

## A APPENDIX

### A.1 DATASETS USED FOR EVALUATION

**COCO** Lin et al. (2014) (Microsoft Common Objects in Context) is a large-scale dataset for object detection and segmentation. In this paper, COCO refers to the 5k images from the 2017 validation set.

**COCO 20K** Lin et al. (2014) contains 19,817 images, a subset of COCO train2014. Many previous unsupervised methods Wang et al. (2022b; 2023); Arica et al. (2024) have used this dataset to evaluate model performance.

**LVIS** Gupta et al. (2019): (Large Vocabulary Instance Segmentation) is a dataset for long-tail instance segmentation. It contains 2.2 million high-quality instance masks of over 1,000 entry-level object categories, collected based on the COCO dataset. In this paper, LVIS refers to the 19,809 images in the validation set.

**VOC** Everingham et al. (2010) (PASCAL Visual Object Classes) is a widely used benchmark for object detection. We evaluate on its trainval07 split.

**KITTI** Geiger et al. (2012) (Karlsruhe Institute of Technology and Toyota Technological Institute) is one of the most popular datasets for mobile robotics and autonomous driving. We evaluate on its trainval split.

**OpenImages V7** Kuznetsova et al. (2020) contains multiple tasks, including image classification, object detection, instance segmentation, and visual relationship detection. We evaluate on over 40K images from the val split.

**Object365 V2** Shao et al. (2019) provides a supervised object detection benchmark with a focus on diverse objects in the natural world. We evaluate on 80K images from the val split.

The summary of these datasets used for zero-shot evaluation is provided in Table 1.

Table 1: **Summary of datasets used for zero-shot evaluation (except ImageNet).** "avg. # obj." denotes the average number of annotations per image.

datasets	testing data	seg label	#images	avg. # obj.
COCO	val2017	✓	5,000	7.4
COCO20K	train2014	✓	19,817	7.3
LVIS	val	✓	19,809	12.4
Pascal VOC	trainval07	×	9,963	3.1
KITTI	trainval	×	7,521	4.7
OpenImages V7	val	×	41,620	7.3
Object365 V2	val	×	80,000	15.5
ImageNet	val	×	50,000	1.6

Table 2: **Unsupervised instance segmentation results on all benchmarks in this work.**

Datasets	$AP^{\text{mask}}$	$AP_{50}^{\text{mask}}$	$AP_{75}^{\text{mask}}$	$AP_S^{\text{mask}}$	$AP_M^{\text{mask}}$	$AP_L^{\text{mask}}$	$AR_1^{\text{mask}}$	$AR_{10}^{\text{mask}}$	$AR_{100}^{\text{mask}}$	$AR_S^{\text{mask}}$	$AR_M^{\text{mask}}$	$AR_L^{\text{mask}}$
COCO	9.6	20.1	8.5	2.3	10.5	21.5	5.6	16.3	25.4	9.5	31.2	44.9
COCO20K	9.8	20.5	8.4	2.6	10.5	21.6	5.6	16.5	25.6	9.7	31.6	44.7
LVIS	3.7	7.3	3.2	1.5	6.9	12.0	2.1	7.9	16.1	6.3	29.2	41.8

Table 3: **Unsupervised object detection results on all benchmarks in this work.**

Datasets	$AP^{\text{box}}$	$AP_{50}^{\text{box}}$	$AP_{75}^{\text{box}}$	$AP_S^{\text{box}}$	$AP_M^{\text{box}}$	$AP_L^{\text{box}}$	$AR_1^{\text{box}}$	$AR_{10}^{\text{box}}$	$AR_{100}^{\text{box}}$	$AR_S^{\text{box}}$	$AR_M^{\text{box}}$	$AR_L^{\text{box}}$
COCO	12.5	23.8	11.9	4.0	13.5	27.7	6.6	19.8	32.0	13.2	38.7	55.8
COCO20K	12.6	24.1	11.8	4.3	13.3	27.6	6.6	20.0	32.2	13.5	38.9	55.8
LVIS	4.6	9.2	4.1	2.4	8.6	15.5	2.4	9.5	20.0	8.7	34.9	51.6
VOC	20.5	39.1	19.6	2.7	8.2	31.5	15.9	33.2	44.6	19.3	36.5	54.5
KITTI	8.8	20.8	6.2	1.2	6.2	17.4	6.3	19.7	29.7	17.4	26.8	42.1
OpenImages	9.3	16.7	9.2	0.2	1.9	14.6	6.6	16.5	27.1	4.1	19.6	34.6
Objects365	11.2	22.6	9.8	2.6	10.5	19.1	2.8	15.0	32.0	11.7	34.6	45.9

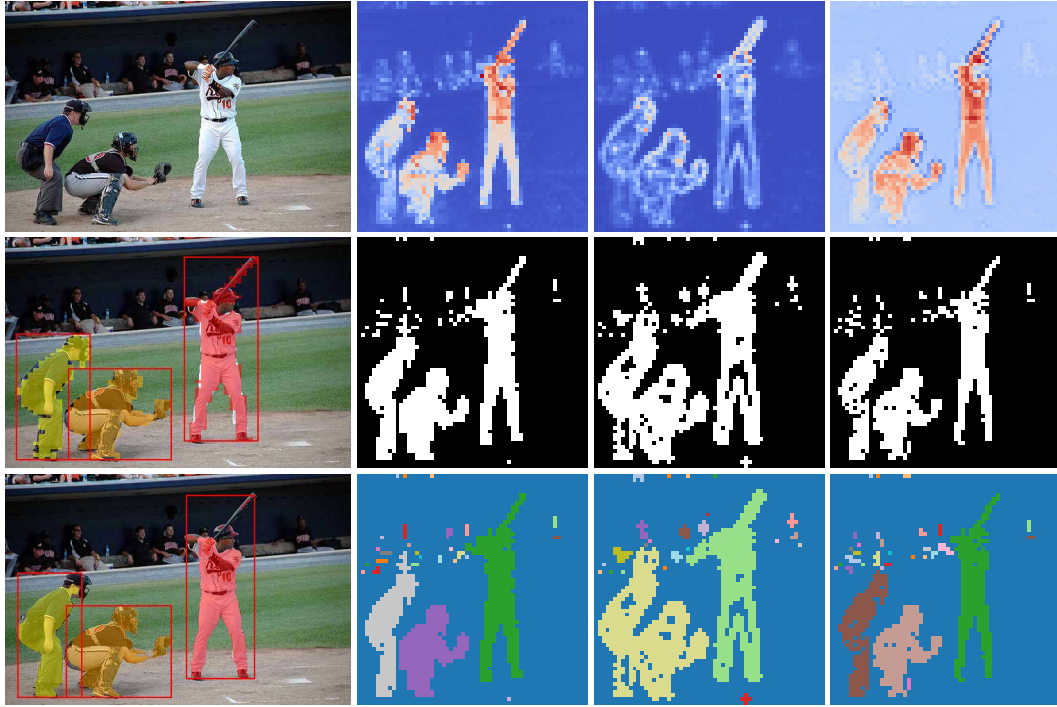


Figure 1: **Visualization of the computation process of CutOnce.** The first column (top to bottom) shows the original image, CutOnce prediction, and CutOnce with post-processing. The second to fourth columns show *raw eigenvector*, *boundary eigenvector*, *difference between the two*, and the corresponding foreground-background *binary maps* and *connected component maps*.

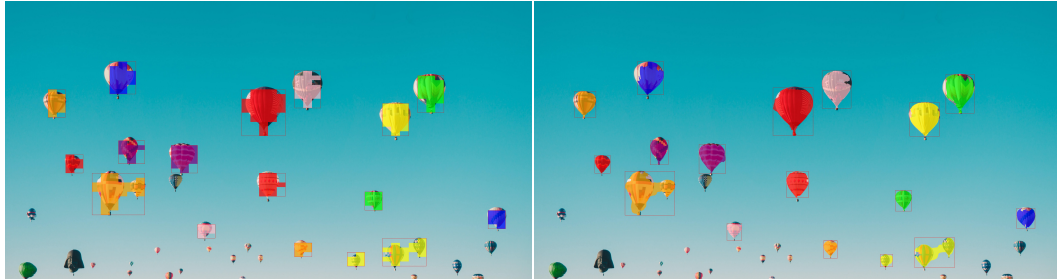


Figure 2: **Detecting many objects with CutOnce.** The first row shows the results of CutOnce, and the second row presents the results of CutOnce with post-processing. Both methods successfully detect 17 objects.

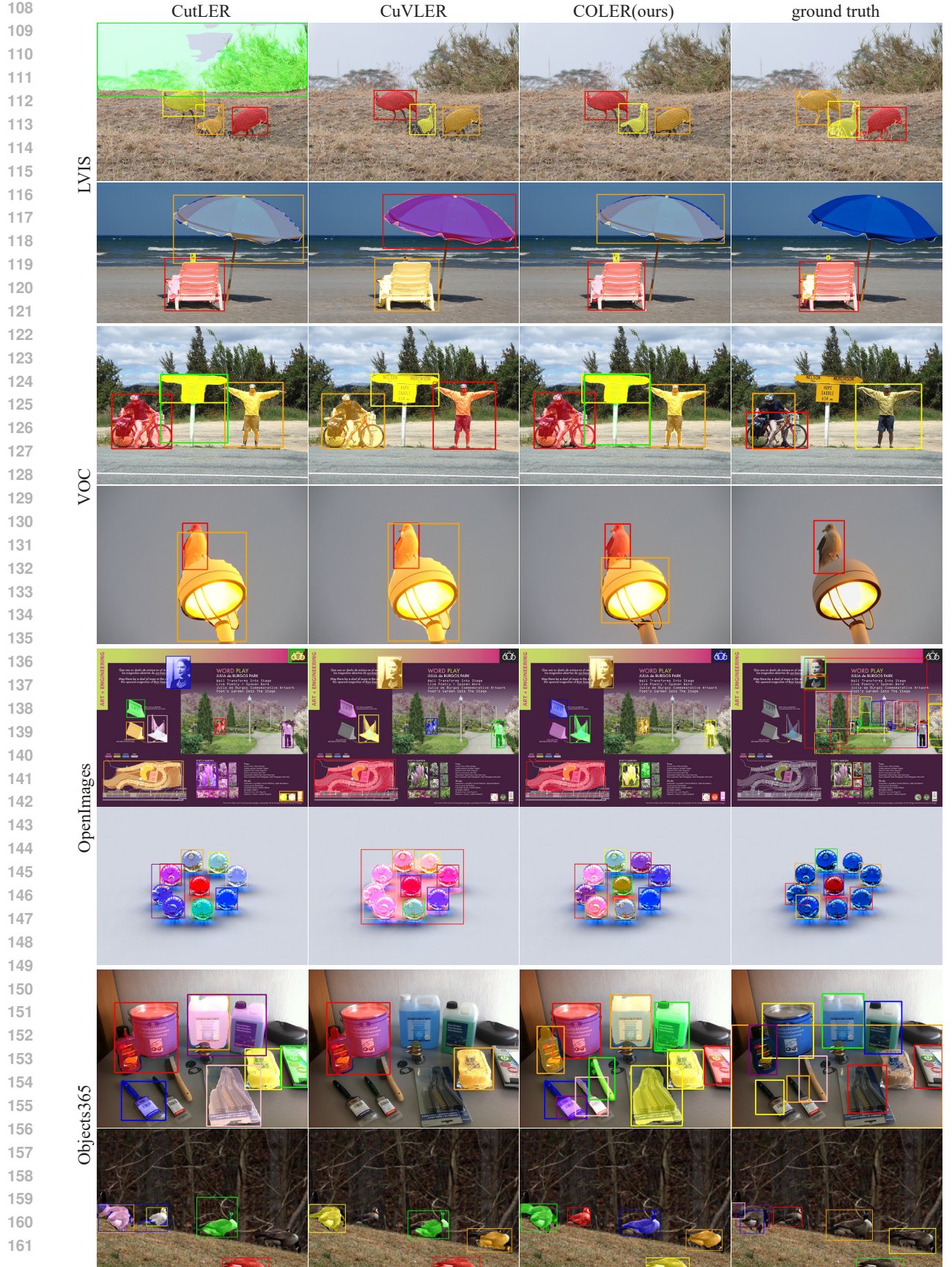


Figure 3: Qualitative comparison of our COLER previous SOTA methods on LVIS, VOC, OpenImages, and Objects365.





Figure 4: Qualitative comparison of our COLER with previous SOTA methods on KITTI.

## A.2 OTHER VISUALIZATIONS

Figure 1 shows a visualization of the intermediate computation process of *CutOnce*'s *boundary enhancement module*. Obviously, the mechanism of this module is easy to understand and shows immediate effectiveness.

Figure 2 demonstrates the strong capability of *CutOnce* in segmenting multiple objects, which previous methods were unable to detect in such quantity.

Table 2 and Table 3 present the zero-shot evaluation results on unsupervised instance segmentation and object detection tasks across various datasets, respectively.

Figure 4 and Figure 3 show additional visualization results of our COLER method compared to previous state-of-the-art approaches. These figures only display predicted results with a confidence score of *no less than 0.5* (ground truth is excluded).