Scene-Centric Unsupervised Panoptic Segmentation

Oliver Hahn[*1] Christoph Reich[*1,2,4,5] Nikita Araslanov[2,4] Daniel Cremers[2,4,5] Christian Rupprecht[3] Stefan Roth[1,5,6]

[1] TU Darmstadt [2] TU Munich [3] University of Oxford [4] MCML [5] ELIZA [6] hessian.AI [*] equal contribution

CVPR Nashville JUNE 11-15, 2025

Project Page

TL;DR

CUPS — Scene-Centric Unsupervised Panoptic Segmentation — leverages self-supervised representations, motion, and depth from stereo image pairs to generate pseudo labels for training a monocular panoptic network, achieving state-of-the-art results across multiple scene-centric benchmarks.

Figure 1. **CUPS pseudo-label generation and training overview.**

Introduction

Unsupervised Panoptic Segmentation aims to partition images into *semantically meaningful regions* and *detect object instances* without any form of human supervision.

Motivation:

- Mitigate limitations of human-labeled data (*e. g.*, high cost, inherent bias, *etc.*)
- Shift "starting point" for solving tasks with supervision
- Single previous work on unsupervised panoptic segmentation – U2Seg [3] – struggles on scene-centric data due to:
  - MaskCut [7] instance pseudo-labeling requires object-centricity
  - Bypass the classification into "thing" and "stuff" categories
  - Low-resolution semantic pseudo labeling

Goal: Propose the first unsupervised panoptic segmentation method that directly learns from scene-centric data.

References & Acknowledgments

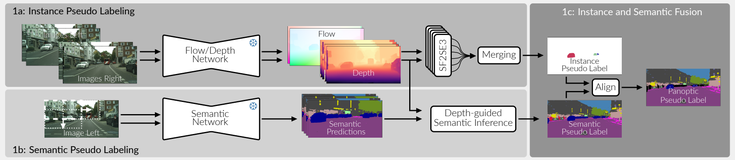
[1] Mark Hamilton et al. Unsupervised semantic segmentation by distilling feature correspondences. In *ICLR*, 2022.
[2] Alexander Kirillov et al. Panoptic feature pyramid networks. In *CVPR*, 2019.
[3] Dantong Niu et al. Unsupervised universal image segmentation. In *CVPR*, 2024.
[4] Leon Sick et al. Unsupervised semantic segmentation through depth-guided feature correlation and sampling. In *CVPR*, 2024.
[5] Leonhard Sommer et al. SF2SE3: Clustering scene flow into SE(3)-motions via proposal and selection. In *GCPR*, 2022.
[6] Yihong Sun and Bharath Hariharan. MOD-UV: Learning mobile object detectors from unlabeled videos. In *ECCV*, 2024.
[7] Xudong Wang et al. Cut and learn for unsupervised object detection and instance segmentation. In *CVPR*, 2023.

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 866008), the ERC Advanced Grant SIMULACRON, the DFG project CR 250/26-1 "4D-YouTube", the GNI Project "AICC", and the State of Hesse within the LOEWE emergenCITY center and the cluster project "The Adaptive Mind". Christoph Reich is supported by the Konrad Zuse School of Excellence in Learning and Intelligent Systems (ELIZA). Finally, we acknowledge the support of the European Laboratory for Learning and Intelligent Systems.

Method

CUPS comprises three stages: *(1)* Generating panoptic pseudo labels; *(2)* bootstrapping a panoptic network with these pseudo labels; and *(3)* self-training of the network.

Stage 1: Pseudo-label generation provides (sparse) high-resolution panoptic pseudo labels for scene-centric data.



1a: Ensembled SF2SE3 [5] motion clustering yields scene-centric instance pseudo labels.

1b: Depth-guided semantic pseudo labels weights sliding window unsupervised semantic segmentation predictions of DepthG [4], obtaining high-resolution semantic pseudo labels.

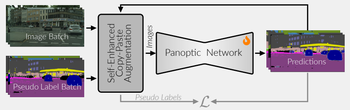
1c: Align semantic prediction inside instance masks. Distinguish between "thing" and "stuff" semantic pseudo classes. Ignore "thing" regions without an instance mask.

Stage 2: Panoptic bootstrapping trains a panoptic network using the pseudo labels.

- Using **DropLoss [7]** we only supervise "thing" instances $\mathbf{R}_j$ with $\mathrm{IoU} > \tau^{\mathrm{IoU}}$ with a pseudo instance mask $\hat{\mathbf{R}}_i$.

$$\mathcal{L}_{\text{drop}}(\mathbf{R}_j, \hat{\mathbf{R}}_i) = \mathbb{1}\left(\mathrm{IoU}_j^{\max} > \tau^{\mathrm{IoU}}\right) \mathcal{L}_{\text{Th}}(\mathbf{R}_j, \hat{\mathbf{R}}_i)$$

- **Self-enhanced copy-paste augmentation:** copy-paste pseudo instance masks and confident model predictions at runtime to augment "thing" masks and discover more static objects.

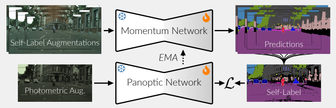


Stage 3: Panoptic self-training applies a teacher-student training on self-labels from ensembling and confidence thresholding of augmented predictions.

- **Self-Labeling:** infer on flipped + multi-scale views, inverse-transform and average soft predictions.
  - *Instance:* keep masks with confidence $\kappa_j > \gamma$ on averaged $\tilde{R} \in [0,1]^{J \times H \times W}$.
  - *Semantic:* for each class k, threshold $\zeta_k = \hat{\zeta} \max(\tilde{P}_k)$ so that

$$L_{h,w}^{\text{sem}} = \begin{cases} \arg\max_k \tilde{P}_{k,h,w}, & \max_k \tilde{P}_{k,h,w} \geq \zeta_k, \\ \text{ignore}, & \text{otherwise.} \end{cases}$$

- **Training:** supervise panoptic network with standard panoptic loss *w. r. t.* self-labels; update student by SGD, update momentum network with EMA; train only heads of the network.



Results

Setup: We train the CUPS Panoptic Cascade Mask R-CNN [2] on pseudo labels generated using Cityscapes training sequences and evaluate on Cityscapes val as well as generalization to five different domains. CUPS predicts semantic pseudo IDs which are aligned to the ground truth via Hungarian matching for evaluation.

Table 1. **Unsupervised panoptic segmentation comparison** using PQ, SQ, and RQ (all in %, ↑).

Method	Cityscapes			KITTI			BDD			MUSES			Waymo			MOTS *(OOD)*		
	PQ	SQ	RQ	PQ	SQ	RQ	PQ	SQ	RQ	PQ	SQ	RQ	PQ	SQ	RQ	PQ	SQ	RQ
Supervised (Cityscapes)	62.3	81.8	75.1	31.9	71.7	40.4	33.0	76.3	42.0	38.1	62.4	49.6	31.5	70.1	40.9	73.8	86.4	84.6
DepthG [4] + CutLER [7]	16.1	45.4	21.1	11.0	34.5	13.8	14.4	41.9	19.2	10.1	30.1	13.1	13.4	37.3	17.0	49.6	78.4	60.6
U2Seg [3]	18.4	55.8	22.7	20.6	52.9	25.2	15.8	57.2	19.2	20.3	45.8	26.5	19.8	50.8	23.4	50.7	79.2	64.3
CUPS *(Ours)*	27.8	57.4	35.2	25.5	58.1	32.5	19.9	60.3	25.9	24.4	48.5	33.0	26.4	60.3	33.0	67.8	86.4	76.9
vs. prev. SOTA	+9.4	+1.6	+12.5	+4.9	+5.2	+7.3	+4.1	+3.1	+6.7	+4.1	+2.7	+6.5	+6.6	+9.5	+9.6	+17.1	+7.2	+12.6

Table 2. **Unsupervised semantic segmentation** on Cityscapes val, using Acc. and mIoU (all in %, ↑).

Method	Acc	mIoU
Supervised	94.7	76.7
STEGO [1]	73.2	21.0
DepthG [4]	81.6	23.1
U2Seg [3]	79.1	21.6
CUPS *(Ours)*	83.2	26.8

Table 3. **Unsupervised instance segmentation results** on Waymo using AP_{50} and AP (all in %, ↑).

Method	Training data	AP_{50}	AP
Supervised	Cityscapes	44.6	27.6
U2Seg [3]	COCO & ImageNet	4.3	2.3
CutLER [7]	ImageNet	9.1	5.2
MOD-UV [6]	Waymo	25.1	11.1
CUPS *(Ours)*	Cityscapes	30.5	12.4

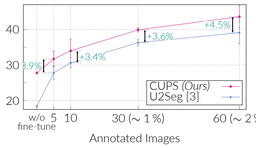


Figure 2. **Label-efficient learning** on Cityscapes val using PQ (in %, ↑).

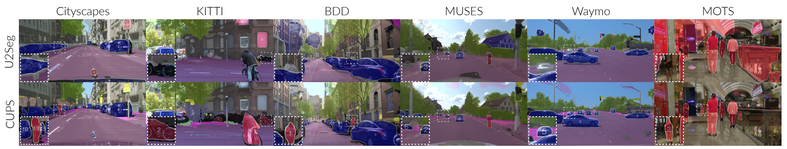


Figure 3. **Qualitative comparison** of the previous SotA U2Seg [3] and CUPS.

Conclusion

- Combine *self-supervised visual representation, unsupervised depth, and motion cues* effectively for unsupervised panoptic scene-understanding
- Significantly *improved* panoptic segmentation *accuracy* on scene-centric data and generalization to various scene-centric datasets
- Achieve *state-of-the-art* accuracy in unsupervised semantic and instance segmentation as well as strong label-efficient learning results