Machine Learning and Data Intensive Computing (Mining) LAB

NEURAL INFORMATION PROCESSING SYSTEMS

# Adaptive Important Region Selection with Reinforced Hierarchical Search for Dense Object Detection

**Dingrong Wang, Hitesh Sapkota, Qi Yu**

Golisano College of Computing and Information Sciences
Rochester Institute of Technology (RIT)

# Background

- Dense object detection enjoys a wide of **applications**, including *surveillance video tracking* by the police and *merchandise recognition* for online shopping.
- An inherently **challenging** is: it requires predicting the bounding boxes for all objects present in a given image irrespective of their shape, size, and number.
- The inborn complexity of images, such as **shadow/occlusion, image size, shape, color, and texture** could also pose a significant ***hindrance*** in the detection process resulting in a lower accuracy.
- Existing efforts have been made to address above key challenges, including two-stage (R-CNN [1]) and one-stage (RetinaNet [2], FCOS [3]) approaches.

[1] Girshick, Ross, et al. "Rich feature hierarchies for accurate object detection and semantic segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2014. [2] Lin, Tsung-Yi et al. “Focal Loss for Dense Object Detection.” 2017 IEEE International Conference on Computer Vision (ICCV) (2017): 2999-3007. [3] Tian, Zhi, et al. "Fully convolutional one-stage 3d object detection on lidar range images." Advances in Neural Information Processing Systems 35 (2022): 34899-34911.

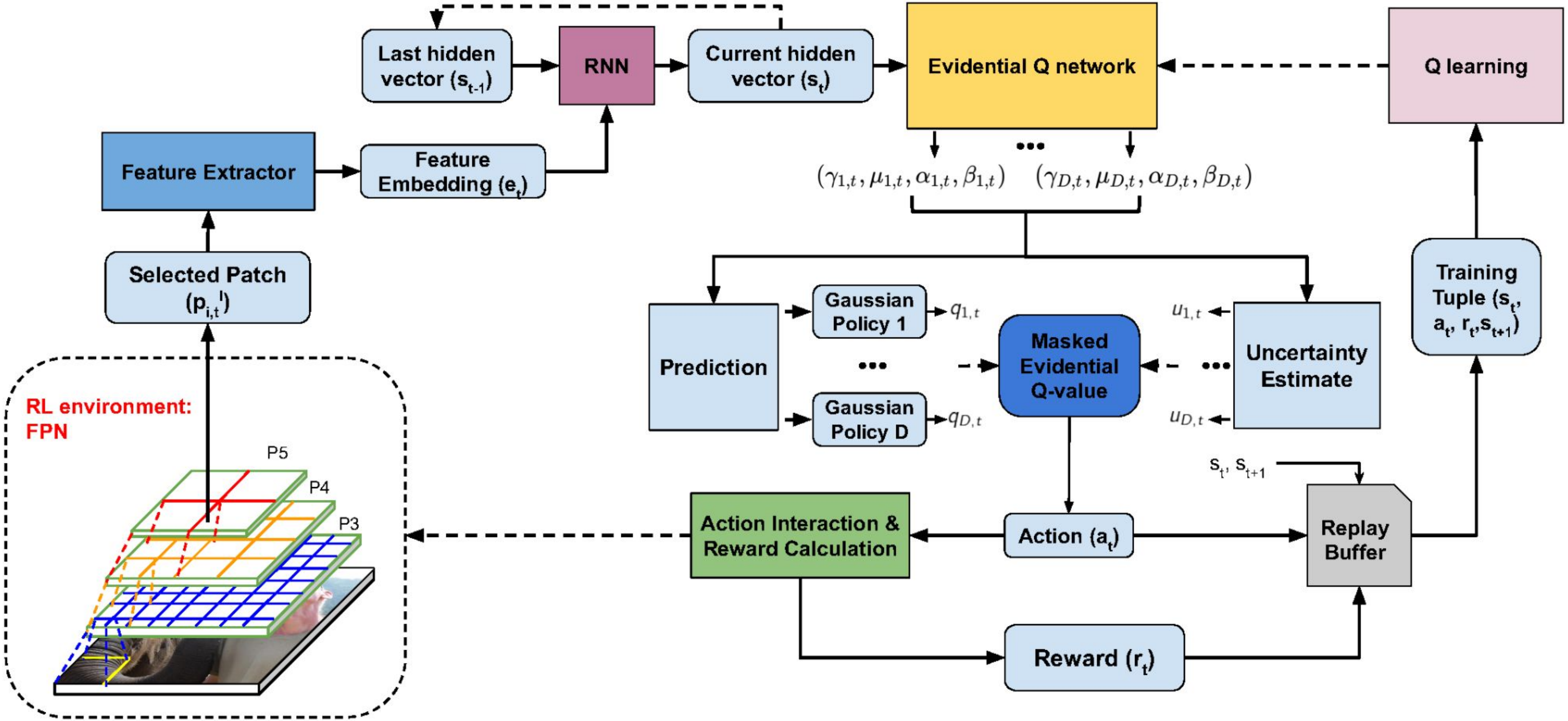
# Challenge


(a) GFocal (b) AIRS (c) GFocal-V2 (d) AIRS

Figure 1: Bounding boxes produced by GFocal [4], GFocal-V2 [5], and AIRS, where GFocal, GFocal-V2 still tend to generate unnecessary bounding boxes resulting from false positive anchors, comparing to the proposed AIRS model.

[4] Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. In NeurIPS, 2020. [5] Xiang Li, Wenhai Wang, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss v2: Learning reliable localization quality estimation for dense object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 11632–11641, June 2021.

# Methodology



# Generation of Masked Evidential Q-value

**We use masked evidential Q-value to select optimal action, and the reward is measured by target patch quality score resulting from that action.**

$$q_{d,t} \sim \mathcal{N}(\cdot|\mu_{d,t}, \sigma^2_{d,t}),\ \mu_{d,t} \sim \mathcal{N}(\cdot|\gamma_{d,t}, \sigma^2_{d,t}\nu^{-1}_{d,t}),\ \sigma^2_{d,t} \sim \text{Inv-Gamma}(\cdot|\alpha_{d,t}, \beta_{d,t}) \quad (1)$$

$$q_{d,t} \sim \mathcal{N}\left(\cdot|\gamma_{d,t}, \frac{\beta_{d,t}}{(\alpha_{d,t}-1)}\right) \quad (2)$$

$$q^e_{d,t} = q_{d,t} + \lambda \text{Var}[\mu_{d,t}], \quad \text{Var}[\mu_{d,t}] = \frac{\mathbb{E}[\sigma^2_{d,t}]}{\nu_{d,t}} = \frac{\beta_{d,t}}{\nu_{d,t}(\alpha_{d,t}-1)} \quad (3)$$

$$\widetilde{\mathbf{q}^e_{d,t}} = \mathbf{q}^e_{d,t} \otimes \mathbf{m}^d_{l,t} \quad (1)$$

# Experiment Results

**Table 1: Detection performance comparison on all three datasets along with their challenging subsets**

| Category | Method | MS COCO | | | | | Pascal VOC 2012 | | | | | Open Image V4 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | AP | $AP^S$ | $AP^M$ | $AP^L$ | $AP^{CH}$ | AP | $AP^S$ | $AP^M$ | $AP^L$ | $AP^{CH}$ | AP | $AP^S$ | $AP^M$ | $AP^L$ | $AP^{CH}$ |
| Two-stage | Faster R-CNN [33] | 36.2 | 18.2 | 39.0 | 48.2 | 19.4 | 73.8 | 25.2 | 75.2 | 78.4 | 26.5 | 37.4 | 19.6 | 38.5 | 42.2 | 20.5 |
| | Cascade R-CNN [7] | 42.8 | 23.7 | 45.5 | 55.2 | 22.5 | 82.7 | 29.5 | 73.6 | 83.5 | 28.6 | 38.6 | 25.4 | 40.4 | 44.8 | 23.7 |
| | RepPoints [41] | 41.0 | 23.6 | 44.1 | 51.7 | 21.2 | 81.3 | 29.1 | 74.4 | 83.0 | 27.6 | 39.1 | 24.2 | 39.1 | 42.5 | 21.5 |
| | TridentNet [24] | 42.7 | 23.9 | 46.6 | 56.6 | 20.5 | 82.5 | 29.5 | 64.3 | 84.7 | 28.4 | 40.5 | 26.2 | 41.9 | 45.8 | 20.4 |
| | DETR [9] | 42.0 | 20.5 | 45.8 | 61.1 | 17.5 | 80.2 | 25.1 | 62.8 | 84.5 | 26.3 | 39.6 | 23.5 | 41.5 | 45.9 | 17.8 |
| | Co-DETR [49] | 42.5 | 20.8 | 46.2 | 61.5 | 17.9 | 80.5 | 25.4 | 63.2 | 84.9 | 26.5 | 39.7 | 23.9 | 41.8 | 46.3 | 18.3 |
| | EVA [14] | 46.7 | 28.5 | 48.2 | 61.9 | 28.8 | 84.7 | 31.5 | 75.4 | 86.5 | 28.7 | 44.1 | 25.8 | 46.5 | 50.8 | 26.7 |
| | DINO-4scale [44] | 47.8 | 30.2 | 50.1 | 62.3 | 29.0 | 86.9 | 33.4 | 77.2 | 88.5 | 30.9 | 46.2 | 29.8 | 47.8 | 52.3 | 28.1 |
| | DINO-5scale [44] | 47.9 | 30.0 | **50.4** | **62.5** | 29.0 | 87.1 | 33.3 | 77.4 | 88.6 | 31.2 | 46.4 | 29.9 | 47.7 | 52.4 | 28.2 |
| One-stage | RetinaNet [26] | 39.1 | 21.8 | 42.7 | 50.2 | 21.6 | 77.0 | 27.8 | 62.9 | 81.5 | 27.3 | 38.5 | 24.8 | 40.2 | 42.4 | 21.3 |
| | FCOS [38] | 41.5 | 24.4 | 44.8 | 51.6 | 23.5 | 83.3 | 31.4 | 64.2 | 85.8 | 30.5 | 40.3 | 26.1 | 41.8 | 45.4 | 23.2 |
| | ATSS [45] | 43.6 | 26.1 | 47.0 | 53.6 | 23.8 | 84.2 | 32.6 | 74.3 | 86.9 | 31.3 | 42.2 | 26.9 | 42.5 | 46.8 | 24.0 |
| | SAPD [48] | 43.5 | 24.9 | 46.8 | 54.6 | 22.4 | 83.8 | 31.5 | 75.3 | 86.2 | 29.5 | 41.1 | 25.9 | 41.6 | 45.8 | 23.5 |
| | SpineNet [11] | 41.5 | 23.3 | 45.0 | 58.0 | 21.2 | 82.6 | 29.3 | 73.5 | 85.7 | 27.4 | 40.2 | 25.8 | 41.2 | 45.3 | 21.6 |
| | GFocal [23] | 45.0 | 27.2 | 48.8 | 54.5 | 25.4 | 86.5 | 35.0 | 78.0 | 90.5 | 32.6 | 45.8 | 29.5 | 46.5 | 51.4 | 26.3 |
| Ours | AIRS | **48.3** | **32.1** | 48.5 | 54.3 | **29.4** | **88.7** | **37.3** | **79.0** | **91.5** | **35.6** | **47.5** | **31.5** | **48.1** | **53.1** | **29.0** |

# More detailed analysis

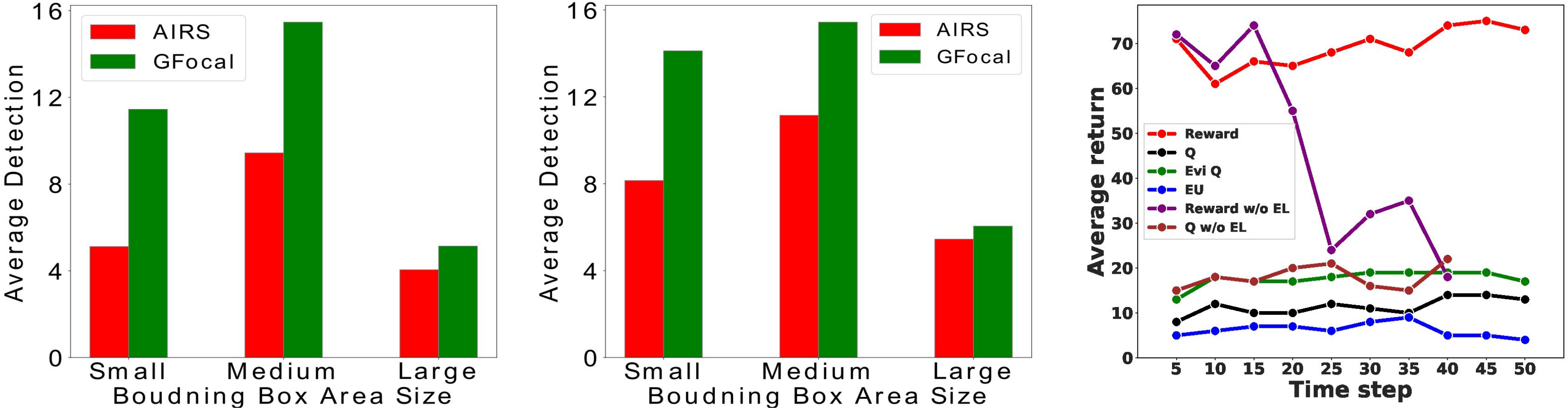



**Figure (a)-(b): Average number of detections per test image based on the bounding box area on MS COCO and OpenImages V4. Figure (c): Ablative study on epistemic uncertainty to deep Q-evaluation.**