Large-scale interactive object segmentation with human annotators, supplementary material

Rodrigo Benenson Stefan Popov Vittorio Ferrari

Google Research

{benenson, spopov, vittoferrari}@google.com

1. Content

This supplementary material provides additional qualitative results and more detailed quantitative results of the experiments described in the main paper. The supplementary material follows the structure of the main paper, and provides supplementary results section by section.

2. Simulations

We provide here more details about the simulations reported in section 3 of the main paper.

2.1. Models training details

Here the details for the M_b and M_{b+c} training mentioned in section 3.1 of the main paper.

Both M_b and M_{b+c} use the same architecture and training procedure. We train Deeplabv2 ResNet101 [2] for a per-pixel binary classification (instance foreground/background). We start with a model pre-trained for ImageNet classification, use a batch size 1, and train for 10k steps with learning rate linearly decreasing from $5 \cdot 10^{-4}$ down to $5 \cdot 10^{-5}$. Training over 20 GPUs takes ~8 hours.

2.2. Boundary or region clicks?

We provide here additional figures to support the results discussed in section 3.2.2 of the main paper.

Figure 1 illustrates corrective clicks on the object boundary (as considered in [7, 6, 3, 1]) and corrective clicks inside the error regions (as considered in [8, 5, 4]).

Figure 3 shows the quantitative results of the 3 × 3 simulation. Both type of clicks bring clear improvements in every round; however region clicks reach higher mIoU values faster. This trend is consistent across different type of input encoding and number of clicks/rounds.

Result. Figure 3 shows the simulation results. Both type of clicks see clear improvements round after round, after three rounds, region clicks reach 80% mIoU while boundary clicks reach only 77% mIoU. This trend is consistent across different type of input encoding and number of clicks/rounds.

See main paper for details of the experimental setup and further discussions.

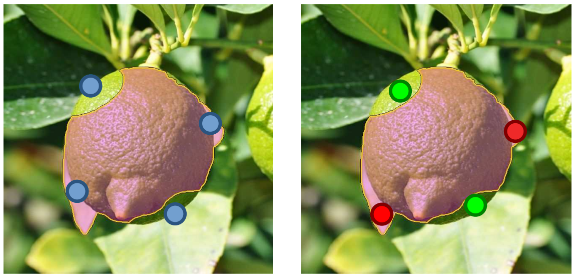
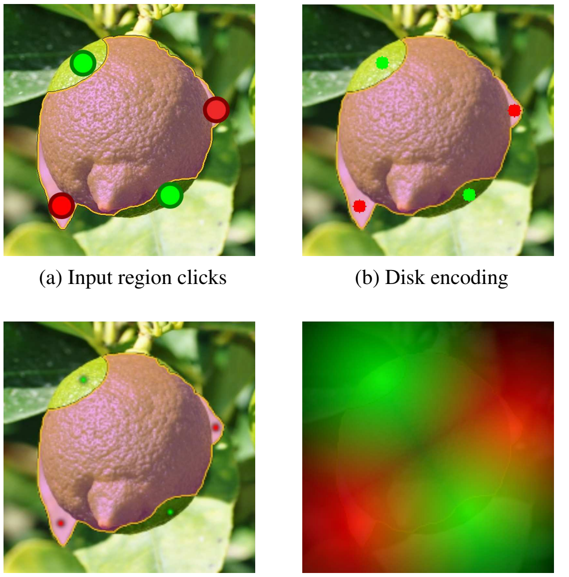


Figure 1: Example of two types of corrective clicks: boundary clicks (left) versus region clicks (right). See section 2.2.



(a) Input region clicks

(b) Disk encoding

(c) Gaussian encoding

(d) Distance transform encoding

Figure 2: (a) Illustrates the user clicks. (b,c,d) Show different encodings for region clicks. Each encoding uses two channels encoded in green and red colours. See section 2.4.

2.3. Annotation noise

We provide here additional figures to support the results discussed in section 3.2.3 of the main paper.

Figure 4 shows the mIoU reached at the end of 3 × 3 simulations; when considering different level of click noise and the minimum region size considered by the simulated annotator. We observe more than 5 pp mIoU fluctuations

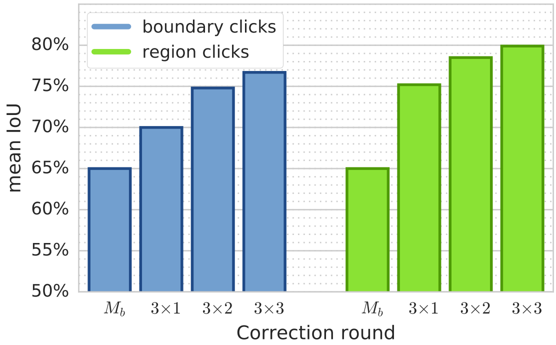


Figure 3: Boundary click versus region clicks simulations. Three rounds of three clicks. M_b indicates the masks obtained with zero clicks (bounding box only). Region clicks reach higher quality masks faster than boundary clicks. See section 2.2.

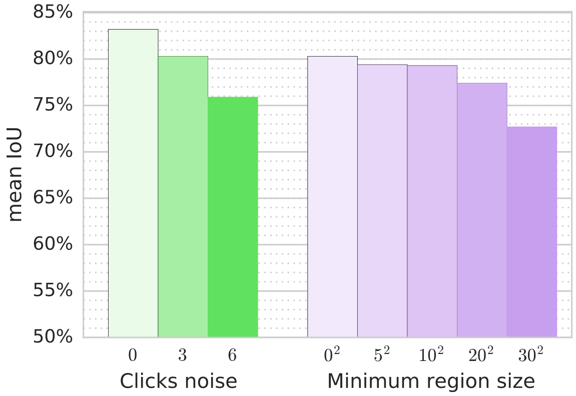


Figure 4: Effect of different annotation noise parameters. Mean IoU reported after at the end of 3 × 3 simulations. We observe more than 5 pp mIoU fluctuations depending on how much noise is assumed. See section 2.3.

depending on how much annotator noise is assumed. See main paper for details of the experimental setup and further discussions.

2.4. Clicks encoding

Figure 2 illustrates the different types of encodings considered in section 3.2.4 (and figure 3) of the main paper.

2.5. Class-agnostic vs class-specific

We describe here the experiments used to support the claims of section 3.2.6 in the main paper.

To understand the effect of class-agnostic versus class-specific M_b and M_{b+c} models we consider an extreme case: the class "giraffe". Since the ADE20k dataset covers mostly indoor and urban scenes, it contains few animals. The $\text{ADE}_\text{L}^\text{train}$ set contains zero giraffes and $< 2\%$ of four-legged animals instances (the closest are 149 cows, out of 0.4M instances). We can thus consider as if $\text{ADE}_\text{L}^\text{train}$ contains no giraffe nor similar instances in the set. In contrast COCO_Lcontains 10% of animals including 10k+ instances of giraffe-like classes like: horse, zebra, sheep, etc.

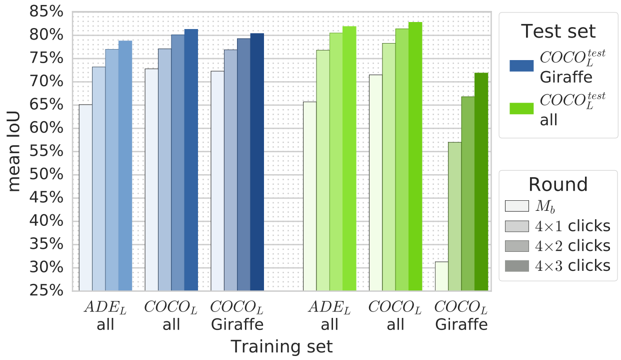


Figure 5: Mask quality after 3 rounds of (up to) 4 simulated clicks; when using different training sets for M_b and M_{b+c} . We evaluate either over giraffes only (left side), or over all COCO classes (right side). We see that after the annotations, there are no significant differences between the class-agnostic or the class-specific models. See section 2.5.

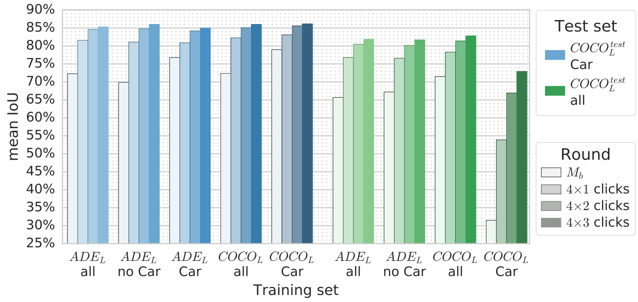


Figure 6: Mask quality after 3 rounds of (up to) 4 simulated clicks; when using different training sets for M_b and M_{b+c} . We evaluate either over car only (left side), or over all COCO classes (right side). We see that after the annotations, there are no significant differences between the class-agnostic or the class-specific models. See section 2.5.

Result. Figure 5 shows result when training models on $\text{ADE}_\text{L}^\text{train}$, $\text{COCO}_\text{L}^\text{train}$, or $\text{COCO}_\text{L}^\text{train}$ giraffes only, and evaluating over either $\text{COCO}_\text{L}^\text{test}$ giraffes (left side) or $\text{COCO}_\text{L}^\text{test}$ all classes (right side). As expected using the in-domain $\text{COCO}_\text{L}^\text{train}$ (that includes giraffes) provides better results than $\text{ADE}_\text{L}^\text{train}$ that does not include giraffes (both from M_b and after 4×3 rounds). However, after annotations the gap is rather small: starting at 65% versus 73% mIoU and ending at 79% versus 81% mIoU. Using all COCO classes, all COCO classes except giraffe (not in the plot), or only giraffe has a negligible effect and all end around $\sim 81\%$ mIoU. Similarly when evaluating over all $\text{COCO}_\text{L}^\text{test}$ classes we see only a minor difference between using $\text{ADE}_\text{L}^\text{train}$ versus $\text{COCO}_\text{L}^\text{train}$ (82% versus 83% mIoU).

When using the giraffe-specific model over all COCO classes (an extreme case of domain-shift), as expected the M_b output quality is dismal (worse than the M_b baselines). However, as soon as one or two rounds of annotations are done; the resulting masks do a significant quality jump. This indicates that even when trained for a single class, the M_{b+c} model learns to account for the input clicks to adapt the output masks well outside of its training domain.
Figure 6 shows results when considering "car" instead of "giraffes". Overall we see the same trends.

From these results we conclude that there is no need to have class-specific models. Instead a model trained with a large number of instances covering diverse classes performs essentially just as well.

3. Large-scale annotation campaign

3.1. Need for annotation policies

This section details the small-scale experiment mentioned in section 4.3 of the main paper.

To understand the need for annotation policies we ran an small experiment with 30 novice annotators. After explaining and demonstrating the task, we exposed them to a sequence of 12 constructed questions presenting common but non-trivial cases. The annotators answer the questions individually without knowing each other's answers.

The sequence of questions was structured as a series of triplets where the first instance is a baseline to confirm the annotators understands the task, while the second and third show two variants of the same non-trivial case. We are interested in observing how the annotators react across these last two.

Figures 12 and 13 shows such question triplets together with the annotators response. We can see that annotators are inconsistent even across consecutive instances, and that when they diverge there is almost a $^{50}/_{50}$ split (majority voting would not help). This anecdotal evidence supports the need for well-defined policies in order to obtain consistent annotator answers over each instance and across instances of the same class.

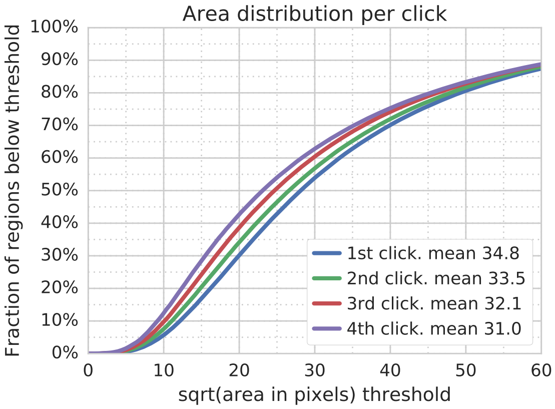


Figure 7: Cumulative distribution of the area of the clicked error region as a function of the click order. We observe the initial clicks aim larger region than the later ones. See section 4.2.

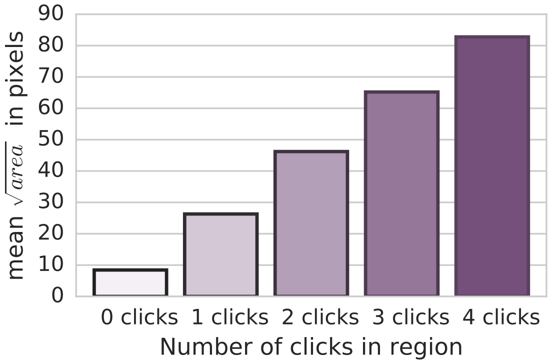


Figure 8: Average $\sqrt{\text{area}}$ of error regions with n clicks. We observe that the smallest regions ($< 10^2$ pixels) are left without clicks, and that the areas of the clicked region grows almost linearly with the number of clicks (at about ~22^2 pixels per click). See section 4.2.

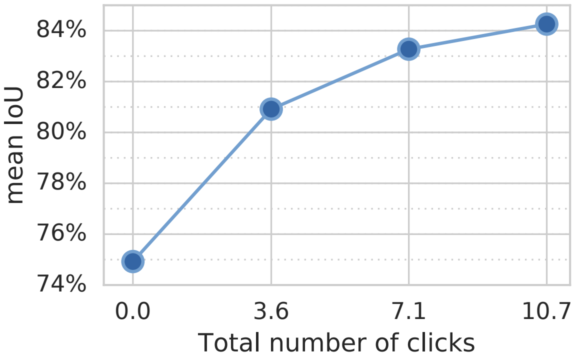


Figure 9: Masks quality per round versus mean number of collected corrective clicks (over COCO$_{\text{L}}$ 65 annotated classes, using free-paintings as evaluation ground truth). See section 4.3.

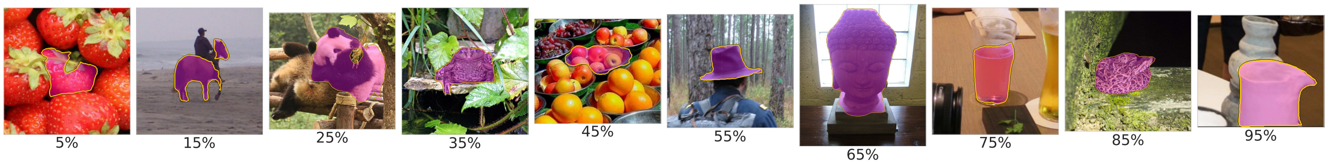


Figure 10: Example of corrective clicks masks ranked by M_r. These are OpenImages masks, each of a different class, covering different percentiles of the M_r ranking (marked below each image). Lower predicted quality on the left, highest predicted quality on the right. See section 4.4.

4. Analysis of human annotations

4.1. Free-painting annotations

Fig. 14 shows random examples of the manually created free-painting masks, expanding Fig. 6 of the main paper.

4.2. Corrective clicks: Annotators behaviour

We present here quantitative results used to support the claims of section 5.2 in the main paper.

Clicks per round. Figure 7 shows the distribution of error region areas for the 1rst, 2nd, etc. click of the corrections. We observe that, as expected, the initial clicks aim larger region than the later ones.

Clicks distribution. Figure 8 shows the mean area of click error regions (over COCO_L), as a function of the number of clicks per region. We observe that indeed only the smallest regions are left without clicks, and that the number of clicks grows almost linearly with the area of the error region. Overall, annotators indeed only do multiple clicks if the region to correct is rather large.

4.3. Corrective clicks: Time versus quality

Figure 9 shows the corrective click mask quality versus the average number of cumulated corrective clicks. This is analogue to the curve in figure 9 of the main paper, but using “number of clicks” as progress indicator (rather than seconds).

Figure 15 shows random examples of the generated corrective clicks masks, expanding figure 6 of the main paper. These are the raw results from the DeeplabV2 M_{b+c} model without any post-processing.

4.4. Corrective clicks: Masks ranking

Figure 11 shows the curves discussed in section 5.4 of the main paper. We see a clear an monotonic trend for both top and bottom n% ranked masks. This shows that M_r is effective at ranking the annotated corrective click masks. Figure 10 gives a qualitative view of the ranking.

References

[1] D. Acuna, H. Ling, A. Kar, and S. Fidler. Efficient interactive annotation of segmentation datasets with polygon-rnn++. In *CVPR*, 2018. 1

[2] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *PAMI*, 2018. 1

[3] H. Le, L. Mai, B. Price, S. Cohen, H. Jin, and F. Liu. Interactive boundary prediction for object selection. In *ECCV*, pages 18–33, 2018. 1

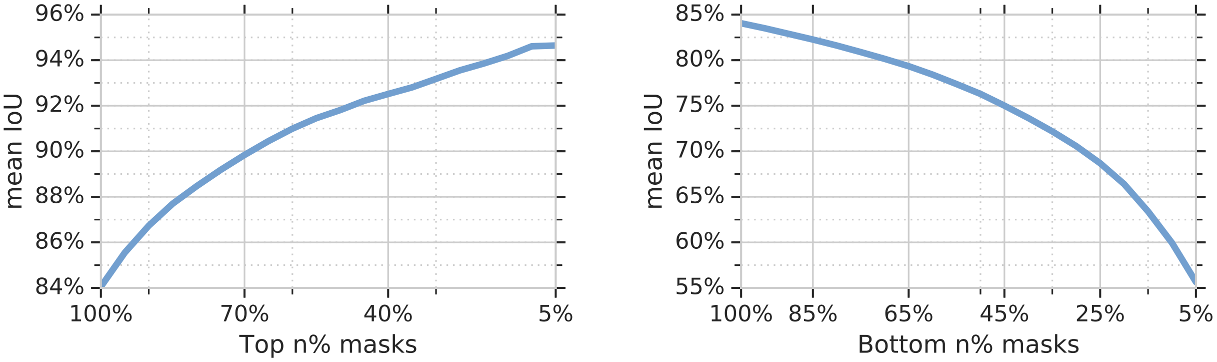
[4] Z. Li, Q. Chen, and V. Koltun. Interactive image segmentation with latent diversity. In *CVPR*, 2018. 1

[5] J. Liew, Y. Wei, W. Xiong, S.-H. Ong, and J. Feng. Regional interactive image segmentation networks. In *ICCV*, 2017. 1

[6] K.-K. Maninis, S. Caelles, J. Pont-Tuset, and L. Van Gool. Deep extreme cut: From extreme points to object segmentation. In *CVPR*, 2018. 1

[7] D. P. Papadopoulos, J. R. Uijlings, F. Keller, and V. Ferrari. Extreme clicking for efficient object annotation. In *ICCV*, 2017. 1

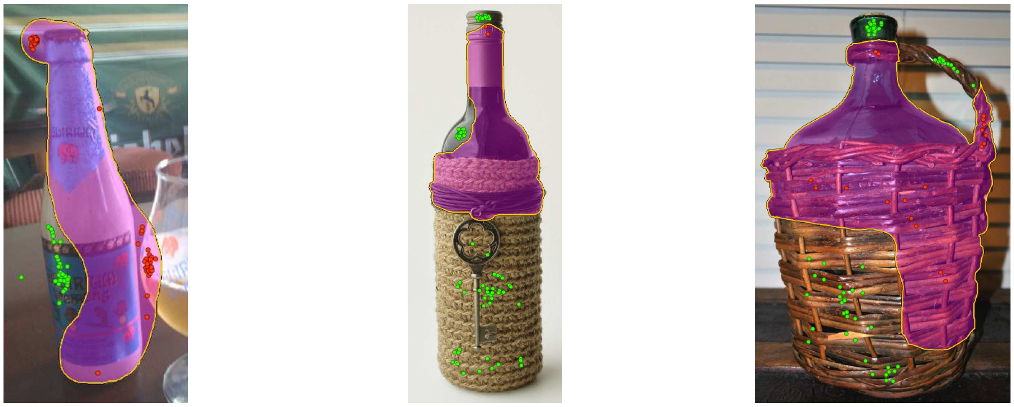
[8] N. Xu, B. Price, S. Cohen, J. Yang, and T. S. Huang. Deep interactive object selection. In *CVPR*, 2016. 1



(a) Top n% ranked masks

(b) Bottom n% ranked masks

Figure 11: mIoU over COCO_L for fraction of top/bottom ranked samples according to M_r. Top ranking subsets are most suitable for high quality models training. Bottom ranking subsets is the data that most requires further corrections. See section 4.4.

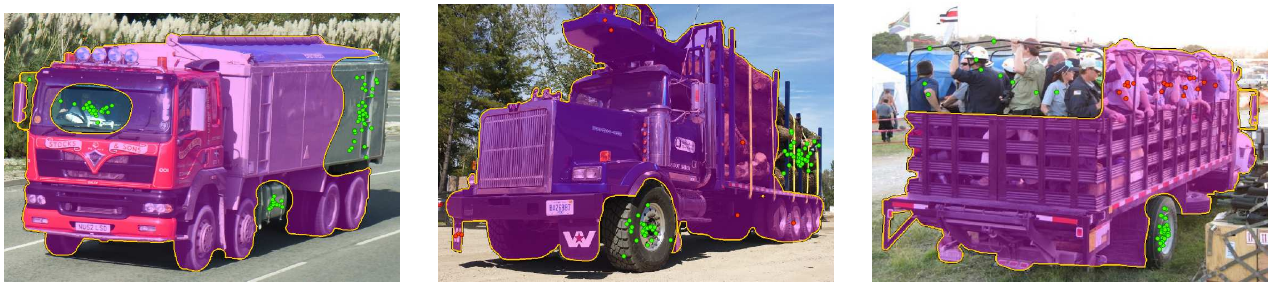


(a) Baseline question, the task is well understood.

(b) Bottle with a sleeve. Most annotators consider the sleeve as part of the bottle.

(c) Bottle with a sleeve, bis. This time, the annotators disagree on whether the sleeve is part of the bottle or not.

Figure 12: Anecdotal evidence of need of annotation policies. Example of question triplets about bottles. See discussion in section 3.1.



(a) Baseline question, the task is well understood.

(b) Loaded truck. Most annotators consider the load part of the truck.

(c) Loaded truck, bis. This time, the annotators disagree on whether the load is part of the truck or not.

Figure 13: Anecdotal evidence of need of annotation policies. Example of question triplets about trucks. See discussion in section 3.1.

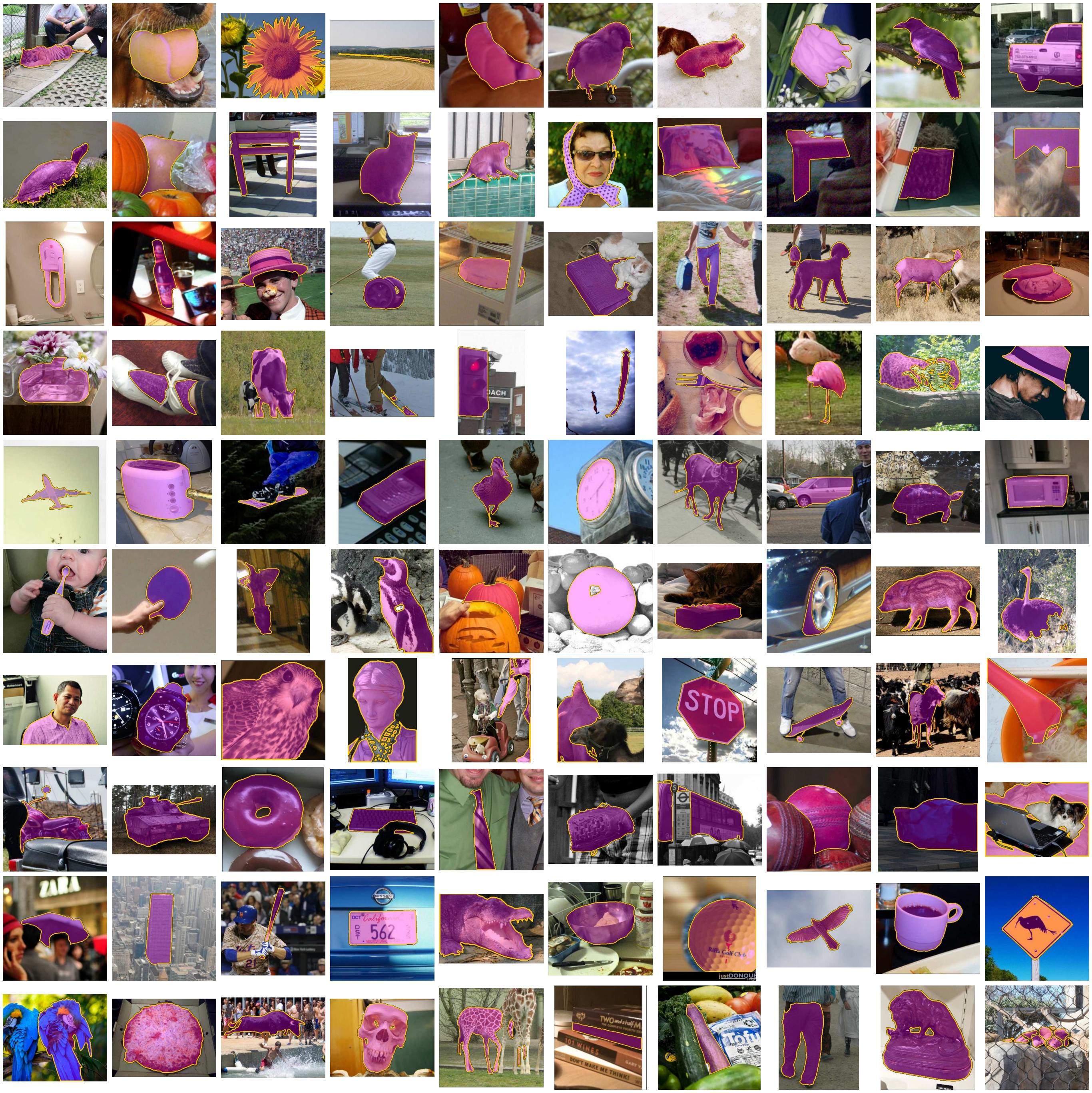


Figure 14: Random examples of free-painting OpenImage masks. 100 classes out of 300, one example per class. See section 4.1.

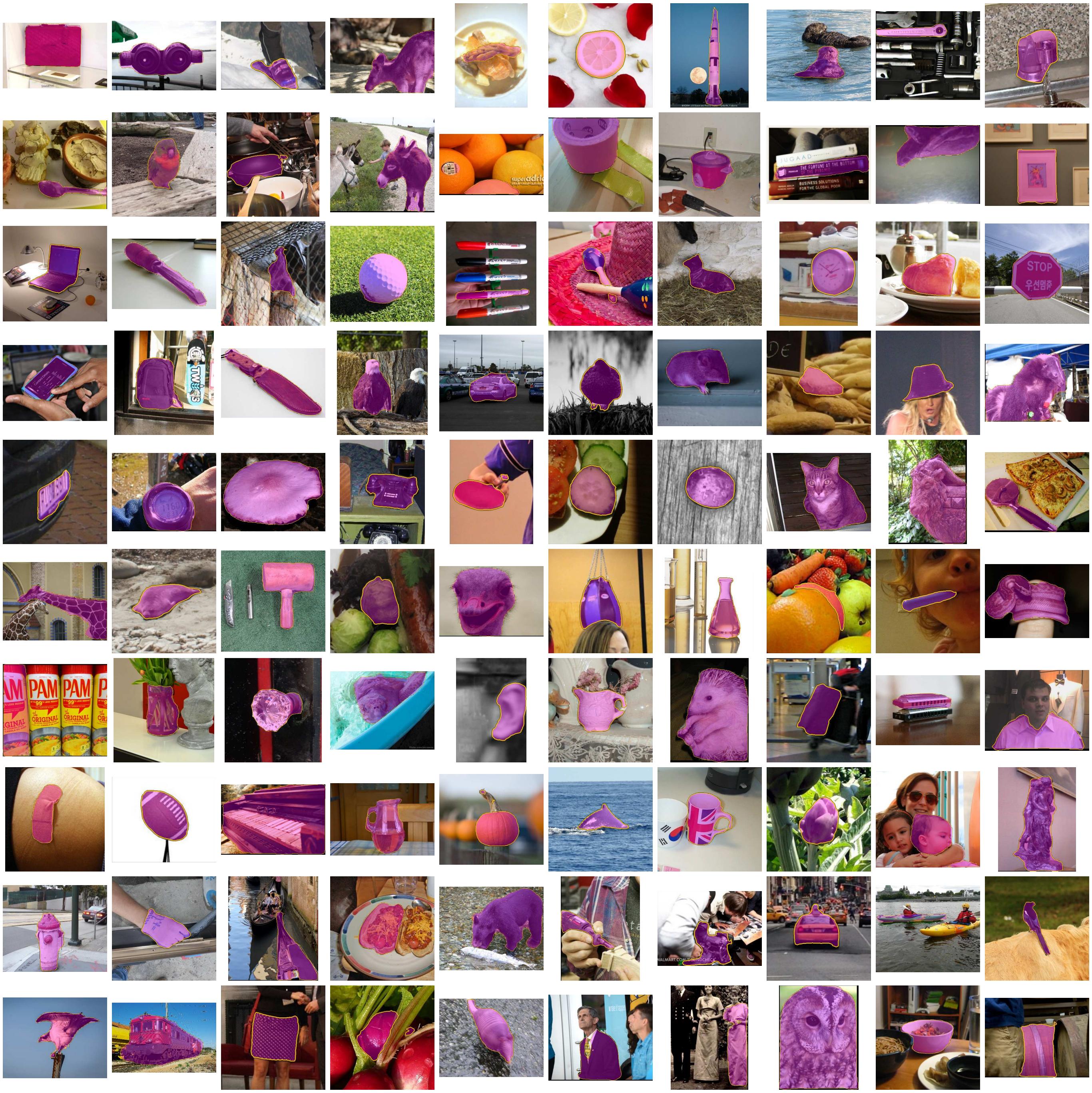


Figure 15: Random examples of corrective clicks OpenImage masks, sampled from the top 50% M_r ranking. 100 classes out of 300, one example per class. See section 4.3.