A. More Visualization of COCO-O

We present the class distribution of COCO-O in Figure 10. The number of instances per image in COCO-O are shown in Figure 8. COCO-O have around 5% more iconic images compared with COCO validation set. To study the impact, we decrease the proportion by randomly deleting iconic images in COCO-O, and re-run the evaluation. The ablation in Table 5 suggests more iconic images in COCO-O will not bring great change in result.

Figure 8: Number of annotated instances per image for COCO and COCO-O.

<table>
<thead>
<tr>
<th>Detectors</th>
<th>Percent of iconic images</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faster R-CNN</td>
<td>24% 22% 20% 18%</td>
</tr>
<tr>
<td>RetinaNet</td>
<td>16.4 16.3 16.5 16.4</td>
</tr>
</tbody>
</table>

Table 5: Impact of the more iconic images in COCO-O.

B. Detailed Results of Compared Detectors

Table 6 presents the detailed results of compared detectors in Figure 1. For compared SOTA detectors, we introduce their details in Table 7.

C. Comparison of Results on COCO-C and COCO-O

Figure 9 suggests COCO-C has a positive growth trend with clean performance, while such trend in COCO-O is more weak. It demonstrates COCO-O measures the robustness aspect out of standard COCO test set.

D. Visualization of Detection Results

We adopt five detectors in Table 6: Faster R-CNN, DETR, YOLOX-S, PVTv2-B2, GLIP-T, and visualize their results on our COCO-O in Figure 12. The score threshold is set as 0.5, which means that only bounding box with classification confidence over 0.5 will be visualized. Among the compared detectors, Faster R-CNN has failed to localize most objects and leads to a low recall rate. Detectors such as DETR, YOLOX-S recall some wrong bounding box with misclassified category or meaningless positions. GLIP-T outputs more accurate detection results. It can even recognize some abstract objects, such as the person with umbrella (the third line in figure).

Figure 9: The results of COCO-C and COCO-O on different detectors.

E. Visualization of Bad Cases

Among COCO-O images, we have found some bad cases which make most detectors fail. They have been shown Figure 11. Although the presented images contain common objects that can be easily recognized by humans, machines are hard to understand them. A vivid example is the anime characters in “The Powerpuff Girls” or “Dora the Explorer”, which should be detected as person while machine looks unaware of such art form of expression. In addition, some traditional factors, e.g., perspective change, severe haze or blur, very poor lighting conditions, will also increase the difficulty of detection task.

F. Usage of COCO-O

- Example code for testing with MMDetection:

```bash
$ git clone https://github.com/open-mmlab/mmdetection.git
$ cd mmdetection
$ mkdir data
$ ln -s /path/to/coco-o/sketch data/coco
$ python tools/test.py ${CONFIG_FILE} \
${CHECKPOINT_FILE} --eval bbox
```

- Example code for testing with Detectron2:

```bash
$ git clone https://github.com/facebookresearch/detectron2.git
$ cd detectron2
$ mkdir datasets
$ ln -s /path/to/coco-o/sketch datasets/coco
$ ./tools/lazyconfig_train_net.py \
--config-file ${CONFIG_FILE} \
train.init_checkpoint=${CHECKPOINT_FILE} \
--eval-only
```
Figure 10: Class distribution of our COCO-O.

<table>
<thead>
<tr>
<th>Detection Methods</th>
<th>BackBone</th>
<th>COCO mAP</th>
<th>COCO rank</th>
<th>COCO-O mAP</th>
<th>COCO-O rank</th>
<th>Effective Rank Diff.</th>
<th>Use Mask</th>
<th>Additional Data</th>
<th>FPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mask R-CNN</td>
<td>ViT-B</td>
<td>51.6</td>
<td>1</td>
<td>24.3</td>
<td>8</td>
<td>+1.06</td>
<td>✓</td>
<td>x</td>
<td>12.7</td>
</tr>
<tr>
<td>YOLOv7</td>
<td>-</td>
<td>51.2</td>
<td>2</td>
<td>29.0</td>
<td>3</td>
<td>+5.99</td>
<td>-1</td>
<td>x</td>
<td>161.0</td>
</tr>
<tr>
<td>FIBER-B</td>
<td>Swin-B</td>
<td>49.5</td>
<td>3</td>
<td>33.7</td>
<td>1</td>
<td>+11.43</td>
<td>+2</td>
<td>✓</td>
<td>3.1</td>
</tr>
<tr>
<td>Mask R-CNN</td>
<td>Swin-S</td>
<td>48.5</td>
<td>4</td>
<td>25.5</td>
<td>6</td>
<td>+3.71</td>
<td>-2</td>
<td>✓</td>
<td>10.1</td>
</tr>
<tr>
<td>Mask R-CNN</td>
<td>MViTV2-T</td>
<td>48.3</td>
<td>5</td>
<td>25.7</td>
<td>5</td>
<td>+3.98</td>
<td>0</td>
<td>✓</td>
<td>14.6</td>
</tr>
<tr>
<td>GLIP-T</td>
<td>Swin-T</td>
<td>46.6</td>
<td>6</td>
<td>29.1</td>
<td>2</td>
<td>+8.11</td>
<td>+4</td>
<td>✓</td>
<td>5.5</td>
</tr>
<tr>
<td>Mask R-CNN</td>
<td>Uniformer-S</td>
<td>45.6</td>
<td>7</td>
<td>26.2</td>
<td>4</td>
<td>+5.65</td>
<td>+3</td>
<td>✓</td>
<td>15.6</td>
</tr>
<tr>
<td>Mask R-CNN</td>
<td>PVTv2-B2</td>
<td>45.3</td>
<td>8</td>
<td>24.9</td>
<td>7</td>
<td>+4.52</td>
<td>+1</td>
<td>✓</td>
<td>15.8</td>
</tr>
<tr>
<td>Deformable-DETR</td>
<td>ResNet-50</td>
<td>44.5</td>
<td>9</td>
<td>18.5</td>
<td>13</td>
<td>-1.49</td>
<td>-4</td>
<td>x</td>
<td>15.0</td>
</tr>
<tr>
<td>DyHead</td>
<td>ResNet-50</td>
<td>42.5</td>
<td>10</td>
<td>19.3</td>
<td>10</td>
<td>+0.16</td>
<td>0</td>
<td>x</td>
<td>13.2</td>
</tr>
<tr>
<td>HTC</td>
<td>ResNet-50</td>
<td>42.3</td>
<td>11</td>
<td>19.1</td>
<td>11</td>
<td>+0.08</td>
<td>0</td>
<td>✓</td>
<td>5.8</td>
</tr>
<tr>
<td>DETR</td>
<td>ResNet-50</td>
<td>42.0</td>
<td>12</td>
<td>17.1</td>
<td>17</td>
<td>-1.82</td>
<td>-5</td>
<td>x</td>
<td>27.8</td>
</tr>
<tr>
<td>VarifocalNet</td>
<td>ResNet-50</td>
<td>41.6</td>
<td>13</td>
<td>18.8</td>
<td>12</td>
<td>+0.05</td>
<td>+1</td>
<td>x</td>
<td>19.4</td>
</tr>
<tr>
<td>YOLOX-S</td>
<td>ResNet-50</td>
<td>40.5</td>
<td>14</td>
<td>20.6</td>
<td>9</td>
<td>+2.48</td>
<td>+5</td>
<td>x</td>
<td>102.0</td>
</tr>
<tr>
<td>Cascade R-CNN</td>
<td>ResNet-50</td>
<td>40.3</td>
<td>14</td>
<td>18.2</td>
<td>14</td>
<td>+0.02</td>
<td>0</td>
<td>x</td>
<td>16.1</td>
</tr>
<tr>
<td>ATSS</td>
<td>ResNet-50</td>
<td>39.4</td>
<td>16</td>
<td>16.8</td>
<td>19</td>
<td>-0.91</td>
<td>-3</td>
<td>x</td>
<td>19.7</td>
</tr>
<tr>
<td>Mask R-CNN</td>
<td>ResNet-50</td>
<td>38.2</td>
<td>17</td>
<td>17.1</td>
<td>17</td>
<td>-0.11</td>
<td>0</td>
<td>✓</td>
<td>16.1</td>
</tr>
<tr>
<td>Sparse R-CNN</td>
<td>ResNet-50</td>
<td>37.9</td>
<td>18</td>
<td>17.5</td>
<td>16</td>
<td>+0.41</td>
<td>+2</td>
<td>✓</td>
<td>18.3</td>
</tr>
<tr>
<td>YOLOv5s</td>
<td>-</td>
<td>37.4</td>
<td>19</td>
<td>17.8</td>
<td>15</td>
<td>+1.00</td>
<td>+4</td>
<td>x</td>
<td>156.3</td>
</tr>
<tr>
<td>Faster R-CNN</td>
<td>ResNet-50</td>
<td>37.4</td>
<td>19</td>
<td>16.4</td>
<td>22</td>
<td>-0.41</td>
<td>-3</td>
<td>x</td>
<td>21.4</td>
</tr>
<tr>
<td>FCOS</td>
<td>ResNet-50</td>
<td>36.6</td>
<td>21</td>
<td>16.7</td>
<td>20</td>
<td>+0.25</td>
<td>+1</td>
<td>x</td>
<td>22.7</td>
</tr>
<tr>
<td>RetinaNet</td>
<td>ResNet-50</td>
<td>36.5</td>
<td>22</td>
<td>16.6</td>
<td>21</td>
<td>+0.18</td>
<td>+1</td>
<td>x</td>
<td>19.0</td>
</tr>
<tr>
<td>YOLOS-S</td>
<td>-</td>
<td>36.1</td>
<td>23</td>
<td>15.5</td>
<td>23</td>
<td>-0.80</td>
<td>0</td>
<td>x</td>
<td>5.7</td>
</tr>
<tr>
<td>YOLOv3</td>
<td>DarkNet-53</td>
<td>33.7</td>
<td>24</td>
<td>14.8</td>
<td>24</td>
<td>-0.37</td>
<td>0</td>
<td>x</td>
<td>48.1</td>
</tr>
<tr>
<td>SSD</td>
<td>VGG16</td>
<td>29.5</td>
<td>25</td>
<td>13.6</td>
<td>25</td>
<td>+0.36</td>
<td>0</td>
<td>x</td>
<td>30.7</td>
</tr>
</tbody>
</table>

Table 6: Detailed results of the compared detectors in Figure 1.
<table>
<thead>
<tr>
<th>Detection Methods</th>
<th>BackBone</th>
<th>COCO mAP</th>
<th>COCO rank</th>
<th>COCO-O mAP</th>
<th>COCO-O rank</th>
<th>Effective Robustness</th>
<th>Rank Diff.</th>
<th>Use Mask</th>
<th>Additional Data</th>
<th>FPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cascade R-CNN</td>
<td>ResNet-101</td>
<td>42.2 42.2</td>
<td>7 7</td>
<td>43.2 43.2</td>
<td>1 1</td>
<td>+1.0 1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Cascade R-CNN</td>
<td>ResNet-101</td>
<td>42.2 42.2</td>
<td>7 7</td>
<td>43.2 43.2</td>
<td>1 1</td>
<td>+1.0 1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 7: Detailed results of the compared SOTA detectors in Figure 6.
Figure 11: Hard examples in COCO-O which make detection failed.
Figure 12: Visualization of detection results on sampled COCO-O images.
G. Queries for Collecting COCO-O

We show the query keywords used for collect sketch images in COCO-O. Similarly, for other test domains, we create queries by just replacing with the corresponding domain keywords.

sketch person
sketch bicycle
sketch car
sketch motorcycle
sketch airplane
sketch bus
sketch train
sketch truck
sketch boat
sketch traffic light
sketch fire hydrant
sketch stop sign
sketch parking meter
sketch bench
sketch bird
sketch cat
sketch dog
sketch horse
sketch sheep
sketch cow
sketch elephant
sketch bear
sketch zebra
sketch giraffe
sketch backpack
sketch umbrella
sketch handbag
sketch tie
sketch suitcase
sketch frisbee
sketch skis
sketch snowboard
sketch sports ball
sketch kite
sketch baseball bat
sketch baseball glove
sketch skateboard
sketch surfboard
sketch tennis racket
sketch bottle
sketch wine glass
sketch cup
sketch fork
sketch knife
sketch spoon
sketch bowl
sketch banana
sketch apple
sketch sandwich
sketch orange
sketch broccoli
sketch carrot
sketch hot dog
sketch pizza
sketch donut
sketch cake
sketch chair
sketch couch
sketch potted plant
sketch bed
sketch dining table
sketch toilet
sketch tv
sketch laptop
sketch mouse
sketch remote
sketch keyboard
sketch cell phone
sketch microwave
sketch oven
sketch toaster
sketch sink
sketch refrigerator
sketch book
sketch clock
sketch vase
sketch scissors
sketch teddy bear
sketch hair drier
sketch toothbrush
sketch person chair
sketch person car
sketch person handbag
sketch person dining table
sketch chair dining table
sketch person cup
sketch person backpack
sketch person bottle
sketch cup dining table
sketch person cell phone
sketch person tennis racket
sketch person bench
sketch person truck
sketch person sports ball
sketch car truck
sketch bowl dining table
sketch person umbrella
sketch person bus
sketch person surfboard
sketch person tie
sketch person bowl
sketch cup chair
sketch bottle cup
sketch person motorcycle
sketch person skateboard
sketch person bicycle
sketch bottle dining table
sketch person skis
sketch knife dining table
sketch cup bowl
sketch car traffic light
sketch person traffic light
sketch fork dining table
sketch car bus
sketch person knife
sketch bottle chair
sketch person book
sketch person baseball glove
sketch bottle bowl
sketch person remote
sketch person baseball bat
sketch person pizza
sketch person couch
sketch chair couch
sketch chair book
sketch chair book
sketch person horse
sketch spoon bowl