Appendix Materials

A. Datasets used for Evaluation

We provide information about the datasets used in this work as shown in Tab. A1

COCO. The COCO dataset, introduced by [3], is used for object detection and instance segmentation. It has 115,000 training images, 5,000 validation images, and a separate batch of 123,000 unannotated images. We test our unsupervised instance segmentation on the COCO `val2017` set with zero-shot setting. We report results using standard COCO metrics, including average precision and recall for detection and segmentation. Also, for unsupervised universal image segmentation, we test the performance on COCO `val2017`. We report results using panoptic segmentation COCO metrics.

PASCAL VOC. The PASCAL VOC dataset [2] is a widely-used benchmark for object detection. We test our model using the `trainval07` split and adopt COCO-style evaluation metrics.

UVO. The UVO dataset [4] is designed for video object detection and instance segmentation. We test our unsupervised instance segmentation on the UVO `val` split, which includes 256 videos with each one annotated at 30 fps. We remove the extra 5 non-COCO categories which are marked as "other" in their official annotations. For evaluation, we employ COCO-style metrics.

Cityscapes. Cityscapes is a dataset dedicated to semantic urban scene understanding, focusing primarily on semantic segmentation of urban scenes. In our research, we tested our unsupervised universal image segmentation on the Cityscapes `val` splits, using COCO-stype panoptic evaluation metrics.

B. Hungarian Matching for Unsupervised Segmentation Evaluation

In unsupervised object detection and instance segmentation, category IDs are predicted without referencing any predefined labels. For convenience, we differentiate the predicted category ID of U2Seg as "cluster ID" while keep the ground truth category ID as "category ID" in the following analysis. To evaluate the segmentation performance, particularly concerning category accuracy, an optimal correspondence between the cluster ID and the ground truth category ID is essential. We leverage a multi-to-one Hungarian matching for evaluation of U2Seg.

Hungarain Matching. Given a set of predicted bounding boxes, masks associated with predicted cluster IDs and the corresponding ground truth, the objective is to find the best match from "cluster ID" to "category ID". To do this, we first use the predicted confidence score conf as a threshold to filter the predicted instance, removing the ones with low confidence. Then, for each predicted instance with its cluster ID, we calculate the IoU of the predicted bounding box or mask with all ground truth instances, then select the one whose IoU is bigger than the predefined threshold, regarding it as the ground truth category ID for this cluster ID. After we get these cluster ID and ground truth category ID pairs, we form a histogram for each kind of cluster ID based on its overlap with all kinds of ground truth category ID. The ground truth category ID that appears most frequently in this histogram becomes the mapping for this cluster ID. This process may result in multiple predicted cluster IDs being mapped to the same ground truth category ID, leading to a multi-to-one matching scenario.

In our experiment, the confidence score threshold conf to filter the predicted instance and the IoU threshold to match predicted instance with its ground truth instance are both hyperparameters, some ablations can be found in Sec. 4.6.

Evaluation Implications. The multi-to-one Hungarian matching method provides a systematic and efficient way to assess the performance of unsupervised segmentation models. By mapping predicted cluster ID to their most likely ground truth counterparts, the method ensures that the evaluation reflects the true categorization capability of the model. This, in turn, allows for a fair and consistent comparison across different unsupervised segmentation techniques.

C. Unsupervised Instance Segmentation

In this section, we provide complete results for the unsupervised instance segmentation of U2Seg. The results are presented over various datasets and classes to furnish a comprehensive evaluation of our model's capability.

Tab. A2 and Tab. A3 display the results for unsupervised object detection and instance segmentation on different datasets. One trend can be observed across the different datasets: as the number of the predicted cluster ID increases (*e.g.*, moving from 300 to 2911), there is a consistent increase for most of the metrics. This trend can be succinctly attributed to the intrinsic properties of the multi-to-one Hungarian matching approach (we also show the parameter IoU and Conf used for Hungarian matching). With an increase of the cluster numbers, the Hungarian matching has a broader set of predictions to associate with a single label. This inherently increases the chances of having at least one correct prediction for the given label, making the matching process more amenable. In essence, larger cluster numbers afford easier matching, thereby boosting the evaluation metrics.

Furthermore, the qualitative results are shown in Fig. A1, with the samples selected in COCO `val2017` and PASCAL VOC `val2012`. After Hungarian matching, we are able to get the real categories of the predicted instances.

Datasets	Domain	Testing Data	#Images	Instance Segmentation Label
COCO [3]	natural images	`val2017` split	5,000	✓
UVO [4]	video frames	`val` split	21,235	✓
PASCAL VOC [2]	natural images	`trainval07` split	9,963	✗
Cityscapes [1]	urban scenes	`val` split	500	✓

Table A1. **Summary of datasets** used for evaluation.

Datasets	# cluster	IoU	Conf	AP^{box}	AP^{box}_{50}	AP^{box}_{75}	AP^{box}_{S}	AP^{box}_{M}	AP^{box}_{L}	AR^{box}_{1}	AR^{box}_{10}	AR^{box}_{100}
UVO	2911	0.6	0.1	9.7	15.1	9.3	0.6	5.2	14.4	18.0	25.3	25.8
	800	0.4	0.1	6.8	10.8	7.2	0.6	2.9	10.2	17.2	24.5	25.0
	300	0.7	0.1	6.5	9.8	6.5	0.8	2.6	9.2	16.0	22.2	22.6
VOC	2911	0.5	0.2	19.2	31.6	19.7	1.0	6.4	26.6	28.6	44.9	48.3
	800	0.8	0.2	19.0	31.0	19.5	0.6	4.8	26.6	28.8	45.2	48.1
	300	0.8	0.4	18.4	29.6	18.8	0.3	3.8	26.0	27.1	41.0	42.8
COCO	2911	0.5	0.3	8.2	13.3	8.4	1.4	7.0	18.2	14.1	21.4	22.1
	800	0.6	0.4	7.3	11.8	7.5	1.2	5.8	15.8	13.3	20.8	21.5
	300	0.6	0.3	5.7	9.3	5.9	0.5	4.6	12.9	11.9	19.5	20.1

Table A2. **Complete results for unsupervised object detection.** We show results on UVO `val`, PASCAL VOC `val2012` and COCO `val2017`, with corresponding clustering numbers. The IoU and Conf are the Hungarian matching parameter we use for evaluation.

Datasets	# cluster	IoU	Conf	AP^{mask}	AP^{mask}_{50}	AP^{mask}_{75}	AP^{mask}_{S}	AP^{mask}_{M}	AP^{mask}_{L}	AR^{mask}_{1}	AR^{mask}_{10}	AR^{mask}_{100}
UVO	2911	0.6	0.1	8.8	13.9	8.4	0.5	6.4	14.4	16.0	21.7	22.1
	800	0.4	0.1	6.2	9.5	6.0	0.5	2.1	9.8	15.7	20.6	21.0
	300	0.7	0.1	6.1	9.5	5.8	0.7	1.0	8.8	14.1	19.2	19.4
COCO	2911	0.5	0.3	7.3	12.4	7.4	0.8	4.9	17.9	12.8	18.7	19.2
	800	0.6	0.4	6.4	11.2	6.4	0.7	3.7	15.0	11.9	18.0	18.5
	300	0.6	0.3	4.9	8.6	5.0	0.3	2.6	11.8	10.7	16.9	17.3

Table A3. Complete results for **unsupervised instance segmentation.** We show results on UVO `val` and COCO `val2017`, with corresponding clustering numbers. The IoU and Conf is the Hungarian matching parameter we use for evaluation.

Datasets	Pretrain	# Cluster	PQ	PQ^{St}	PQ^{Th}	SQ	SQ^{Th}	SQ^{St}	RQ	RQ^{Th}	RQ^{St}
COCO	IN	300	11.1	9.5	19.3	60.1	60.3	59.0	13.7	11.6	25.0
	IN	800	11.9	10.5	19.6	65.9	67.4	58.2	14.8	12.8	25.3
	COCO	300	15.3	14.2	21.6	66.5	67.2	62.4	19.1	17.5	27.5
	COCO	800	15.5	14.6	20.5	69.7	71.1	62.6	19.1	17.8	26.1
	IN+COCO	300	15.5	14.4	21.2	67.1	67.7	64.3	19.2	17.8	26.9
	IN+COCO	800	16.1	15.1	21.2	71.1	72.5	63.8	19.9	18.6	26.8
Cityscapes	IN	300	15.3	4.1	23.4	48.8	54.7	44.6	19.5	5.4	29.7
	IN	800	15.7	4.3	24.0	46.6	47.5	45.9	19.8	5.5	30.2
	COCO	300	18.4	7.8	26.1	47.4	47.3	47.4	22.6	9.8	31.9
	COCO	800	15.4	5.8	22.3	51.5	62.9	43.2	19.0	7.5	27.4
	IN+COCO	300	16.5	6.2	24.1	44.1	45.2	43.3	20.5	7.9	29.7
	IN+COCO	800	17.6	8.4	24.2	52.7	67.5	42.0	21.7	10.5	29.9

Table A4. Complete results for **unsupervised universal image segmentation.** We show results for different models pretrained on various dataset and test on COCO `val2017`, Cityscapes `val`, with corresponding cluster numbers.

D. Unsupervised Universal Image Segmentation

Our model's performance for unsupervised universal image segmentation closely mirrors the trends observed in instance segmentation. Specifically, as the number of the predicted clusters increases, the performance of the panoptic segmentation also improves. Detailed universal segmentation results are shown in Tab. A4 and Fig. A2.

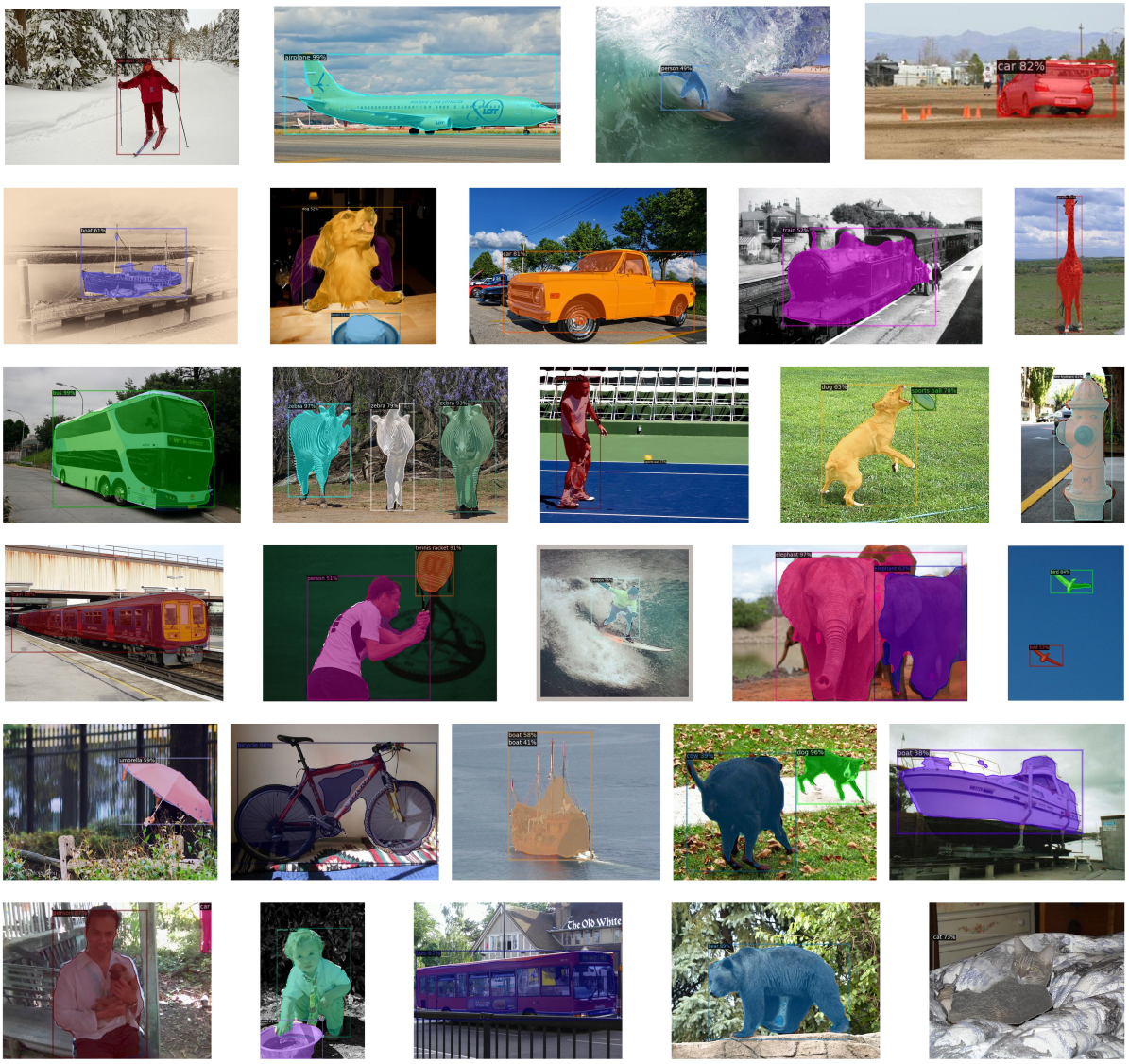

Figure A1. **Unsupervised object detection and instance segmentation visualization** of COCO `val2017` and PASCAL VOC `val2012` (**after Hungarian matching**).

Model	AP^{box}	AP^{box}_{50}	AP^{mask}	AP^{mask}_{50}
CutLER+	5.9	9.0	5.3	8.6
Panoptic	6.1	9.8	5.8	9.0
Instance	7.3	11.8	6.4	11.2

Table A5. **Limitation of U2Seg.** We show the zero-shot unsupervised instance segmentation results on COCO `val2017`. CutLER+ is evaluated on the combination of CutLER and offline clustering, Panoptic is trained on both *"stuff"* and *"things"* pseudo labels, Instance is trained solely on *"things"* labels.

E. Limitation

The primary goal of our research is to develop a comprehensive model capable of excelling in all areas of unsupervised segmentation. As shown in Tab. A5, in terms of the individual sub-task, the universal model exhibits a slight underperformance compared to its counterpart model trained with task-specific annotations. This suggests that U2Seg is adaptable to various tasks, yet it requires task-specific training to achieve the best outcomes for a specific sub-task. Looking ahead, we aim to develop a more versatile model that can be trained once to effectively handle multiple tasks.

Figure A2. **Unsupervised universal image segmentation** visualizations of COCO `val2017` (**after Hungarian matching**).

References

[1] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. *CoRR*, abs/1604.01685, 2016.

[2] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[3] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[4] Weiyao Wang, Matt Feiszli, Heng Wang, and Du Tran. Unidentified video objects: A benchmark for dense, open-world segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10776–10785, 2021.