Appendix

In this document, we provide:

- An extensive study on the LVIS [18] dataset where UFO^2 learns to detect more than 1k different objects without bounding box annotation (Sec. A), which demonstrates the generalizability and applicability of UFO^2
- Details regarding the annotation policies mentioned in Sec. 5.4 'Budget-aware Omni-supervised Detection' of the main paper
- Additional visualizations of the simulated partial labels
- Additional qualitative results on COCO (complementary to main paper Sec. 5.2)

A Extensions: Learning to Detect Everything

In this section we show that without any architecture changes UFO^2 can be generalized to detect any objects given image-level tags. Hence we follow [21] and refer to this setting as 'learning to detect everything.'

Specifically, we aim to detect objects from the LVIS [18] dataset: it contains 1239 categories and we only use tags as annotation. We use COCO*[†] with boxes and train UFO^2 on it first. Our model achieves 32.7%AP and 52.3%AP-50 on minival. We then jointly fine-tune this model using tags from LVIS and boxes from COCO*. The final model performs comparably on minival (31.6%AP, 50.1%AP-50) and also decent on LVIS validation of over 1k classes (3.5%AP, 6.3%AP-50 where a supervised model achieves 8.6%AP, 14.8%AP-50). To the best of our knowledge, no numbers have been reported on this dataset using weak labels. Our results are also not directly comparable to strongly supervised results [18] as we don't use the bounding box annotation on LVIS.

Qualitative results are shown in Fig. 6. We observe that UFO^2 is able to detect objects accurately even though no bounding box supervision is used for the new classes (*e.g.*, short pants, street light, parking meter, frisbee, *etc.*). Specifically, UFO^2 can (1) detect spatially adjacent or even clustered instances with great recall (*e.g.*, goose, cow, zebra, giraffe); (2) recognize some obscure or hard objects (*e.g.*, wet suite, short pants, knee pad); (3) localize different objects with tight and accurate bounding boxes. Importantly, note that we don't need to change the architecture of UFO^2 at all to integrate both boxes and tags as supervision.

B Annotation Policies

To study budget-aware omni-supervised object detection, we defined the following policies: 80%B, 50%B, 20%B motivated by the following findings: (1) among the three partial labels (tags, points, and scribbles), labeling of points (88.7 s/img; see Sec. 5.4 in the main paper) is roughly as efficient as annotating tags (80

[†] COCO*: because LVIS is a subset of COCO-115, we construct COCO* by taking COCO-115 images excluded from LVIS.

s/img), both of which require half the time/cost of scribbles (160.4 s/img); (2) using points achieves a consistent performance boost compared to using tags (12.4 over 10.8 AP in Tab. 2; 30.1 over 29.4 AP in Tab. 3); (3) using scribbles is just slightly better than using points (13.7 over 12.4 AP in Tab. 2; 30.9 over 30.1 AP in Tab. 3) but twice as expensive to annotate; and (4) strong supervision (boxes) is still necessary to achieve good results (strongly supervised models are significantly better than others in Tab. 2 and Tab. 3).

Therefore, we choose to combine points and boxes as a new annotation policy which we found to work well under the fixed-budget setting as shown in Tab. 6: `80%B` is slightly better than `STRONG` and `50%B` also performs better than `EQUAL-NUM`. These results suggest that spending some amount of cost to annotate more images with points is a better annotation strategy than the commonly-adopted bounding box only annotation (`STRONG`). Meanwhile, the optimal annotation policy remains an open question and better policies may exist if more accurate scribbles are collected or advanced algorithms are developed to utilize partial labels.

C Additional Visualization of Partial Labels

We show additional results together with the ground-truth bounding boxes in Figs. 7–10. Fig. 7 and Fig. 8 show labels for single objects (*e.g.*, car, motor, sheep, chair, person, and bus) and Fig. 9 and Fig. 10 visualize labels for all the instances in the images.

We observe: (1) both points and scribbles are correctly located within the objects; (2) points are mainly located around the center area of the objects and with a certain amount of randomness, which aligns with our goal to mimic human labeling behavior as discussed in Sec. 4 of the main paper; (3) the generated scribbles are relatively simple yet effective in capturing the rough shape of the objects. Also, they exhibit a reasonable diversity. These partial labels serve as a proof-of-concept to show the effectiveness of the proposed UFO^2 framework.

D Additional Qualitative Results

In Fig. 11 and Fig. 12 we show additional qualitative results. We compare the same VGG-16 based model trained on `COCO-80` with different forms of supervision. From left to right we show predicted boxes and their confidence scores when using boxes, scribbles, points, and tags. Similar to the results in Sec. 5.2 of the main paper, we find that stronger labels better reduce false positive predictions and better localize true positive predictions.

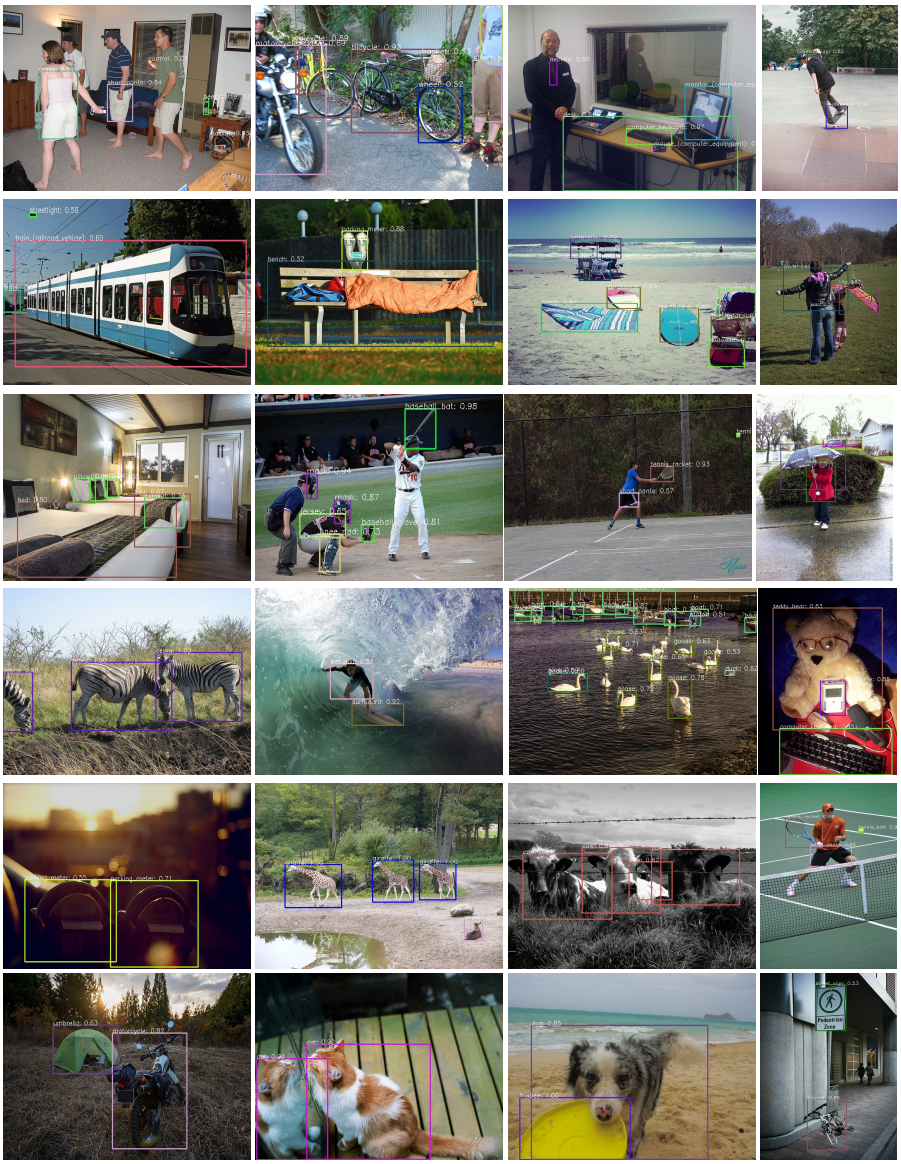

Fig. 6. Visualization of results on `LVIS` data.

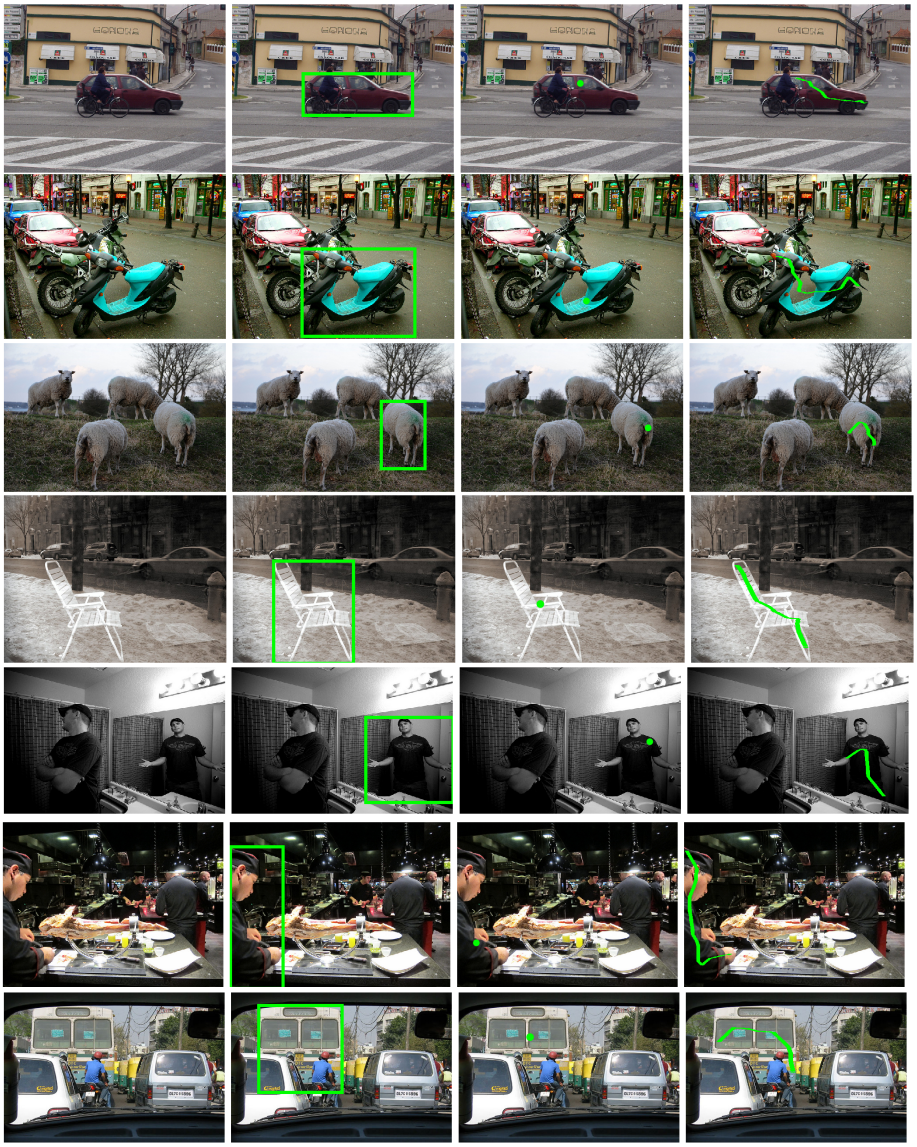

Fig. 7. Additional visualization of the ground-truth boxes and the simulated partial labels.

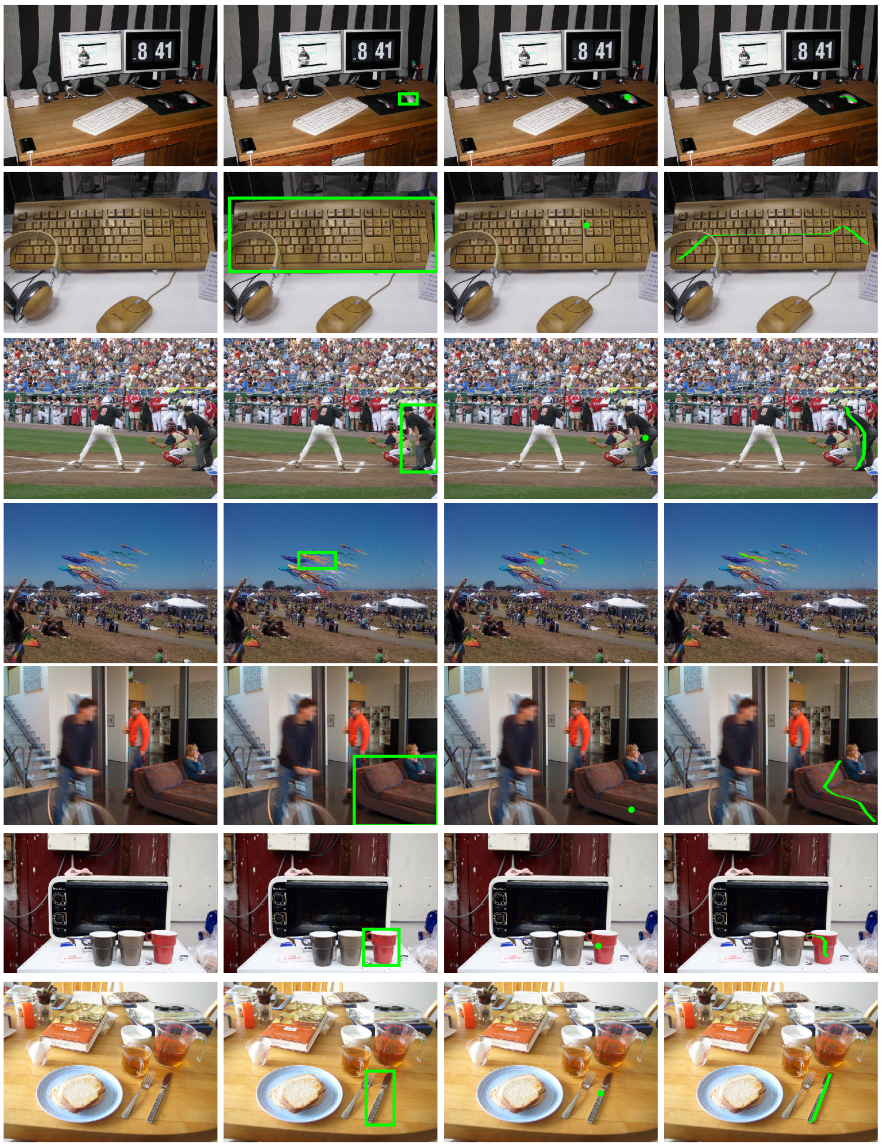

Fig. 8. Additional visualization of the ground-truth boxes and the simulated partial labels.

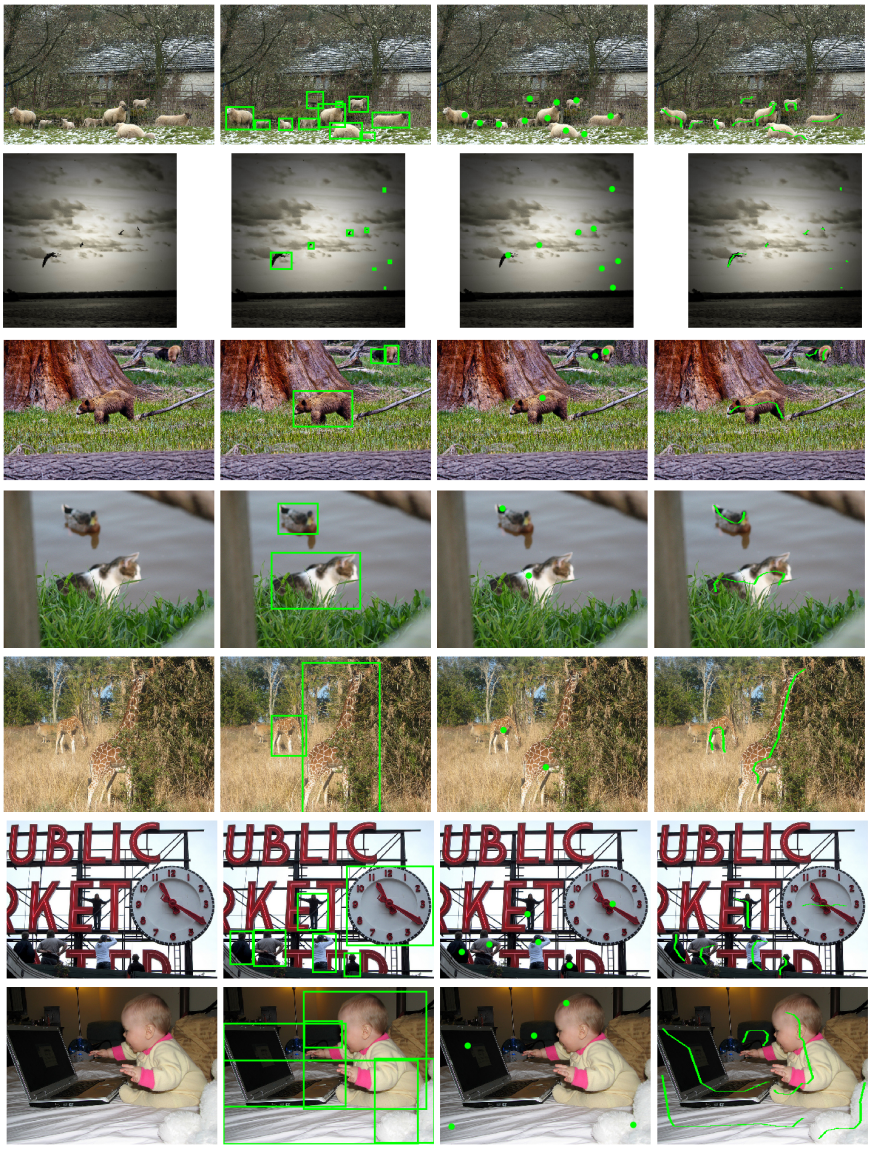

Fig. 9. Additional visualization of the ground-truth boxes and the simulated partial labels.

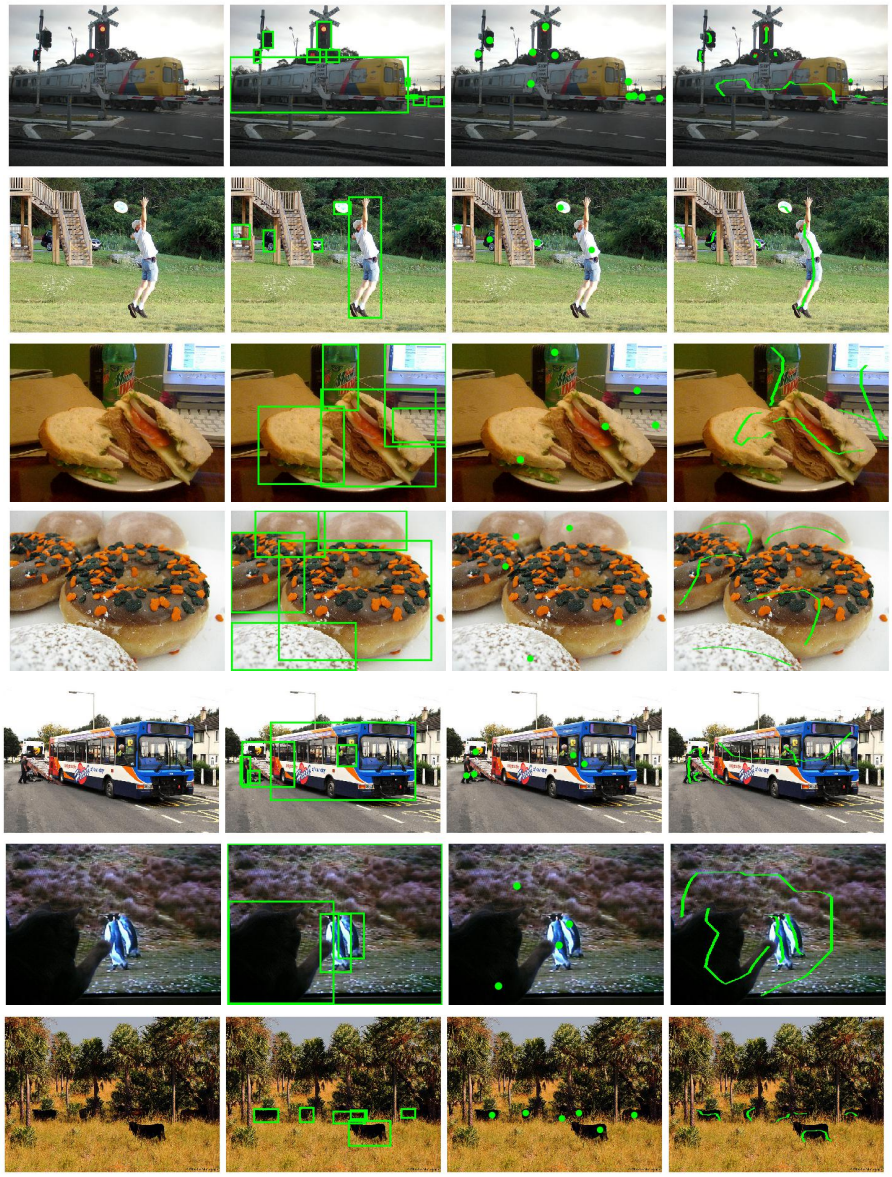

Fig. 10. Additional visualization of the ground-truth boxes and the simulated partial labels.

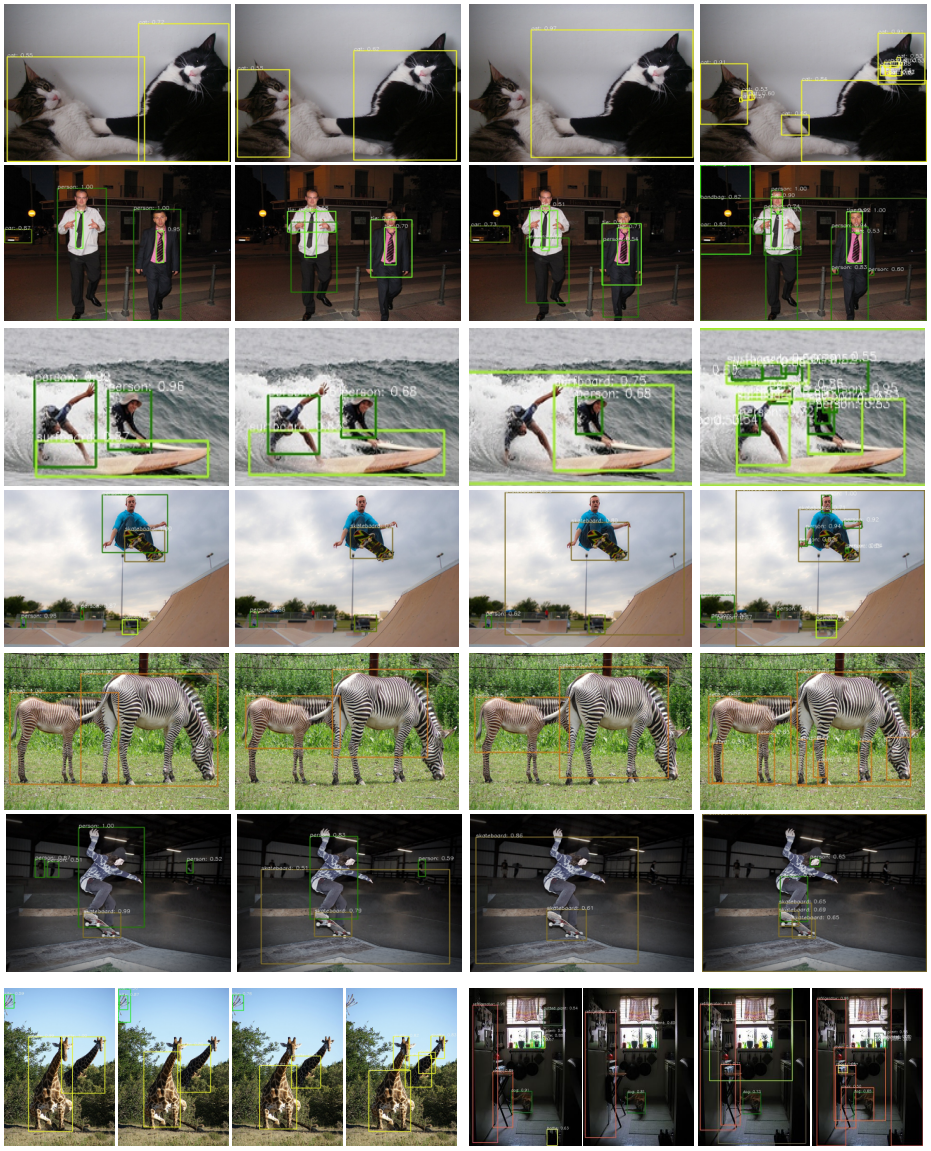

Fig. 11. Additional qualitative comparison of models trained with different labels on COCO (left to right: boxes, scribbles, points, tags).

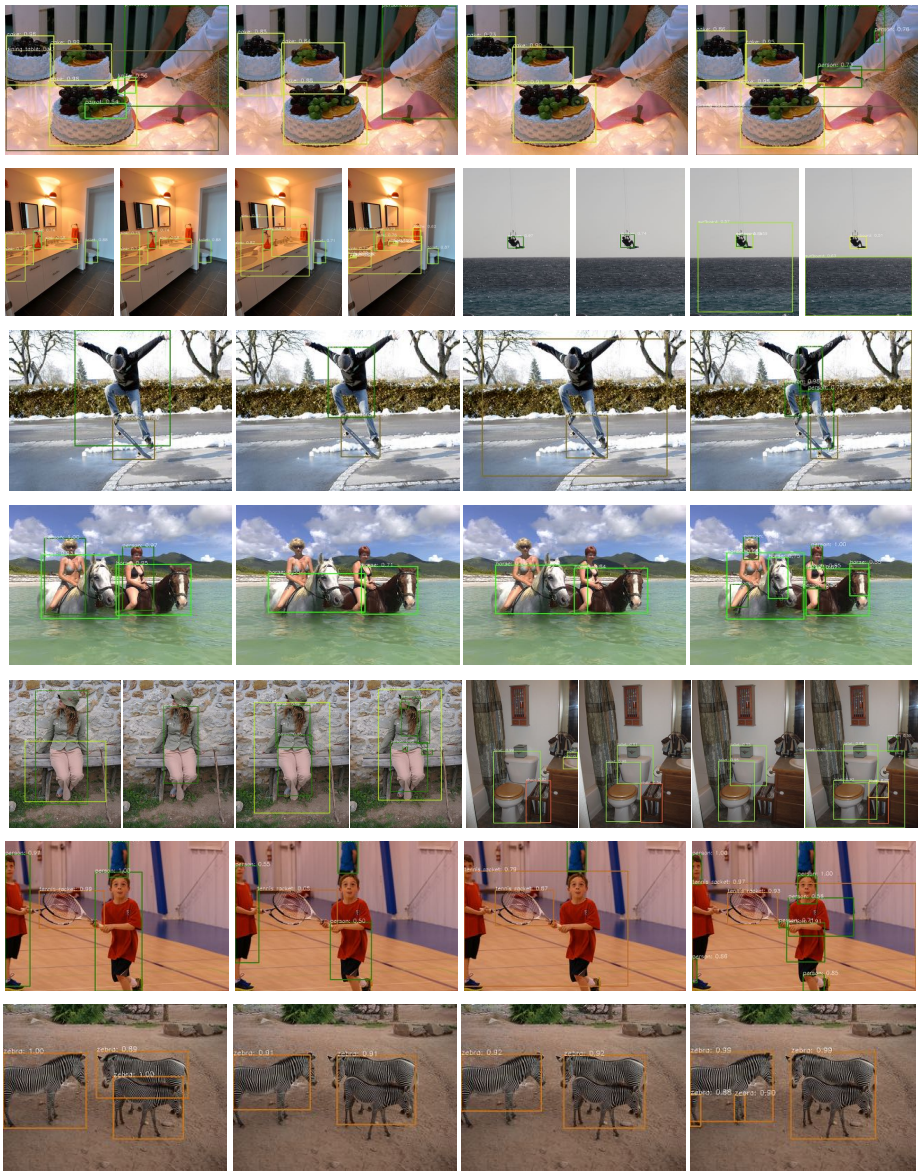

Fig. 12. Additional qualitative comparison of models trained with different labels on COCO (left to right: boxes, scribbles, points, tags).