UNIVERSITY OF OXFORD

Unsupervised Object Detection with Theoretical Guarantees

Marian Longa, João F. Henriques

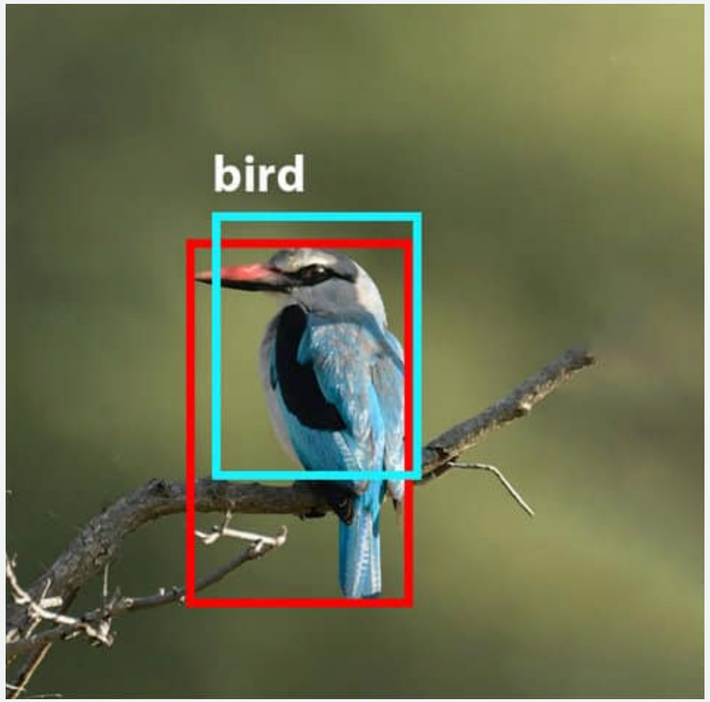
VGG Group, University of Oxford

mlonga@robots.ox.ac.uk, joao@robots.ox.ac.uk

arxiv.org/abs/2406.07284

Motivation

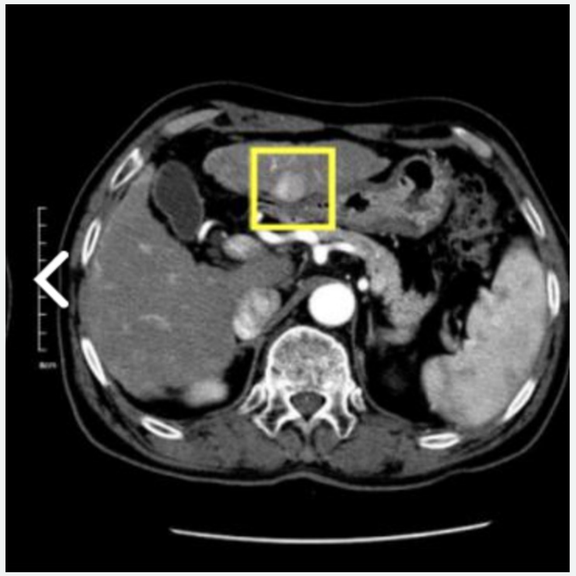
Objects not detected.[1]



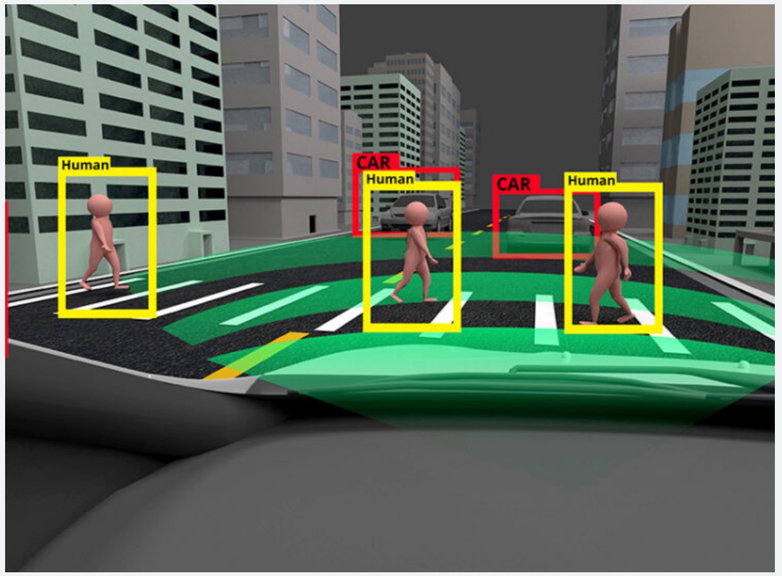
Objects detected inaccurately.[2]

1: Katyal, Sarthak et al. “Object Detection in Foggy Conditions by Fusion of Saliency Map and YOLO.” *2018 12th International Conference on Sensing Technology (ICST)* (2018): 154-159.
2: https://learnopencv.com/intersection-over-union-iou-in-object-detection-and-segmentation/

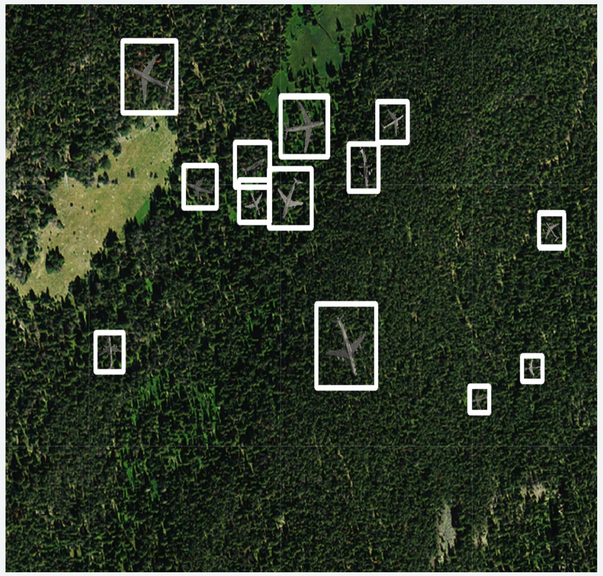
Motivation



Healthcare.[1]



Autonomous driving.[2]



Defence.[3]

1: https://paperswithcode.com/task/medical-object-detection
2: https://intellias.com/cost-effective-3d-object-detection-for-autonomous-vehicles/
3: https://edgeforce.in/defence-aerospace.php

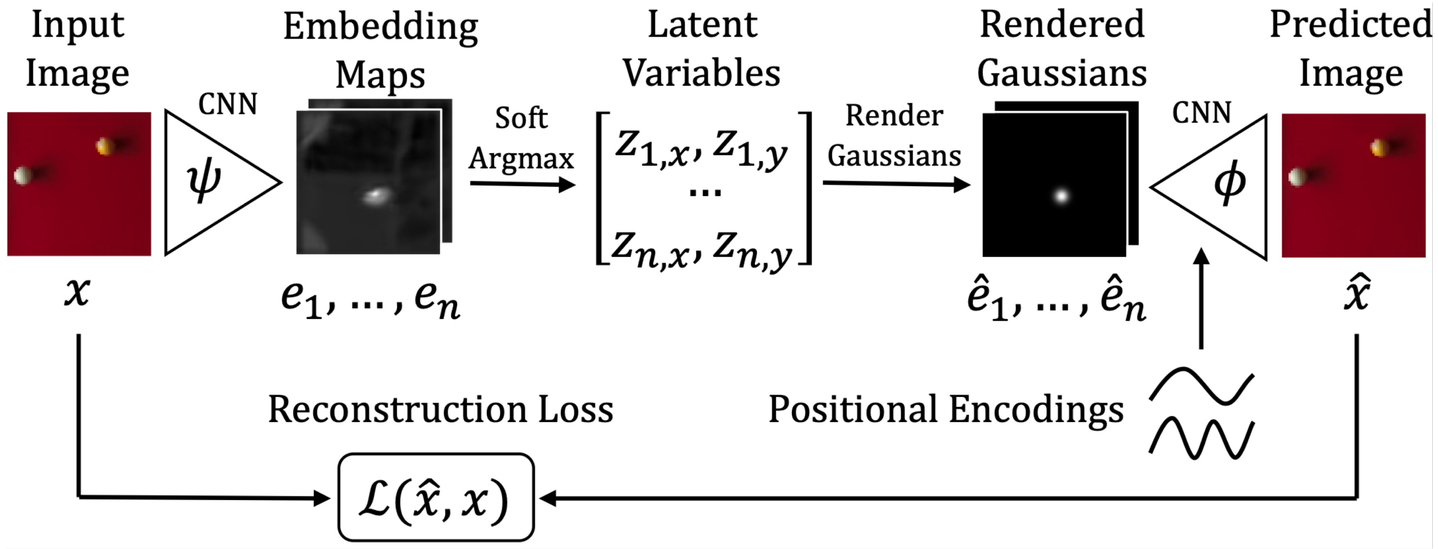
UNIVERSITY OF OXFORD

Can we develop an unsupervised method that is guaranteed to detect objects?

Contributions

1. The first unsupervised object detection method guaranteed to learn true object positions up to small shifts.
2. Proof and derivation of maximum error bounds, using encoder and decoder RF sizes, object sizes, rendering Gaussians widths.
3. Experiments on synthetic, CLEVR, and real images and video, validating our theory up to precisions of individual pixels.

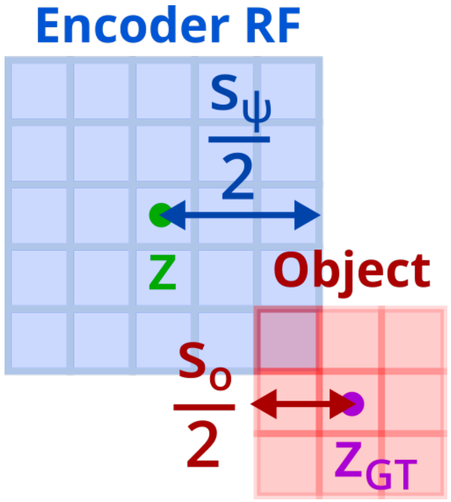
Method

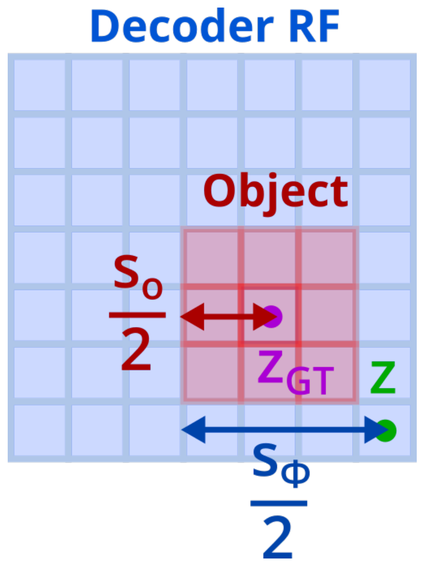


Encoder: (1) an image x is passed through a CNN ψ to obtain n embedding maps e_1, ..., e_n, (2) a maximum of each map is found using softargmax to obtain latent variables [$z_{1,x}$, $z_{1,y}$, ..., $z_{n,x}$, $z_{n,y}$].

Decoder: (1) Gaussians $\hat{e}_1$, ..., $\hat{e}_n$ are rendered at the positions given by the latent variables, (2) the Gaussian maps are concatenated with positional encodings and passed through a CNN φ to obtain the predicted image $\hat{x}$. Finally, x and $\hat{x}$ are used to compute reconstruction loss L($\hat{x}$, x).

Detection Uncertainty





Maximum error due to encoder, $\Delta_\psi = s_\psi / 2 + s_o / 2 - 1$. Maximum error occurs when the encoder and the object are as far away from each other as possible while still overlapping by one pixel.

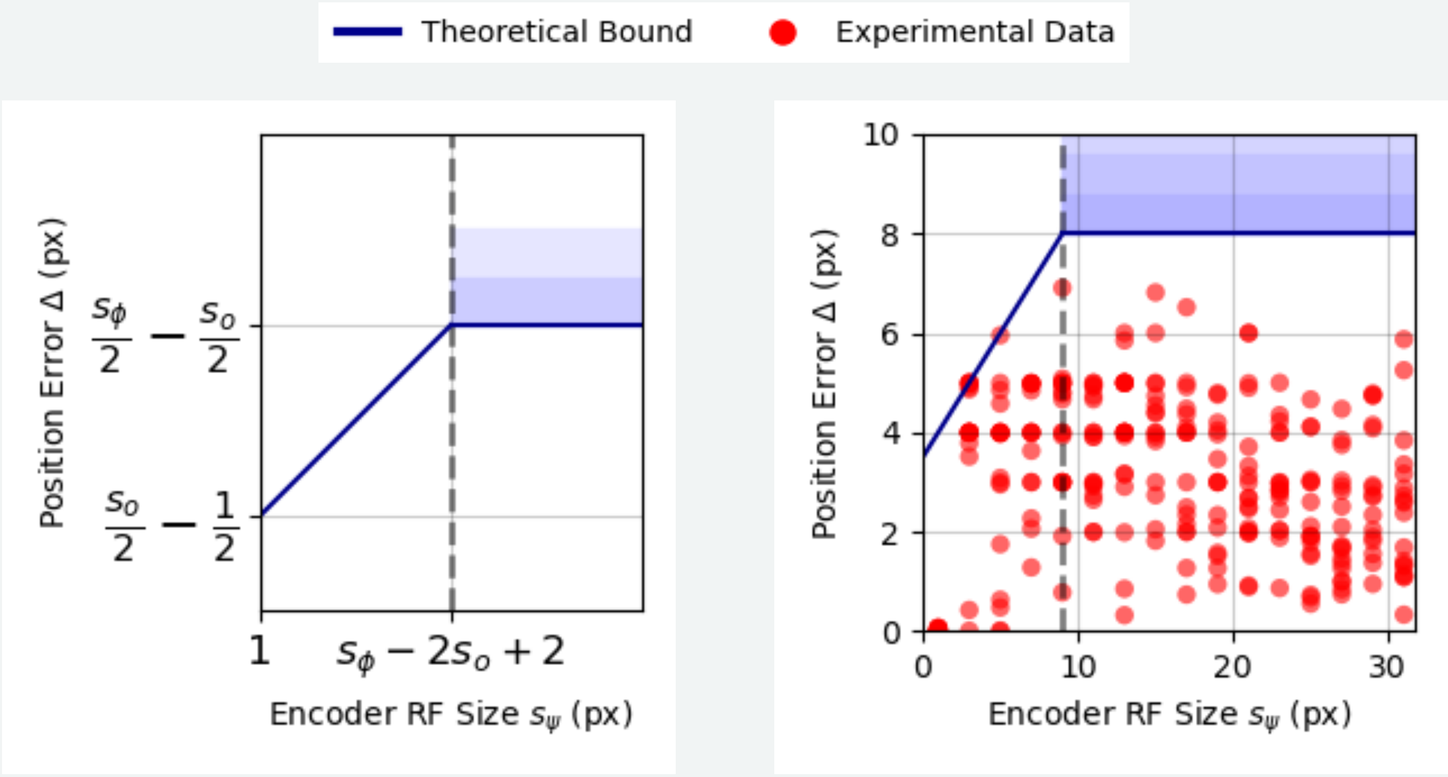
Maximum error due to decoder, $\Delta_\varphi = s_\varphi/2 - s_o/2 + \Delta G$. Maximum error occurs when some part of the Gaussian at position $z + \Delta G$ is within the decoder receptive field (RF) but is as far away from the rendered object as possible.

Maximum Error Bound

Theorem 4.1. Error Bound. *Consider a set of images $x \sim X$ with objects of size s_o, CNN encoder ψ with receptive field size s_ψ, CNN decoder ϕ with receptive field size s_ϕ, soft argmax function* softargmax*, rendering function* render *with Gaussian standard deviation σ_G and $\Delta_G \sim \mathcal{N}(0, \sigma_G^2)$, and latent variables z, composed as $z = \text{softargmax} \circ \psi \circ x$ and $\hat{x} = \phi \circ \text{render} \circ z$ (fig. 1). Assuming (1) the objects are reconstructed at the same positions as in the original images, (2) each object appears in at least two different positions in the dataset, and (3) there are no two identical objects in any image, then the learned latent variables z correspond to the true object positions up to object permutations and maximum position errors Δ of*

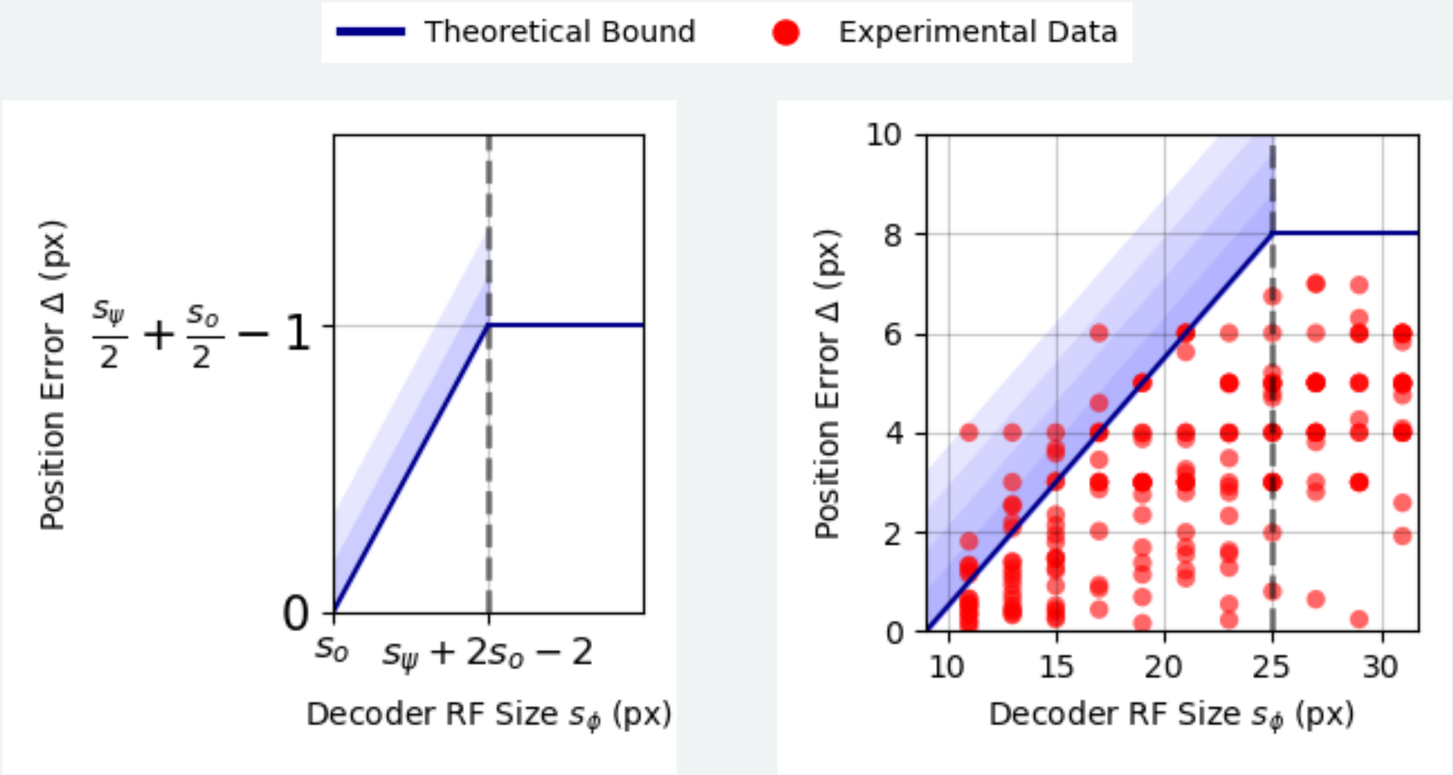
$$\Delta = \min\left(\frac{s_\psi}{2} + \frac{s_o}{2} - 1, \frac{s_\phi}{2} - \frac{s_o}{2} + \Delta_G\right). \tag{9}$$

Maximum Error Bound vs Encoder RF



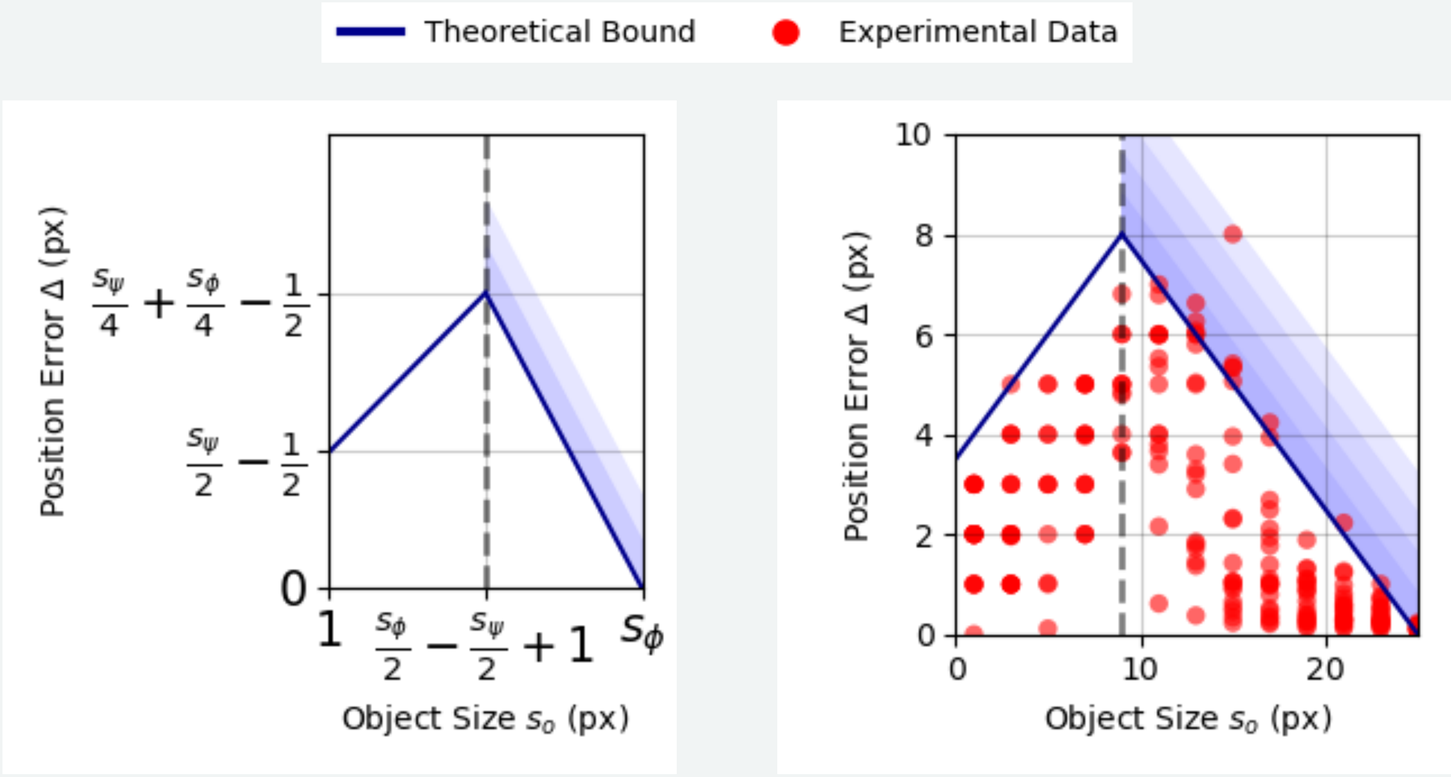
Maximum position error Δ as a function of the encoder receptive field size s_ψ, while fixing the decoder receptive field size $s_\varphi = 25$, object size $s_o = 9$, and Gaussian s.d. $\sigma_G = 0.8$.

Maximum Error Bound vs Decoder RF



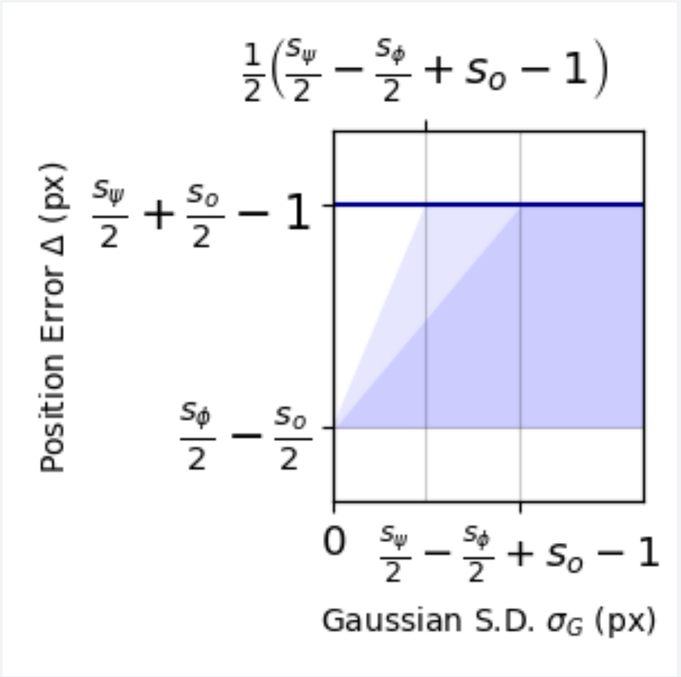
Maximum position error Δ as a function of the decoder receptive field size s_ϕ, while fixing the encoder receptive field size $s_\psi = 9$, object size $s_o = 9$, and Gaussian s.d. $\sigma_G = 0.8$.

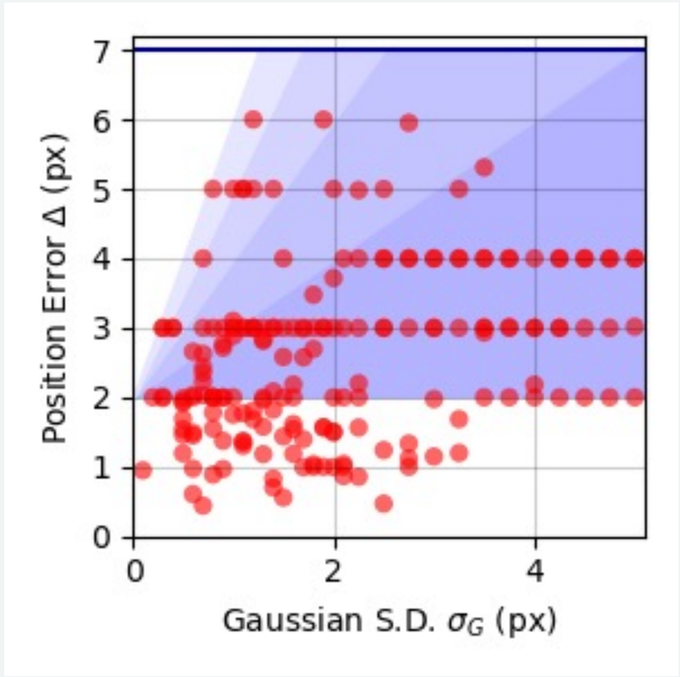
Maximum Error Bound vs Object Size



Maximum position error Δ as a function of the object size s_o, while fixing the encoder receptive field size $s_\psi = 9$, decoder receptive field size $s_\phi = 25$, and Gaussian s.d. $\sigma_G = 0.8$.

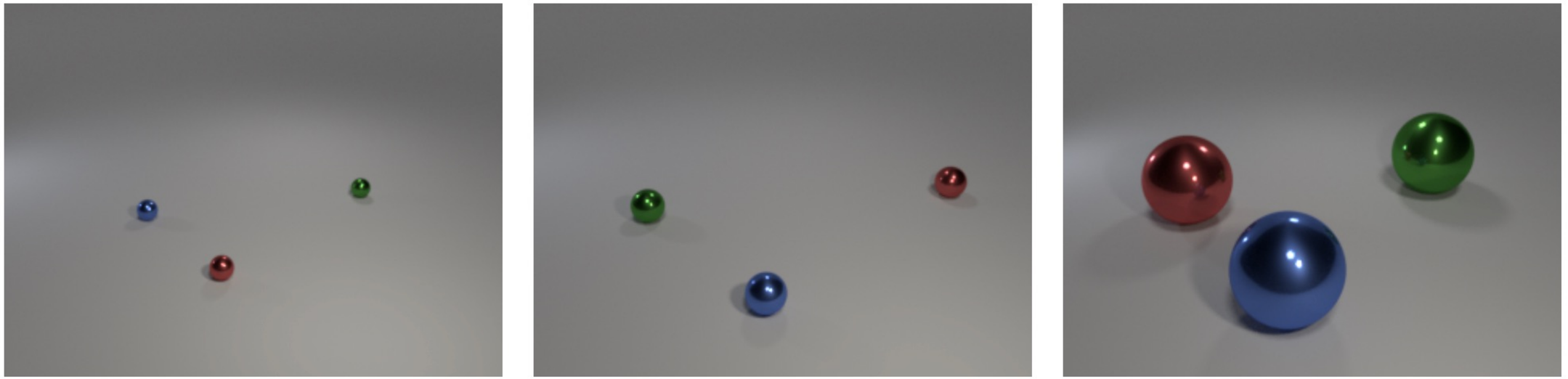
Maximum Error Bound vs Gaussian S.D.





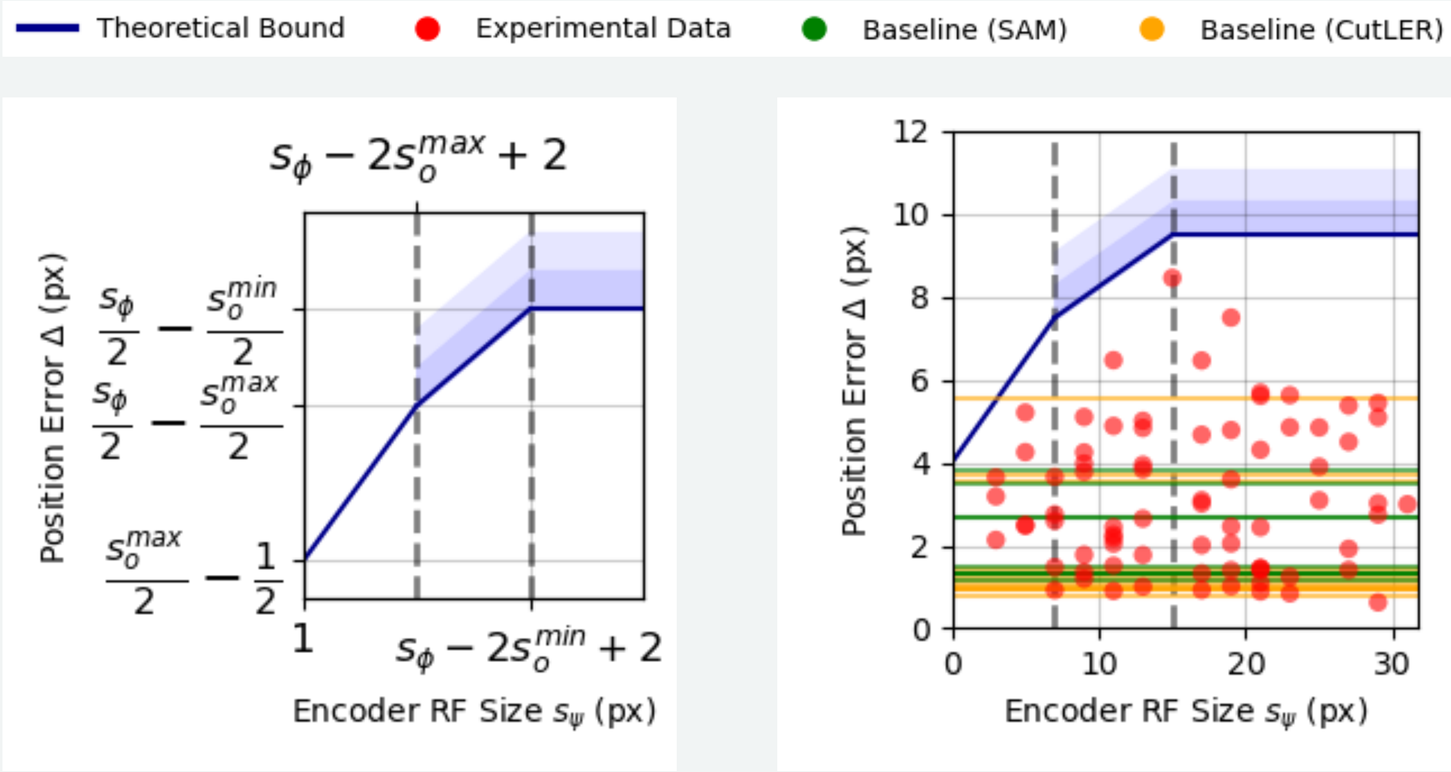
Maximum position error Δ as a function of the Gaussian s.d. σ_G, while fixing the encoder receptive field size $s_\psi = 9$, decoder receptive field size $s_\varphi = 11$, and object size $s_o = 7$.

CLEVR Experiments



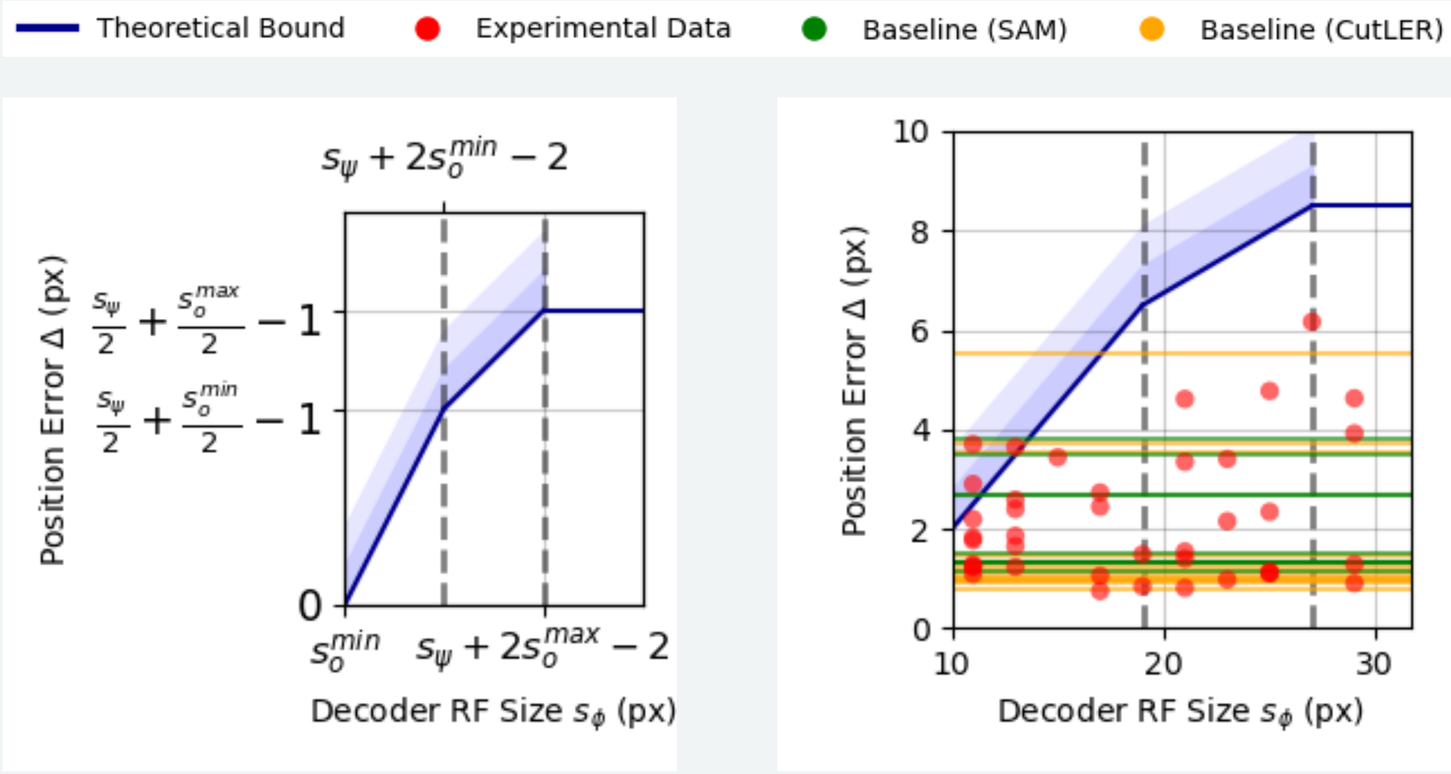
Samples from some of our CLEVR datasets, with object sizes (a) 4-6 px, (b) 6-10 px, (c) 17-27 px.

Maximum Error Bound vs Encoder RF – CLEVR



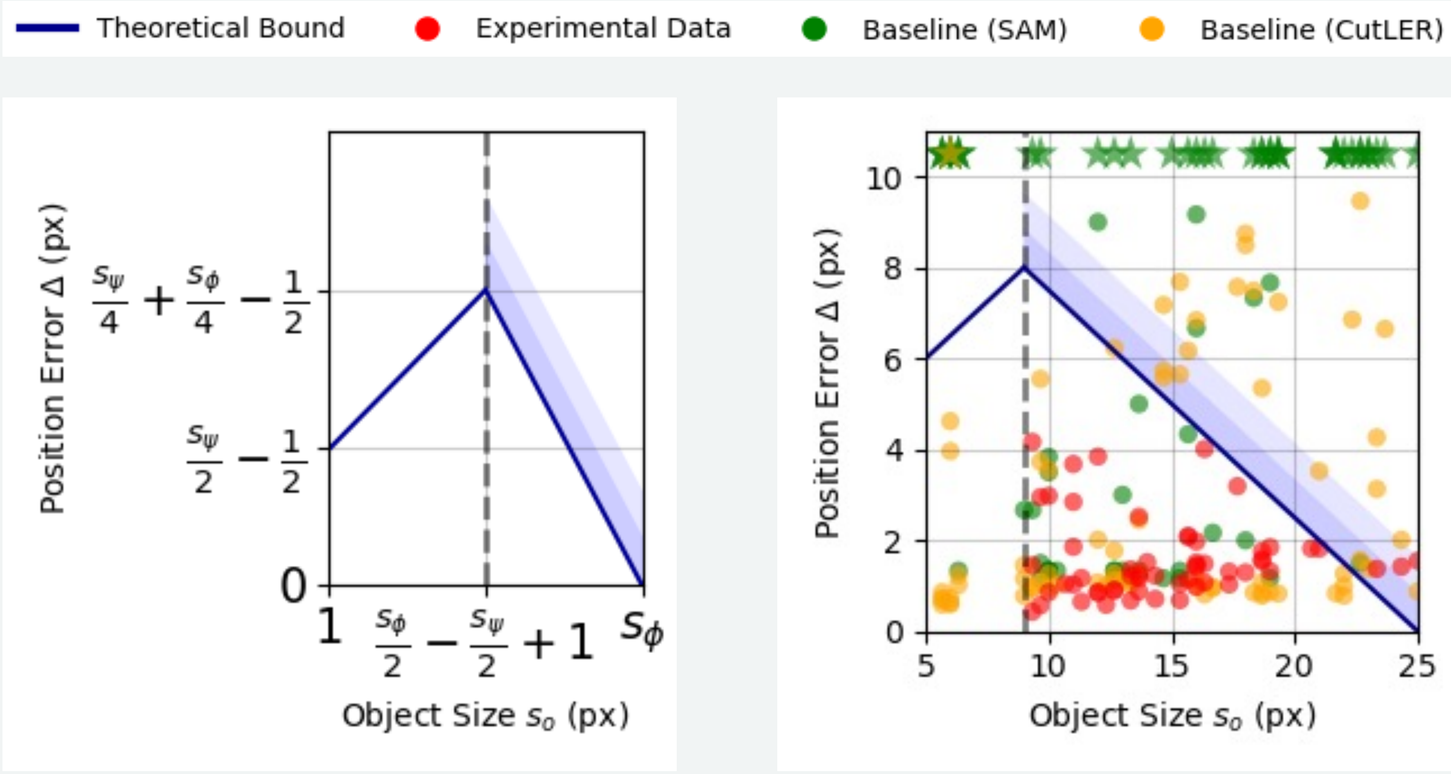
Maximum position error Δ as a function of the encoder receptive field size s_ψ, while fixing the decoder receptive field size $s_\varphi = 25$, object sizes $s_o^{min} = 6$, $s_o^{max} = 10$, and Gaussian s.d. $\sigma_G = 0.8$.

Maximum Error Bound vs Decoder RF – CLEVR



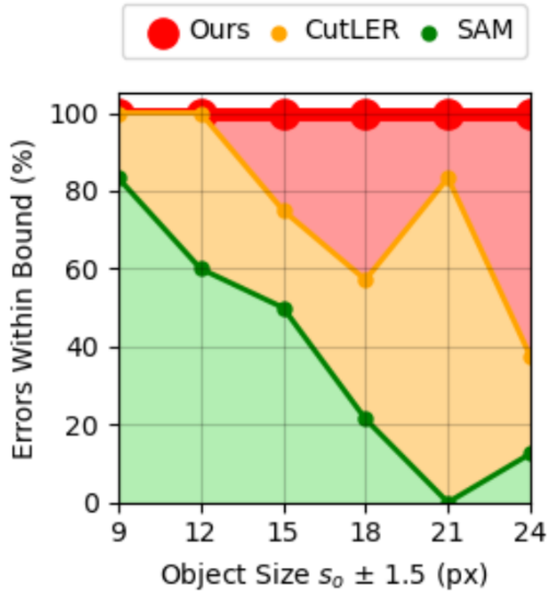
Maximum position error Δ as a function of the decoder receptive field size s_ϕ, while fixing the encoder receptive field size $s_\psi = 9$, object sizes $s_o^{min} = 6$, $s_o^{max} = 10$, and Gaussian s.d. $\sigma_G = 0.8$.

Maximum Error Bound vs Object Size – CLEVR



Maximum position error Δ as a function of the decoder receptive field size s_ϕ, while fixing the encoder receptive field size $s_\psi = 9$, decoder receptive field size $s_\varphi = 25$, and Gaussian s.d. $\sigma_G = 0.8$.

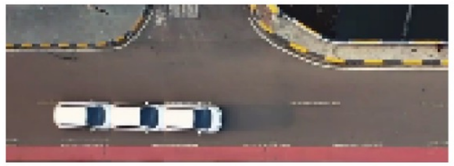
Proportion of Errors Within Bound – CLEVR



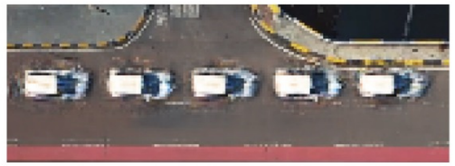
Errors Within Bound (%)							
	Object Size ± 1.5 (px)						
Method	All	9	12	15	18	21	24
Ours	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
CutLER	78.4	**100.0**	**100.0**	75.0	57.1	83.3	37.5
SAM	37.0	83.3	60.0	50.0	21.4	0.0	12.5

Proportion of position errors within 2 standard deviations of the theoretical bound (%), reported for different object sizes and methods. Results from table (right) are visualised in plot (left).

Real Video Experiments

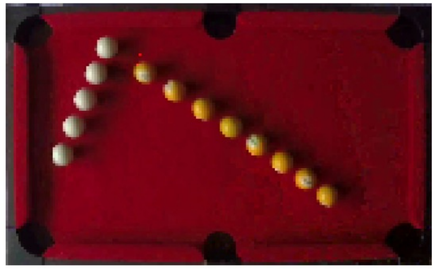


(a) Training data.



(b) Generated data (steady speed in a different lane).

(c) Generated data (lane change and acceleration).



(a) Training data.

(b) Generated data (linear motion at unseen positions).

(c) Generated data (collision and slowing down).

Real videos used for training (left column), together with two videos generated after training by modifying and decoding the learned latent variables (middle, right columns). Video frames are superimposed.

Conclusion

- Presented the first unsupervised object detection method guaranteed to detect objects up to small shifts.
- Proved theoretical bounds for error in position depending on encoder and decoder RF sizes, object sizes, and rendering Gaussian widths.
- Validated method on synthetic, CLEVR and real image and video experiments.
- We hope this work helps open up an avenue of research into object detection methods possessing theoretical guarantees.

UNIVERSITY OF OXFORD

Unsupervised Object Detection with Theoretical Guarantees

Marian Longa, João F. Henriques

VGG Group, University of Oxford

mlonga@robots.ox.ac.uk, joao@robots.ox.ac.uk

arxiv.org/abs/2406.07284