CuVLER: Enhanced Unsupervised Object Discoveries through Exhaustive Self-Supervised Transformers

Supplementary Material

7. Additional details

7.1. Utilized models

- DINO ViT-B/8 [8]
- DINO ViT-B/16 [8]
- DINO ViT-S/8 [8]
- DINO ViT-S/16 [8]
- DINOv2 ViT-S/14 [27]
- DINOv2 ViT-B/14 [27]

7.2. Ablation results

In Table 6, Table 7 and Table 8 we present the specific results used to create Figure 3, Figure 4 and Figure 5, respectively.

Model count	AP^{box}	AP^{box}_{50}	AP^{box}_{75}	AR^{box}_{100}
1	7.4	15.3	6.2	29.4
2	14.2	27	12.7	38
3	15.3	29.3	13.8	38.7
4	17.4	29.3	13.8	38.7
5	19.6	35.0	18.4	44
6	**20.9**	**36.3**	**20.1**	**44.9**

Table 6. Model count ablation results

k_{max}	AP^{box}	AP^{box}_{50}	AP^{box}_{75}	AR^{box}_{100}
2	19.5	33.5	19	41.9
3	**20.9**	**36.3**	**20.1**	44.9
4	20.4	36.1	19.3	**45.2**
5	19.4	35	18.1	44.9

Table 7. k_{max} ablation results

τ^m	AP^{box}	AP^{box}_{50}	AP^{box}_{75}	AR^{box}_{100}
0.1	17.7	30.6	17.1	37.2
0.2	**20.9**	**36.3**	**20.1**	**44.9**
0.3	20.3	35.8	19.2	44.4
0.4	19.2	35.1	17.8	43.3
0.5	17.6	33.9	15.7	41.3

Table 8. τ^m ablation results

8. Datasets

COCO and COCO20K [25] is a large-scale instance segmentation and object detection dataset containing approximately $118K$ images for training and another $5K$ for validation. Additionally, COCO has an unannotated split of $123K$ images. COCO `20K` is a subset of the COCO `trainval2014` [25], accommodated from 19817 randomly sampled images, used for evaluation in [29, 32, 34, 35]. We evaluate our models in a class-agnostic manner on COCO `val2017` and COCO `20K`. We use COCO-style average precision (AP) and average recall (AR) from object detection and segmentation tasks for evaluation.

Pascal VOC [16] is an object detection dataset widely used as a benchmark.

OpenImages V6 [24] Unifies instance segmentation, object detection and image classification, visual relationship detection, and more, in a single dataset. We evaluate our method on its $42K$ images from the `val` split.

Clipart1k [22] , which we refer to as 'Clipart', is an object detection dataset consisting of 1000 images from a clip art domain. We evaluate our model using all annotated images from this dataset, $traintest$.

Watercolor2K [22] , which we refer to as 'Watercolor', is an object detection dataset consisting of 2000 images from a watercolor painting domain. We evaluate our model using all annotated images from this dataset, $traintest$.

Comic2K [22] , which we refer to as 'Comic', is an object detection dataset consisting of 2000 images from a comic domain. We evaluate our model using all annotated images from this dataset, $traintest$.

LVIS [19] (Large Vocabulary Instance Segmentation) is a large-scale dataset comprising $164K$ images featuring $2.2M$ high-quality instance segmentation masks. It covers over 1000 entry-level object categories, naturally forming a long-tail distribution of categories.

9. Addtinal results

On Table 9, we can see additional results of a zero-shot evaluation of our CuVLER method and the results of CutLER, the previous SOTA. This table complement Table 2, with additional AR^{box}_{100} and AP^{box}_{75} metrics. We can see that our model suppressed CutLER across all the datasets in almost all metrics.

In Figure 6, we can visually observe the capabilities of VoteCut in generating superior detection and mask proposals, surpassing the previous SOTA methods. It is noteworthy that VoteCut succeeds in discovering multiple objects and assigns a confidence score to each proposal, a significant feature leveraged for subsequent training.

Figure 6. Additinoal Visual illustration of VoteCut performance vs. SOTA NCut based object-discovery methods on the ImageNet validation set. The VoteCut bounding box score is calculated according to Eq. (4)

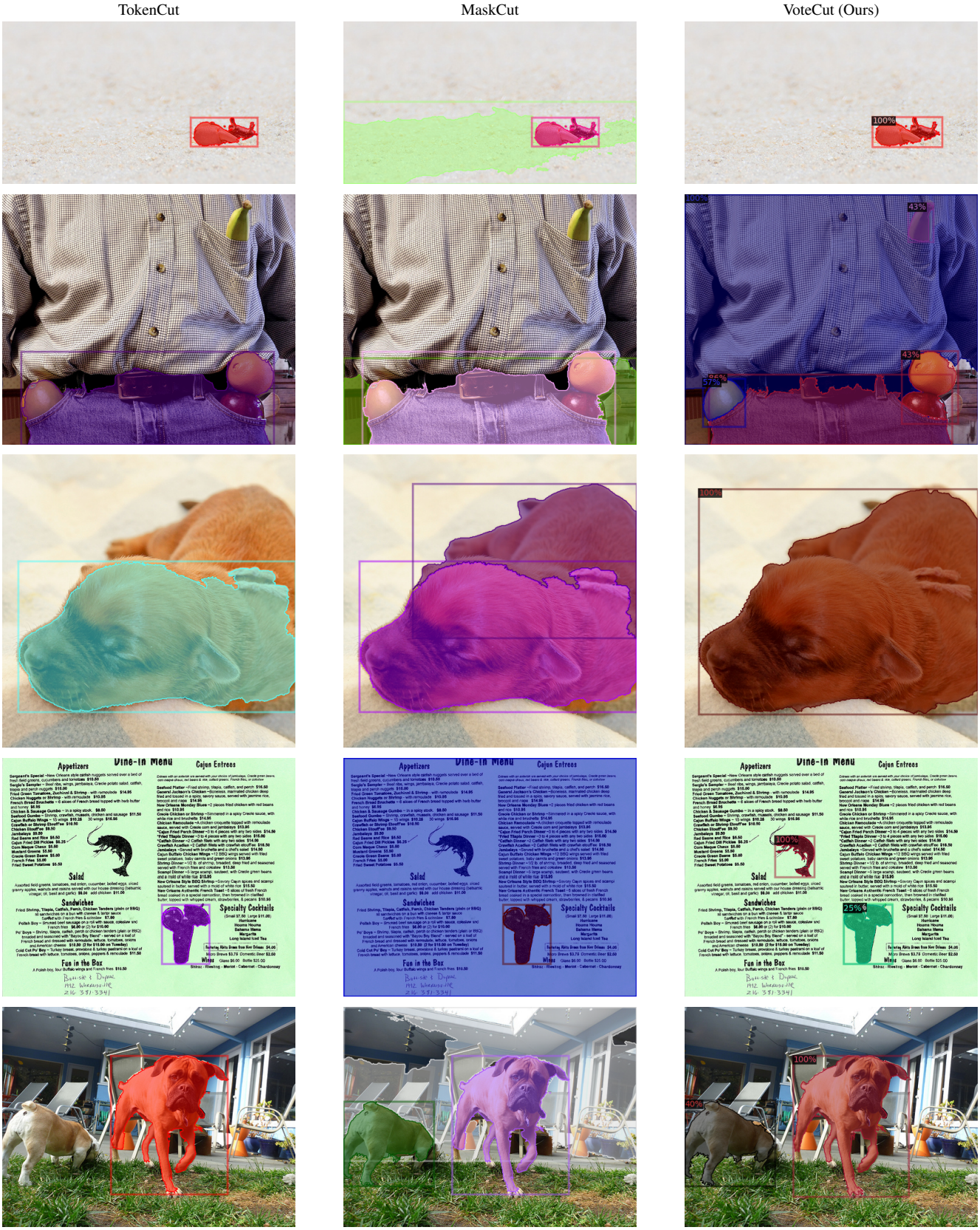

Figure 6. Additinoal Visual illustration of VoteCut performance vs. SOTA NCut based object-discovery methods on the ImageNet validation set. The VoteCut bounding box score is calculated according to Eq. (4)

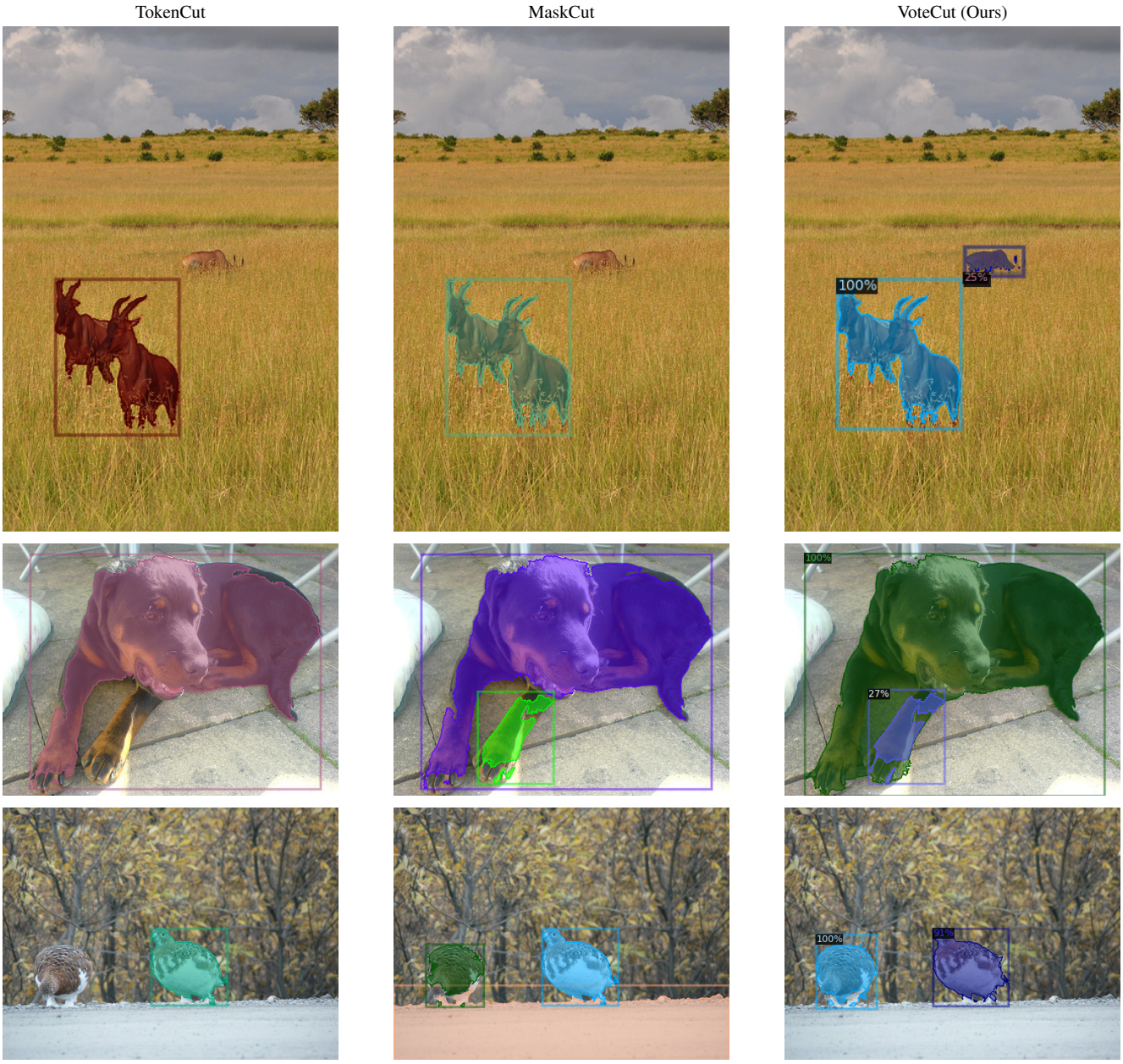

Figure 6. Additinoal Visual illustration of VoteCut performance vs. SOTA NCut based object-discovery methods on the ImageNet validation set. The VoteCut bounding box score is calculated according to Eq. (4)

Method	COCO		COCO20K		VOC		OpenImages		Clipart		Watercolor		Comic	
	AP_{75}	AR	AP_{75}	AR	AP_{75}	AR	AP_{75}	AR	AP_{75}	AR	AP_{75}	AR	AP_{75}	AR
Prev. SOTA [34]	11.8	32.8	11.9	33.1	19.2	44	9.5	29.6	6	40.7	10.9	44.2	7.7	38.4
CuVLER (ours)	11.8	32.8	11.9	33	21.2	46.5	11.4	30.5	6.9	42	15.9	47.4	11.5	41
vs. prev. SOTA	+0	+0	+0	-0.1	**+2**	**+2.5**	**+1.9**	**+0.9**	**+0.9**	**+1.3**	**+5**	**+3.2**	**+3.8**	**+2.6**

Table 9. SOTA zero-shot unsupervised object detection performance on seven datasets. The reported results are based on the COCO metrics, encompassing both Average Recall (AR) and AP_{75} scores. The presented models are trained in an unsupervised manner solely on ImageNet. Results of [34] are produced with official code and checkpoint. AR refers to AR^{box}_{100} metric.