Evaluating Weakly Supervised Object Localization Methods Right

Junsuk Choe et al. (CVPR 2020)

Cheng Jiang, EECS 598-012 Unsupervised Visual Learning
March 31, 2021

Agenda

- What is weakly supervised object localization (WSOL)?
- When is WSOL ill-posed?
- How to evaluate WSOL?
- Authors' experiments
- Discussion

Object Localization

Input: $H \times W$ image $\mathbf{X}$

Goal: identify binary mask
$\mathbf{T} = (T_{11}, \ldots, T_{HW})$

Weakly Supervised Object Localization (WSOL)

$$\mathbf{X} \in \mathbb{R}^{H \times W} \quad \mathbf{T} = (T_{11}, \ldots, T_{HW})$$

Supervision	Data available for Training
Fully Supervised	$(\mathbf{X}, \mathbf{T})$ Pairs
Weakly Supervised	$\mathbf{X}$, and image-level label $Y \in \{0,1\}$

Multiple Instance Learning (MIL)

Training instances are arranged in sets, or **bags:**

$$X = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$$

Label of the bag:

$$Y = \begin{cases} +1 & \text{if } \exists y_i = +1 \\ -1 & \text{if } \forall y_i = -1 \end{cases}$$

Positive bags contain at least one positive instance

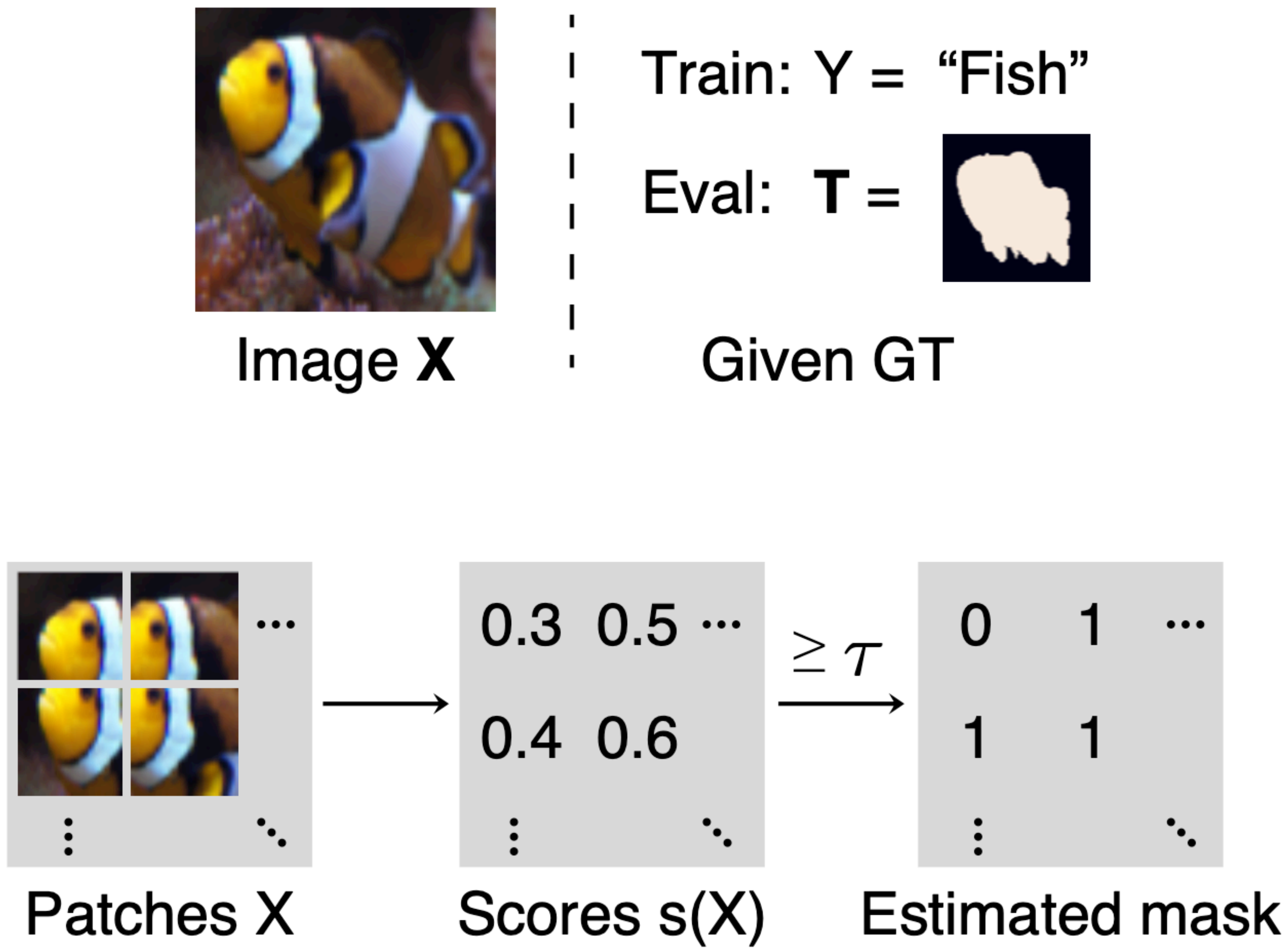
Negative bags contain all negative instances

WSOL as Multiple Instance Learning

Treat input image as a bag of $h \times w$ sliding window patches $(X_{11}, \ldots, X_{HW})$, collectively with a single label $Y \in \{0,1\}$

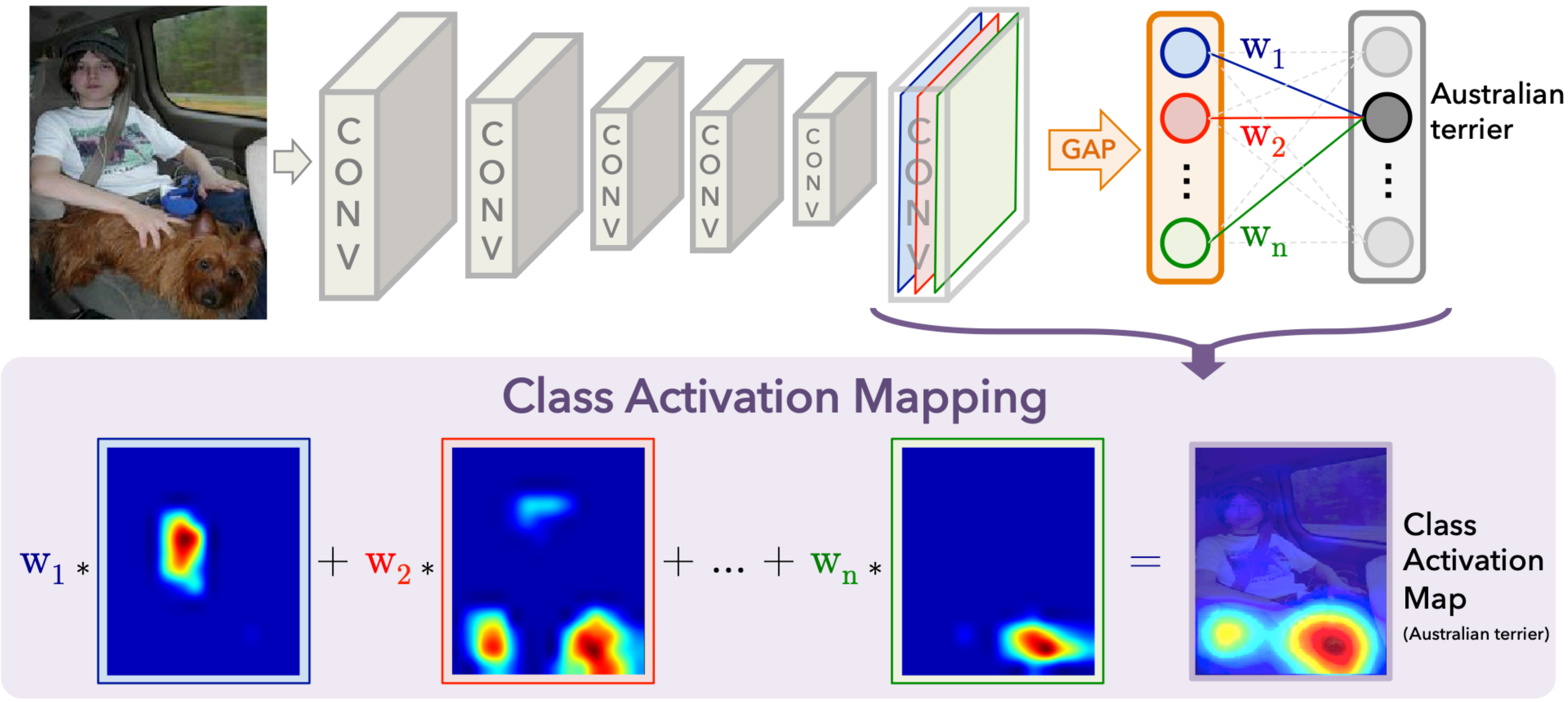
WSOL Task: predict object presence T_{ij} at each patch X_{ij} using a scoring function s and a threshold τ

$$T = \begin{cases} 1 & \text{if } s(X) \geq \tau \\ 0 & \text{if } s(X) < \tau \end{cases}$$



Class Activation Mapping (CAM)
Zhou et al. (CVPR 2016)

Idea: project output layer weights back on to the convolutional feature maps



Improvements since CAM

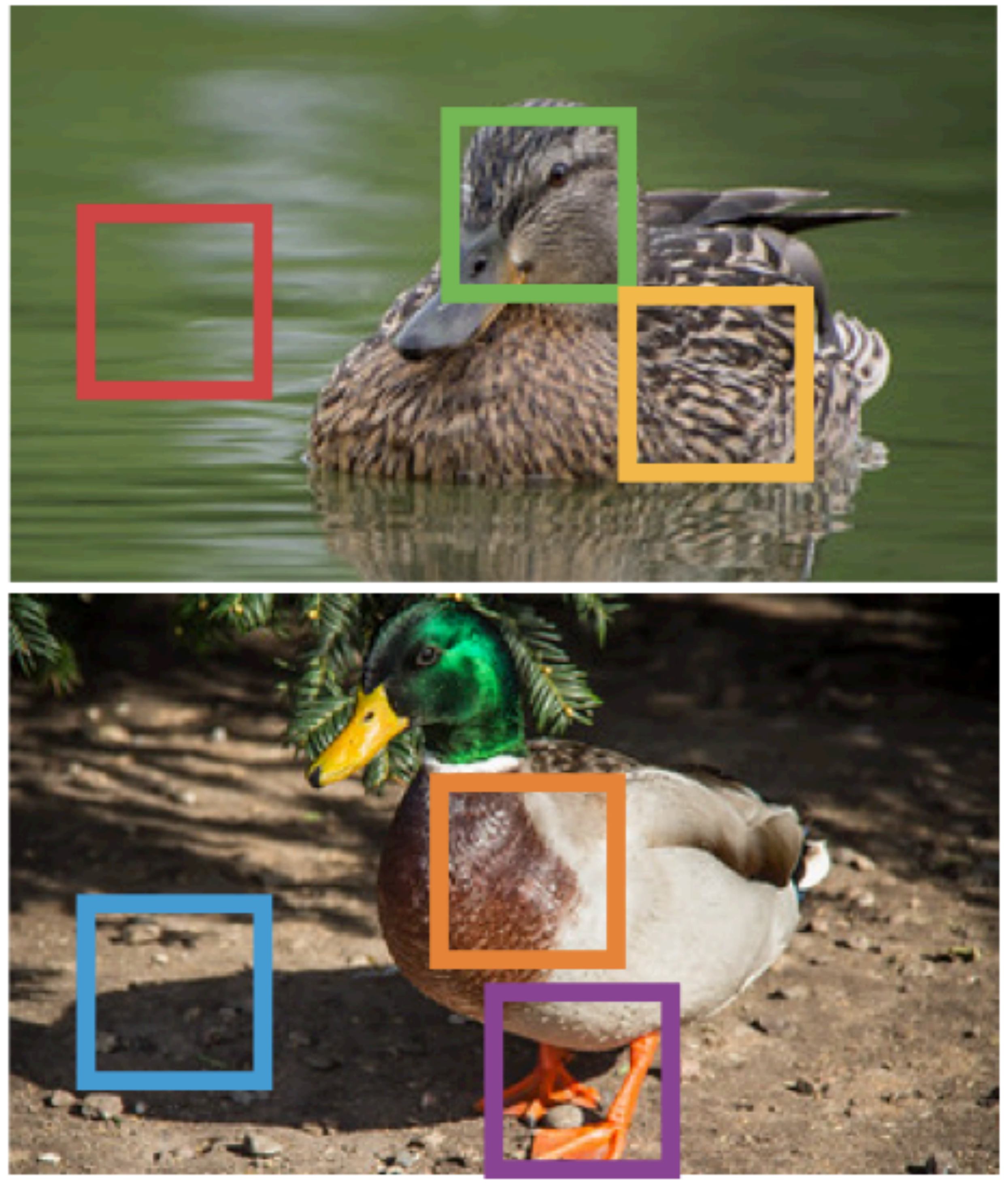
- Architectural improvements
- Data augmentation
- Parameter search

Existing approaches:
Choose scoring rule $s(X) = p(Y | X)$

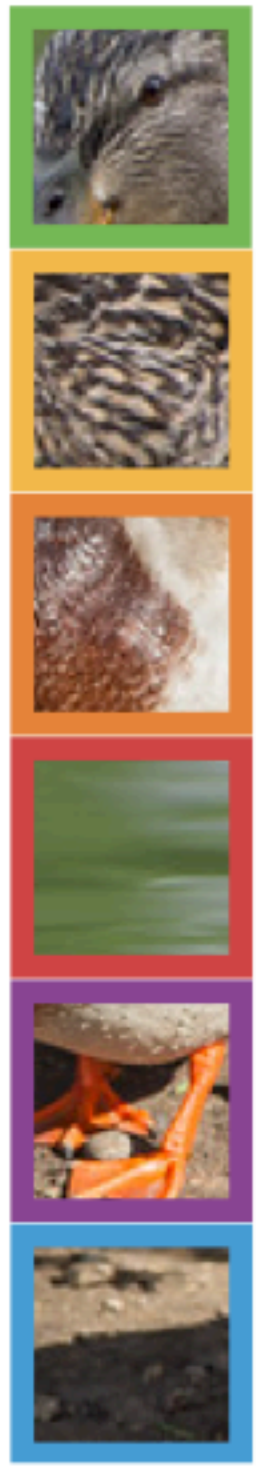
Problem: What if the background cues are more strongly associated with the target label than some foreground cues?

What if the background cues are more strongly associated with the target label than some foreground cues?

Image



X



M	$p(Y \mid M)$	T	Evaluation
duck's head	0.8	1	TP
duck's body	0.7	1	TP
duck's body	0.7	1	TP
water	0.4	0	FP
duck's feet	0.3	1	FN
dirt	0.1	0	TN

threshold $\tau = 0.35$

Claim: If background cues are more strongly associated with the target labels than some foreground cues, the localization task CANNOT be solved

Assumptions

- There exists a finite set of cue labels $\mathscr{M}$ containing all patch-level concepts in natural images
- Every patch X is equivalently represented by its cue label $M(X) \in \mathscr{M}$
- *We have access to the joint distribution* $p(Y, M)$

Formally Speaking...

Lemma 3.1 Assume that the true posterior $p(Y|M)$ with a continuous pdf is used as the scoring rule $s(M) = p(Y|M)$. Then, there exists a scalar $\tau \in \mathbb{R}$ such that $s(M) \geq \tau$ is identical to T if and only if the foreground-background posterior ratio $\dfrac{p(Y=1|M^{fg})}{p(Y=1|M^{bf})} \geq 1$ almost surely, conditionally on the event $\{T(M^{fg}) = 1$ and $T(M^{bf}) = 0\}$

Lemma 3.1 Assume that the true posterior $p(Y|M)$ with a continuous pdf is used as the scoring rule $s(M) = p(Y|M)$. Then, there exists a scalar $\tau \in \mathbb{R}$ such that $s(M) \geq \tau$ is identical to T if and only if the foreground-background posterior ratio $\dfrac{p(Y=1|M^{fg})}{p(Y=1|M^{bf})} \geq 1$ almost surely, conditionally on the event $\{T(M^{fg}) = 1 \text{ and } T(M^{bf}) = 0\}$

Suppose we choose $s(X) = p(Y|X)$ ...

Lemma 3.1 Assume that the true posterior $p(Y \mid M)$ with a continuous pdf is used as the scoring rule $s(M) = p(Y \mid M)$. Then, there exists a scalar $\tau \in \mathbb{R}$ such that $s(M) \geq \tau$ is identical to T if and only if the foreground-background posterior ratio $\frac{p(Y=1 \mid M^{fg})}{p(Y=1 \mid M^{bf})} \geq 1$ almost surely, conditionally on the event $\{T(M^{fg}) = 1$ and $T(M^{bf}) = 0\}$

... suppose $T(\text{foreground}) = 1, T(\text{background}) = 0$ ...

Lemma 3.1 Assume that the true posterior $p(Y|M)$ with a continuous pdf is used as the scoring rule $s(M) = p(Y|M)$. Then, there exists a scalar $\tau \in \mathbb{R}$ such that $s(M) \geq \tau$ is identical to T if and only if the foreground-background posterior ratio $\dfrac{p(Y=1|M^{fg})}{p(Y=1|M^{bf})} \geq 1$ almost surely, conditionally on the event $\{T(M^{fg}) = 1$ and $T(M^{bf}) = 0\}$

… we can achieve object localization …

Lemma 3.1 Assume that the true posterior $p(Y|M)$ with a continuous pdf is used as the scoring rule $s(M) = p(Y|M)$. Then, there exists a scalar $\tau \in \mathbb{R}$ such that $s(M) \geq \tau$ is identical to T if and only if the foreground-background posterior ratio $\dfrac{p(Y=1\,|\,M^{fg})}{p(Y=1\,|\,M^{bf})} \geq 1$ almost surely, conditionally on the event $\{T(M^{fg}) = 1$ and $T(M^{bf}) = 0\}$

… iff $p(Y=1\,|\,M^{fg}) \geq p(Y=1\,|\,M^{bf})$

Lemma 3.1 Assume that the true posterior $p(Y|M)$ with a continuous pdf is used as the scoring rule $s(M) = p(Y|M)$. Then, there exists a scalar $\tau \in \mathbb{R}$ such that $s(M) \geq \tau$ is identical to T if and only if the foreground-background posterior ratio $\dfrac{p(Y=1 \mid M^{fg})}{p(Y=1 \mid M^{bf})} \geq 1$ almost surely, conditionally on the event $\{T(M^{fg}) = 1$ and $T(M^{bf}) = 0\}$

… iff foreground posterior $\geq$ background posterior

In (mostly) English:

Lemma 3.1

Suppose we choose $s(X) = p(Y \mid X)$, we can achieve object localization iff foreground posterior $\geq$ background posterior.

Takeaway: WSOL is ill-posed when background cues are more strongly associated with the target labels than some foreground cues

Where to go from here?

Data augmentations focused on:

- Positive samples with less represented foreground features
- Negative samples with more target correlated background features

Pixel Precision and Recall

WSOL evaluations when masks are available

Pixel precision

$$PxPrec(\tau) = \frac{\left| \left\{ s_{ij}^{(n)} \geq \tau \right\} \cap \left\{ T_{ij}^{(n)} = 1 \right\} \right|}{\left| \left\{ s_{ij}^{(n)} \geq \tau \right\} \right|}$$

Prediction

Ground Truth

Pixel recall

$$PxRec(\tau) = \frac{\left| \left\{ s_{ij}^{(n)} \geq \tau \right\} \cap \left\{ T_{ij}^{(n)} = 1 \right\} \right|}{\left| \left\{ T_{ij}^{(n)} = 1 \right\} \right|}$$

Area under the pixel precision-recall curve

Pixel average precision

$$PxAP = \sum_l PxPrec(\tau_l) \cdot \left[PxRec(\tau_l) - PxRec(\tau_{l-1}) \right]$$

Box Accuracy

WSOL evaluations when only bounding boxes are available

Score map threshold

IoU threshold

IoU threshold

Box Accuracy

$$BoxAcc(\tau, \delta) = \frac{1}{N}\sum_n \mathbf{1}_{IoU\left(box\left(s\left(\mathbf{X}^{(n)}\right),\tau\right),B^{(n)}\right)\geq\delta}$$

Tightest box around the largest connected component of the predicted mask

Ground truth box

Max Box Accuracy

$$MaxBoxAcc(\delta) = \max_{\tau} BoxAcc(\tau, \delta)$$

Experiments

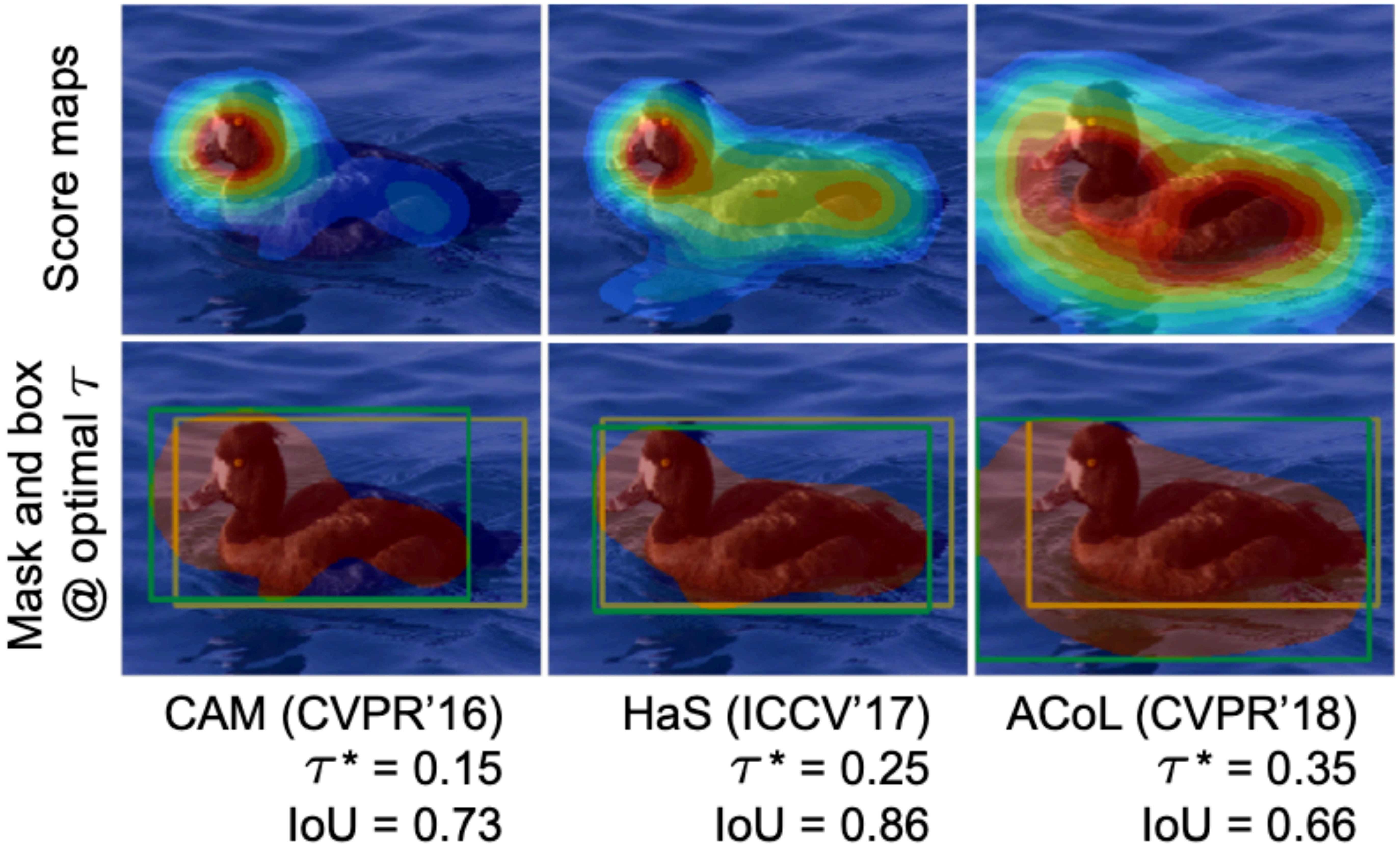
Dataset

		ImageNet	CUB	OpenImages
# Classes		1000	200	100
# Images per Class	Training, Weakly supervised	1.2k	30	300
	Training, Fully supervised	10	5	25
	Test	10	29	50

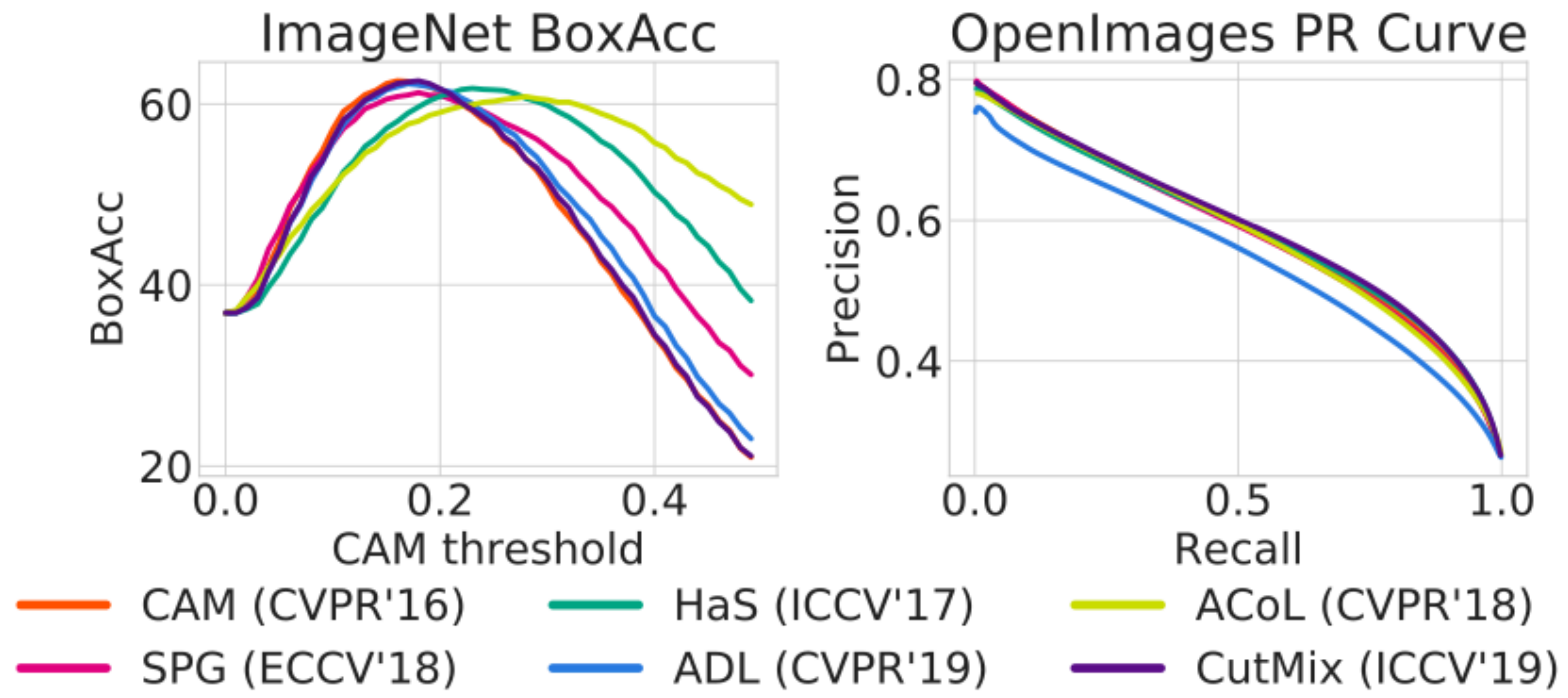
Results

Methods	ImageNet (MaxBoxAcc) VGG	Inception	ResNet	Mean	CUB (MaxBoxAcc) VGG	Inception	ResNet	Mean	OpenImages (PxAP) VGG	Inception	ResNet	Mean	Total Mean
CAM [60]	61.1	65.3	64.2	63.5	71.1	62.1	73.2	68.8	58.1	61.4	58.0	59.1	63.8
HaS [26]	+0.7	+0.1	-1.0	-0.1	+5.2	-4.4	+4.9	+1.9	-1.2	-2.9	+0.2	-1.3	+0.2
ACoL [58]	-0.8	-0.7	-2.5	-1.4	+1.2	-2.5	-0.5	-0.6	-3.4	+1.6	-0.2	-0.7	-0.9
SPG [59]	+0.5	+0.1	-0.7	+0.0	-7.4	+0.7	-1.8	-2.8	-2.2	+1.0	-0.3	-0.5	-1.1
ADL [6]	-0.3	-3.8	+0.0	-1.4	+4.6	+1.3	+0.3	+2.0	+0.2	+0.7	-3.7	-0.9	-0.1
CutMix [56]	+1.0	+0.1	-0.3	+0.3	+0.8	+3.4	-5.4	-0.4	+0.1	+0.3	+0.7	+0.4	+0.1
Best WSOL	62.2	65.5	64.2	63.8	76.2	65.5	78.1	70.8	58.3	63.0	58.6	59.5	64.0
FSL baseline	62.8	68.7	67.5	66.3	86.3	94.0	95.8	92.0	61.5	70.3	74.4	68.7	75.7
Center baseline	52.5	52.5	52.5	52.5	59.7	59.7	59.7	59.7	45.8	45.8	45.8	45.8	52.3

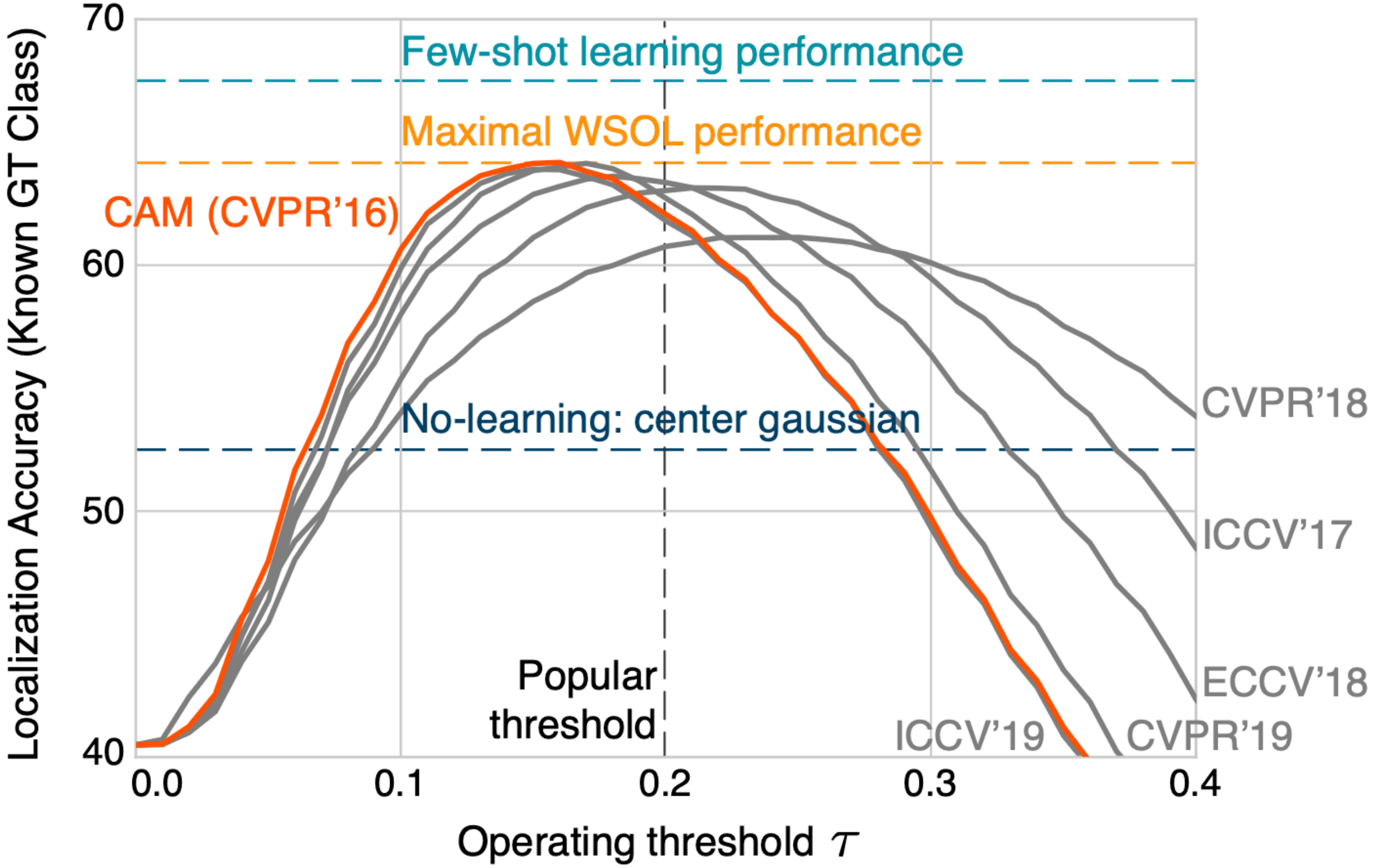
Visualization



Performance at Varying Operating Thresholds



Performance at Varying Operating Thresholds



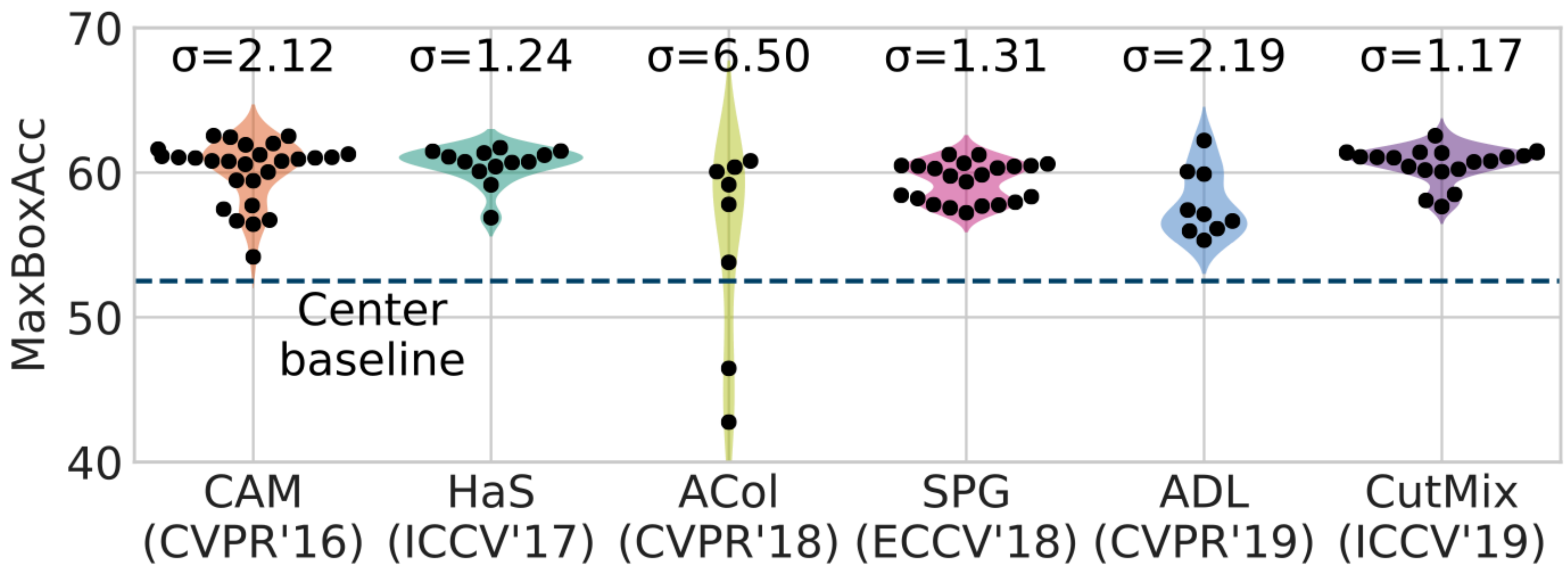
Hyperparameter Trials

Random search on 30 hyperparameter sets

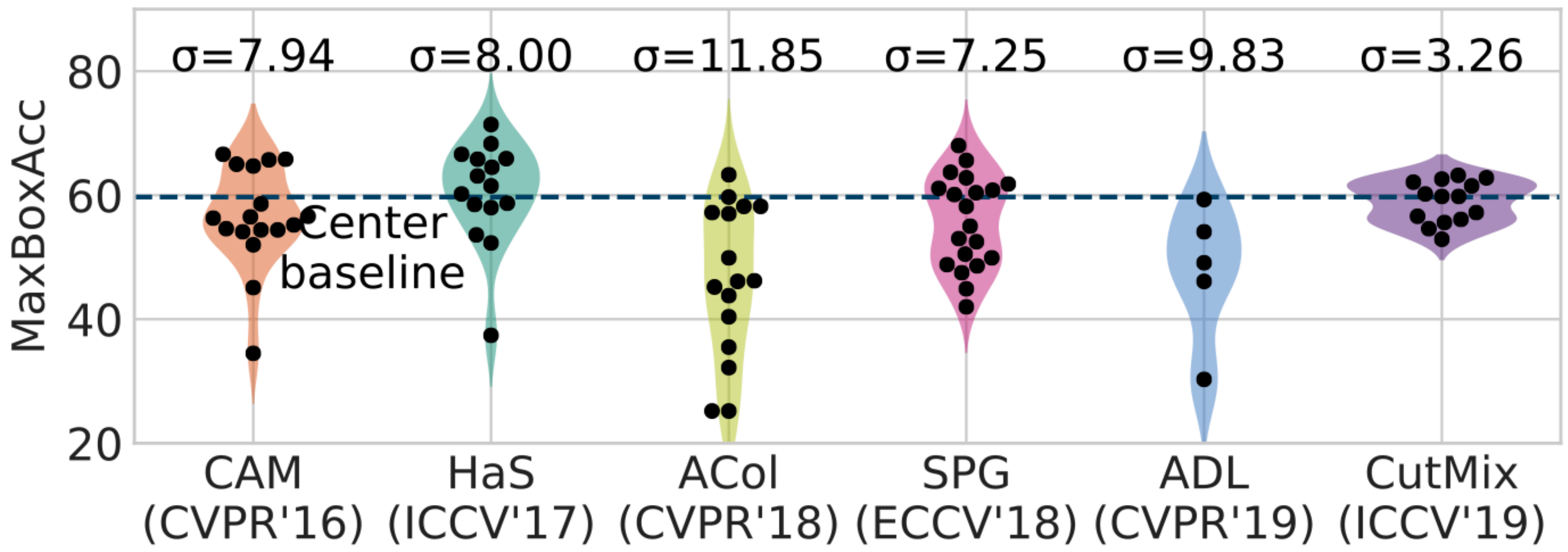
Trained on weakly supervised training set

Validated on fully supervised training set

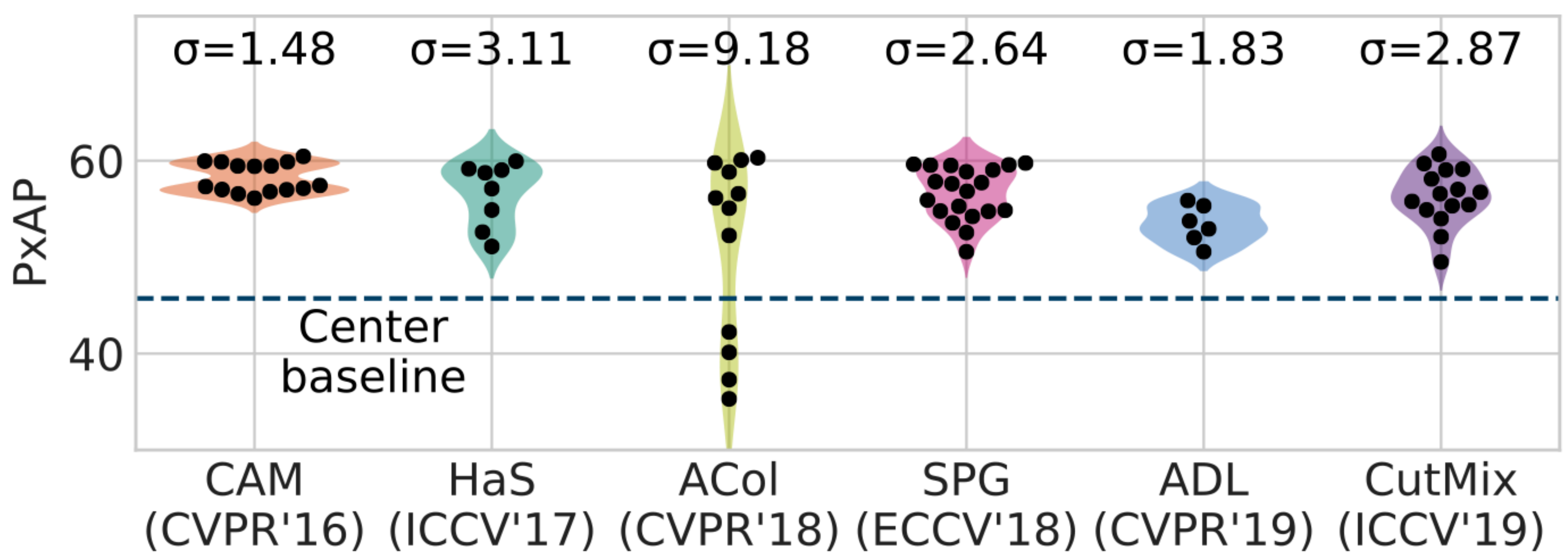
Hyperparameter Trial Results



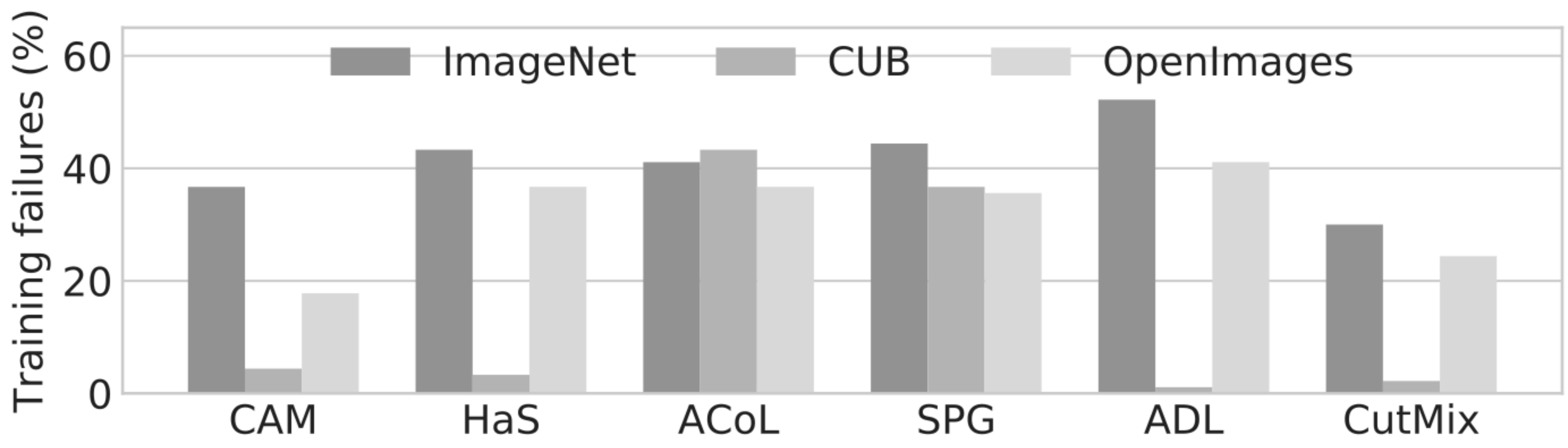
(a) ResNet50 architecture, ImageNet dataset.



(b) ResNet50 architecture, CUB dataset.



(c) ResNet50 architecture, OpenImages dataset.

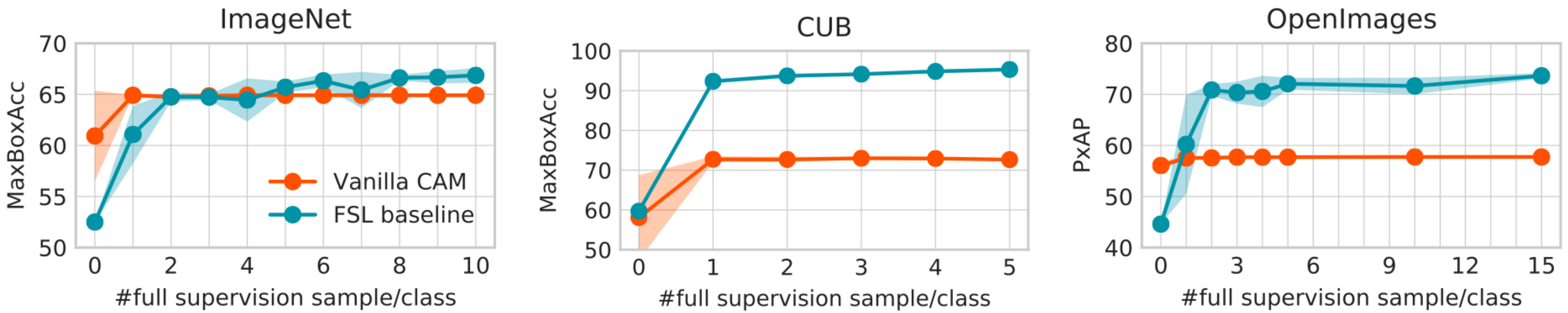


(d) Ratio of training failures in different WSOL methods.

Compare to Few-shot Learning Baseline

Using a modified foreground saliency mask predictor (a fully convolutional network)

Trained using the fully supervised training set



Discussion

Pros:

- Well written
- Good analysis of the WSOL task
- Proposed evaluation methods that make a lot of sense
- Open source code + data annotation

Cons:

- Feasibility of the proposed future direction
- Number of object localization annotation is relatively small

Next Up

Contrastive Learning for Weakly Supervised Phrase Grounding

References

- Choe J, Oh SJ, Lee S, Chun S, Akata Z, Shim H. Evaluating weakly supervised object localization methods right. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2020 (pp. 3133-3142).
- Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. InProceedings of the IEEE conference on computer vision and pattern recognition 2016 (pp. 2921-2929).
- Carbonneau MA, Cheplygina V, Granger E, Gagnon G. Multiple instance learning: A survey of problem characteristics and applications. Pattern Recognition. 2018 May 1;77:329-53.