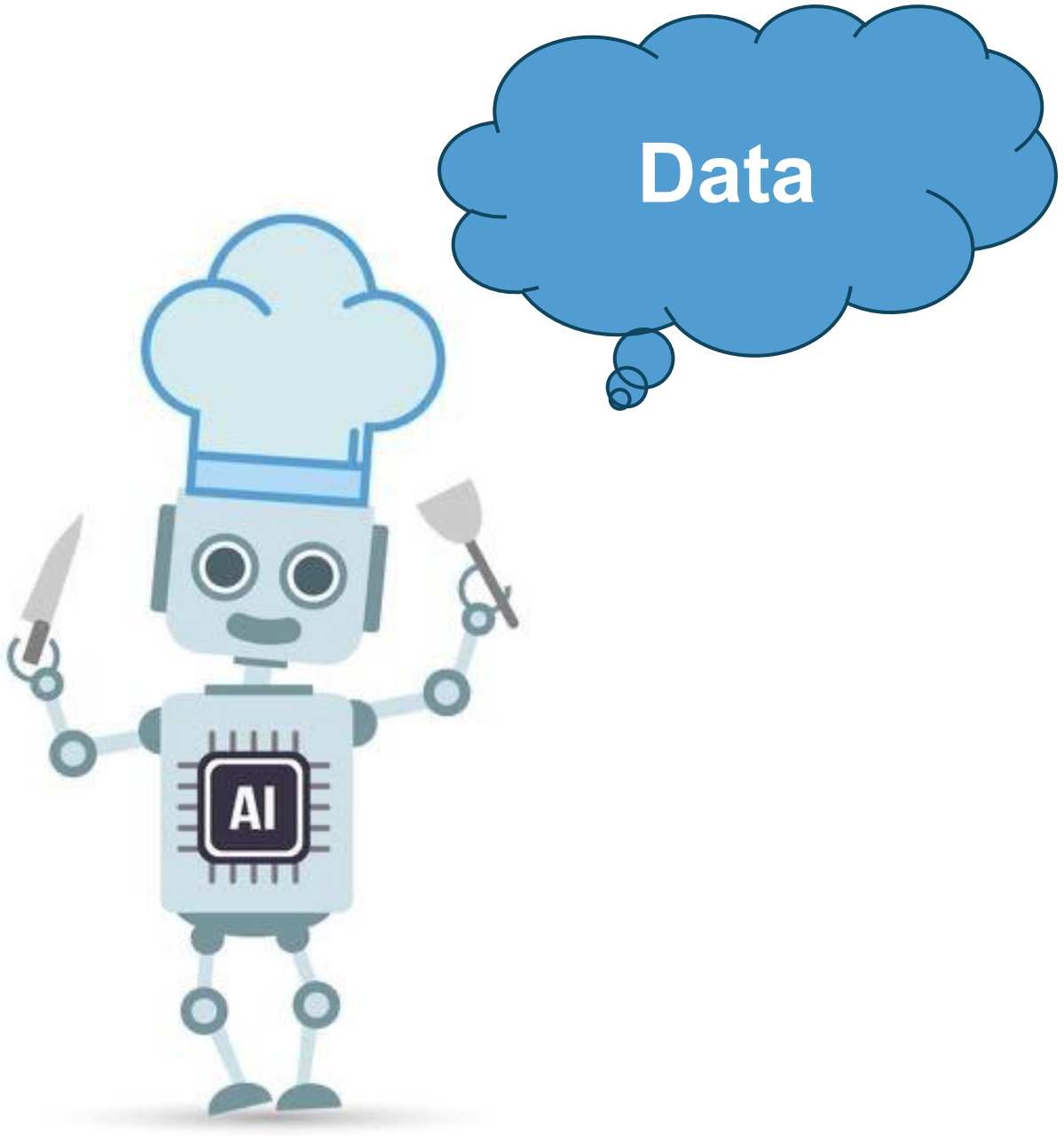
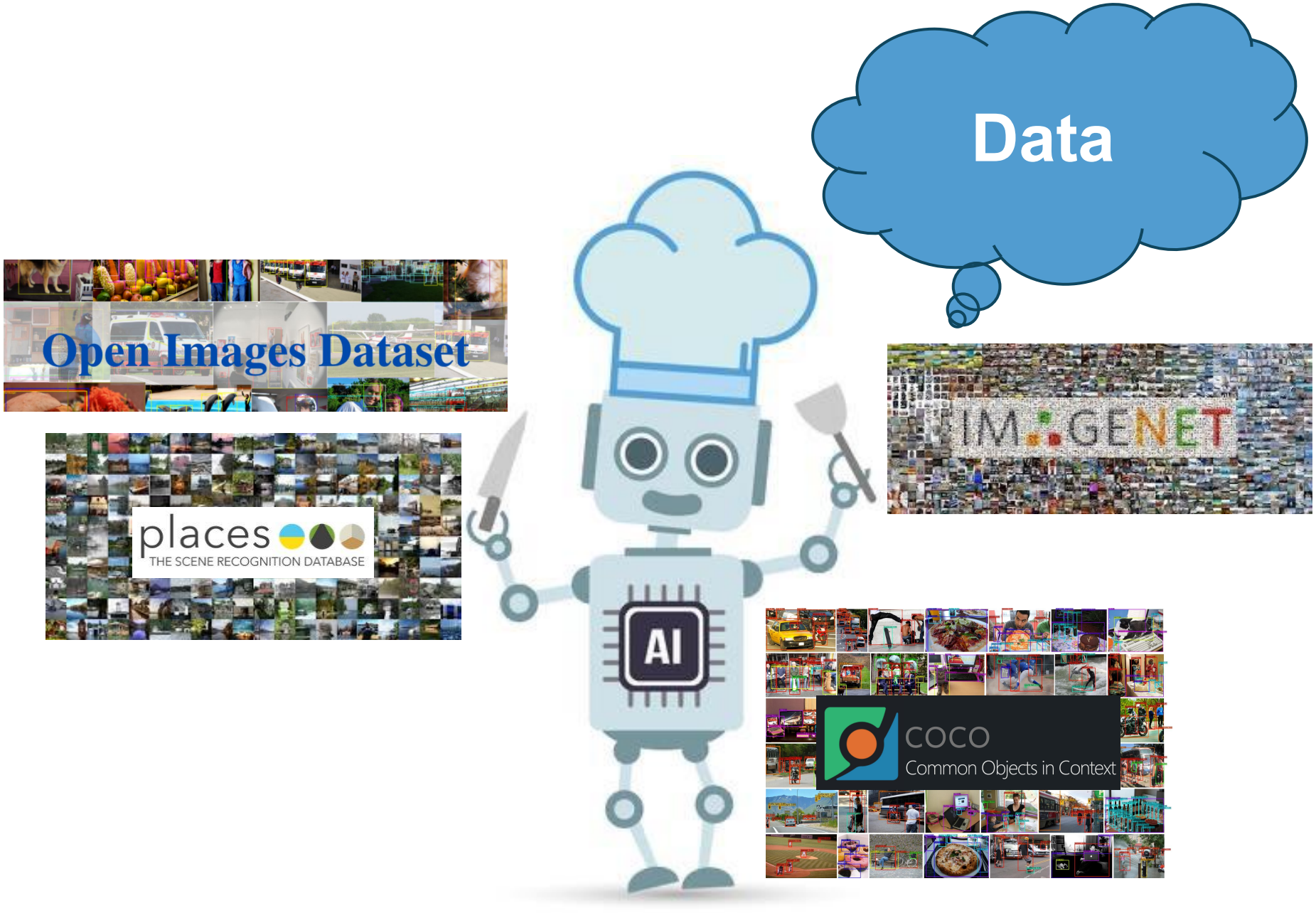
Semantic segmentation with limited labels

Dimitrios Papadopoulos
Associate Professor, DTU Compute
Edited by: J. Miguel Valverde
Postdoc, DTU Compute

Data-hungry deep learning



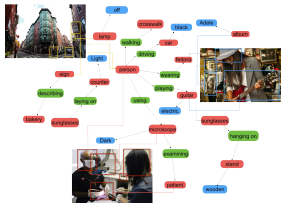
Data-hungry deep learning

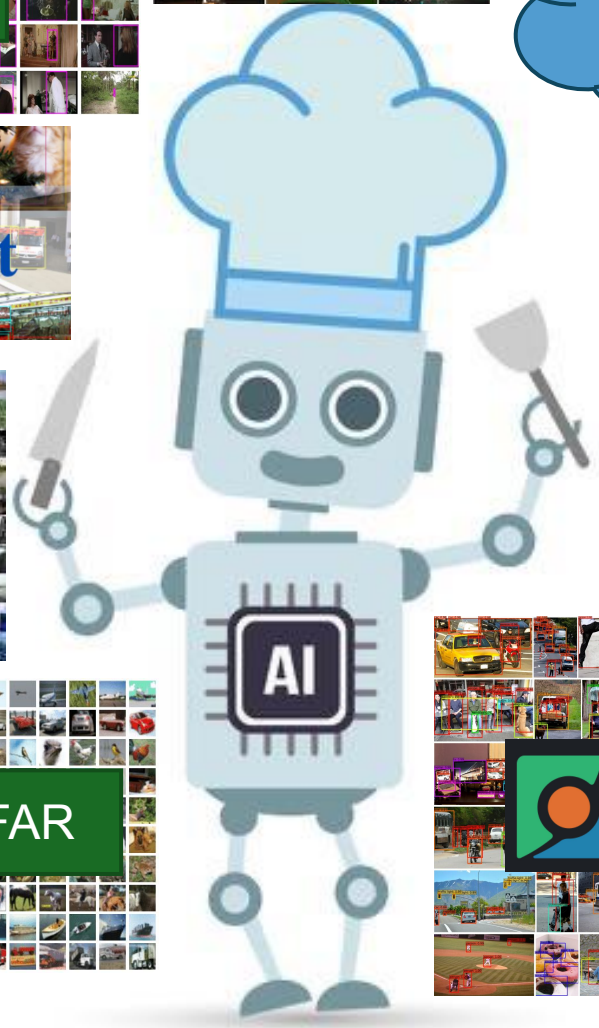


Data-hungry deep learning

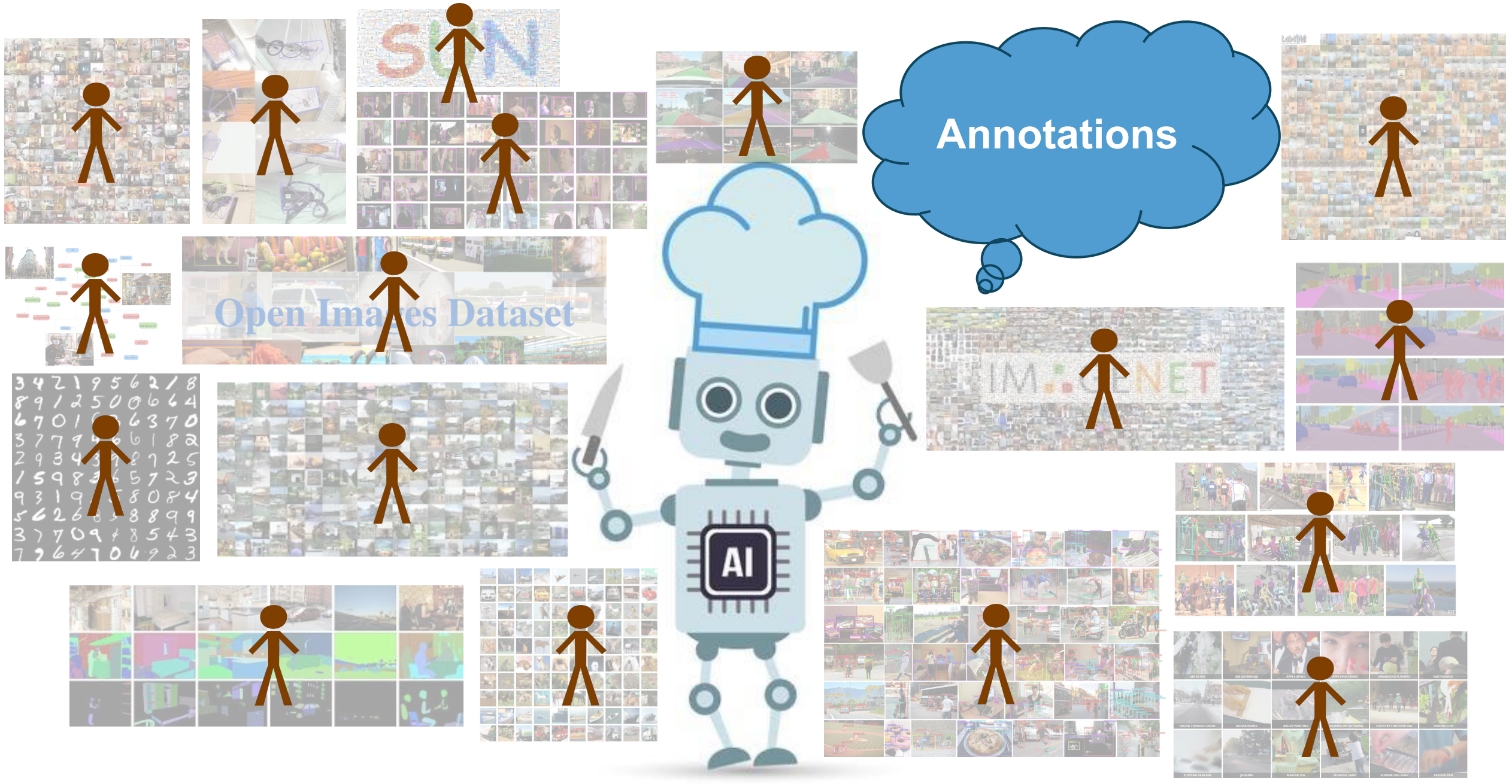
Data-hungry deep learning

Even more data !!!

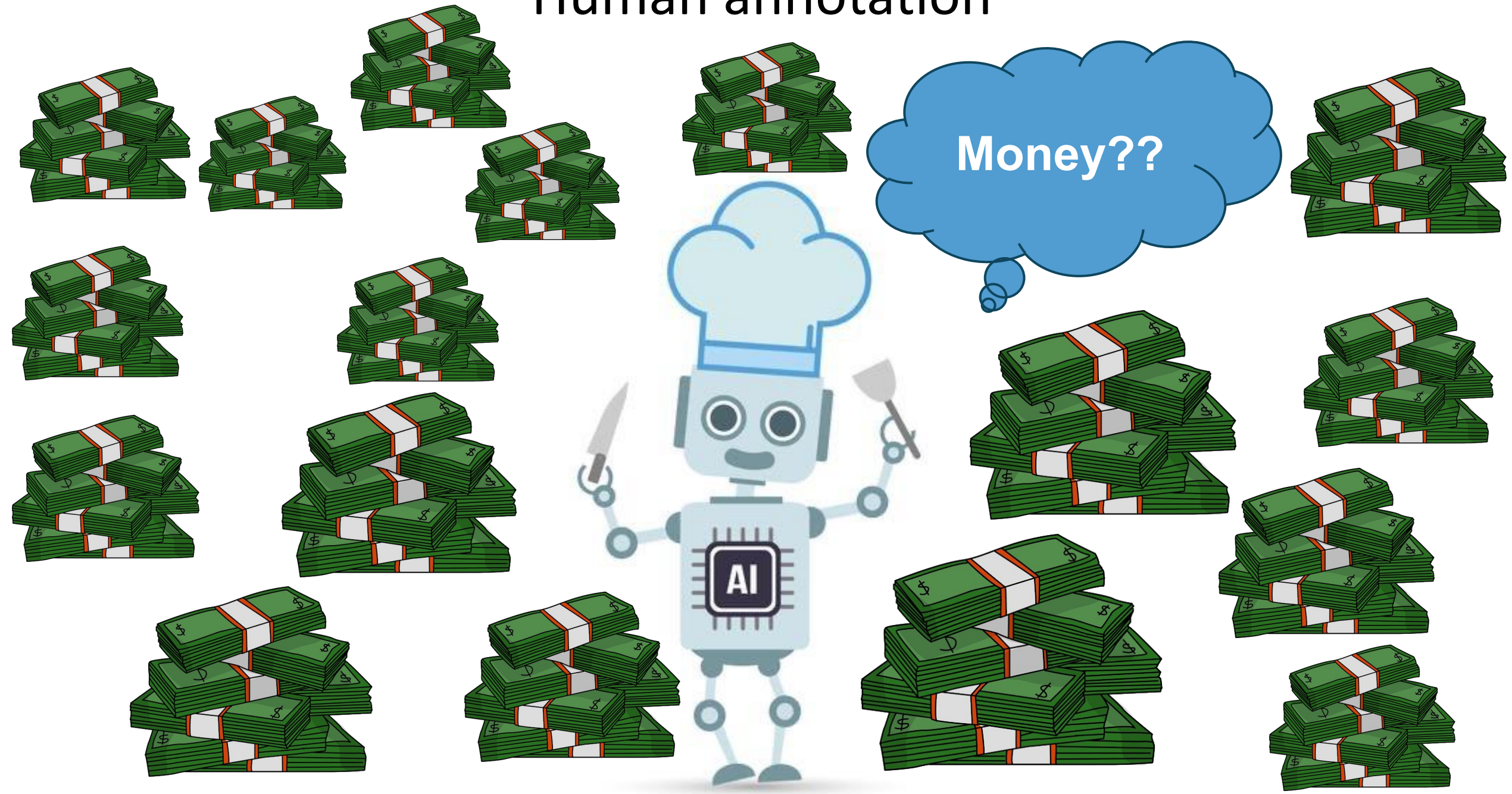




Human annotation

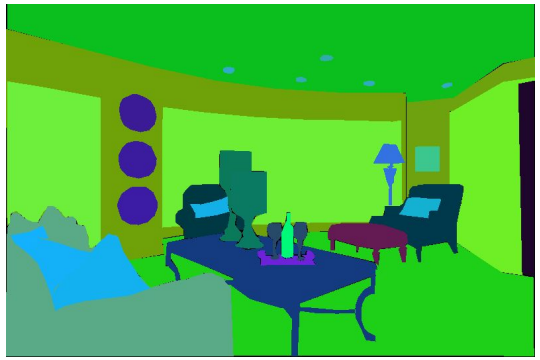


Human annotation

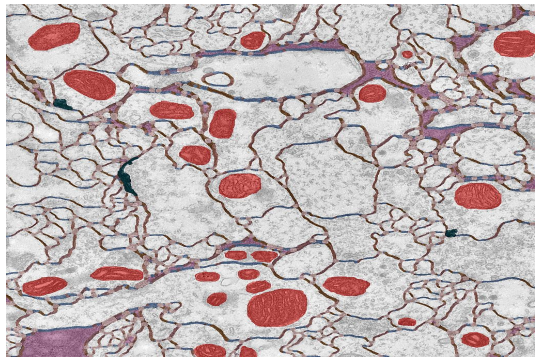


Manual annotations for computer vision applications

Scene understanding

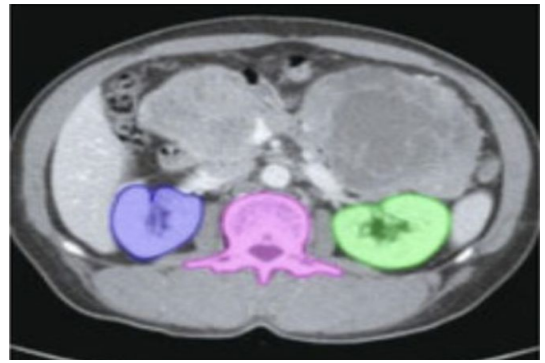


Microscopy

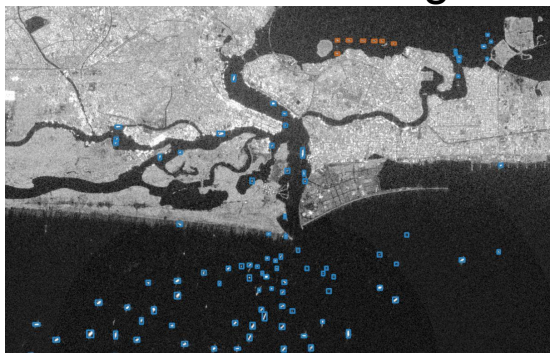


Autonomous driving

Medical imaging



Remote sensing



Biodiversity/ Ecology

Cityscapes dataset

Cityscapes dataset

A common pipeline

- Choose your favorite deep learning architecture
- Load pretrained weights from models trained on ImageNet
- Label/Annotate some data
- Finetune your model

A common pipeline - Problems

- Choose your favorite deep learning architecture
- Load pretrained weights from models trained on ImageNet
- Label/Annotate some data
- Finetune your model

(A) Transfer learning – ImageNet weights

(B) Limited budget – Limited task labels

(B) Limited Data - Solutions

Human Machine Collaboration

Label propagation

Synthetic data

Semi-supervised learning

Active Learning

Transfer learning

Weakly-supervised learning

Few-shot learning

Interactive annotation

Unsupervised learning

Human in the loop

Domain adaptation

Self-supervised learning

Multimodal Learning

(B) Limited Data - Solutions

Human Machine Collaboration

Label propagation

Synthetic data

Semi-supervised learning

Active Learning

Transfer learning

Weakly-supervised learning

Few-shot learning

Interactive annotation

Unsupervised learning

Human in the loop

Domain adaptation

Self-supervised learning

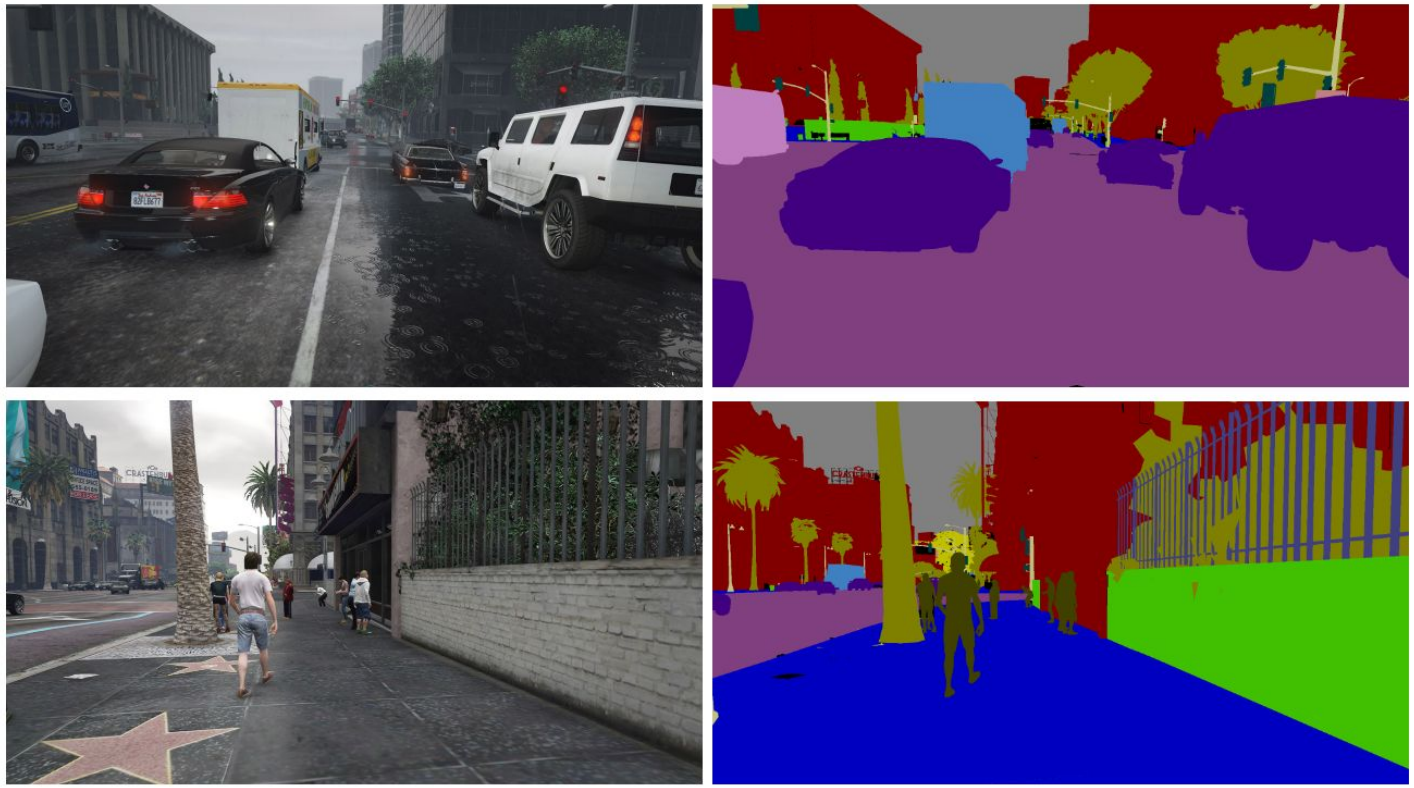
Multimodal Learning

Synthetic data

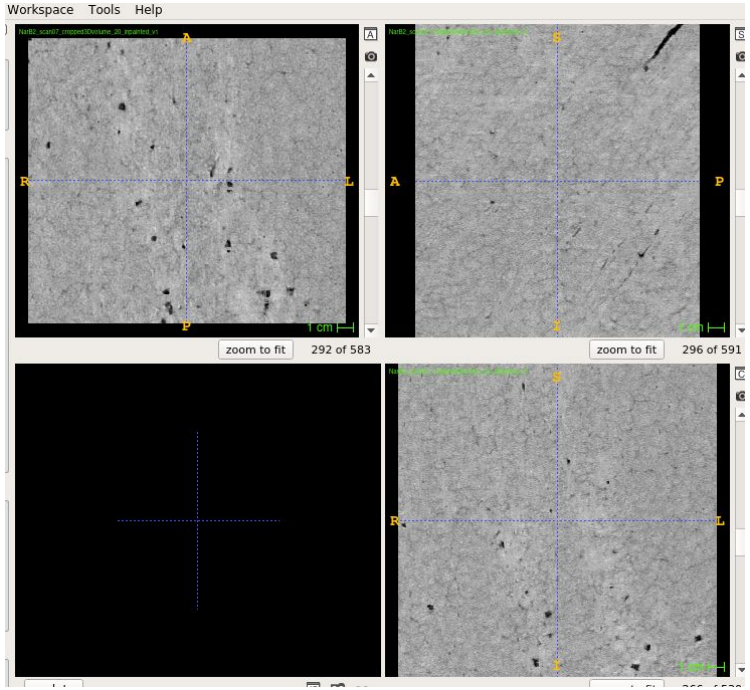
Playing for Data: Ground Truth from Computer Games

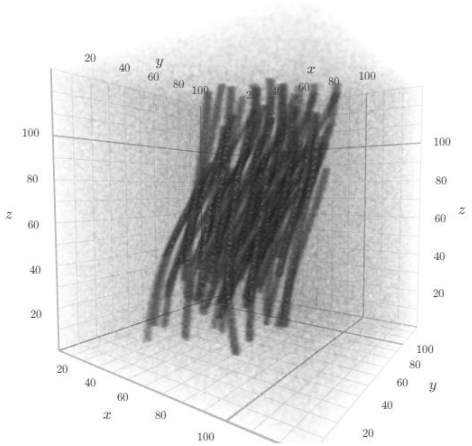
Stephan R. Richter*[1] Vibhav Vineet*[2] Stefan Roth[1] Vladlen Koltun[2]

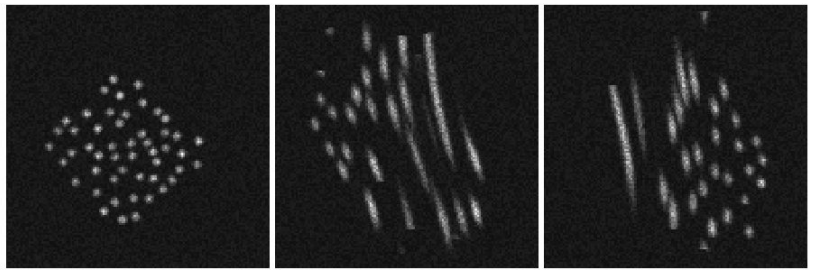
[1]TU Darmstadt [2]Intel Labs

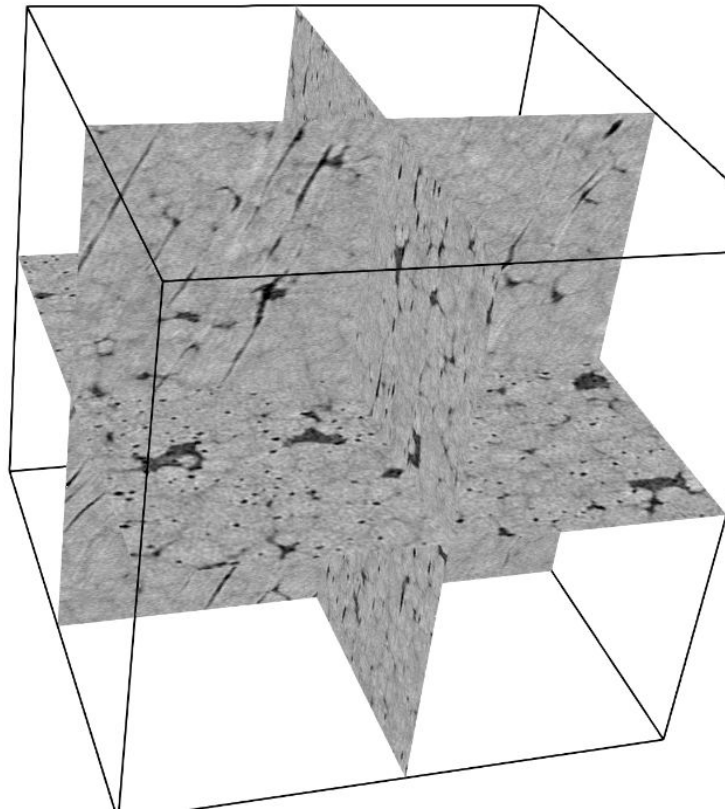


Synthetic data









Domain adaptation

Example:

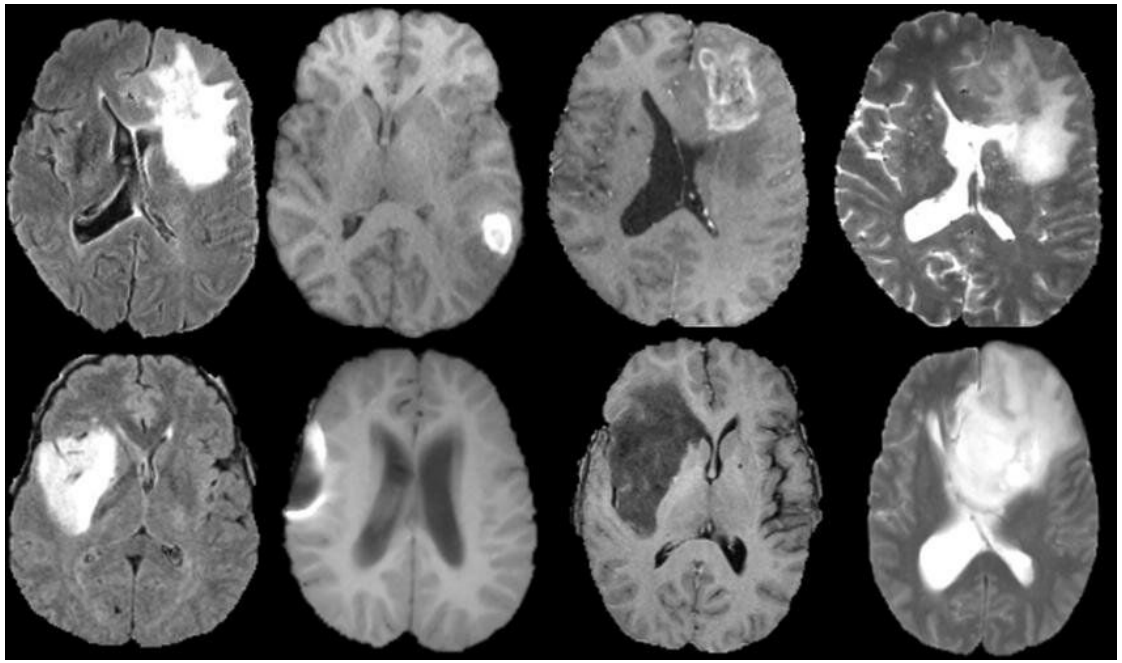
Dataset 1

Images and Labels

Dataset 2

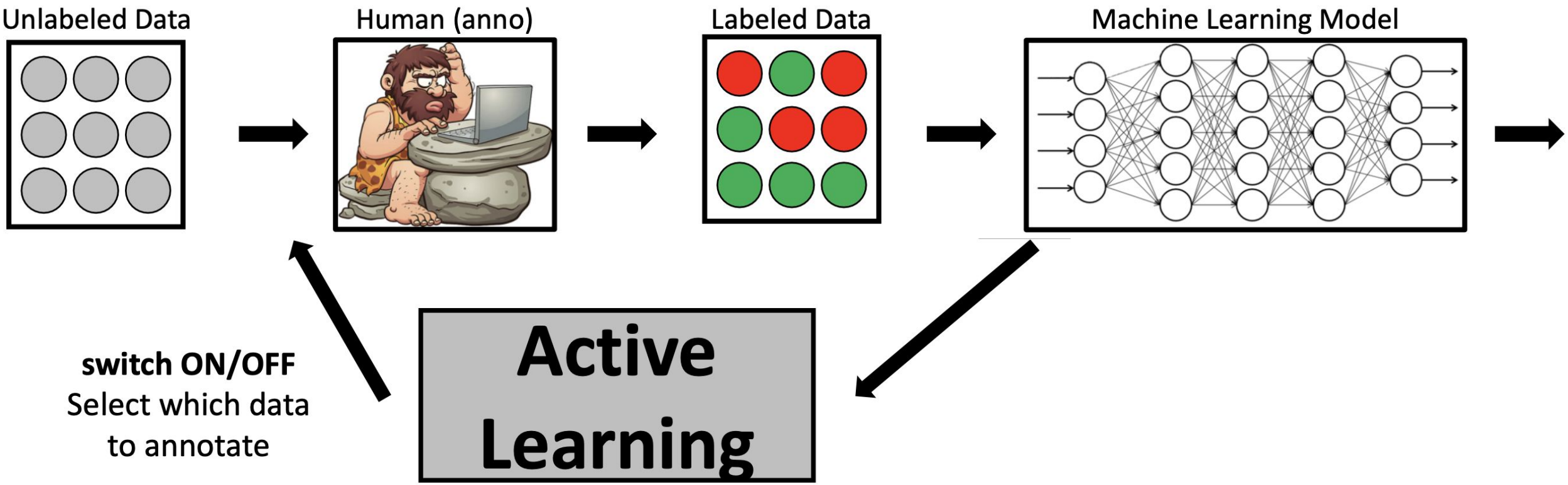
Images

Goal: Segment Dataset 2

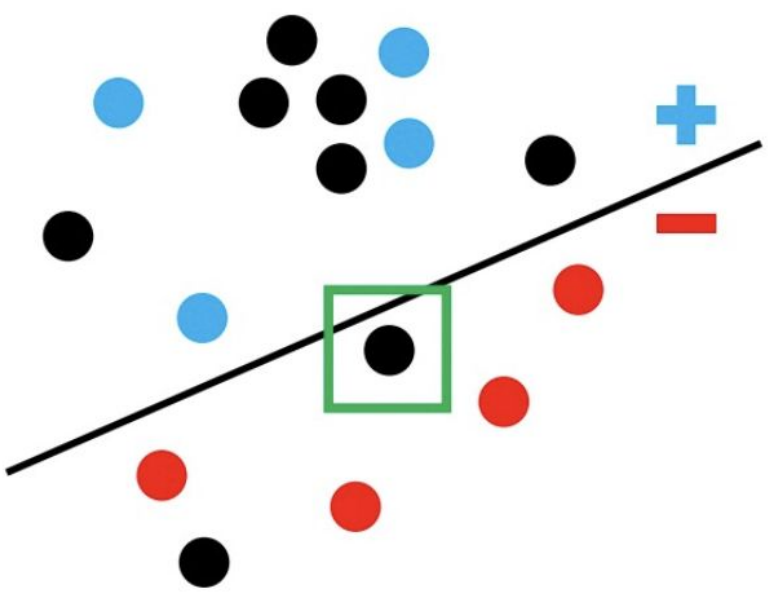


BraTS dataset

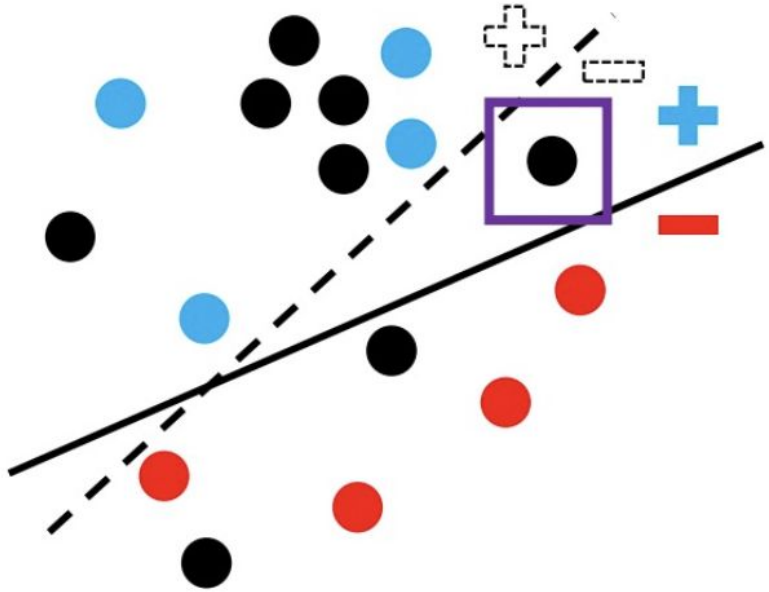
Active Learning



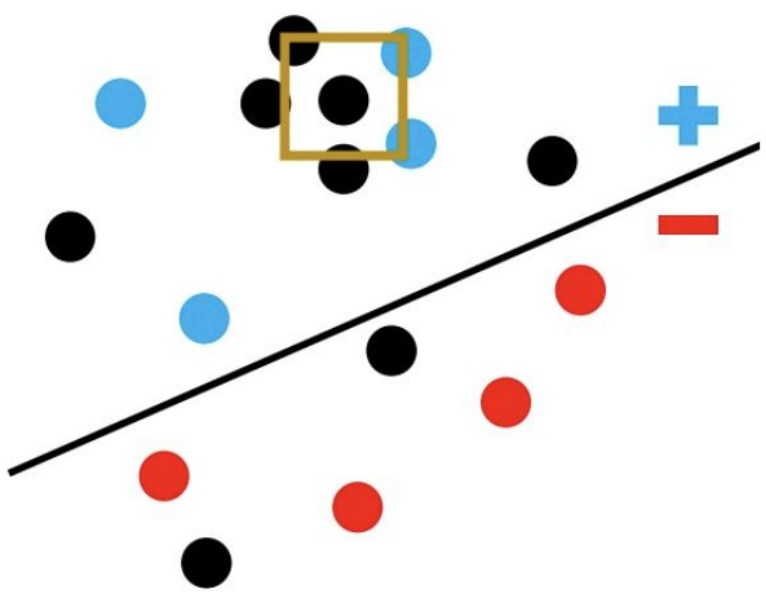
Active Learning



(a) Uncertainty sampling
e.g. [Tong and Koller, 2002; Kapoor et al., 2010]



(b) Query by committee
e.g. [Seung et al., 1992; Loy et al., 2012]



(c) Sampling from dense region
e.g. [Li and Guo, 2013]

[Kovashka, Russakovsky, Fei-Fei, Grauman, Foundations and Trends in Computer Vision and Graphics, 2016]

Active Learning

Active learning	
General	
Paper	Comments
[Kapoor ICCV07]	image classification (image-level labels)
[Qi CVPR08]	image classification (image-level labels)
[Collins ECCV08]	image classification (image-level labels)
[Joshi CVPR09]	image classification (image-level labels)
[Vijayanarasimhan CVPR09]	general framework (tested for segmentation only)
[Vijayanarasimhan NIPS09]	segmentation
[Jain CVPR09]	image classification (image-level labels)
[Siddiquie CVPR10]	segmentation
[Kovashka ICCV11]	image-level labels (objects and attributes)
[Luo NIPS 11]	results for layout3D only
[Vijayanarasimhan CVPR11]	object detection
[Yao CVPR12]	object detection
[Kao arxiV18]	object detection
[Roy BMVC18]	object detection
[Vezhnevets CVPR12]	semantic segmentation
[Russakovsky CVPR15]	actually not AL but related
[Sun CVPR15]	semantic segmentation
[Jain CVPR16]	semantic segmentation
[Konyushkova CVPR18]	object detection + active learning (yes/no verification or draw?)
[Sinha ICCV19]	Variational Adversarial Active Learning

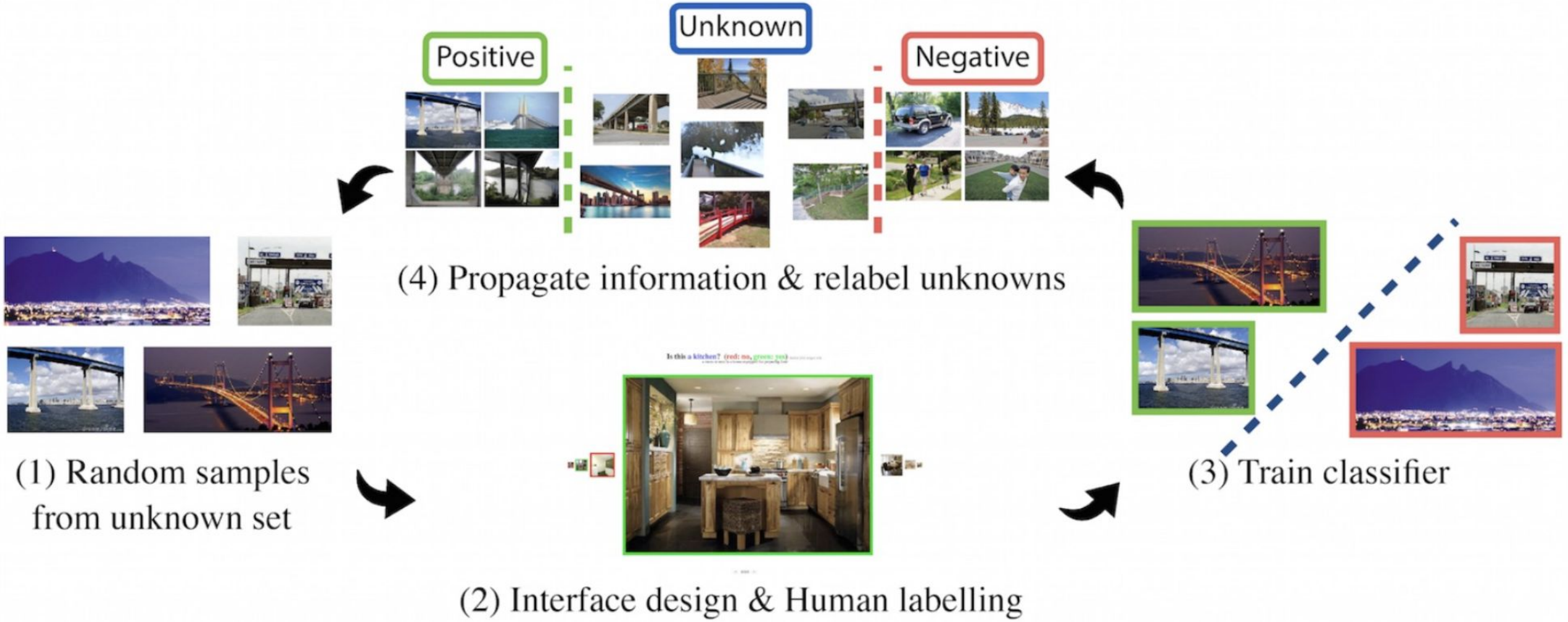
Semantic segmentation	
Paper	Comments
[Andriluka ACMMM18]	Fluid annotation
[Agustsson arxiv18]	Bounding-boxes plus scribbles to segments
[Maninis CVPR18]	Extreme clicks to segments
[Lin CVPR16]	Scribbles to segments
[Xu CVPR15]	scribbles to segments
[Bearman ECCV16]	Point-clicks to segments
[Jain HCOMP16]	Point-clicks to segments
[Wang CVIU14]	Point-clicks to segments
[Castrejon CVPR17]	Polygon RNN
[Acuna CVPR18]	Polygon RNN++
[Ling CVPR19]	Curve GCN
[Rother SIGGRAPH 04]	Original GrabCut (bounding box to segments)
[Lempitsky CVPR09]	GrabCut-style (bounding box to segments)
[Wu CVPR04]	GrabCut-style (bounding box to segments)
[Bai IJCV09]	GrabCut-style (scribbles to segments)
[Duchenne CVPR08]	GrabCut-style (scribbles to segments)
[Gulshan CVPR10]	GrabCut-style (scribbles to segments)
[Price CVPR10]	GrabCut-style (scribbles to segments)
[Liang arxiV19]	Polytransform

Annotation propagation	
Paper	Comments
[Rubinstein ECCV12]	semantic segmentation
[Kuettel ECCV12]	segmentation propagation Imagenet
[Jain CVPR16]	active learning

Annotation tools	
Paper	Comments
[LabelMe]	LabelMe
[Scalabel]	Scalabel
[Andriluka ACMMM18]	Fluid Annotation (just a demo)
[Caesar arxiV17]	COCOstuff 10K annotation tool
[Toronto]	Toronto Annotation suite

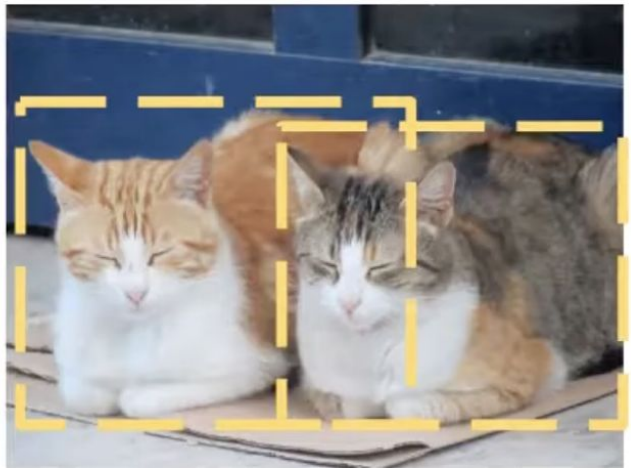
Efficient annotation	
Object detection	
Paper	Comments
[Branson CVPR17]	crowdsourcing protocol for image-level labels and bounding-boxes
[Su AAAIW12]	crowdsouce bounding boxes for ImageNet
[Papadopoulos CVPR16]	Human verification (Yes/No type of questions)
[Papadopoulos CVPR17]	Center-clicking (Click on the center)
[Papadopoulos ICCV17]	Extreme clicking (another way to get the bounding box, also helps on segmentation)
OpenImages dataset	Google Research annotated about 4 million bounding boxes using two of my above schemes

Label propagation



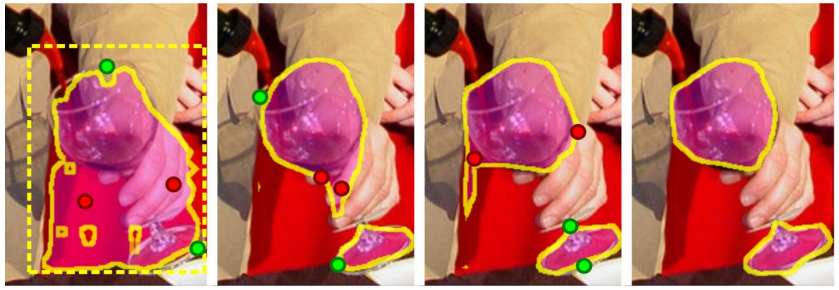
[LSUN, Fisher Yu, arXiv 2015]

Typical computer vision tasks

Classification	Detection	Semantic segmentation	Instance segmentation
			
Is there a cat?	Where is the cat?	Which pixels is cat?	Which are the cat's pixels?

Annotation time

Cheaper annotations



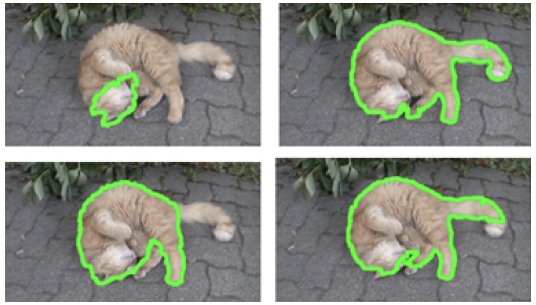
[Xu CVPR 16, Li CVPR 18, Chen CVPR 18, Benenson CVPR 19]

[Lin CVPR 16, Bearman ECCV 16, Agustsson CVPR 19]

cheap → *expensive*

Binary | Binary ++ | Bounding box | Pick among N | Click/Scribble/ Edit polygon

[Papadopoulos CVPR 16]

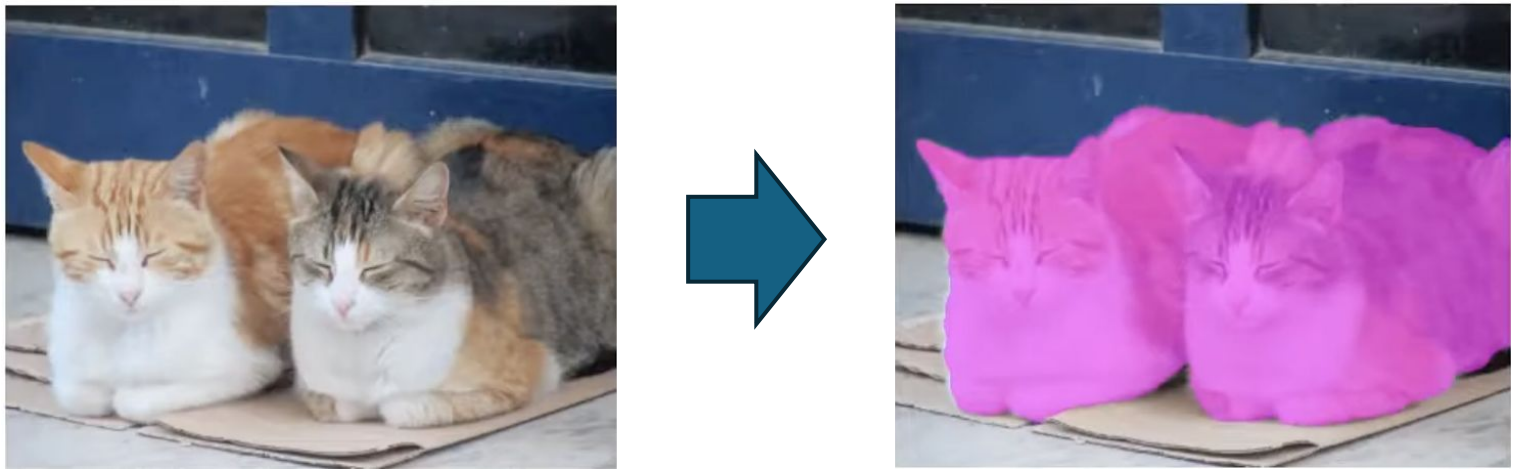


[Jain IJCV 19]

[Castrejon CVPR 17, Acuna CVPR 18, Ling CVPR 19]

Fully supervised vs Weakly supervised (TRAINING)

Fully supervised **Directly** supervised



Weakly supervised **Indirectly** supervised

Weaker annotation
Faster annotation

How do you train a segmentation model with weak labels?

Bounding boxes
Class-level labels
Points

[Example] Weak annotations: Bounding boxes

Use bounding boxes rather than manual segmentations
Initialize with bounding boxes as current segmentations and iterate:

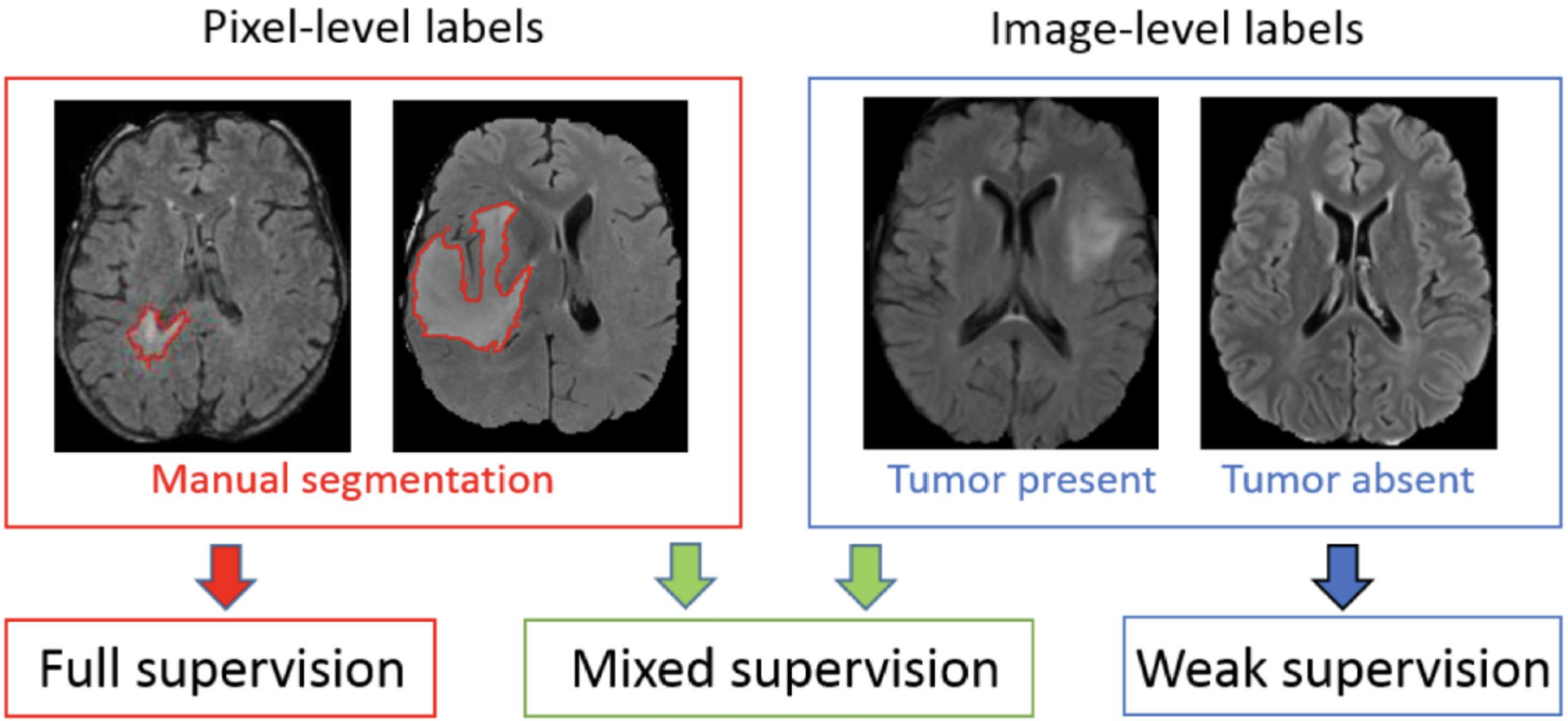
- Use current segmentations as labels to train a segmentation network
- Use the segmentation network to predict a new set of segmentations

Example input rectangles | Output after 1 training round | After 5 rounds | After 10 rounds | Ground truth

This approach is “naive” – the paper includes additional tricks to clean up the current segmentation

https://openaccess.thecvf.com/content_cvpr_2017/papers/Khoreva_Simple_Does_It_CVPR_2017_paper.pdf

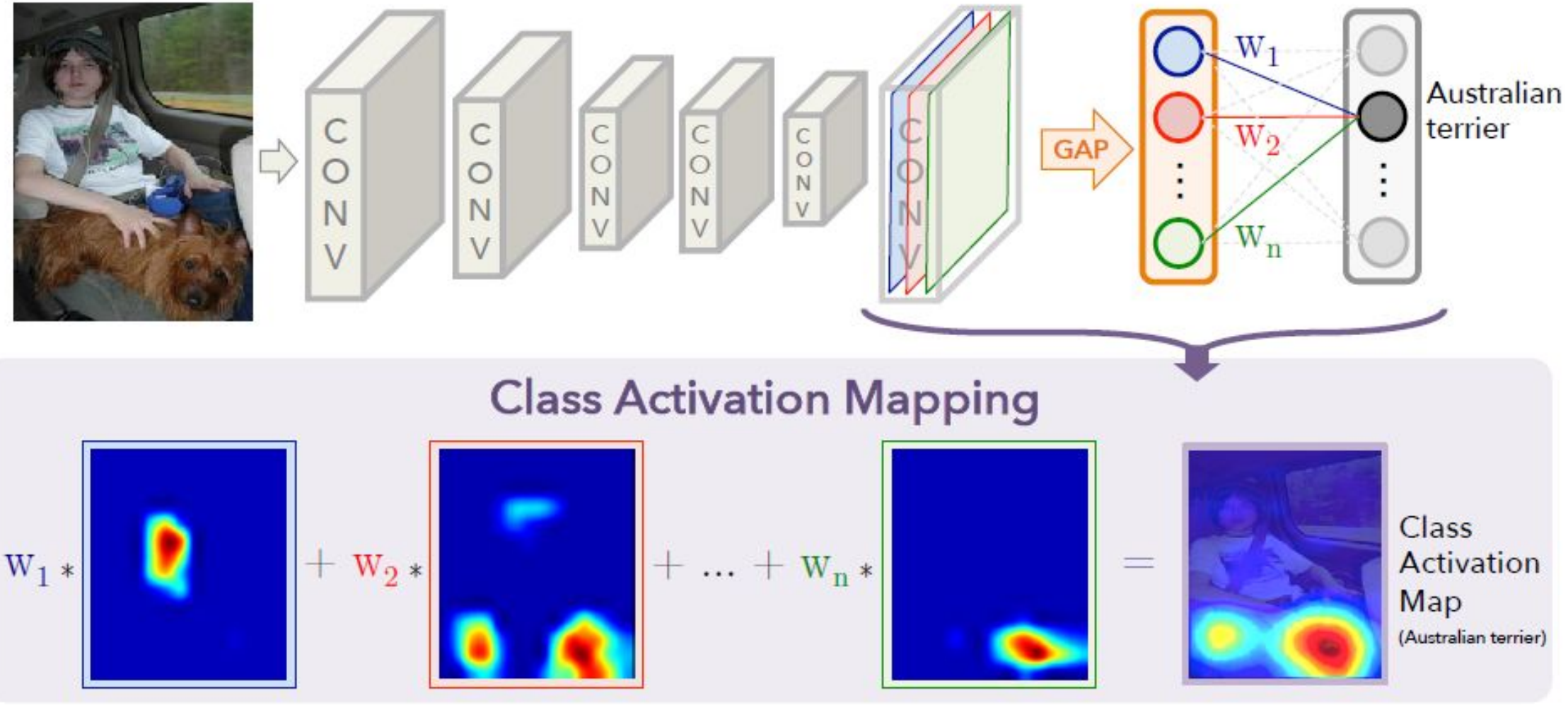
[Example] Weak annotations: Image level labels



How can you handle this?

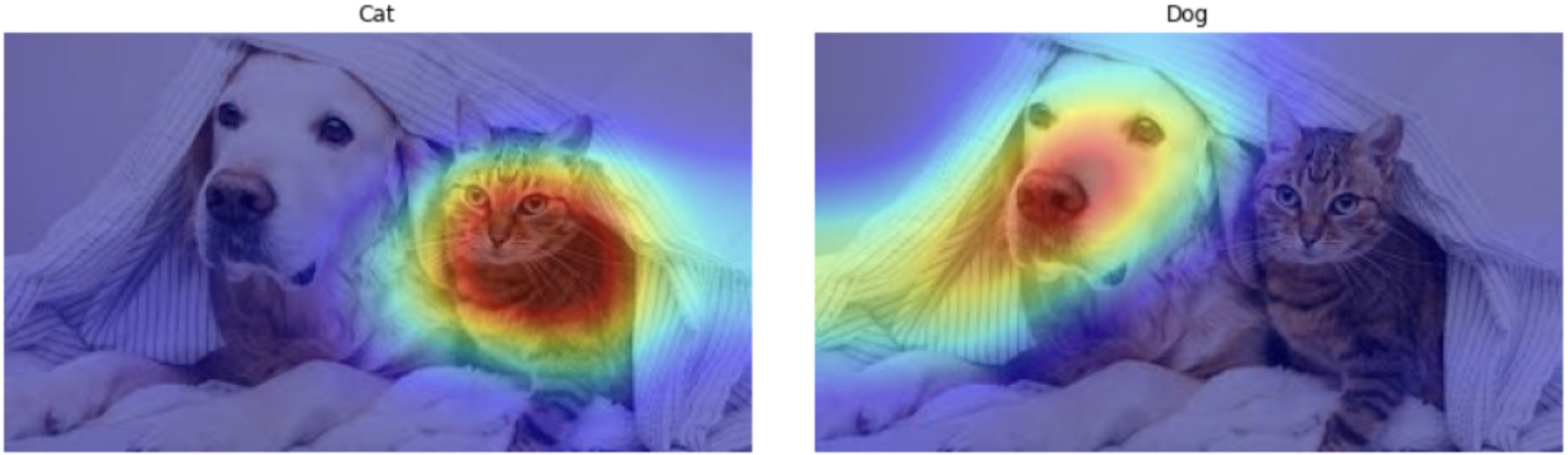
[Figure from Mlynarski et al, 2018]

From image labels to segmentation



[Zhou et al 2016]

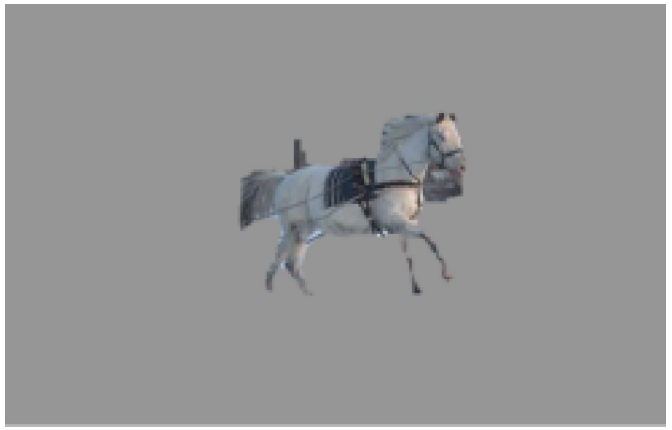
Using saliency maps from image classification for segmentation



Saliency maps provide (when they work well) some level of localization –how to utilize this for segmentation is an active research field

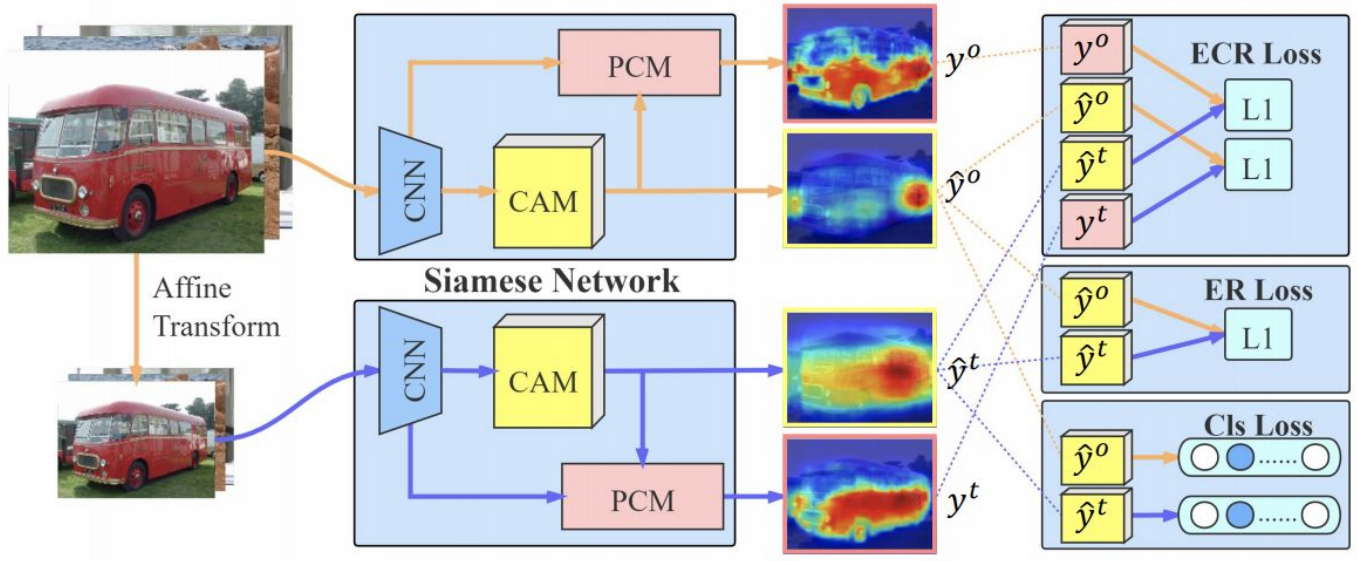
How do we approach this?

(A) Estimate segmentation masks in training set



Pseudo GT masks are used to train a segmentation model with standard CE

(B) Learn a model directly from weak labels



Learn a model directly from weak labels using loses appropriate for each task

Priors and Hints

Priors: what do I believe to be true independent of any particular image sample

Size, shape, location, number of instances, contrast, motion, similarity across images, similarity with external images, etc, etc.

Hints: Indirect supervision received for each image sample

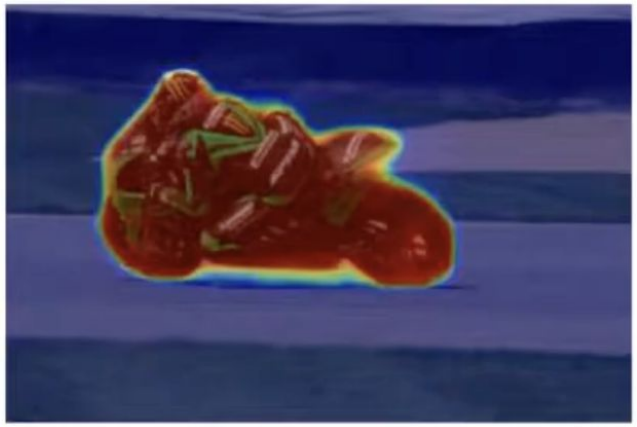
Image labels, image captions, transfer across images, click inside the object, size from center clicks, object bounding boxes, extreme clicks, scribbles, eye gaze, localized narratives

Hints and Priors provide constraints

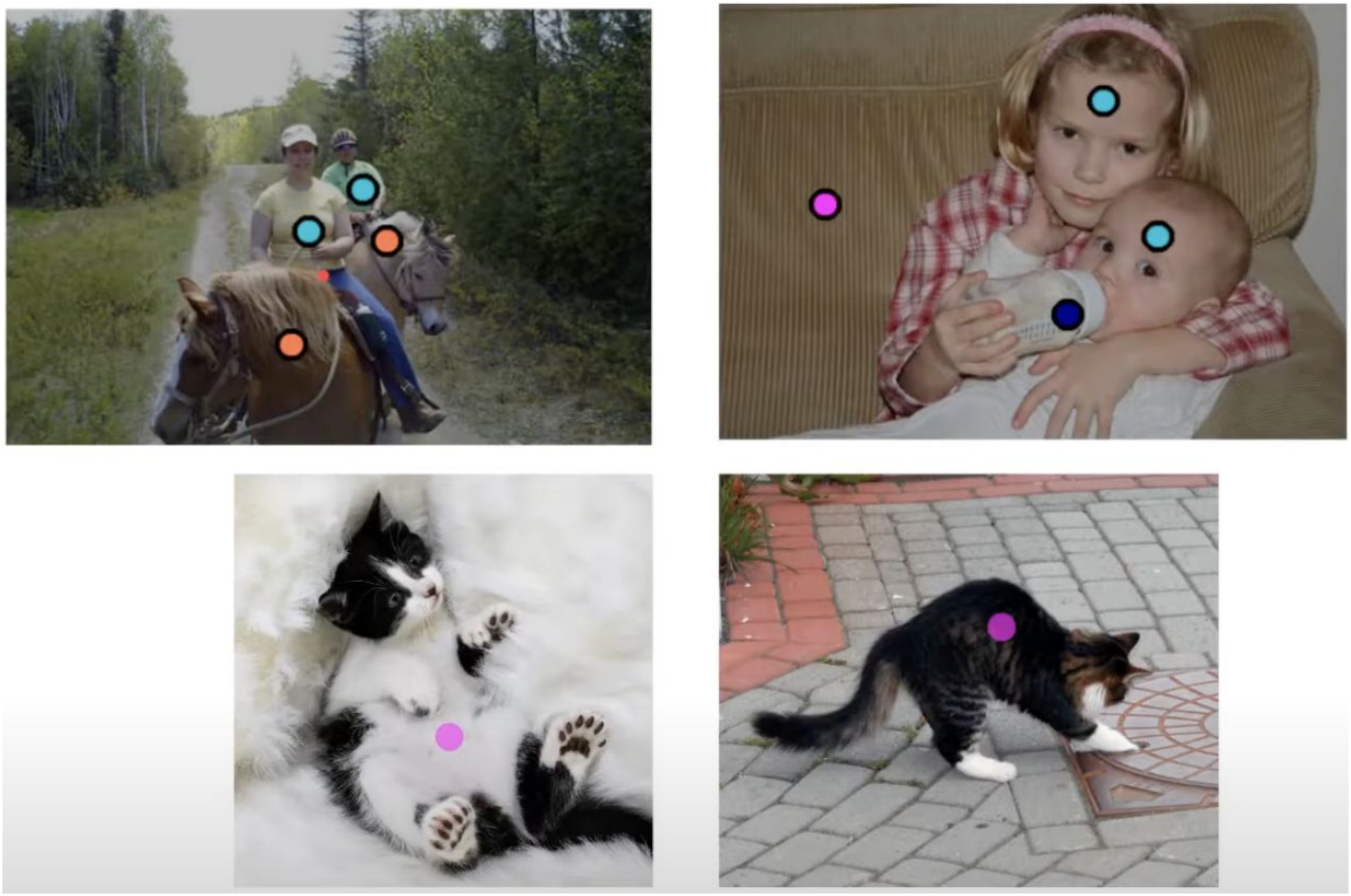
Priors: Boundaries

Which pixels separate objects?

Priors: Saliency

Hints: Clicks



[Bearman ECCV 2016]

Hints: Center clicks

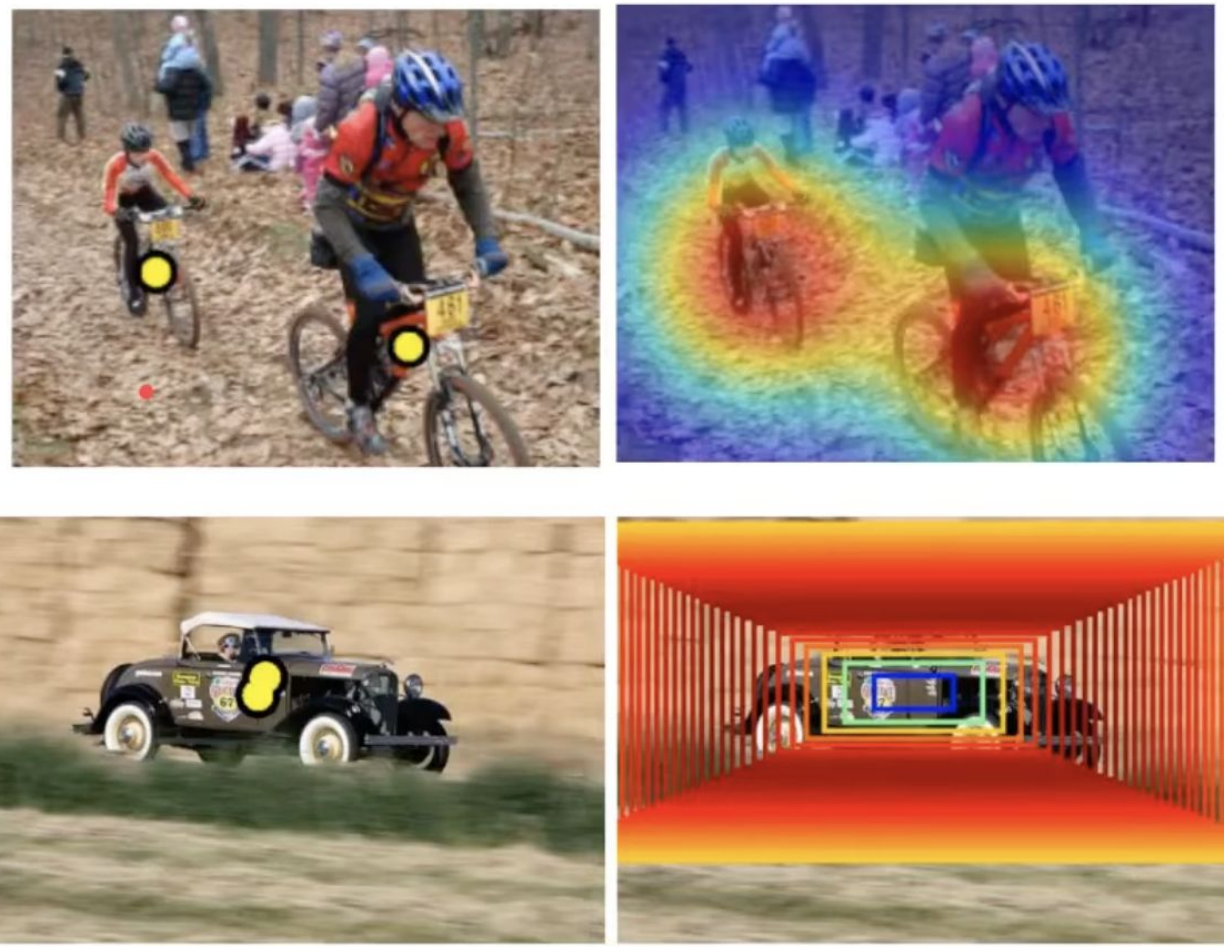
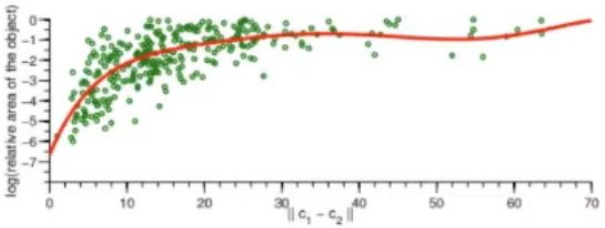
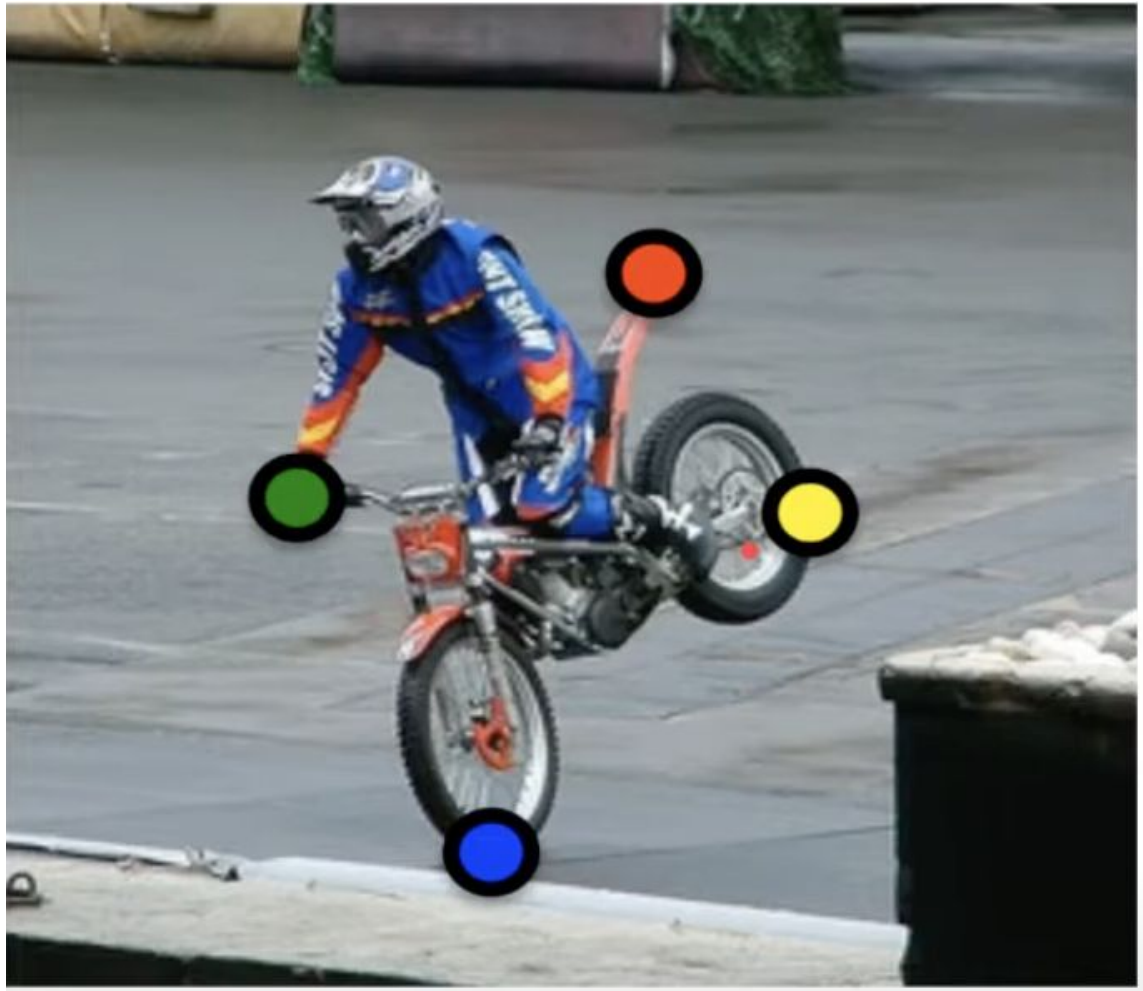


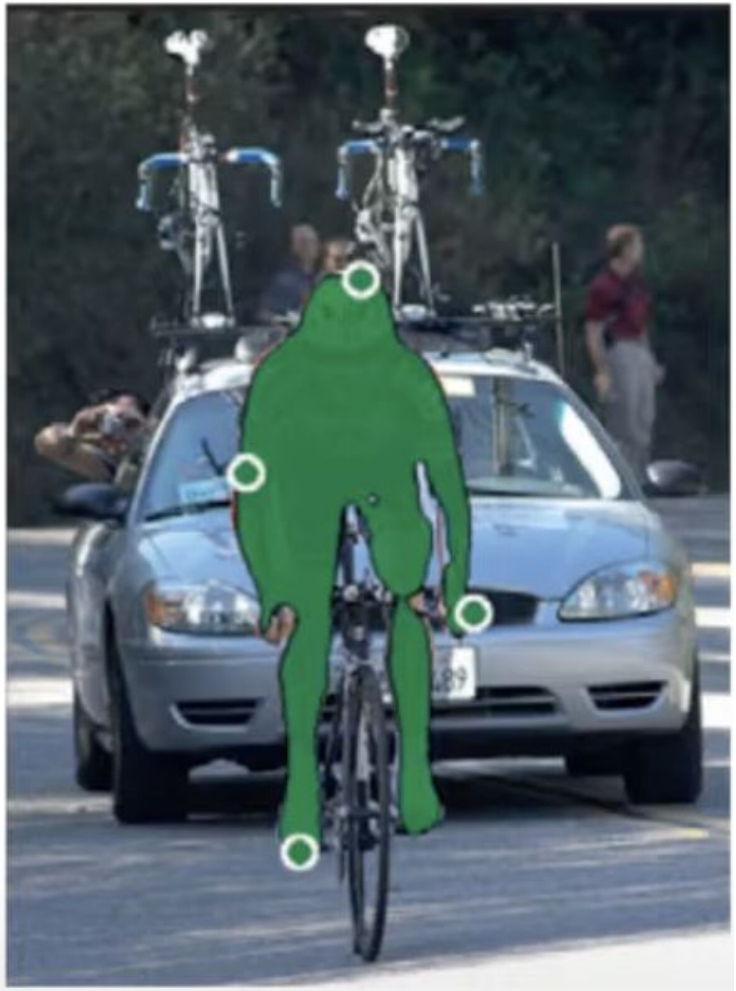
Figure 6. **Box area score** S_{ba}. All windows used here have fixed aspect ratio and are centered on the center of the object.



[Papadopoulos CVPR 2017]

Hints: Extreme clicks

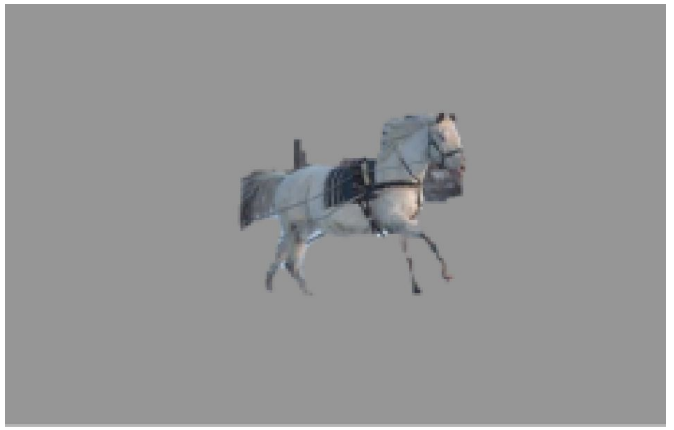




[Papadopoulos ICCV 2017, Maninis CVPR 2018]

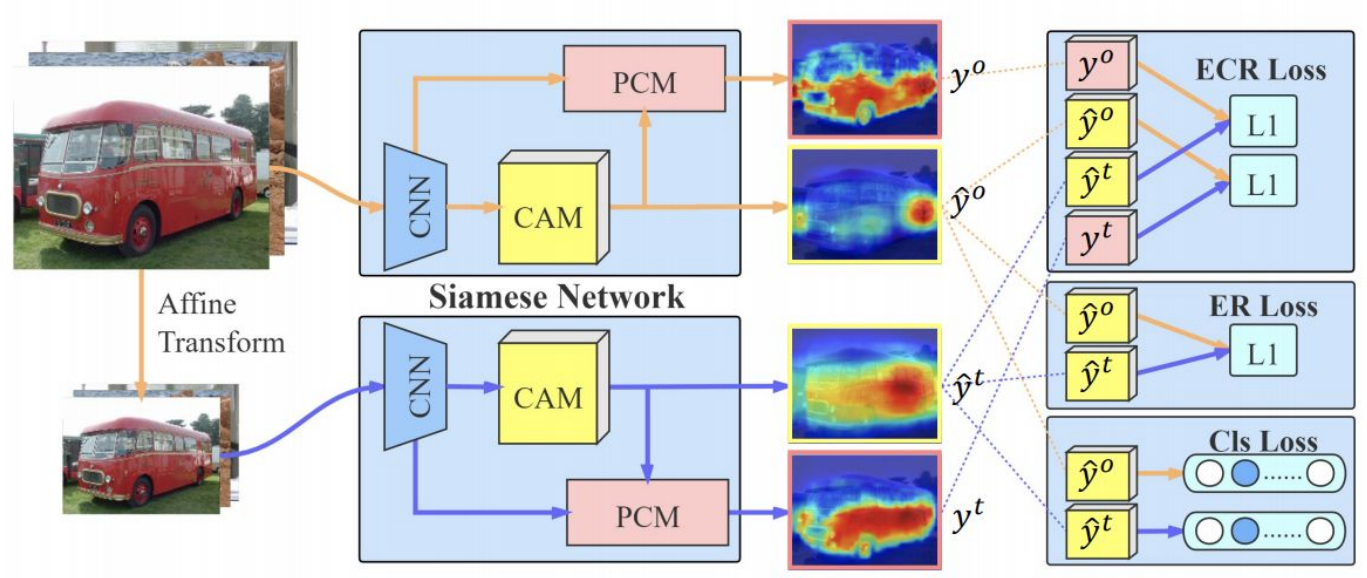
How do we approach this?

(A) Estimate segmentation masks in training set



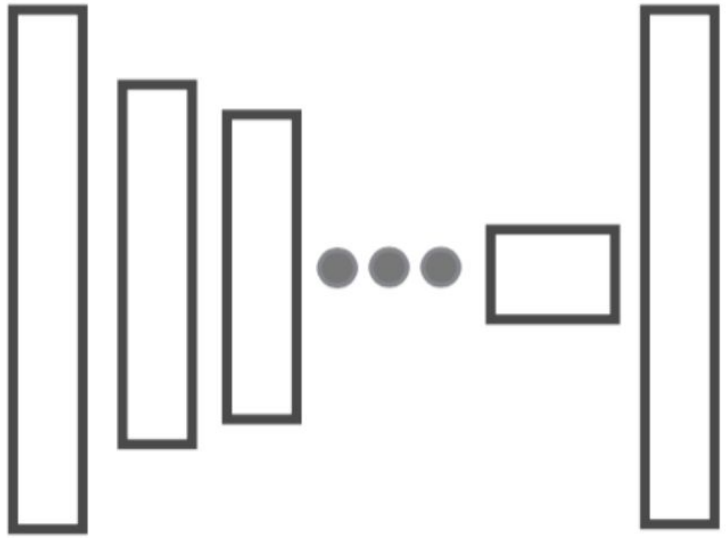
Pseudo GT masks are used to train a segmentation model with standard CE

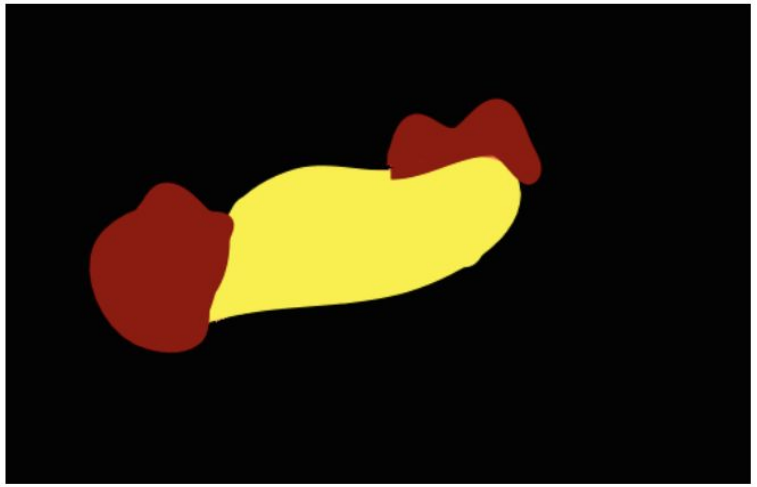
(B) Learn a model directly from weak labels



Learn a model directly from weak labels using loses appropriate for each task

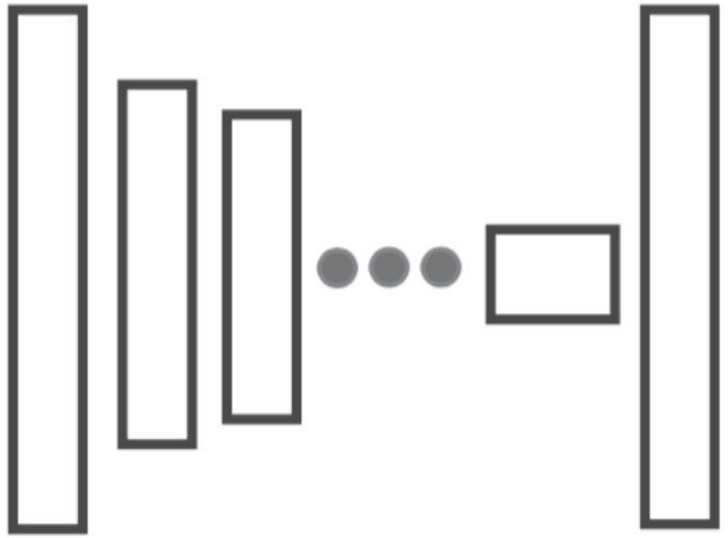
Fully supervised (TRAINING)

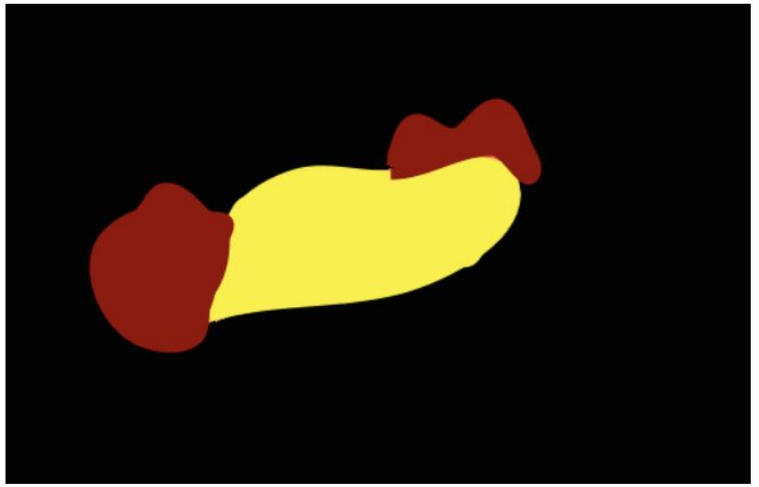




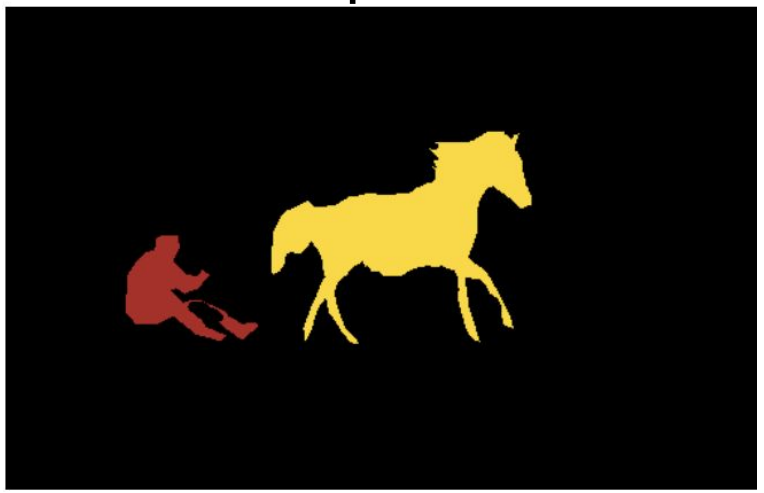
[Bearman ECCV 2016]

Fully supervised (TRAINING)



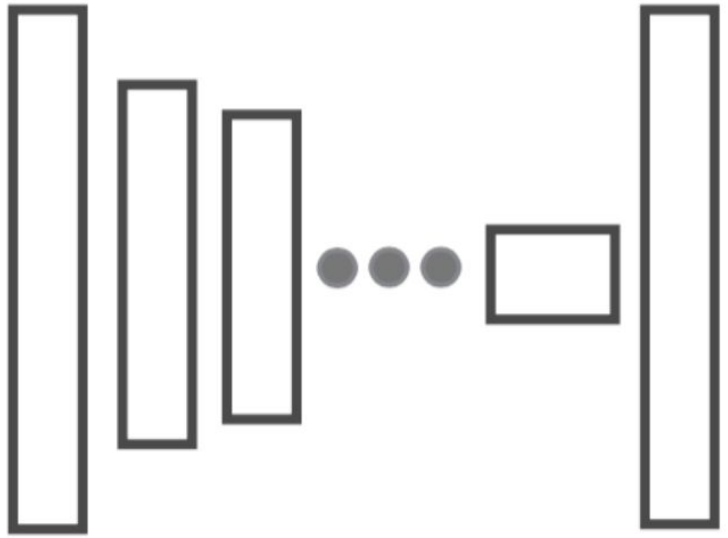


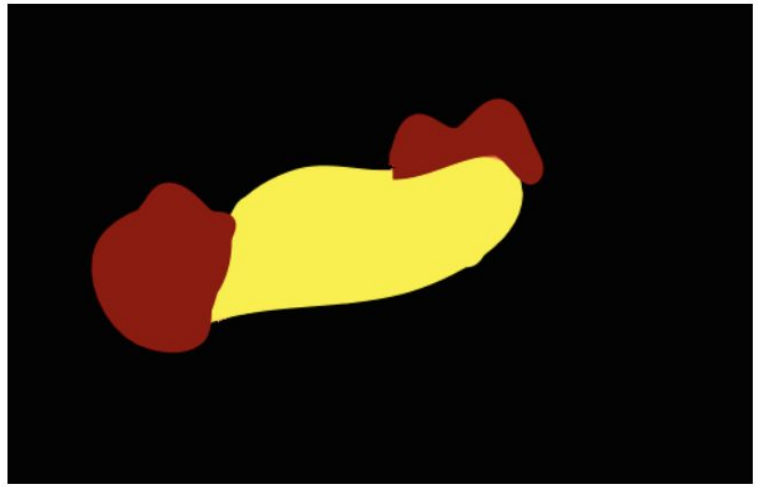
Full supervision



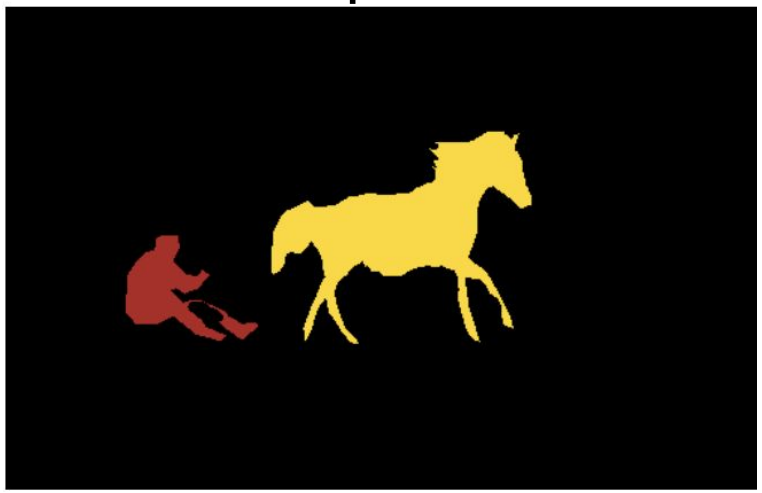
[Bearman ECCV 2016]

Fully supervised (TRAINING)





Full supervision

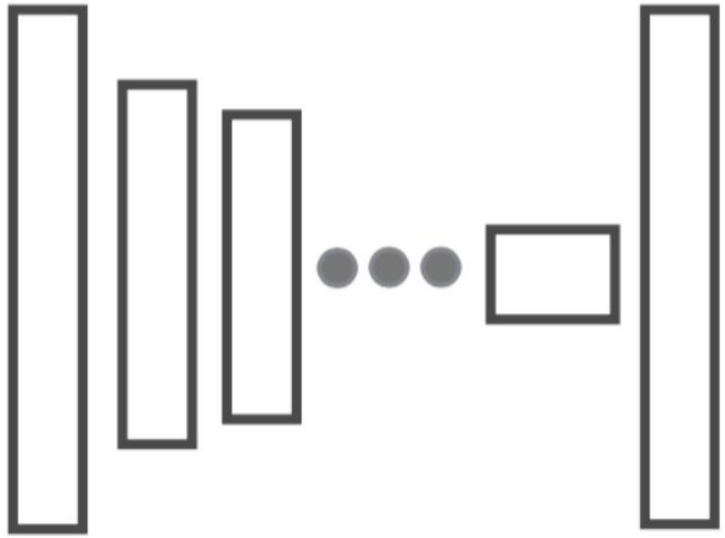


Full Supervision. When the class label is available for every pixel during training, the CNN is commonly trained by optimizing the sum of per-pixel cross-entropy terms [5,37]. Let $\mathcal{I}$ be the set of pixels in the image. Let s_{ic} be the CNN score for pixel i and class c. Let $S_{ic} = \exp(s_{ic}) / \sum_{k=1}^{N} \exp(s_{ik})$ be the softmax probability of class c at pixel i. Given a ground truth map G indicating that pixel i belongs to class G_i, the loss on a single training image is:

$$\mathcal{L}_{pix}(S, G) = -\sum_{i \in \mathcal{I}} \log(S_{iG_i}) \qquad (1)$$

[Bearman ECCV 2016]

Weakly supervised (TRAINING)



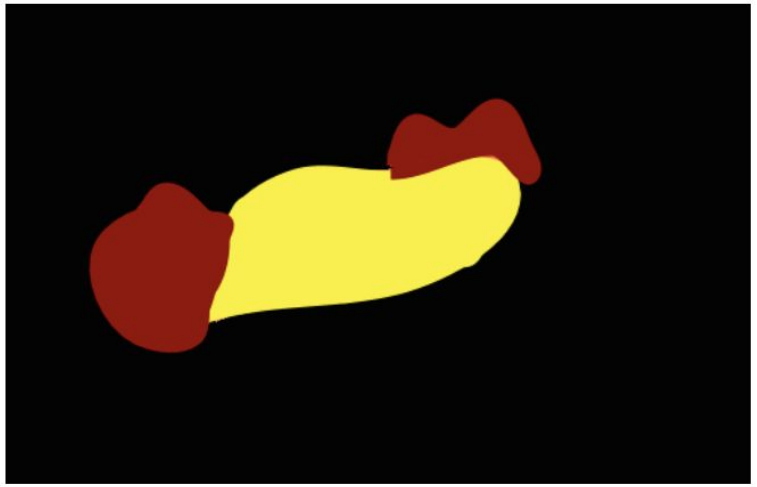
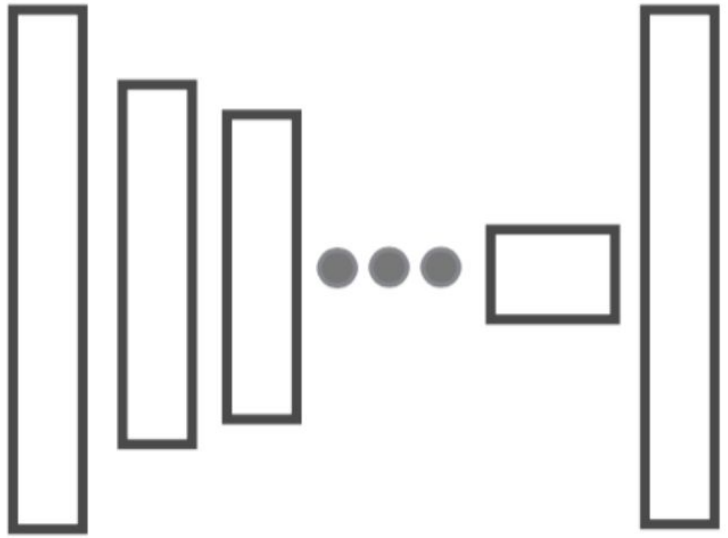


Image-level supervision

horse
person

[Bearman ECCV 2016]

Weakly supervised (TRAINING)



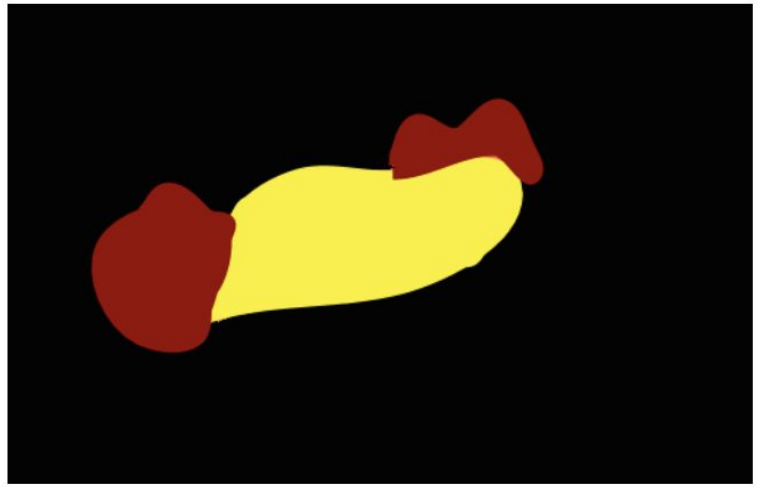


Image-level supervision

horse
person

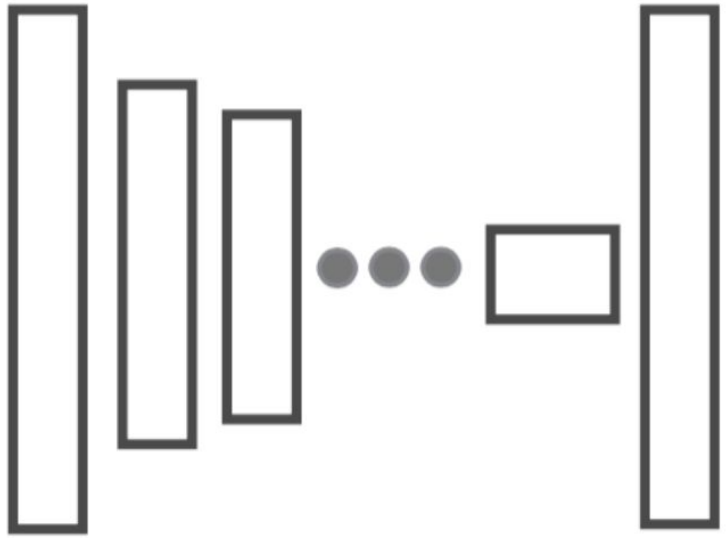
Image-Level Supervision. In this case, the only information available during training are the sets $L \subseteq \{1, \ldots, N\}$ of classes present in the image and $L' \subseteq \{1, \ldots, N\}$ of classes not present in the image. The CNN model can be trained with a different cross-entropy loss:

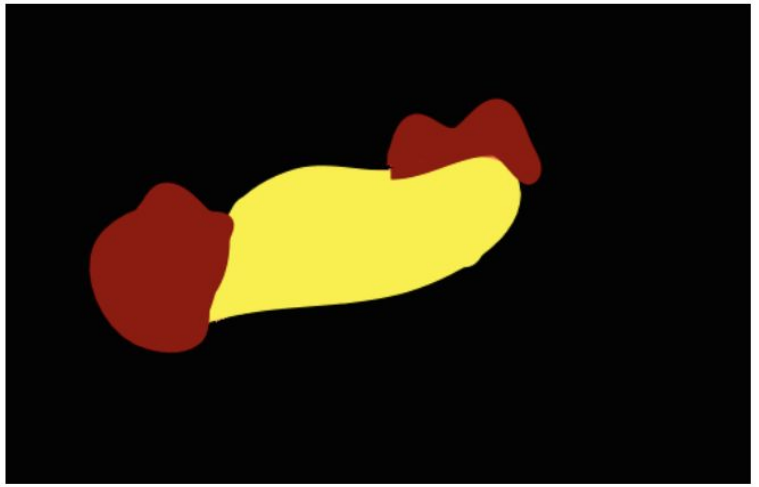
$$\mathcal{L}_{img}(S, L, L') = -\frac{1}{|L|}\sum_{c \in L} \log(S_{t_c c}) - \frac{1}{|L'|}\sum_{c \in L'} \log(1 - S_{t_c c}) \qquad (2)$$

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

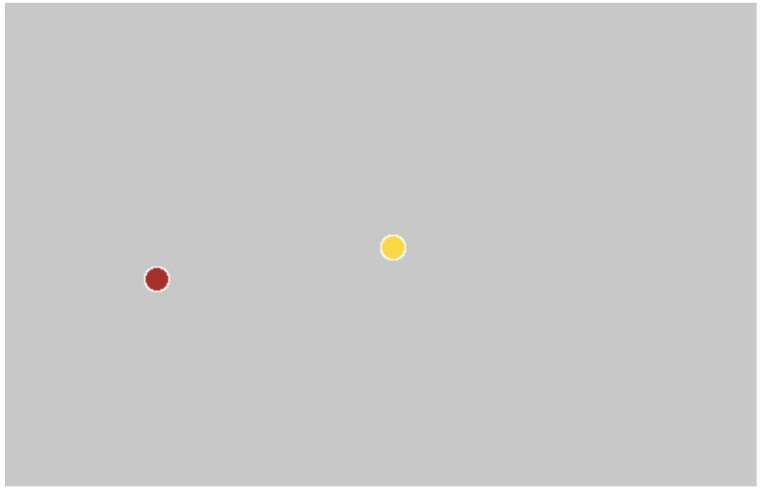
[Bearman ECCV 2016]

Weakly supervised (TRAINING)



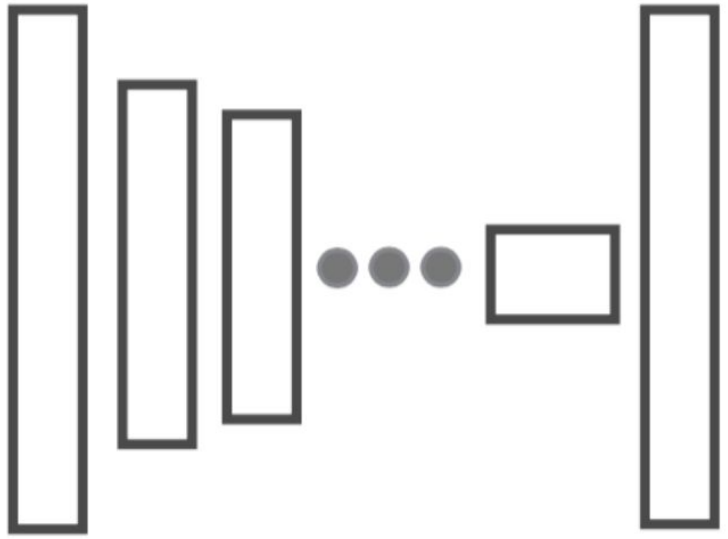


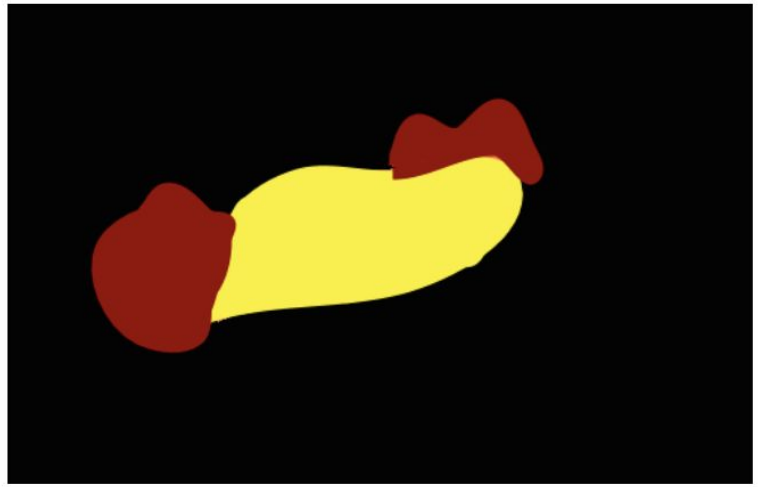
Point supervision



[Bearman ECCV 2016]

Weakly supervised (TRAINING)

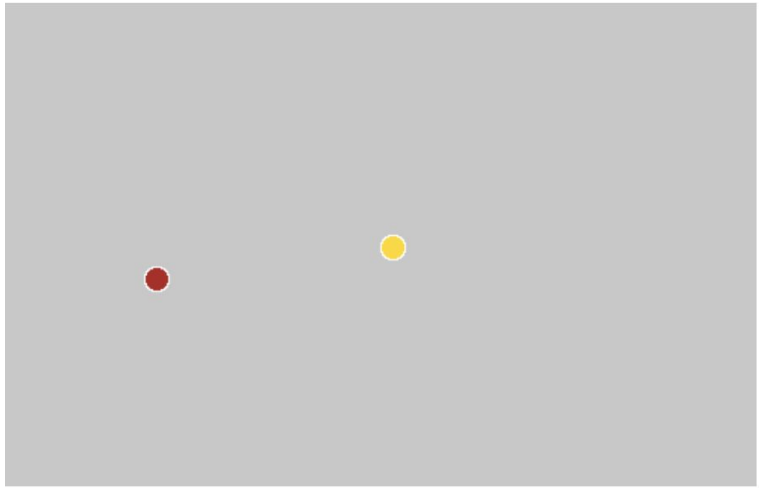




Point supervision

Point-Level Supervision. We study the intermediate case where the object classes are known for a small set of supervised pixels $\mathcal{I}_s$, whereas other pixels are just known to belong to some class in L. We generalize Eqns. (1) and (2) to:

$$\mathcal{L}_{point}(S, G, L, L') = \mathcal{L}_{img}(S, L, L') - \sum_{i \in \mathcal{I}_s} \alpha_i \log(S_{iG_i}) \quad (3)$$



[Bearman ECCV 2016]

Next: Image segmentation project

PH2 dataset:

DRIVE dataset: