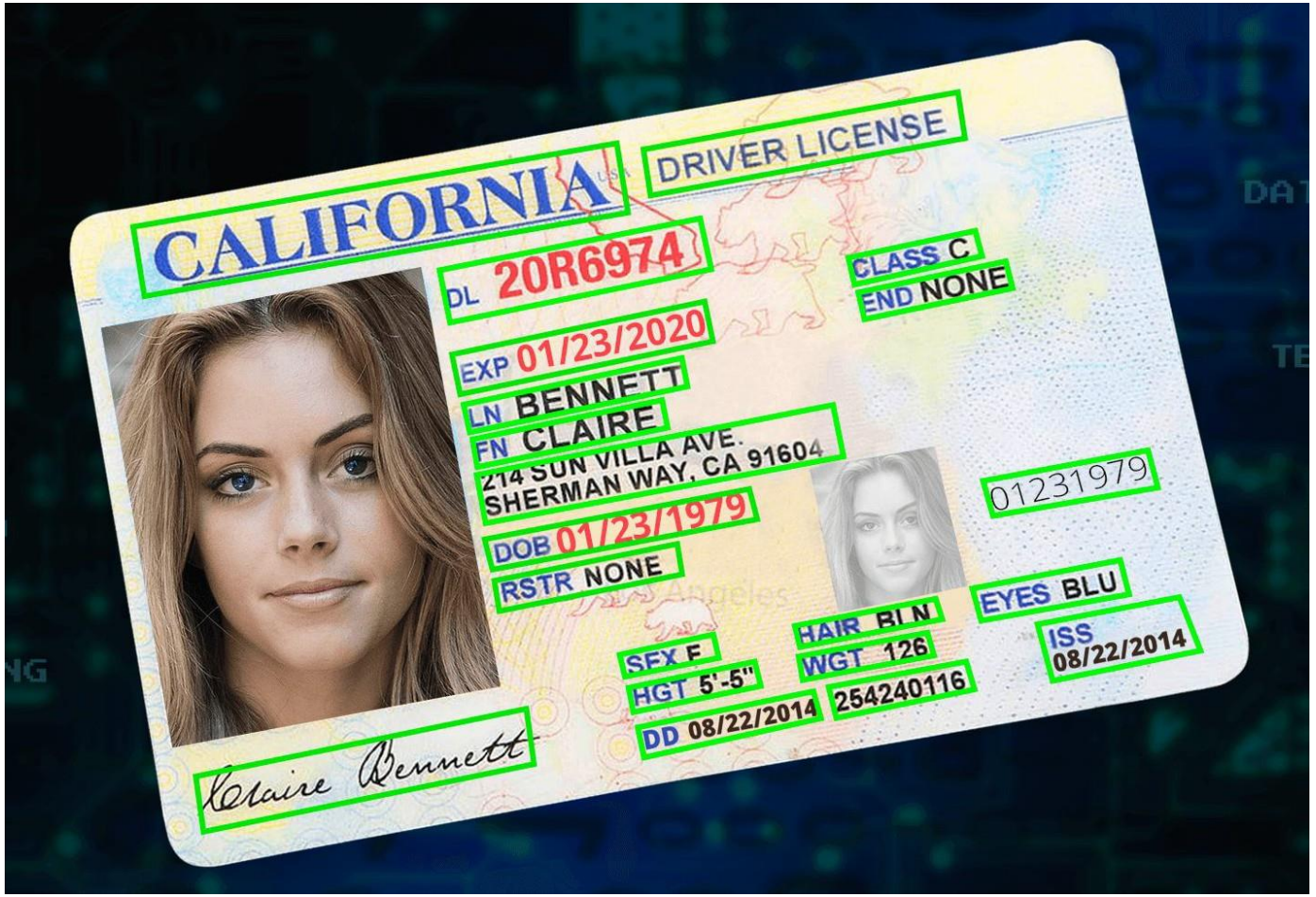
Weakly Supervised Learning in Computer Vision

Speaker: Xinyu Yang
(Senior Research Associate)

Computer vision is working in real life!

https://cloud.google.com/vision/automl/object-detection/docs

Computer vision is working in real life!

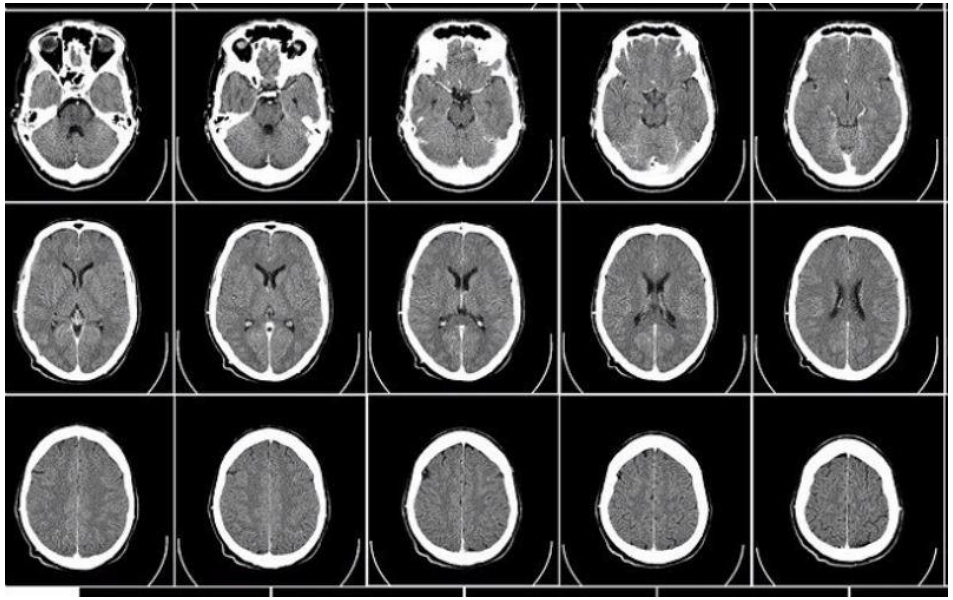


https://mobidev.biz/blog/ocr-machine-learning-implementation

Computer vision is working in real life!

https://www.biometricupdate.com/202001/airport-biometrics-predications-deployments-upgrades-and-plans-for-future-services

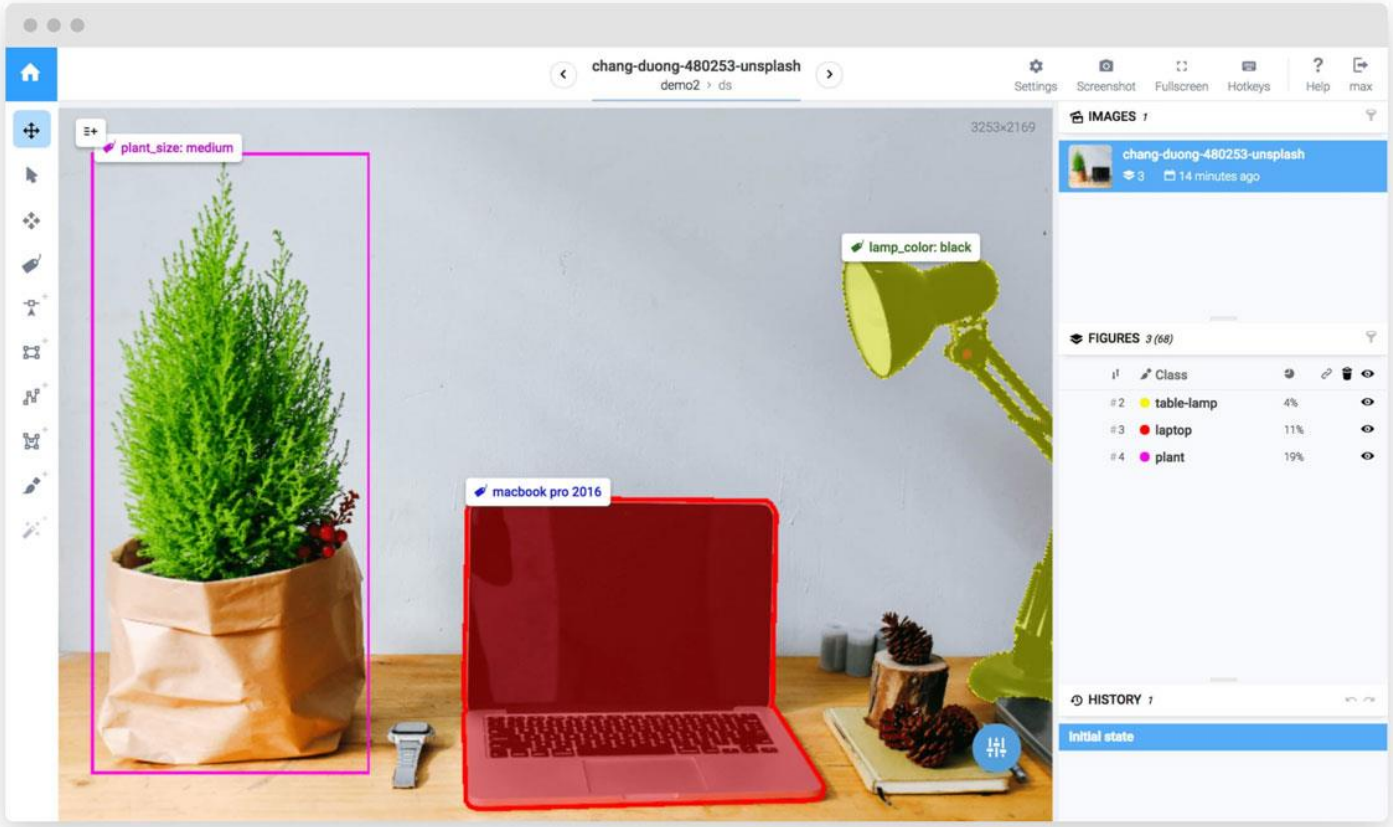
Computer vision is working in real life!



Medical AI

Self-driving cars

Behind the success··· Huge annotation costs.



ImageNet1K

Multi-label annotation takes

26 seconds / image

14M images in total!

Shankar et al. Evaluating Machine Performance on ImageNet. ICML 2020.

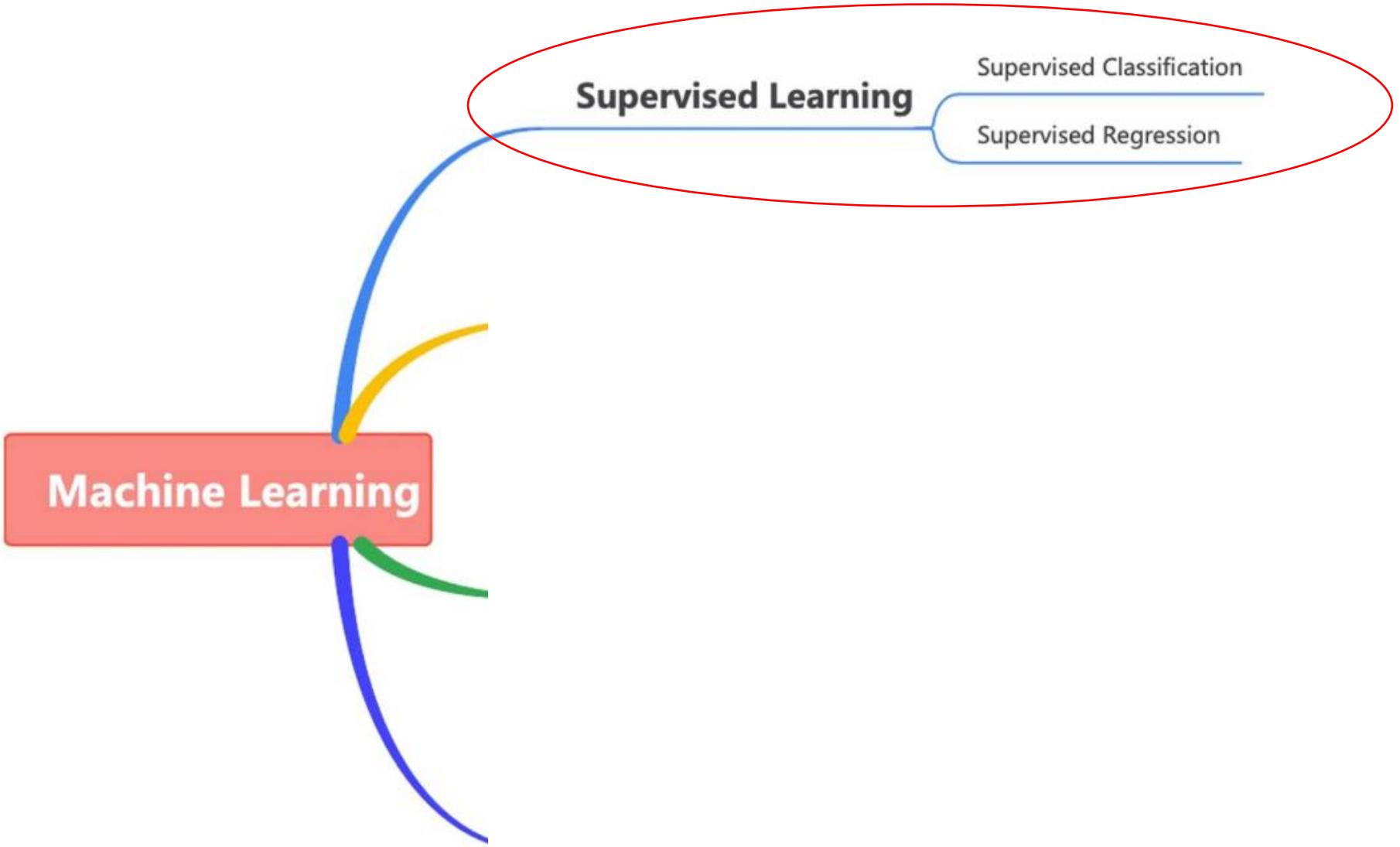
Behind the success… Huge annotation costs.

Cityscapes

1.5 hour / image.

Cordts et al. The Cityscapes Dataset for Semantic Urban Scene Understanding. CVPR 2016.

Supervision Types

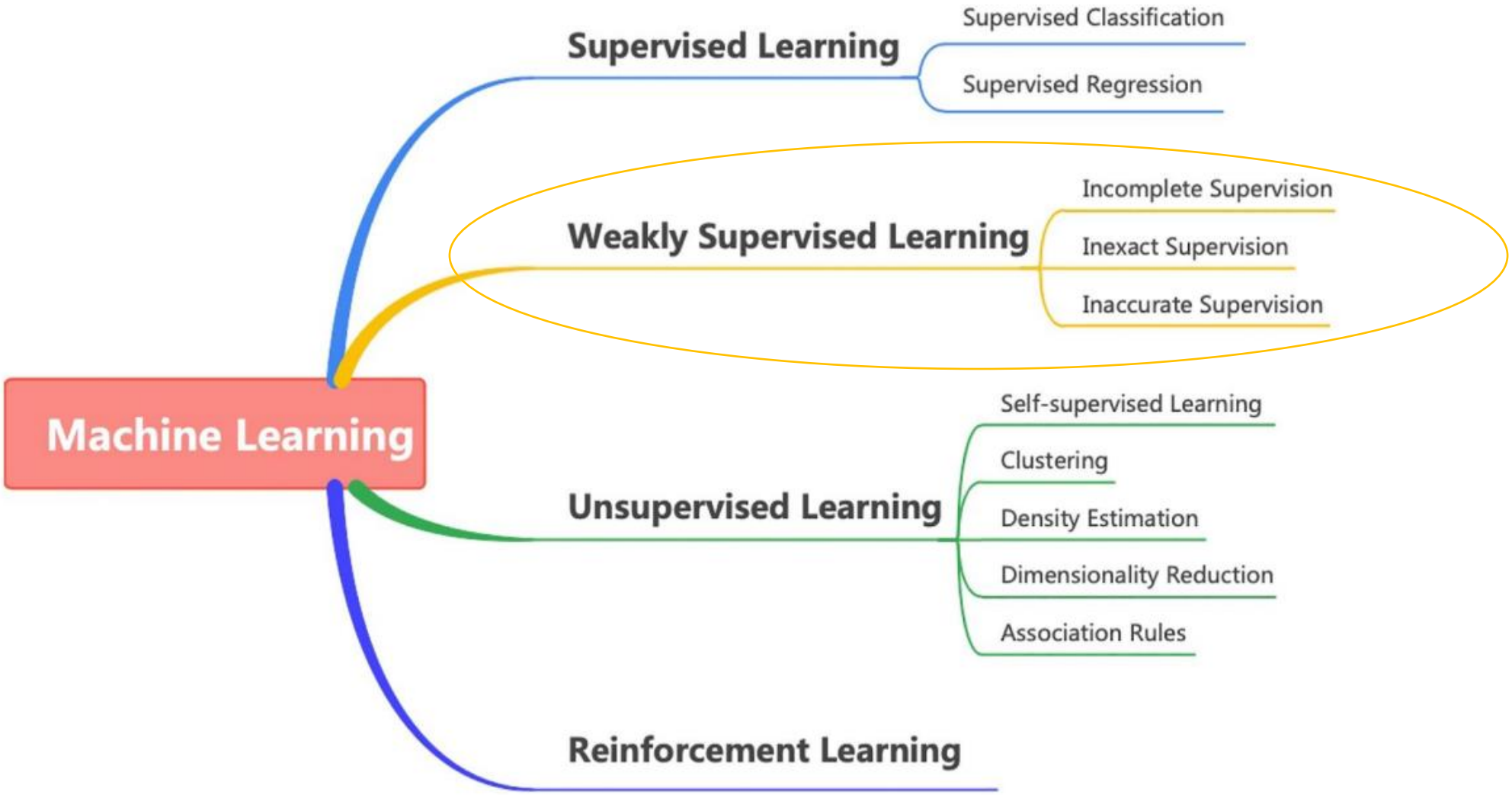


Supervised Classification

Label: Yes

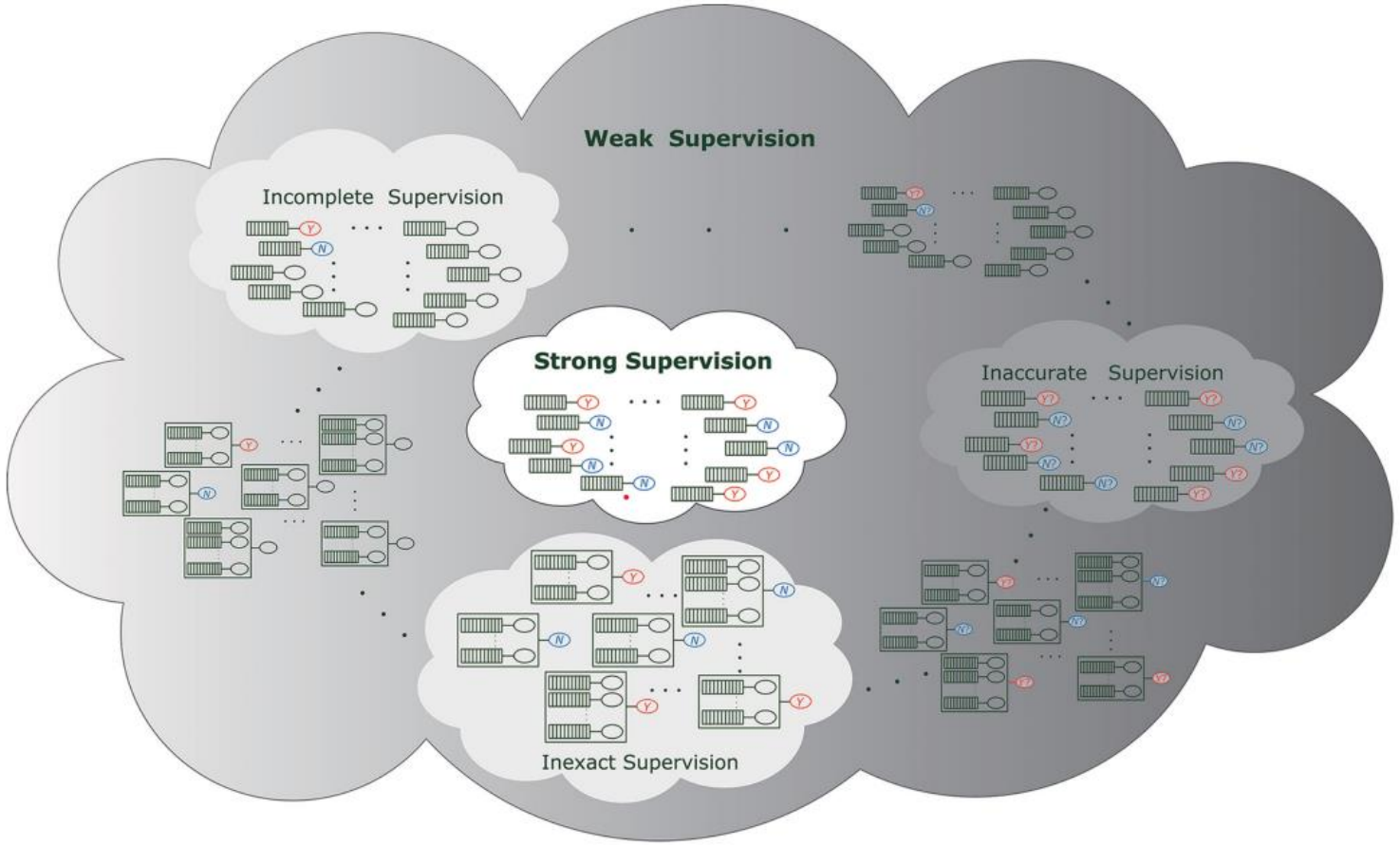
Is there a cat?

Supervision Types



Weak Supervision Types

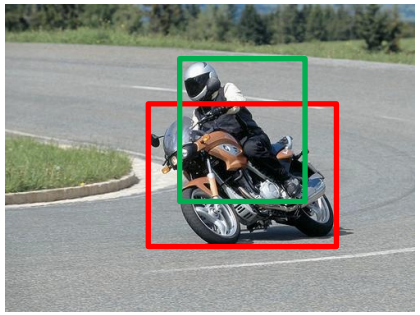
- Incomplete supervision: labeled + unlabeled
- Inaccurate supervision: noisy labels
- Inexact supervision: coarse labels



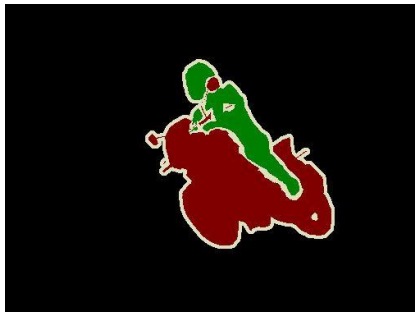
Weak supervision: How to save costs on labelling.

{motorbike,person}

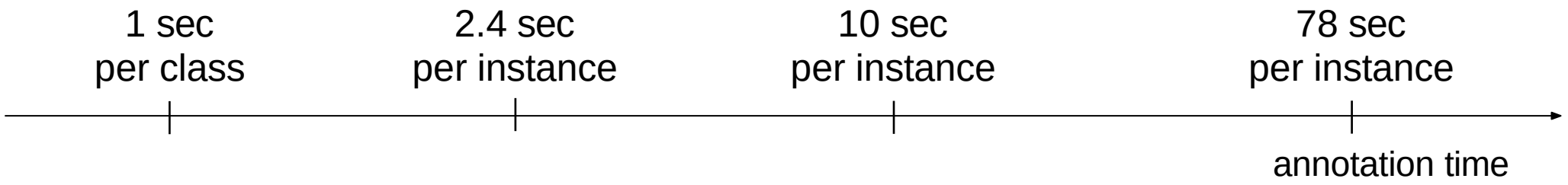
{motorbike (point), person (point)}



{motorbike (b-box), person (b-box)}



{motorbike (pixel labels), person (pixel labels)}



Berman et al., What's the Point: Semantic Segmentation with Point Supervision, ECCV'16.


sky
umbrella
umbrella
umbrella
umbrella
person
person
person
person
person
umbrella
tree
car
grass
pavement

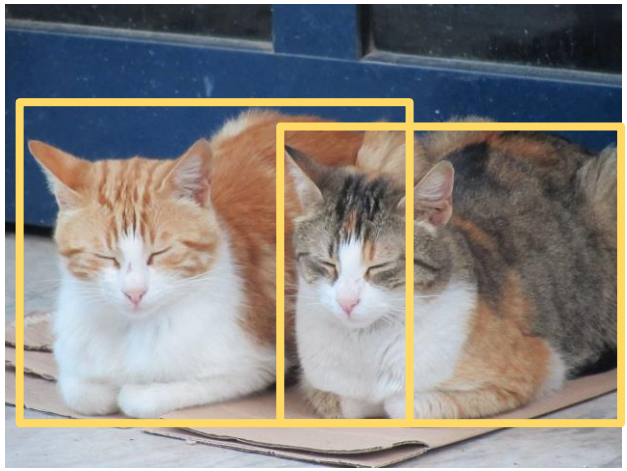
Classification

Is there a cat?

Classification

Is there a cat?

Detection

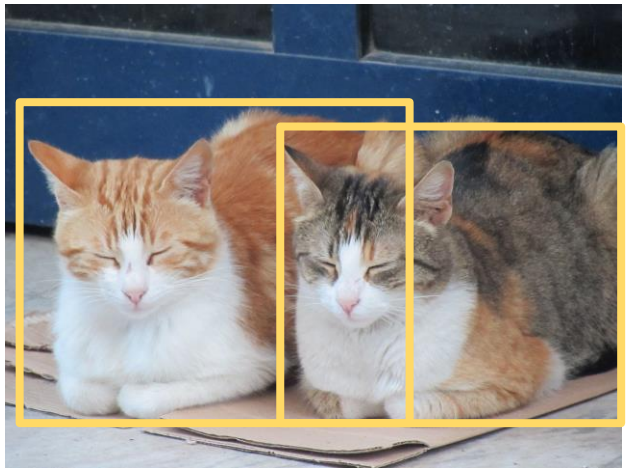


Where is the cat?

Classification

Is there a cat?

Detection



Where is the cat?

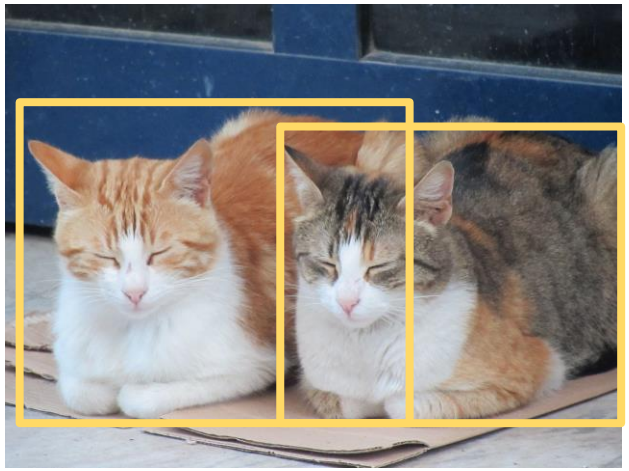
Semantic segmentation

Which pixel is cat ?

Classification

Is there a cat?

Detection



Where is the cat?

Semantic segmentation

Which pixel is cat ?

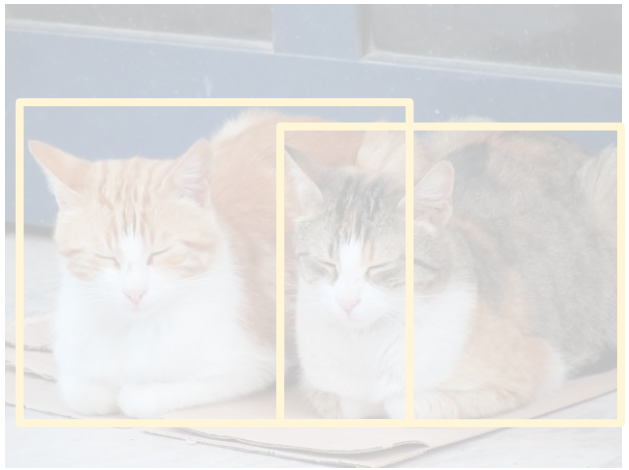
Instance segmentation

Which are the cat’s pixels?

Classification

Is there a cat?

Detection



Where is the cat?

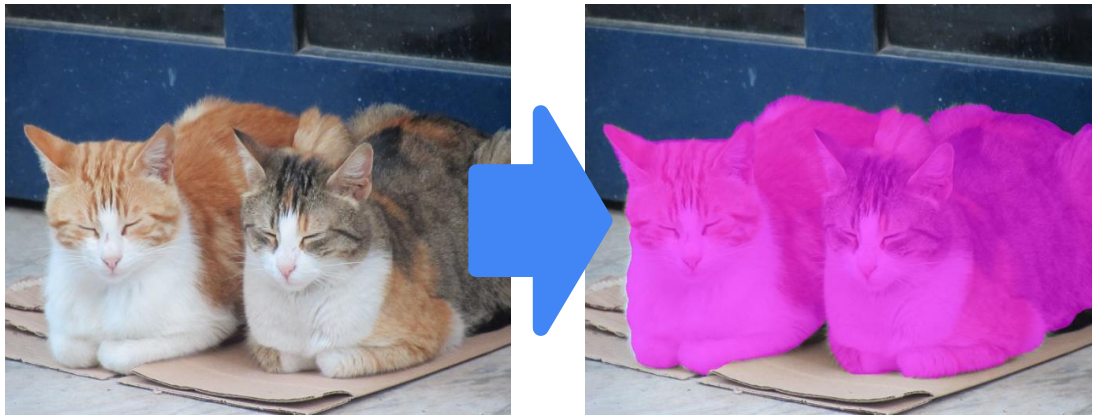
Semantic segmentation

Which pixel is cat ?

Instance segmentation

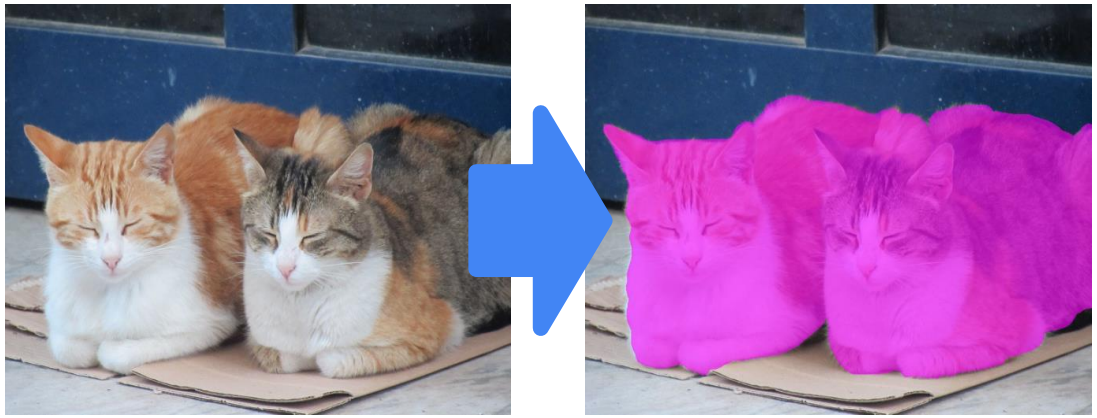
Which are the cat's pixels?

Fully supervised → **Directly** supervised

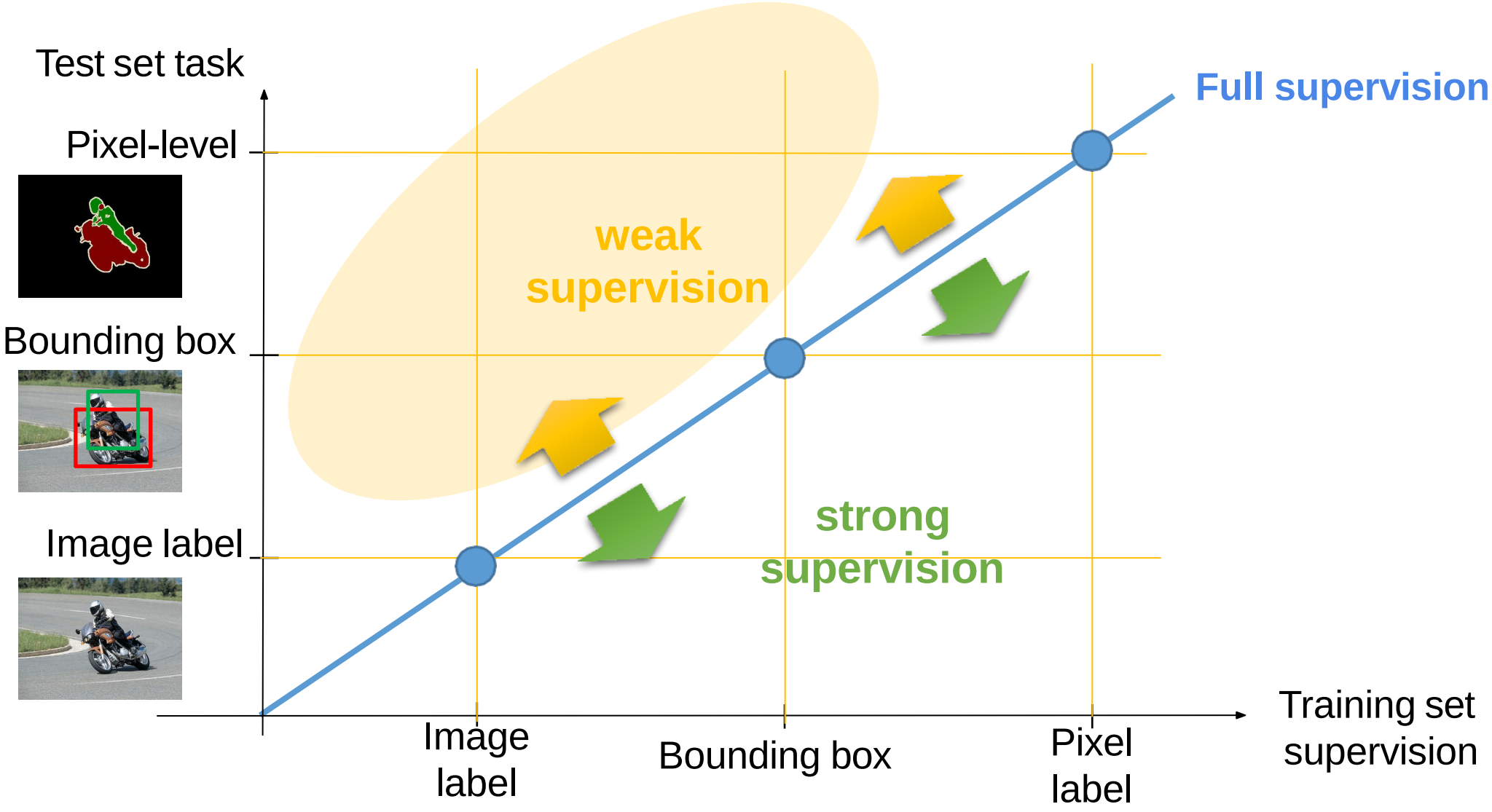


For each input, the desired output is provided

Fully supervised → **Directly** supervised



Weakly supervised → **Indirectly** supervised


Test set task
Pixel-level
Bounding box
Image label
weak
supervision
strong
supervision
Full supervision
Image
label
Bounding box
Pixel
label
Training set
supervision

Priors + Hints =

Weakly Supervised Algorithms

1. Size of the Object

Which cat is in the right size?

2. Shape of the Object

Which cat is in the right shape?

3. Object location

Which cat is in the right location?

4. Colour Contrast



Which pixels separate cats?

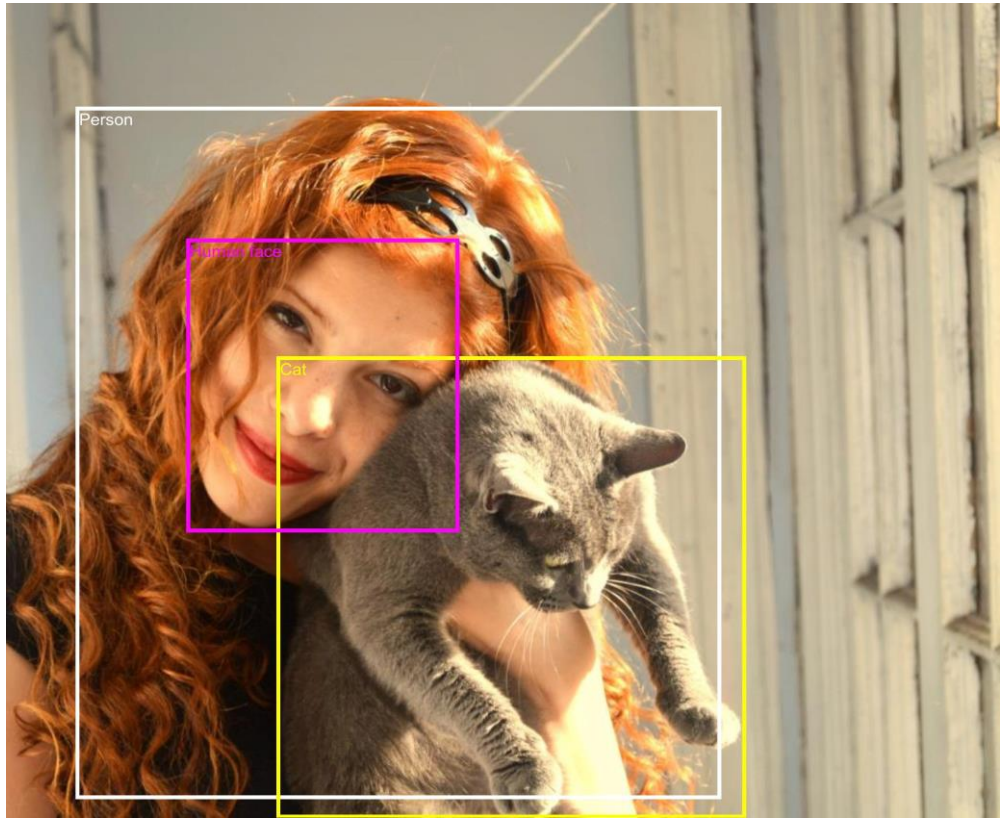
1. Image Labels

Person, cat, door

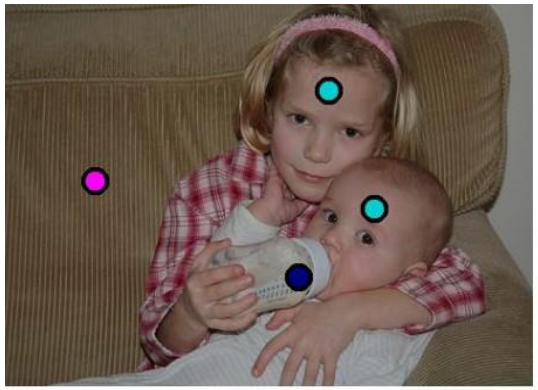
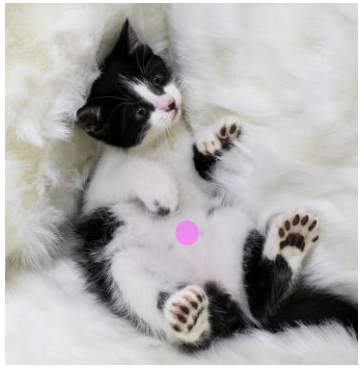
2. Image Captions

Kattie holds the cat next to the door

3. Object Bounding Boxes



4. Object Points


Click inside the objects

5. Scribbles

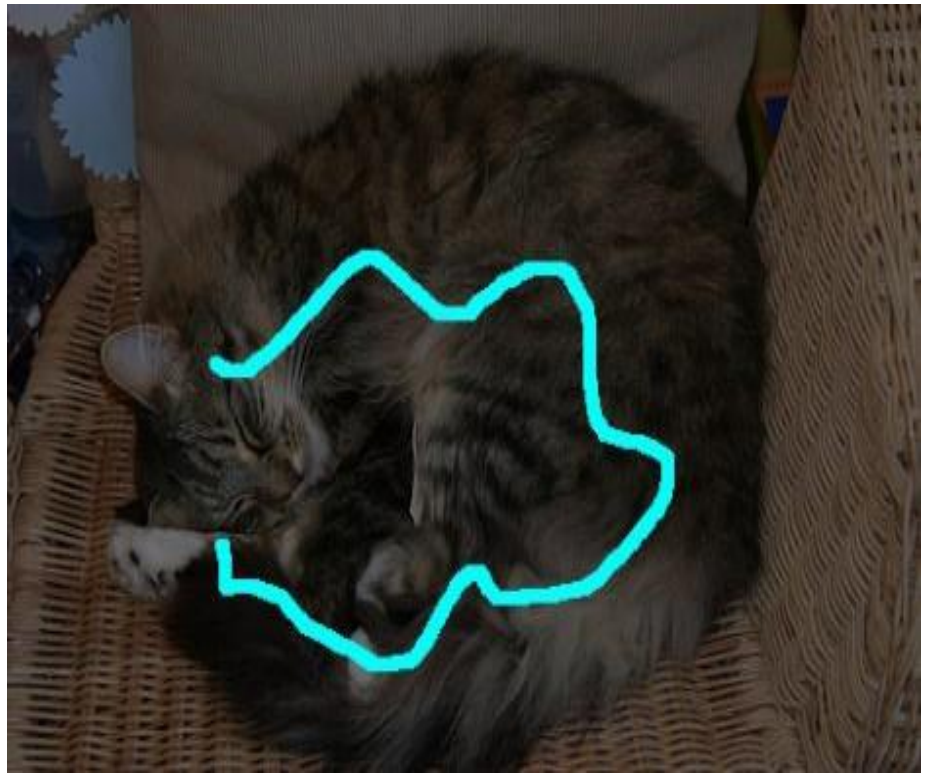


Image-level weakly-supervised semantic segmentation

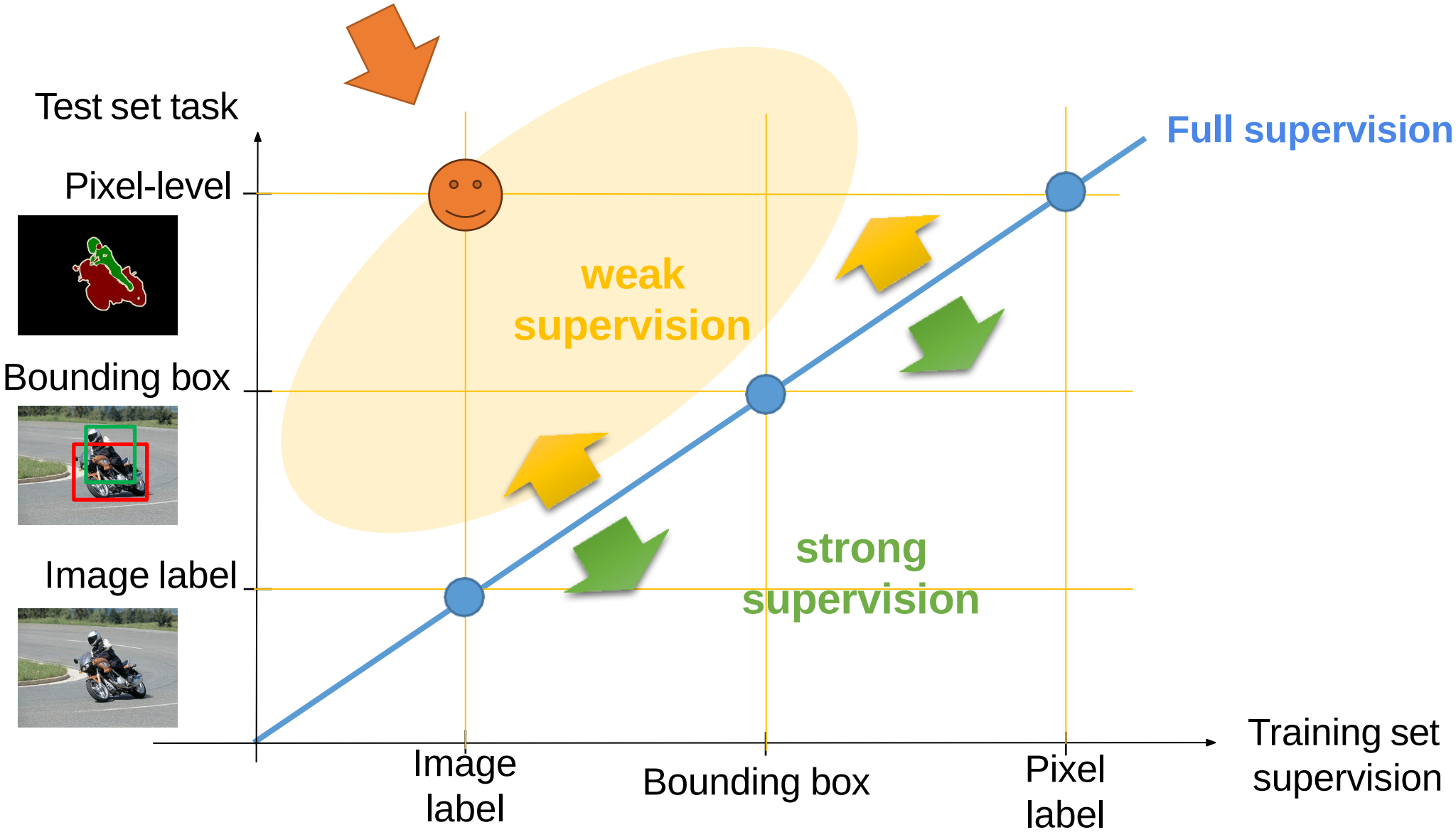
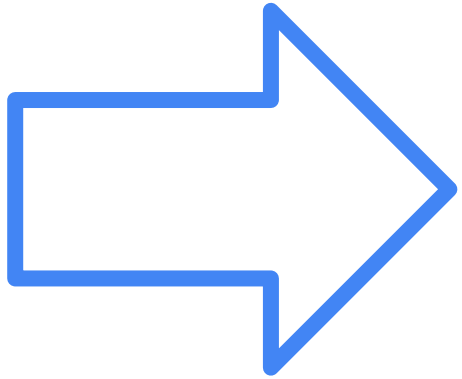
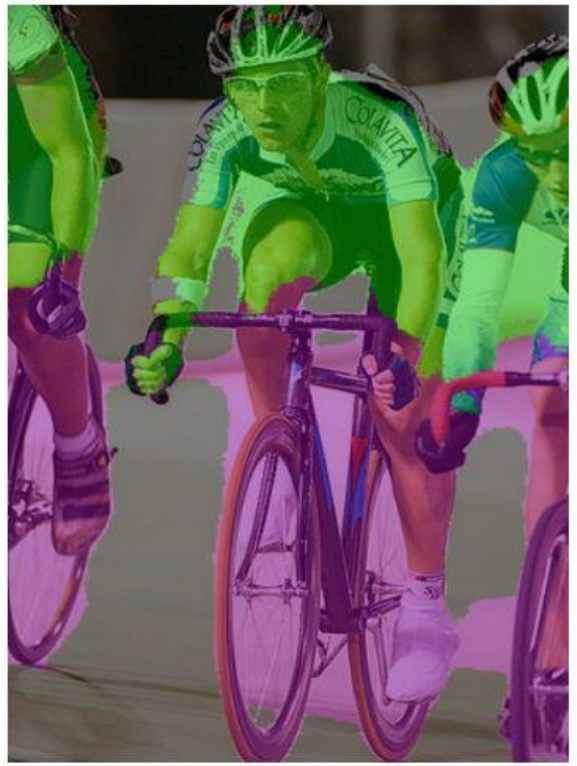


Image-level weakly-supervised semantic segmentation

Image-level labels
at training time





Pixel-level labels
at test time

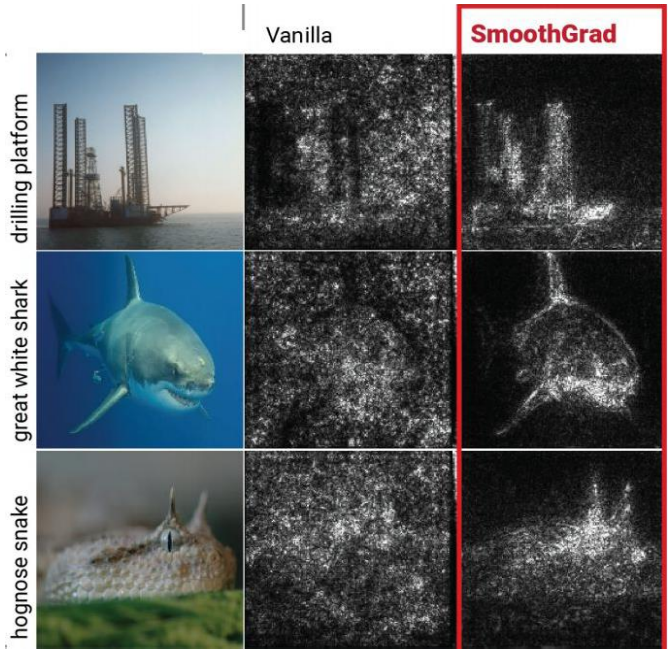
Smilkov et al. 2017

Step 1: Train an image classifier
Step 2: **Add small random noise** to image Step 3: given known label,
compute

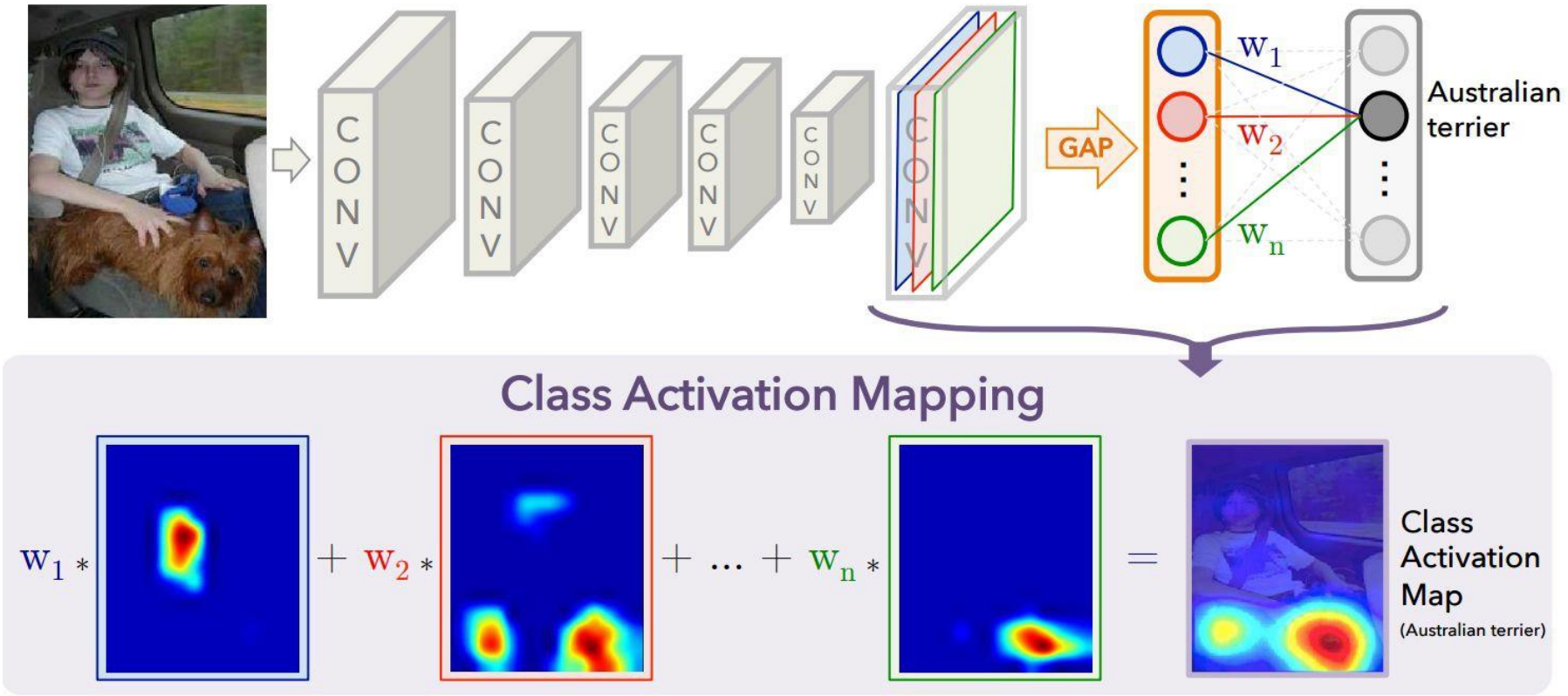
$$\left.\frac{\partial S_c}{\partial I}\right|_{I_0}$$

("how much does changes in the input image affect the score?")

Step 4: **Repeat** 50 times step 2&3, **average result**



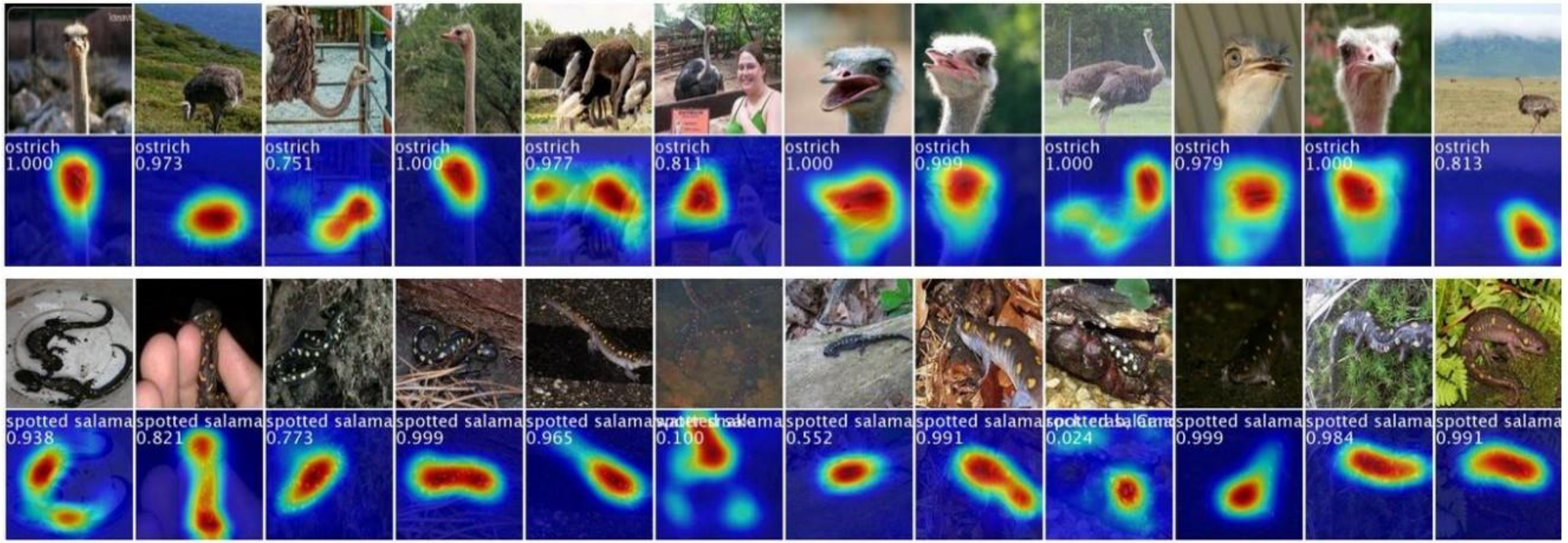
Zhou et al. 2016



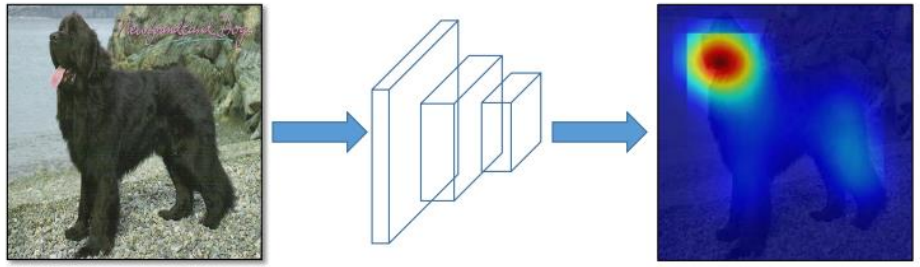
CAM: $M_c(x, y) = \sum_k w_k^c f_k(x, y).$

Good, but not good enough:

1. Resulting masks are not sharp
2. Focused on discriminative area only



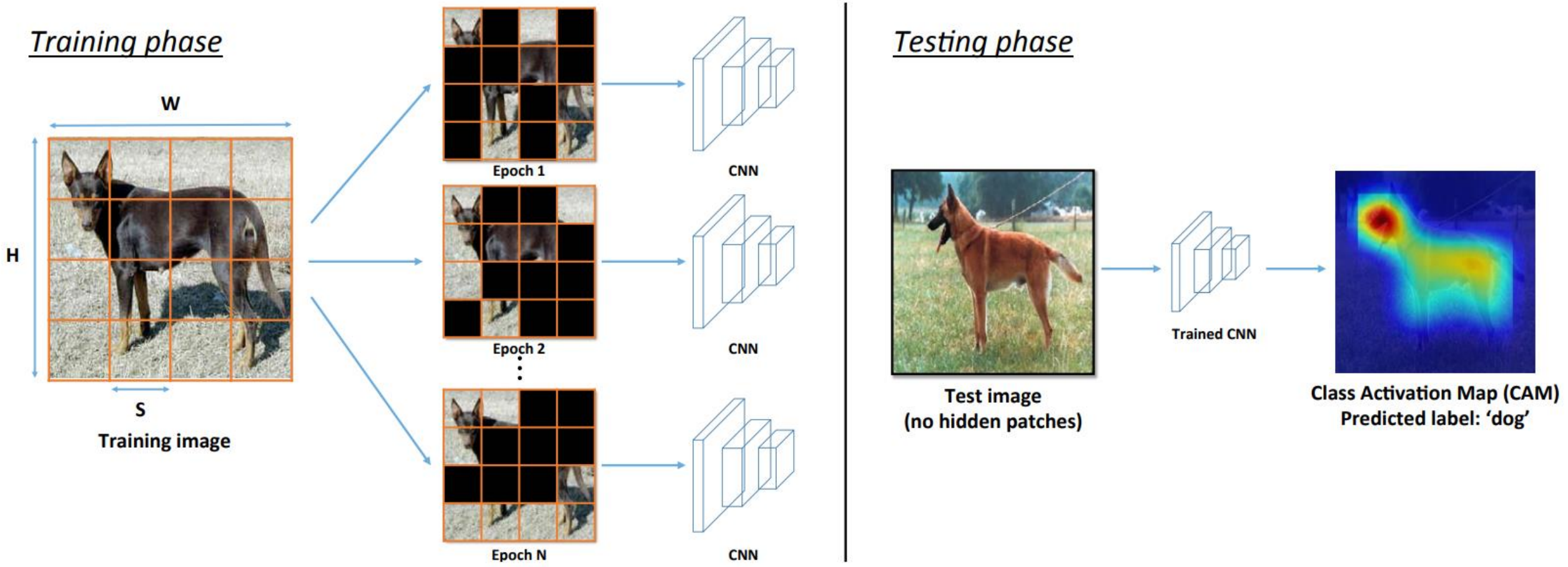
Improvement



Full image

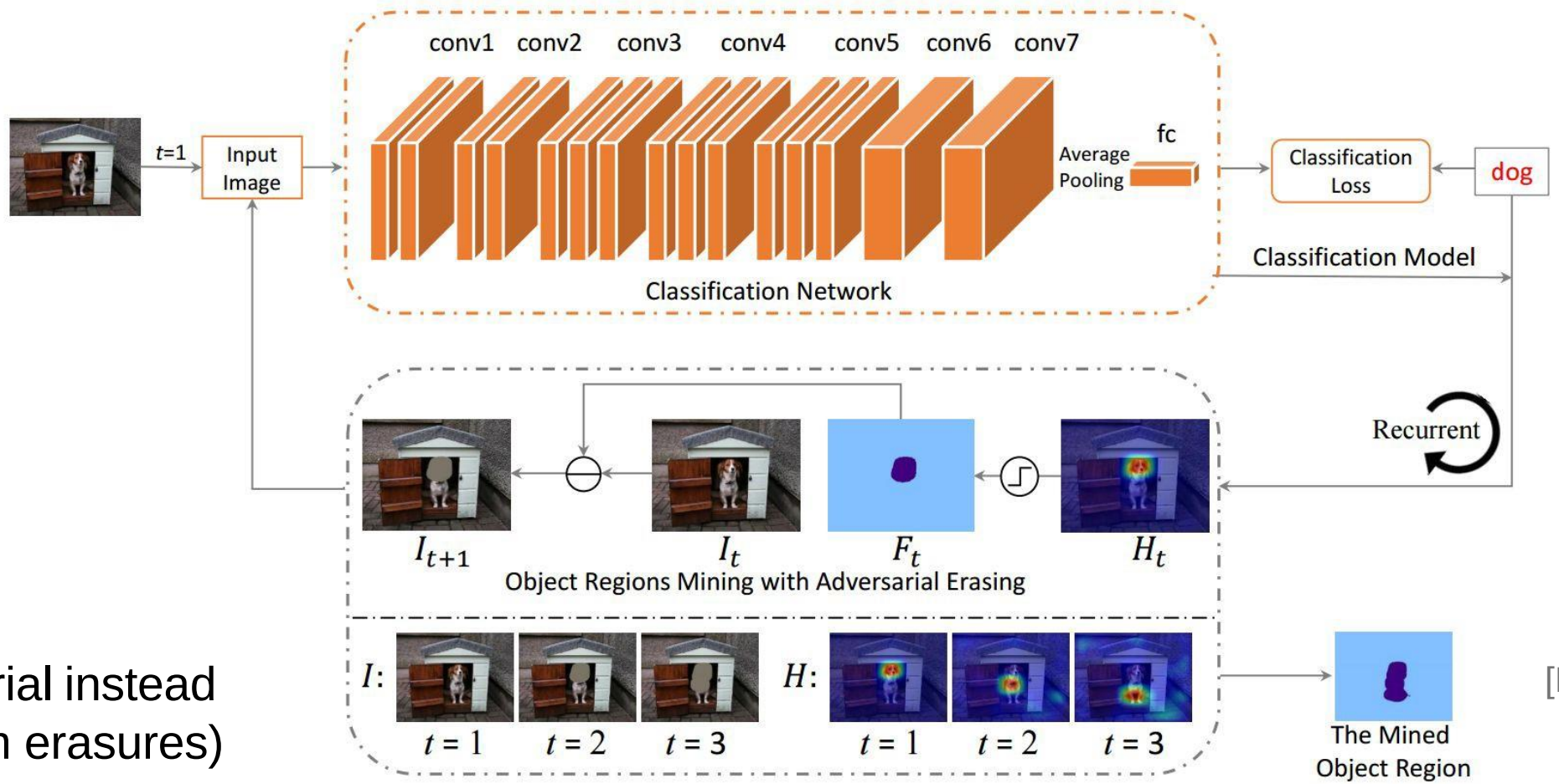
[Singh & Lee 2017]

Hide-and-Seek Algorithm



[Singh & Lee 2017]

Adversarial Erasing



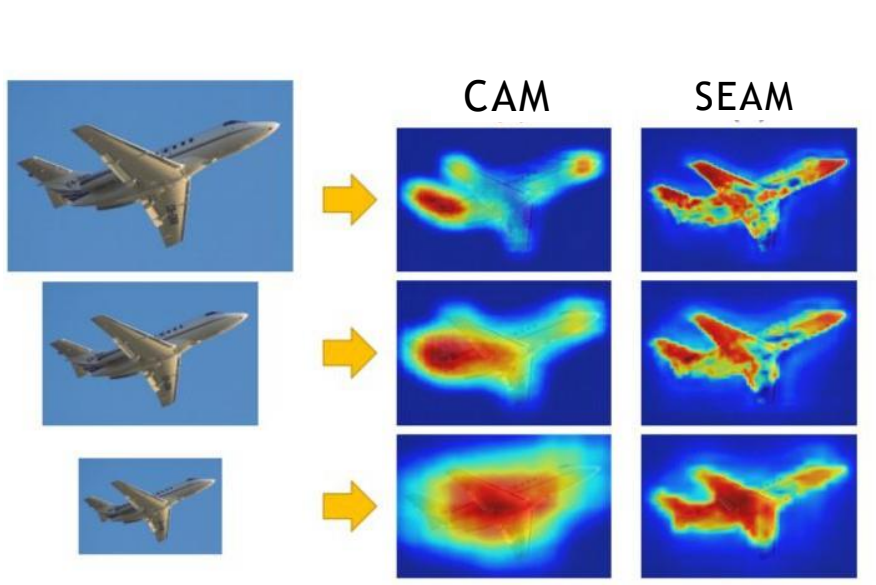
(adversarial instead of random erasures)

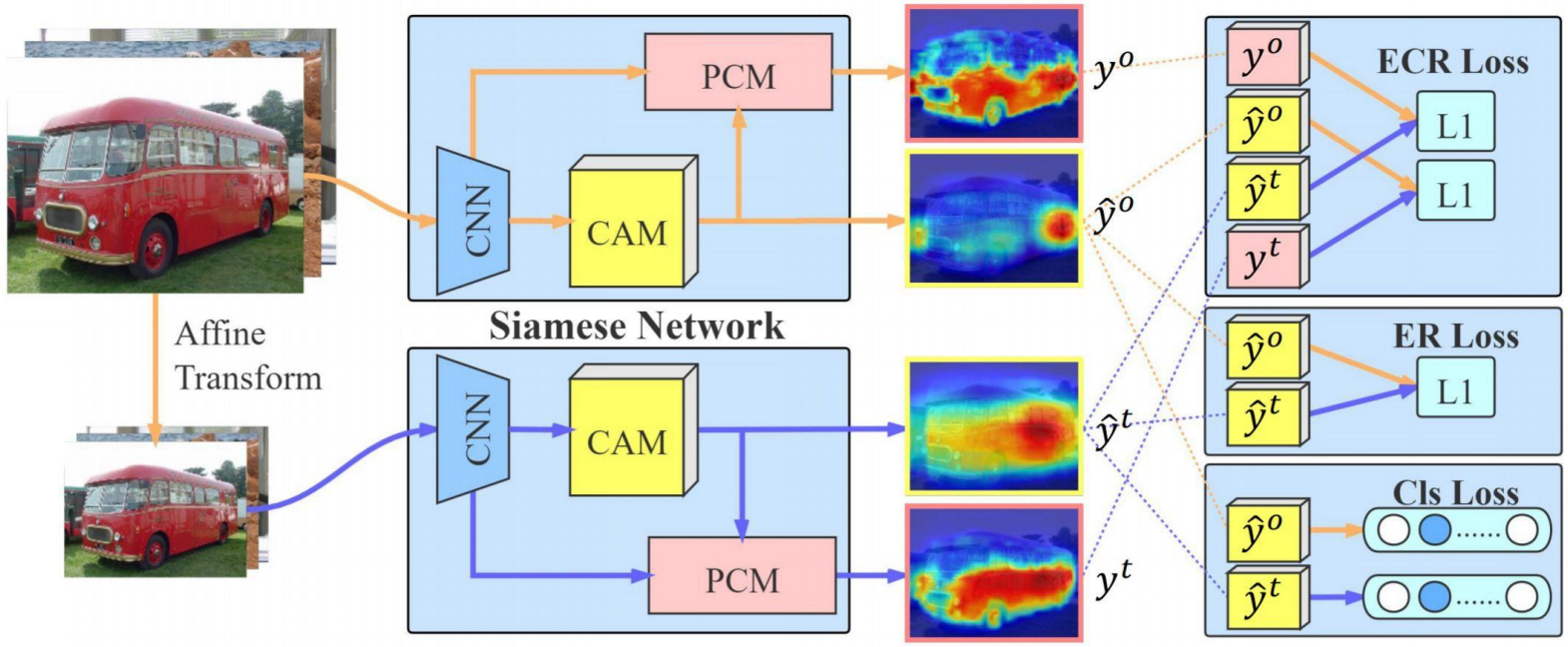
(Similar in spirit in to [Fong & Vedaldi 2017])

[Wei et al. 2017]

Invariant to Geo-transformation

"Activation maps should be equivariant to the input"





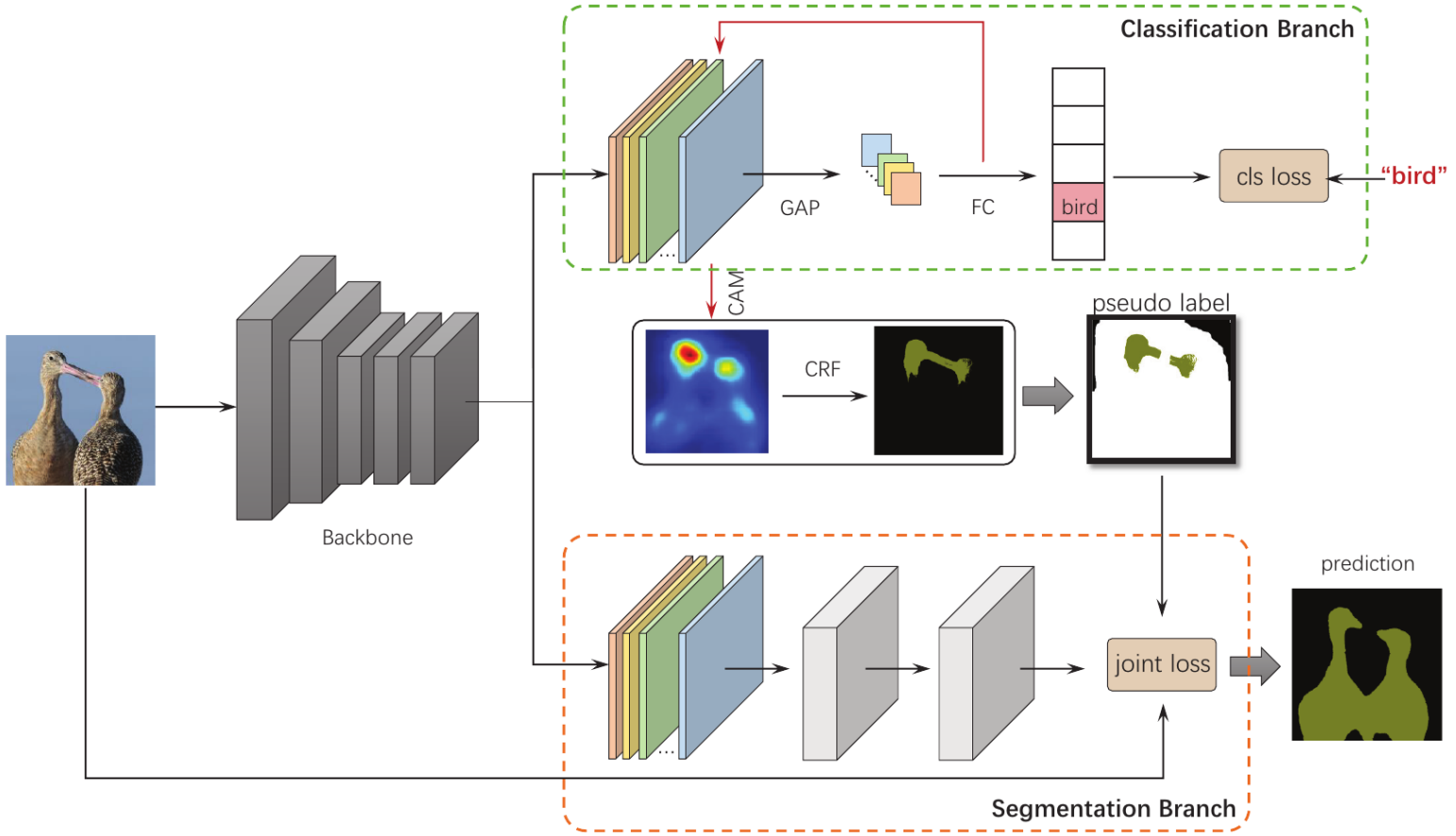
$$\mathcal{R}_{ER} = ||F(A(I)) - A(F(I))||_1.$$

[Wang et al. CVPR 2020]

How to do weakly-supervised segmentation

- Step1: train an image classifier.
- Step2: generate activation maps.
- Step3: refine the activation maps and produce pseudo-labels
- Step4: train segmentation model using pseudo-labels

Unified Model for weakly supervised segmentation

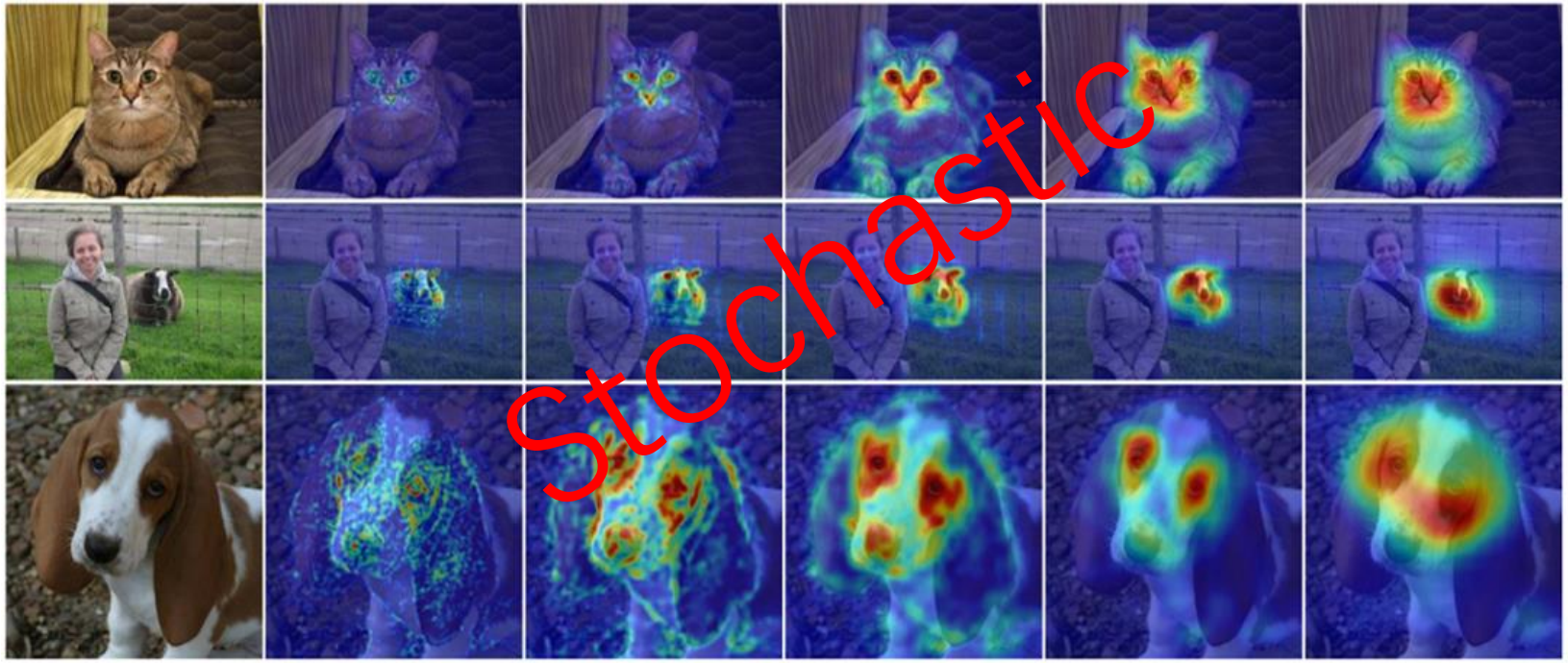


[Zhang et al. AAAI 2020]

Further improvement

CAM is

- Incomplete activation
- Not robust to input variation



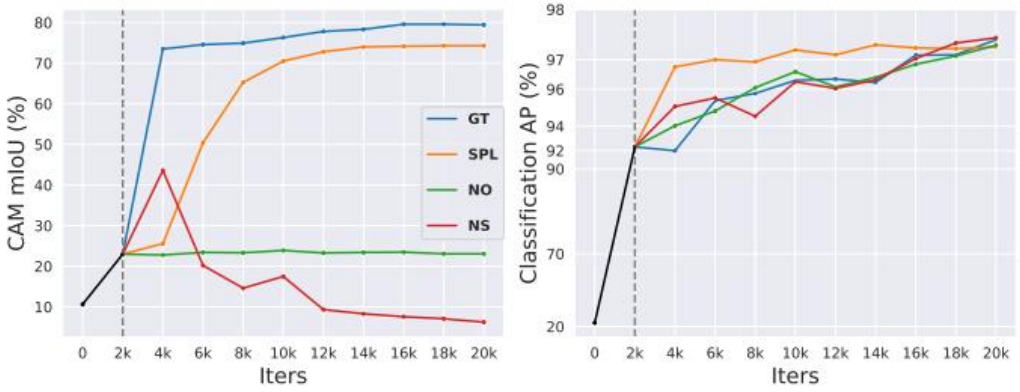
Proof of Concept

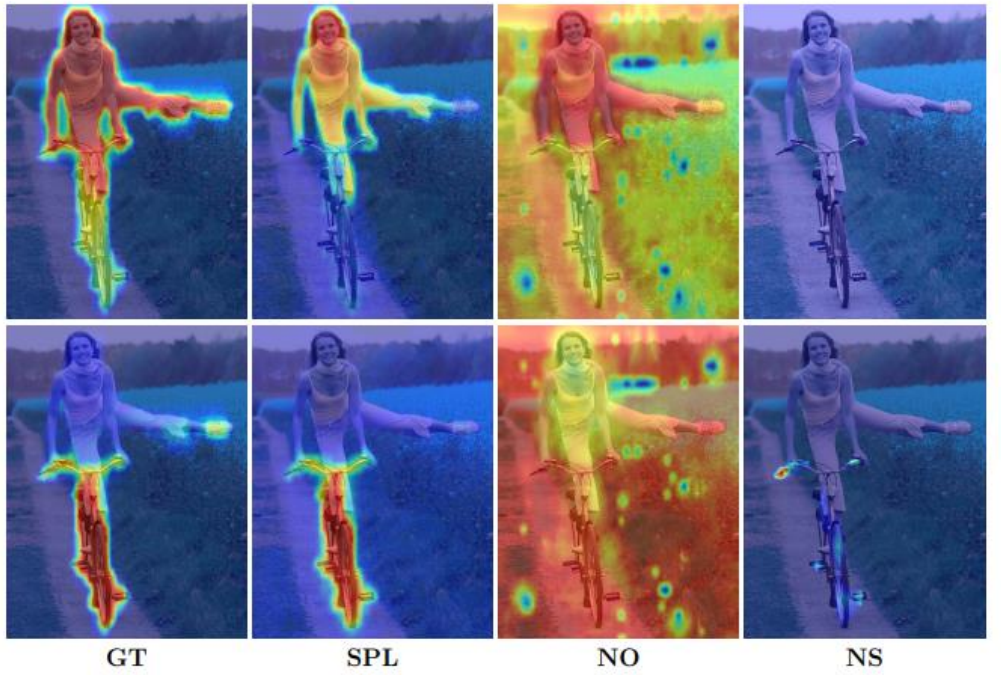
- CAM is differentiable
- But not fully optimized
- Thus it shows stochastic behaviours

Experiments:

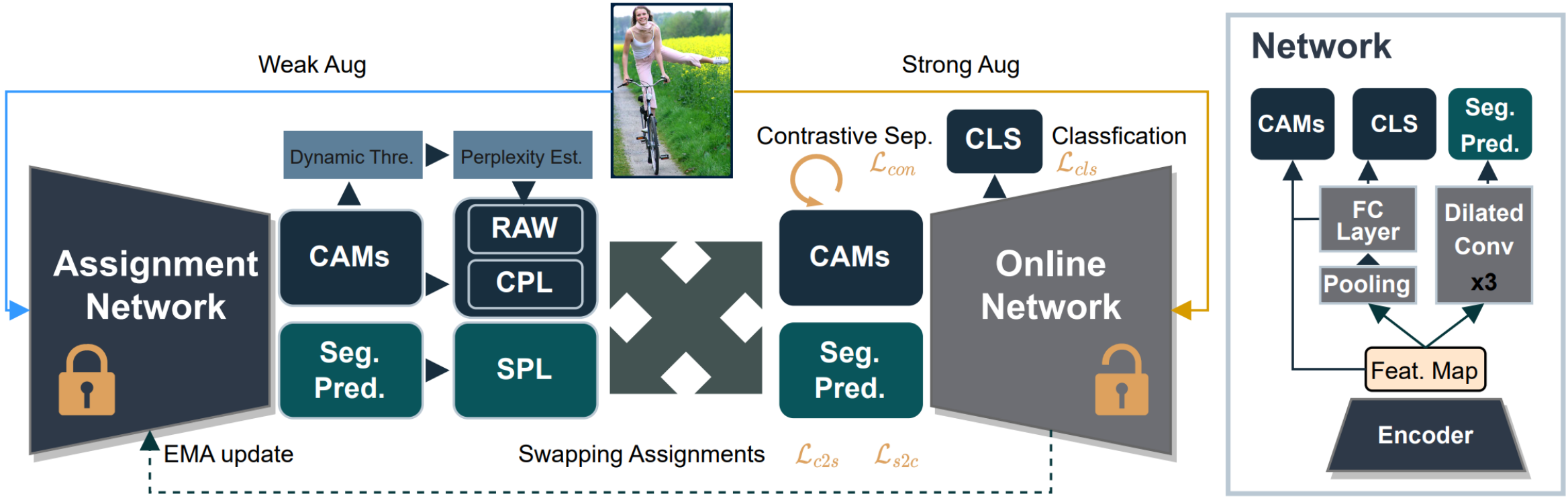
- GT: guided by ground truth (pixel-label)
- NO: no guidance
- NS: guided by random noise.

- But cannot use pixel-label
- Segmentation pseudo-label (SPL) as the substitute



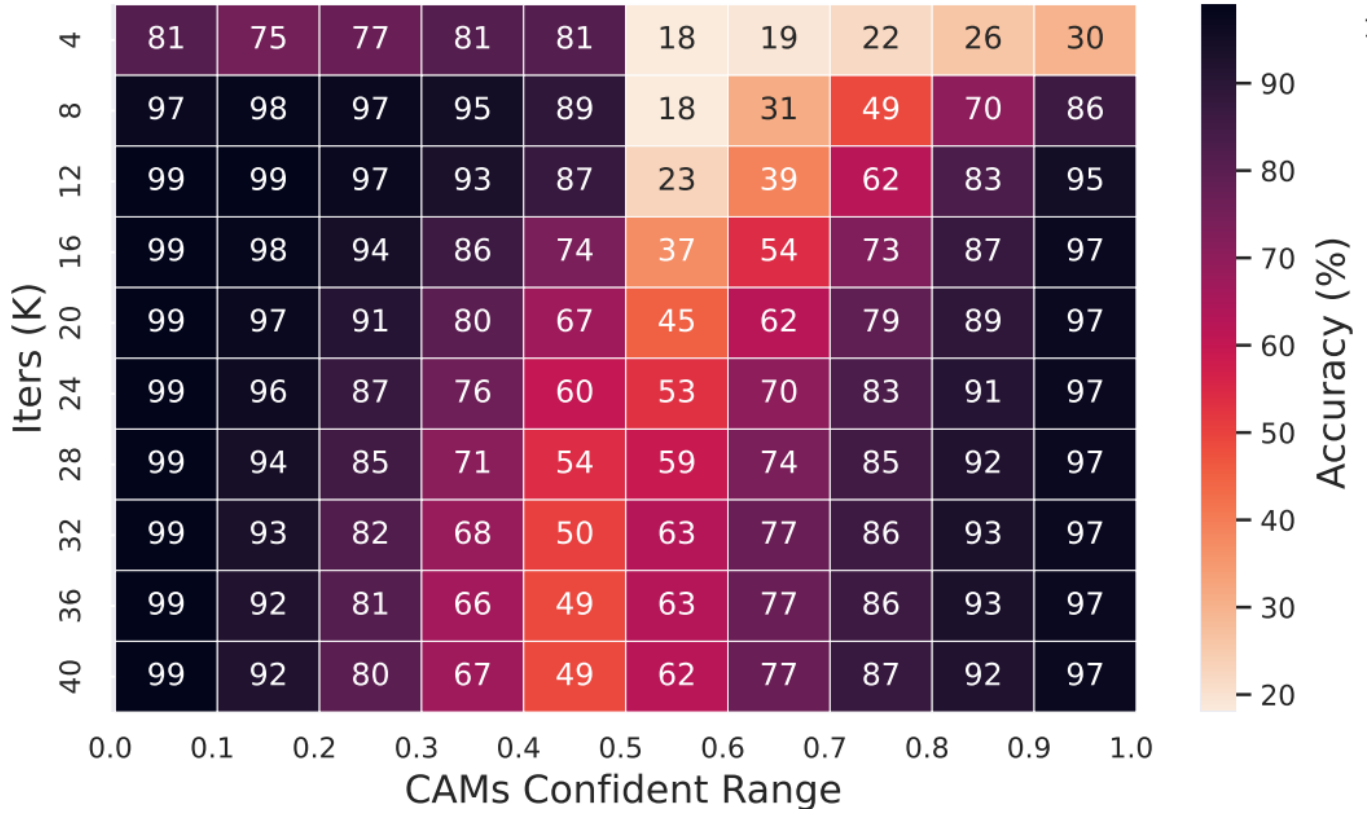


Our Recent Work



Reliability based Adaptive Weighting

- Pseudo-label is noisy label
- Need to be accessed before usage



Reliability based Adaptive Weighting (RAW)

- We can define the perplexity as

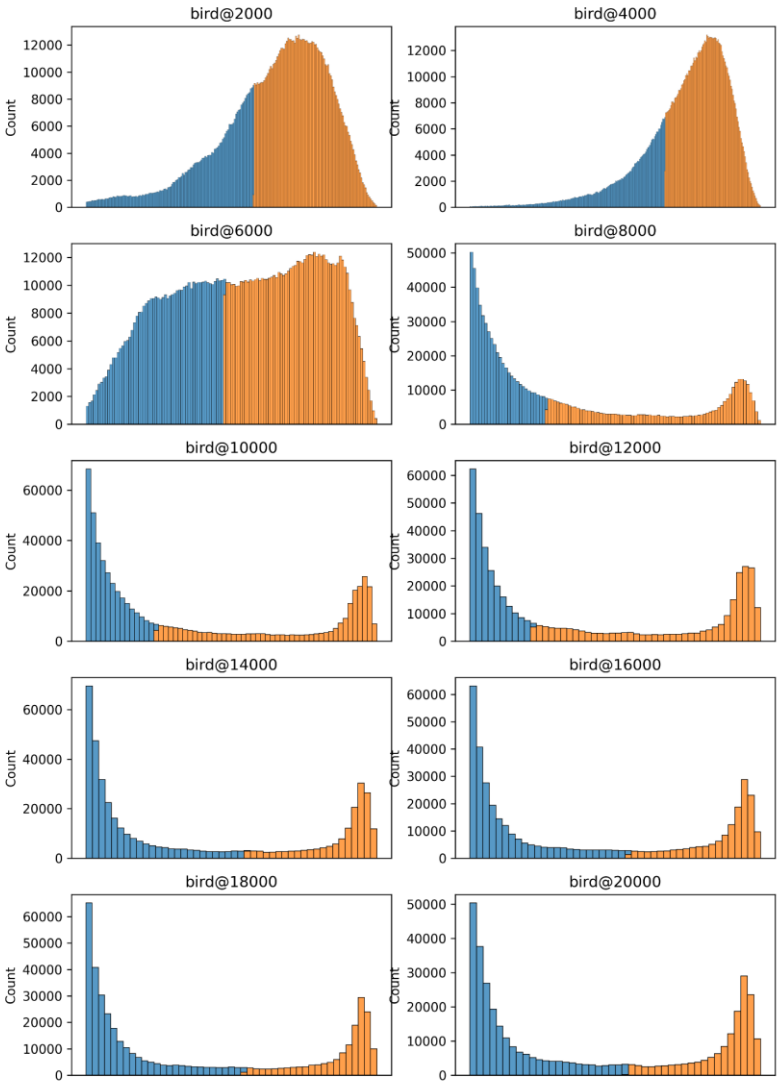
$$\mathcal{P}_{x,y} = \begin{cases} \left[-\log\left(\lambda_\alpha \dfrac{\max(\mathcal{M}'_{x,y}) - \xi}{1-\xi}\right)\right]^{\lambda_\beta} & \text{if } \max(\mathcal{M}'_{x,y}) \geq \xi, \\ \left[-\log\left(\lambda_\alpha \dfrac{\xi - \max(\mathcal{M}'_{x,y})}{\xi}\right)\right]^{\lambda_\beta} & \text{if } \max(\mathcal{M}'_{x,y}) < \xi, \end{cases}$$

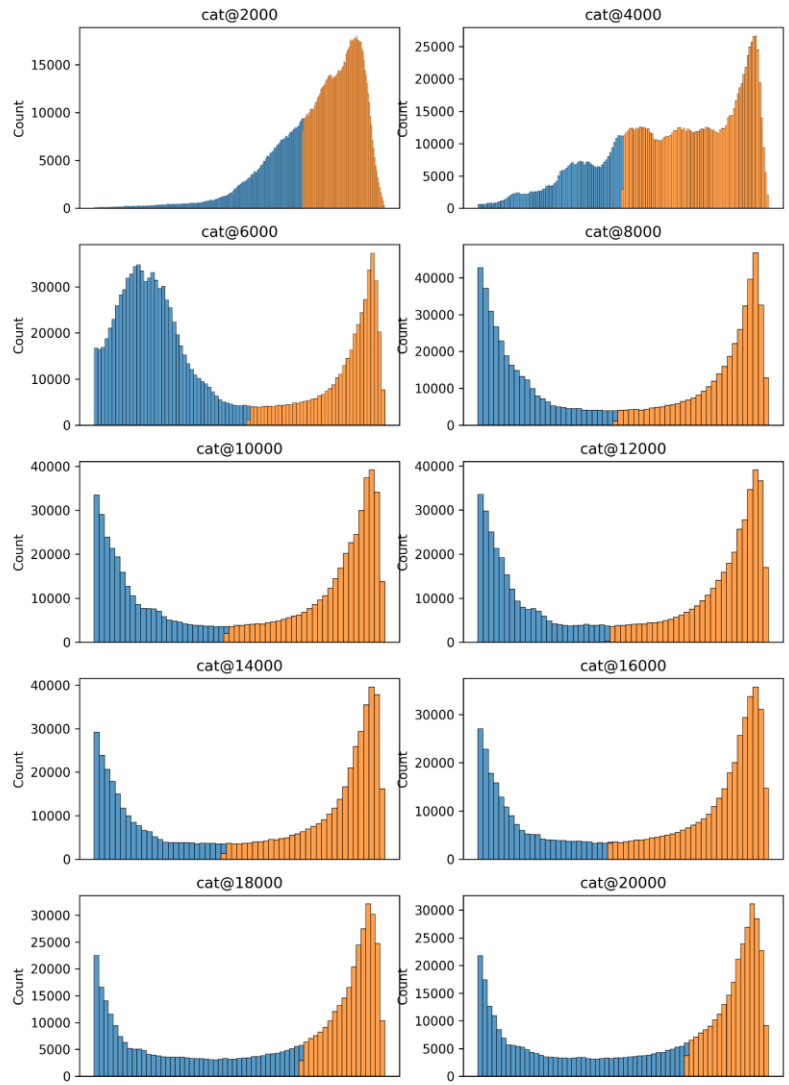
- RAW is the normalized reciprocal of perplexity

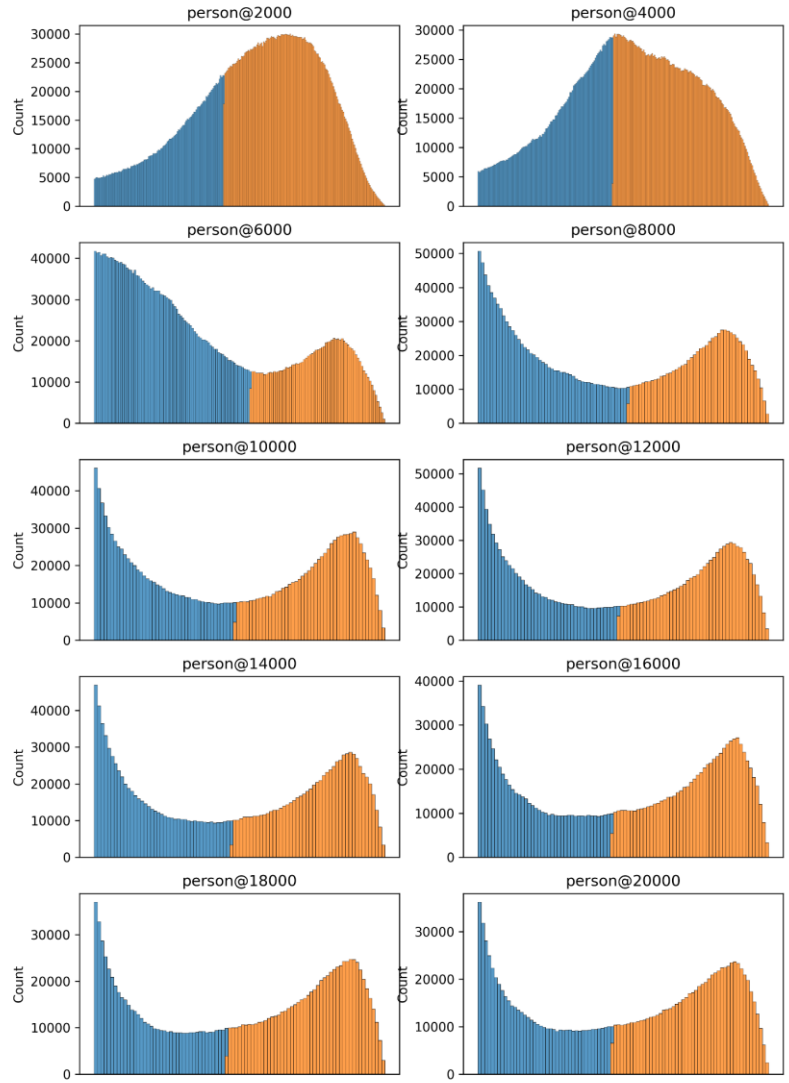
$$\mathcal{W}^{\text{raw}}_{x,y} = |\mathcal{R}| / \sum_{i,j \in \mathcal{R}} \frac{1}{\mathcal{P}_{i,j}\mathcal{P}_{x,y}},$$

Dynamic Threshold

- To generate CAMs pseudo-labels (CPL) a threshold is need.
- Previous works use a fixed threshold.

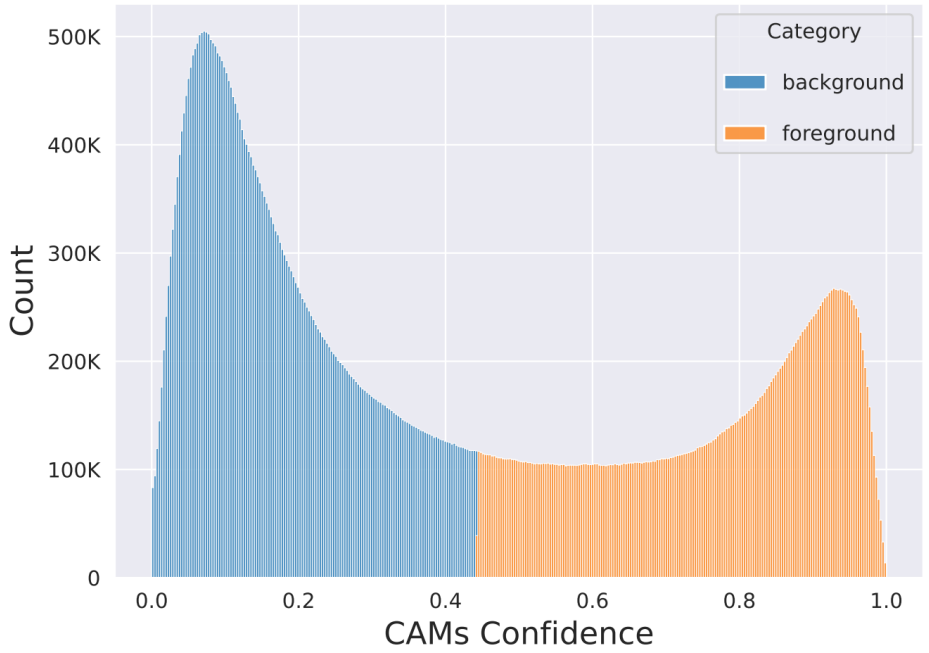

bird@2000
bird@4000
bird@6000
bird@8000
bird@10000
bird@12000
bird@14000
bird@16000
bird@18000
bird@20000
Count


cat@2000
cat@4000
cat@6000
cat@8000
cat@10000
cat@12000
cat@14000
cat@16000
cat@18000
cat@20000
Count


person@2000
person@4000
person@6000
person@8000
person@10000
person@12000
person@14000
person@16000
person@18000
person@20000
Count

Dynamic Threshold

- To generate CAMs pseudo-labels (CPL) a threshold is need.
- Previous works use a fixed threshold.
- The bi-modal distribution



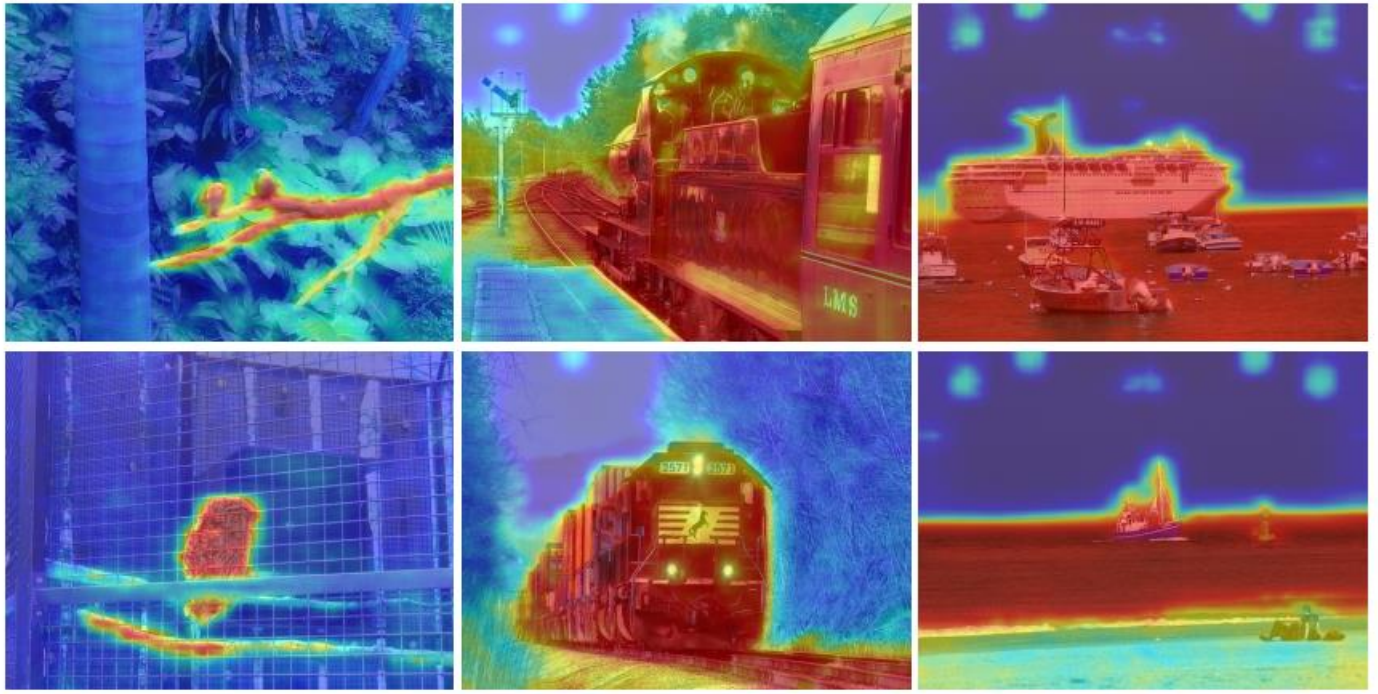
Dynamic Threshold

- To generate CAMs pseudo-labels (CPL) a threshold is need.
- Previous works use a fixed threshold.
- The bi-modal distribution
- Propose to use Gaussian Mixture Model (GMM) to find optimal threshold

$$\xi^{\star} = \underset{\xi}{\texttt{argmax}} \prod_{x \in \{\mathcal{M}' \geq \xi\}} \tilde{\pi}_{fg} \mathcal{N}\left(x \mid \tilde{\mu}_{fg}, \tilde{\Sigma}_{fg}\right) + \prod_{x \in \{\mathcal{M}' < \xi\}} \tilde{\pi}_{bg} \mathcal{N}\left(x \mid \tilde{\mu}_{bg}, \tilde{\Sigma}_{bg}\right) ,$$

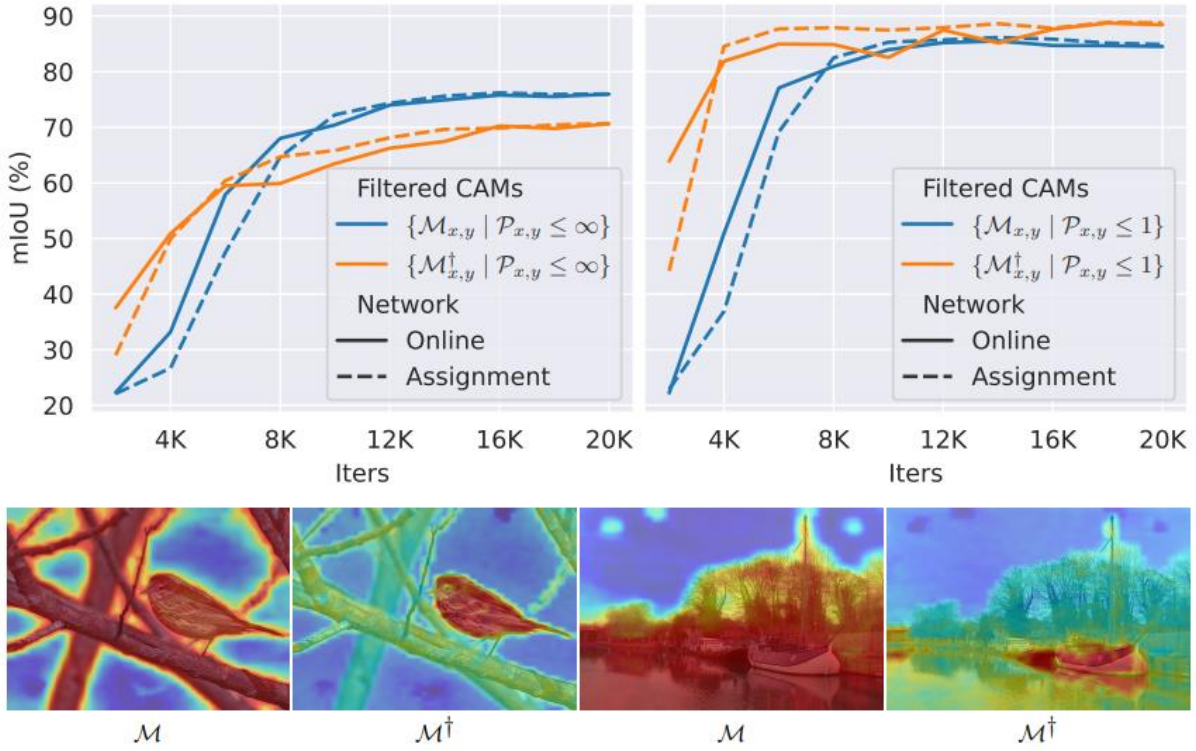
Contrastive Separation in CAMs

- Coexistence Problem in CAMs



Contrastive Separation in CAMs

- Auxiliary CAMs (from early stage of the model)



Contrastive Separation in CAMs

- Use Auxiliary CAMs filter by low perplexity to define some positive regions and negative regions

$$\mathcal{R}^{+}_{i,j} = \left\{ (x,y) \mid \mathcal{P}_{x,y} \leq \epsilon,\ \hat{y}^{\text{CPL}}_{x,y} = \hat{y}^{\text{CPL}}_{i,j},\ x \neq i,\ y \neq j \right\},$$
$$\mathcal{R}^{-}_{i,j} = \left\{ (x,y) \mid \mathcal{P}_{x,y} \leq \epsilon,\ \hat{y}^{\text{CPL}}_{x,y} \neq \hat{y}^{\text{CPL}}_{i,j} \right\},$$

- Contrastively separate these regions in CAMs by minimising the distance among the positives and maximising the distance among negatives.

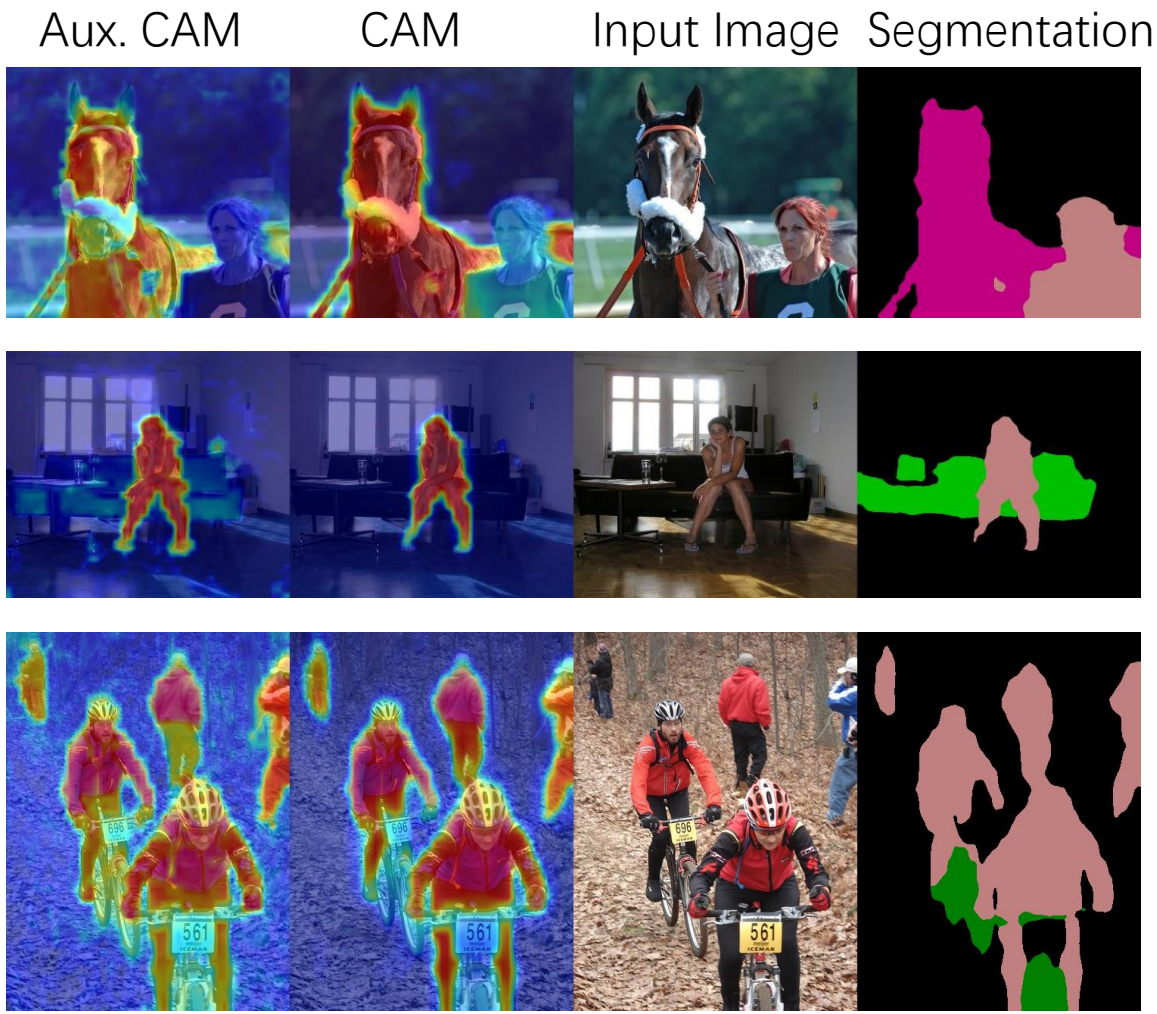
$$\mathcal{L}_{con} = -\frac{1}{|\{(x,y) \mid \mathcal{P}_{x,y} \leq \epsilon\}|} \sum_{i,j \in \{(x,y) | \mathcal{P}_{x,y} \leq \epsilon\}} \frac{1}{|\mathcal{R}^{+}_{i,j}|} \sum_{x,y \in \mathcal{R}^{+}_{i,j}}$$
$$\left[\log \frac{\exp\left(l_d\left(\mathcal{M}_{i,j}, \mathcal{M}_{x,y}\right)/\tau\right)}{\exp\left(l_d\left(\mathcal{M}_{i,j}, \mathcal{M}_{x,y}\right)/\tau\right) + \sum_{n,m \in \mathcal{R}^{-}_{i,j}} \exp\left(l_d\left(\mathcal{M}_{i,j}, \mathcal{M}_{n,m}\right)/\tau\right)} \right]$$

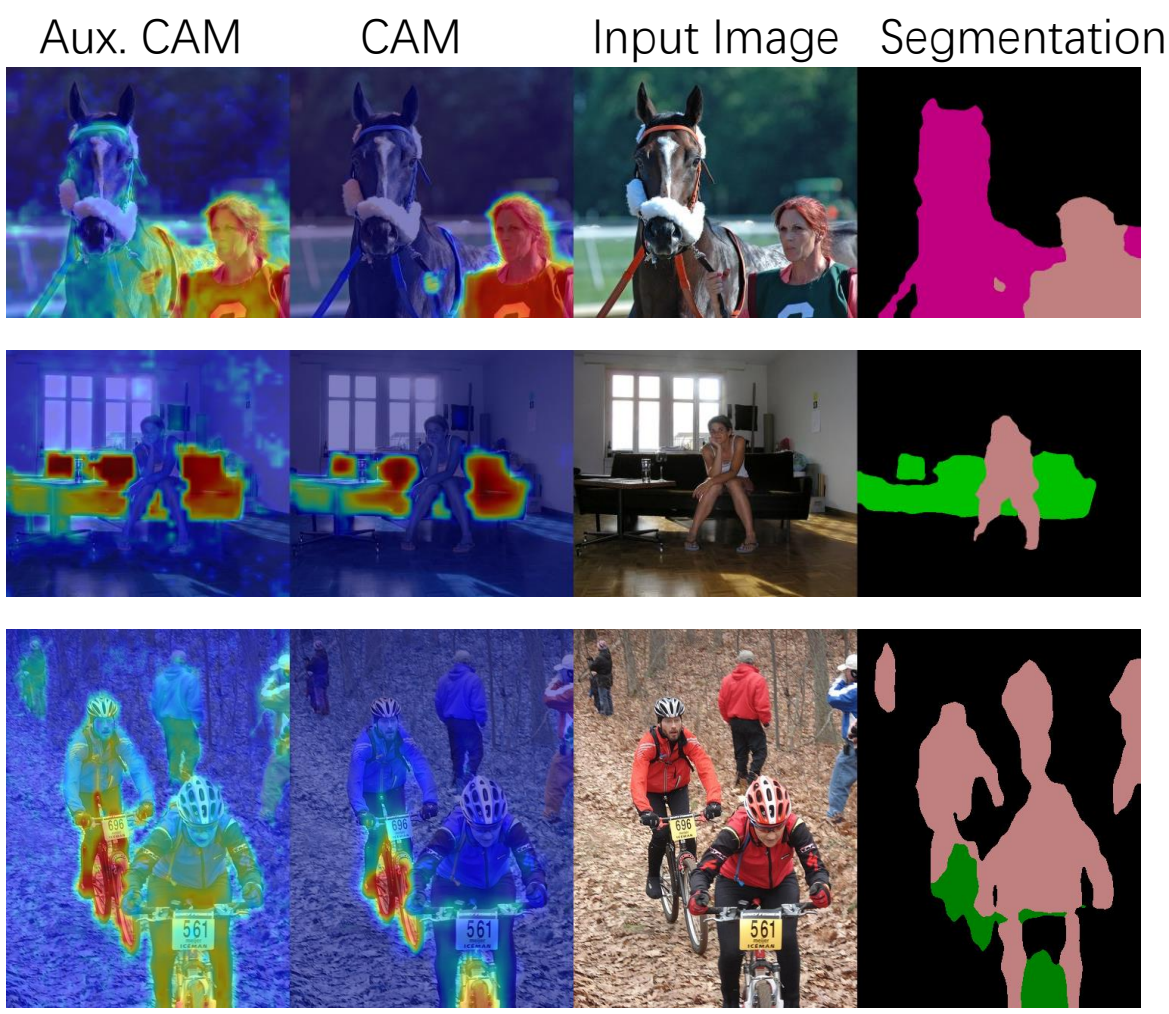
Results on VOC & COCO

- State-of-the-art performance compared with the recent methods.
- Our single-stage model outperforms multi-stage models and models that use additional modalities.

Methods	*Sup.*	*Net.*	VOC		COCO
			val	test	val
Supervised Upperbounds.					
Deeplab [10] TPAMI'2017	$\mathcal{F}$	R101	77.6	79.7	–
WideRes38 [59] PR'2019	$\mathcal{F}$	WR38	80.8	82.5	–
ViT-Base [17] ICLR'2021	$\mathcal{F}$	ViT-B	80.5	81.0	–
UperNet-Swin [42] ICCV'2021	$\mathcal{F}$	SWIN	83.4	83.7	–
Multi-stage Methods.					
L2G [26] CVPR'2022	$\mathcal{I}+\mathcal{S}$	R101	72.1	71.7	44.2
Du *et al.* [18] CVPR'2022	$\mathcal{I}+\mathcal{S}$	R101	72.6	73.6	–
CLIP-ES [40] CVPR'2023	$\mathcal{I}+\mathcal{L}$	R101	73.8	73.9	45.4
ESOL [37] NeurIPS'2022	$\mathcal{I}$	R101	69.9	69.3	42.6
BECO [48] CVPR'2023	$\mathcal{I}$	R101	72.1	71.8	45.1
Mat-Label [57] ICCV'2023	$\mathcal{I}$	R101	73.0	72.7	45.6
CoSA-MS	$\mathcal{I}$	R101	**76.5**	**75.3**[1]	**50.9**
Xu *et al.* [63] CVPR'2023	$\mathcal{I}+\mathcal{L}$	WR38	72.2	72.2	45.9
W-OoD [33] CVPR'2022	$\mathcal{I}$	WR38	70.7	70.1	–
MCT [62] CVPR'2022	$\mathcal{I}$	WR38	71.9	71.6	42.0
ex-ViT [66] PR'2023	$\mathcal{I}$	WR38	71.2	71.1	42.9
ACR-ViT [31] CVPR'2023	$\mathcal{I}$	WR38	72.4	72.4	–
MCT+OCR [15] CVPR'2023	$\mathcal{I}$	WR38	72.7	72.0	42.0
CoSA-MS	$\mathcal{I}$	WR38	**76.6**	**74.9**[2]	**50.1**
ReCAM [14] CVPR'2022	$\mathcal{I}$	SWIN	70.4	71.7	47.9
LPCAM [12] CVPR'2023	$\mathcal{I}$	SWIN	73.1	73.4	48.3
CoSA-MS	$\mathcal{I}$	SWIN	**81.4**	**78.4**[3]	**53.7**
Single-stage (End-to-end) Methods.					
1Stage [3] CVPR'2020	$\mathcal{I}$	WR38	62.7	64.3	–
RRM [67] AAAI'2020	$\mathcal{I}$	WR38	62.6	62.9	–
AFA [50] CVPR'2022	$\mathcal{I}$	MiT-B1	66.0	66.3	38.9
RRM [67]† AAAI'2020	$\mathcal{I}$	ViT-B	63.1	62.4	–
ViT-PCM [49] ECCV'2022	$\mathcal{I}$	ViT-B	69.3	–	45.0
ToCo [51] CVPR'2023	$\mathcal{I}$	ViT-B	71.1	72.2	42.3
CoSA	$\mathcal{I}$	ViT-B	76.2	75.1[4]	51.0
CoSA*	$\mathcal{I}$	ViT-B	**76.4**	**75.2**[5]	**51.1**

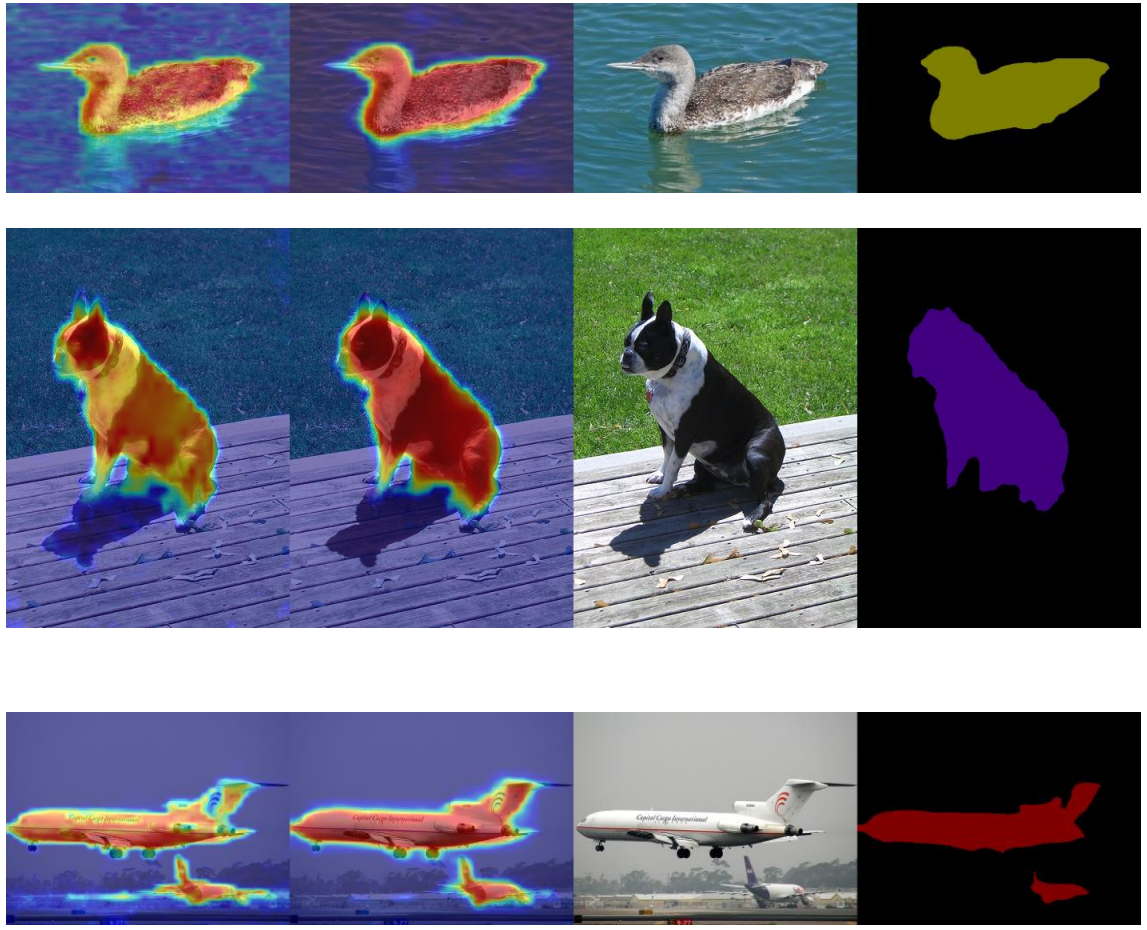
Results on VOC



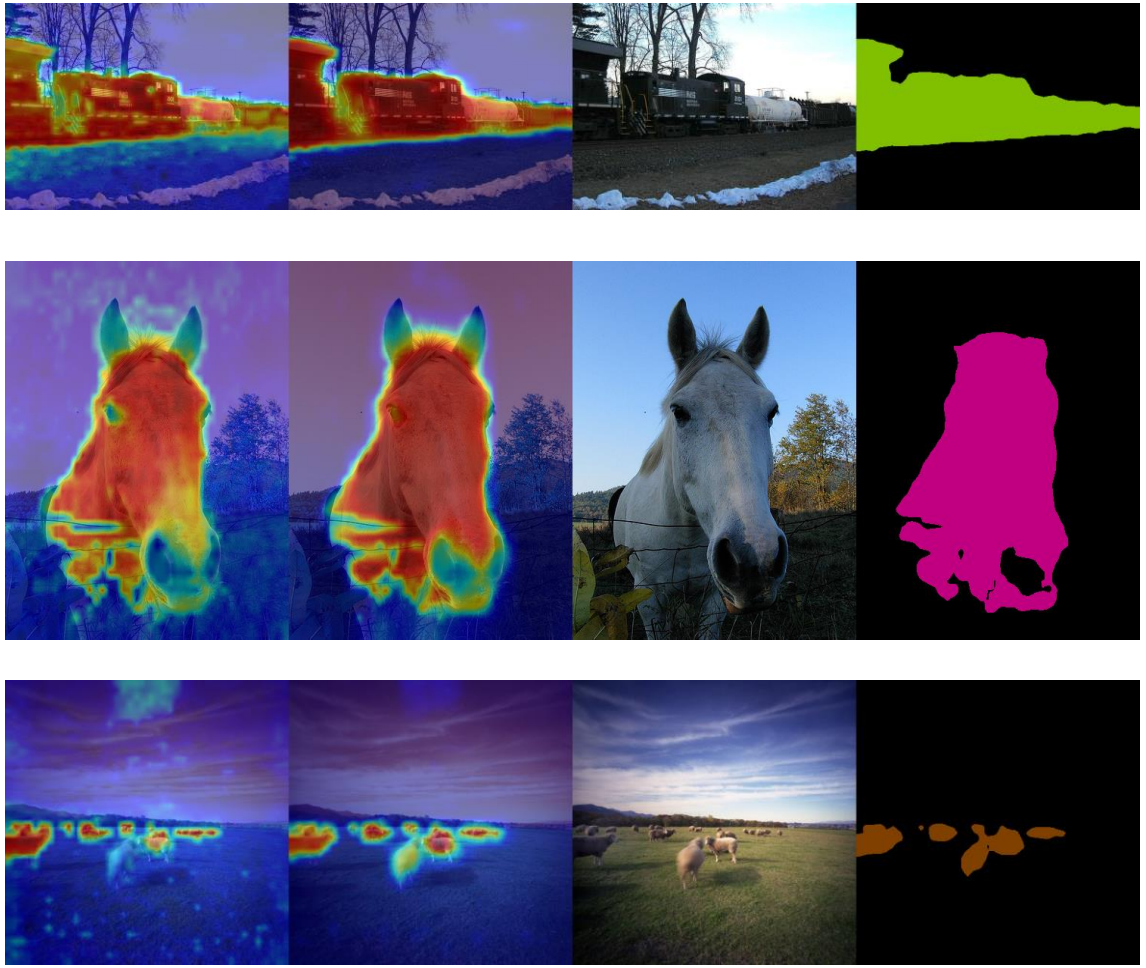


Results on VOC

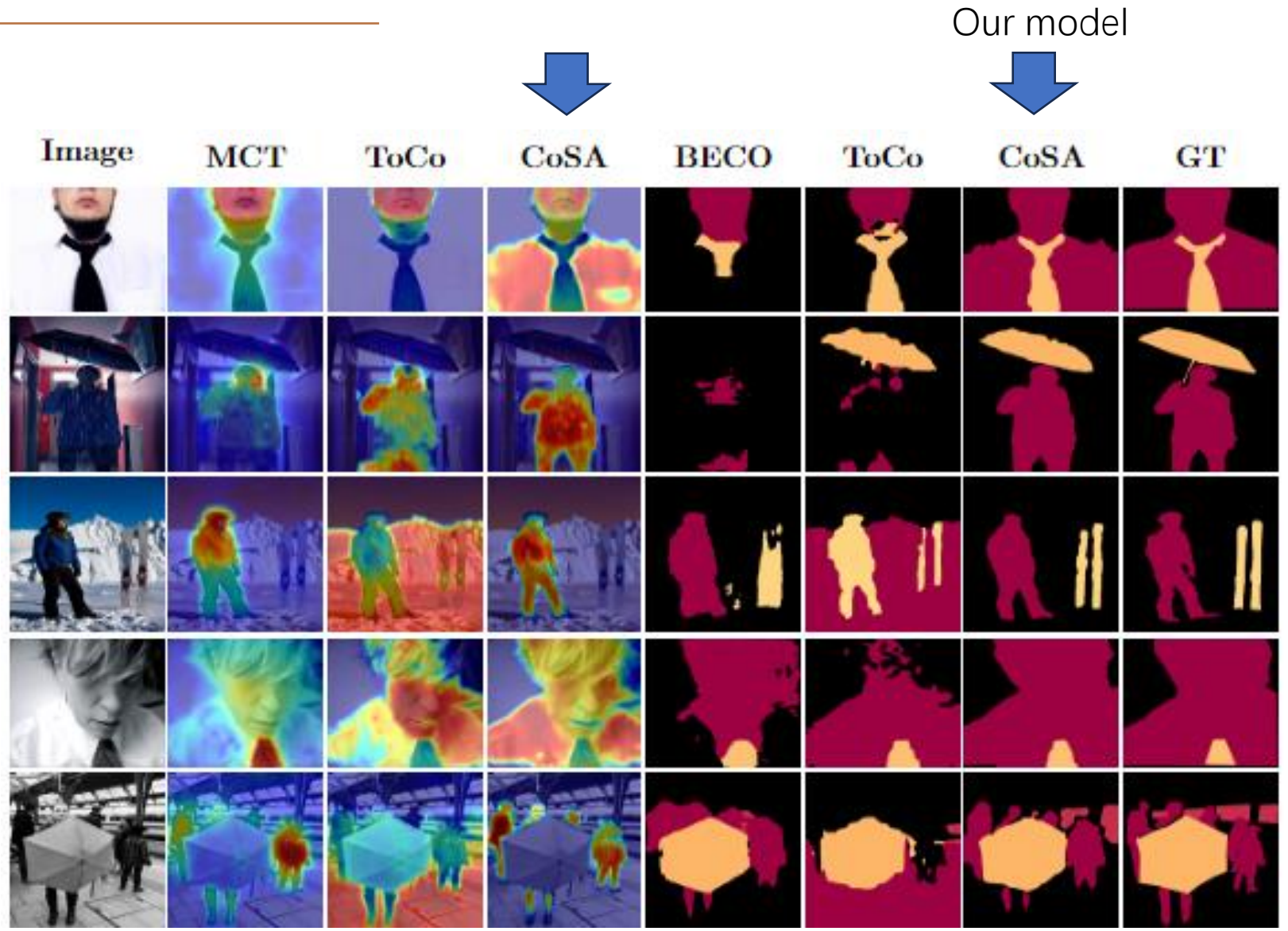
Aux. CAM CAM Input Image Segmentation



Aux. CAM CAM Input Image Segmentation



Results on COCO



H-Unique Datasets

H-Unique
Lend a Hand To Fight Child Abuse

The task is to segment the interesting part in a hand:

Tattoo

Ring

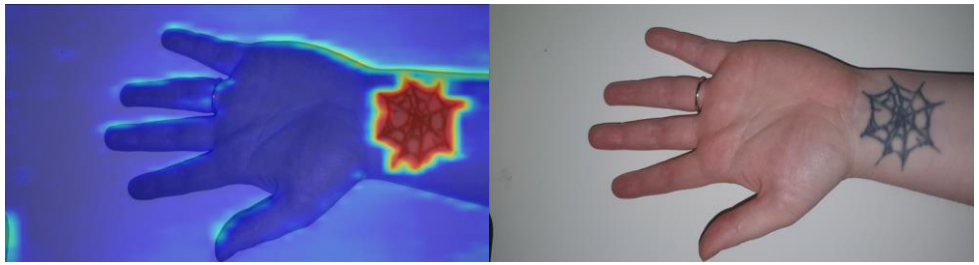
Lunule

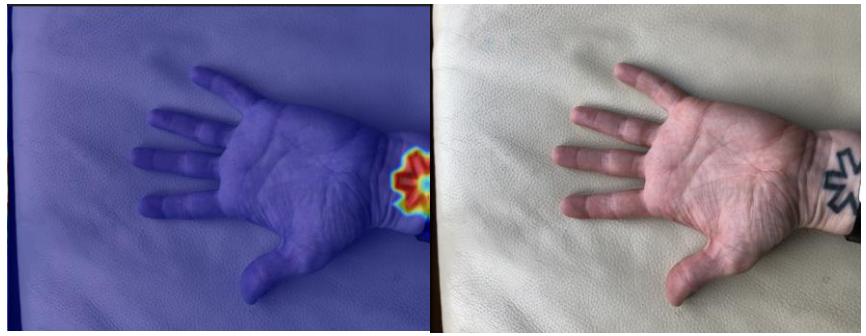
Polished nail

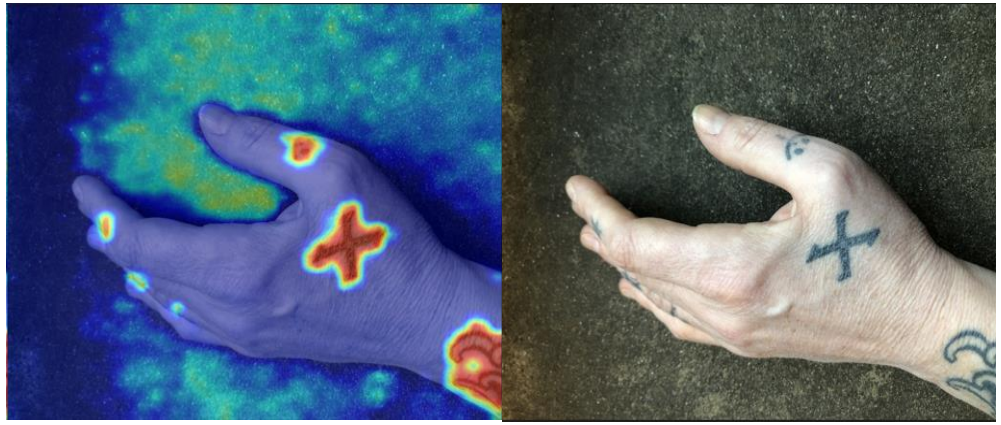
Credit to Ricki for collecting and annotating data

Results on Tattoo segmentation

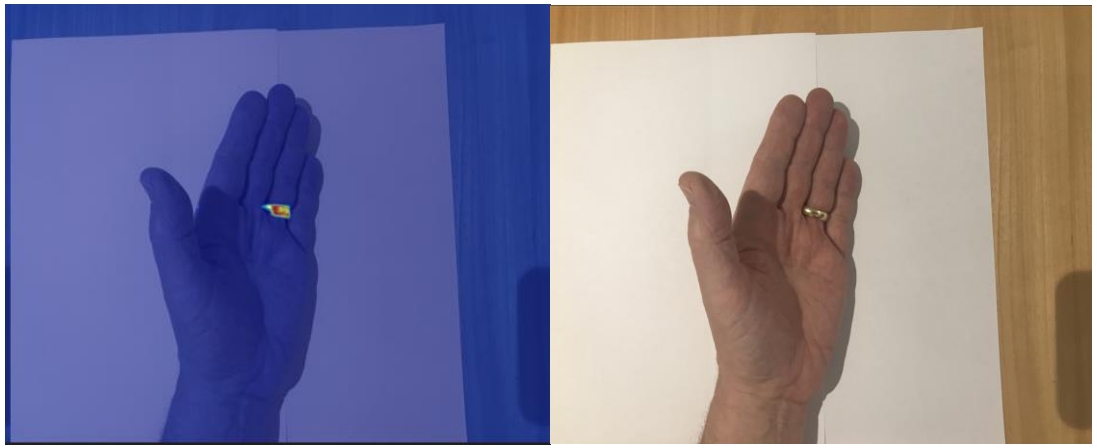
Results on Tattoo segmentation

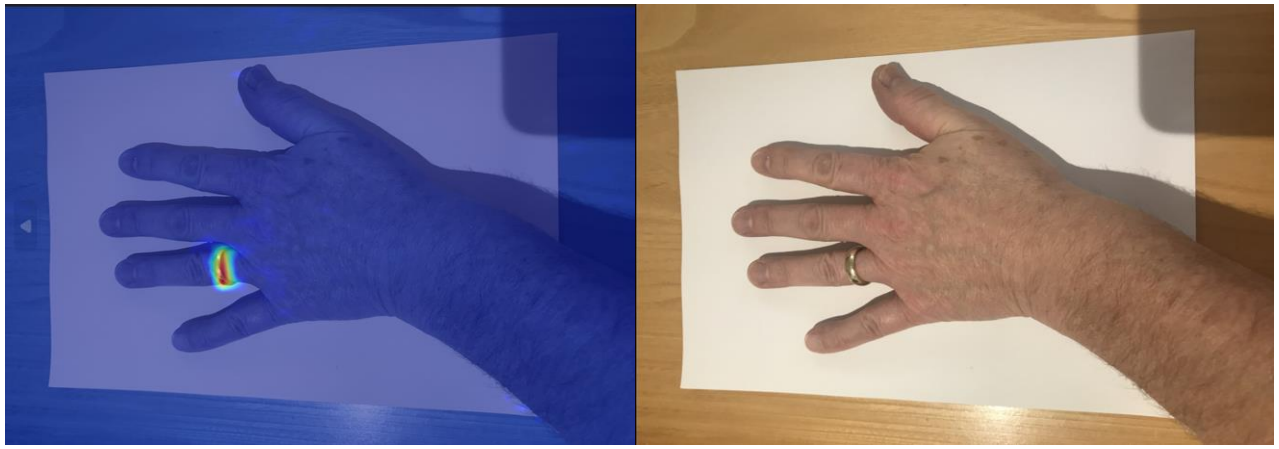


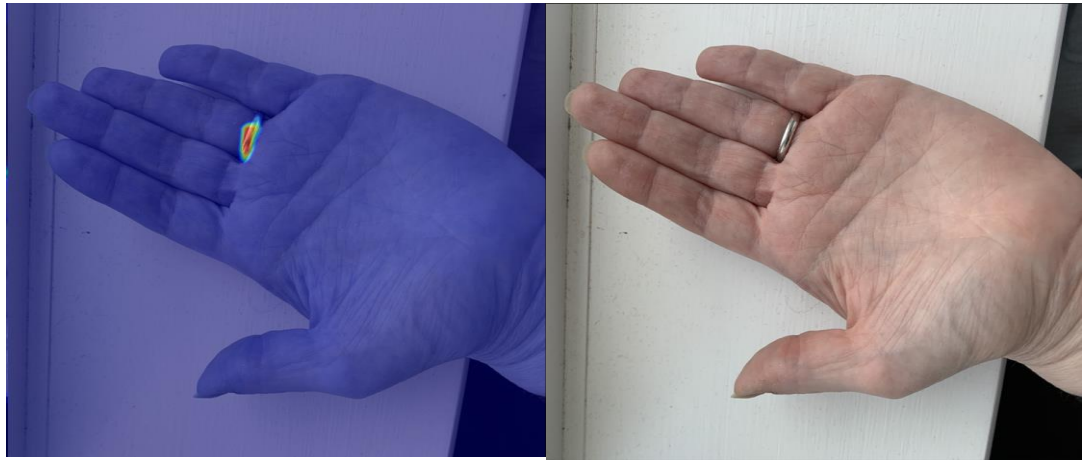




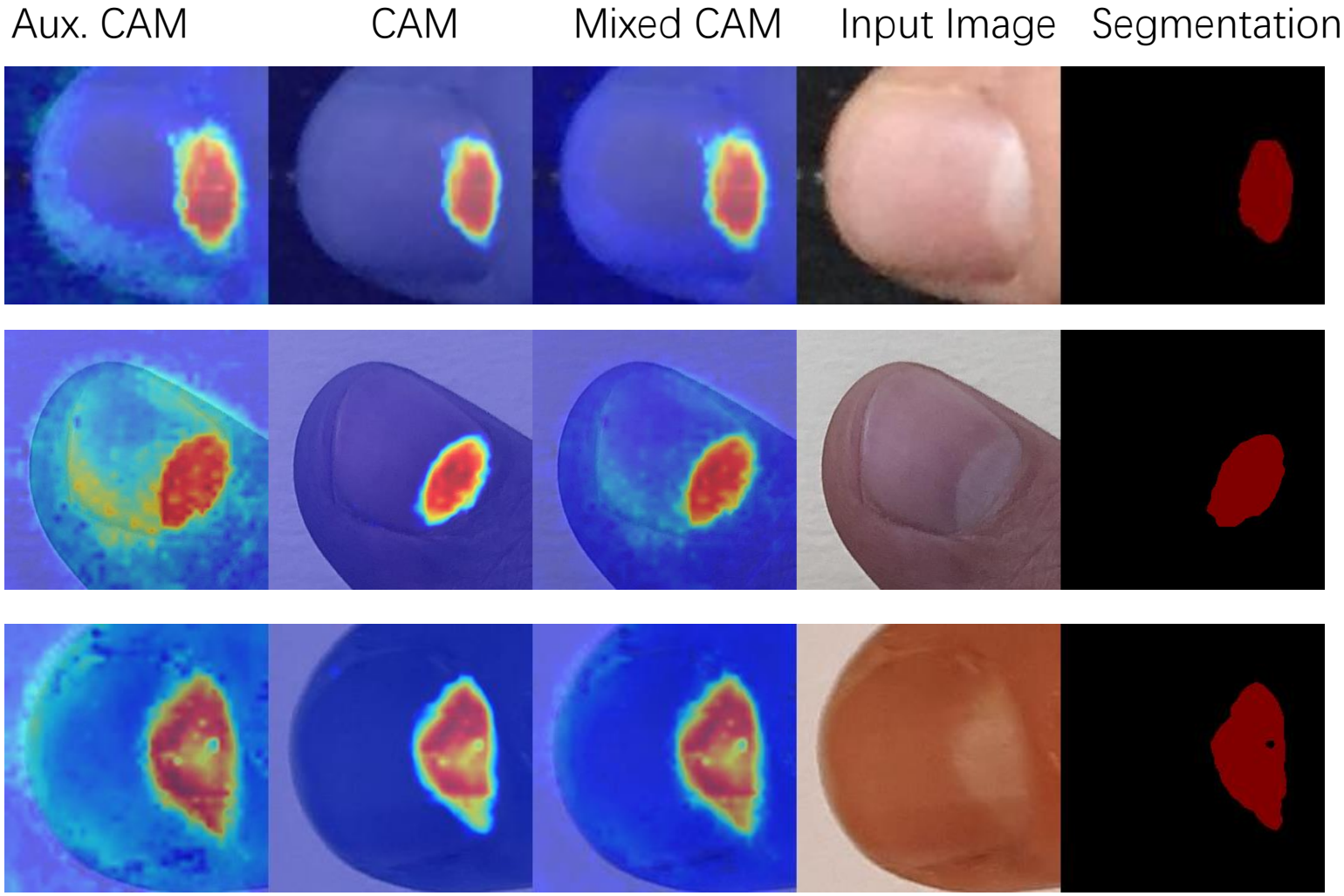
Results on Ring segmentation



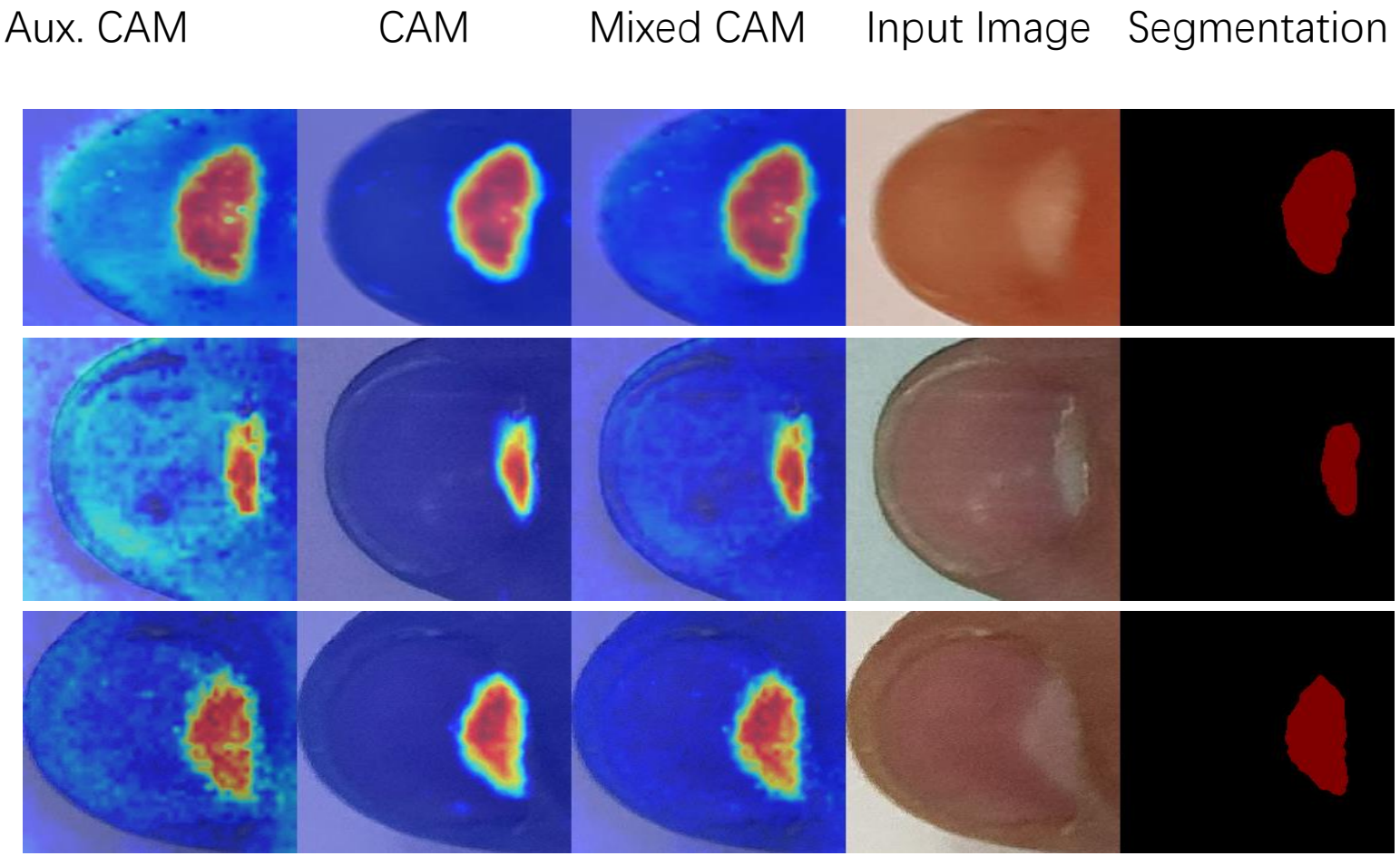




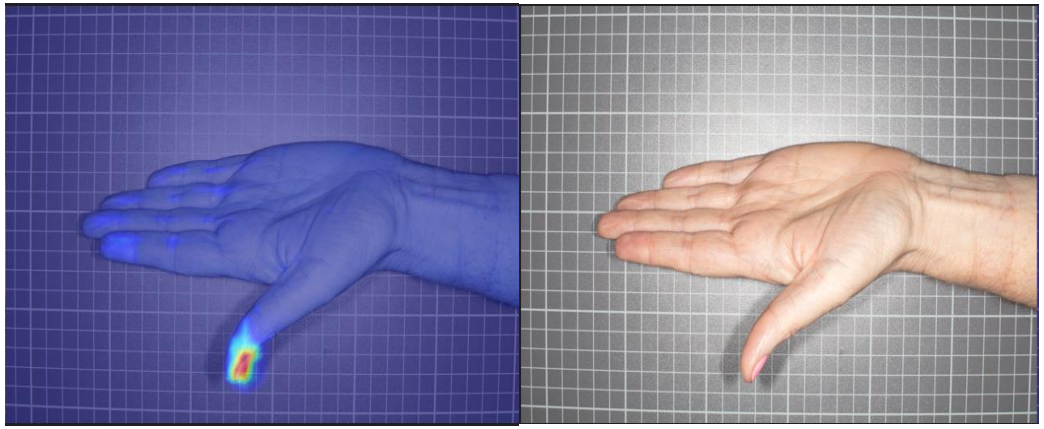
Results on Lunule segmentation

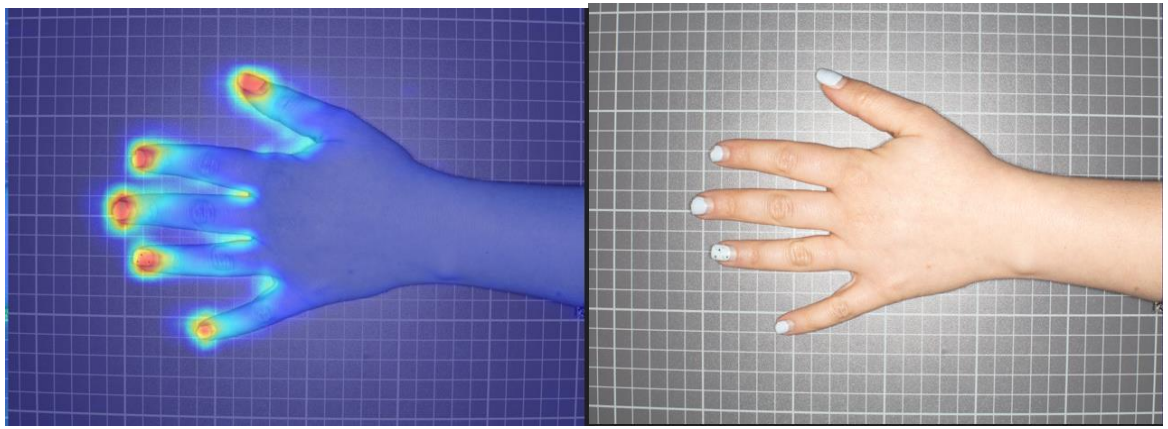


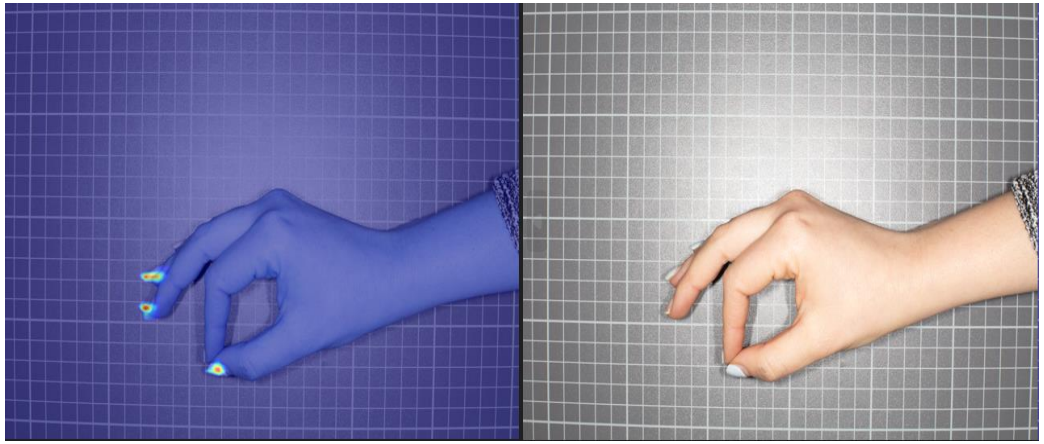
Results on Lunule segmentation

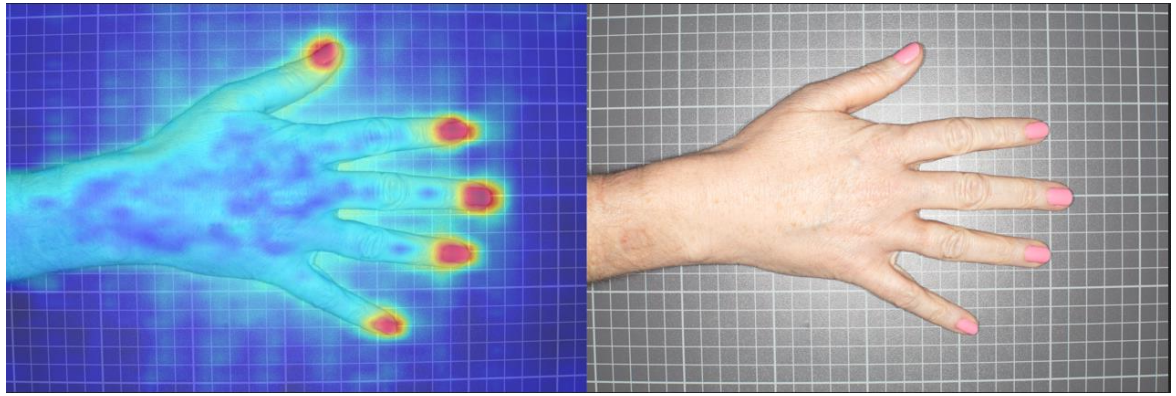


Results on Polished Nail segmentation









Takeaways

- Computer vision requires lots of data; data is expensive.
- Weakly-supervised learning is an active research area and required in practice.
- You use priors and hints to develop weakly-supervised learning algorithm.
- It is a good time to join the field and make contributions in this field!

Thanks