A Bottom-Up Approach to Class-Agnostic Image Segmentation

Supplementary Material

Sebastian Dille[1], Ari Blondal[1,2], Sylvain Paris[3], and Yağız Aksoy[1]

[1] Simon Fraser University, Canada
[2] McGill University, Canada
[3] Adobe Research, United States

In this supplementary document, we present (i) extended qualitative analysis and discussion in Section A and in Figures 2 and 3, and (ii) extended discussion and a pseudo-code for the multi-resolution refinement process in Section B. Additionally, we show a visualization of our attraction and repulsion term from Section 3.4 for the three-dimensional case in Figure 1.

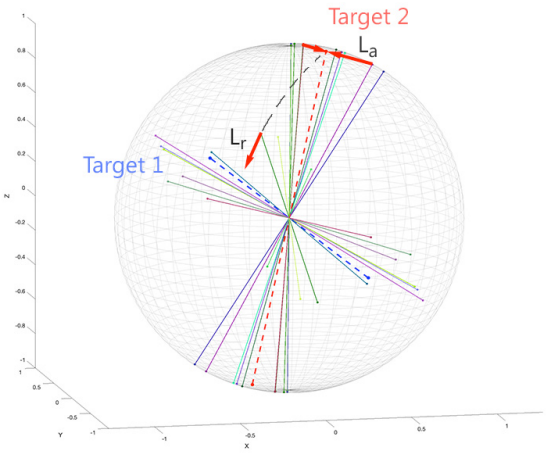

Fig. 1: We visualize the functionality of our attraction and repulsion losses exemplary on a 3D sphere. The features in this simple scenario form two bundles, concentrated around Target 1 (dotted blue) and Target 2 (dotted red). L_a attracts features corresponding to Target 2 closer towards its main orientation line, while L_r pushes other features further away and towards Target 1.

A Qualitative Analysis

We extend the quantitative experiments from our main paper with a set of qualitative comparisons. We demonstrate the generalization capability of our

approach on images in the wild and provide qualitative examples for the difficult cell-segmentation task.

A.1 Qualitative Analysis in-the-wild

We show comparisons on a wide variety of in-the-wild images in Figure 2. We have named each example for easy referral. There are several interesting conclusions we can make about the limited generalization of top-down approaches.

The Strawberries example represent arguably a very easy image for segmentation, being composed of distinct repetition of simple objects. Despite the simplicity of the input, we observe that other methods fail to successfully segment the images, some producing a single segment for foreground or background, some only segmenting parts of the strawberries, and others not generating any segments. Apart from our method, only Segment Anything [3] with the small ViT-B encoder successfully segments the strawberries, but over-segments the silhuette image due to color differences in the foreground. The Wall example is another simple image to segment, only complicated by the shadows of the palm trees. We again see a similar behavior of generating a single background segment, oversegmenting or no segment at all from others. While this can be explained by the lack of strawberry, or window blind classes in the training datasets of panoptic segmentation methods, the class-agnostic approaches by Qi *et al.* [7] and Qi *et al.* [6] also perform similarly. We believe that this demonstrates the limited generalization ability of top-down approaches for image segmentation when trained on standard class-based datasets.

The Cars example shows a sleigh and reindeers in the middle of a road full of cars. We see that panoptic segmentation methods end up missing the unusual objects in the scene, while class-agnostic methods are able to segment Santa's ride. Kirillov *et al.* [3] misses the sleigh but is the only method that provides segments for the intricate car windows and rear lights.

Despite the variety in the style and complexity of the input images, our bottom-up approach is able to differentiate the objects in the scene, demonstrating our generalization ability. In some cases such as in the Painting example, while detecting most of the entities in the scene, it fails to differentiate between some of the walnuts, showing similar performance to Kirillov *et al.* [3].

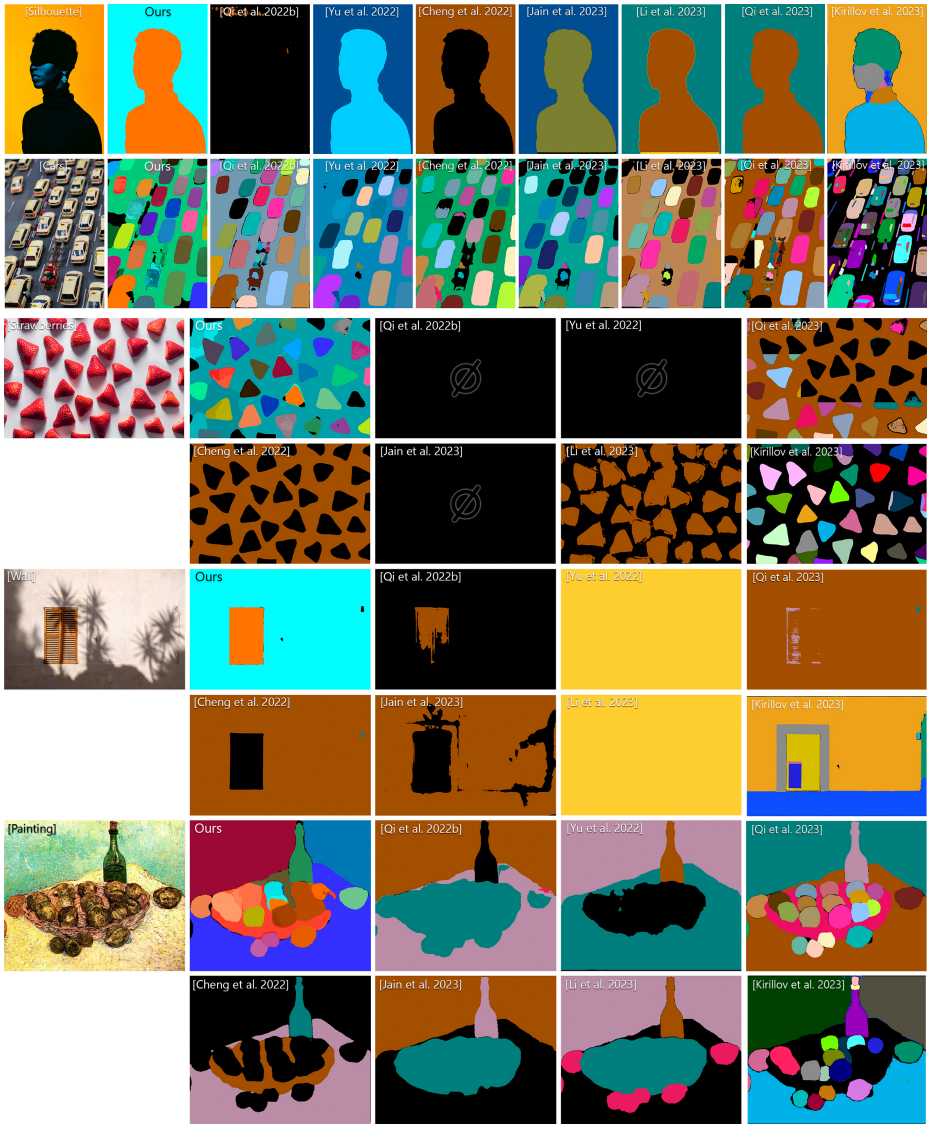

Fig. 2: We show the results of our method on images in the wild in comparison with [7], [9] [1], [2], [4], [6] and [3]. Please refer to Section A.1 for discussion.

A.2 Qualitative Results on Cell Segmentation

Additionally, we provide example results for cell and nucleus segmentation on the EVICAN dataset [8] in Figure 3. Note that none of the baselines that we compare against publicly share their code. Instead, we compare our results with the provided ground truth annotations in the dataset. Our method manages to segment the cells and nuclei from the three difficulty levels well. Even for challenging low-contrast input images such as in the middle row or for cell clusters as in the top row, the segments closely match the ground truth annotations.

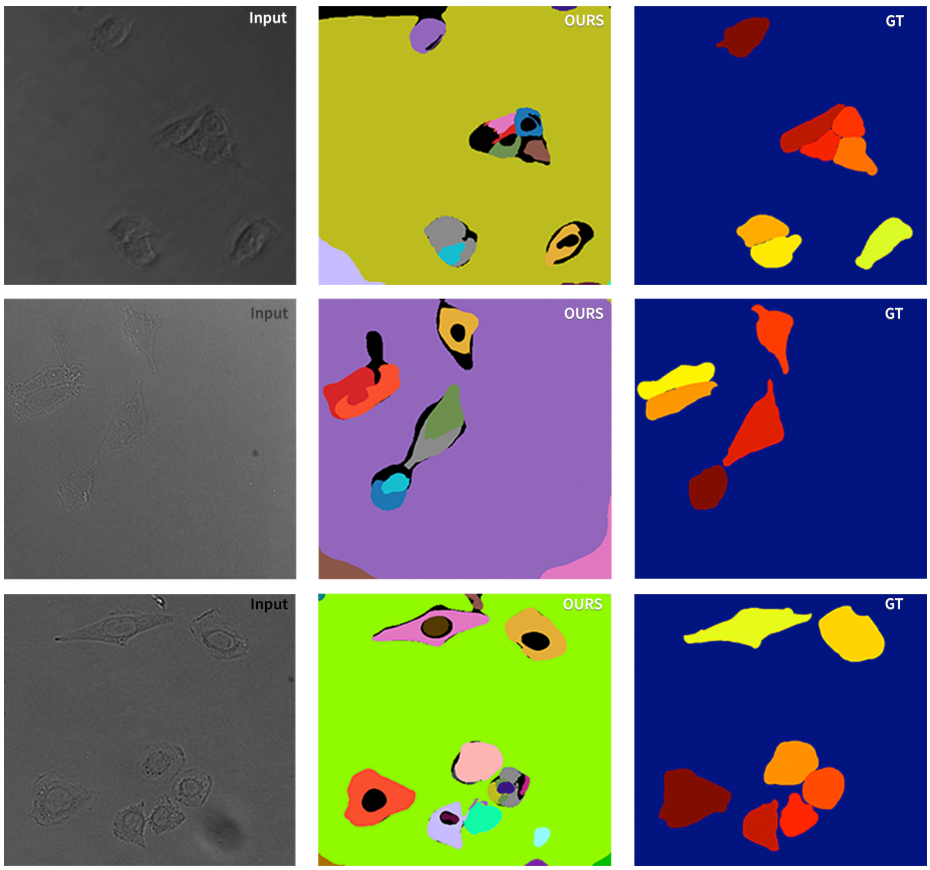

Fig. 3: We show example images from EVICAN [8] with our results in comparison with the ground truth from the hard (upper row), the medium (middle row) and the easy subset (lower row).

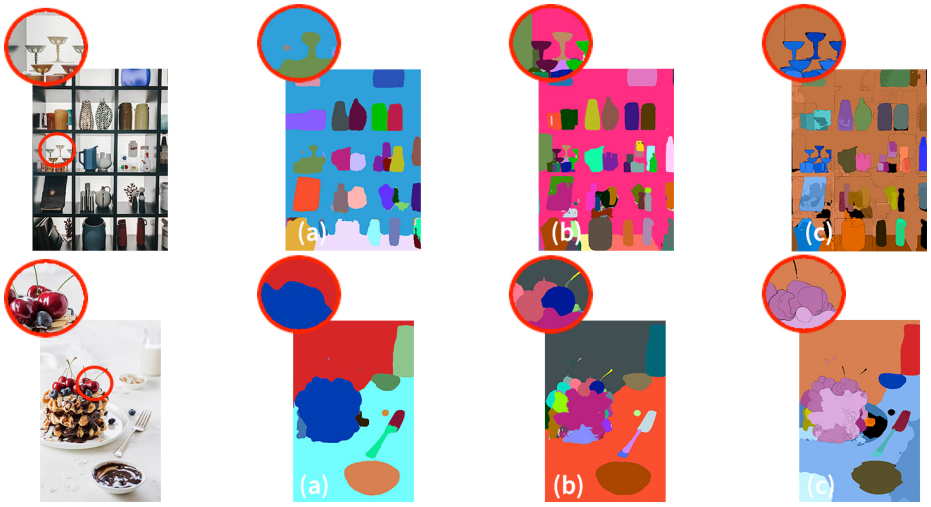

Fig. 4: For a complex scene, the initial estimation in (a) contains merged segments and inaccurate boundaries. Inputs at higher resolution in (b) retrieve more intricate details that can interatively be merged as shown in (c).

B Multi-resolution refinement

For class-agnostic segmentation, the behavior of the network at different resolutions directly corresponds to the context within the receptive field. As Figure 4 shows, when looking at the entire image, the meaningful regions roughly correspond to the shelf units in the upper row and the plates and the entire stack of waffles in the lower row. At higher resolutions, though, as the network focuses on local regions, the context changes, and the contents of the shelf units can be separated into individual objects. For the waffles, we even retrieve the cherries on top. This is similar to how humans divide the scene they are looking at into meaningful regions.

We hence employ a multi-resolution refinement process at inference. Starting at the receptive field size of the network, we increase the input resolution by a factor of 1.25 until we reach the maximum resolution that is likely to result in over-segmentation. We adopt the gradient-based ideal input resolution $\mathcal{R}_{20}$ from [5] to estimate this maximum level on a per-image base. For every inference step, we generate labels via mean-shift clustering and compare them with the root segments. At an $IoU > 0.85$, we assume that both segments are identical but the old segment is likely less accurate and hence replaced by the new one. We assume a merging error if the old segment is covered by more than 50% but the new segment is not, in which case we disregard the new one. If the new segment is covered by more than 50%, we assume it's a child and recursively repeat the comparison with all previously detected children to find the direct parent. Algorithm 1 shows the pseudo-code for this procedure.

Algorithm 1 Multi-Resolution Refinement Process

Input: Network with initial receptive field size, Image
Output: Refined segmentation
$currentR \leftarrow$ initial receptive field size
$maxR \leftarrow$ Ideal input resolution $\mathcal{R}_{20}$ [5]
while $currentR \leq maxR$ **do**
    Perform inference at $currentR$
    Generate labels using mean-shift clustering
    Compare labels with root segments
    for each pair (root, new) **do**
        $iou \leftarrow IoU(root, new)$
        $coverageRoot \leftarrow iou/Area(root)$
        $coverageNew \leftarrow iou/Area(new)$
        if $iou > 0.85$ **then**
            $root \leftarrow new$
        else if $coverageNew < 0.5$ **then**
            if $coverageRoot > 0.5$ **then**
                $delete(new)$
            end if
        else
            recursively compare against children
            assign to direct parent
        end if
    end for
    $currentR \leftarrow currentR * 1.25$
end while
return Refined segmentation

References

1. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proc. CVPR (2022)
2. Jain, J., Li, J., Chiu, M., Hassani, A., Orlov, N., Shi, H.: OneFormer: One Transformer to Rule Universal Image Segmentation. In: Proc. CVPR (2023)
3. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollar, P., Girshick, R.: Segment anything. In: Proc. ICCV (2023)
4. Li, F., Zhang, H., xu, H., Liu, S., Zhang, L., Ni, L.M., Shum, H.Y.: Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In: Proc. CVPR (2023)
5. Miangoleh, S.M.H., Dille, S., Mai, L., Paris, S., Aksoy, Y.: Boosting monocular depth estimation models to high-resolution via content-adaptive multi-resolution merging. In: Proc. CVPR (2021)
6. Qi, L., Kuen, J., Shen, T., Gu, J., Guo, W., Jia, J., Lin, Z., Yang, M.H.: High quality entity segmentation. In: Proc. ICCV (2023)
7. Qi, L., Kuen, J., Wang, Y., Gu, J., Zhao, H., Torr, P., Lin, Z., Jia, J.: Open world entity segmentation. IEEE Trans. Pattern Anal. Mach. Intell. (2022)
8. Schwendy, M., Unger, R.E., Parekh, S.H.: Evican—a balanced dataset for algorithm development in cell and nucleus segmentation. Bioinformatics **36**(12), 3863–3870 (2020)
9. Yu, Q., Wang, H., Qiao, S., Collins, M., Zhu, Y., Adam, H., Yuille, A., Chen, L.C.: k-means mask transformer. In: Proc. ECCV (2022)