Improving Open-World Object Localization by Discovering Background (Supplemental material)

Ashish Singh*
Univ. of Mass.-Amherst
ashishsingh@cs.umass.edu

Michael J. Jones
Mitsubishi Electric Research Labs
mjones@merl.com

Kuan-Chuan Peng
Mitsubishi Electric Research Labs
kpeng@merl.com

Anoop Cherian
Mitsubishi Electric Research Labs
cherian@merl.com

Moitreya Chatterjee
Mitsubishi Electric Research Labs
chatterjee@merl.com

Erik Learned-Miller
Univ.of Mass.-Amherst
elm@cs.umass.edu

In this supplemental material, we first provide more qualitative results comparing BOWL with competing baseline methods: OLN [1] and GGN [5]. We also provide visualization of the patches selected as exemplars to represent non-object information. Finally we evaluate the accuracy of our exemplar set in identifying non-object regions in unseen images and show visualizations of negative anchor boxes used for training BOWL.

1. More qualitative results

We show localization results of BOWL, OLN [1] and GGN [5] on some of the MS-COCO [2] validation set image in Figures 1, 2 and 3. The results are generated using models trained on the 20 VOC categories. For each method, we show all the predicted boxes with objectness score greater than 0.75. From all the qualitative results, we observe that BOWL provides significantly better results. Specifically, as discussed in the main paper, we see that while both OLN and GGN are able to localize unseen objects, both methods suffer from high false-positive (for GGN) and false-negative predictions (for OLN). These qualitative results further support our hypothesis that non-object supervision can boost objectness learning and improve open-set object localization.

2. Non-object exemplar set

Figure 4 shows all the non-object exemplar patches used to identify negative anchor boxes when training BOWL. Specifically, 86436 patches were selected in total using our exemplar selection method. Out of the total set, we further selected the top 1000 patches based on the nearest-neighbor count of each exemplar patch *i.e.* how many patches in the total set of all patches are similar to a given exemplar patch, to create non-object exemplar set. In Figure 4, we show these 1000 non-object patches, in the descending order of their nearest-neighbor count, with patches in the first row representing the most common patches found in MS-COCO [2] training set, which were used to represent the total set of patches. As can be observed from Figure 4, all the patches represent regions generally characterized as background, for example sky, forest *etc*. Furthermore, the order of the exemplar patches also aligns with the frequency of the background semantic regions in the dataset, for example, sky is more common than grassy regions in the dataset. We can also see that the selected exemplar set is visually and semantically diverse, leading to a compact model of non-objectness in the dataset.

3. Ablation study: Number of exemplar samples to model non-objectness

Table 1. Our model's performance with varying sizes of the exemplar set (constructed from the MS-COCO training set) for identifying non-object regions on MS-COCO validation set.

Num Exemplars	Percentage	Average precision (%)
N = 432	0.5%	95.23
N = 864	1%	95.10
N = 1000	1.15%	95.08
N = 4321	5%	94.57
N = 8643	10%	94.50
N = 21609	25%	94.48

As mentioned in the main paper and in the previous section, after extracting the exemplar set of patches, we further subsample exemplars based on the number of nearest-neighbor counts of exemplars. Specifically we select the top N exemplar samples that are most similar to other patches in the original set of all patches. For training BOWL, we selected $N = 1000$ exemplar samples to create a non-object exemplar subset representing a compact model of non-objectness. To further validate our design choice, we conduct an experiment to measure the precision of non-object regions identified in unseen images by varying

*Ashish did most of this work while an intern at MERL.

Figure 1. **Qualitative results of GGN [4], OLN [1], and BOWL on MS-COCO validation images.** Here For each image, (Row 1, Column 1) → Ground Truth bounding boxes; (Row 1, Column 2) → BOWL predictions; (Row 2, Column 1) → GGN predictions; (Row 2, Column 2) → OLN predictions. Green colored boxes refer to ground-truth bounding boxes, while red colored boxes are model predictions with objectness score greater than 0.75

the size of the non-object exemplar set. Concretely, for a given set of non-object exemplar patch embeddings, we categorize patches in a test image as object or non-object by computing the similarity of the test image patch with the exemplar set. Given the binary segmentation of the test image, we compute anchor boxes of a fixed resolution that overlaps with non-object regions. We then compute the overlap of the non-object anchor boxes with ground-truth object boxes from all classes. Based on the overlap between predicted non-object anchor box and ground-truth object bounding box, we calculate the precision of the non-object anchor boxes. The above setup simulates our process of identifying negative anchor boxes used for training BOWL. For the above experiment, we measure the overlap with IoU threshold of 0.1 and fix the anchor box size to 128×128. We conduct the above experiment on MS-COCO [2] validation set with ground-truth bounding boxes from all the 80 COCO categories. We report our results in Table 1. From the results, we can see that with a smaller subset we obtain the highest average precision. As we increase the size of the subset, the precision reduces by a small margin but saturates quickly. This result validates our design choice and confirms our hypothesis that a small subset of most common exemplar patches is sufficient to accurately identify non-object regions in unseen images.

4. Negative anchor box visualizations

We show examples of negative anchor boxes selected using non-object exemplar set on MS-COCO training images in Figure 5. These negative anchors are directly used in training BOWL. For a given training image, the negative anchorboxes are selected for all scales used in general Faster-RCNN architecture [1, 3]. In Figure 5 we show negative anchorboxes computed at two scales for brevity. Selecting negative anchors in multi-scale fashion allows us to accurately localize larger spatial region as non-objects, providing richer supervision during model training. We can observe from the figure that, while the negative anchorboxes donot cover all the non-object regions, it is highly precise in categorizing a region as non-object.

References

[1] Dahun Kim, Tsung-Yi Lin, Anelia Angelova, In So Kweon, and Weicheng Kuo. Learning open-world object proposals without learning to classify. *IEEE Robotics and Automation Letters (RA-L)*, 2022. 1, 2, 3, 4

[2] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1, 2

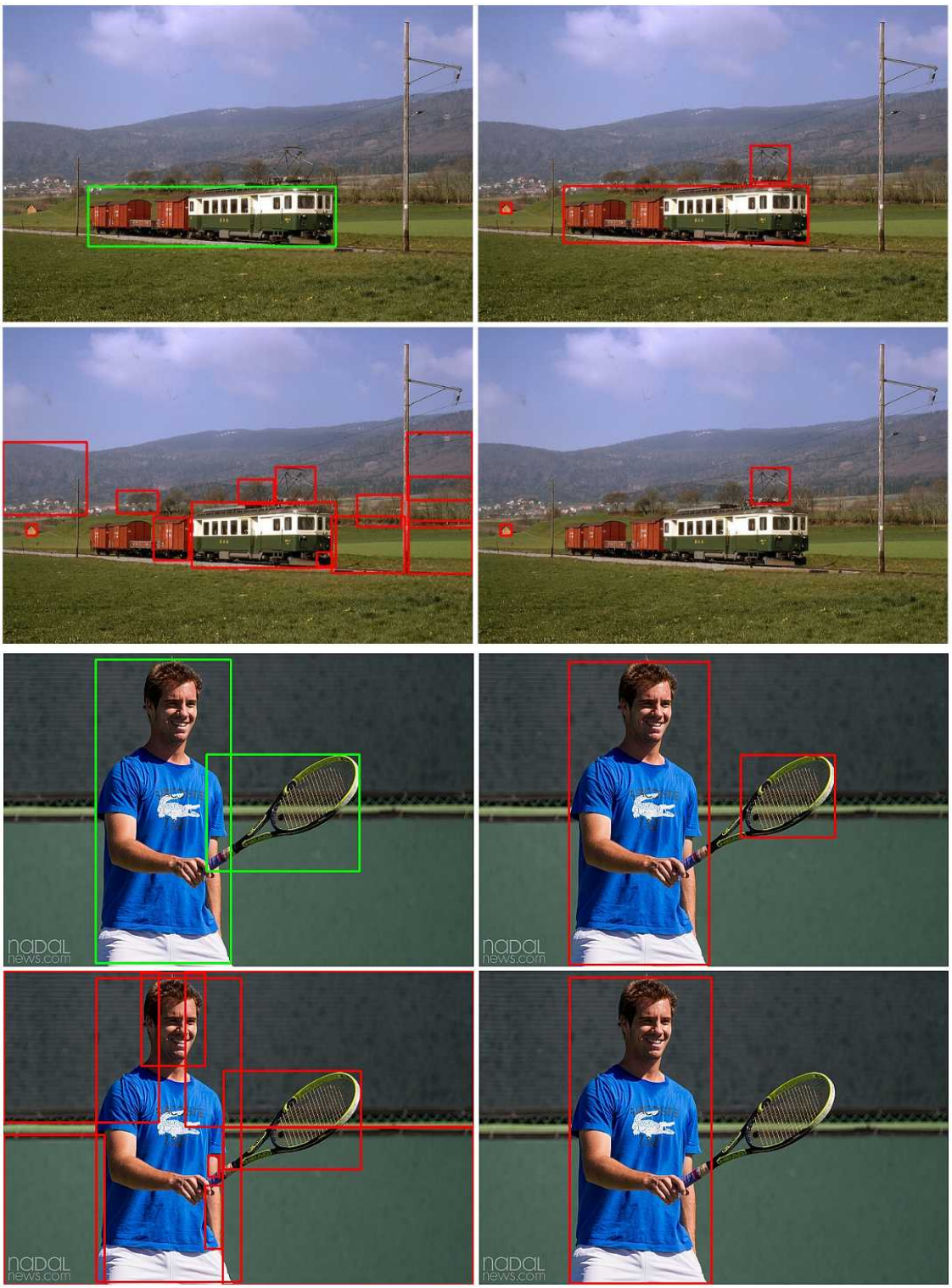

Figure 2. **Qualitative results of GGN [4], OLN [1], and `BOWL` on MS-COCO validation images.** Here For each image, (Row 1, Column 1) $\rightarrow$ Ground Truth bounding boxes; (Row 1, Column 2) $\rightarrow$ `BOWL` predictions; (Row 2, Column 1) $\rightarrow$ GGN predictions; (Row 2, Column 2) $\rightarrow$ OLN predictions. Green colored boxes refer to ground-truth bounding boxes, while red colored boxes are model predictions with objectness score greater than 0.75

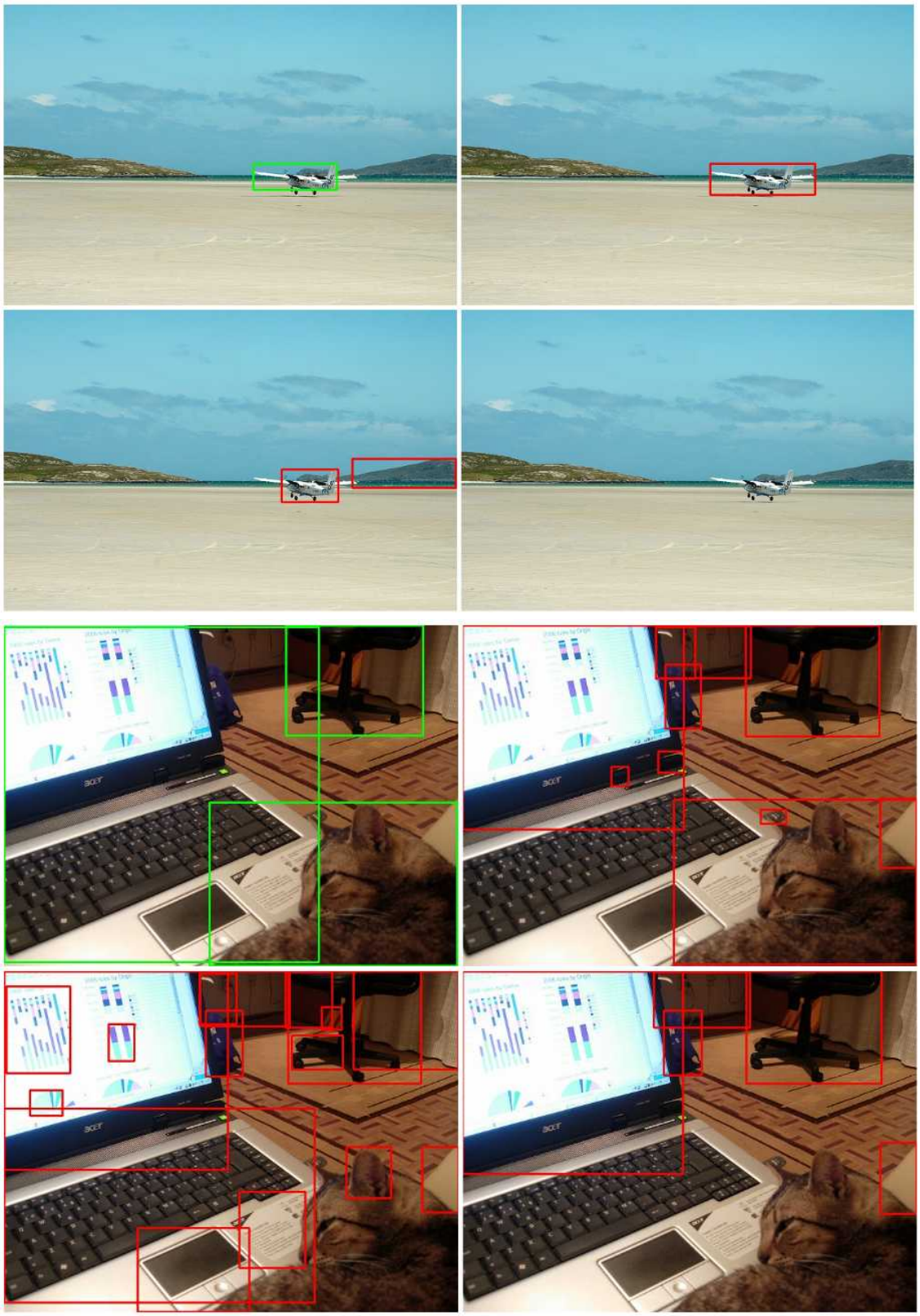

Figure 3. **Qualitative results of GGN [4], OLN [1], and `BOWL` on MS-COCO validation images.** Here For each image, (Row 1, Column 1) $\rightarrow$ Ground Truth bounding boxes; (Row 1, Column 2) $\rightarrow$ `BOWL` predictions; (Row 2, Column 1) $\rightarrow$ GGN predictions; (Row 2, Column 2) $\rightarrow$ OLN predictions. Green colored boxes refer to ground-truth bounding boxes, while red colored boxes are model predictions with objectness score greater than 0.75

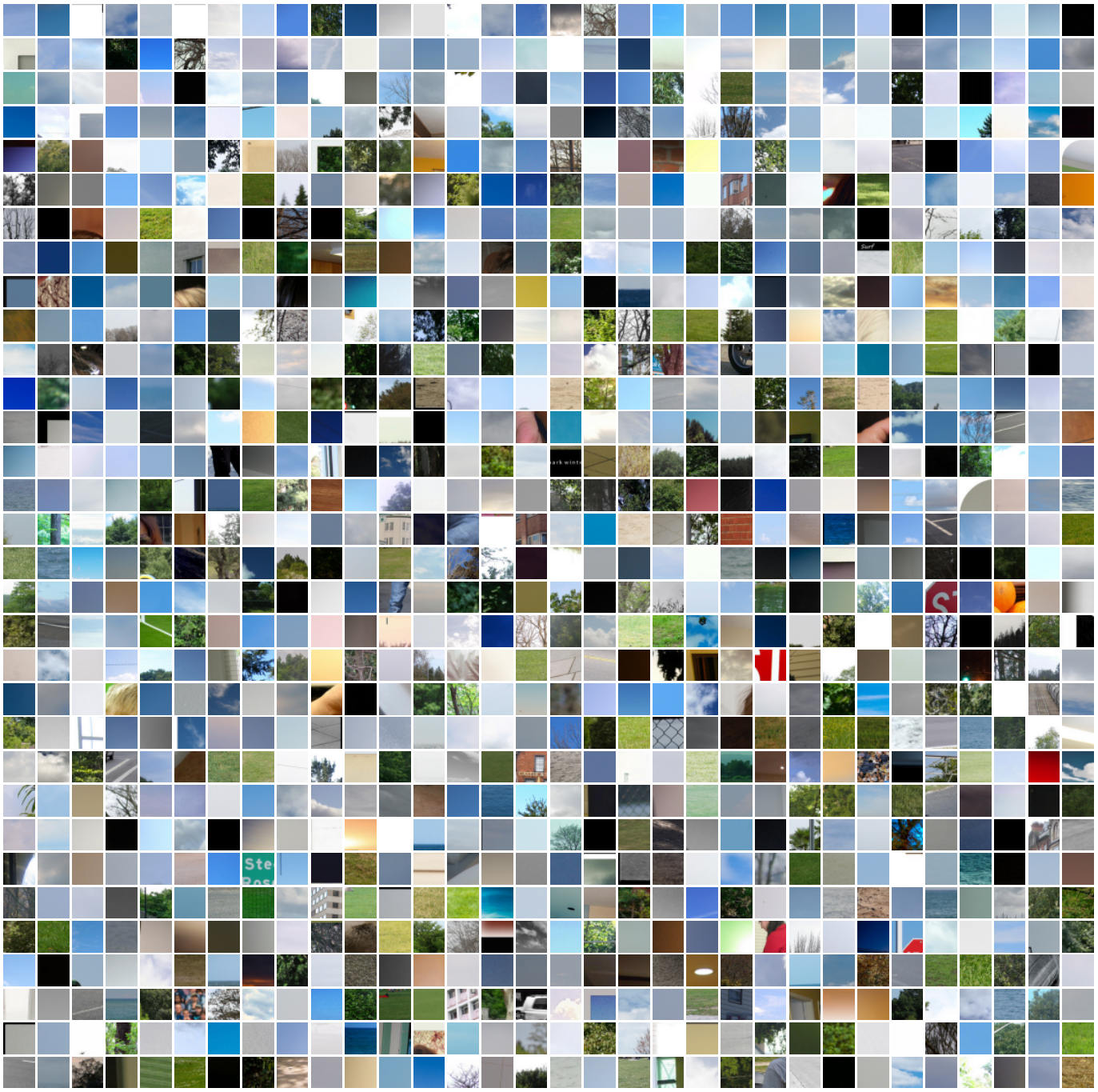

Figure 4. Non-object exemplar patches, constructed from the training set of the COCO dataset

[3] Shaoqing Ren. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015. 2, 6

[4] Jiaqi Wang, Kai Chen, Shuo Yang, Chen Change Loy, and Dahua Lin. Region proposal by guided anchoring. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2965–2974, 2019. 2, 3, 4

[5] Weiyao Wang, Matt Feiszli, Heng Wang, Jitendra Malik, and Du Tran. Open-world instance segmentation: Exploiting pseudo ground truth from learned pairwise affinity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4422–4432, 2022. 1

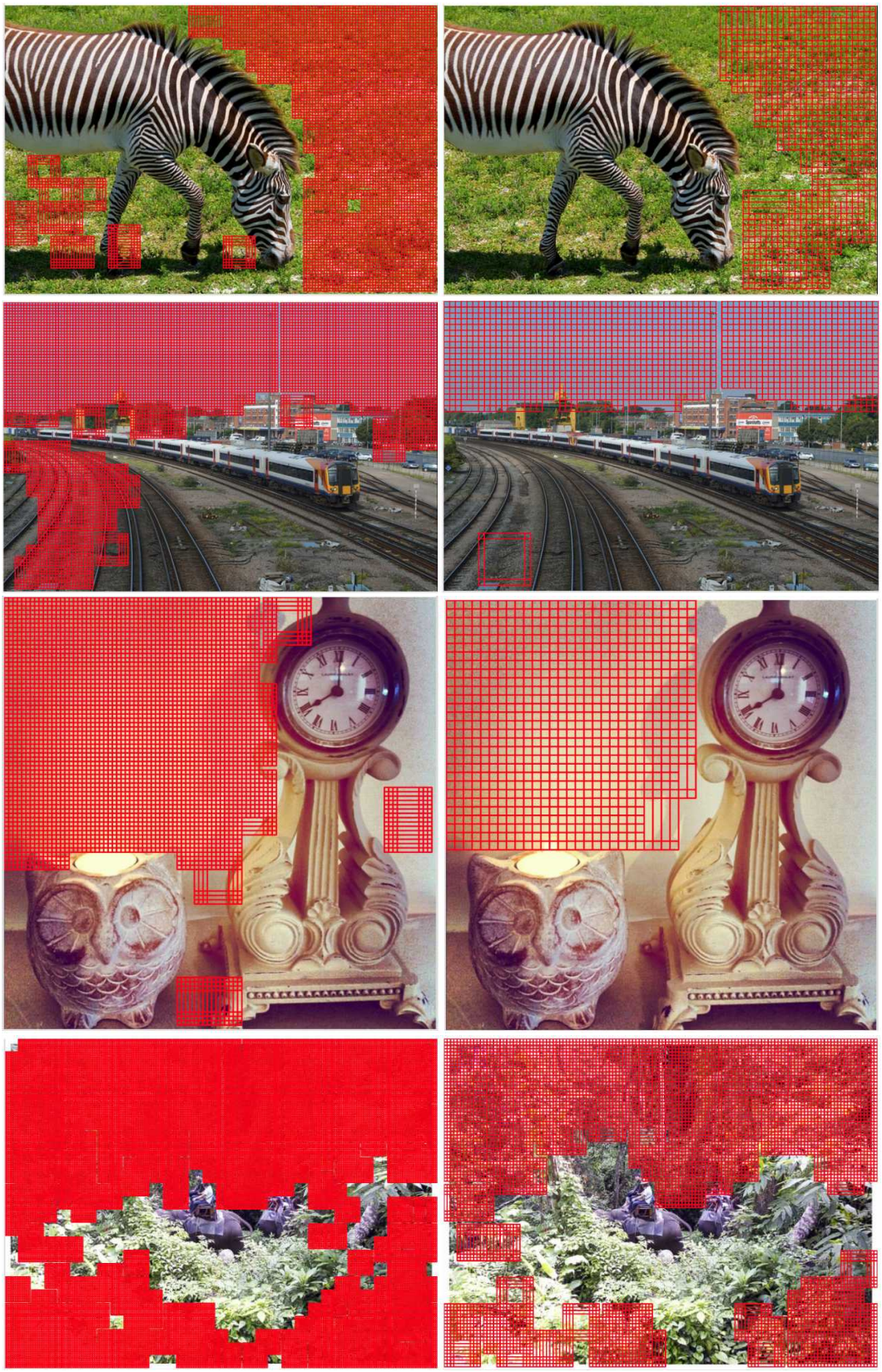

Figure 5. Negative anchorboxes on MS-COCO training set. Here we show negative anchors generated at two scales. For model training 5 scales are used, a common design choice followed in Faster-RCNN [3] type architectures.