Exploring Foveation and Saccade for Improved Weakly-Supervised Localization

Timur Ibrayev, Manish Nagaraj, Amitangshu Mukherjee, Kaushik Roy

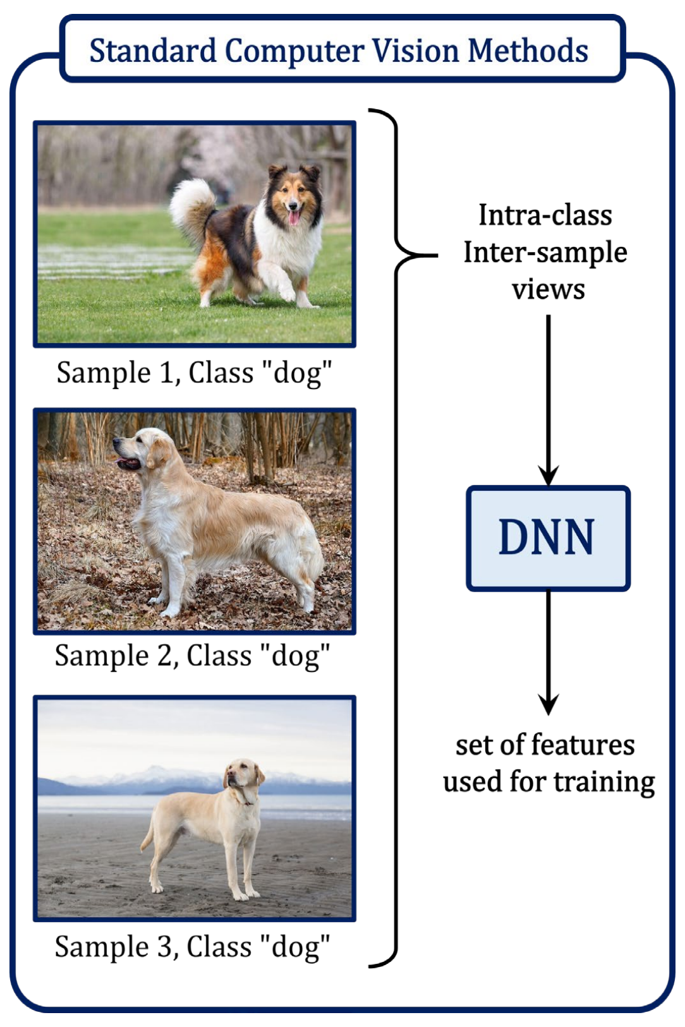
Purdue University

NEURAL INFORMATION PROCESSING SYSTEMS

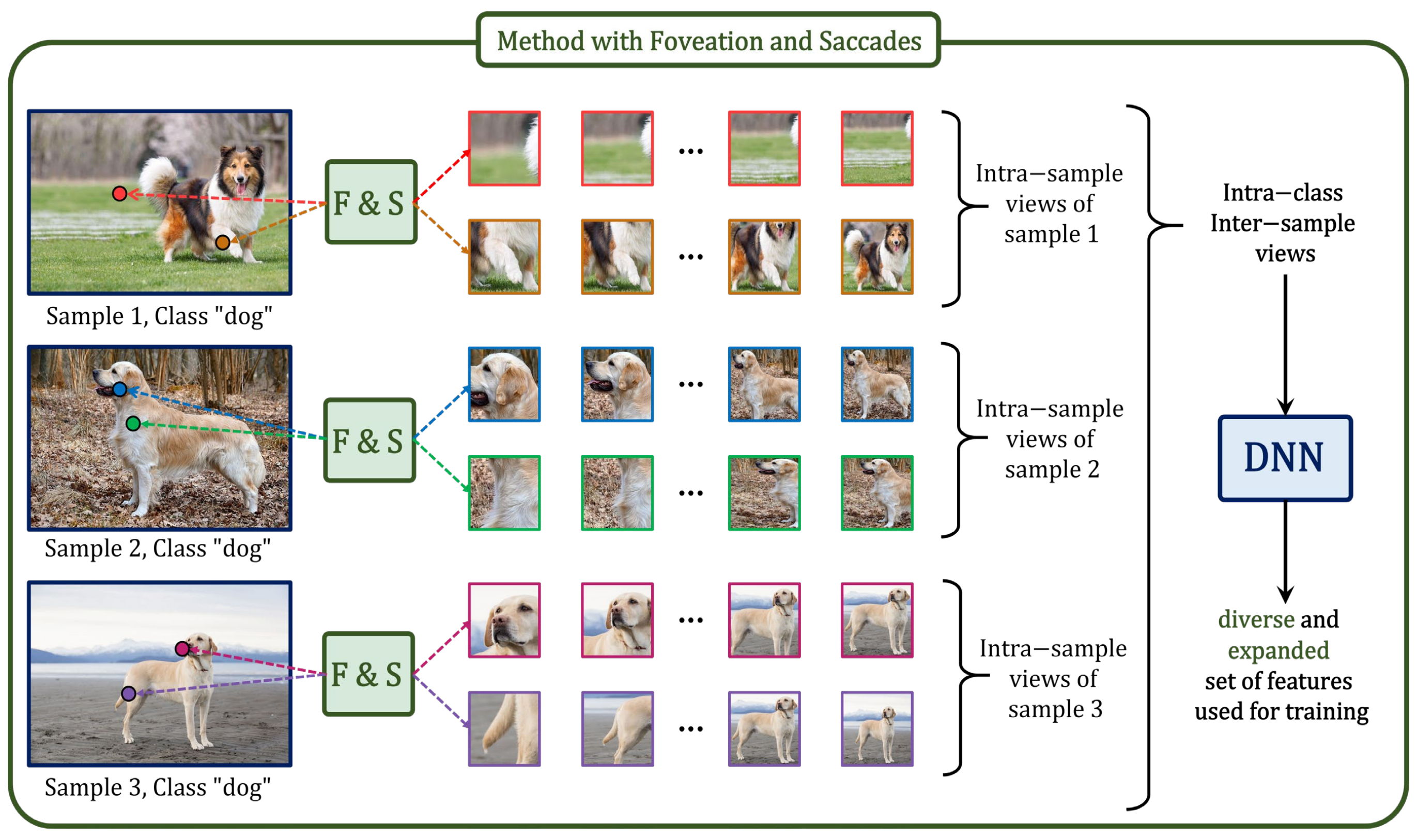
Gaze Meets ML
Workshop
@ NeurIPS 2023

1. MOTIVATION

- Weakly supervised object localization (WSOL) requires prediction of both the object class (*image-level label*) and the object location (*instance-level label*), while being trained only on the image-level labels.
- Such methods can benefit from a large number of views obtained from within each sample.

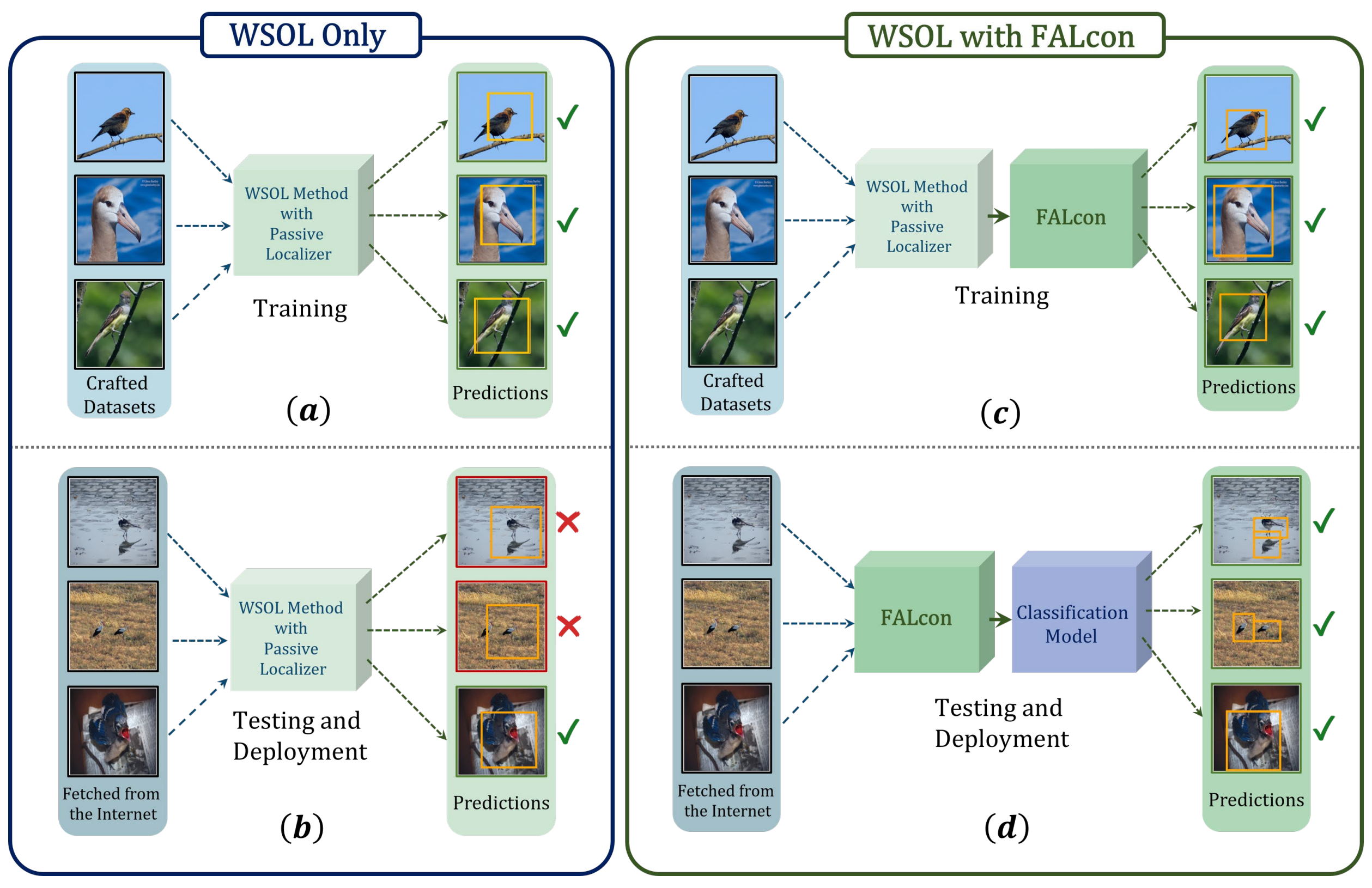

Standard Computer Vision Methods
Sample 1, Class "dog"
Sample 2, Class "dog"
Sample 3, Class "dog"
Intra-class Inter-sample views
DNN
set of features used for training

Standard methods process inputs with uniform resolution.


Method with Foveation and Saccades
Sample 1, Class "dog"
Sample 2, Class "dog"
Sample 3, Class "dog"
F & S
F & S
F & S
Intra−sample views of sample 1
Intra−sample views of sample 2
Intra−sample views of sample 3
Intra−class Inter−sample views
DNN
diverse and expanded set of features used for training

Methods with Foveation and Saccades (F&S) obtain different views from within the sample at different fixation points with different degrees of detail.

2. APPLICATION



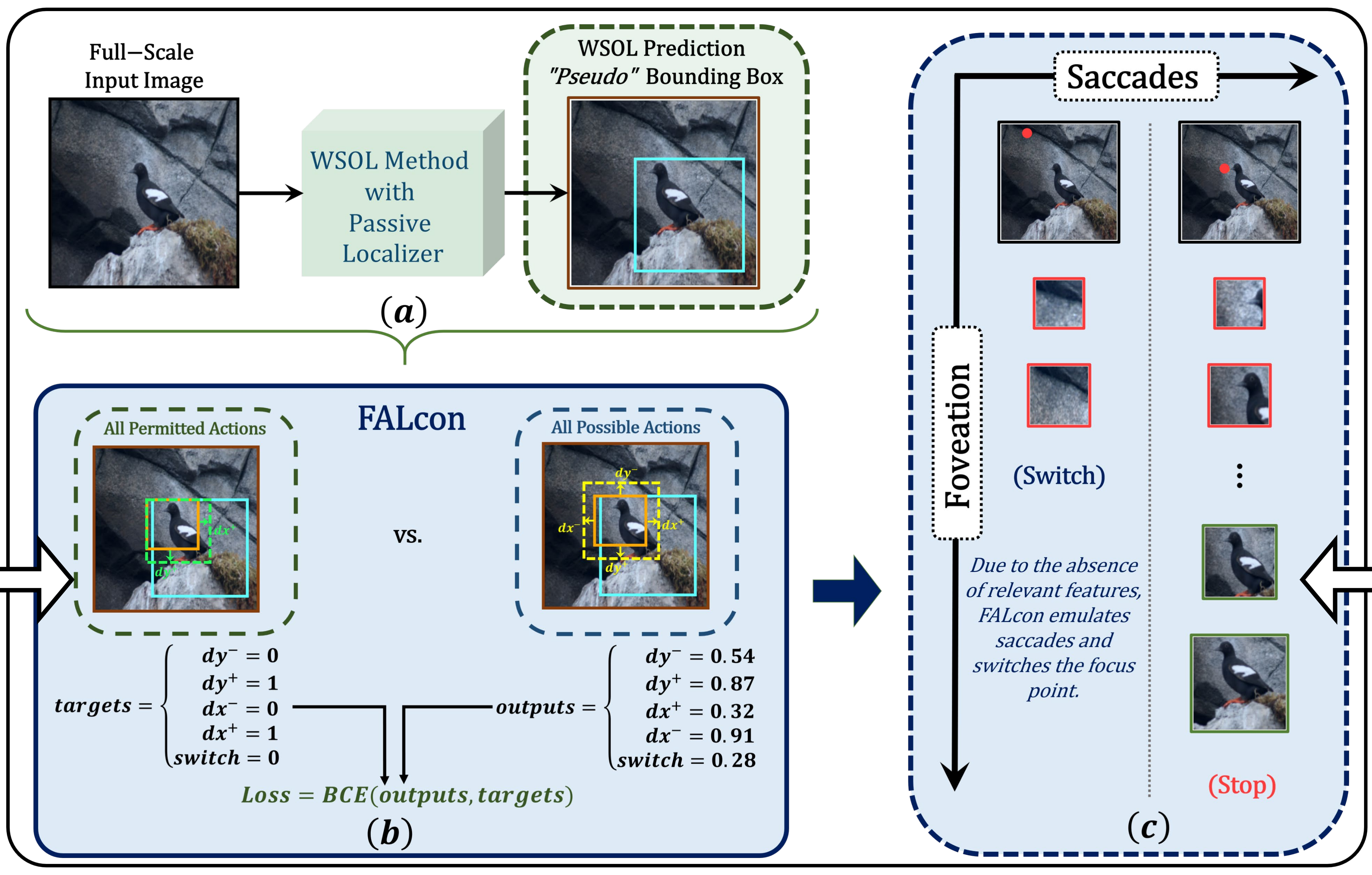
Benefits

- Improved localization performance.
- Resilient WSOL pipeline: capable of detecting multiple objects while being trained only on images with a single object.
- Neuro-inspired design, capable of producing bounding boxes similar to humans.

3. METHODOLOGY

Training

- ***Foveation*** is trained by making the model predict four possible expansion actions (yellow dashed box) with the set of permitted actions (green dashed box) serving as training targets
- ***Saccades*** is trained by making the model predict whether the current *foveated region* is within the "pseudo" bounding box or not



$$targets = \begin{cases} dy^- = 0 \\ dy^+ = 1 \\ dx^- = 0 \\ dx^+ = 1 \\ switch = 0 \end{cases} \qquad outputs = \begin{cases} dy^- = 0.54 \\ dy^+ = 0.87 \\ dx^+ = 0.32 \\ dx^- = 0.91 \\ switch = 0.28 \end{cases}$$

$$Loss = BCE(outputs, targets)$$

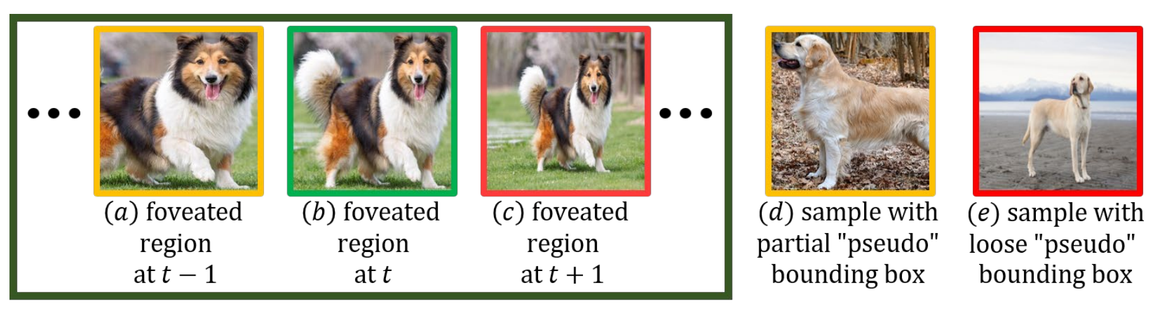
Inference

- Starting from different fixation points, modeled ***foveation*** sequentially extracts *foveated views* from within the sample
- Concurrently, modeled ***saccades*** estimates the relevance of each *foveated observation* and decides whether further exploration is necessary

4. RESULTS

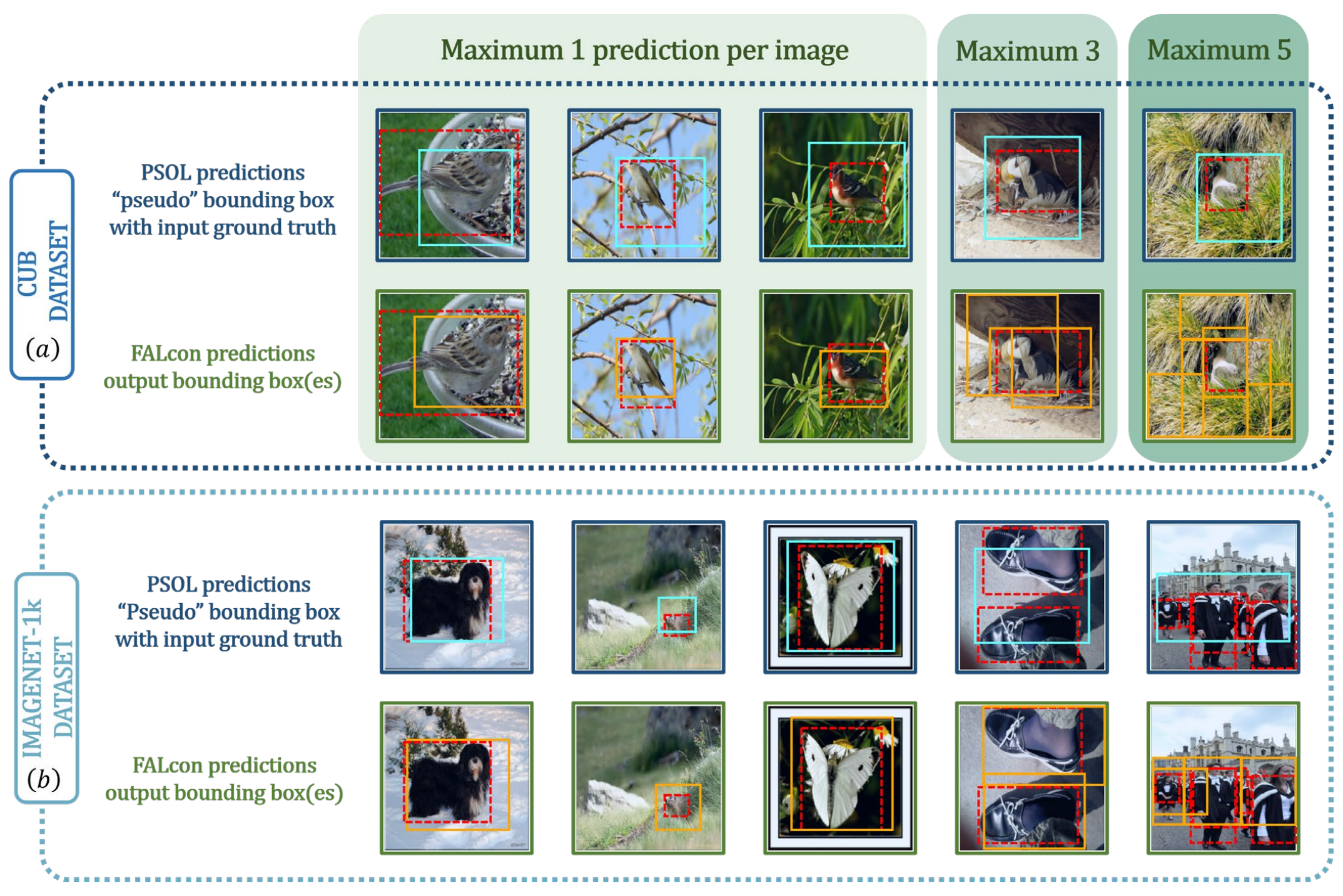
FALcon Improving Localization

- Bounding box predictions are improved by *transferring the knowledge from foveated regions*:
- Correct required expansions from (a) to (b) is transferred to ***complete*** the regions like (d).
- Wrong excessive expansions from (b) to (c) is transferred to ***tighten*** the regions like (e).


(a) foveated region at $t-1$
(b) foveated region at t
(c) foveated region at $t+1$
(d) sample with partial "pseudo" bounding box
(e) sample with loose "pseudo" bounding box

Weakly supervised object localization (WSOL) results

Method	# of predictions per image	CUB		ImageNet	
		GT Loc	Top-1 Loc	GT Loc	Top-1 Loc
VGG16 CAM (Zhou et al., 2016)	1	57.96	36.13	59.00	42.80
InceptionV3 SPG (Zhang et al., 2018c)	1	60.50	46.64	64.49	48.60
VGG16 SLT-Net (Guo et al., 2021)	1	87.60	67.80	67.20	51.20
DenseNet161 C^{2}AM (Xie et al., 2022)	1	94.46	83.28	68.20	59.28
PSOL (baseline) (Zhang et al., 2020a)	**1**	**77.41**	**63.56**	**66.28**	**55.31**
FALcon + PSOL (Ours)	max 1	88.30	62.82	62.45	49.39
	max 3	89.35	63.50	67.38	53.31
	max 5	**89.35**	**63.50**	**67.51**	**53.50**


Maximum 1 prediction per image
Maximum 3
Maximum 5
CUB DATASET
(a)
IMAGENET-1k DATASET
(b)
PSOL predictions "pseudo" bounding box with input ground truth
FALcon predictions output bounding box(es)

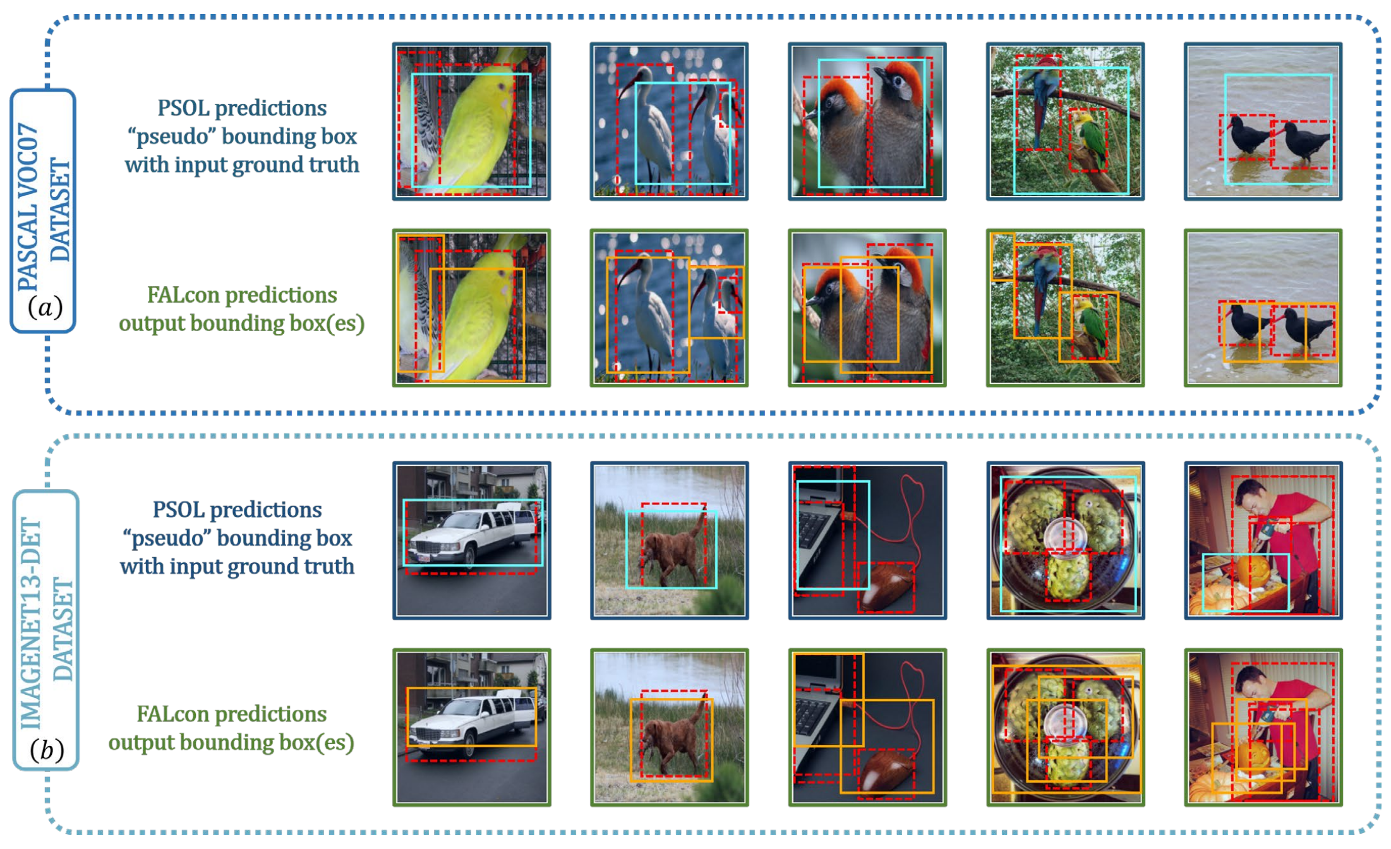
Qualitative results obtained on the (a) CUB and (b) ImageNet-1k datasets.

FALcon Enabling Resiliency

- The end goal of WSOL-based systems would be to operate in an *unconstrained environment*, like the internet, without human supervision.
- Hence, it is crucial for them to be able either to flag unexpected input or to detect multiple objects *(even if trained only on data with a single object)*.

Results of applying localization models trained on images containing a single object to datasets containing multiple objects per image

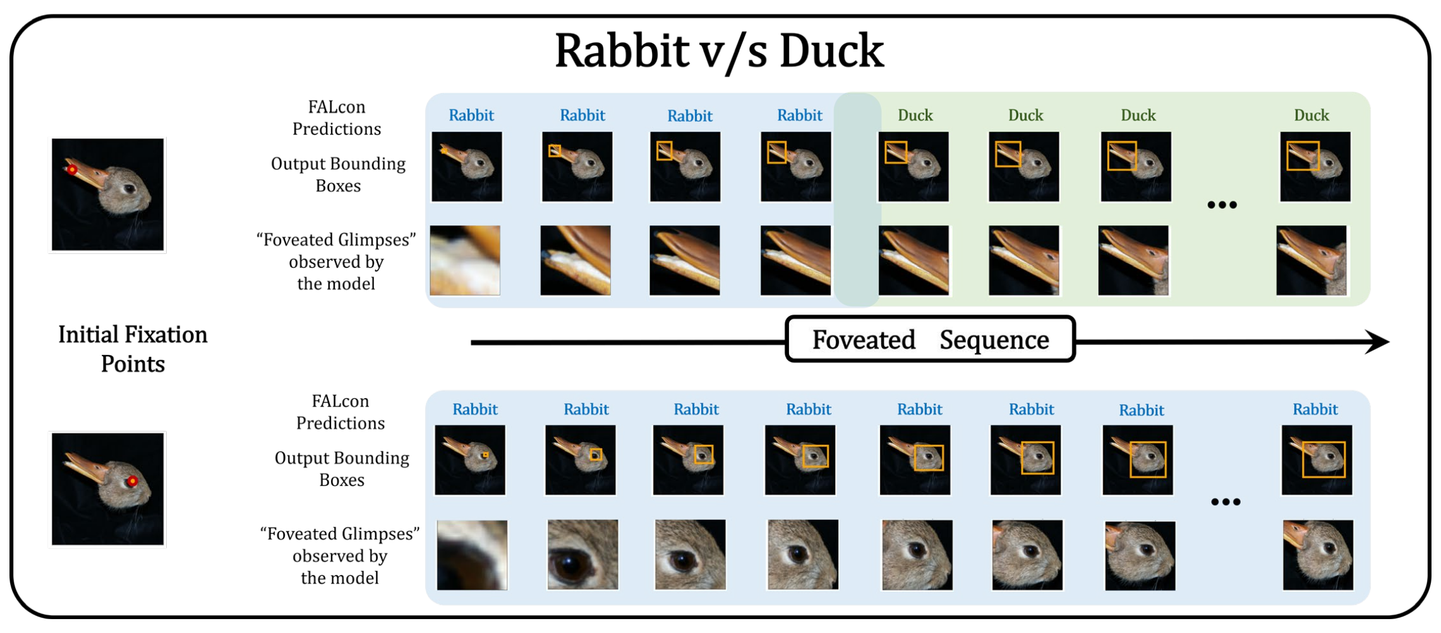
Method	Birds			All ImageNet Classes	
	Training Dataset	Testset (VOC07) $AP_{@0.5}$	Testset (VOC12) $AP_{@0.5}$	Training Dataset	Testset (ImageNet13-Det) $mAP_{@0.5}$
PSOL (baseline) (Zhang et al., 2020a)	CUB	0.32	0.42	ImageNet-1k	9.89
FALcon + PSOL (Ours)	**CUB**	**12.56**	**7.01**	**ImageNet-1k**	**10.32**


PASCAL VOC07 DATASET
(a)
IMAGENET13-DET DATASET
(b)
PSOL predictions "pseudo" bounding box with input ground truth
FALcon predictions output bounding box(es)

Qualitative results obtained on the (a) VOC07 and (b) ImageNet13-Det datasets.

Towards Neuro-inspired Algorithms

- We explored the alignment of our method with human perception using an optical illusion.
- FALcon *observes and focuses* on objects of different classes based on the different fixations.



Key Points

✓ This work models ***foveation*** with extreme cutoff as the method of iterative hard attention, and ***saccades*** as the method of estimating the relevance of each foveated observation.

✓ These mechanisms allow the task of predicting object locations to be reformulated from being a *"passive"* process to an iterative *"active"* process.

✓ The augmented WSOL methods with *"active"* localization enjoy **the benefits of improved and resilient localization.**

CoCoSys
CENTER FOR THE CO-DESIGN OF COGNITIVE SYSTEMS
NSF
DARPA
SRC Semiconductor Research Corporation

SCAN ME