Figures and Tables

Table 1: Quantitative results on COCO* validation dataset based on object count. "# of GT obj." refers to the average number of ground truth objects per image, while "# of pred obj." refers to that of predicted objects.

# of GT obj.		Direct Object Discovery										
		AP^{box}_{50}	AP^{box}_{75}	AP^{box}	AR^{box}_{100}	AR^{box}	AP^{mask}_{50}	AP^{mask}_{75}	AP^{mask}	AR^{mask}_{100}	AR^{mask}	# pred obj
[0,4]	MaskCut (K=3)	25.1	12.3	13.3	28.5	28.5	22.5	8.9	10.6	24.2	24.2	1.8
	MaskCut (K=10)	24.5	11.7	12.9	29.3	29.3	21.9	8.8	10.3	24.8	24.8	1.9
	VoteCut	38.9	21.1	22.0	**39.1**	**39.1**	37.0	17.4	19.0	34.3	34.3	8.5
	OCN$_{disc}$ (ours)	**42.1**	**21.2**	**23.2**	38.7	38.7	**42.1**	**22.0**	**22.8**	**37.8**	**37.8**	5.9
[5,9]	MaskCut (K=3)	10.7	4.8	5.3	10.4	10.4	9.4	3.4	4.3	9.0	9.0	1.9
	MaskCut (K=10)	11.4	5.0	5.5	11.4	11.4	9.6	3.6	4.4	9.9	9.9	2.1
	VoteCut	17.2	7.6	8.6	17.5	17.5	15.4	6.7	7.4	15.0	15.0	9.0
	OCN$_{disc}$ (ours)	**25.2**	**12.7**	**13.7**	**25.8**	**25.8**	**24.0**	**12.3**	**12.8**	**24.2**	**24.2**	7.9
[10,14]	MaskCut (K=3)	5.1	2.3	2.7	4.9	4.9	4.4	1.6	1.8	4.3	4.3	1.9
	MaskCut (K=10)	5.4	2.4	2.7	5.5	5.5	4.7	1.4	1.9	4.8	4.8	2.3
	VoteCut	8.8	3.0	3.9	9.5	9.5	7.2	3.6	3.3	8.1	8.1	9.2
	OCN$_{disc}$ (ours)	**18.0**	**8.2**	**9.4**	**19.3**	**19.3**	**16.7**	**7.7**	**8.6**	**18.5**	**18.5**	9.4
[15, +)	MaskCut (K=3)	1.8	0.5	0.7	1.9	1.9	1.6	0.4	0.6	1.6	1.6	2.0
	MaskCut (K=10)	1.7	0.5	0.8	2.1	2.1	1.5	0.4	0.7	1.9	1.9	2.3
	VoteCut	4.2	1.4	1.8	4.6	4.6	3.2	1.2	1.4	4.0	4.0	9.3
	OCN$_{disc}$ (ours)	**13.6**	**6.5**	**7.1**	**14.0**	**14.0**	**12.3**	**5.6**	**6.3**	**13.4**	**13.4**	12.2

# of GT obj.		Training Detectors										
		AP^{box}_{50}	AP^{box}_{75}	AP^{box}	AR^{box}_{100}	AR^{box}	AP^{mask}_{50}	AP^{mask}_{75}	AP^{mask}	AR^{mask}_{100}	AR^{mask}	# pred obj
[0,4]	UnSAM	15.5	10.5	10.2	**66.4**	**73.3**	15.9	10.4	10.1	**60.1**	**65.5**	244.1
	CutLER	55.2	35.4	34.4	61.2	61.2	51.2	29.0	28.6	52.4	52.4	100.0
	CuVLER	**56.9**	**36.1**	**35.1**	60.5	60.5	**53.6**	**30.1**	**29.9**	52.6	52.6	99.9
	OCN (ours)	55.3	33.7	33.1	59.6	59.6	52.9	29.5	29.5	52.8	52.8	100.0
[5,9]	UnSAM	13.3	8.6	8.7	**49.9**	**62.0**	13.5	8.5	8.5	46.2	**56.3**	317.9
	CutLER	37.7	21.4	21.7	49.3	49.3	33.3	16..7	17.5	42.5	42.5	100.0
	CuVLER	39.0	21.1	21.9	48.4	48.4	34.1	16.6	17.8	41.3	41.3	99.6
	OCN (ours)	**40.8**	**21.8**	**22.8**	**49.9**	49.9	**37.0**	**18.6**	**19.6**	**44.4**	44.4	100.0
[10,14]	UnSAM	11.2	6.7	6.9	38.5	**52.7**	11.3	6.7	6.8	36.6	**48.6**	378.1
	CutLER	26.3	13.2	14.3	40.3	40.3	22.8	10.2	11.5	34.9	34.9	100.0
	CuVLER	28.5	13.7	15.2	39.7	39.7	24.9	11.1	12.4	34.0	34.0	99.7
	OCN (ours)	**33.4**	**16.7**	**17.9**	**43.2**	43.2	**30.5**	**14.3**	**15.7**	**38.6**	38.6	100.0
[15, +)	UnSAM	8.9	5.2	5.5	24.8	**40.9**	8.6	4.9	5.2	24.2	**38.2**	475.7
	CutLER	19.3	9.0	10.0	29.0	29.0	15.5	6.7	7.6	25.1	25.1	100.0
	CuVLER	21.4	9.7	10.9	28.4	28.4	17.0	7.1	8.3	24.4	24.4	99.6
	OCN (ours)	**29.0**	**14.2**	**15.3**	**33.4**	33.4	**24.8**	**11.0**	**12.5**	**30.0**	30.0	100.0

Table 2: Quantitative comparison between OCN$_{disc}$ and rough masks from VoteCut. Evaluation for ImageNet val only includes metrics on bounding box predictions, as ImageNet does not have ground truth masks.

	COCO train2017										ImageNet val									
	AP^{box}_{50}	AP^{box}_{75}	AP^{box}	AR^{box}_{100}	AR^{box}	AP^{mask}_{50}	AP^{mask}_{75}	AP^{mask}	AR^{mask}_{100}	AR^{mask}	AP^{box}_{50}	AP^{box}_{75}	AP^{box}	AR^{box}_{100}	AR^{box}	AP^{mask}_{50}	AP^{mask}_{75}	AP^{mask}	AR^{mask}_{100}	AR^{mask}
VoteCut	11.1	4.7	5.5	12.2	12.2	9.7	3.8	4.6	10.3	10.3	36.2	**20.0**	20.9	**45.0**	**45.0**	-	-	-	-	-
OCN$_{disc}$	**15.7**	**6.7**	**7.8**	**18.3**	**18.3**	**15.2**	**6.5**	**7.4**	**17.5**	**17.5**	**38.9**	19.8	**21.1**	44.4	44.4	-	-	-	-	-

Table 3: Training and inference time of different methods. For a fair comparison, all methods are evaluated on the same hardware configurations.

	Training Time (hours in total)				Inference Efficiency (seconds per image)			
Direct Object Discovery	MaskCut (N=3)	MaskCut (N=10)	VoteCut	OCN$_{disc}$	MaskCut (N=3)	MaskCut (N=10)	VoteCut	OCN$_{disc}$
	-	-	-	10.1	11.3	33.7	5.1	45.3
Training Detectors	UnSAM	CutLER	CuVLER	OCN	UnSAM	CutLER	CuVLER	OCN
	90.0	75.0	60.0	30.0	3.0	0.1	0.1	0.1

Table 4: Quantitative results of zero-shot detection. Each method uses its best model in Group 3. Since KITTI/ VOC/ Object365/ OpenImages datasets do not have ground truth masks, only bounding box metrics are calculated.

	COCO20K									
	AP^{box}_{50}	AP^{box}_{75}	AP^{box}	AR^{box}_{100}	AR^{box}	AP^{mask}_{50}	AP^{mask}_{75}	AP^{mask}	AR^{mask}_{100}	AR^{mask}
UnSAM	6.3	3.2	3.4	29.7	**42.5**	6.3	3.1	3.3	27.5	**38.0**
CutLER	22.4	11.9	12.5	33.1	33.1	19.6	9.2	10.0	27.2	27.2
CuVLER	24.1	12.3	13.1	32.6	32.6	21.1	9.7	10.7	27.2	27.2
OCN (ours)	**25.9**	**13.0**	**13.9**	**35.4**	35.4	**23.6**	**11.1**	**12.0**	**30.5**	30.5

	LVIS									
	AP^{box}_{50}	AP^{box}_{75}	AP^{box}	AR^{box}_{100}	AR^{box}	AP^{mask}_{50}	AP^{mask}_{75}	AP^{mask}	AR^{mask}_{100}	AR^{mask}
UnSAM	4.4	2.5	2.7	23.1	**35.7**	4.5	2.8	2.8	22.9	**34.2**
CutLER	8.5	3.9	4.5	21.8	21.8	6.7	3.2	3.5	18.7	18.7
CuVLER	8.9	4.1	4.7	20.8	20.8	7.2	3.4	3.8	17.9	17.9
OCN (ours)	**10.4**	**5.0**	**5.6**	**24.1**	24.1	**8.9**	**4.5**	**4.9**	**21.4**	21.4

	KITTI									
	AP^{box}_{50}	AP^{box}_{75}	AP^{box}	AR^{box}_{100}	AR^{box}	AP^{mask}_{50}	AP^{mask}_{75}	AP^{mask}	AR^{mask}_{100}	AR^{mask}
UnSAM	1.9	0.6	0.8	17.0	21.7	-	-	-	-	-
CutLER	20.8	7.4	9.5	28.9	28.9	-	-	-	-	-
CuVLER	18.8	5.9	8.0	27.9	27.9	-	-	-	-	-
OCN (ours)	**26.7**	**12.6**	**13.7**	**34.8**	**34.8**	**-**	**-**	**-**	**-**	**-**

	VOC									
	AP^{box}_{50}	AP^{box}_{75}	AP^{box}	AR^{box}_{100}	AR^{box}	AP^{mask}_{50}	AP^{mask}_{75}	AP^{mask}	AR^{mask}_{100}	AR^{mask}
UnSAM	5.1	2.3	2.6	38.8	**51.9**	**-**	**-**	**-**	**-**	**-**
CutLER	36.8	19.3	20.2	44.0	44.0	-	-	-	-	-
CuVLER	39.4	20.1	21.5	43.7	43.7	-	-	-	-	-
OCN (ours)	**40.4**	**21.5**	**22.7**	**47.4**	47.4	-	-	-	-	-

	Object365									
	AP^{box}_{50}	AP^{box}_{75}	AP^{box}	AR^{box}_{100}	AR^{box}	AP^{mask}_{50}	AP^{mask}_{75}	AP^{mask}	AR^{mask}_{100}	AR^{mask}
UnSAM	9.1	4.9	5.3	30.5	**47.9**	**-**	**-**	**-**	**-**	**-**
CutLER	21.7	10.3	11.5	34.2	34.2	-	-	-	-	-
CuVLER	21.9	9.4	10.9	32.5	32.5	-	-	-	-	-
OCN (ours)	**24.7**	**11.0**	**12.4**	**35.9**	35.9	-	-	-	-	-

	OpenImages									
	AP^{box}_{50}	AP^{box}_{75}	AP^{box}	AR^{box}_{100}	AR^{box}	AP^{mask}_{50}	AP^{mask}_{75}	AP^{mask}	AR^{mask}_{100}	AR^{mask}
UnSAM	6.6	3.7	4.0	**34.6**	**48.7**	**-**	**-**	**-**	**-**	**-**
CutLER	17.2	9.5	9.7	29.6	29.6	-	-	-	-	-
CuVLER	18.6	**11.3**	**11.4**	29.8	29.8	-	-	-	-	-
OCN (ours)	**19.0**	10.9	11.2	29.5	29.5	-	-	-	-	-

Table 5: Quantitative results on COCO* validation dataset. "# of pred obj." refers to the average number of predicted objects per image.

			Trainable Module	AP^{box}_{50}	AP^{box}_{75}	AP^{box}	AR^{box}_{100}	AR^{box}	AP^{mask}_{50}	AP^{mask}_{75}	AP^{mask}	AR^{mask}_{100}	AR^{mask}	# of pred obj.
Direct Object Discovery	w/o Learnable Modules	FreeMask	-	3.7	0.6	1.3	4.6	4.6	3.1	0.3	0.9	3.5	3.5	3.7
		MaskCut (K=3)	-	6.0	2.4	2.9	6.7	6.7	5.1	1.8	2.3	5.8	5.8	1.8
		MaskCut (K=10)	-	6.2	2.6	2.9	7.2	7.2	5.3	2.0	2.3	6.2	6.2	2.1
		VoteCut	-	10.8	4.9	5.5	11.3	11.3	9.5	4.0	4.6	9.8	9.8	8.9
	w/ Learnable Modules	DINOSAUR	Recon. SlotAtt	2.0	0.2	0.6	4.8	4.8	1.1	0.1	0.3	2.9	2.9	7.0
		FOUND	Seg. Head	4.4	1.8	2.1	3.6	3.6	3.3	1.3	1.5	3.0	3.0	1.0
		OCN_{disc}	Obj. Net	19.1	9.0	10.1	19.6	19.6	17.8	8.7	9.5	18.9	18.9	8.2
Training Detectors	-	UnSAM	Detector x 4	10.2	6.3	6.4	36.1	**50.1**	10.2	6.2	6.3	34.1	**46.1**	332.2
		CutLER	Detector x 3	26.0	14.2	14.7	37.9	37.9	22.7	11.2	11.8	32.7	32.7	100.0
		CuVLER	Detector x 2	28.0	14.8	15.5	37.8	37.8	24.4	11.7	12.6	32.1	32.1	99.7
		OCN	**Obj. Network + Detector x 1**	**32.6**	**17.2**	**18.0**	**40.9**	40.9	**29.6**	**14.4**	**15.5**	**36.5**	36.5	100.0

Table 6: Ablation results of different choices of object-centric representations on COCO* validation.

Object Existence	Object Center Field	Object Boundary Distance Field	AP^{box}_{50}	AP^{box}_{75}	AP^{box}	AR^{box}_{100}	AR^{box}	AP^{mask}_{50}	AP^{mask}_{75}	AP^{mask}	AR^{mask}_{100}	AR^{mask}
-	-	-	23.4	10.7	11.8	33.8	33.8	19.6	8.0	9.4	35.7	35.7
✓	-	-	27.2	13.0	14.2	35.6	35.6	23.0	9.8	11.3	30.9	30.9
-	✓	-	29.2	14.9	15.8	37.3	37.3	25.6	11.8	13.0	32.5	32.5
✓	✓	-	29.0	14.4	15.4	36.3	36.3	25.0	11.1	12.5	31.0	31.0
-	-	✓	30.7	16.1	16.9	40.7	40.7	28.1	13.9	14.8	37.0	37.0
✓	-	✓	31.4	16.2	17.1	40.1	40.1	28.4	13.6	14.7	35.9	35.9
-	✓	✓	30.1	16.3	17.0	40.6	40.6	28.3	13.9	14.9	36.8	36.8
✓	✓	✓	**32.6**	**17.2**	**18.0**	**40.9**	**40.9**	**29.6**	**14.4**	**15.5**	**36.5**	**36.5**

Table 7: Quantitative results for video instance segmentation on YouTube-VIS2021 Val# dataset.

	Training Dataset	AP^{box}_{50}	AP^{box}_{75}	AP^{box}	AR^{box}_{100}	AR^{box}	AP^{mask}_{50}	AP^{mask}_{75}	AP^{mask}	AR^{mask}_{100}	AR^{mask}
VideoCutLER	ImageNet	24.6	8.7	11.1	21.7	21.7	23.4	8.0	10.6	20.9	20.9
	YouTubeVIS-2021 Train#	22.9	7.1	10.4	21.2	21.2	22.0	6.6	9.6	19.5	19.5
VideoCuVLER	ImageNet	26.0	11.2	13.3	**26.2**	**26.2**	24.4	10.8	12.4	24.3	24.3
	YouTubeVIS-2021 Train#	29.5	13.5	15.3	26.1	26.1	27.3	10.7	13.3	23.9	23.9
VideoOCN	YouTubeVIS-2021 Train#	**29.7**	**14.5**	**15.6**	25.7	25.7	**29.2**	**13.6**	**14.8**	**24.9**	**24.9**

Table 8: Ablation study for rough masks.

	Rough Masks	SSL features / Supervision	AP^{box}_{50}	AP^{box}_{75}	AP^{box}	AR^{box}_{100}	AR^{box}	AP^{mask}_{50}	AP^{mask}_{75}	AP^{mask}	AR^{mask}_{100}	AR^{mask}
OCN_{disc}	SelfMask	DINO_b16, MoCov2, SwAV	13.2	6.1	4.8	16.4	16.4	12.0	5.0	5.6	15.3	15.3
	MaskCut	DINO_b8	16.3	7.3	6.4	17.7	17.7	14.3	5.7	6.1	18.7	18.7
	VoteCut	DINO_b8, DINO_s8, DINO_b16, DINO_s16, DINOv2_s14, DINOv2_b14	19.1	9.0	10.1	19.6	19.6	17.8	8.7	9.5	18.9	18.9
	VoteCut + SAM	supervised on SA-1B dataset	**21.9**	**9.1**	**10.7**	**19.7**	**19.7**	**18.4**	**9.2**	**9.9**	**19.1**	**19.1**

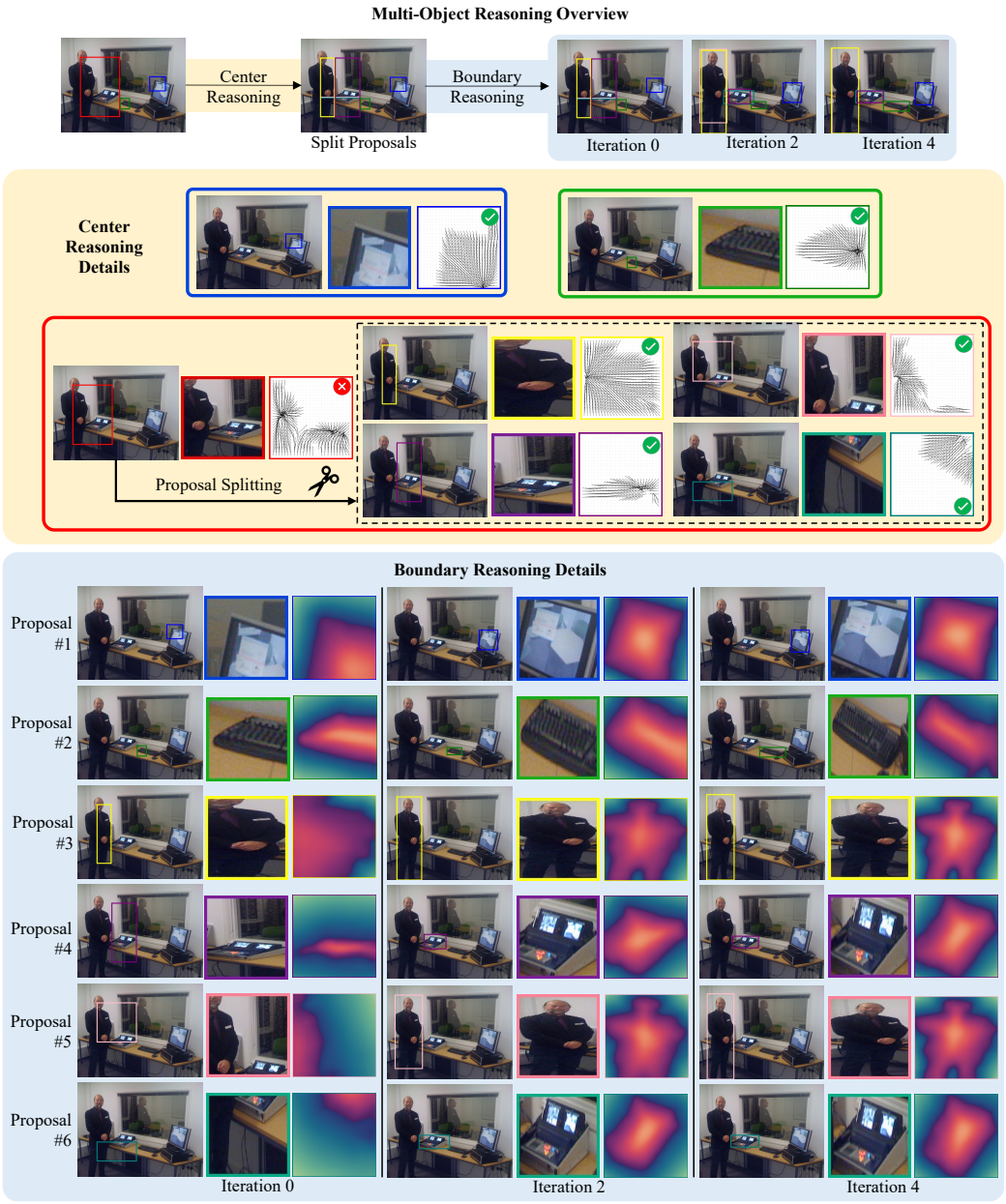

Figure 1: Multi-object reasoning with object center and boundary representations on a multi-object image.

(a) COCO train2017 (b) ImageNet val

Figure 2: Qualitative results of OCN$_{disc}$ and VoteCut on COCO train2017 and ImageNet val. OCN$_{disc}$ has learned (refined) objectness and can better deal with undersegmentation on multi-object images, whereas VoteCut fails to do so.

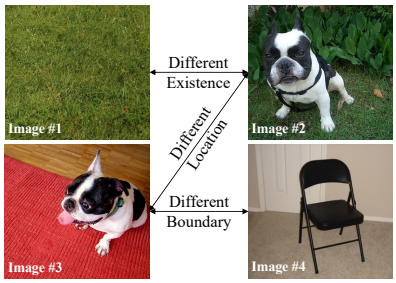

Figure 3: Object images.

Figure 4: Qualitative results of Direct Object Discovery without CAD on COCO* validation set as discussed in Section 4.1 Group 1&2. (Adapted from Figures 11-12 in the original paper.)

Figure 5: Qualitative results from trained detectors on COCO* validation set as discussed in Sec 4.1 Group 3. (Adapted from Figure 13 in the original paper.)

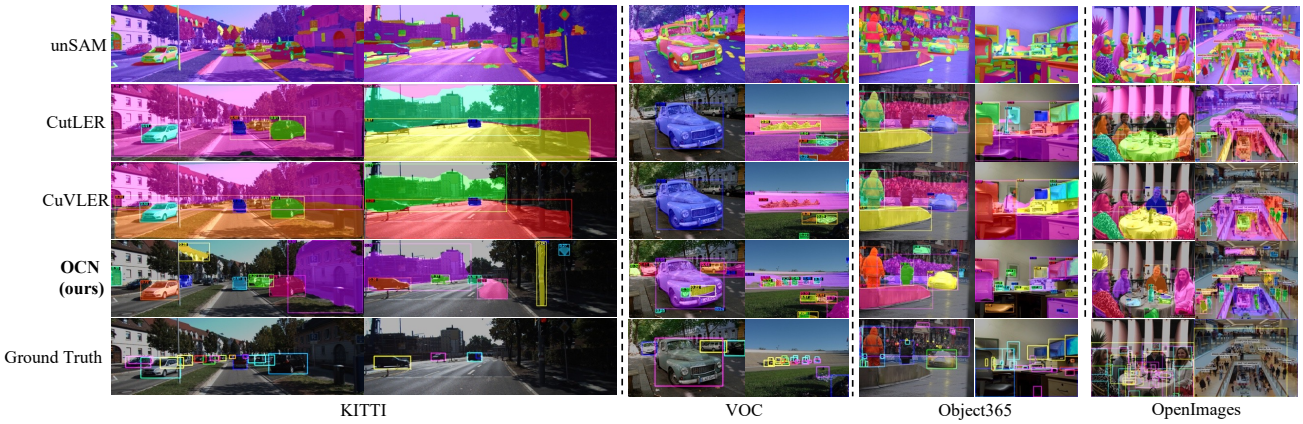

Figure 6: Qualitative results for zero-shot detection as discussed in Sec 4.2. (Adapted from Figures 14-17 in the original paper.)

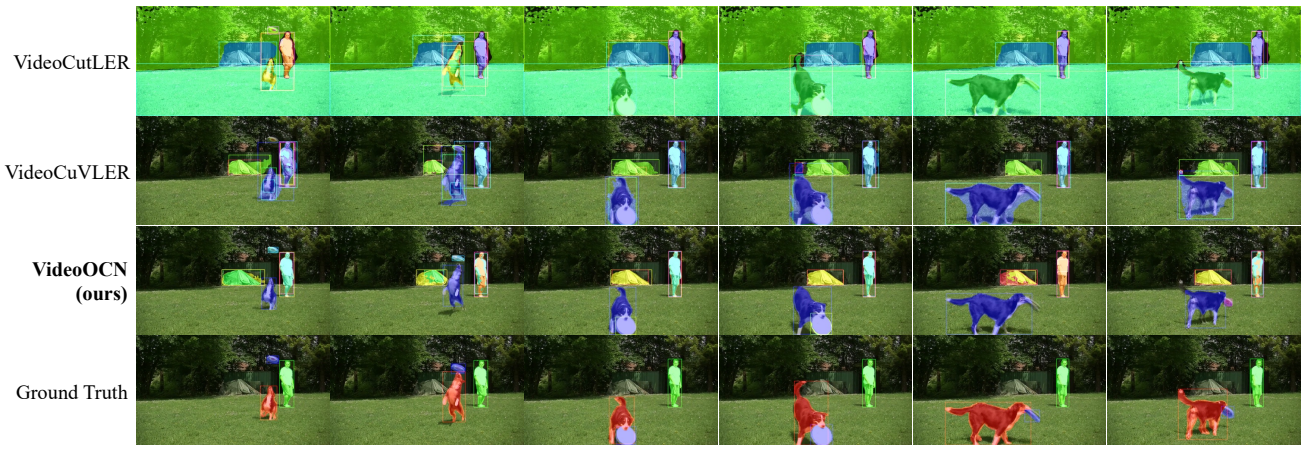

Figure 7: Qualitative results for video instance segmentation on YouTube-VIS2021 Val$^{\#}$ set.

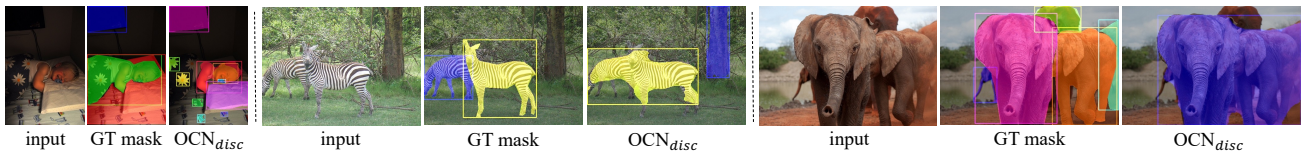

Figure 8: Failure cases of OCN$_{disc}$.