A. Co-occurrence matrix

Figure 10 illustrates all class co-occurrence statistics for the ground truth objects in COCO `train2017`. Each entry represents the expected number of instances of the co-occurrent class given that there is at least one object from the observed class. If the co-occurrent and observed classes are the same, the entry represents how many co-occurrent instances of that class will be observed *in addition* to the observed one. Mathematically, for the set of classes $C = \{c_1, \ldots, c_k\}$, entry $(i, j) \in [k] \times [k]$ is computed as

$$\frac{\sum_{q \in S_i} |c_j(G_q^*)|}{|S_i|} - \mathbb{1}\{c_i = c_j\}, \tag{10}$$

where S_i is the set of images containing at least one object of class c_i, i.e., $S_i = \{q \in [n] \mid |c_i(G_q^*)| \geq 1\}$. Row i iterates over observed classes, column j iterates over co-occurrent classes, G_q^* is the set of ground truth bounding boxes for image $q \in [n]$, where n is the number of available images to compute the statistics (in this case, the number of images in `train2017`). $c(G_q^*) \subseteq G_q^*$ is the subset of bounding boxes in G_q^* with class $c \in C$, and $\mathbb{1}\{\cdot\}$ is the indicator function.

B. Additional ablations and results

Model comparison Table 7 compares improvements obtained by using a bidirectional model and self-attention. The base model is an unidirectional RNN with $n_r = 3$ stacked layers and a hidden state of size $n_h = 256$ trained with shuffling instance with probability 0.75. We compare the performance improvement of a bidirectional model and the addition of self-attention both with a GRU and a LSTM. To compare to a model that does not use RNNs, we replace the RNN with a fully-connected layer (Linear(85,128) + ReLU) followed by self-attention (using "general" attention from [24]) and the regressor (Linear(256,128) + ReLU + Linear(128,80) + ReLU + Linear(80,1) + Sigmoid).

The choice of LSTM or GRU has little impact on performance. GRU achieves higher performances with smaller models. Predictions made with an attention module or a bidirectional RNN conditions on the whole set of detections. The results using a linear layer with self-attention demonstrate the attention mechanism's ability to capture context with fewer parameters.

Class AP improvements In Table 8 we aggregate the improvements on the per-class AP for the tested baseline architectures on COCO `test-dev2017`. Our rescoring model produces consistent AP improvements for most classes, while few have a small decrease. The mean of each column is the improvement on the final AP metric for the model associated to that column. Faster R-CNN with a ResNet-50 backbone has the largest improvement

RNN	attention	bidirectional	# params	AP
baseline				42.1
Linear	✓		0.1 M	42.6
LSTM			1.4 M	42.6
LSTM		✓	3.9 M	42.8
LSTM	✓		1.5 M	42.6
LSTM	✓	✓	4.0 M	42.7
GRU			1.1 M	42.6
GRU		✓	2.9 M	42.8
GRU	✓		1.2 M	42.7
GRU	✓	✓	3.0 M	**42.8**

Table 7: Ablation study of model components comparison. 'Linear' replaces the RNN by a fully-connected layer.

C. Rescored examples

To systematically explore the results of rescoring, we compare, for each image, the vectors of confidences for the detections before and after rescoring. We sort images in decreasing order of the change in confidences, as measured by the cosine distance between the vectors of confidences before and after rescoring, i.e., for image $q \in [n]$,

$$d(v_q, v_q') = 1 - \frac{v_q^T v_q'}{||v_q||_2 ||v_q'||_2}, \tag{11}$$

where $v_q, v_q' \in \mathbb{R}^{|\hat{G}_q|}$ are the vectors of confidences before and after rescoring, respectively, and $\hat{G}_q$ is the set of detections being rescored. This analysis uses the detections produced by Cascade R-CNN with a ResNet-101 backbone on `val2017`.

We present the top 16 images according to this metric in two different ways. In Figure 11 we only consider images that have at most 4 detections (i.e., $q \in [n] \mid |\hat{G}_q| \leq 4\}$) as their detections and changes in confidence can be visualized clearly. In Figure 12, we consider all images but only show detections that have confidence above 0.2. An image is shown three times annotated with, left to right, predicted bounding boxes and their confidences before rescoring, predicted bounding boxes and their confidences after rescoring, and ground truth bounding boxes. The bounding box line width is proportional to its confidence. Images are ordered left to right, top to bottom.

In Figure 11, we see mostly successful suppressions: a rock classified as a sheep in an image with a zebra (left, row 1); duplicate tie detections (left, row 4 and right, row 6); duplicate toilet detections (left, row 2); duplicate train detections (left, row 8); duplicate kite detections (right, row 8); superimposed horse and zebra (right, row 2); duplicate bed detections (left, row 5); the moon classified as a frisbee

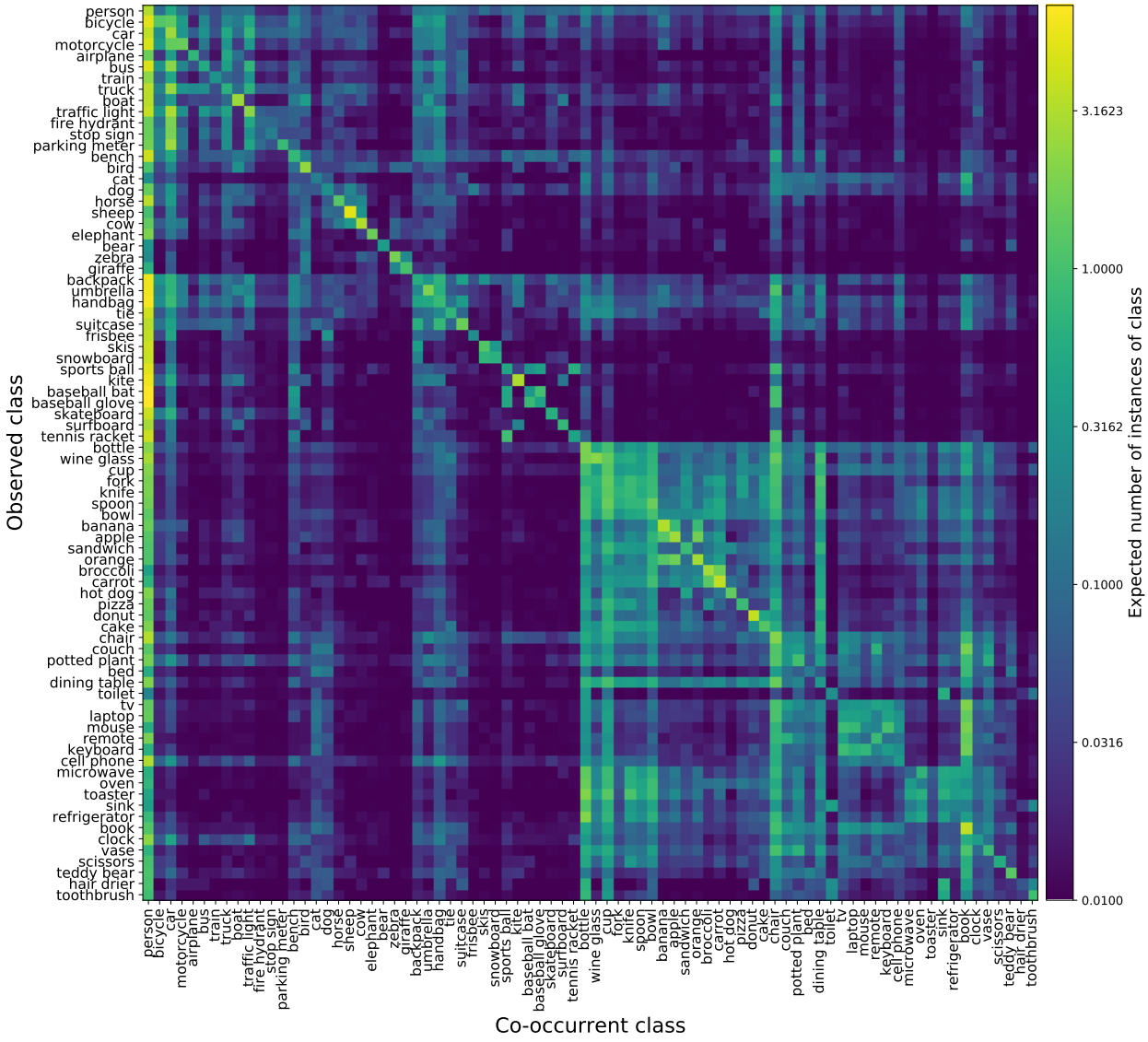

Figure 10: Co-occurrence matrix for COCO `train2017` annotations. See Equation 10 to see how these co-occurrences were computed.

(right, row 4); a sink and a toilet near a horse (right, row 5); bird and umbrella in the zebra's reflection (right, row 7).

In Figure 12, we have examples with many detections: either for small background objects (left, rows 2, 5 and 8; right, row 4), or multiple duplicate detections of skateboard (right, row 3), banana (right, row 6), and scissors (left, rows 4 and 7). While for most cases we have observed, the model suppresses detections, on the left, on row 3, the model has increased the confidence of its most central object (scissors). In this instance, all original confidences are low (smaller than 0.7) compared to what happens in most images where there is at least a detection which has more than 0.85 confidence.

The behavior of the model shown here can be explained from the point of view of AP computation — suppressing detections might be useful if we are not confident on their location or existence in the ground truth. This is frequently observed in images containing many (often small) objects of the same class (e.g., apples, bananas, cars, books, and people). The ground truth annotations often do not contain many of the instances in the image. For example, in Figure 12 (left, row 8), an airplane flies over a parking lot containing hundreds of cars and trucks, out of which only 15 are in the ground truth annotations. The set of detections contains many of these cars with medium confidence (most ranging from 0.3 to 0.7). After rescoring these detections

class	C-101	C-50	F-101	F-50	Mean
couch	+1.7	+1.4	+1.9	+2.5	+1.88
toaster	+3.2	+0.9	+1.2	+1.5	+1.70
frisbee	+1.4	+0.3	+1.4	+3.1	+1.55
cake	+1.1	+1.4	+1.3	+1.7	+1.38
pizza	+1.1	+1.0	+1.9	+1.3	+1.32
donut	+0.9	+1.1	+1.1	+1.8	+1.22
sandwich	+1.2	+1.2	+1.0	+1.1	+1.13
orange	+1.1	+1.3	+0.4	+1.7	+1.13
toilet	+0.6	+0.7	+1.4	+1.7	+1.10
bed	+0.4	+0.3	+1.7	+2.0	+1.10
refrigerator	+0.6	+0.8	+1.1	+1.9	+1.10
microwave	+1.4	+1.0	+0.7	+0.9	+1.00
vase	+1.0	+0.7	+1.4	+0.9	+1.00
hair drier	+0.1	+0.2	+2.9	+0.7	+0.98
laptop	+0.6	+1.0	+1.0	+0.8	+0.85
carrot	+1.0	+1.0	+0.5	+0.9	+0.85
mouse	+0.8	+0.8	+0.6	+1.2	+0.85
cow	+0.8	+0.8	+0.5	+1.2	+0.83
surfboard	+0.6	+0.9	+0.5	+1.3	+0.83
baseball glove	+0.8	+0.5	+0.9	+1.1	+0.83
snowboard	+0.4	+0.7	+0.6	+1.6	+0.83
cell phone	+0.8	+0.6	+0.9	+0.9	+0.80
dining table	+0.4	+0.3	+0.6	+1.9	+0.80
baseball bat	+1.4	+0.6	-0.2	+1.3	+0.78
cat	+0.2	+0.3	+0.8	+1.7	+0.75
fork	+0.6	+0.8	+0.3	+1.3	+0.75
backpack	+0.6	+0.5	+0.7	+1.0	+0.70
broccoli	+0.8	+0.4	+0.4	+1.2	+0.70
toothbrush	+0.6	+1.2	+0.4	+0.6	+0.70
bench	+0.5	+0.6	+0.8	+0.8	+0.68
suitcase	+0.4	+0.7	+0.3	+1.2	+0.65
oven	+0.3	+1.1	+0.3	+0.8	+0.63
remote	+0.7	+0.8	+0.3	+0.6	+0.60
kite	+0.5	+0.3	+0.4	+1.1	+0.58
apple	+1.3	+0.5	+0.6	-0.2	+0.55
knife	+0.6	+0.7	+0.3	+0.6	+0.55
teddy bear	+0.7	+1.1	-0.1	+0.5	+0.55
scissors	+0.3	+1.1	-0.3	+1.0	+0.53
skis	+0.4	+0.6	+0.1	+0.9	+0.50
hot dog	+1.6	+0.3	+0.3	-0.2	+0.50
truck	+0.3	+0.2	-0.1	+1.4	+0.45
horse	+0.4	+0.4	+0.0	+1.0	+0.45
umbrella	+0.4	+0.4	+0.5	+0.5	+0.45
skateboard	+0.4	+0.2	+0.3	+0.8	+0.43
spoon	+0.7	+0.2	+0.4	+0.4	+0.43
chair	+0.4	+0.4	+0.3	+0.5	+0.40
parking meter	+0.4	+0.8	-0.6	+1.0	+0.40
train	+0.1	+0.5	+0.0	+1.0	+0.40
cup	+0.3	+0.4	+0.3	+0.5	+0.38
bear	+0.6	+0.4	+0.0	+0.4	+0.35
tv	+0.4	+0.1	+0.1	+0.7	+0.33
car	+0.3	+0.1	+0.3	+0.5	+0.30
bowl	+0.3	+0.4	+0.3	+0.1	+0.28
fire hydrant	+0.1	+0.5	+0.0	+0.5	+0.28
airplane	+0.1	+0.3	-0.1	+0.6	+0.23
dog	+0.8	+0.2	-0.1	-0.1	+0.20
sports ball	+0.3	+0.3	-0.1	+0.2	+0.18
sheep	+0.2	+0.1	+0.2	+0.2	+0.18
keyboard	+0.0	+0.6	-0.3	+0.3	+0.15
handbag	+0.1	+0.2	+0.1	+0.0	+0.10
bottle	+0.1	+0.0	+0.4	-0.1	+0.10
tie	+0.6	+0.0	-0.1	-0.2	+0.08
banana	-0.3	+0.4	-0.2	+0.3	+0.05
stop sign	+0.1	+0.3	-0.3	+0.1	+0.05
book	+0.0	+0.0	+0.1	+0.0	+0.03
potted plant	+0.1	-0.1	-0.4	+0.5	+0.03
bus	-0.1	-0.1	-0.2	+0.5	+0.03
boat	+0.0	-0.2	+0.1	+0.1	+0.00
bird	+0.1	-0.3	-0.1	+0.3	+0.00
motorcycle	+0.1	-0.1	-0.2	+0.1	-0.03
tennis racket	-0.2	+0.1	-0.5	+0.4	-0.05
traffic light	+0.0	+0.0	-0.3	-0.1	-0.10
zebra	-0.2	-0.2	-0.2	+0.2	-0.10
sink	+0.2	+0.2	-0.4	-0.7	-0.18
bicycle	-0.1	-0.3	-0.4	+0.0	-0.20
person	-0.3	-0.3	-0.2	+0.0	-0.20
wine glass	-0.4	-0.2	-0.2	+0.0	-0.20
giraffe	-0.1	-0.4	-0.7	-0.1	-0.33
elephant	-0.3	-0.3	-0.8	-0.4	-0.45
clock	-0.3	-0.5	-0.8	-0.3	-0.48
Mean	+0.5	+0.5	+0.4	+0.7	+0.53

Table 8: Per-class AP improvement on `test-dev2017`. **C:** Cascade R-CNN, **F:** Faster R-CNN, **101:** ResNet-101, **50:** ResNet-50.

have been mostly suppressed (lower than 0.2 confidence).

The reason for this omission in the ground truth annotations is two-fold: perceptually, the exact number of cars is not important and annotating this many cars would be tedious. Due to this, suppressing them during rescoring should lead to improvements as most of these would be considered false positives. The same motivation is valid for the images with books (left, row 5) and bananas (right, row 6). Our approach successfully captures the risk associated with detections being false positives.

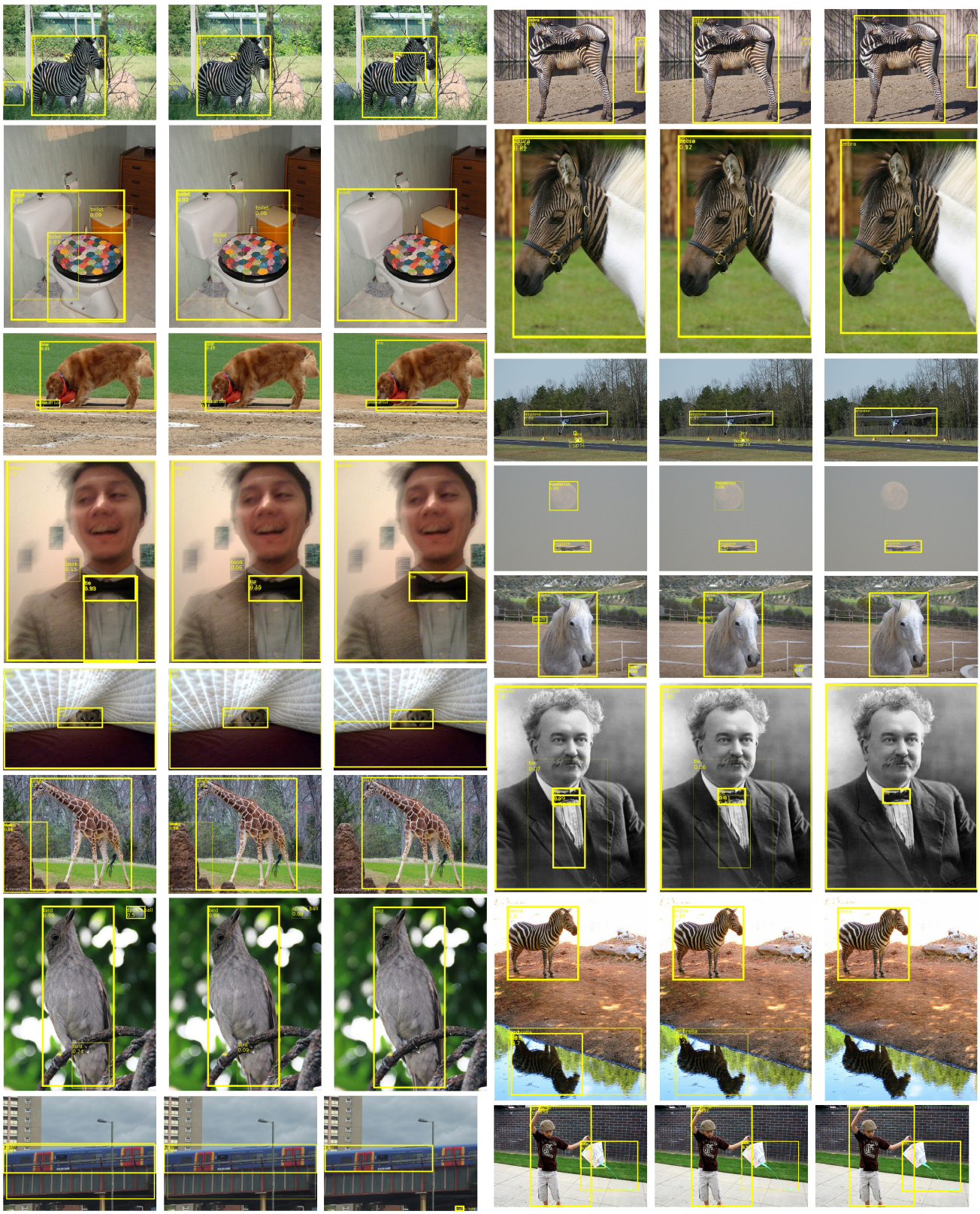

Figure 11: Top 16 images with at most 4 detections which had the largest change in confidences as a result of rescoring. For each image, left to right: detections with initial confidences, detections with rescored confidences, and ground truth bounding boxes.

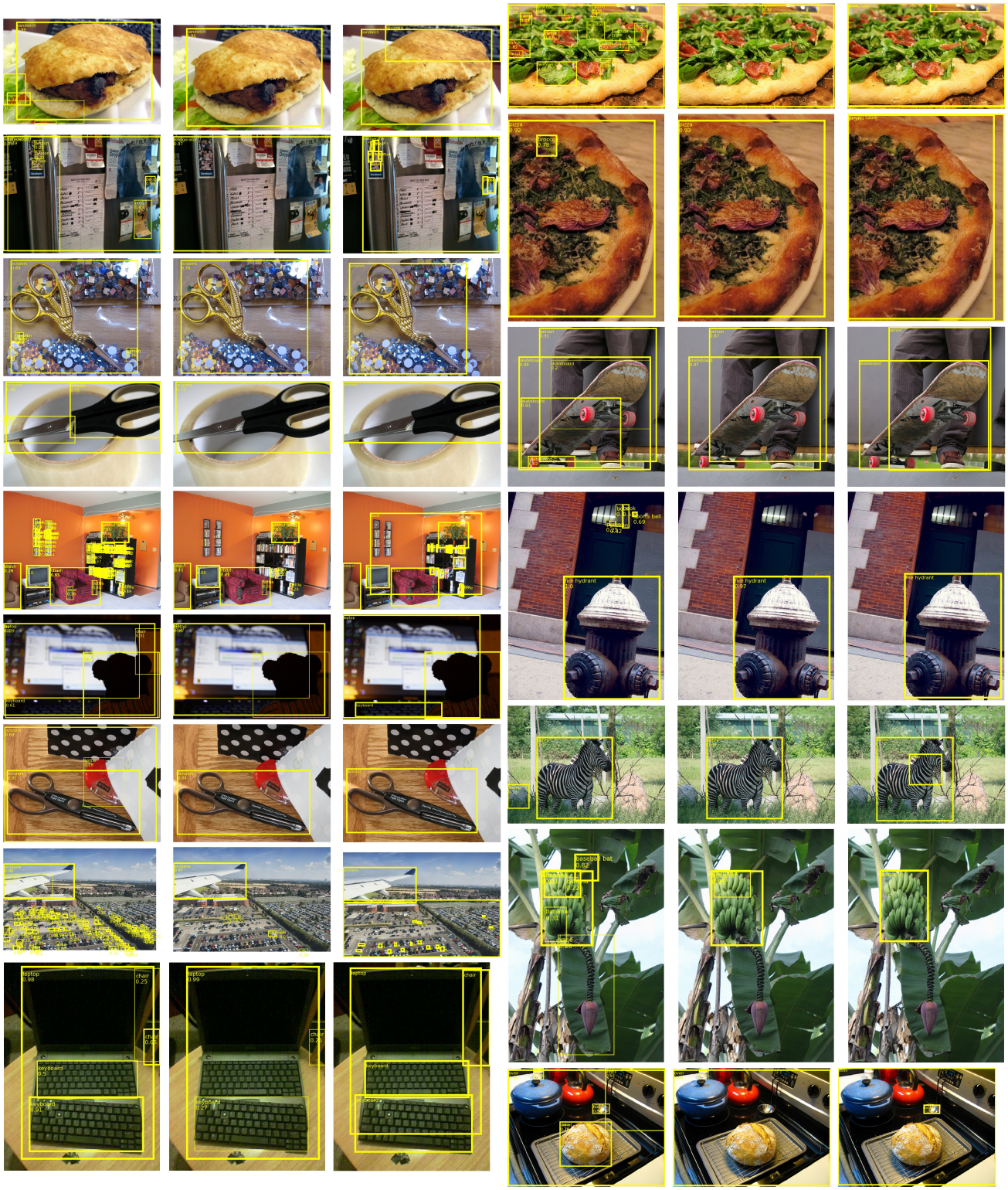

Figure 12: Top 16 images on which had the largest change in confidences as a result of rescoring. Detections with confidence lower than 0.2 are omitted. For each image, left to right: detections with initial confidences, detections with rescored confidences, and ground truth bounding boxes.