Unsupervised Segmentation by Diffusing, Walking and Cutting

Supplementary Material

A. A note on evaluation strategies

Zero-shot unsupervised segmentation presents unique evaluation challenges compared to traditional supervised approaches. This section discusses three complementary evaluation strategies that better capture different aspects of zero-shot performance.

Traditional Global Averaging vs. Per-Image Evaluation While previous non-zero-shot methods typically use global averaging (accumulating statistics across all images before calculating metrics) [9, 22], this may not be ideal for zero-shot models. Global averaging emphasizes performance on frequent classes, but zero-shot models have no prior knowledge of class distributions. Instead, per-image averaging, which computes metrics independently for each image before averaging, treats each instance of class discovery equally - better reflecting the zero-shot nature of the task.

The Challenge of Granularity A key challenge in evaluating zero-shot models is their tendency to discover more fine-grained segmentations than present in the ground truth. For example, in Figure 7, row 7, where the ground truth groups a bear and bird into a single ”animal” class, models trained on the CocoStuff dataset e.g. STEGO [22] also group them together; in contrast, our model distinguishes them as separate classes. While this more detailed segmentation may be semantically meaningful, it is penalized by both global and per-image metrics.

Oracle-Merged Evaluation To address this, we introduce an oracle-merged evaluation strategy that uses ground truth to merge oversegmented areas based on their primary class overlap. This is conceptually similar to EmerDiff’s approach [37], though they merge regions based on embedding similarity. However, EmerDiff requires pre-specifying the number of clusters (K=30), which leads to extreme oversegmentation as shown in Figure 9. While this hurts their performance under standard metrics, they benefit significantly from merged evaluation strategies, as discussed below.

Our extended evaluation framework combines all three approaches. As shown in Table 6, our method outperforms other diffusion-based zero-shot baselines across all metrics. EmerDiff achieves higher accuracy only in the oracle-merged setting, likely due to their cross-attention modulating upsampling approach. The qualitative examples in Figure 9 demonstrate why our approach still outperforms EmerDiff in the more strict F1 and mIoU metrics - despite benefiting from a more sophisticated upsampling strategy, EmerDiff produces in noisy clusters. Furthermore, even our ”oversegmented” versions produce high-quality, semantically meaningful segments, without the need to specify the number of clusters or any other hyperparameter a priori, unlike EmerDiff and DiffSeg [46].

Importantly, this extended evaluation framework further underscores that metrics don’t necessarily measure segmentation ”correctness” (which is inherently subjective), but rather alignment with human-labeled ground truths - which can vary in granularity across datasets and images.

	Globally averaged (traditional)			Per-image averaged			Merged and per-image averaged		
	Acc	F1	mIoU	Acc	F1	mIoU	Acc	F1	mIoU
STEGO	56.9	-	28.2	73.3	-	23.3	-	-	-
EmerDiff	28.8	22.8	13.3	28.9	47.8	0.1	92.5	34.9	24.0
DiffSeg	72.5	58.5	43.6	72.6	71.1	34.1	79.2	73.0	60.6
Ours	74.1	60.8	46.6	74.0	73.3	38.0	83.6	76.0	64.5

Table 6. We calculate metrics on predictions from Coco-Stuff-27 validation set under three different settings: globally averaged, per-image averaged, and per-image averaged with an oracle which merges oversegmented regions; our approach excels in all three.

B. Additional qualitative examples

Besides qualitative comparison for all methods on the two standard segmentation benchmark datasets CocoStuff-27 [5] in Figure 7 and Cityscapes [11] in Figure 8, we also visualize segmentations obtained via our proposed method on anomalous objects from MVTec [4] in Figure 10; Brain MRI scans from BraTS 2017 [36] in Figure 11, damaged analogue media [24] in Figure 12 and real-life images captured in the wild in Figure 13.

C. Ablation on feature extraction timestep

Regarding DDIM inversion and choice of noising step, we conduct ablation experiments, shown in Table 7. Our method performs robustly across different timesteps [30,45], and maintains performance even without DDIM inversion.

Step	25	30	35	40	45	40 no inv.
mIoU	43.1	45.7	46.6	46.6	45.3	46.6

Table 7. Ablation over noising step and importance of DDIM Inversion (mIoU) on Coco-Stuff-27.

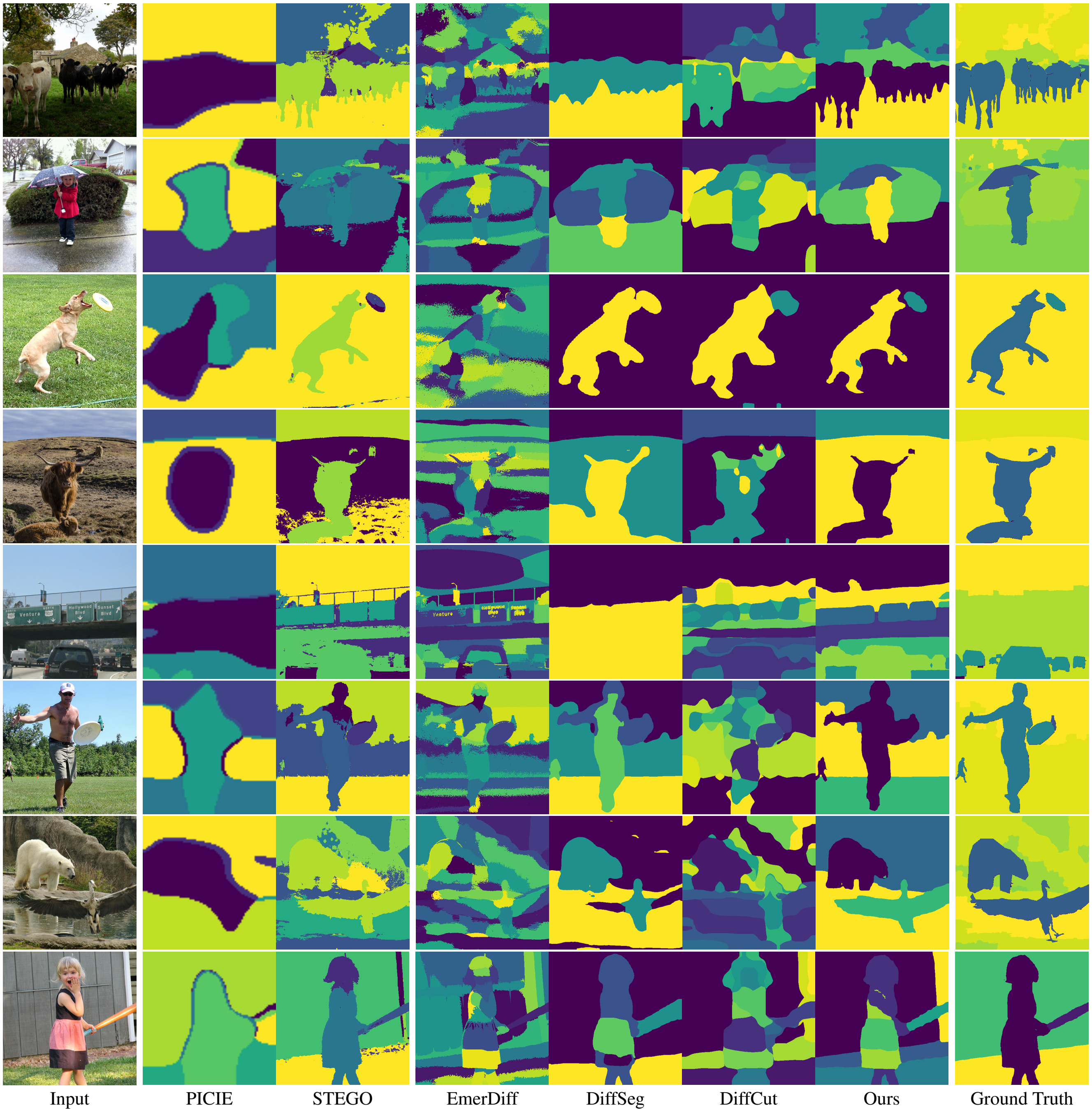

Figure 7. Qualitative comparison of segmentation results on Coco-Stuff-27 [5] Validation across multiple methods. All approaches are unsupervised; PICIE and STEGO are fit on the training set of Coco-Stuff-27, while EmerDiff, DiffSeg, DiffCut and Ours are all zero-shot and rely only on features extracted from Stable Diffusion. Our approach, in contrast to DiffSeg, DiffCut and EmerDiff, is also hyperparameter-free.

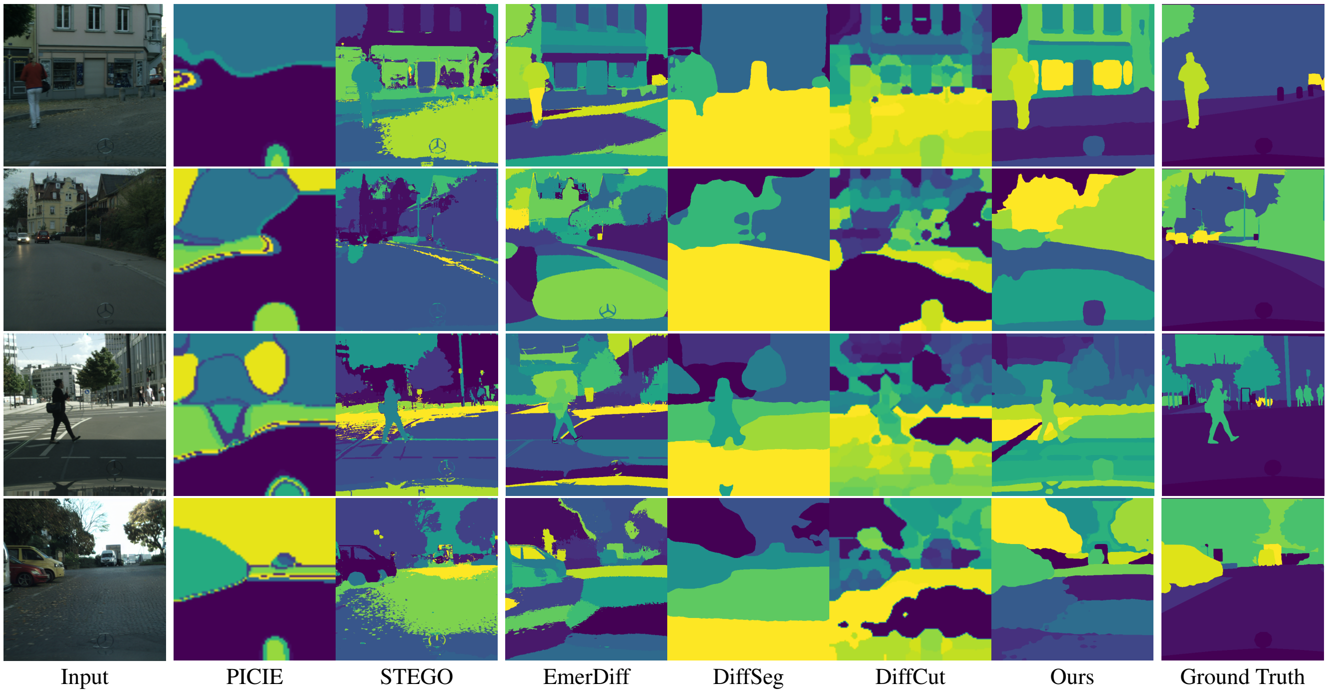

Figure 8. Qualitative comparison of segmentation results on Cityscapes [11] Validation across multiple methods. All approaches are unsupervised; PICIE and STEGO are fit on the training set of Cityscapes, while EmerDiff, DiffSeg, DiffCut and Ours are all zero-shot and rely only on features extracted from Stable Diffusion. Our approach, in contrast to DiffSeg, DiffCut and EmerDiff, is also hyperparameter-free.

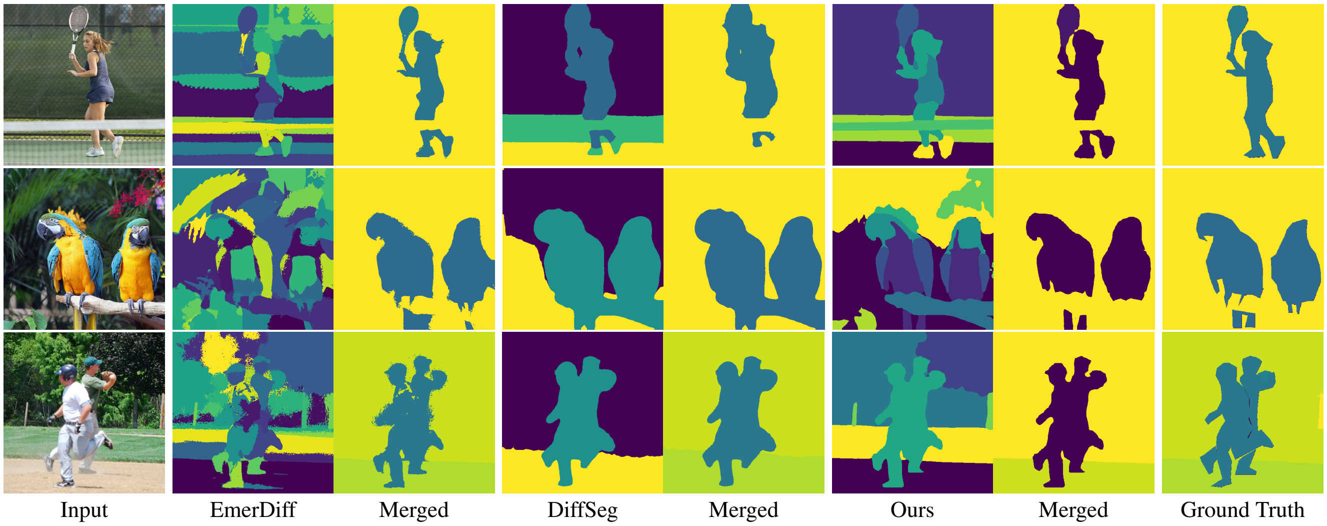

Figure 9. Examples of predictions used in the traditional evaluation setting (Hungarian matching) from the zero-shot diffusion feature-based approaches (columns 2, 4, and 6), and the corresponding "merged" predictions (columns 3, 5, and 7), designed to account for the models segmenting more granular classes compared to what is given in the ground truth. Notice that predictions from our model can be even more precise (baseball players example, row 3) or more accurate (tennis net covering the player's legs, row 1).

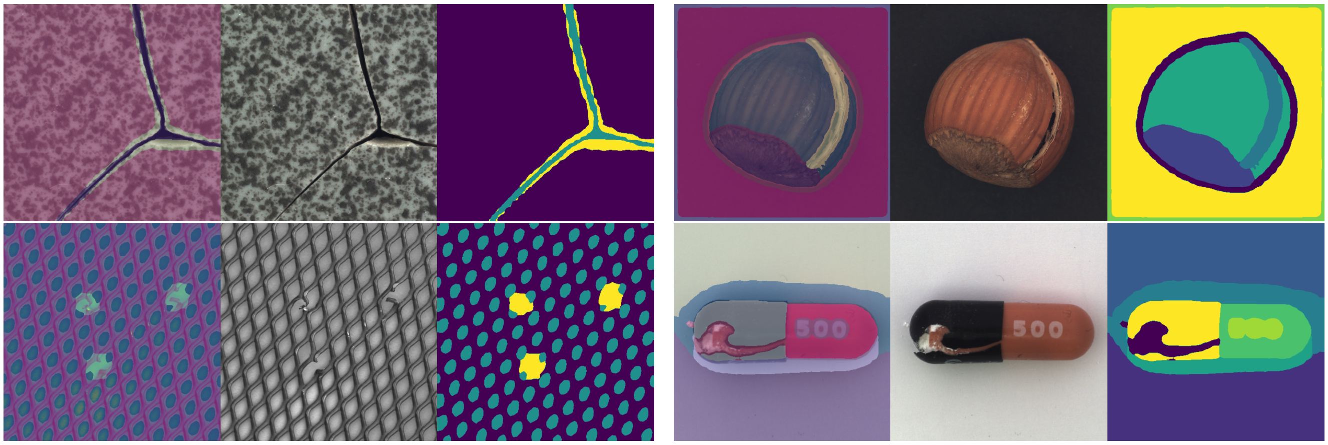

Figure 10. Segmentations of various types of anomalous objects from the MVTec Dataset [4].

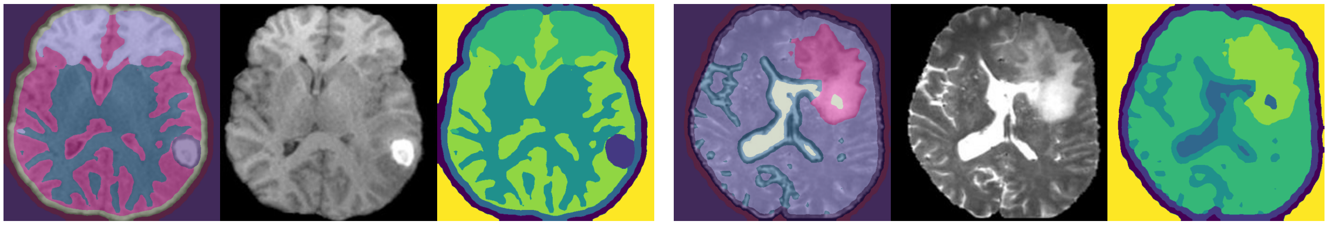

Figure 11. Segmentations of Brain MRI scans from the BraTS 2017 Dataset [36].

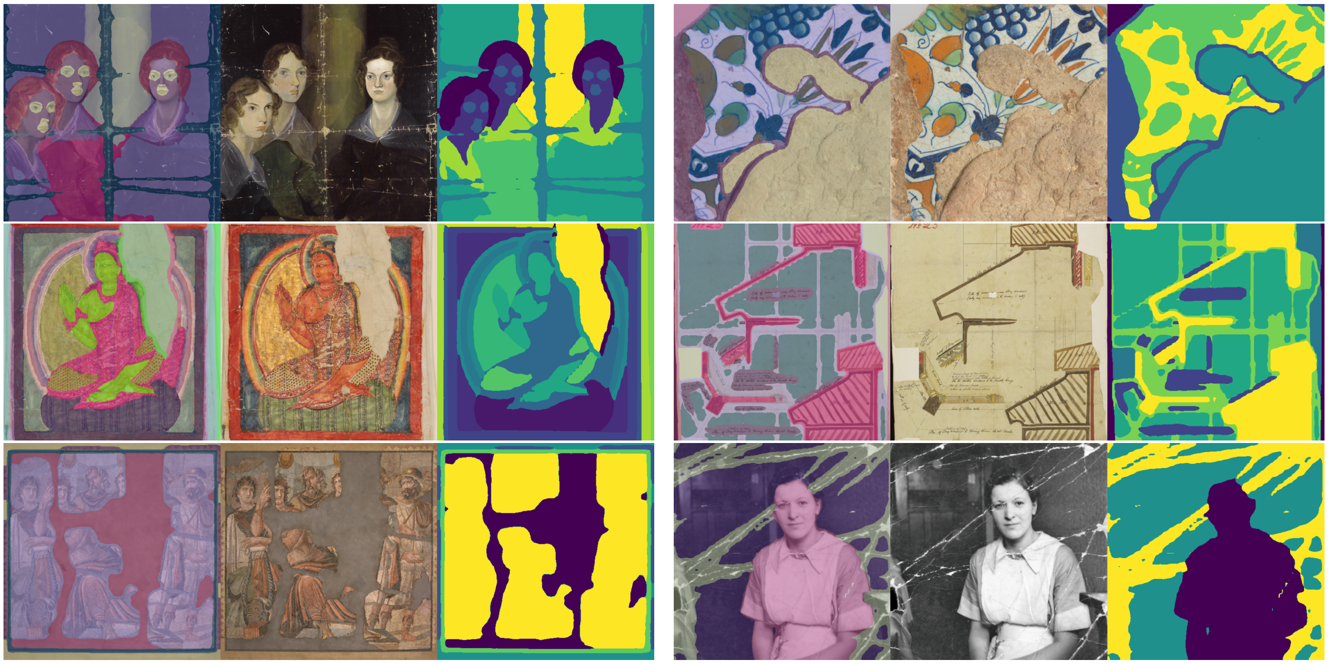

Figure 12. Segmentations of varying granularities on different types of damaged analogue media [24].

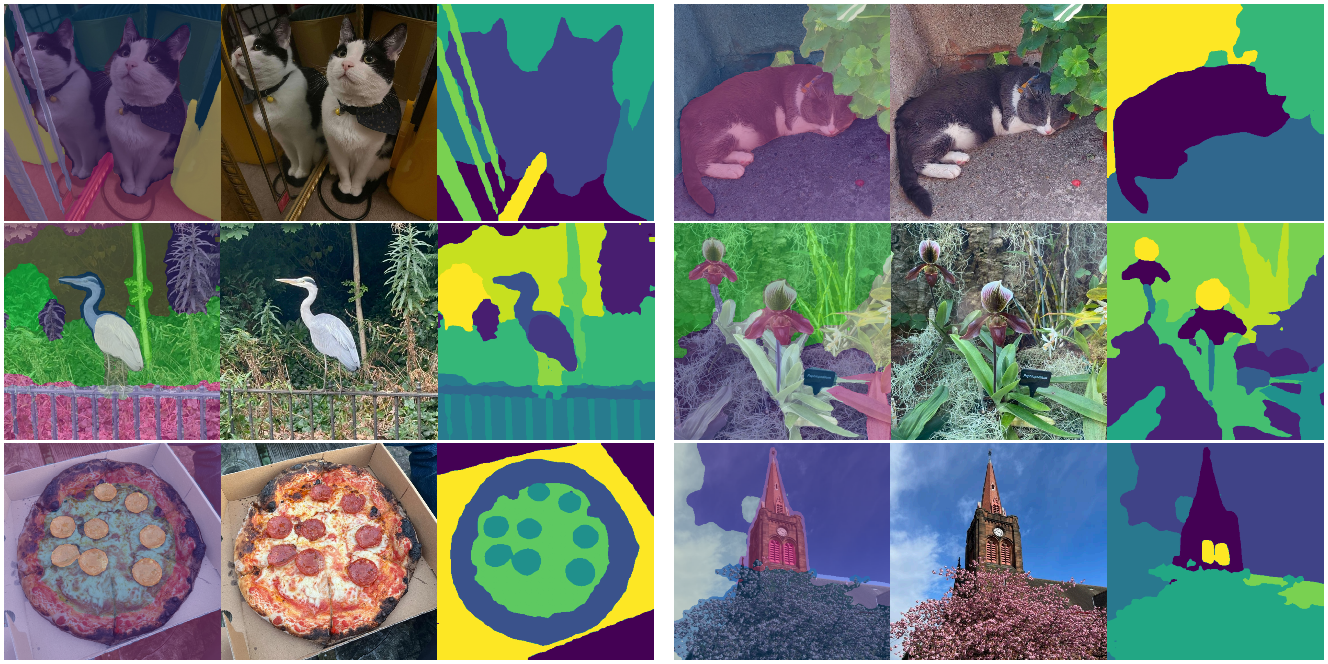

Figure 13. Segmentations of varying granularities of in-the-wild images captured by a smartphone: overlay (left), input (middle), and segmentation (right).