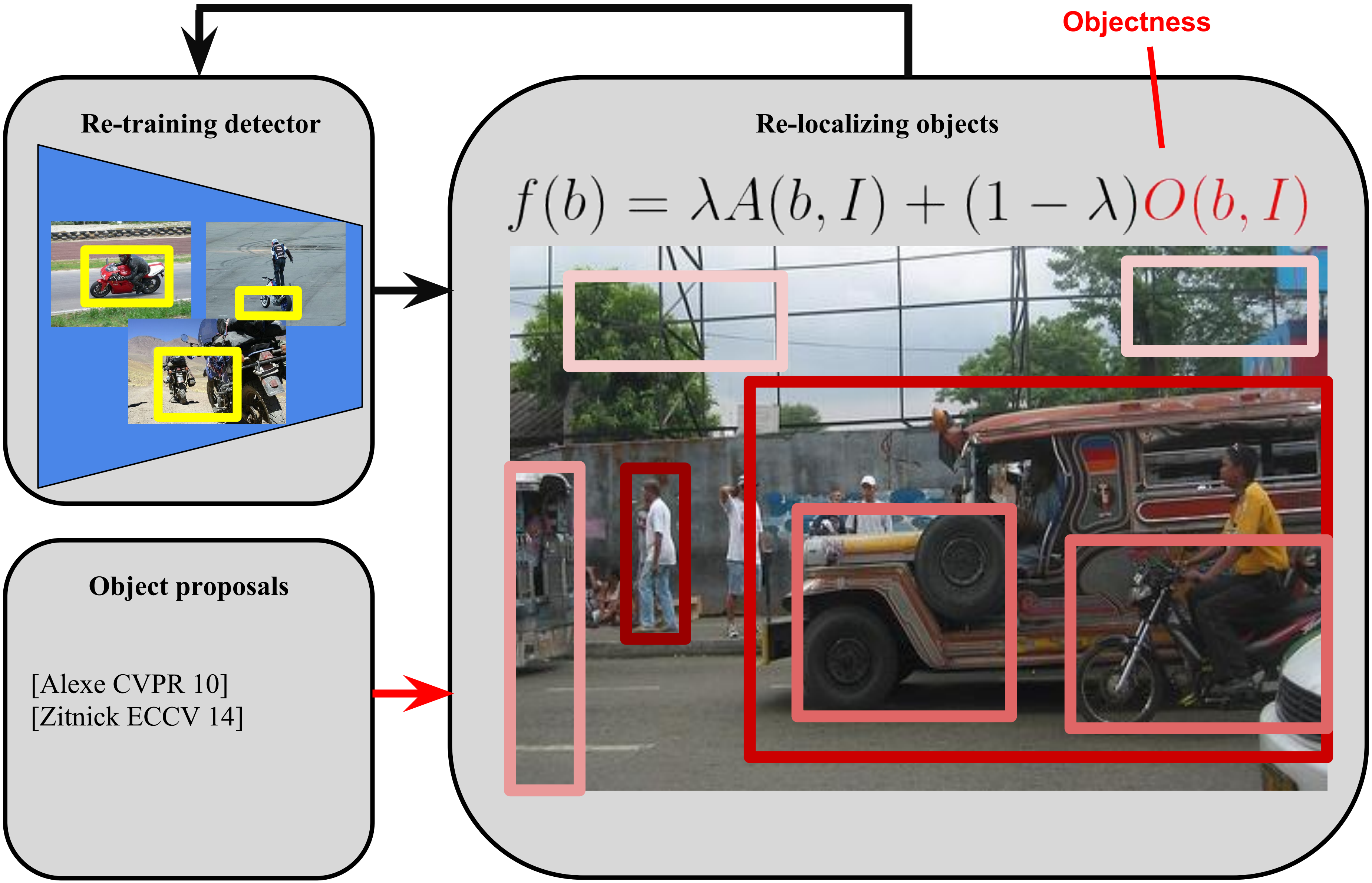
[Blaschko NIPS 10, Cinbis PAMI 16, Deselaers ECCV 10, Nguyen ICCV 09, Russakovsky ECCV 12, Shi PAMI 15, Shi BMVC 12, Siva ICCV 11 & ECCV 12, Song NIPS 14 & ICML 14, Bilen BMVC 14, Sangineto PAMI 18, Li BMVC 17]

MIL-style WSOL



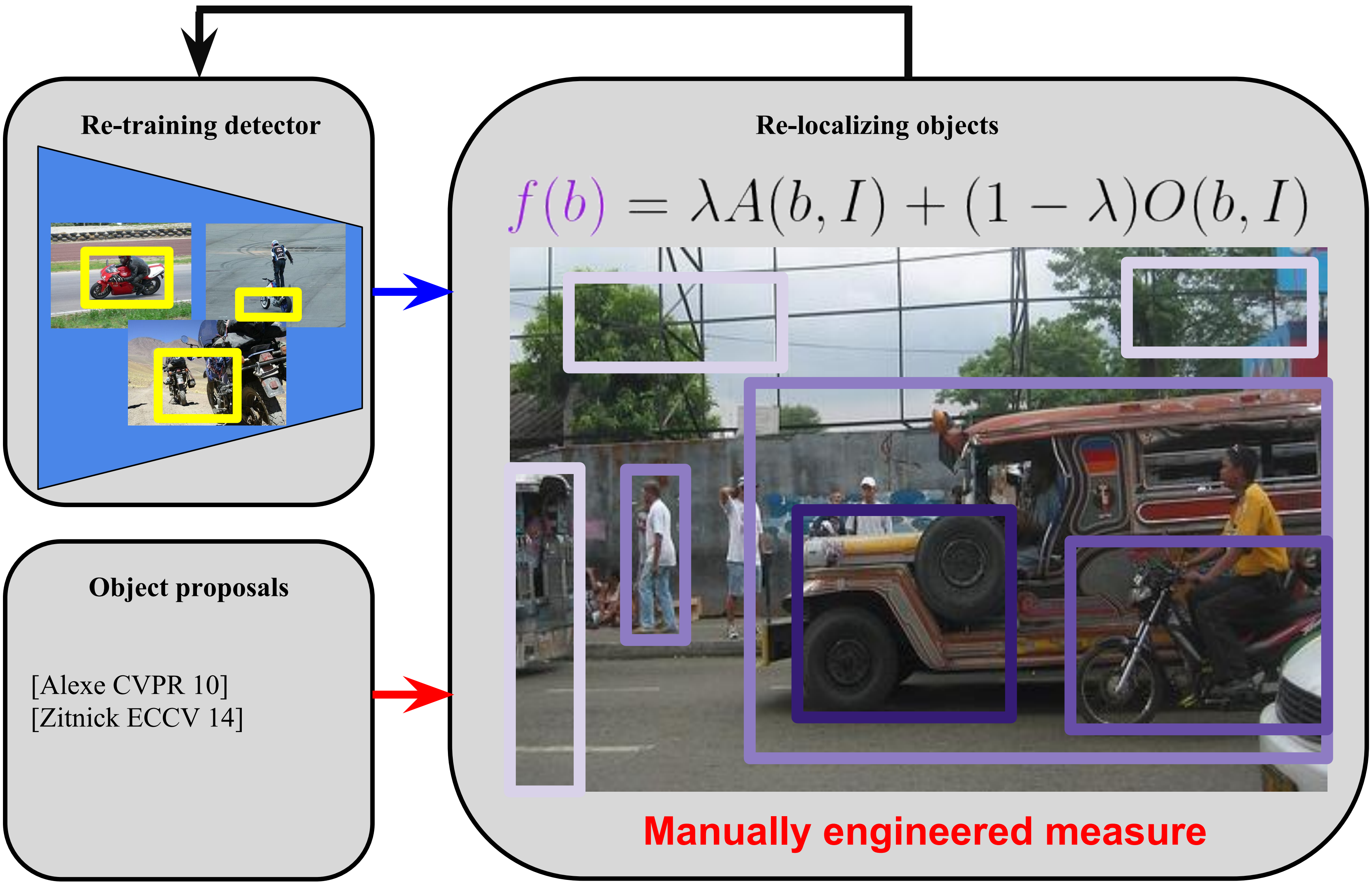
MIL-style WSOL

[**Deselaers ECCV 10**, Prest CVPR 12, Shapovalova ECCV 12, Shi BMVC 12, Siva ICCV 11, Cinbis CVPR 14, Bilen CVPR 16, Tang CVPR 14, Wang ECCV 14]



MIL-style WSOL

[**Deselaers ECCV 10**, Prest CVPR 12, Shapovalova ECCV 12, Shi BMVC 12, Siva ICCV 11, Cinbis CVPR 14, Bilen CVPR 16, Tang CVPR 14, Wang ECCV 14]

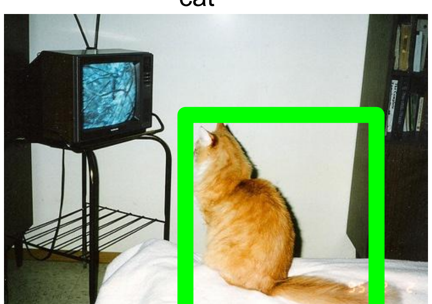


Problem setting: knowledge transfer

Source train set (bounding boxes)

bicycle

cat



knowledge transfer

Target train set (image-level labels)

motorbike

motorbike

no motorbike

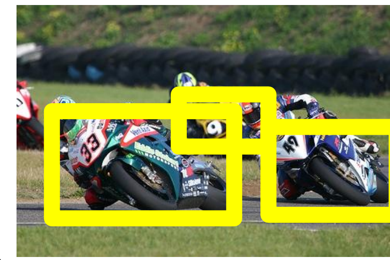
Train

Detection model

Apply

Target test set (no labels)

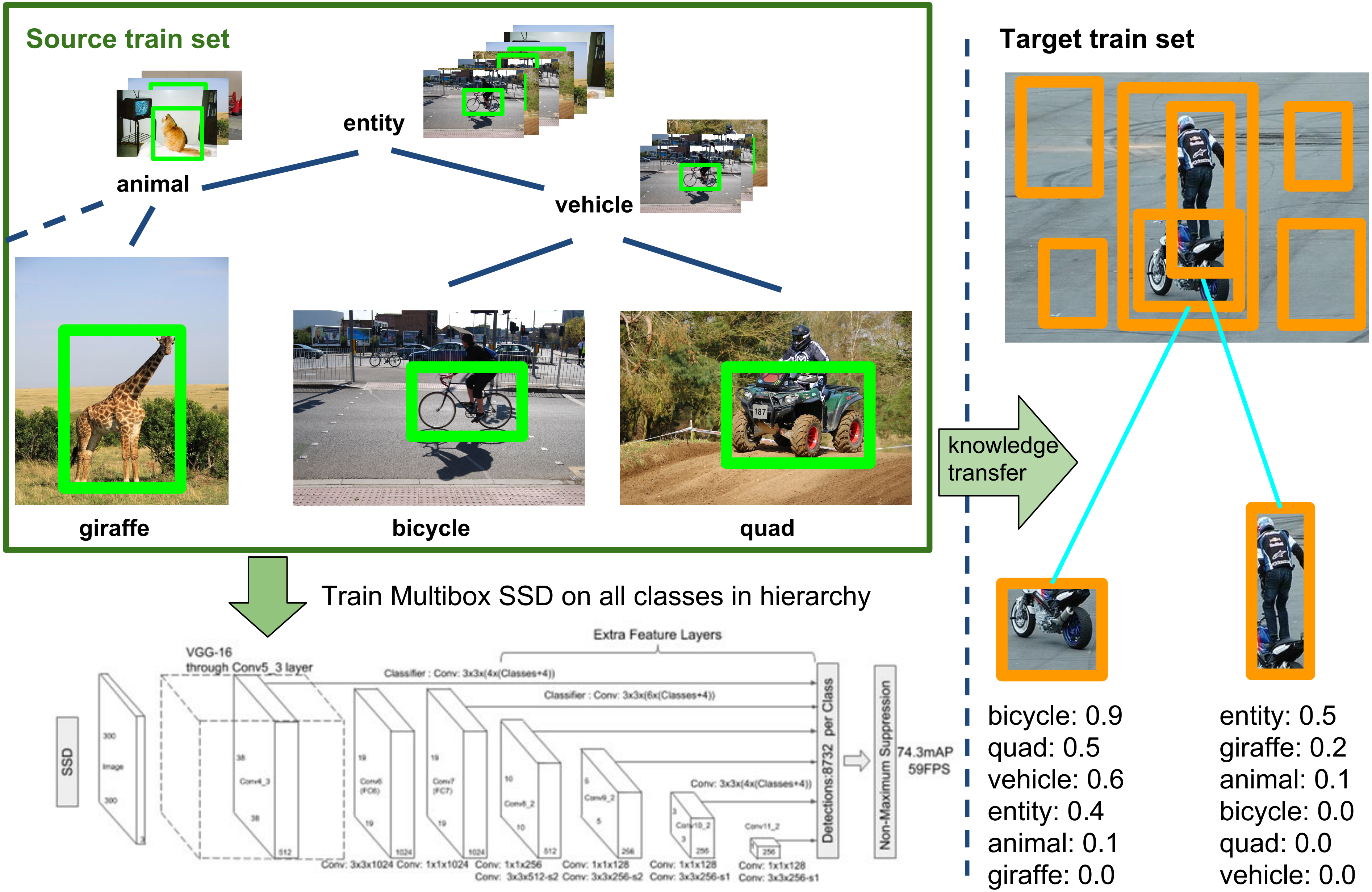
Goal: detect target class in test set

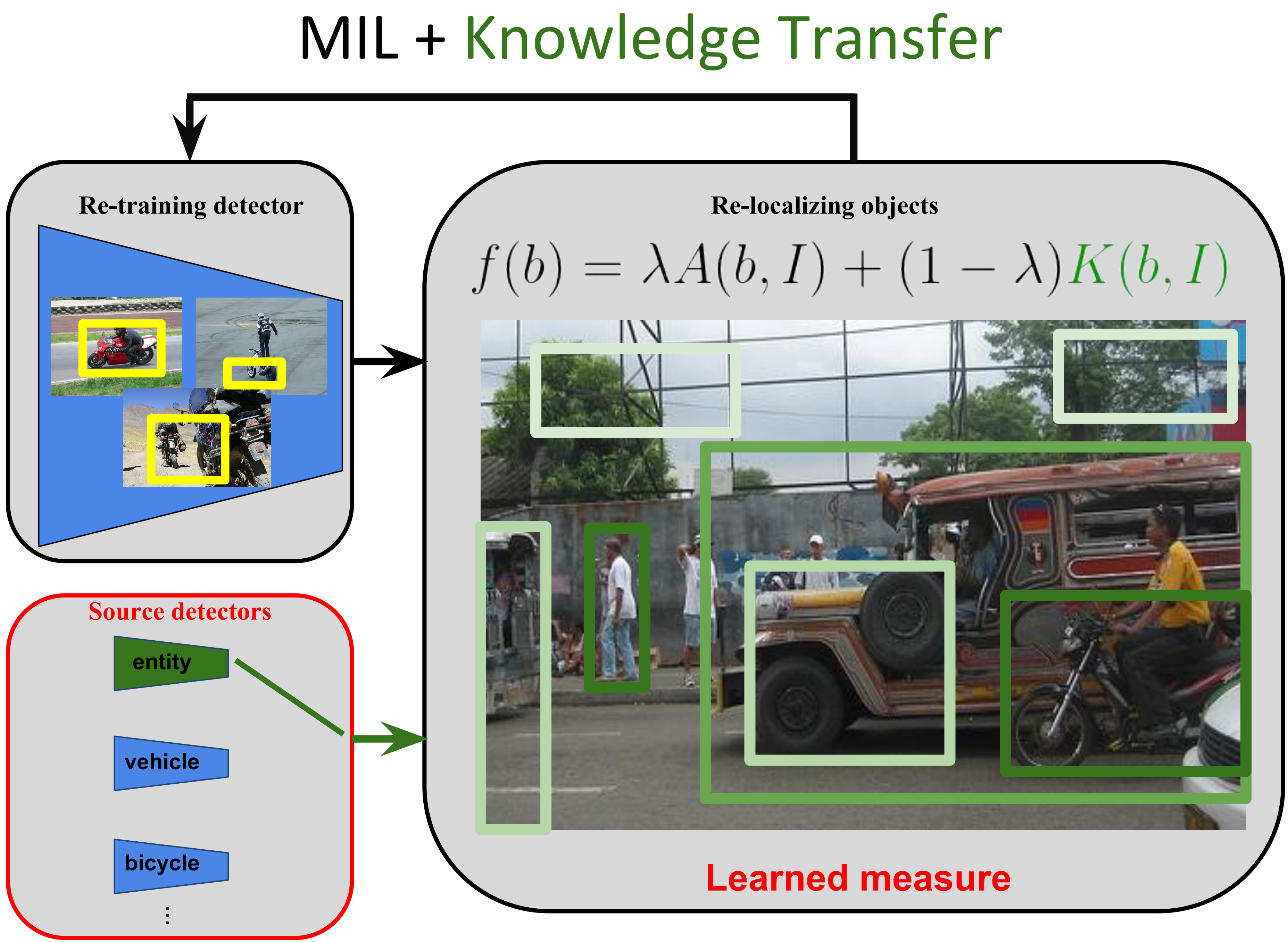


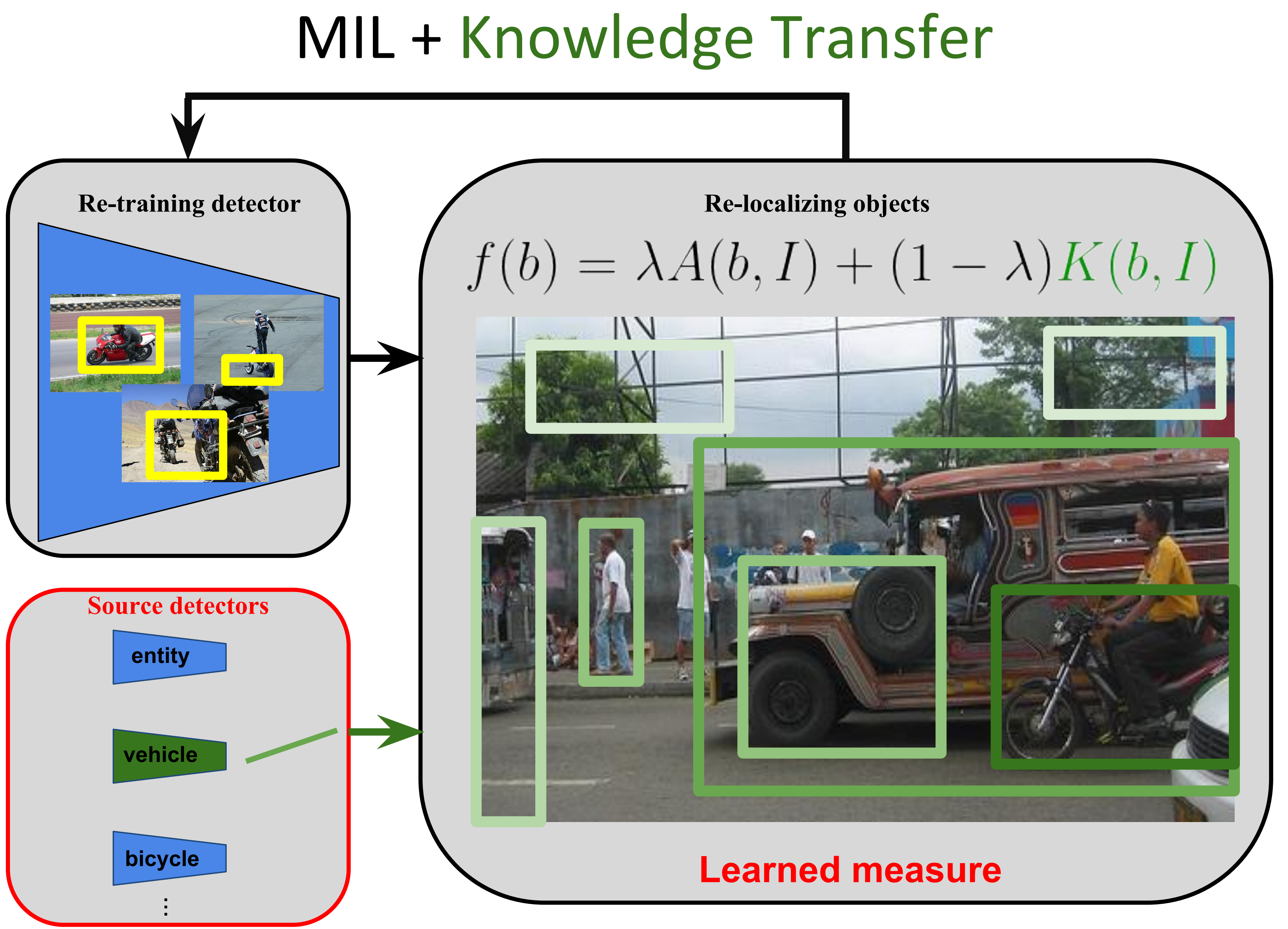
[Guillaumin CVPR 12, Hoffman NIPS 14, Rochan CVPR 15, Tang CVPR 16, Redmon CVPR 17]

Related:
[Guillaumin CVPR 12,
Hoffman NIPS 14]

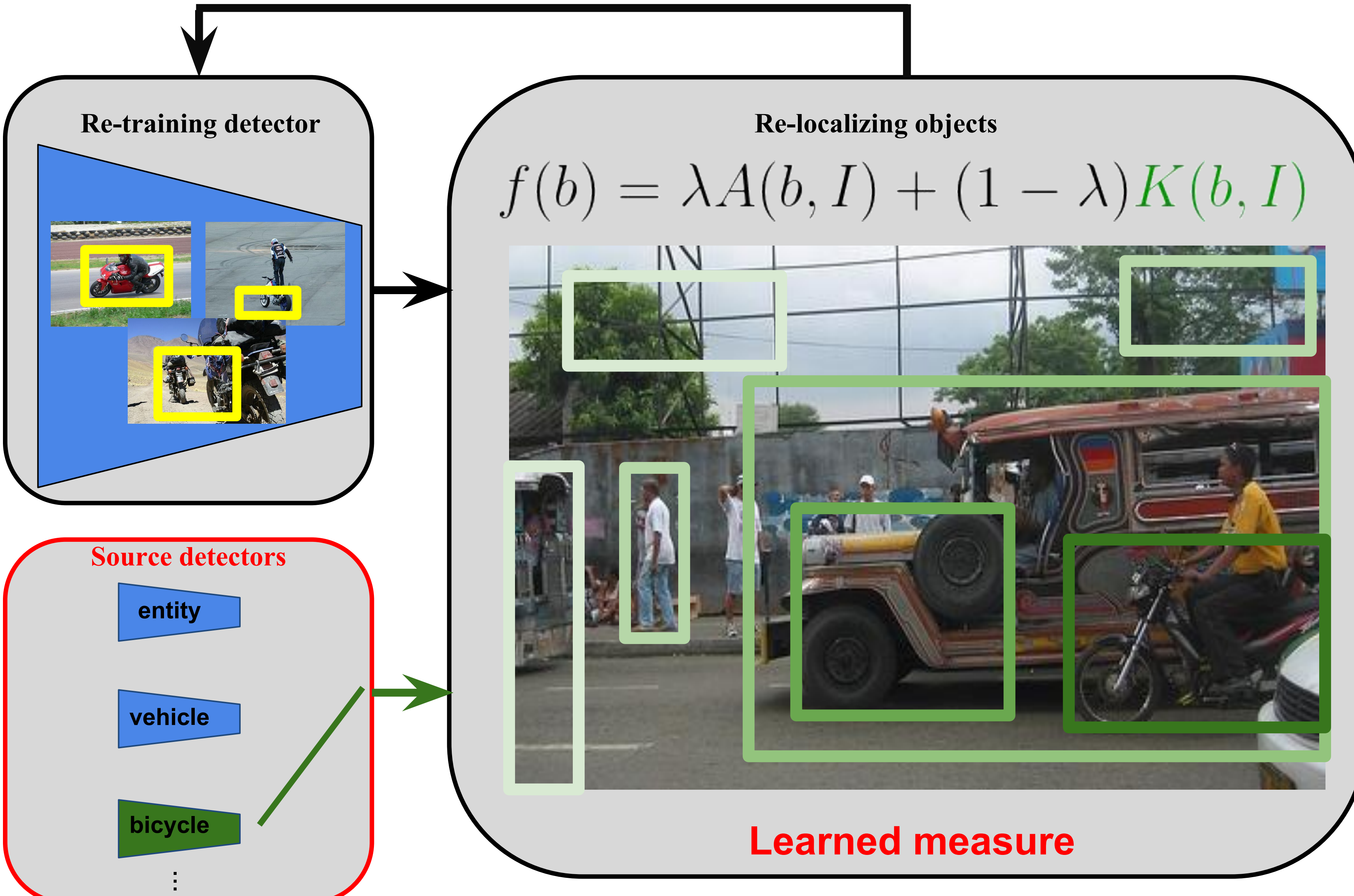
Knowledge Transfer




MIL + Knowledge Transfer
Re-training detector
Re-localizing objects
$f(b) = \lambda A(b, I) + (1 - \lambda) K(b, I)$
Source detectors
entity
vehicle
bicycle
Learned measure


MIL + Knowledge Transfer
Re-training detector
Re-localizing objects
$f(b) = \lambda A(b, I) + (1 - \lambda) K(b, I)$
Source detectors
entity
vehicle
bicycle
Learned measure

MIL + Knowledge Transfer



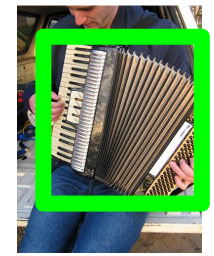
Dataset: ILSVRC 2013

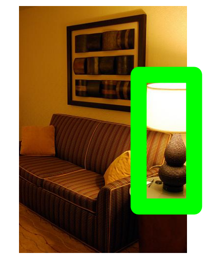
Setup following LSDA [Hoffman NIPS 14]

source training set

- val1 augmented to 1000 boxes per class [Girshick CVPR 14]
- classes 1-100
- bounding box annotations
- for each knowledge transfer function, optimize λ on 80/20 class split (and all other hyper-parameters ;)

Classes from accordion to lamp





target training set

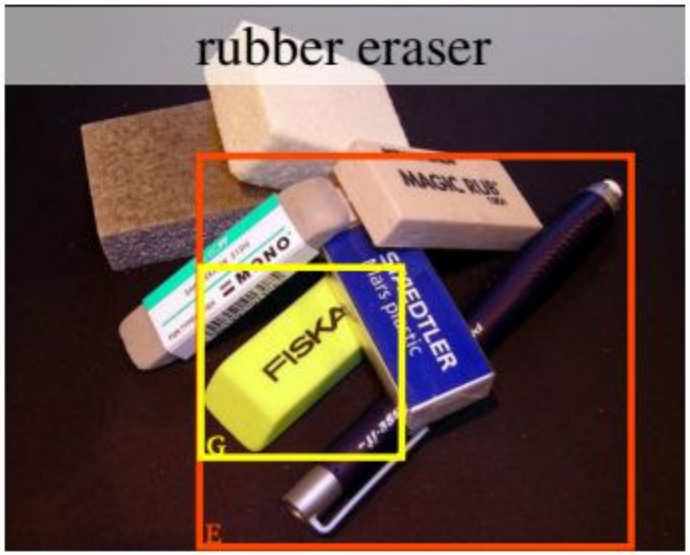
- val1 augmented to 1000 boxes per class [Girshick CVPR 14]
- classes 101-200
- Image-labels only

Classes from laptop to zebra

target test set

- Complete val2 set
- Accuracy measured on class 101-200 only

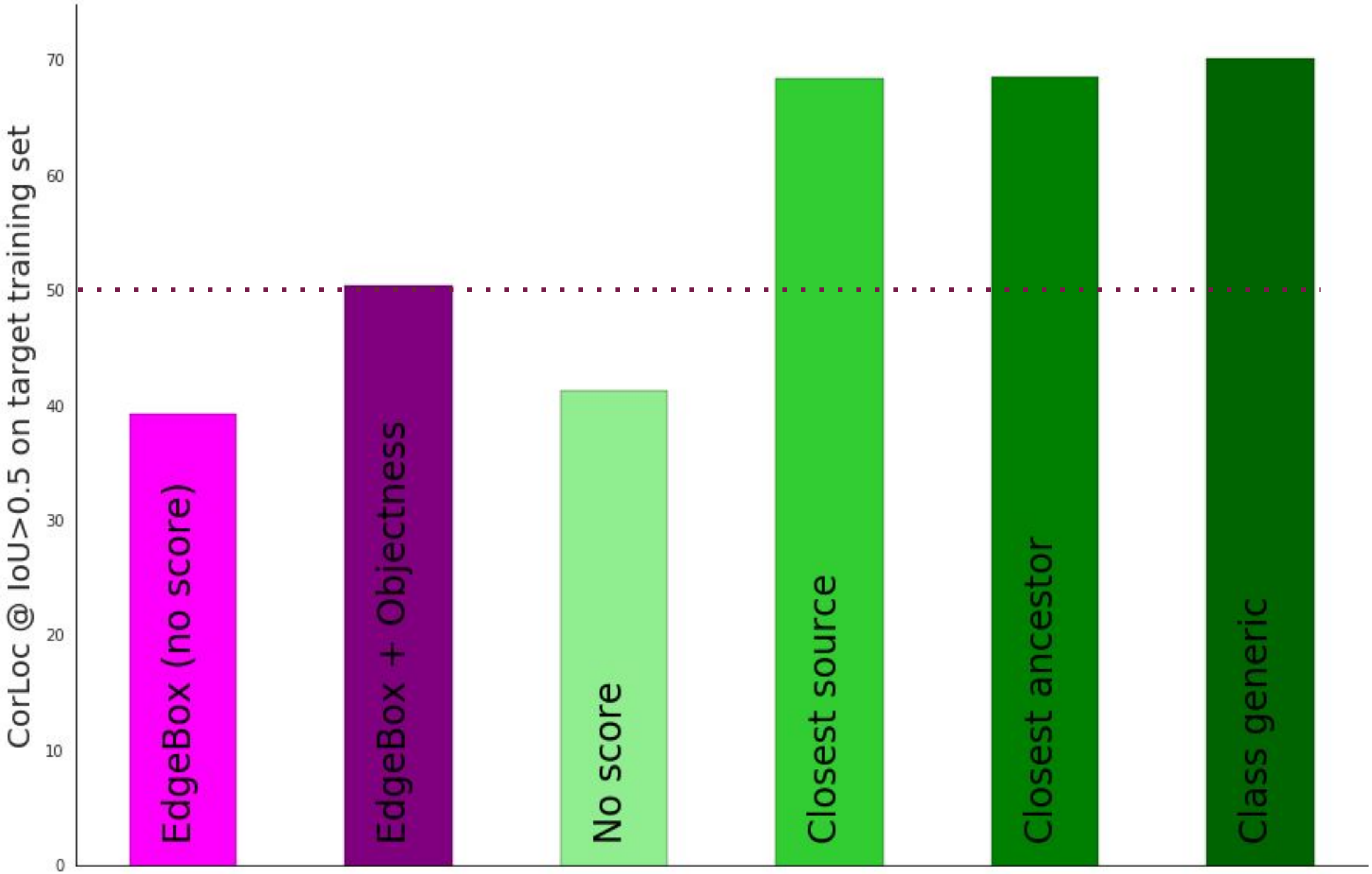
Results: Qualitative localizations on target train



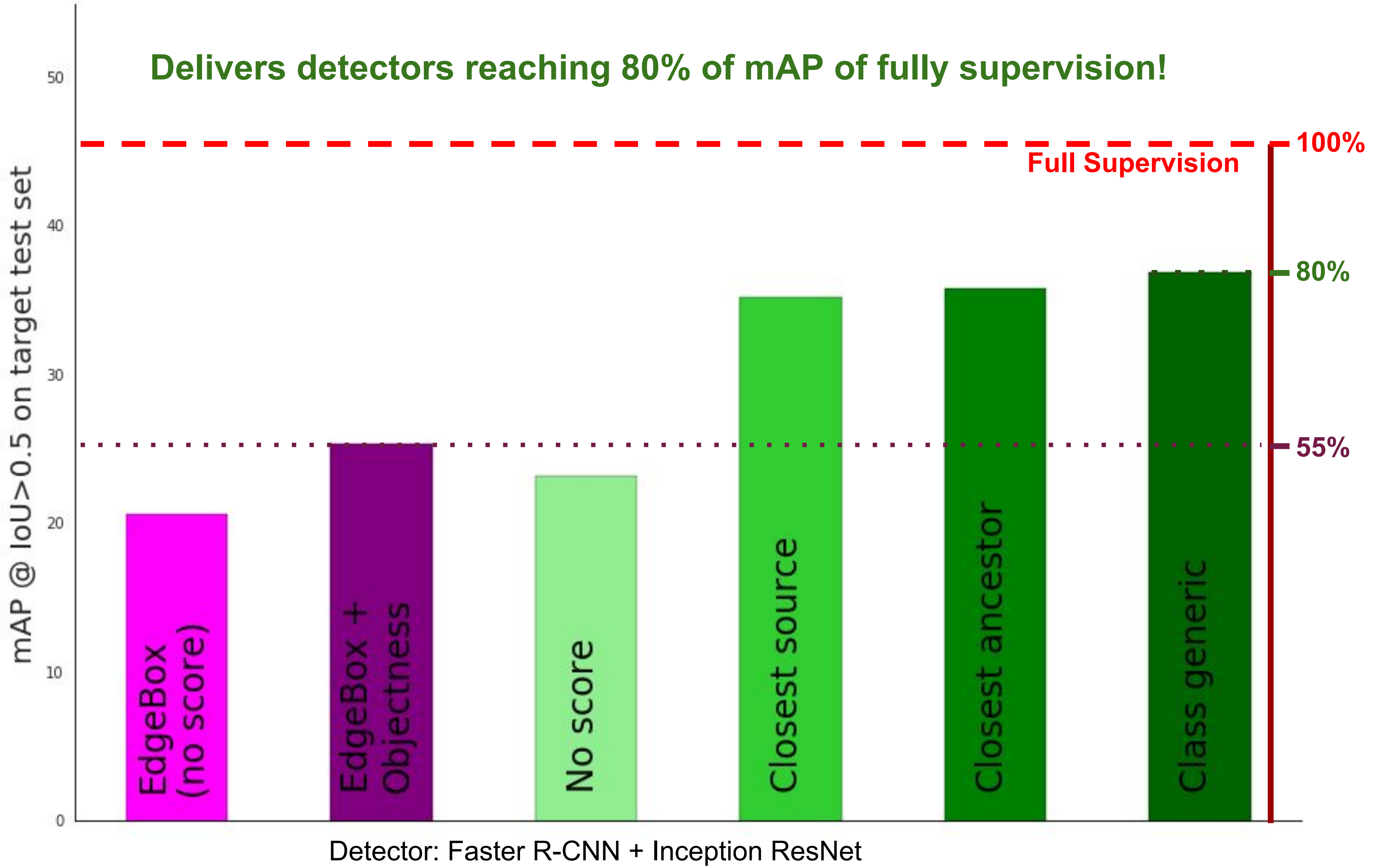
EdgeBox + objectness baseline

Knowledge Transfer (class-generic)

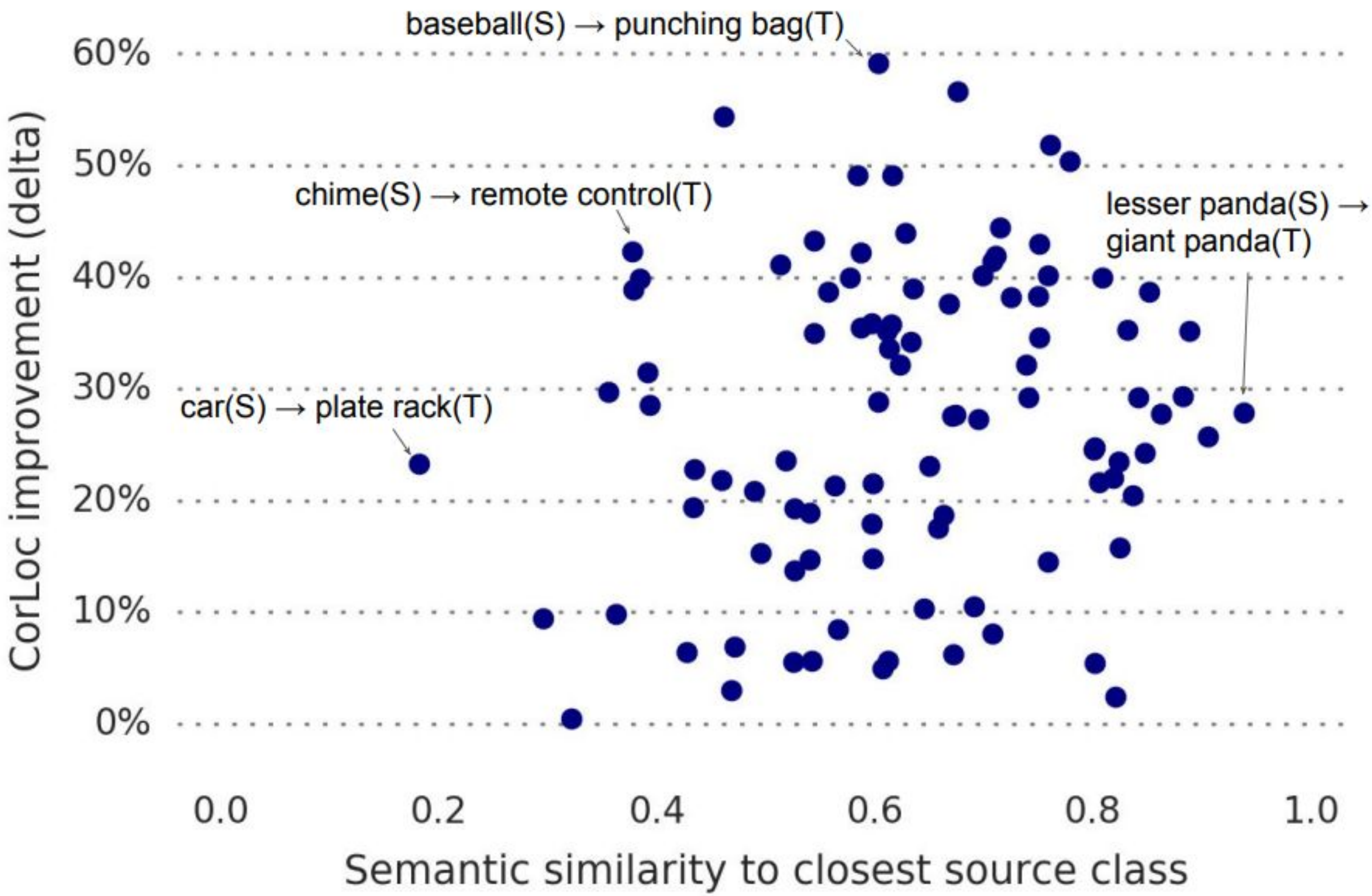
CorLoc@0.5 on target training set



mAP@ 0.5 on target test set

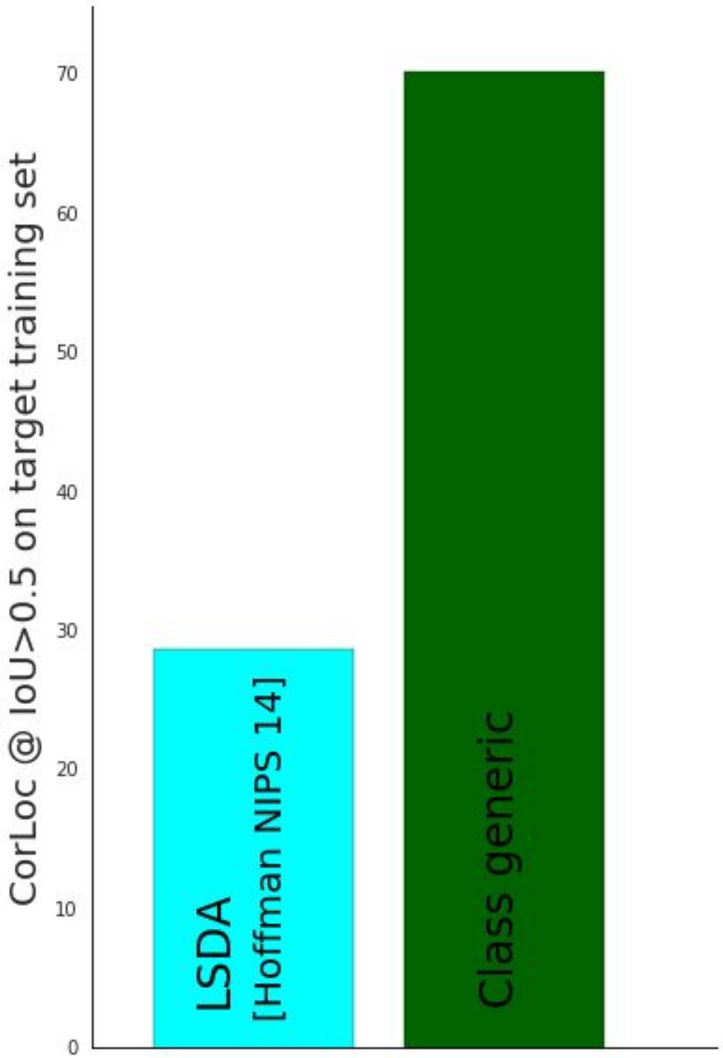


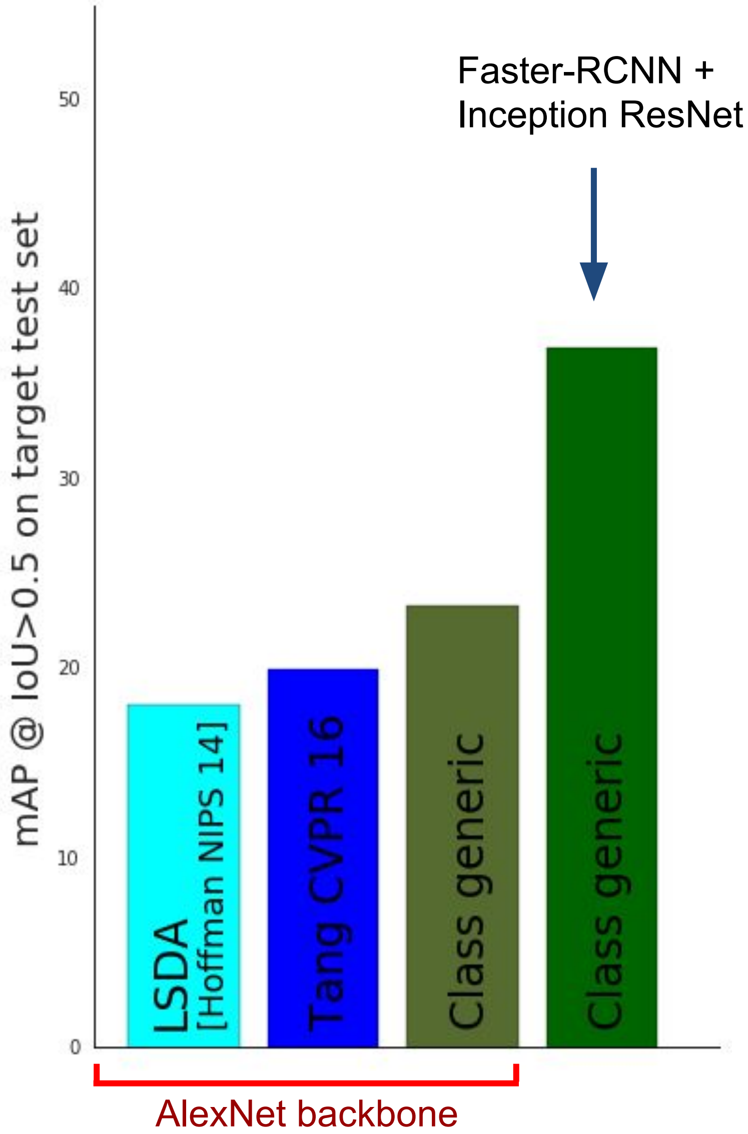
Semantic similarity vs improvement



✓ improvement uncorrelated with semantic similarity

Comparison to state-of-the-art





Comparison to YOLOv2 [Redmon ICCV 17]

Source training set

COCO
Common Objects in Context

Target training set

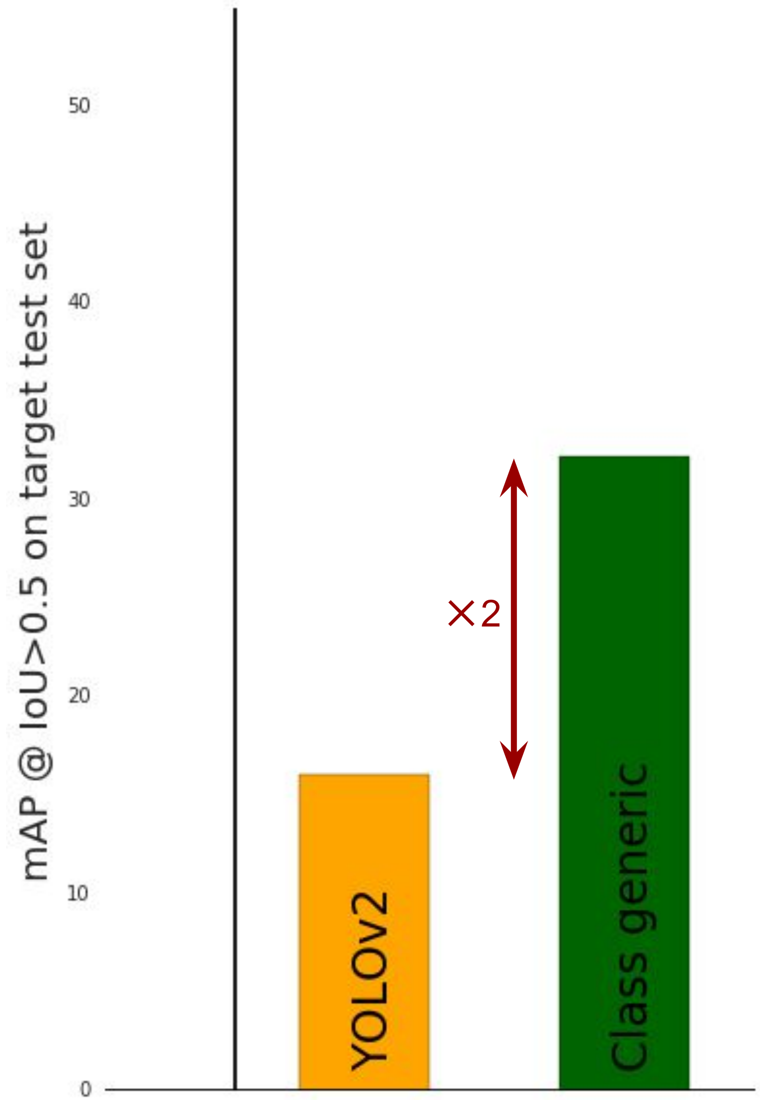
IMAGENET

classification-style image,
9000 classes,
no ILSVRC detection

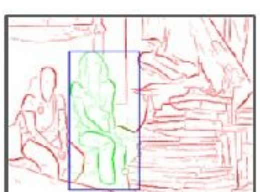
Target test set

IMAGENET

ILSVRC detection validation set:
156 Non-COCO classes



Generalization across datasets: CorLoc @ 0.5

Source training set \ Target training set	IMAGENET Augmented val 1, class 101-200	COCO Common Objects in Context 2014 train	OPEN IMAGES V2, val+test
IMAGENET Augmented val 1, class 1-100 100 classes, 65k images, 81k boxes	**74.2**	**34.5**	**62.0**
COCO Common Objects in Context 2014 train 80 classes, 83k images, 605k boxes	67.7	-	59.5
PASCAL2 Pattern Analysis, Statistical Modelling and Computational Learning **VOC 2007 trainval** 20 classes, 5k images, 13k boxes	59.5	26.2	55.3
 EdgeBox baseline	50.5	20.6	32.4

Conclusions

- Knowledge Transfer substantially improves Weakly Supervised Object Detection

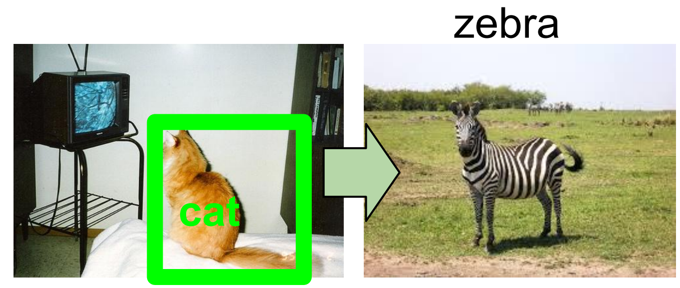
- Delivers detectors performing at 80% of the mAP of their fully supervised counterparts

- Generalize across a wide range of source-target dataset pairs

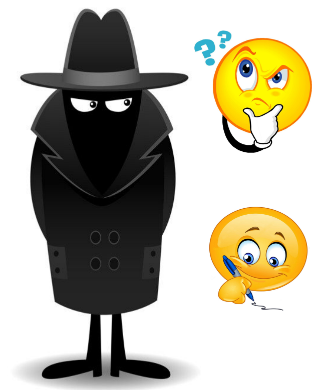
- Simple modification to standard MIL pipelines

This talk

- Revisiting knowledge transfer for training object class detectors
  Uijlings, Popov, Ferrari
  CVPR 2018

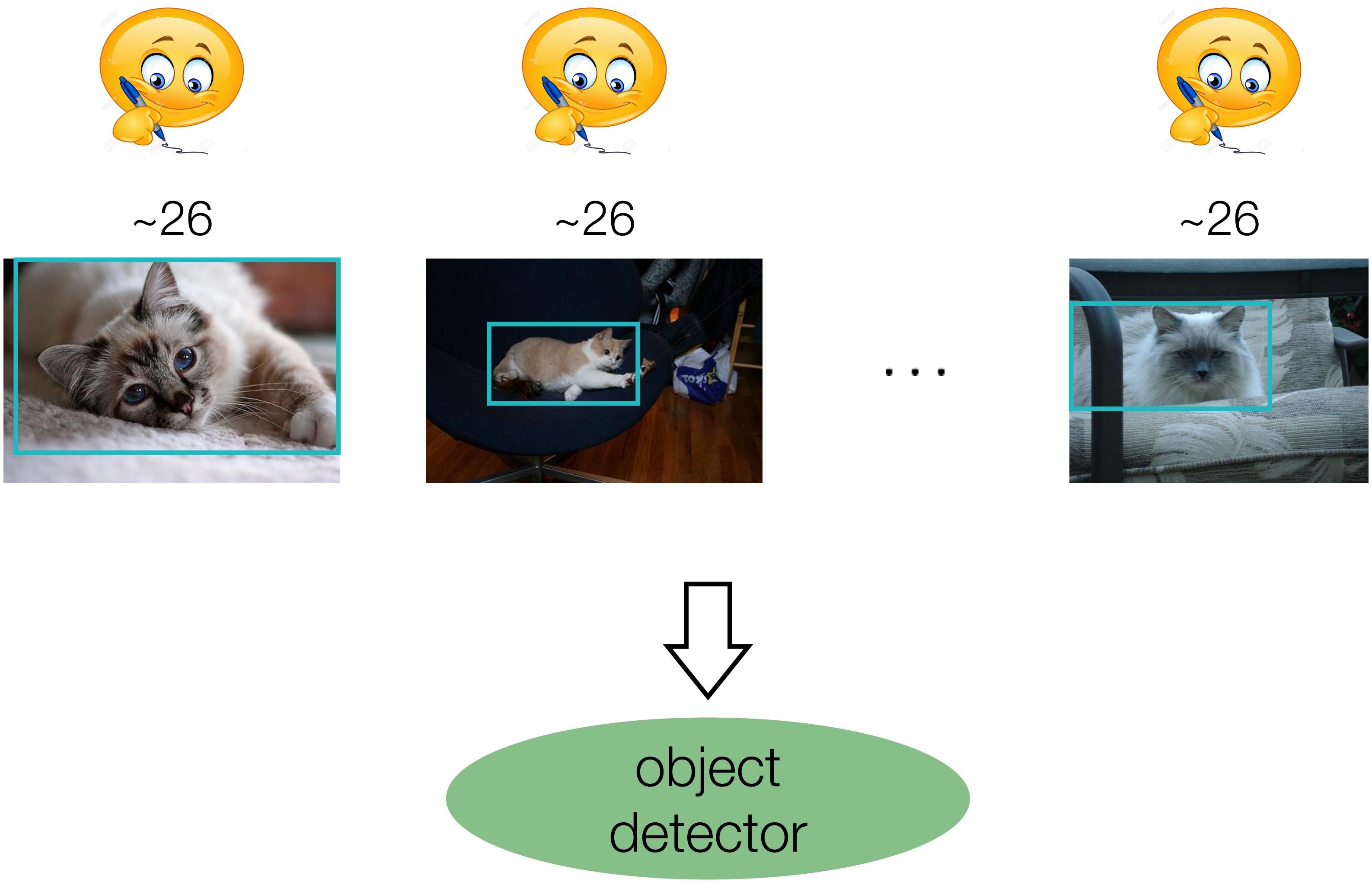


- **Learning intelligent dialogs for bounding box annotation**
  Konyushkova, Uijlings, Lampert, Ferrari
  CVPR 2018



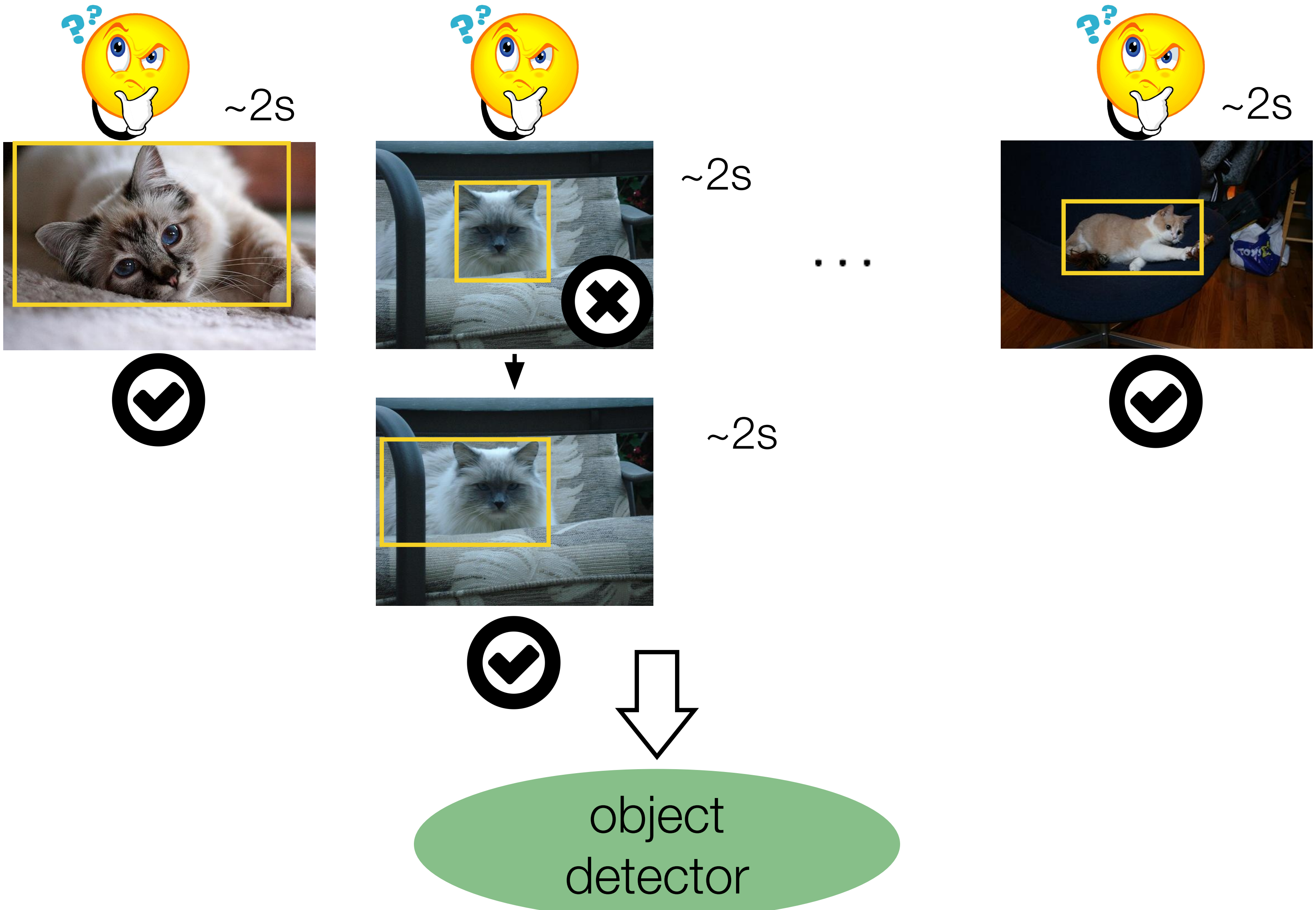
- Fluid Annotation: human-machine collaboration for full image annotation
  Andriluka, Uijlings, Ferrari,
  arXiv June 2018

Manual box drawing



ImageNet annotation timings: [Su AAAI 2012]
Crowdsourcing annotations for visual object detection. AAAI Human Computation Workshop 2012

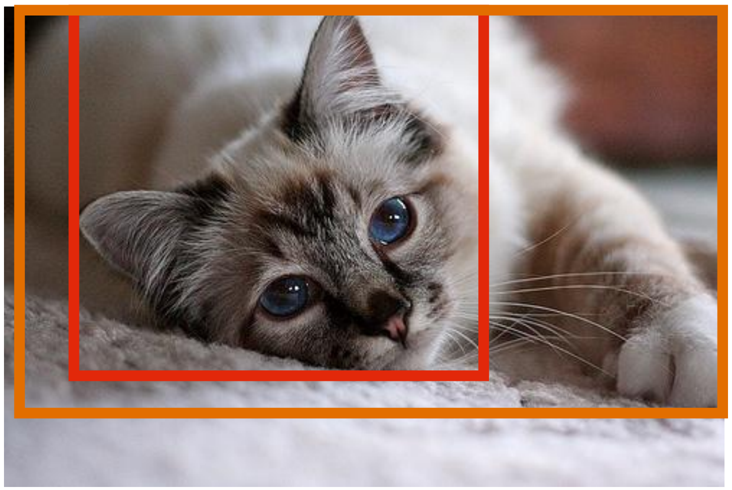
Box verification series



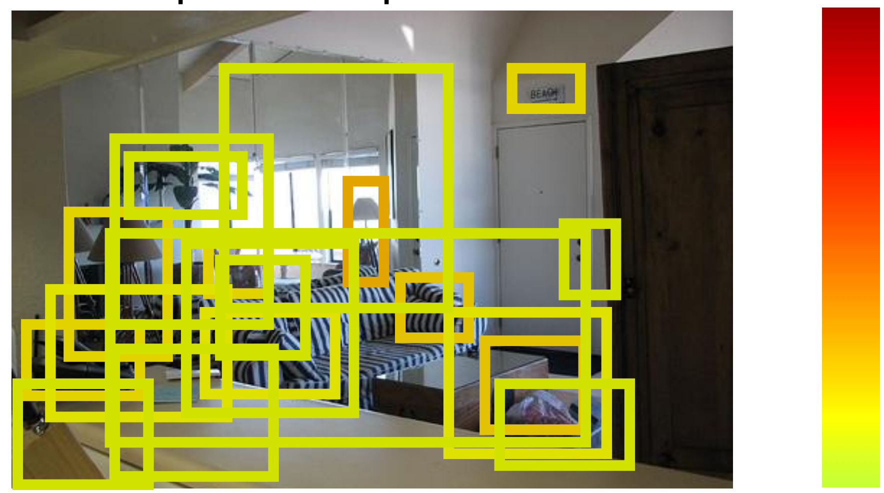
D. P. Papadopoulos, J. R. R. Uijlings, F. Keller, and V. Ferrari.
We don't need no bounding-boxes: Training object class detectors using only human verification. CVPR 2016

Success, failure and motivation

cat



potted plant

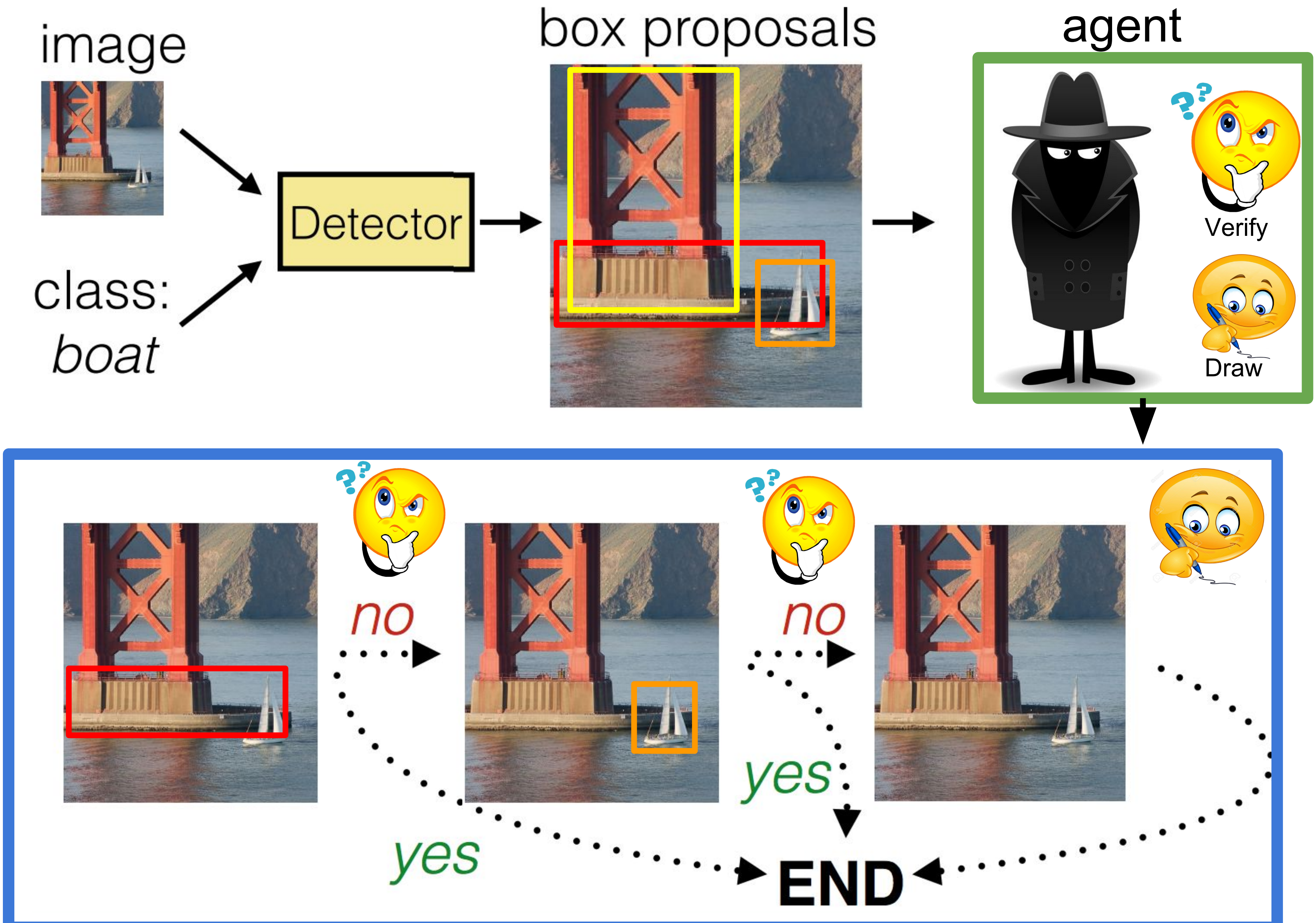


Verify

Draw

Depends on image difficulty, detector strength, desired box quality

Intelligent Annotation Dialog (IAD)



Agent 1: Model-based

Given:

- time for drawing

  t_{draw}

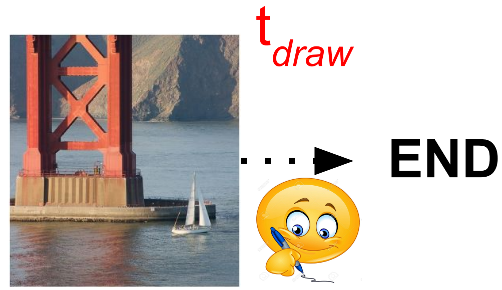
- time for verification

  t_{verify}

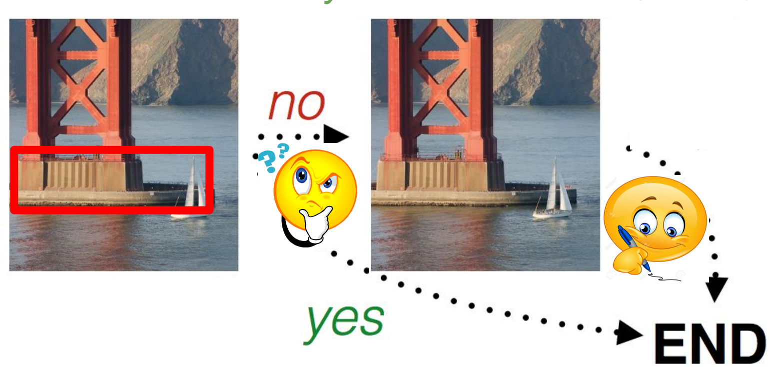
- probability of acceptance of box *i*

  p_i

we can estimate the *expected annotation time* for any dialog

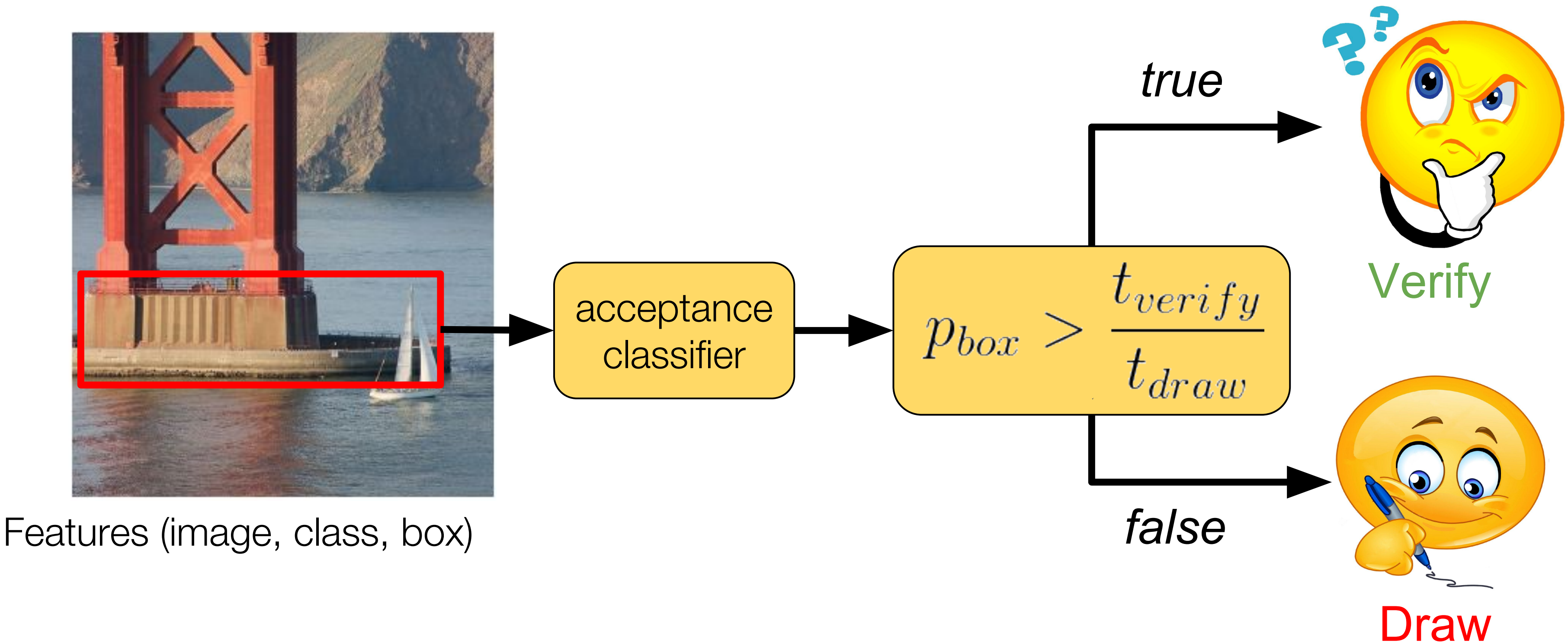


$$t_{verify} \quad + \quad (1-p_0) * t_{draw}$$



[...]

Agent 1: Model-based



Optimal strategy: sort boxes in order of acceptance probability; verify all boxes for which $p > t_{\text{verify}}/t_{\text{draw}}$

Proof

Algorithm IAD Prob

Input: $B_0 = (b_1, \dots, b_n); p(b_1), \dots, p(b_n); t_V; t_D$
$S_n = (\bar{s}_1, \bar{s}_2, \dots, \bar{s}_n) \leftarrow \mathrm{sort}(B_0)$ by $p(b_i)$
$\pi = ()$
$A_k = ()$
while $p(\bar{s}_i) > t_V / t_D$ **do**
  $A_k \leftarrow A_k \cdot s_i$
  select action $V : \pi \leftarrow \pi \cdot V$
select action $D : \pi \leftarrow \pi \cdot D$
return sequence of actions π, sequence of boxes A_k

Theorem 1. *If probabilities of acceptance $p(b_i)$ are known, the strategy of applying a sequence of actions $V^k D$ defined by IAD-Prob to a sequence of boxes A_k minimizes the annotation time, i.e. for all $m \in \{0, \dots, n\}$ and for all box sequences S_m:*

$$\mathbb{E}[t(V^k D, A_k)] \leq \mathbb{E}[t(V^m D, S_m)] \qquad (1)$$

Sketch of the proof. The proof consists of two parts. First, we show that for any strategy $V^m D$, the best box sequence is obtained by sorting the available boxes by their probability of acceptance and using the first m of them. Second, we show that the number of verification steps found by IAD-Prob, k, is indeed the optimal one.

We start by rewriting the expected episode length in closed form. For a strategy $V^m D$ and any sequence of boxes, $S_m = (s_1, \dots, s_m)$, we obtain

$$\begin{aligned}\mathbb{E}[t(V^m D, S_m)] &= t_V + q(s_1)t_V + q(s_1)q(s_2)t_V + \dots \\ &+ q(s_1)q(s_2)\cdots q(s_{m-1})t_V + q(s_1)q(s_2)\cdots q(s_m)t_D \\ &= t_V \sum_{i=0}^{m-1} \prod_{j=1}^{i} q(s_j) + t_D \prod_{j=1}^{m} q(s_j). \end{aligned} \qquad (2)$$

Our first observation is that (2) is monotonically decreasing as a function of $q(s_1), \dots, q(s_m)$. Consequently, the smallest value is obtained by selecting the set of m boxes that have the smallest rejection probabilities. To prove that their optimal order is sorted in decreasing order, assume that S_m is not sorted, i.e. there exists an index $l \in \{1, \dots, m-1\}$ for which $q(s_l) > q(s_{l+1})$. We compare the expected episode length of S_m to that of a sequence $\hat{S}_m$ in which s_l and s_{l+1} are at switched positions. Using (2) and noticing that many of the terms cancel out, we obtain

$$\begin{aligned}&\mathbb{E}[t(V^m D, S_m)] - \mathbb{E}[t(V^m D, \hat{S}_m)] \\ &= t_V \big(q(s_l) - q(s_{l+1})\big)\Big(\prod_{j=1}^{l-1} q_j\Big) > 0. \end{aligned} \qquad (3)$$

This shows that $\hat{S}_m$ has strictly smaller expected episode length than S_m, so S_m cannot have been the optimal order.

Consequently, for any strategy $V^m D$, the optimal sequence is to sort the boxes by decreasing probability of rejection, *i.e.* increasing acceptance probability. We denote it by $\bar{S}_m = (s_1, \dots, s_m)$.

Next, we show that the number, k, of verification actions found by the IAD-Prob algorithm is optimal, *i.e.* $V^k D$ is better or equal to $V^m D$ for any $m \neq k$. As we already know that the optimal box sequence for any strategy $V^m D$ is $\bar{S}_m$, it is enough to show that

$$\mathbb{E}[t(V^{m-1} D, S_{m-1})] \geq \mathbb{E}[t(V^m D, S_m)], \qquad (4)$$

for all $m \in \{1, \dots, k\}$, and

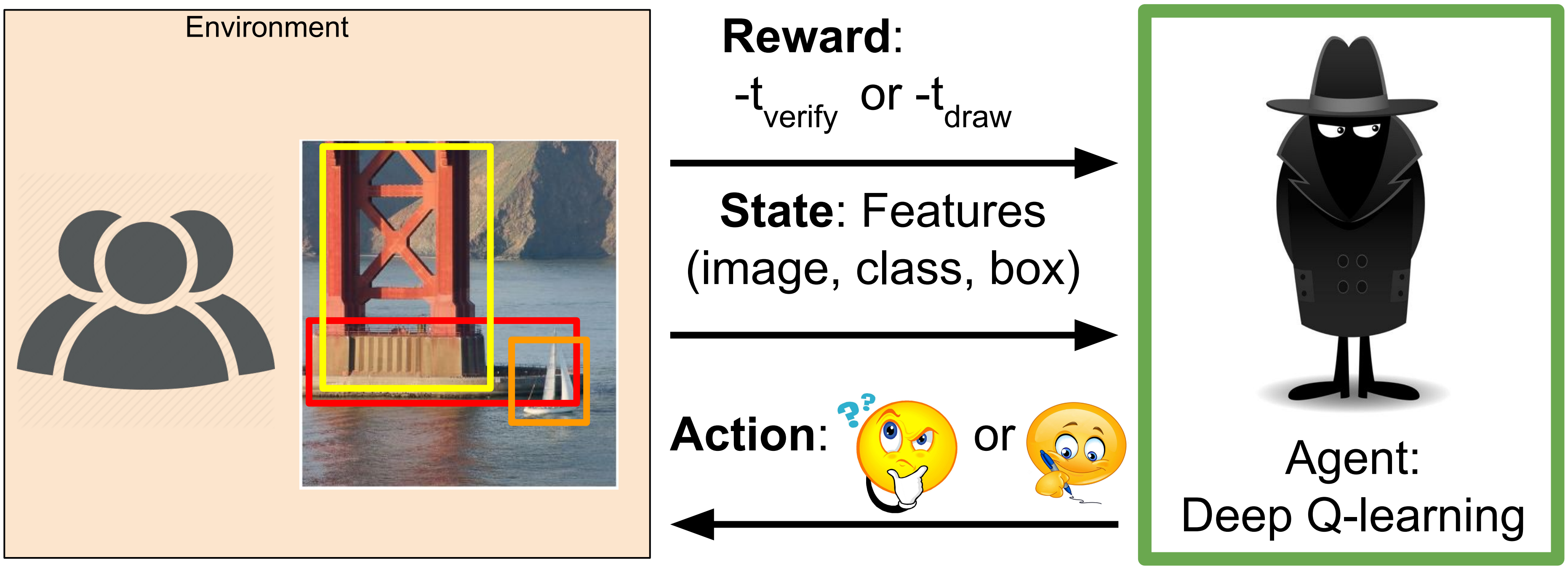
$$\mathbb{E}[t(V^{m-1} D, S_{m-1})] \leq \mathbb{E}[t(V^m D, S_m)], \qquad (5)$$

for all $m \in \{k+1, \dots, n\}$. To prove these inequalities, we again make use of expression (2). For any $m \in \{1, \dots, n-1\}$ we obtain

$$\begin{aligned}&\mathbb{E}[t(V^m D, S_m)] - \mathbb{E}[t(V^{m-1} D, S_{m-1})] \\ &= t_V \prod_{j=1}^{m-1} q(s_j) + t_D \prod_{j=1}^{m} q(s_j) - t_D \prod_{j=1}^{m-1} q(s_j) \\ &= \Big(\prod_{j=1}^{m-1} q(s_j)\Big)\big(t_V + q(s_m)t_D - t_D\big). \end{aligned} \qquad (6)$$

For $m \in \{1, \dots, k\}$, we know that $p(s_m) > t_V / t_D$ by construction of the strategy. This is equivalent to $t_V + q(s_m)t_D - t_D > 0$. Consequently, (6) is nonnegative in this case, and inequality (4) is confirmed. For $m \in \{k+1, \dots, n\}$, we know $p(s_m) \leq t_V / t_D$, again by construction. Consequently, $t_V - q(s_m)t_D - t_D < 0$, which shows that (6) is nonpositive in this case, confirming (5). $\square$

Agent 2: Reinforcement Learning



Experimental Settings

- Annotate PASCAL VOC 2007 trainval dataset
- Image-level labels are available
- 10% reserved for training the agent
- Faster-RCNN object detector [Ren *NIPS* 2015]
- Detector either weak (MIL on VOC07) or strong (fully supervised on VOC12)
- DQN (simplified) as a reinforcement learning algorithm [Mnih Nature 2015]
- RL is unstable for non-linear function approximation, so: experience replay, periodically updated Q-function

Different scenarios

$\alpha = 0.5$

Desired quality

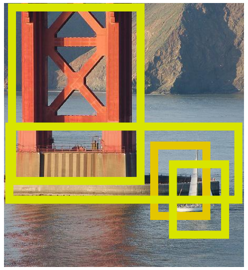
low — high

$\alpha = 0.7$

Train from boxes

Detector strength

high — low



MIL from image labels

26s

[Su AAAI 12] ImageNet

Drawing time

high — low

7s

[Papadopoulos ICCV 17] X-Click

V(erify)

D(raw)

Different scenarios

Agent vs standard strategies

- D: Draw
- V*D: Box verification series

Drawing technique	Slow drawing		Fast drawing			
Detector	Weak detector		Weak detector		Strong detector	
Quality level	$\alpha = 0.7$	$\alpha = 0.5$	$\alpha = 0.7$	$\alpha = 0.5$	$\alpha = 0.7$	$\alpha = 0.5$
D	25.50 ± 0.00	25.50 ± 0.00	$\mathbf{7.00} \pm 0.00$	7.00 ± 0.00	7.00 ± 0.00	7.00 ± 0.00
V*D	42.29 ± 0.07	17.37 ± 0.07	31.82 ± 0.11	11.46 ± 0.04	8.83 ± 0.09	3.18 ± 0.02
Model-based agent	23.07 ± 0.23	12.64 ± 1.29	6.81 ± 0.02	5.86 ± 0.04	3.42 ± 0.18	2.73 ± 0.08
RL agent	23.62 ± 0.38	16.30 ± 0.09	6.83 ± 0.03	5.89 ± 0.05	3.60 ± 0.07	2.66 ± 0.06

Agent outperforms standard strategies in all scenarios

Different scenarios

Fixed strategies

- VD
- VVD
- VVVD

Drawing technique	Slow drawing		Fast drawing			
Detector	Weak detector		Weak detector		Strong detector	
Quality level	$\alpha = 0.7$	$\alpha = 0.5$	$\alpha = 0.7$	$\alpha = 0.5$	$\alpha = 0.7$	$\alpha = 0.5$
D	25.50 ± 0.00	25.50 ± 0.00	$\mathbf{7.00} \pm 0.00$	7.00 ± 0.00	7.00 ± 0.00	7.00 ± 0.00
VD	$\mathbf{23.01} \pm 0.07$	17.30 ± 0.07	7.62 ± 0.02	$\mathbf{6.05} \pm 0.02$	$\mathbf{3.45} \pm 0.01$	2.50 ± 0.01
VVD	23.79 ± 0.06	16.67 ± 0.06	8.92 ± 0.02	6.67 ± 0.02	3.48 ± 0.01	$\mathbf{2.45} \pm 0.01$
VVVD	24.67 ± 0.07	$\mathbf{16.38} \pm 0.07$	10.21 ± 0.02	7.32 ± 0.03	3.65 ± 0.02	2.48 ± 0.01
V*D	42.29 ± 0.07	17.37 ± 0.07	31.82 ± 0.11	11.46 ± 0.04	8.83 ± 0.09	3.18 ± 0.02

No single fixed strategy works best in all scenarios

Different scenarios

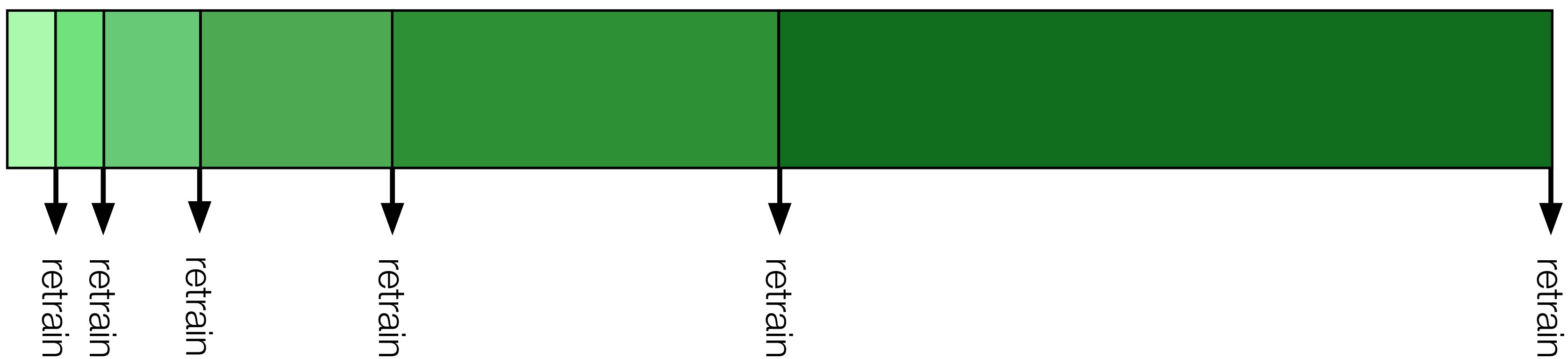
Agent vs fixed strategies

Drawing technique	Slow drawing		Fast drawing			
Detector	Weak detector		Weak detector		Strong detector	
Quality level	$\alpha = 0.7$	$\alpha = 0.5$	$\alpha = 0.7$	$\alpha = 0.5$	$\alpha = 0.7$	$\alpha = 0.5$
D	25.50 ± 0.00	25.50 ± 0.00	$\mathbf{7.00} \pm 0.00$	7.00 ± 0.00	7.00 ± 0.00	7.00 ± 0.00
VD	$\mathbf{23.01} \pm 0.07$	17.30 ± 0.07	7.62 ± 0.02	$\mathbf{6.05} \pm 0.02$	$\mathbf{3.45} \pm 0.01$	2.50 ± 0.01
VVD	23.79 ± 0.06	16.67 ± 0.06	8.92 ± 0.02	6.67 ± 0.02	3.48 ± 0.01	$\mathbf{2.45} \pm 0.01$
VVVD	24.67 ± 0.07	$\mathbf{16.38} \pm 0.07$	10.21 ± 0.02	7.32 ± 0.03	3.65 ± 0.02	2.48 ± 0.01
V*D	42.29 ± 0.07	17.37 ± 0.07	31.82 ± 0.11	11.46 ± 0.04	8.83 ± 0.09	3.18 ± 0.02
Model-based agent	23.07 ± 0.23	12.64 ± 1.29	6.81 ± 0.02	5.86 ± 0.04	3.42 ± 0.18	2.73 ± 0.08
RL agent	23.62 ± 0.38	16.30 ± 0.09	6.83 ± 0.03	5.89 ± 0.05	3.60 ± 0.07	2.66 ± 0.06

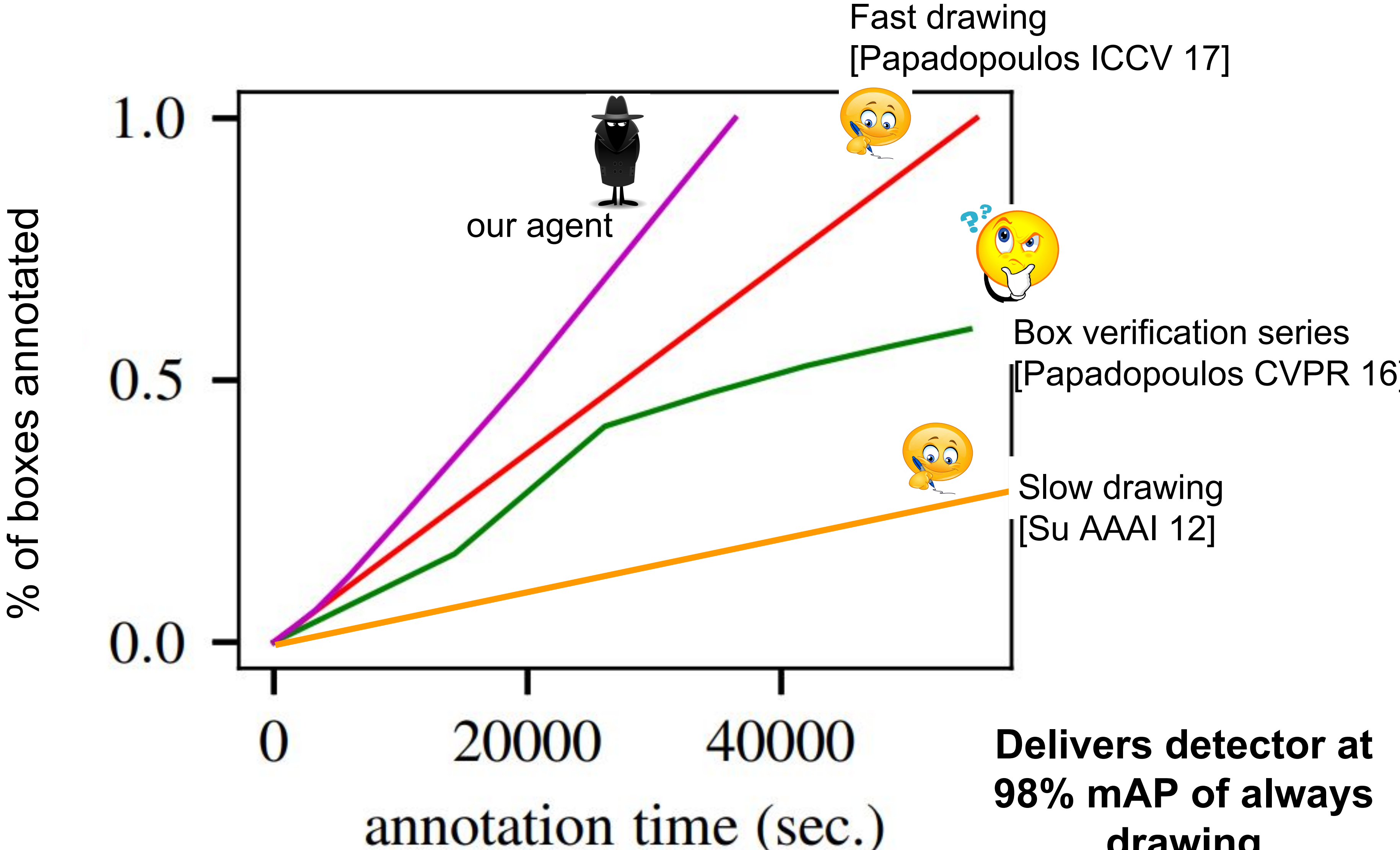
Agent is (almost) always best, *adapting* to scenario at hand

IAD with iteratively improving detector

In real annotation task we want to take advantage of the growing amount of data



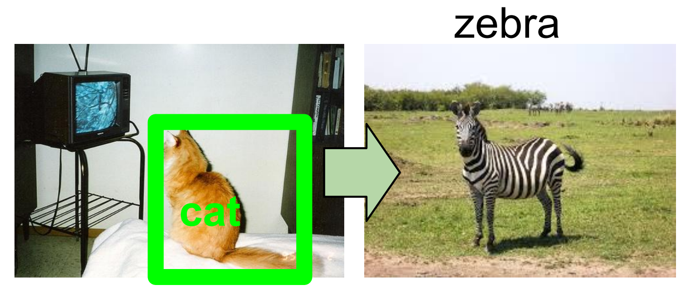
Performance at @ IoU 0.7



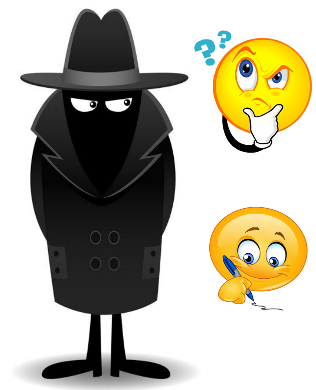
Delivers detector at 98% mAP of always drawing

This talk

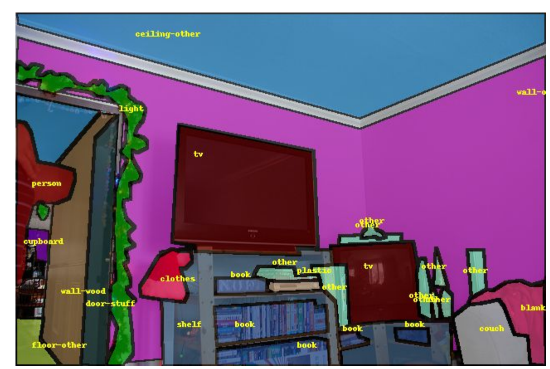
- Revisiting knowledge transfer for training object class detectors
  Uijlings, Popov, Ferrari
  CVPR 2018



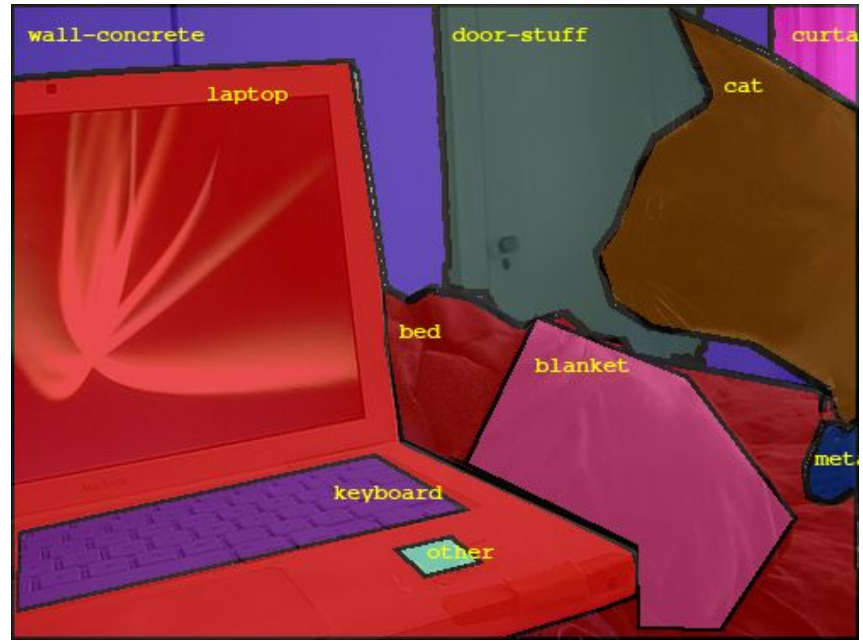
- Learning intelligent dialogs for bounding box annotation
  Konyushkova, Uijlings, Lampert, Ferrari
  CVPR 2018



- **Fluid Annotation: human-machine collaboration for full image annotation**
  Andriluka, Uijlings, Ferrari,
  arXiv June 2018



Task: Full Image Annotation



- Annotate outline and class of every object and background region
- Extremely time consuming (19 min per image for COCO)

Lin et al., Microsoft COCO: common objects in context, ECCV 2014
Caesar et al., COCO-Stuff: Things and Stuff classes in context, CVPR 2018

Traditional Annotation System (e.g. LabelMe)

Input image

Traditional Annotation System (e.g. LabelMe)

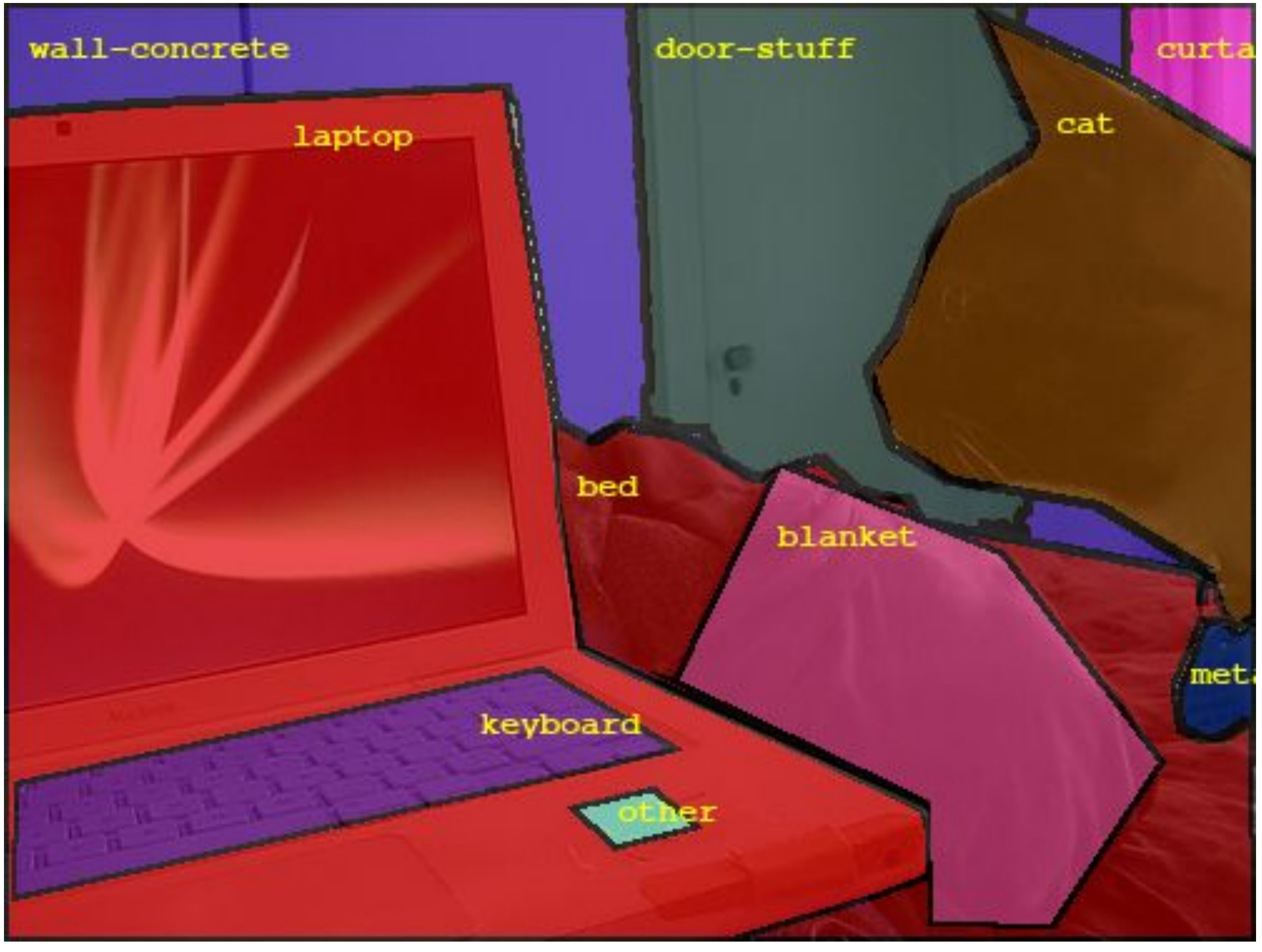
Manually draw a polygon ...

Traditional Annotation System (e.g. LabelMe)

CLICK CLICK CLICK CLICK CLICK CLICK CLICK CLICK CLICK CLICK
CLICK CLICK CLICK CLICK CLICK CLICK CLICK CLICK CLICK CLICK
CLICK CLICK CLICK CLICK CLICK CLICK CLICK CLICK CLICK CLICK
CLICK CLICK CLICK CLICK CLICK CLICK CLICK CLICK CLICK CLICK
CLICK CLICK CLICK CLICK CLICK CLICK CLICK CLICK CLICK CLICK
CLICK CLICK CLICK CLICK CLICK CLICK CLICK CLICK CLICK CLICK
CLICK CLICK CLICK CLICK CLICK CLICK CLICK CLICK CLICK CLICK
CLICK CLICK CLICK CLICK CLICK CLICK CLICK CLICK CLICK CLICK
CLICK CLICK CLICK CLICK CLICK CLICK CLICK CLICK CLICK CLICK
CLICK CLICK CLICK CLICK CLICK CLICK CLICK CLICK CLICK CLICK
CLICK CLICK CLICK CLICK CLICK CLICK CLICK CLICK CLICK CLICK
CLICK CLICK CLICK CLICK CLICK CLICK CLICK CLICK CLICK CLICK
CLICK CLICK CLICK CLICK CLICK CLICK CLICK CLICK CLICK CLICK
CLICK CLICK CLICK CLICK CLICK CLICK CLICK CLICK CLICK CLICK
CLICK CLICK CLICK CLICK CLICK CLICK CLICK CLICK CLICK CLICK
CLICK CLICK CLICK CLICK CLICK CLICK CLICK CLICK CLICK CLICK
CLICK CLICK CLICK CLICK CLICK CLICK CLICK CLICK CLICK CLICK
CLICK CLICK CLICK CLICK CLICK CLICK CLICK CLICK CLICK CLICK
CLICK CLICK CLICK CLICK ……………………….

Hundreds of mouse-clicks later...

Traditional Annotation System (e.g. LabelMe)

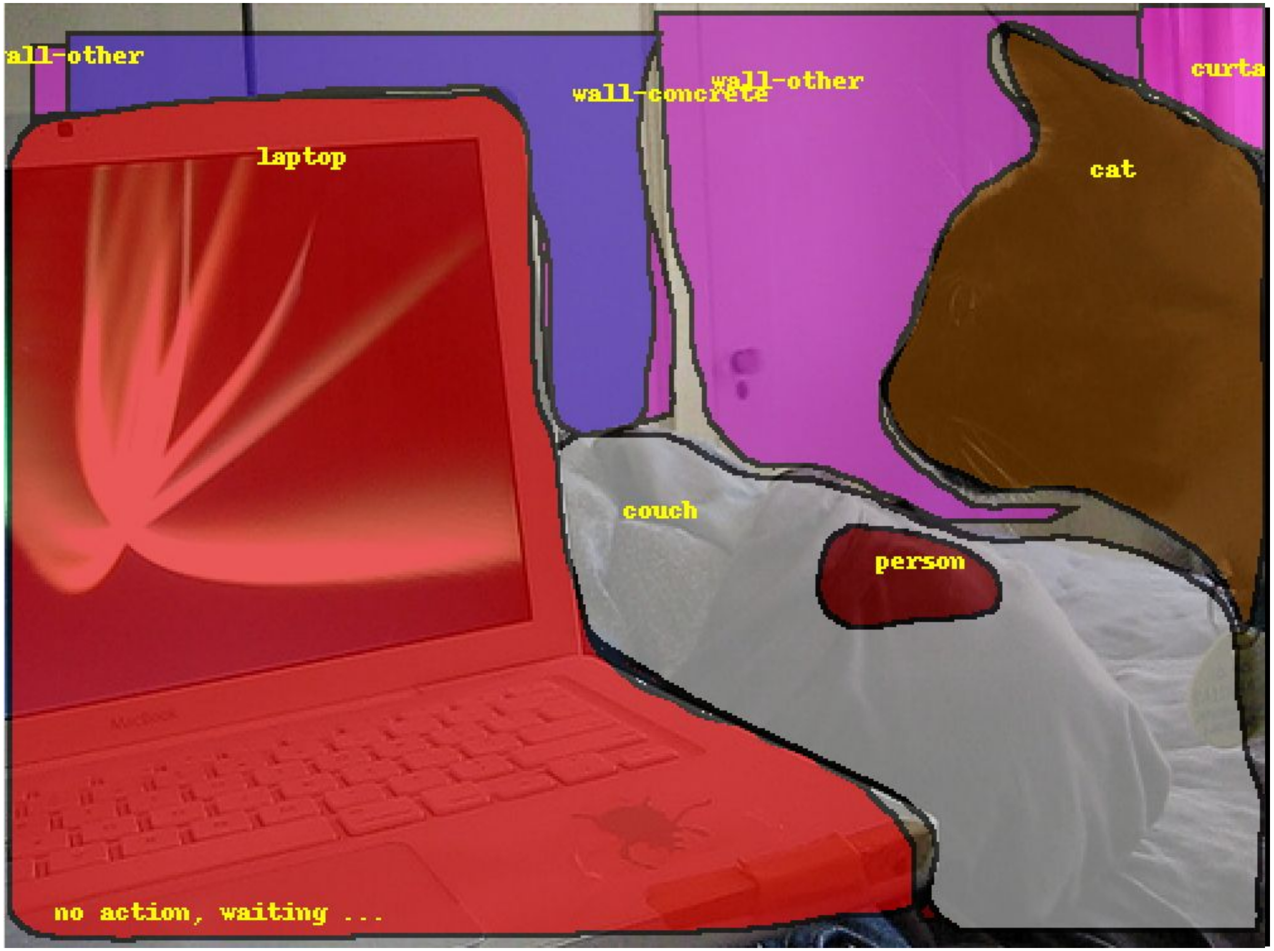


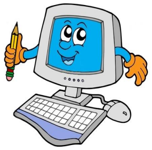
Hundreds of mouse-clicks later… Done!

Our Fluid Annotation System

Input image

Our Fluid Annotation System





Automatic initialization

Our Fluid Annotation System

tens of mouse clicks later… Done!

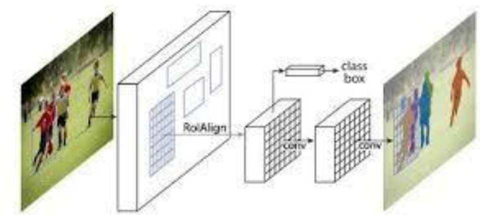
Design Principles

1. Strong Machine Learning Aid
2. Unified interface for full image annotation
3. Empower the annotator

Annotator focuses on what the machine *does not already know*

Method

apply Mask R-CNN



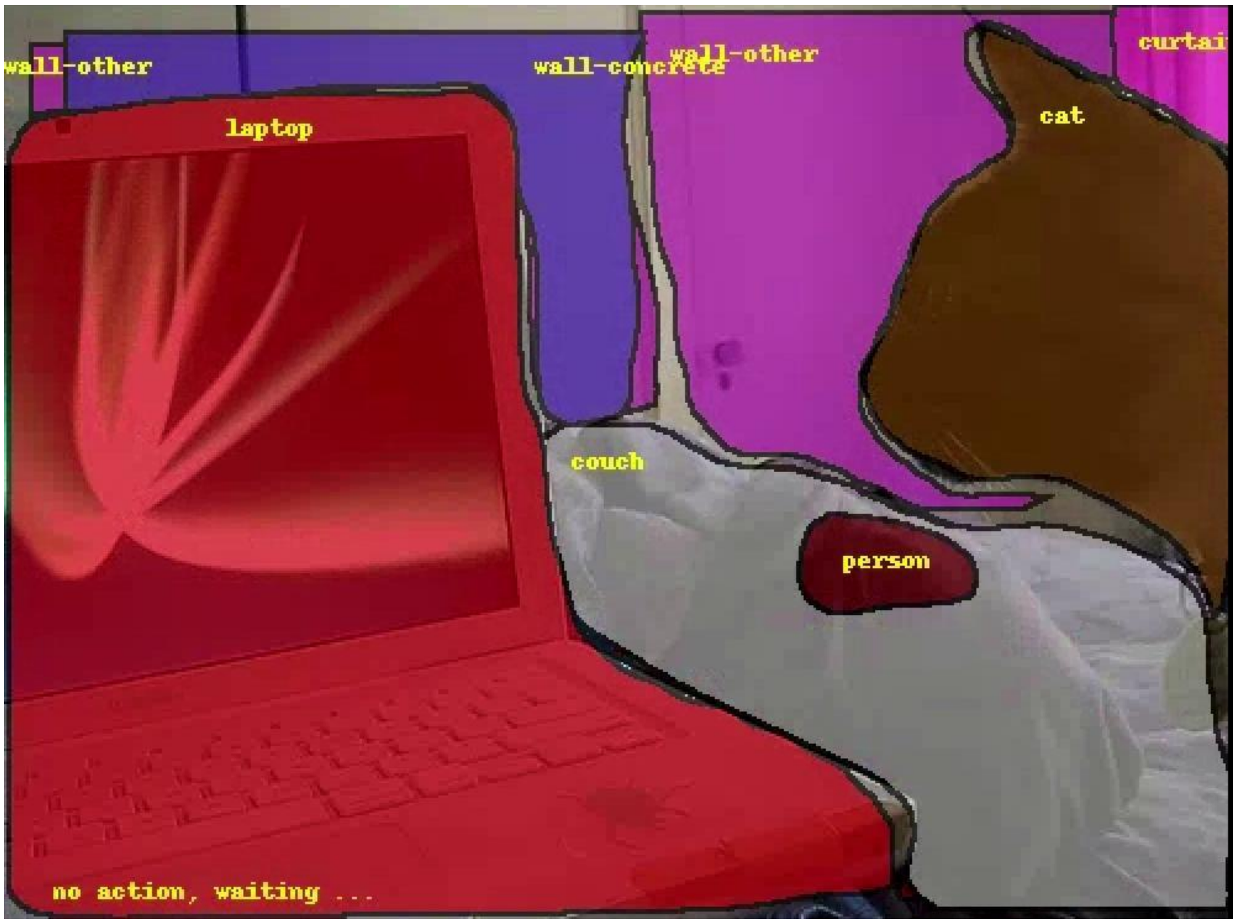
[He ICCV 2017]

Method

Segments (~1000) with Labels and detection score

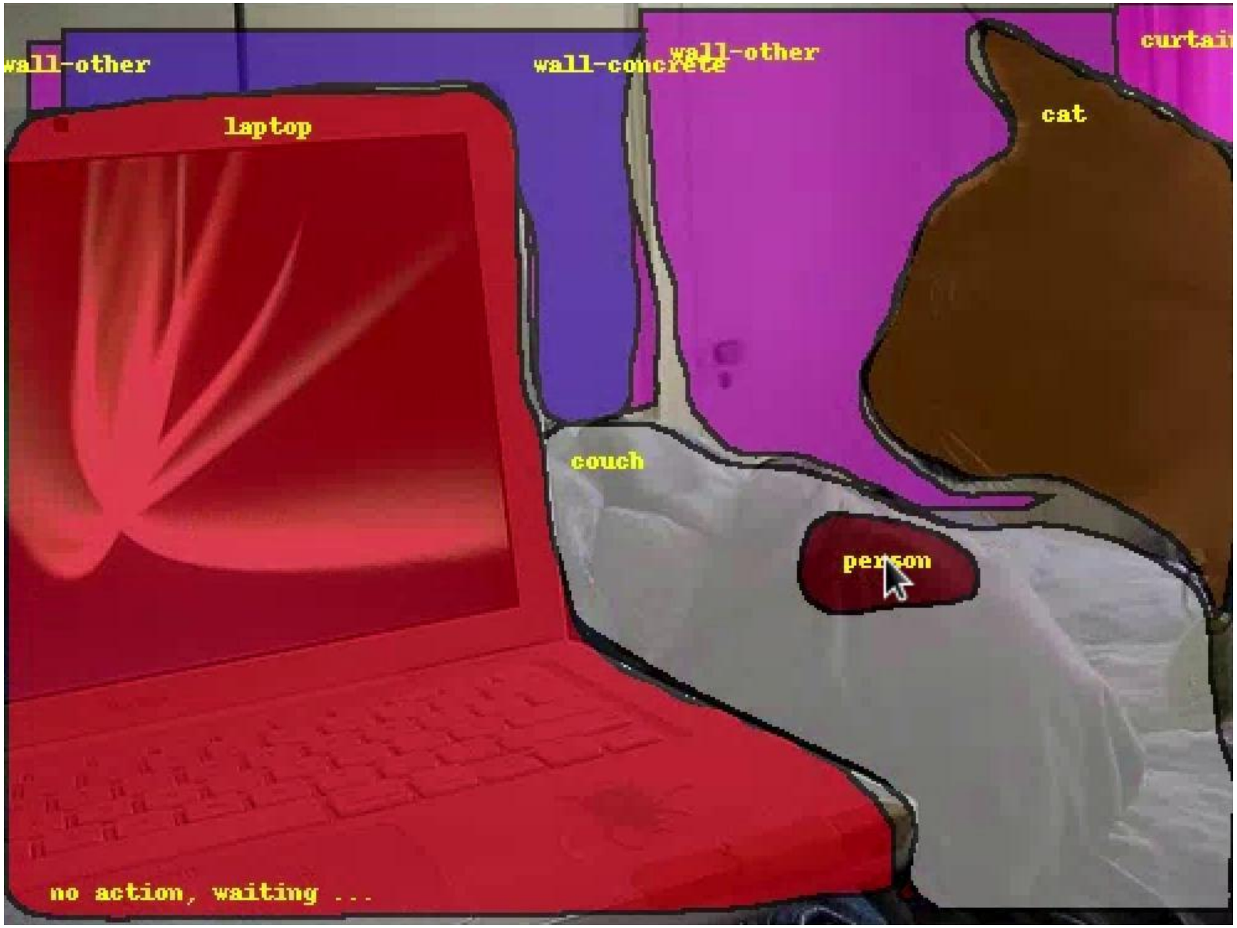
Automatic Initialization

Most likely interpretation of scene

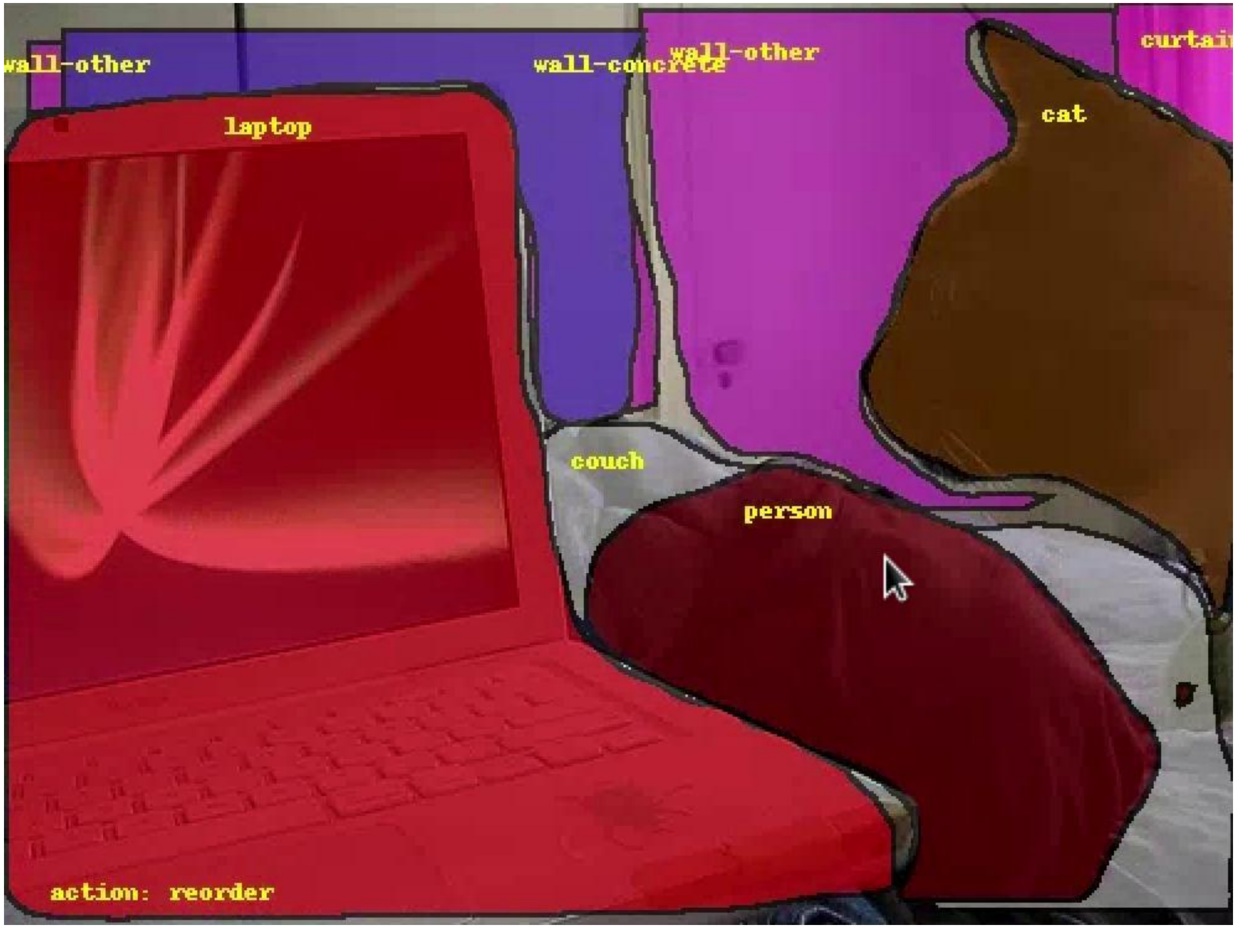


Present to annotator in simple interface

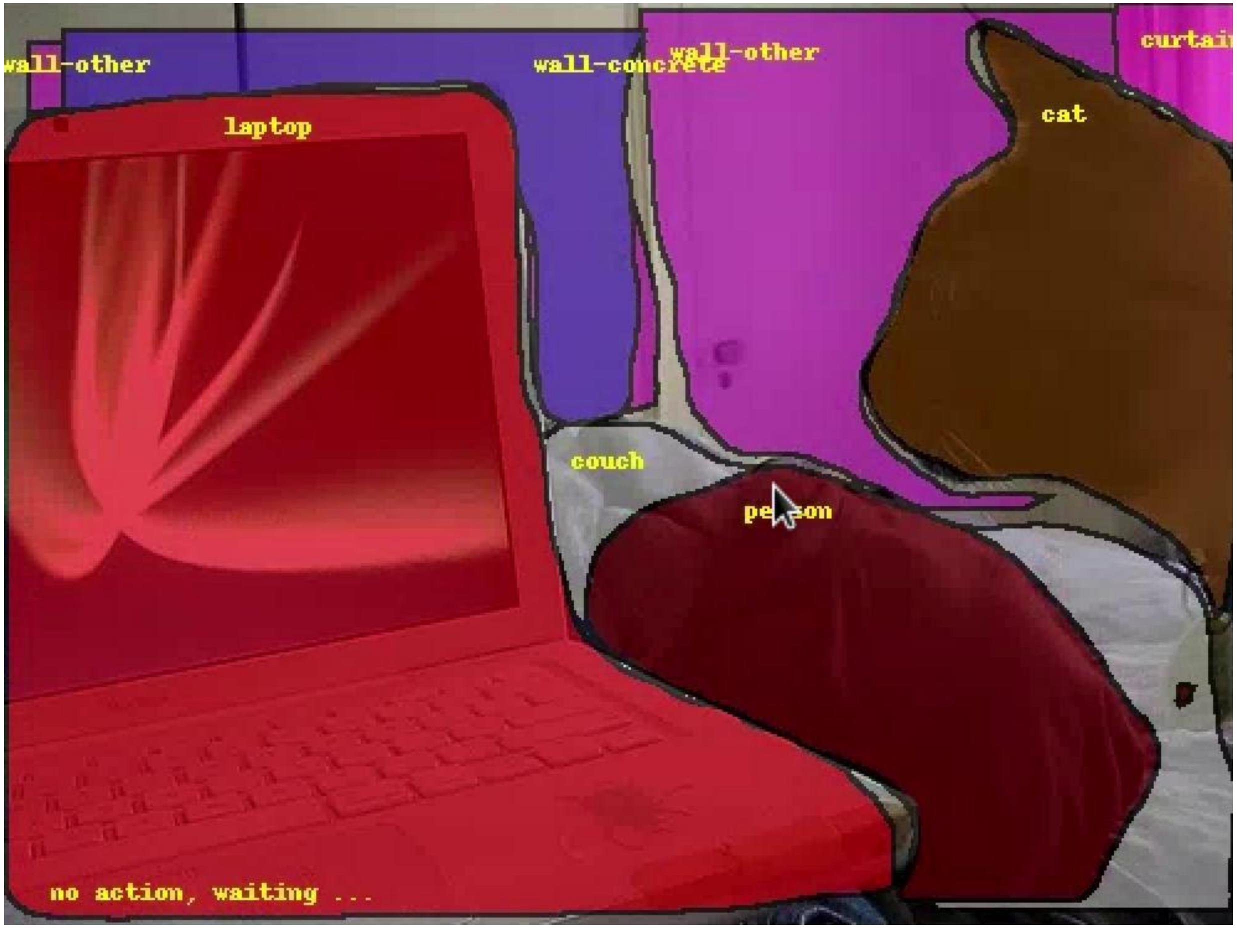
“Reorder” Action



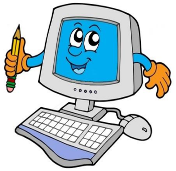
"Reorder" Action



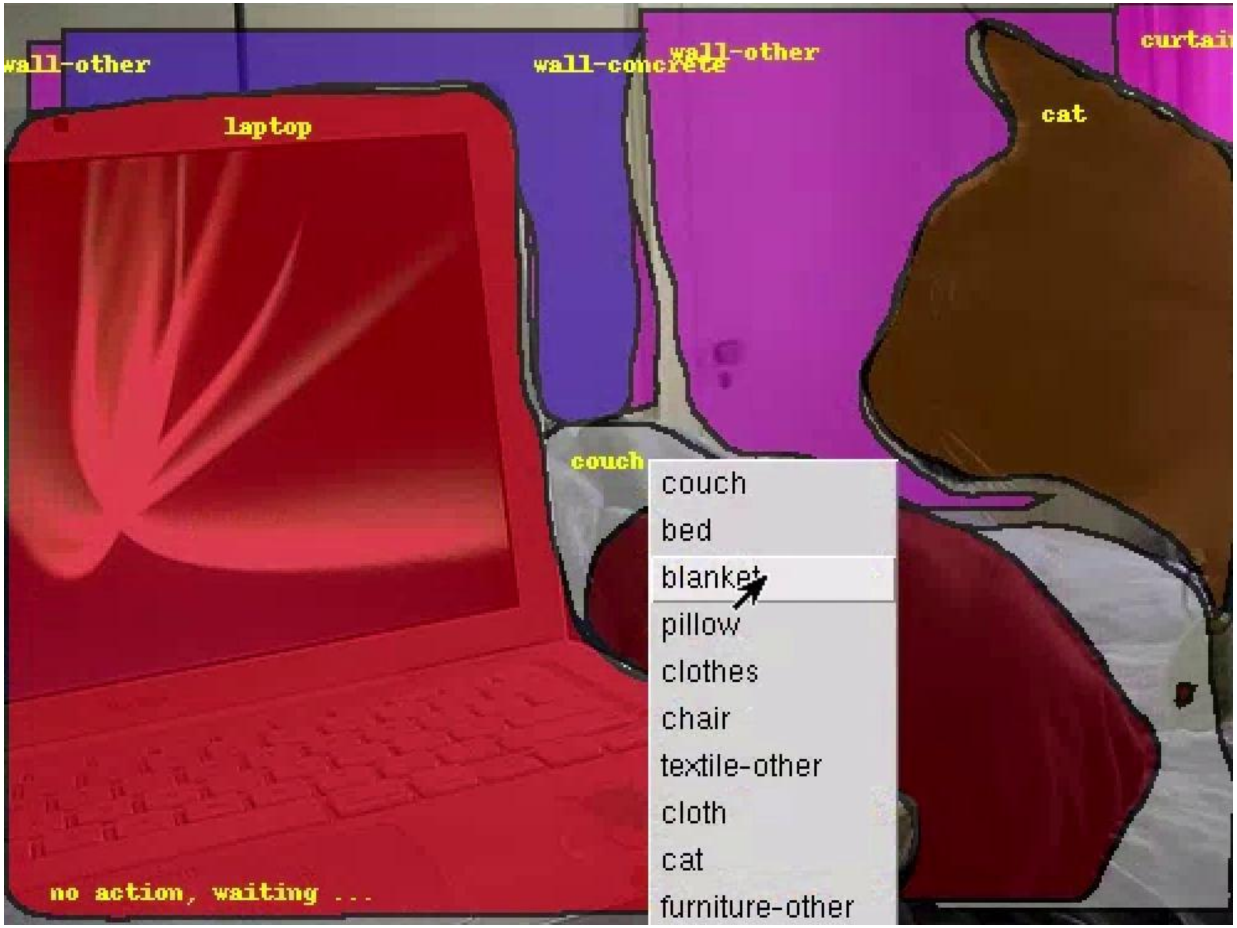
“Change label” Action

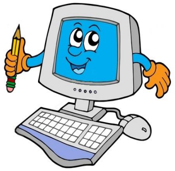


“Change label” Action

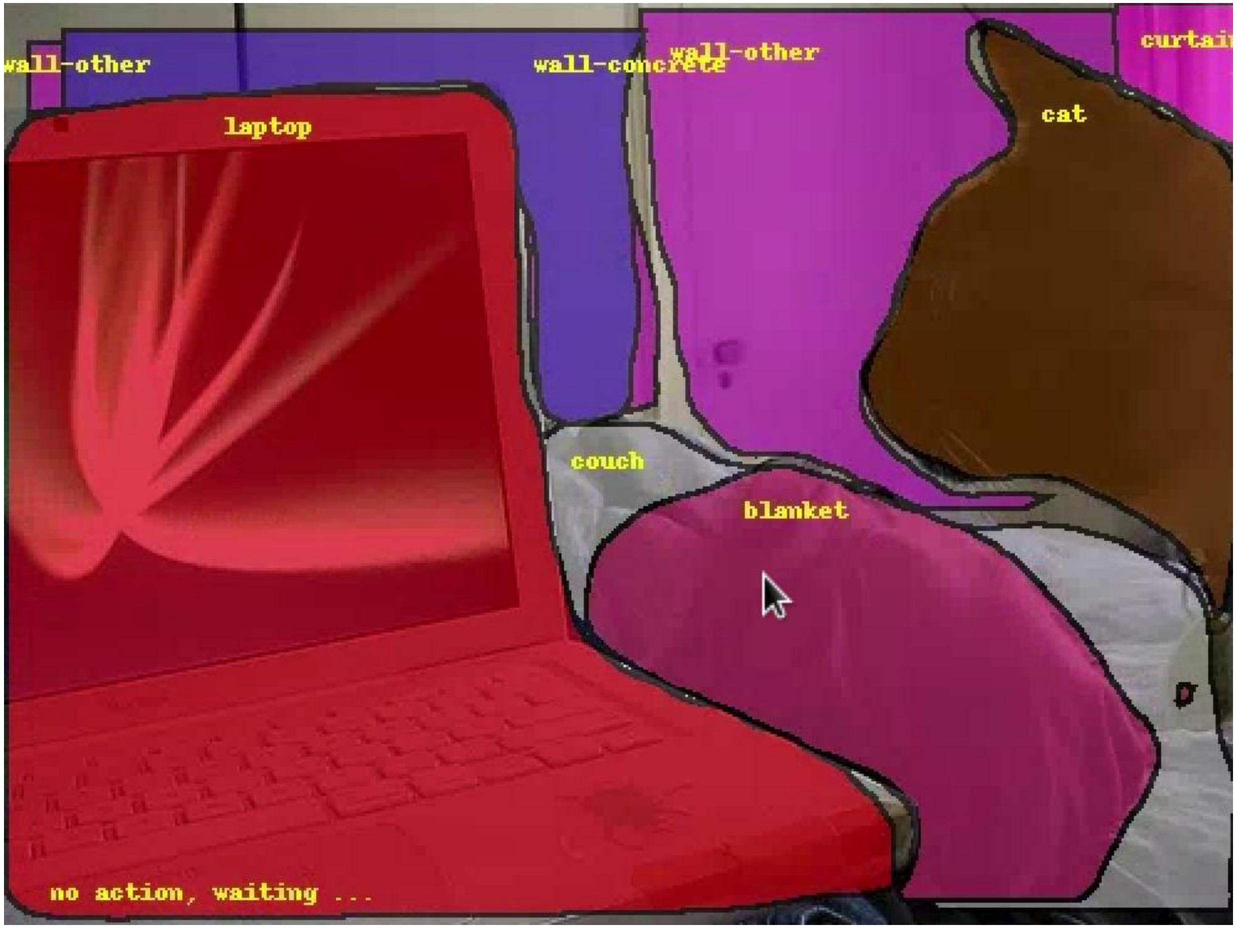


“Change label” Action

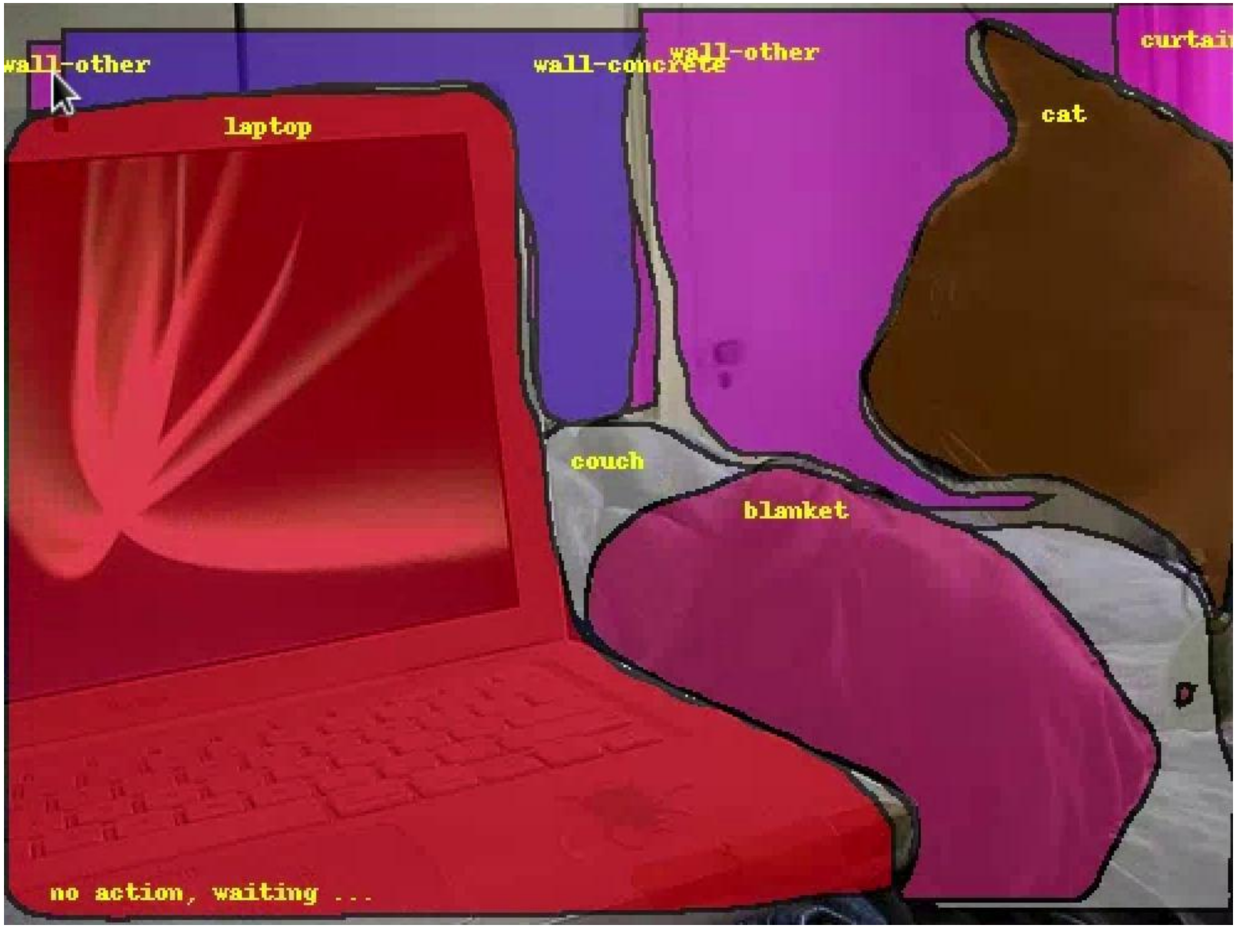




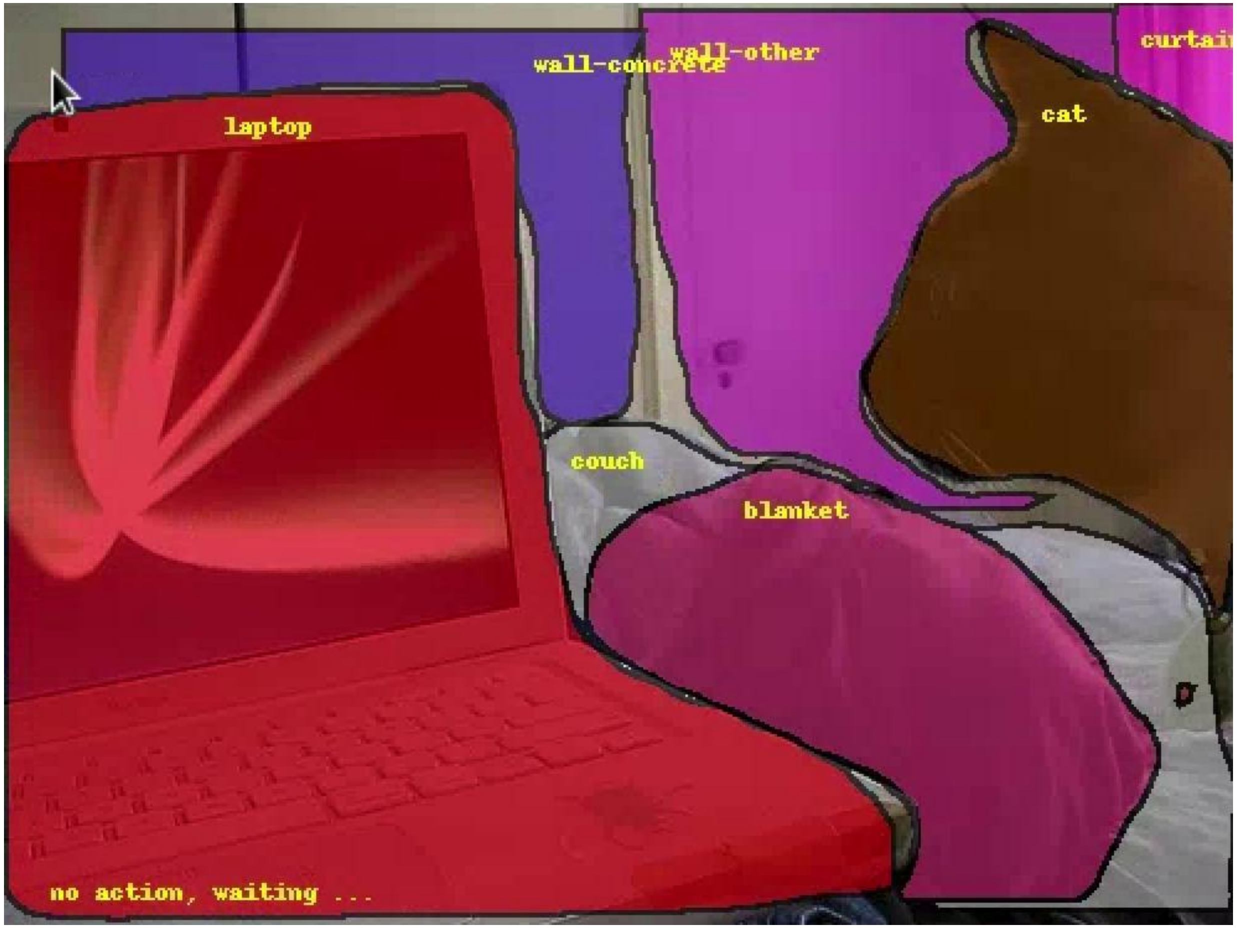
“Change label” Action



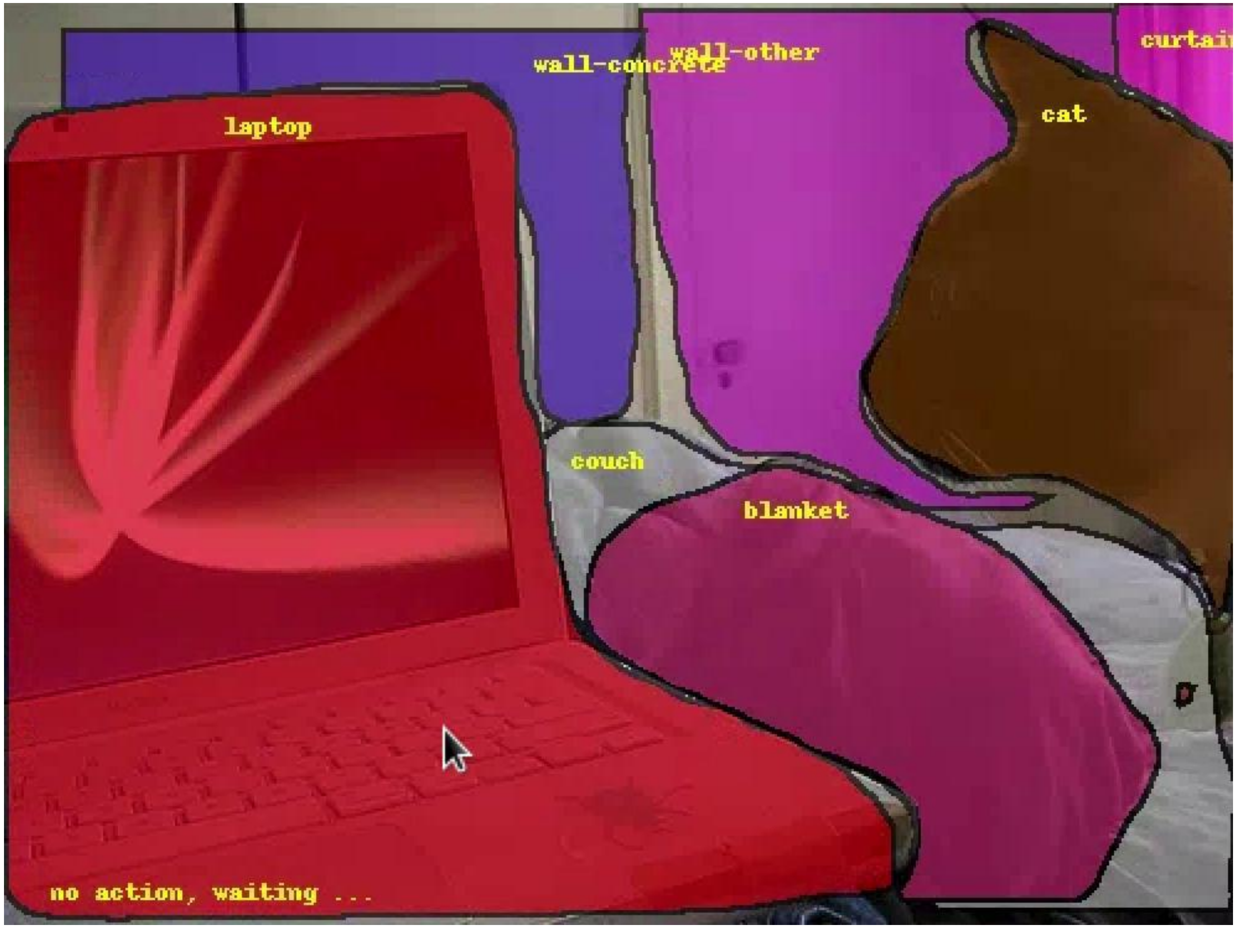
“Remove” Action



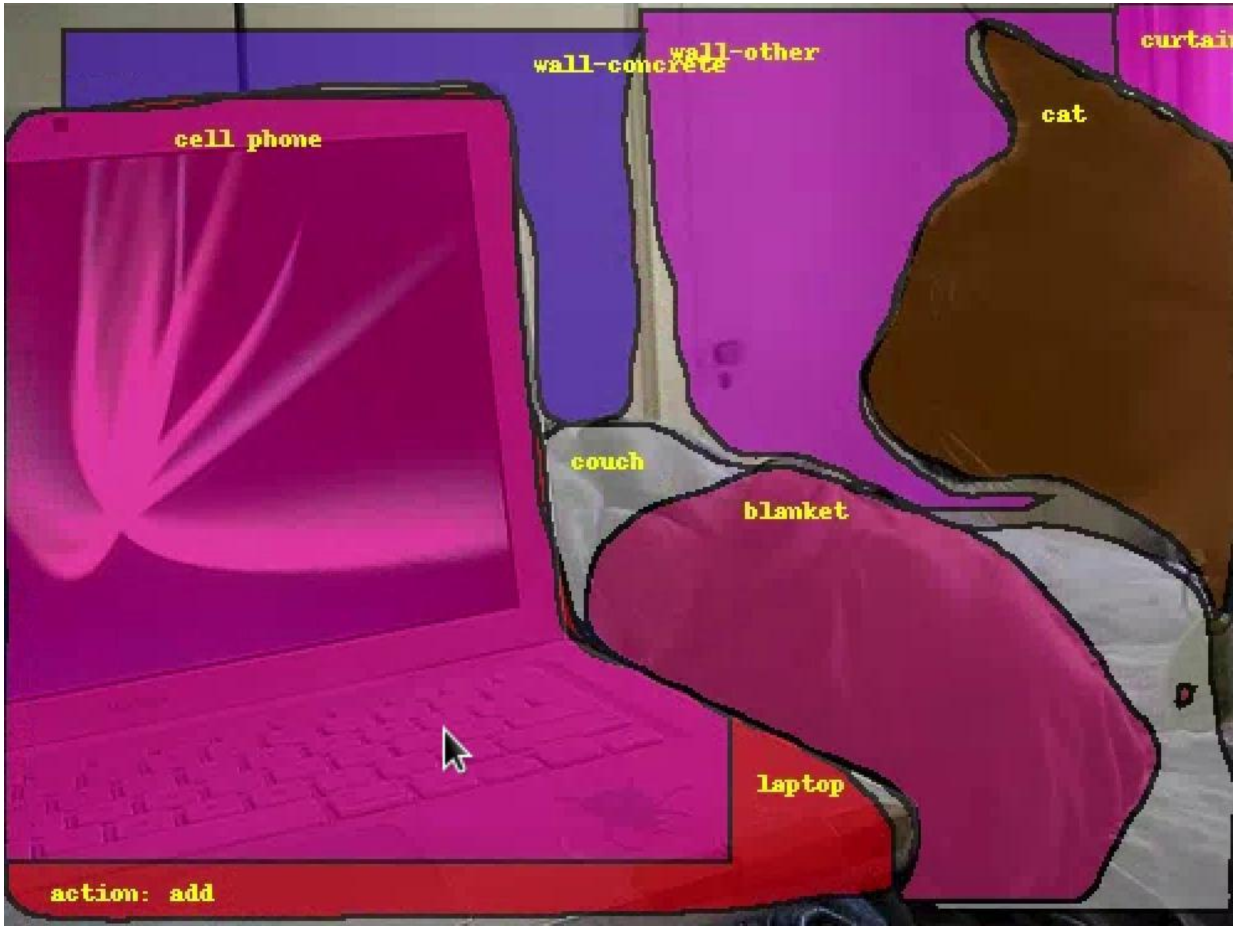
“Remove” Action

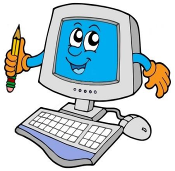


“Add” Action

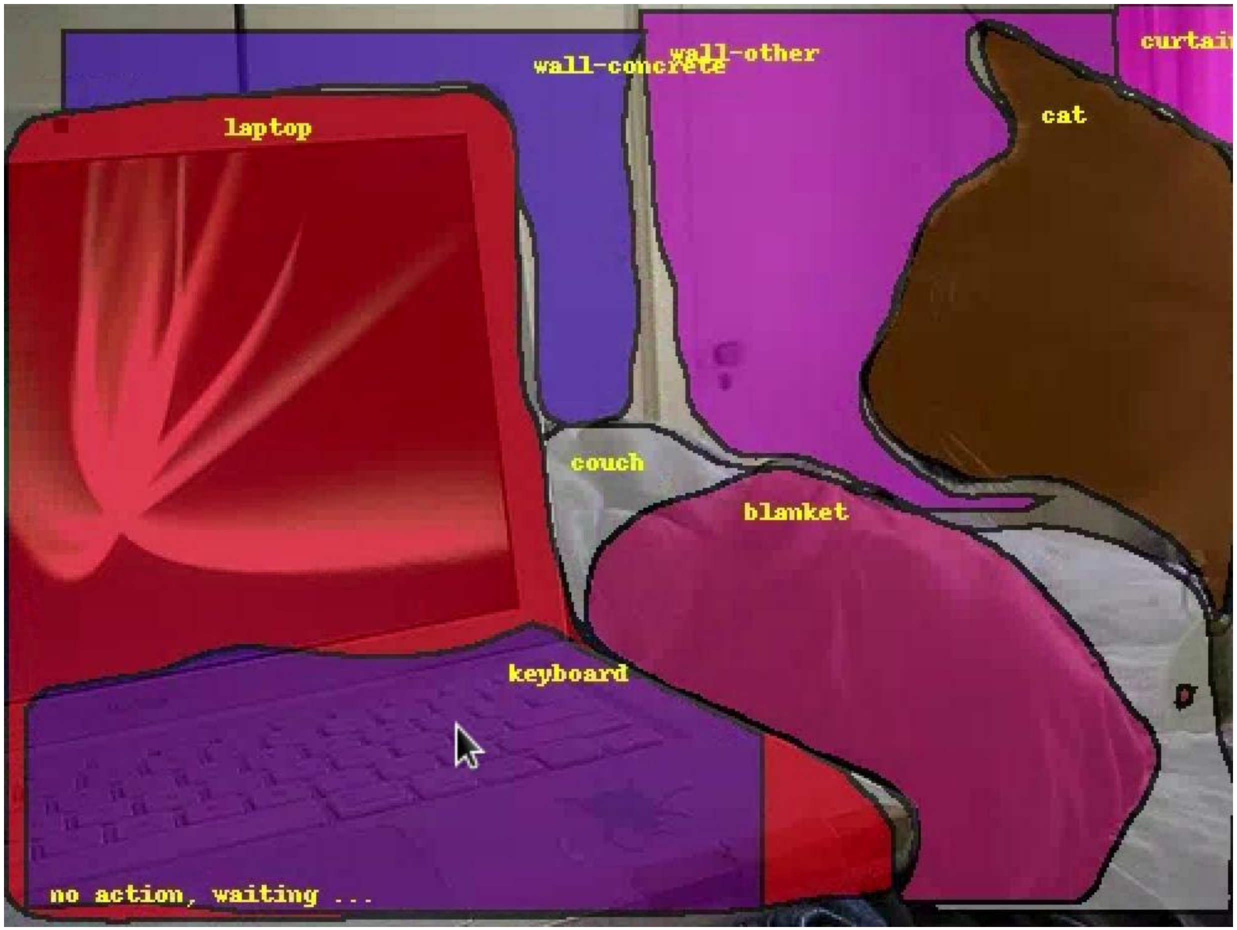

wall-concrete
wall-other
curtain
laptop
cat
couch
blanket
no action, waiting ...

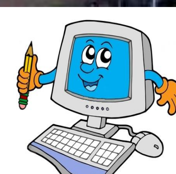
"Add" Action


wall-concrete
wall-other
curtain
cell phone
cat
couch
blanket
laptop
action: add

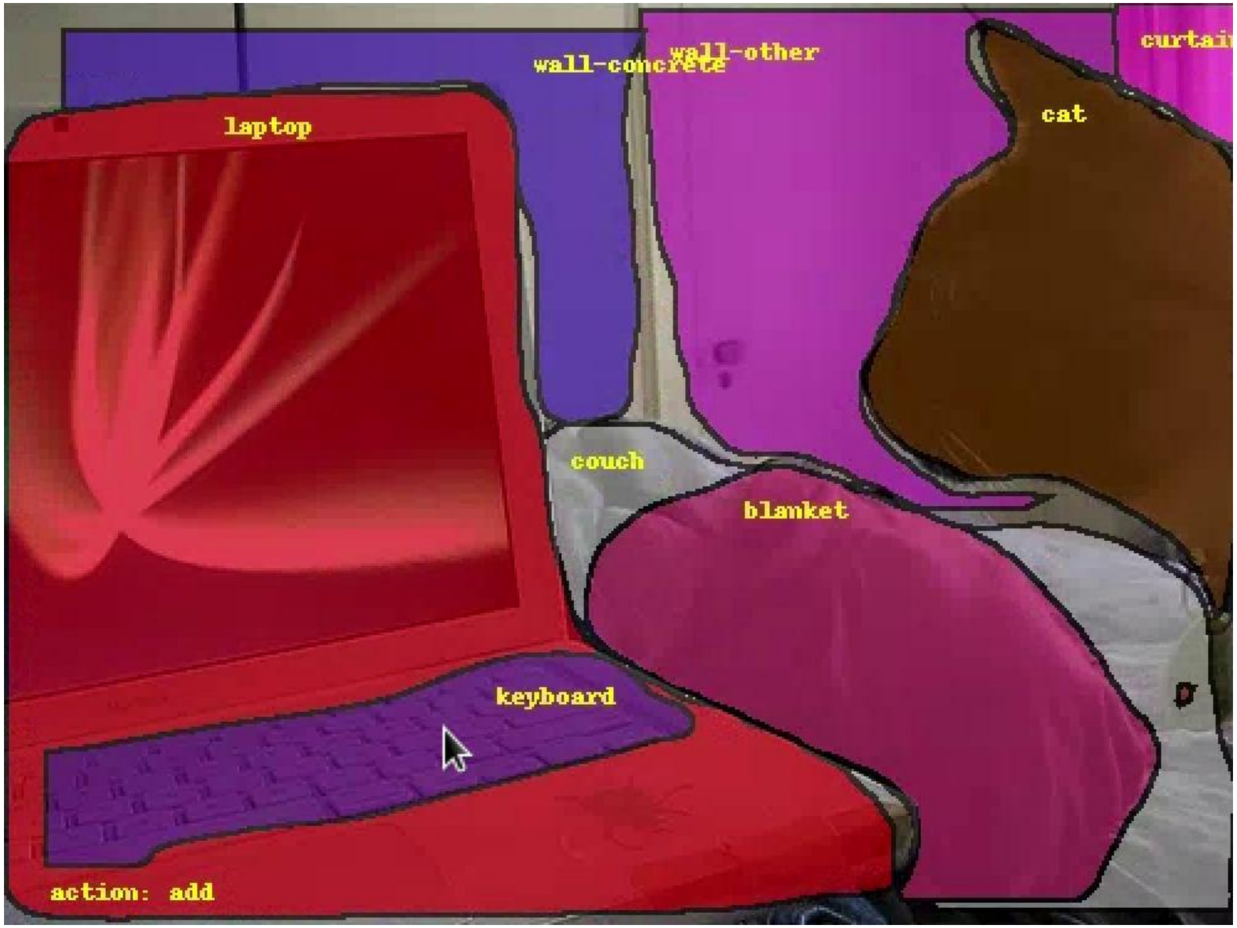


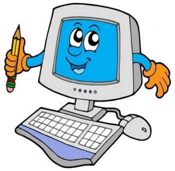
“Add” Action


wall-concrete
wall-other
laptop
cat
couch
blanket
keyboard
no action, waiting ...



“Add” Action

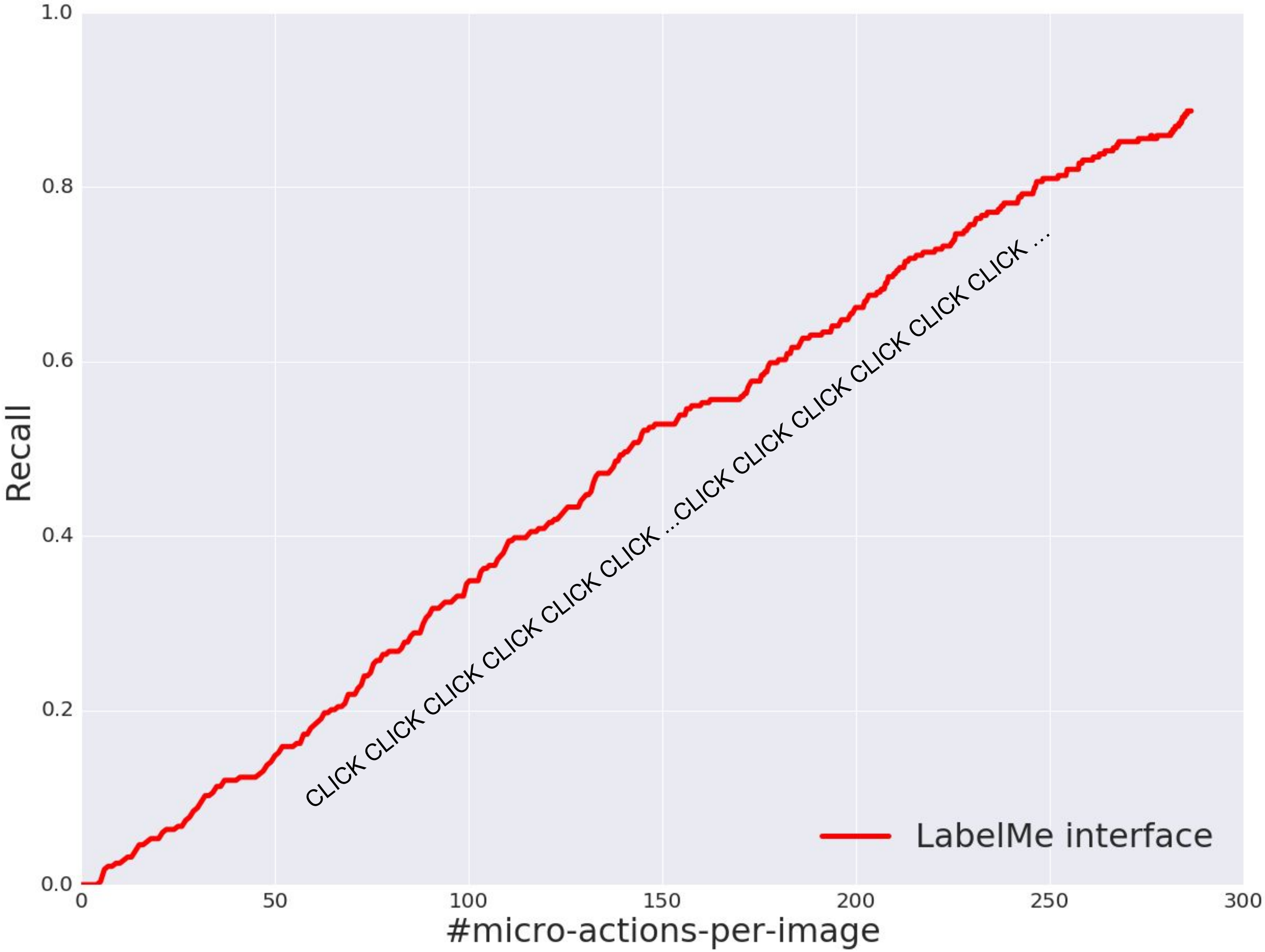

wall-concrete
wall-other
curtain
laptop
cat
couch
blanket
keyboard
action: add



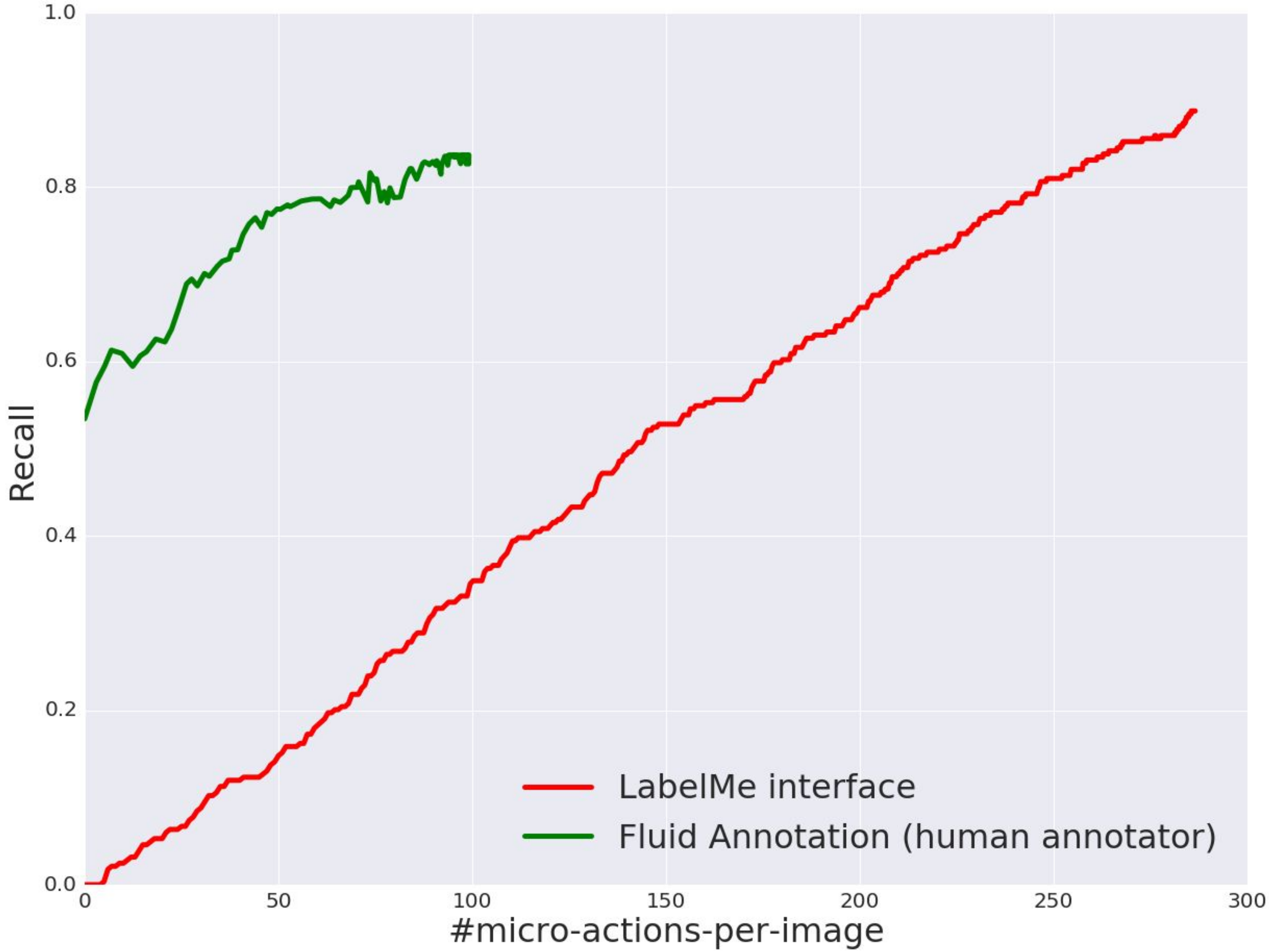
Final Result

Example Annotation Result

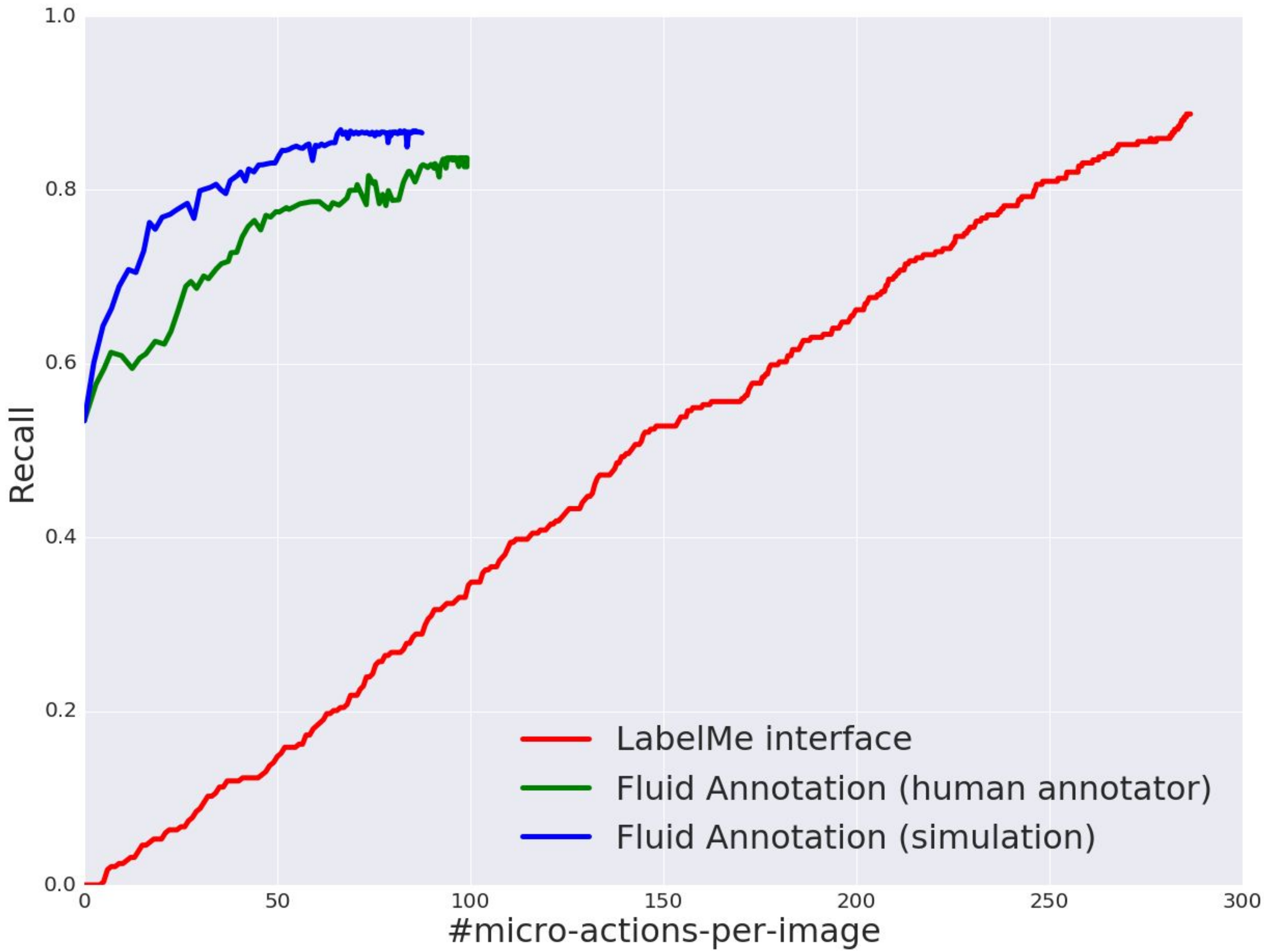
Results: Human Annotators



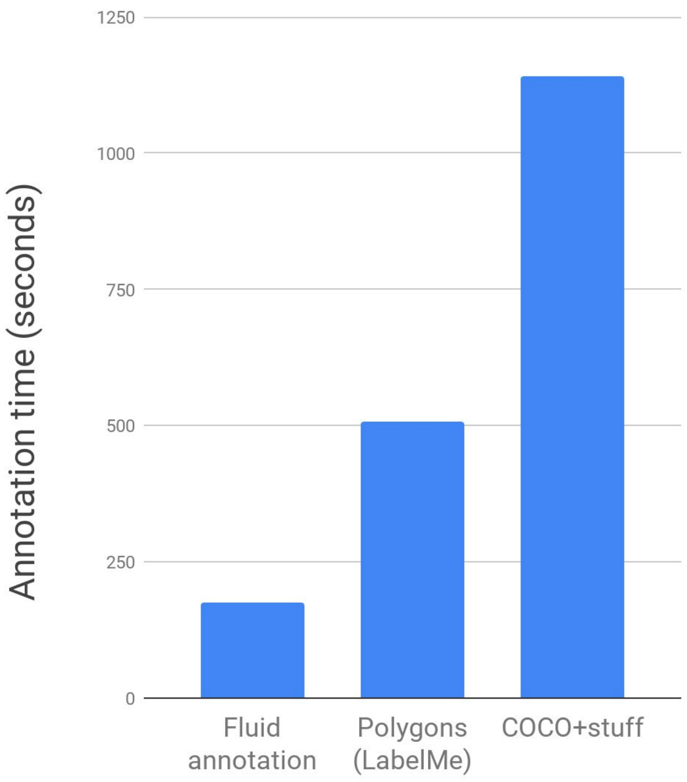
Results: Human Annotators

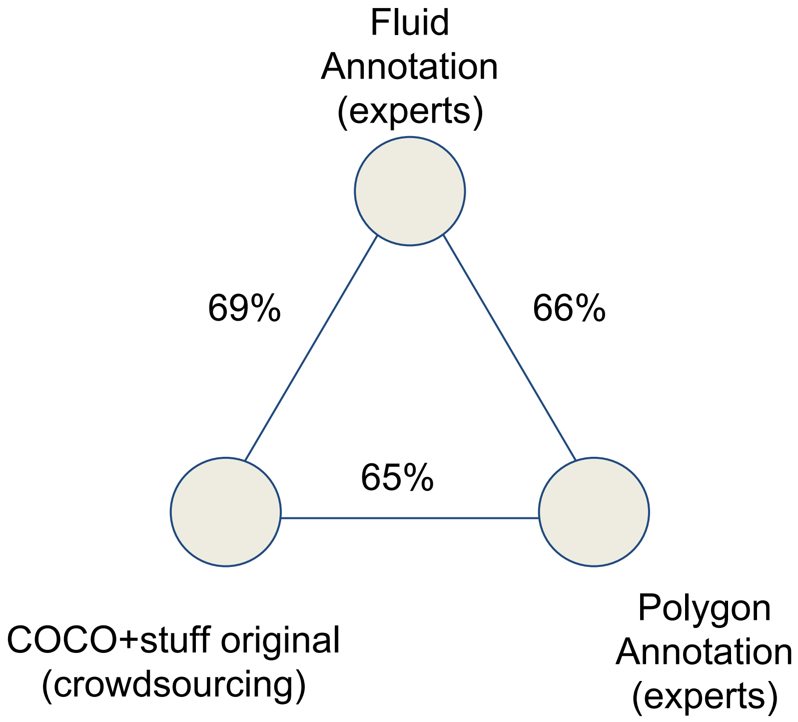


Results: Human Annotators



Annotation time and Label Agreement





Lin et al., Microsoft COCO: common objects in context, ECCV 2014

Caesar et al., COCO-Stuff: Things and Stuff classes in context, CVPR 2018

Example Results

COCO+Stuff original

Fluid Annotation

Polygons

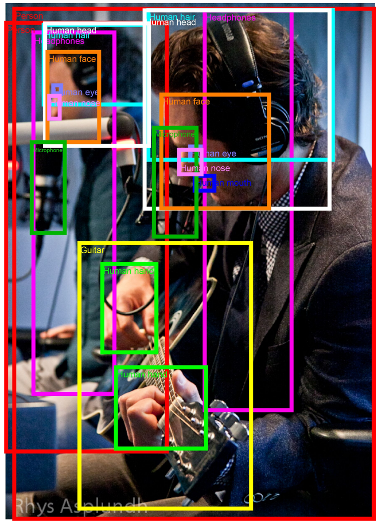
Example Results

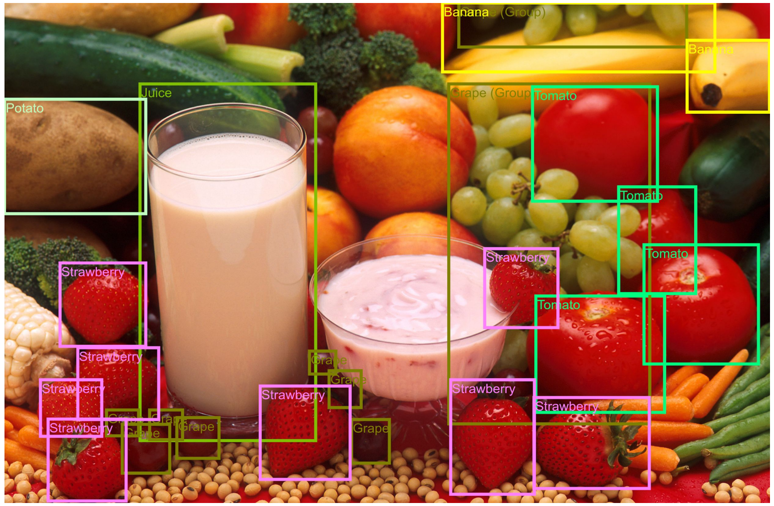
COCO+Stuff original

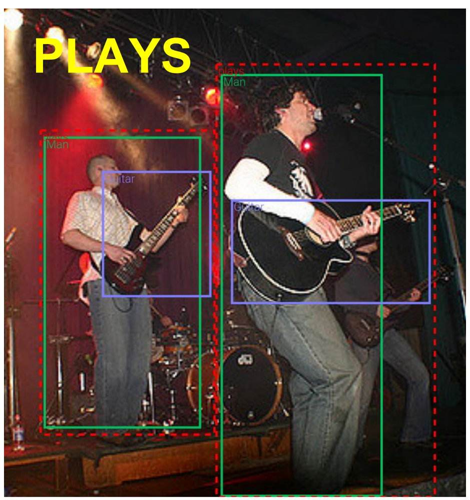
Fluid Annotation

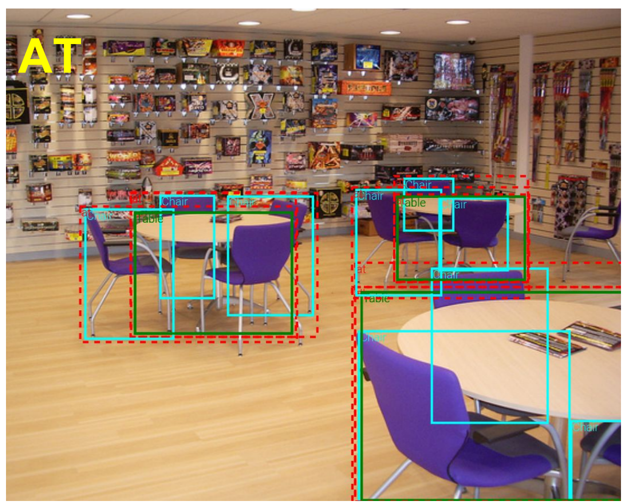
Polygons

Open Images V4 and Challenge









- 600 object classes
- **15.4M** bounding-boxes on 1.9M images
- 10x over existing datasets
- Complex images (average 8 boxes)
- Visual Relationship Detection annotations
- Challenge at ECCV 2018

https://storage.googleapis.com/openimages/web/index.html