Interactive Object Segmentation with Inside-Outside Guidance Supplementary Materials

Shiyin Zhang[1,2], Jun Hao Liew[3], Yunchao Wei[4], Shikui Wei[1,2]*, Yao Zhao[1,2]
[1]Institute of Information Science, Beijing Jiaotong University
[2]Beijing Key Laboratory of Advanced Information Science and Network Technology
[3]National University of Singapore [4]ReLER, University of Technology Sydney
https://github.com/shiyinzhang/Inside-Outside-Guidance

1. Network Architecture

As described in the main paper, our segmentation network adopts a coarse-to-fine structure similar to [2]. In overall, it consists of a CoarseNet for coarse prediction, and a FineNet that aims at recovering the missing boundary details. We also append a lightweight refinement branch to accept the additional user inputs for interactive refinement. The overall architecture is depicted in Figure 1. The details of each subnetwork will be illustrated below.

1.1. CoarseNet

The CoarseNet employs a feature pyramid-style structure [9]. Specifically, it takes ResNet [7] as its backbone, with the global average pooling and `FC` layers removed. To avoid excessive loss in details due to downsampling, we set the stride and dilation rate of the last block in ResNet (`Conv5`) to 1 and 2, respectively. In addition, we also append a PSP module [14] (Figure 1(b)) after `Conv5` to enrich its representation with global contextual information. At each decoder stage, the features are progressively upsampled, followed by a 1×1 `conv` layer before being fused with the features from earlier layers via concatenation. Finally, we also introduce an auxiliary classifier at each level and apply side losses as a form of deep supervision [14, 15].

1.2. FineNet

The FineNet fuses the information across different levels in the CoarseNet via upsampling and concatenation operations. We apply more residual blocks (Figure 1(c)) for features at deeper layers following the same idea in [2]. Lastly, another residual block is applied, followed by a 3×3 `conv` layer before producing the final segmentation output.

1.3. Refinement branch

Our IOG approach also allows additional clicks input for further refinement if the user is not satisfied with the current segmentation result. To achieve this, we append a lightweight refinement branch consisting of 5 `conv` layers before the PSP module. In particular, the refinement branch takes in the two-channel Gaussian heatmaps whose resolution is the same as the input image (512×512). To match the input size of the features before the PSP module (32×32), we set the stride rate of the first three `conv` layers to 2. Then, the features are concatenated with the features extracted from the backbone ResNet (`Conv5`).

Note that we do **not** update the inputs for the CoarseNet, *i.e.* the CoarseNet always takes the same 3 clicks (augmented with the RGB image) as inputs while the refinement branch takes all the clicks input. Given this setting, the encoder features (from `Conv1` to `Conv5` in ResNet) only needs to be computed once, thus making the refinement process very fast. Furthermore, as discussed in our main paper, this setting also performs better than modifying the inputs for the CoarseNet.

2. More qualitative results

Due to space constraints, we only managed to show a few qualitative results in our main paper. Here, we would like to present more qualitative examples. Furthermore, we also recommend our readers to watch the videos in the supplementary materials where a real-time demo is recorded.

2.1. General scenes

Figure 2 and 3 show some qualitative results of our IOG on the PASCAL [5] and COCO [10] dataset. In Figure 2, we can see that our IOG performs well even on challenging scenes, such as occlusion (2nd row, 1st column), complex background (4th row, 4th column) and presence of small objects (1st row, 3rd column and 4th row, 2nd column). Similar observation can be made in Figure 3.

---

*Corresponding author

2.2. Cross-domain performance

In addition to the results presented in the main paper, we show more qualitative results of our proposed IOG on different imagery types, including street scenes (Cityscapes [4]), aerial imagery (Rooftop [13]) and medical images (ssTEM [6]). The results are shown in Figure 4. Despite not trained on those dataset before, our IOG often produces good segmentation results even **without** fine-tuning, demonstrating the superior generalization ability of our method. In addition, we also present some qualitative results of our IOG on the more challenging Agriculture-Vision dataset [3]. As shown in Figure 4, our model still achieves satisfactory performance when using only small amounts of data for fine-tuning. Moreover, we also provide some qualitative results of our IOG fine-tuned on PASCAL-Context [11] and COCO-Stuff [1] dataset in Figure 5 and 6. The results suggests that our proposed IOG not only performs well in segmenting "things", but also generalizes well to "stuff" categories such as buildings, ground, wall *etc*., demonstrating its effectiveness as an annotation tool.

2.3. Interactive refinement

As described previously in the main paper, our proposed IOG naturally supports annotation of extra clicks for further refinement. Here, we also show some examples of the interactive refinement process given more clicks from the user. As shown in Figure 7, the interactive correction process is very efficient in correcting the erroneous regions.

2.4. Comparison with "Object Selection Tool" from Photoshop CC

We also compare our IOG with the latest "object selection tool" from Photoshop CC 2020. Interestingly, our IOG performs better than "object selection tool" when there are overlapping instances, such as the birds (1st column of top row) and kangaroos (3rd column of top row) in Figure 8.

2.5. Extension to datasets with box annotations

In the main paper, we have discussed a potential application of our IOG where it can be used for harvesting high-quality instance masks from existing datasets with off-the-shelf bounding box annotations, such as ImageNet [12] and Open Image [8]. Some qualitative results are visualized in Figure 9 and 10. Note that our IOG performs well even when segmenting instances from unseen classes, such as tire, starfish, snail *etc*.

References

[1] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. pages 1209–1218, 2018. 2

[2] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 1

[3] Mang Tik Chiu, Xingqian Xu, Yunchao Wei, Zilong Huang, Alexander Schwing, Robert Brunner, Hrant Khachatrian, Hovnatan Karapetyan, Ivan Dozier, Greg Rose, et al. Agriculture-vision: A large aerial image database for agricultural pattern analysis. In *CVPR*, 2020. 2

[4] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 2

[5] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *IJCV*, 2010. 1

[6] Stephan Gerhard, Jan Funke, Julien Martel, Albert Cardona, and Richard Fetter. Segmented anisotropic sstem dataset of neural tissue. *figshare*, pages 0–0, 2013. 2

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[8] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv preprint arXiv:1811.00982*, 2018. 2

[9] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 1

[10] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 1

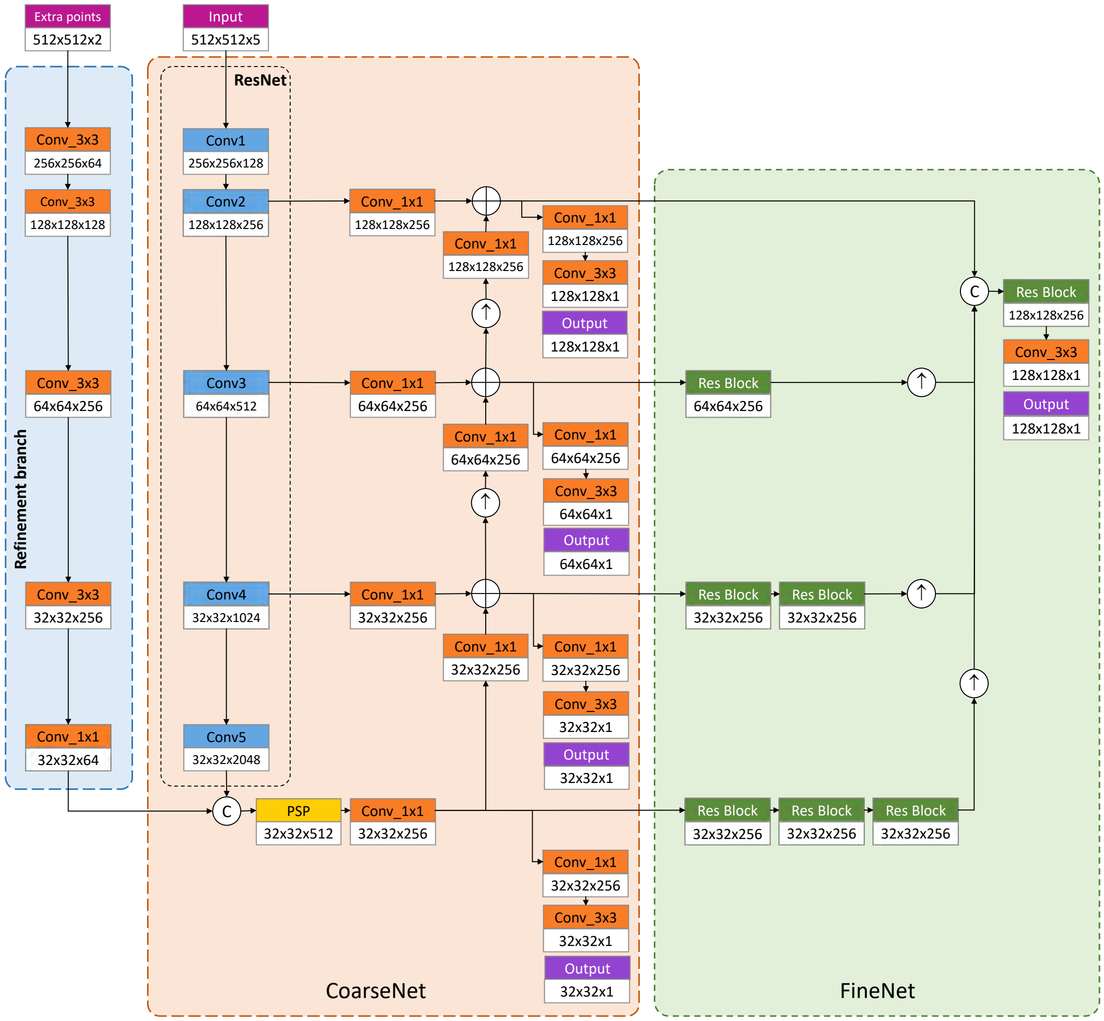
[11] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Namgyu Cho, Seongwhan Lee, Sanja Fidler, Raquel Urtasun, and Alan L Yuille. The role of context for object detection and semantic segmentation in the wild. pages 891–898, 2014. 2

[12] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 2

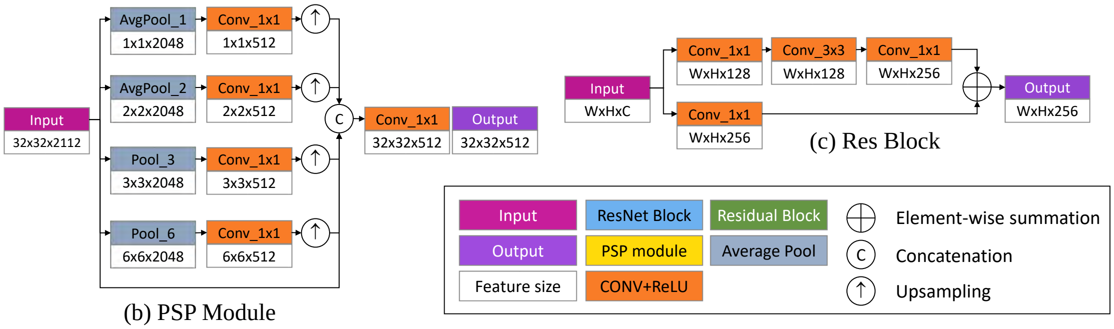
[13] Xiaolu Sun, C Mario Christoudias, and Pascal Fua. Free-shape polygonal object localization. In *ECCV*, pages 317–332. Springer, 2014. 2

[14] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 1

[15] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. Psanet: Point-wise spatial attention network for scene parsing. In *ECCV*, 2018. 1



(a) Coarse-to-fine network architecture



(b) PSP Module

(c) Res Block

Figure 1. **Overall network architecture of our segmentation network.**

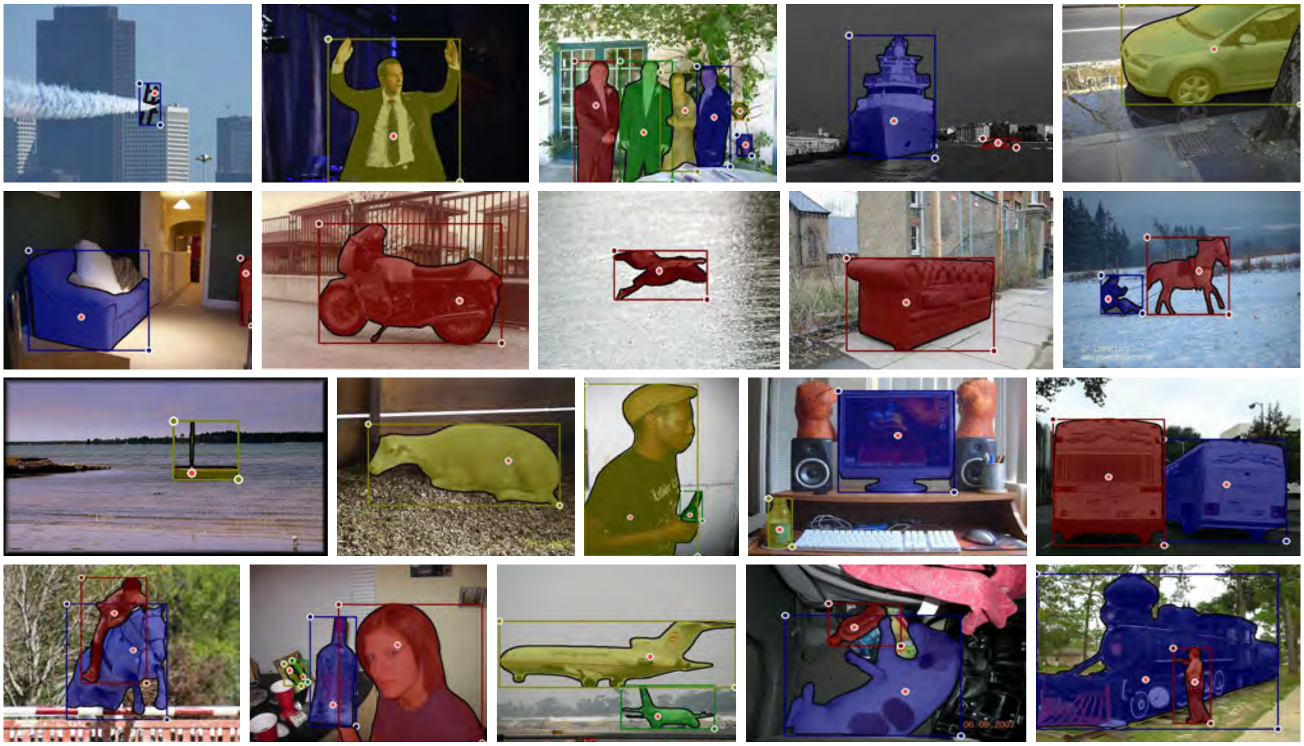


Figure 2. **Qualitative results on PASCAL.**

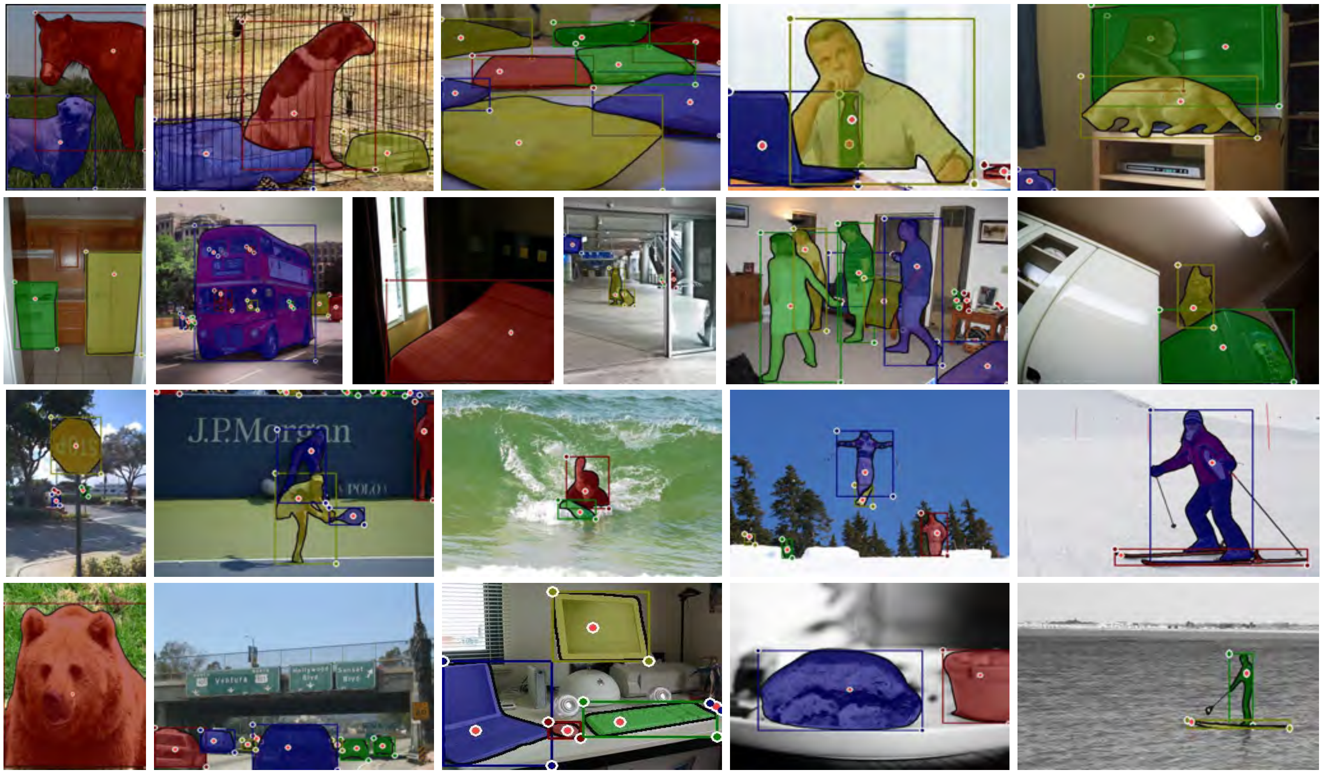


Figure 3. **Qualitative results on COCO.**

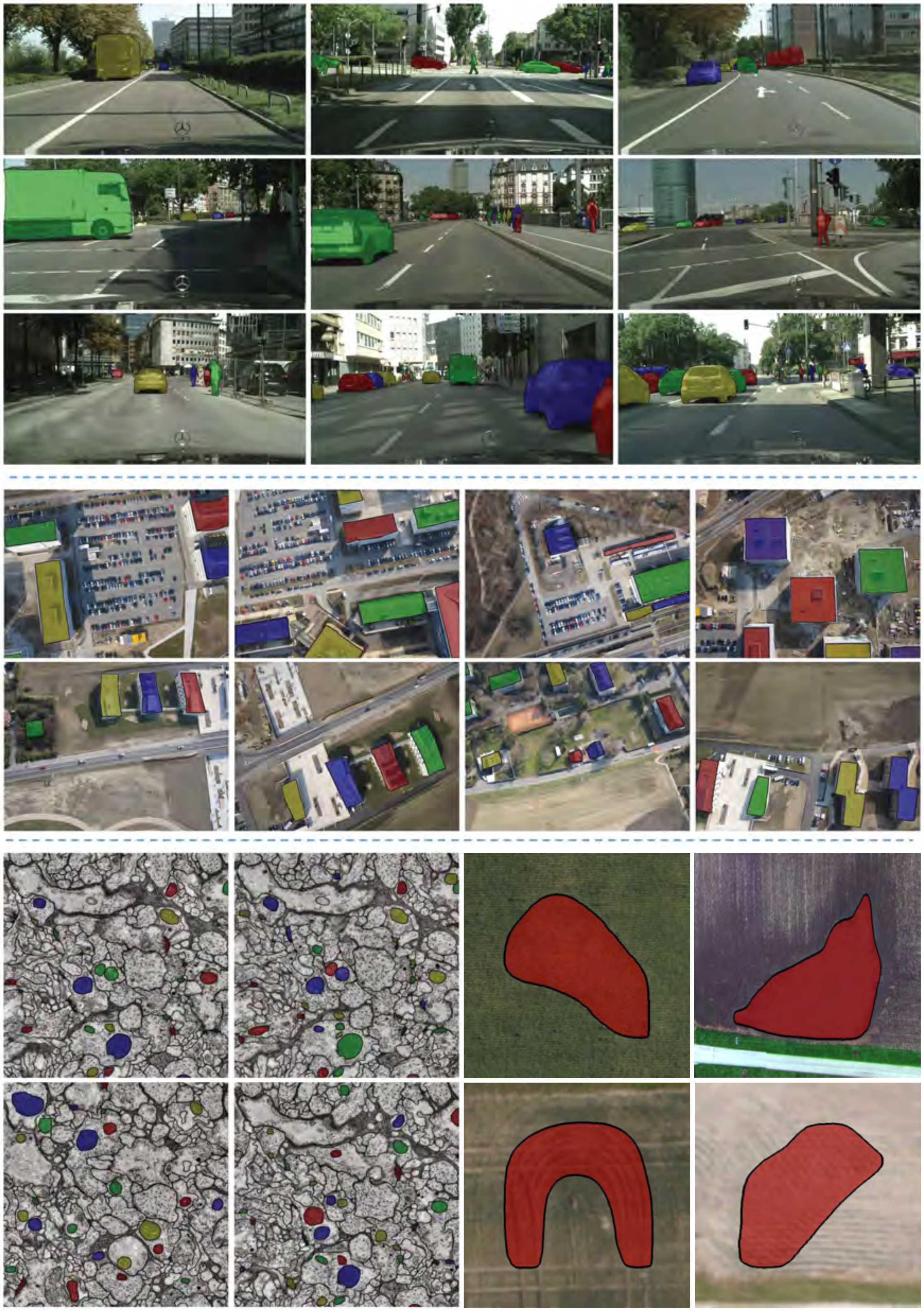


Figure 4. **Cross-domain performance.** Qualitative results of our IOG on street scenes (top), aerial imagery (middle), medical images (bottom left) and Agriculture-Vision (bottom right).

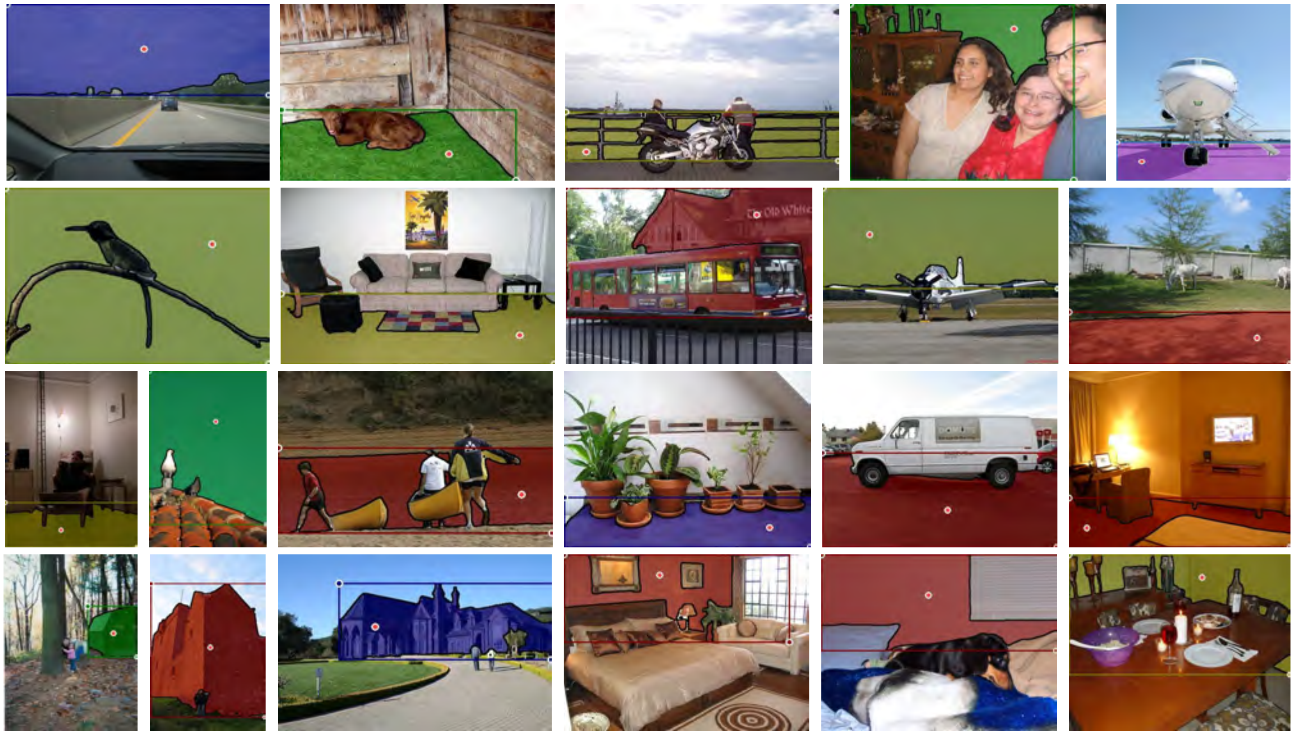


Figure 5. **Qualitative results on PASCAL-Context.**

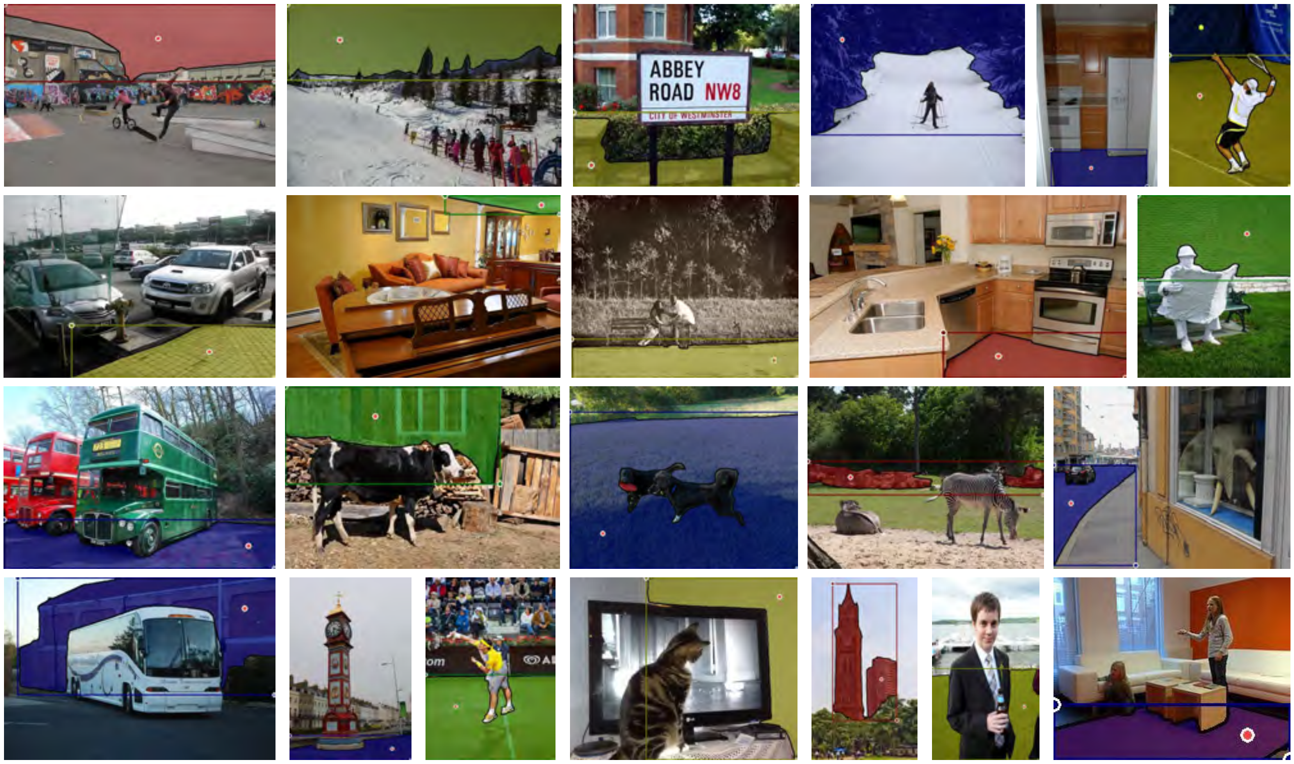


Figure 6. **Qualitative results on COCO-Stuff.**

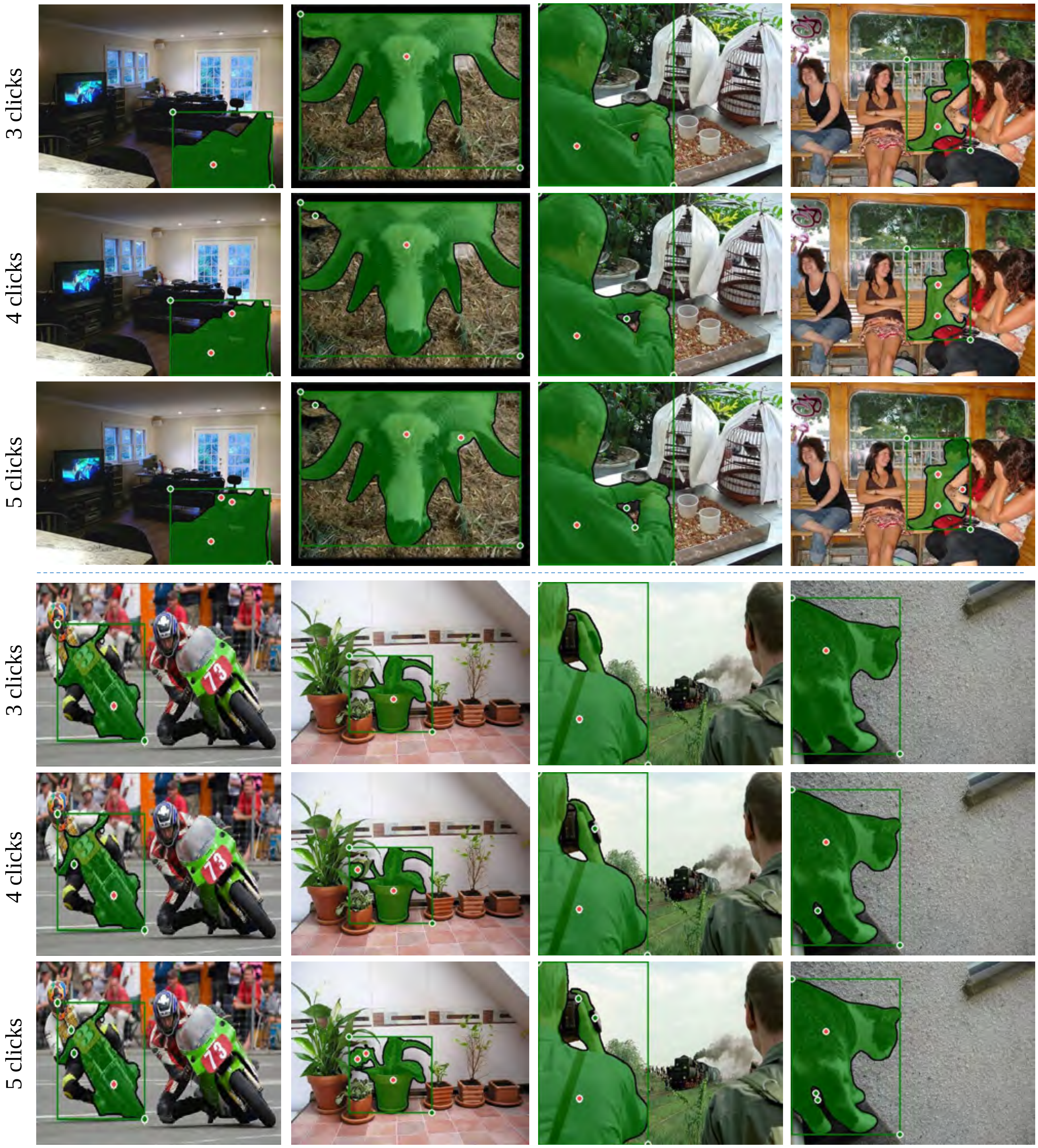


Figure 7. **Interactive refinement.** The red and green clicks denote the foreground and background clicks, respectively.

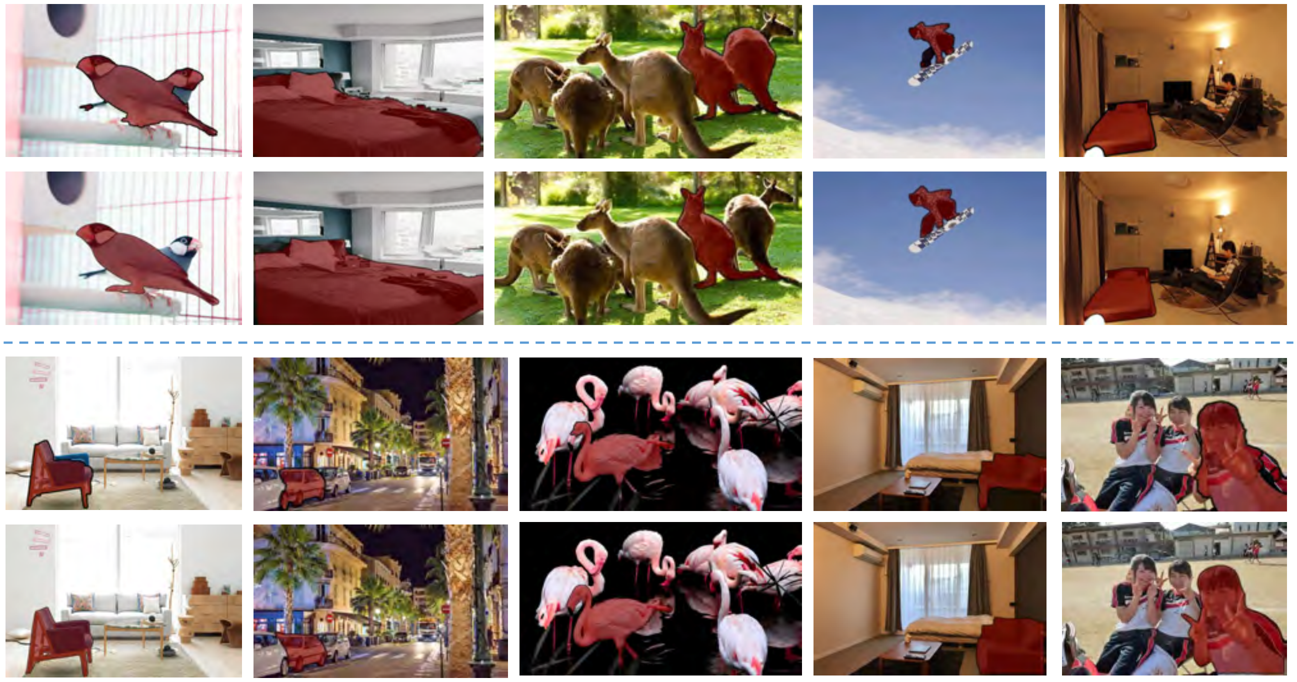


Figure 8. **Qualitative comparison between "object selection tool" from Photoshop CC 2020 (top) and our IOG (bottom).** Note that the test images are taken from the internet.

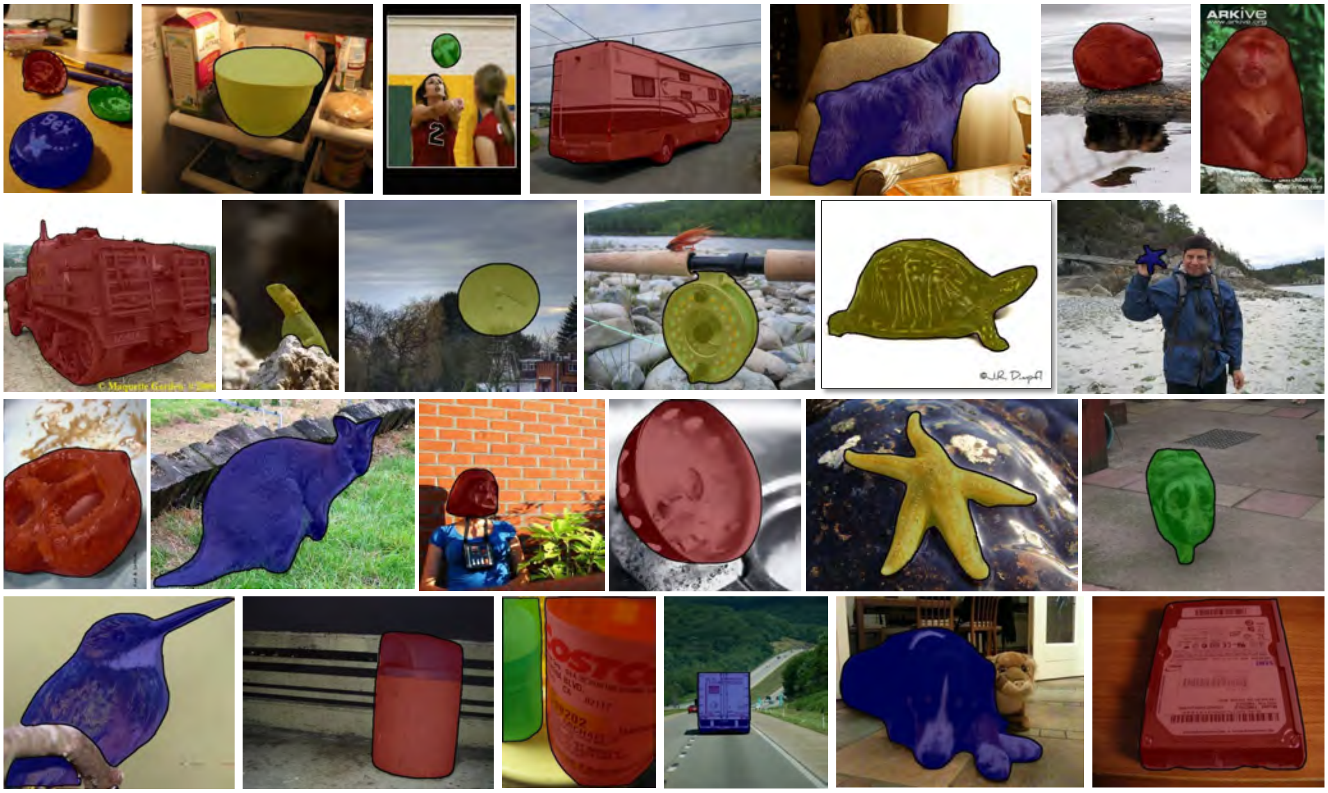


Figure 9. **Qualitative results on ImageNet using our proposed 2-stage approach.** Note that only bounding box annotations are provided. Please refer to our main paper for more details.

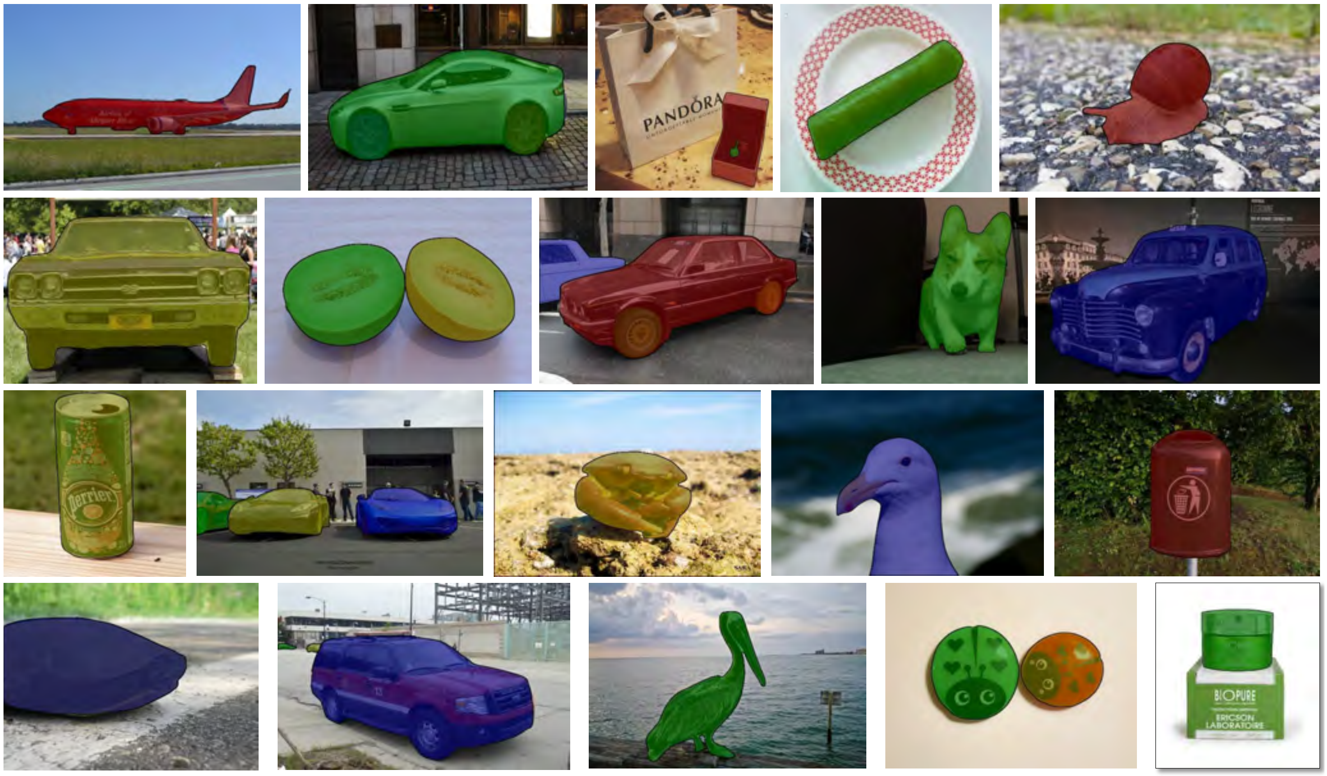


Figure 10. **Qualitative results on Open Image using our proposed 2-stage approach.** Note that only bounding box annotations are provided. Please refer to our main paper for more details.