A. Human Study on Part-level Object Detection

Due to the lack of annotation, a quantitative evaluation of part-level object detection is infeasible. Nonetheless, we measure the real-world usability of our sketch-enabled object detection framework using Mean Opinion Score (MOS) by asking 10 people to draw 20 part-level sketches and rate from 1 to 5 (bad $\rightarrow$ excellent) based on their opinion of how closely the queried object part was detected. Accordingly, we obtain a MOS (mean $\pm$ variance of 200 responses) of 3.67 $\pm$ 0.6.

B. Preliminary Study on Occluded Objects

In addition to category-level, fine-grained, and part-level object detection, we further qualitatively test the generalisability of the system to detect occluded objects as:

While we show some successful, failed, and partially detected cases, future works can further investigate the role of sketch and foundation models like CLIP [4] for occluded object detection.

C. Relation to Open World setup

In open world setup, a model trained on $C$ known classes can recognise the unknown class and update the base model via incremental learning [1, 3]. Our method already works in open world setup as it detects in zero-shot, open-vocab setup, i.e., it works regardless of whether the query sketch is in the train set or not.

D. Detection across Different Poses

Our object detection has multiple setups: (i) for category-level OD, the sketch of object $O_1$ (“zebra”) in image $I_1$ will detect the same object $O_1$ in a different image $I_2$ if it has the same pose, e.g., detect only “zebras sitting down” amongst a herd of “zebras”. Figure below shows qualitative results for clarity.

E. Additional Ablation Study

(i) Varying prompt length $P = \{1, 3, 5\}$ in $\{v_s, v_p\} \in \mathbb{R}^{P \times 768}$ changes $AP_5$ to 16.5, 17.1, and 15.9 on Sketchy-COCO [2] respectively. (ii) Replacing CLIP with VGG-based sketch encoder $F_S$ sharply drops $AP_5$ to 9.1 (iii) Increasing tiling from $n \in [1, 7]$ to $n \in [1, 17]$ reduces $AP_5$ to 11.3 due to high occlusion ($n \rightarrow 17$).

F. Robustness to Tiling

To test robustness, we generate occluded photos by randomly masking (10%, 30%, 50%) of GT object boundaries with zero pixel values and measure the respective drop in accuracy ($AP_5$) on [2]. Performance drop being less with tiling for E-WSSDDN (by $\{1.7, 3.4, 5.7\}$) or our method (by $\{1.6, 3.3, 5.4\}$), than without tiling in WSDDN (by $\{3.1, 5.2, 7.5\}$) verifies robustness due to tiling on object detection.

G. Clarification on CutMix [5] vs. our Tiling

(i) Our novelty lies in adapting well-known modules (CLIP, SBIR) to train an object detector from only object-level sketch-photo pairs (each photo has only one object) without any bounding-box annotations. (ii) Despite sharing a common technical implementation, CutMix [5] is a data augmentation tool that typically replaces a patch in one existing scene-photo with that from another. Contrarily, tiling is a data synthesis tool that combines multiple object-level photos in the SBIR dataset to newly create a scene photo for subsequent training.
Figure A. Additional qualitative results for fine-grained and part-level object detection on SketchyCOCO. Note both the Blue and Yellow boxes are network predictions and not ground truth. The Blue boxes are predictions from the network prior to using Non-maximum suppression (NMS) with the confidence score of the predicted box $\omega_k \geq 0.7$. The Yellow boxes are the resulting predictions after applying NMS with IoU $\geq 0.3$.

References


