CVPR Nashville JUNE 11-15, 2025

Scene-Centric Unsupervised Panoptic Segmentation

CVPR 2025 Highlight


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]

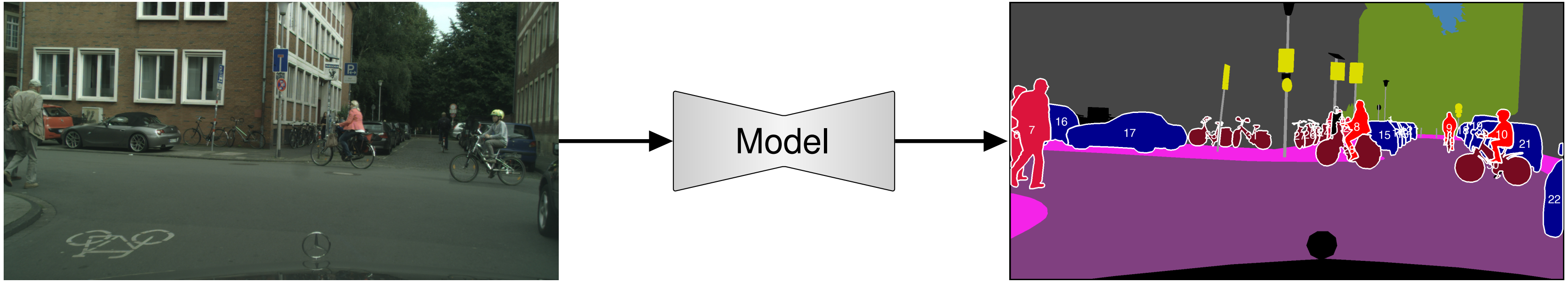
*equal contribution

1 TECHNISCHE UNIVERSITÄT DARMSTADT
2 TUM
3 UNIVERSITY OF OXFORD
4 MCML Munich Center for Machine Learning
5 ZUSE SCHOOL ELIZA
6 hessian.AI

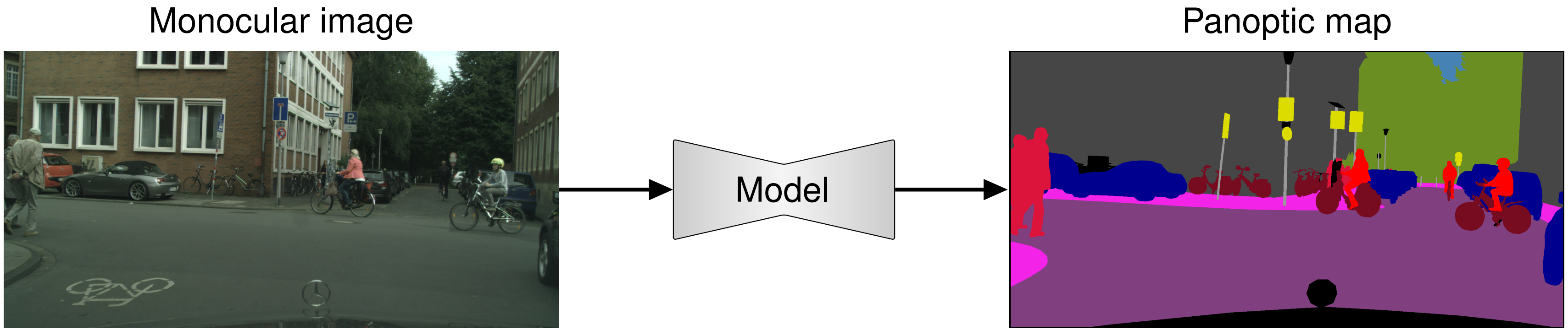
Unsupervised Panoptic Segmentation

Monocular image

Panoptic map



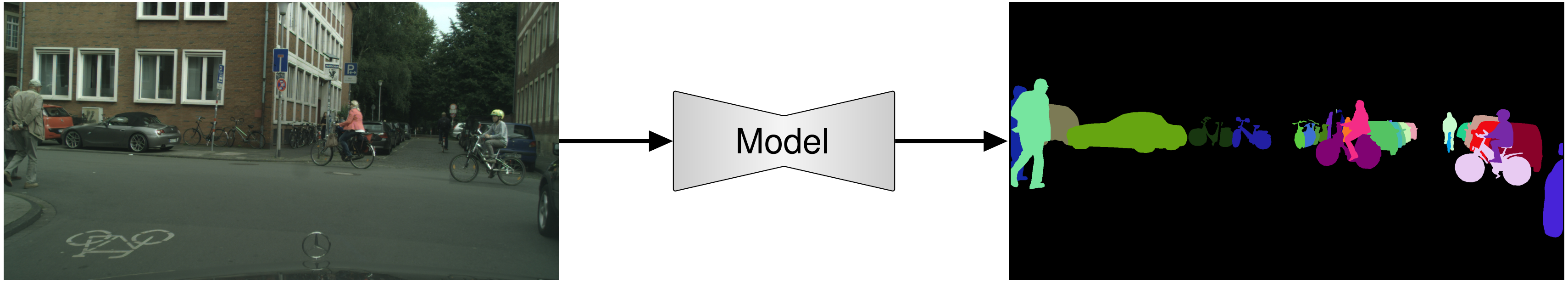
Unsupervised Panoptic Segmentation



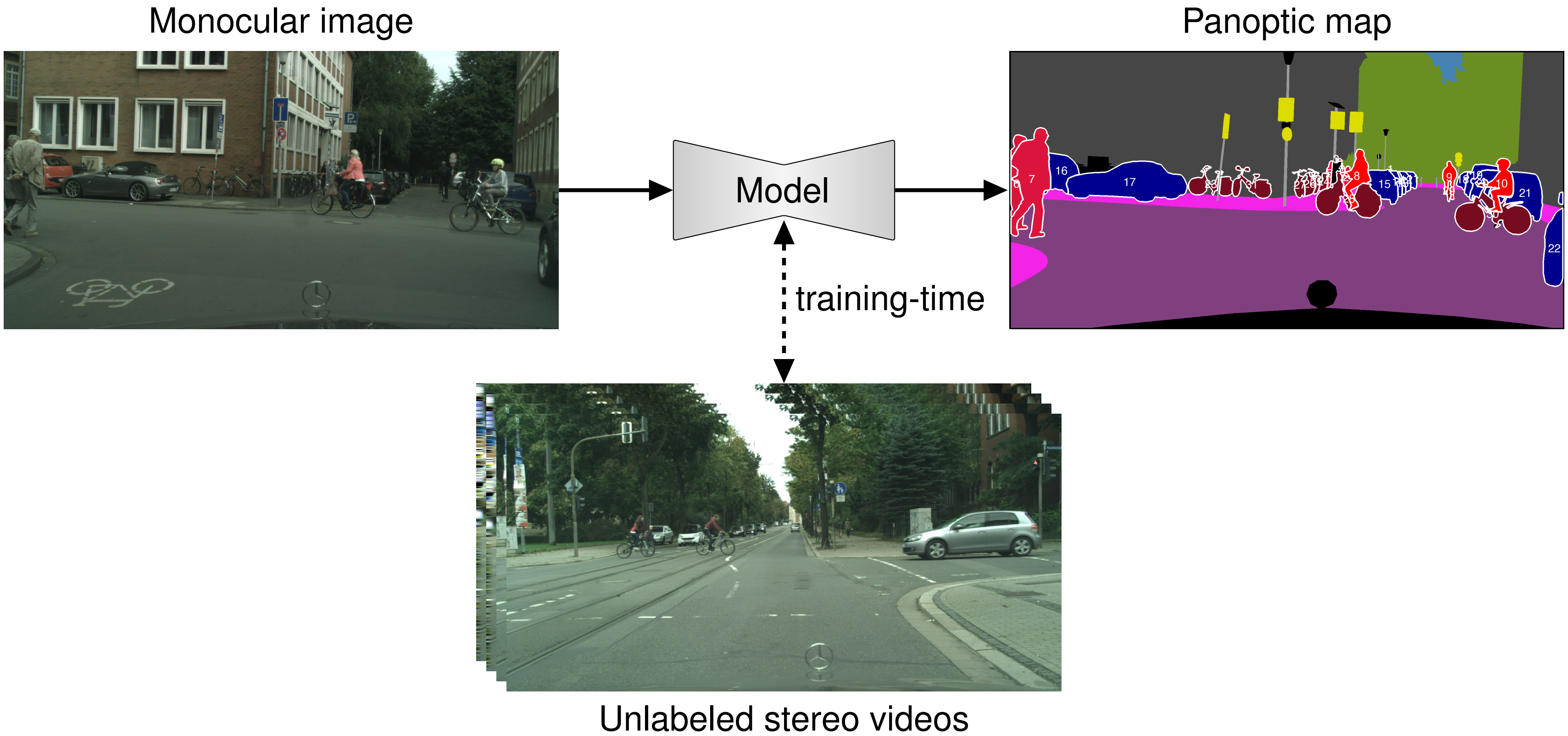
Unsupervised Panoptic Segmentation

Monocular image

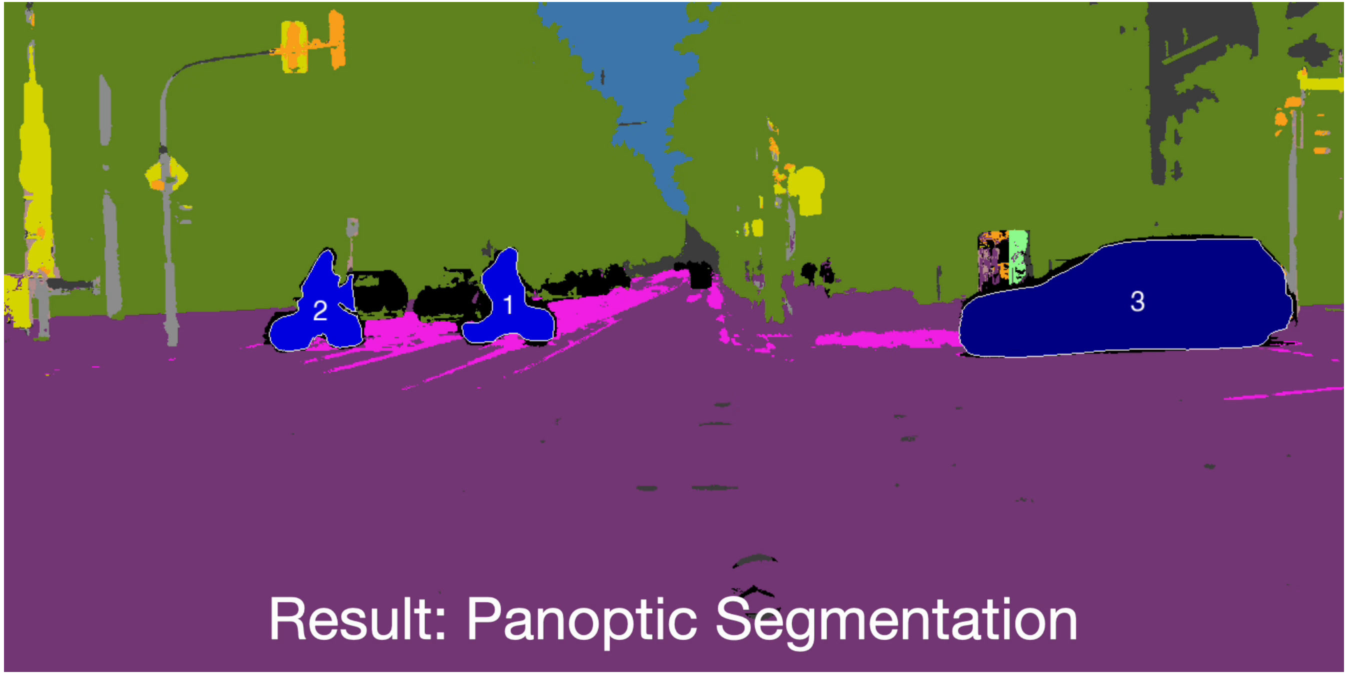
Panoptic map



Unsupervised Panoptic Segmentation



High-Level Idea



Visual representations + depth + optical flow for pseudo labeling

Relation to Previous Work

Features	U2Seg [1]	CUPS *(Ours)*
Unsupervised panoptic segmentation	✓	✓
High-resolution pseudo labels	✗	✓
Thing-stuff separation	∼	✓
Scene-centric training	✗	✓

[1] D. Niu et al., “Unsupervised universal image segmentation,” in *CVPR*, 2024.

Relation to Previous Work

Features	**U2Seg** [1]	**CUPS** *(Ours)*
Unsupervised panoptic segmentation	✓	✓
High-resolution pseudo labels	✗	✓
Thing-stuff separation	∼	✓
Scene-centric training	✗	✓

[1] D. Niu et al., “Unsupervised universal image segmentation,” in *CVPR*, 2024.

Relation to Previous Work

Features	**U2Seg** [1]	**CUPS** *(Ours)*
Unsupervised panoptic segmentation	✓	✓
High-resolution pseudo labels	✗	✓
Thing-stuff separation	∼	✓
Scene-centric training	✗	✓

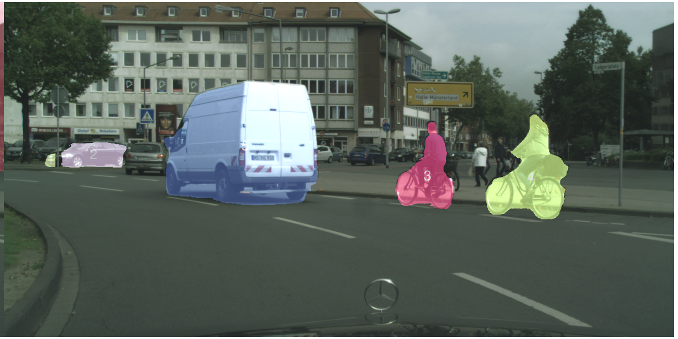
[1] D. Niu et al., “Unsupervised universal image segmentation,” in *CVPR*, 2024.

Relation to Previous Work

Features	**U2Seg** [1]	**CUPS** *(Ours)*
Unsupervised panoptic segmentation	✓	✓
High-resolution pseudo labels	✗	✓
Thing-stuff separation	∼	✓
Scene-centric training	✗	✓

MaskCut

Ours (motion-based)



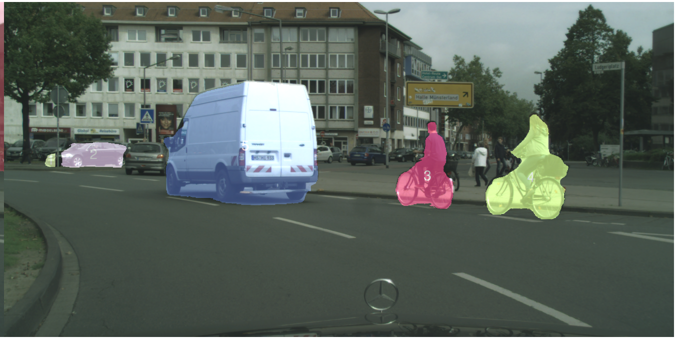
[1] D. Niu et al., "Unsupervised universal image segmentation," in *CVPR*, 2024.

Relation to Previous Work

Features	**U2Seg** [1]	**CUPS** *(Ours)*
Unsupervised panoptic segmentation	✓	✓
High-resolution pseudo labels	✗	✓
Thing-stuff separation	∼	✓
Scene-centric training	✗	✓

MaskCut

Ours (motion-based)

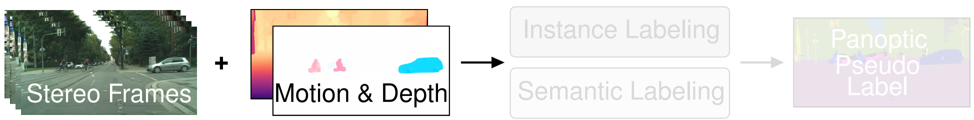


U2Seg performs poorly on **scene-centric data** (*e.g.*, Cityscapes and KITTI)

[1] D. Niu et al., "Unsupervised universal image segmentation," in *CVPR*, 2024.

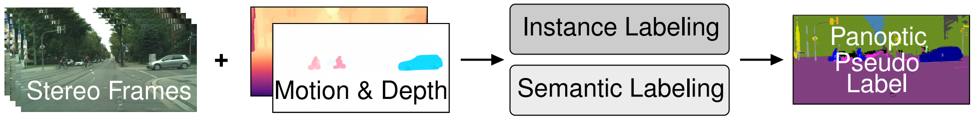
CUPS🥤🥤: Framework Overview

Pseudo Label Generation

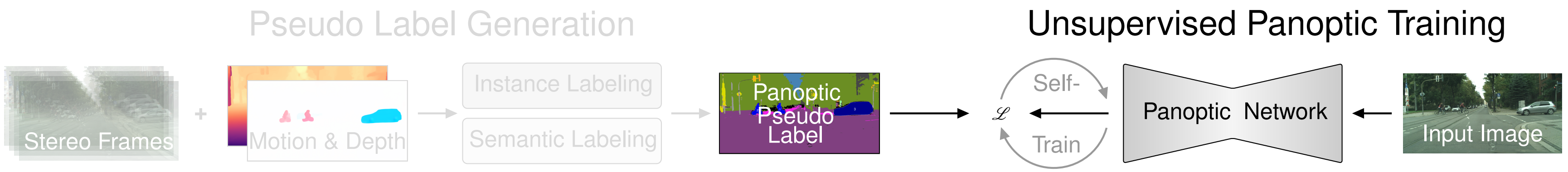


CUPS🥤🥤: Framework Overview

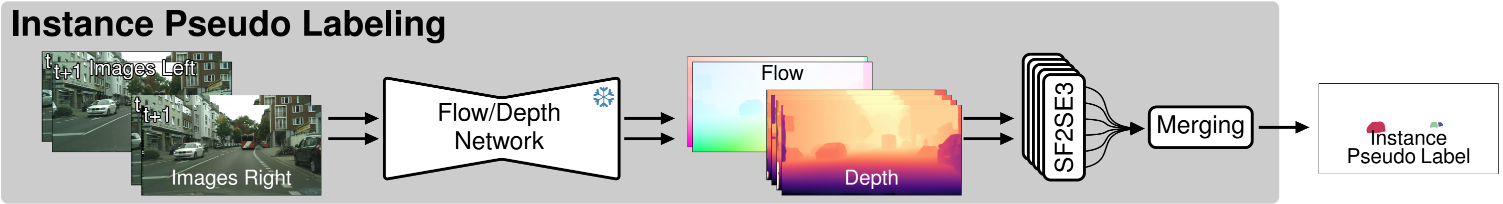
Pseudo Label Generation



CUPS🥤🥤: Framework Overview

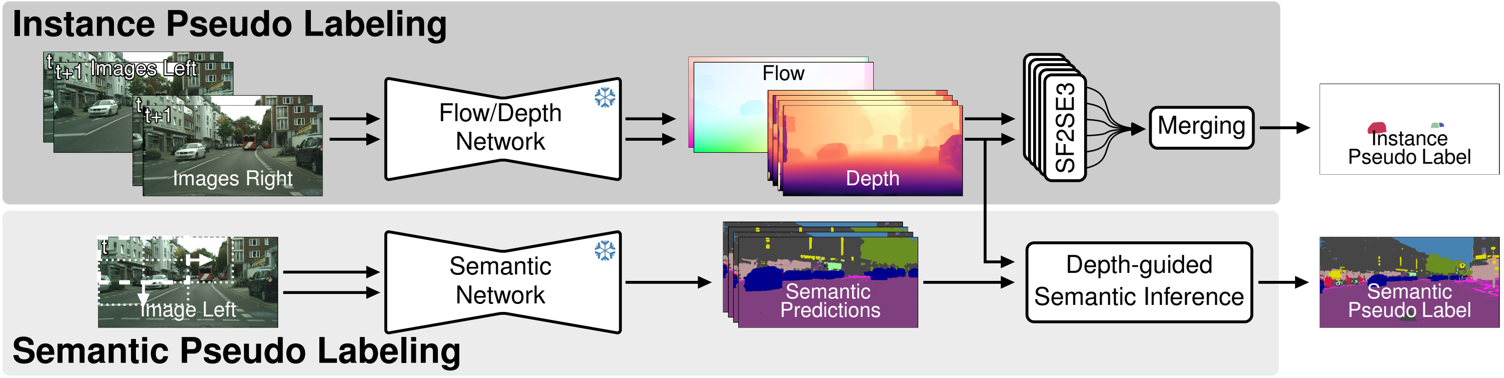


CUPS🥤🥤: Pseudo Label Generation



✓ Scene-centric instance pseudo labels

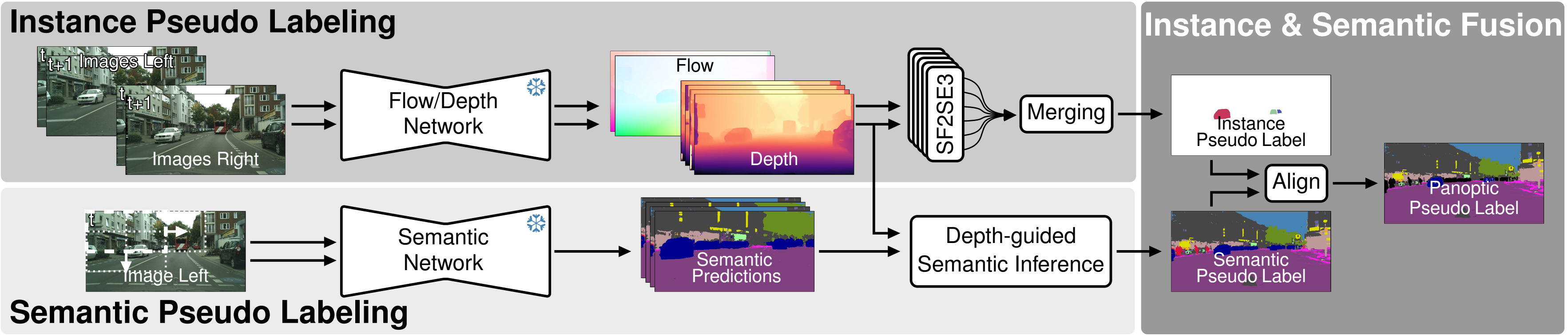
CUPS🥤🥤: Pseudo Label Generation



✓ Scene-centric instance pseudo labels

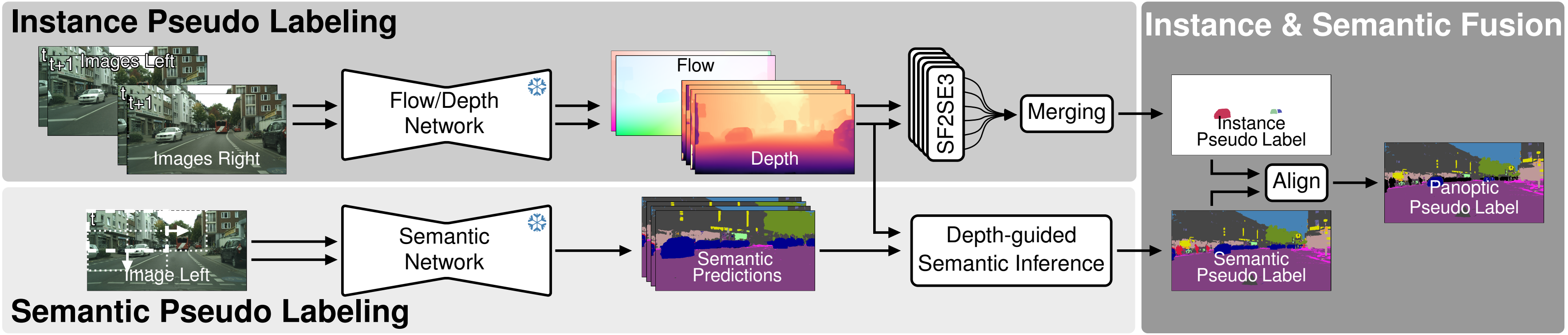
✓ High-resolution semantic pseudo labels

CUPS🥤🥤: Pseudo Label Generation



- ✓ Scene-centric instance pseudo labels
- ✓ High-resolution semantic pseudo labels
- ✓ **High-precision (sparse) panoptic pseudo labels with “thing” and “stuff” split**

CUPS🥤🥤: Pseudo Label Generation



✓ Scene-centric instance pseudo labels

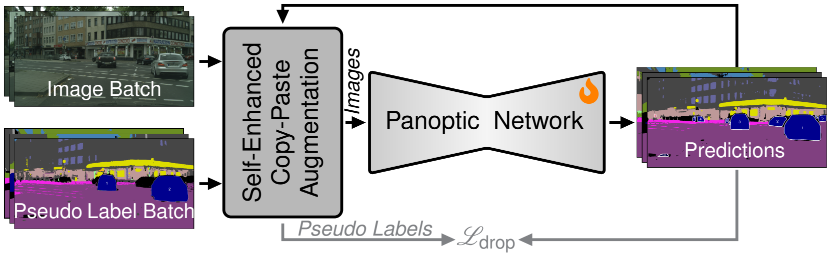
✓ High-resolution semantic pseudo labels

✓ **High-precision (sparse) panoptic pseudo labels with “thing” and “stuff” split**

✓ **Fully unsupervised pseudo labels**

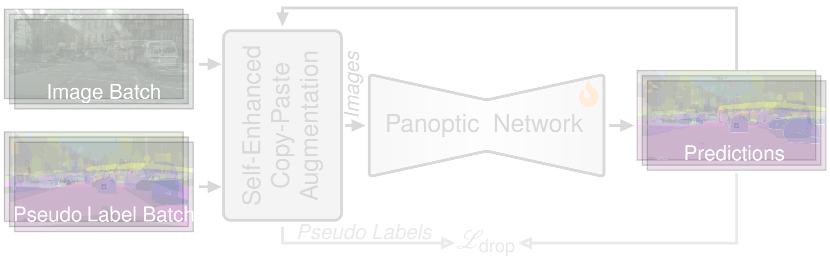
CUPS🥤🥤: Unsupervised Panoptic Training

Panoptic Pseudo Label Training

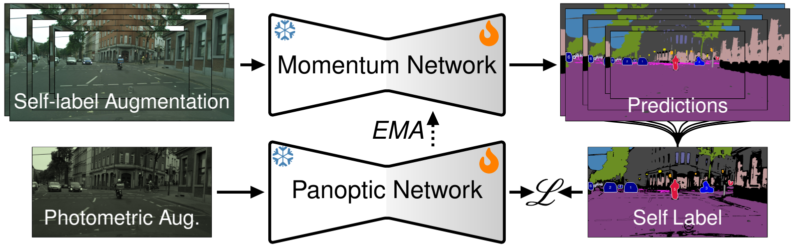


CUPS🥤🥤: Unsupervised Panoptic Training

Panoptic Pseudo Label Training



Panoptic Self-Training



CUPS🥤🥤: Results Panoptic

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 [3] + CutLER [4]	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 [2]	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

PQ: panoptic quality   SQ: segmentation quality   RQ: recognition quality   (all in %, ↑)

[2] L. Sick et al., "Unsupervised semantic segmentation through depth-guided feature correlation and sampling," in *CVPR*, 2024.
[3] X. Wang et al., "Cut and learn for unsupervised object detection and instance segmentation," in *CVPR*, 2023.

CUPS🥤🥤: Results Panoptic

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 [3] + CutLER [4]	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 [2]	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

PQ: panoptic quality SQ: segmentation quality RQ: recognition quality (all in %, ↑)

✓ **Outperform SOTA by a significant margin**

[2] L. Sick et al., “Unsupervised semantic segmentation through depth-guided feature correlation and sampling,” in *CVPR*, 2024.
[3] X. Wang et al., “Cut and learn for unsupervised object detection and instance segmentation,” in *CVPR*, 2023.

CUPS🥤🥤: Results Panoptic

CVPR Nashville JUNE 11-15, 2025

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 [3] + CutLER [4]	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 [2]	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

PQ: panoptic quality SQ: segmentation quality RQ: recognition quality (all in %, $\uparrow$)

✓ **Outperform SOTA by a significant margin**

✓ Generalize to different datasets, including an OOD setting

[2] L. Sick et al., “Unsupervised semantic segmentation through depth-guided feature correlation and sampling,” in *CVPR*, 2024.
[3] X. Wang et al., “Cut and learn for unsupervised object detection and instance segmentation,” in *CVPR*, 2023.

CUPS🥤🥤: Results Panoptic

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 [3] + CutLER [4]	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 [2]	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

PQ: panoptic quality SQ: segmentation quality RQ: recognition quality (all in %, ↑)

- ✓ **Outperform SOTA by a significant margin**
- ✓ Generalize to different datasets, including an OOD setting
- ✓ Performance across datasets is stable, different from supervised learning

[2] L. Sick et al., “Unsupervised semantic segmentation through depth-guided feature correlation and sampling,” in *CVPR*, 2024.
[3] X. Wang et al., “Cut and learn for unsupervised object detection and instance segmentation,” in *CVPR*, 2023.

CUPS🥤🥤: Qualitative Results

Conclusion

We presented CUPS for unsupervised scene-centric panoptic segmentation

Conclusion

We presented CUPS for unsupervised scene-centric panoptic segmentation

- Motion & depth cues, combined with self-supervised visual representations, are effective for unsupervised panoptic scene understanding

Conclusion

CVPR Nashville JUNE 11-15, 2025

We presented CUPS for unsupervised scene-centric panoptic segmentation

- Motion & depth cues, combined with self-supervised visual representations, are effective for unsupervised panoptic scene understanding
- Significantly improved unsupervised panoptic accuracy on scene-centric data

Conclusion

We presented CUPS for unsupervised scene-centric panoptic segmentation

- Motion & depth cues, combined with self-supervised visual representations, are effective for unsupervised panoptic scene understanding
- Significantly improved unsupervised panoptic accuracy on scene-centric data
- CUPS generalizes to various scene-centric datasets

Conclusion

We presented CUPS for unsupervised scene-centric panoptic segmentation

- Motion & depth cues, combined with self-supervised visual representations, are effective for unsupervised panoptic scene understanding
- Significantly improved unsupervised panoptic accuracy on scene-centric data
- CUPS generalizes to various scene-centric datasets
- State-of-the-art performance in unsupervised semantic & instance segmentation

Conclusion

CVPR Nashville JUNE 11-15, 2025

We presented CUPS for unsupervised scene-centric panoptic segmentation

- Motion & depth cues, combined with self-supervised visual representations, are effective for unsupervised panoptic scene understanding
- Significantly improved unsupervised panoptic accuracy on scene-centric data
- CUPS generalizes to various scene-centric datasets
- State-of-the-art performance in unsupervised semantic & instance segmentation
- Strong label-efficient learning results

CVPR Nashville JUNE 11-15, 2025

Paper | **Project Page** | **Code & Weights**

https://visinf.github.io/cups/

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.

European Research Council
Established by the European Commission