HUNAN UNIVERSITY

ACCV HANOI VIETNAM 2024 DEC 8-12

CV:PU

Towards Holistic Scene Understanding for Autonomous Driving

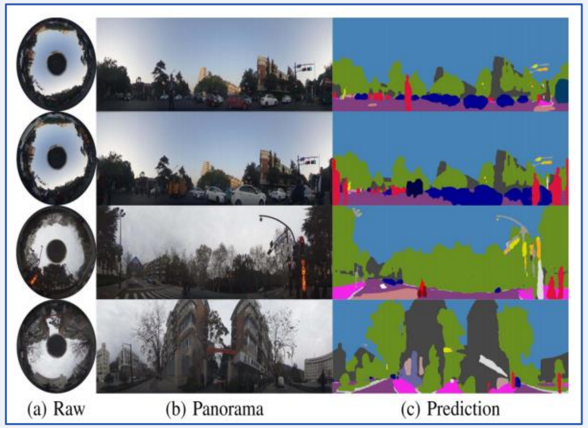
Kailun Yang

Hunan University

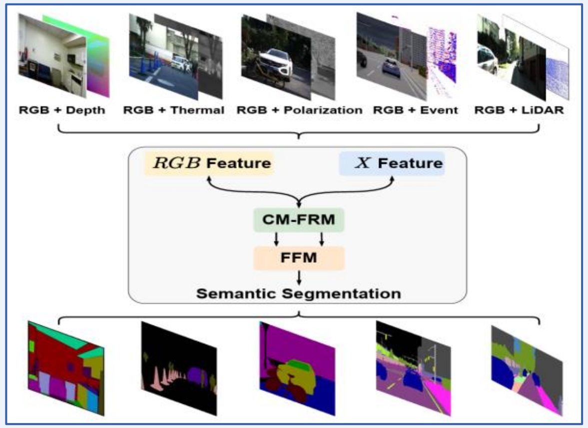
Computer Vision for Panoramic Understanding (Lab)

CV:PU

Panoramic Vision for Scene Understanding

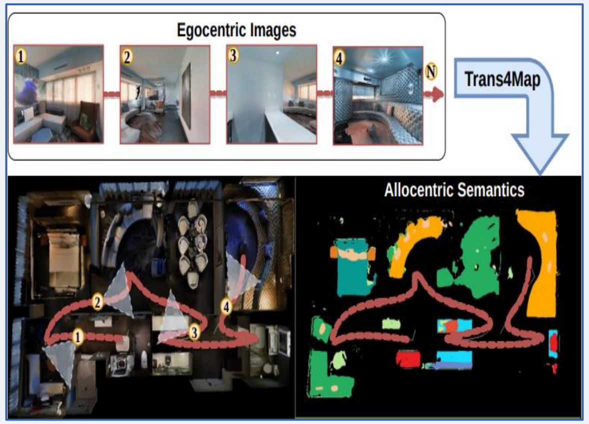


Multimodal Perception and Computational Imaging



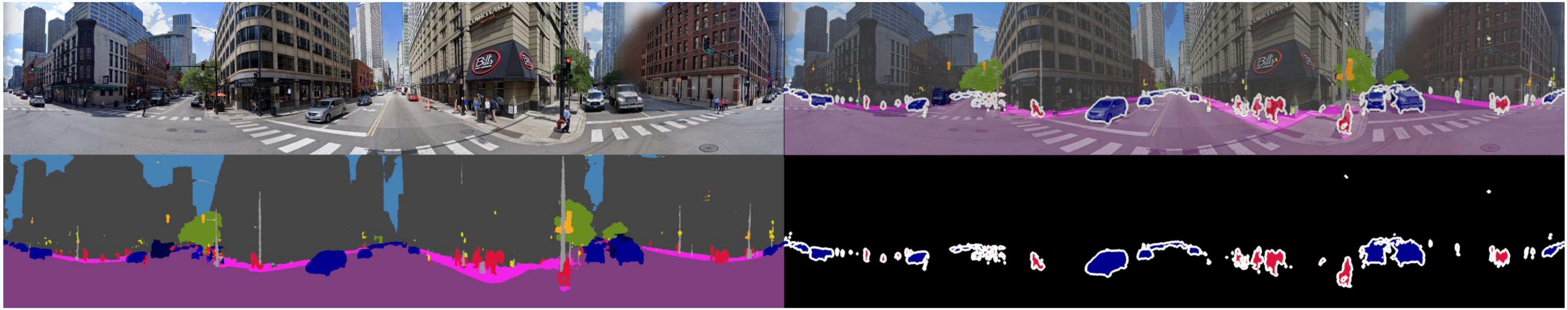
Autonomous Driving and World Models

Visual Assistance and Human-Computer Interaction



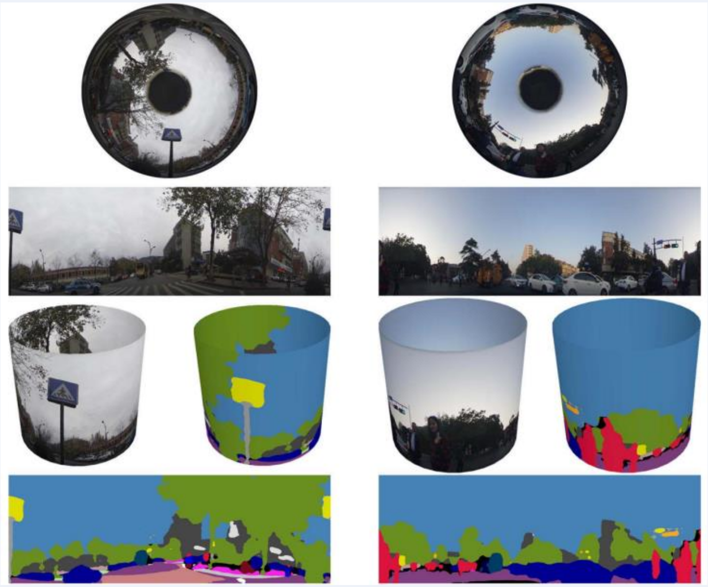
Holistic Scene Understanding

- **Overcoming the limit in Field of View (FoV): Panoramic Scene Segmentation**
- **Overcoming the limit in annotations: Label-efficient Occupancy Prediction**
- **Overcoming the limit in cross-modal fusion: Arbitrary-modal Segmentation**



1. Panoramic Scene Segmentation

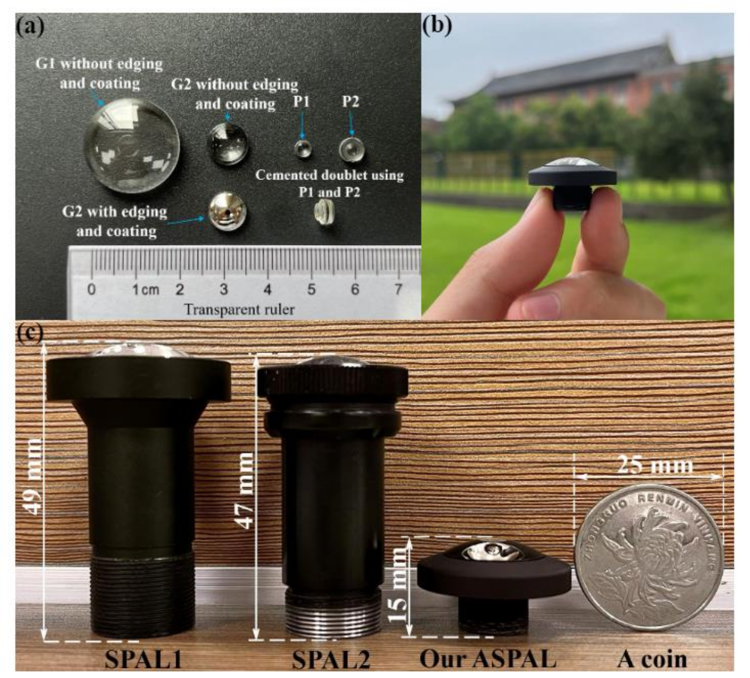
- Panoramic Annular Lens (PAL), applied in GOAT G1
- Panoramic Annular Scene Segmentation (PASS) models
- PASS, WildPASS, WildPPS, and DensePASS benchmarks
- WildPASS: 2500 images collected from 65 cities, 6 continents [1]

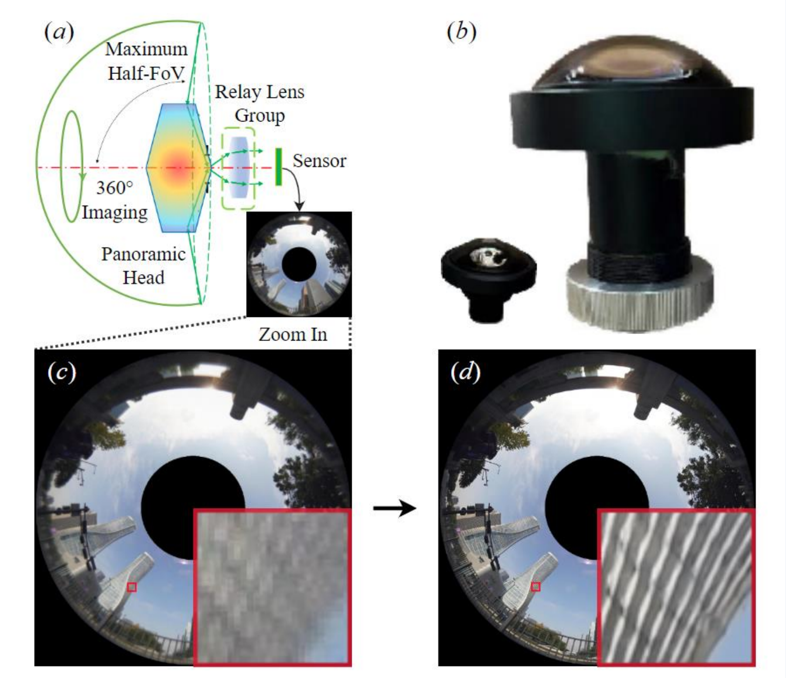


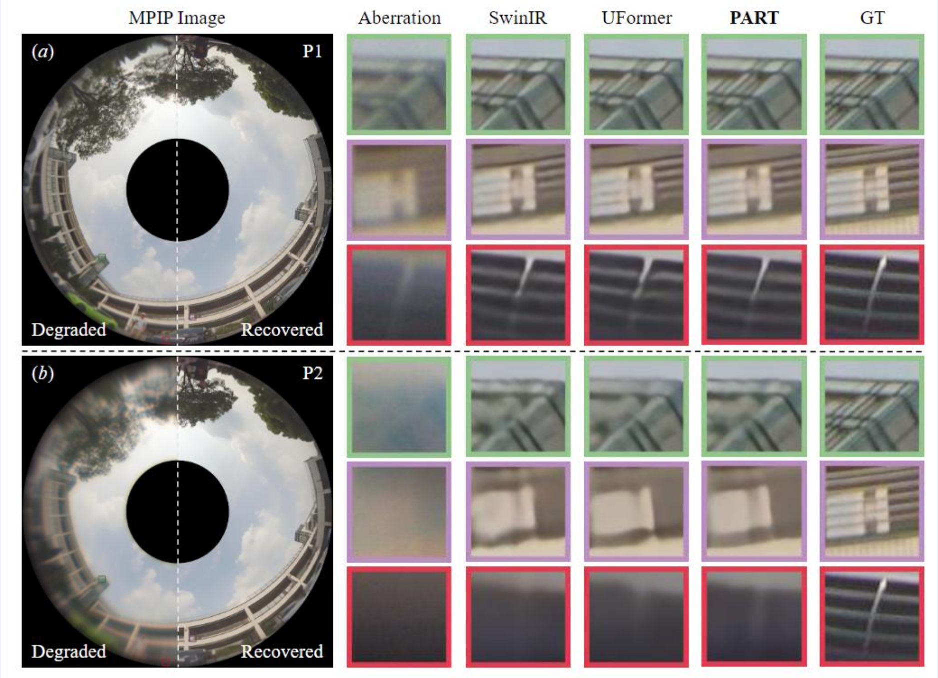
[1] Yang, Kailun, Xinxin Hu, and Rainer Stiefelhagen. "Is context-aware CNN ready for the surroundings? Panoramic semantic segmentation in the wild." IEEE Transactions on Image Processing 30 (2021): 1866-1881.

1. Panoramic Scene Segmentation

- Glass-plastic hybrid minimalist aspheric panoramic lens [1]
- Minimalist and high-quality computational imaging engine [2]



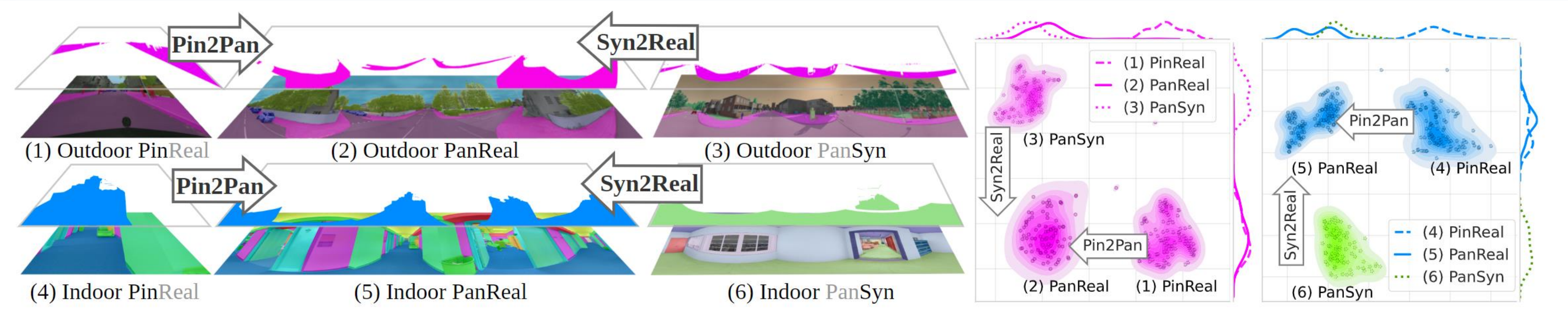




[1] Gao, Shaohua, et al. "Design, analysis, and manufacturing of a glass-plastic hybrid minimalist aspheric panoramic annular lens." Optics & Laser Technology 177 (2024): 111119.
[2] Jiang, Qi, et al. "Minimalist and high-quality panoramic imaging with PSF-aware transformers." *IEEE Transactions on Image Processing* (2024).

1. Panoramic Scene Segmentation

- Unsupervised Domain Adaptation (UDA)
- Outdoor and indoor UDA benchmarks, TPAMI 2024 [1]
- Distortion-aware panoramic segmentation transformers

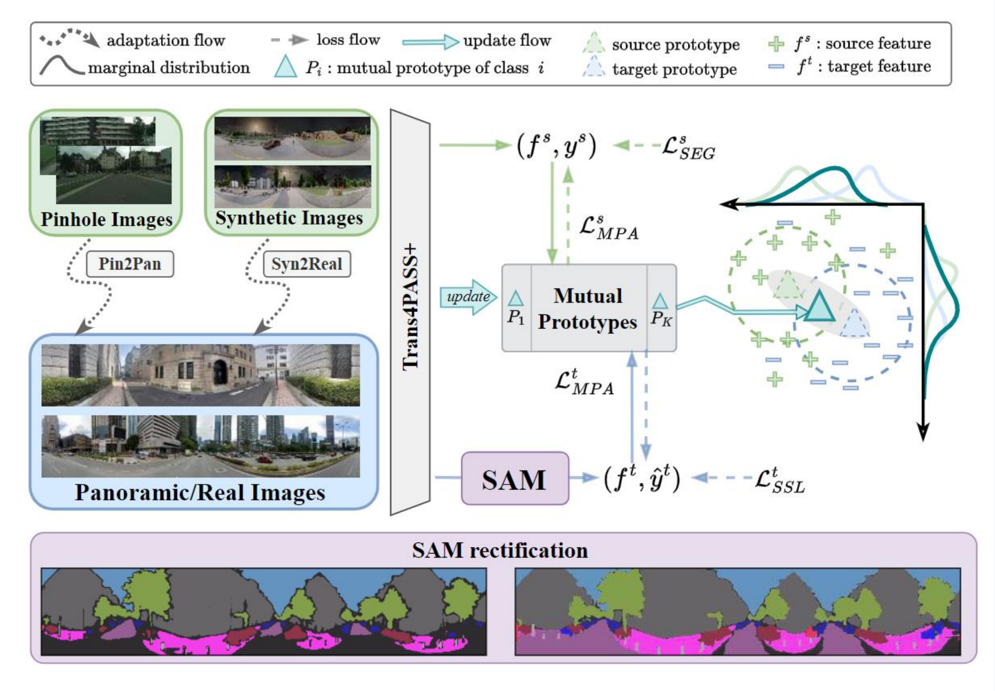


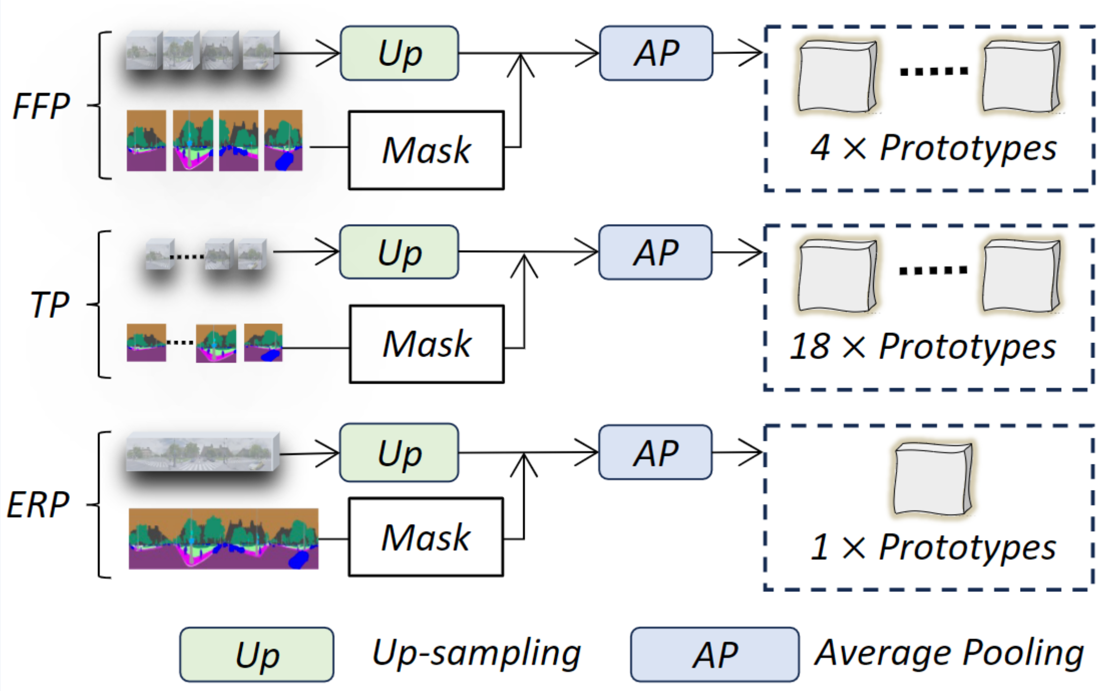
(a) Domain adaptation paradigms (b) Distribution *sidewalk* (c) Distribution *floor*

[1] Zhang, Jiaming, et al. "Behind every domain there is a shift: Adapting distortion-aware vision transformers for panoramic semantic segmentation." IEEE Transactions on Pattern Analysis and Machine Intelligence (2024).

1. Panoramic Scene Segmentation

- SAM-rectified prototypical adaptation, Zhang *et al.*, TPAMI 2024
- Both semantics, distortion, and style matter in source-free UDA, panoramic prototypes, Zheng *et al.*, CVPR 2024 [1]

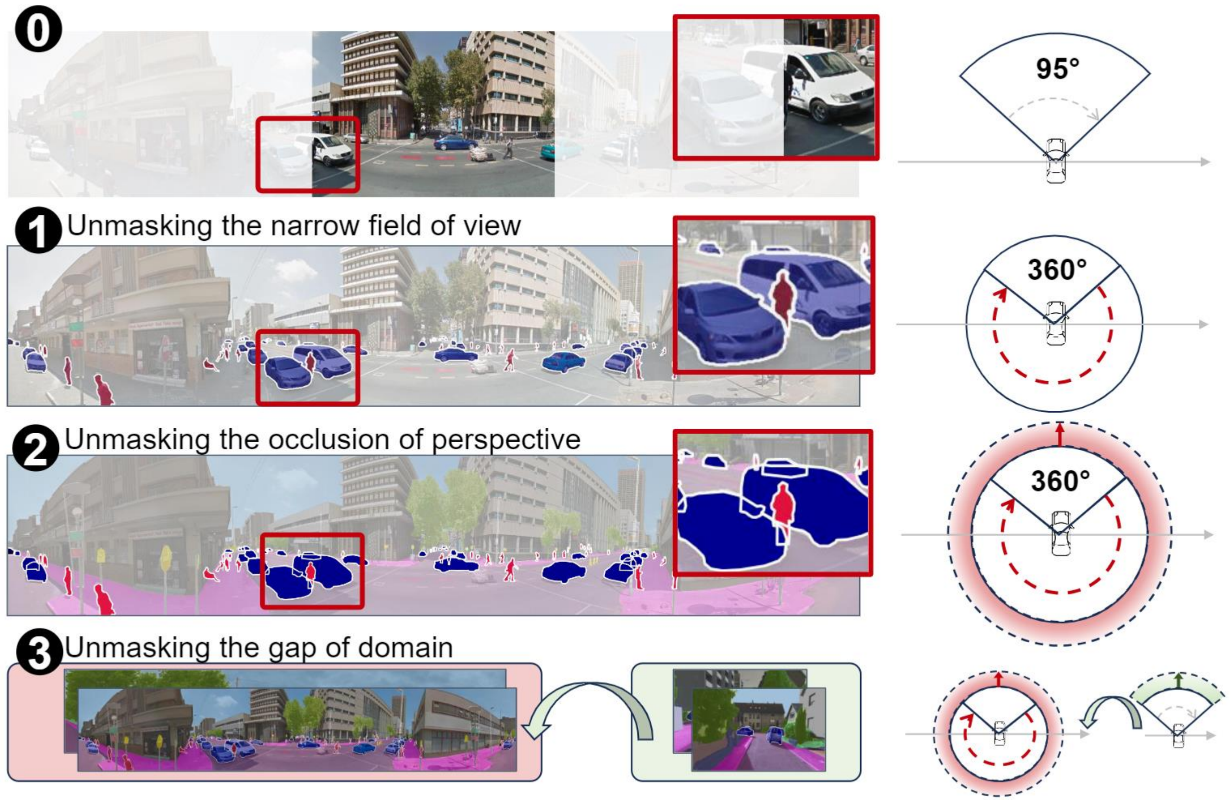


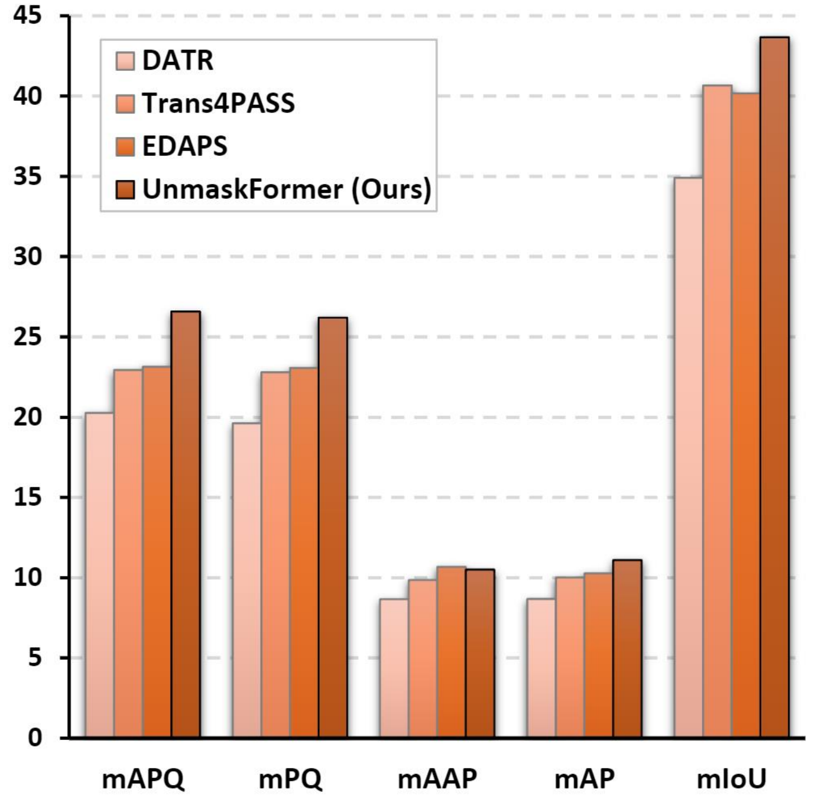


[1] Zheng, Xu, et al. "Semantics Distortion and Style Matter: Towards Source-free UDA for Panoramic Segmentation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

1. Panoramic Scene Segmentation

- Occlusion-Aware Seamless Segmentation, ECCV 2024 [1]

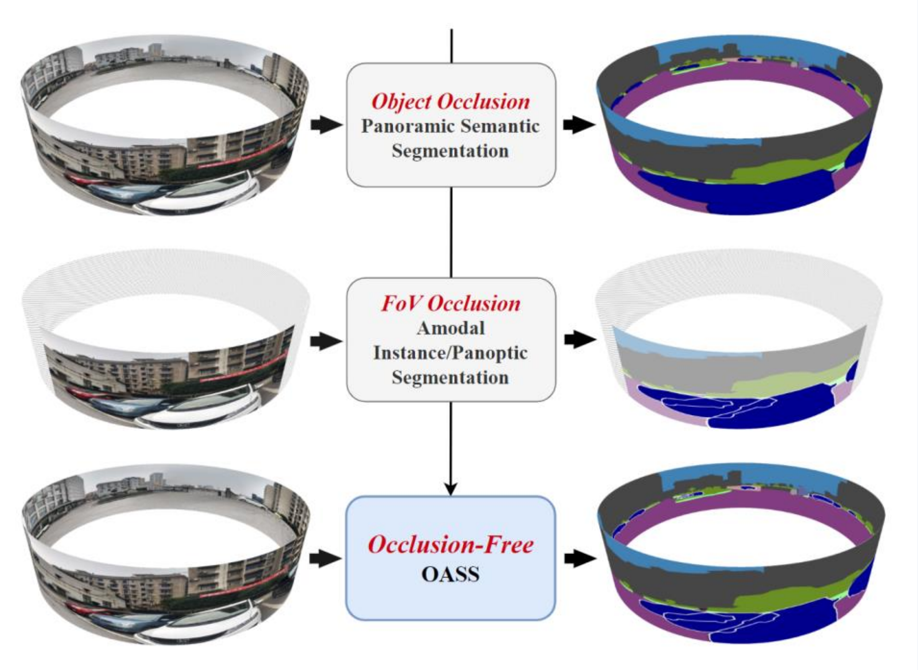


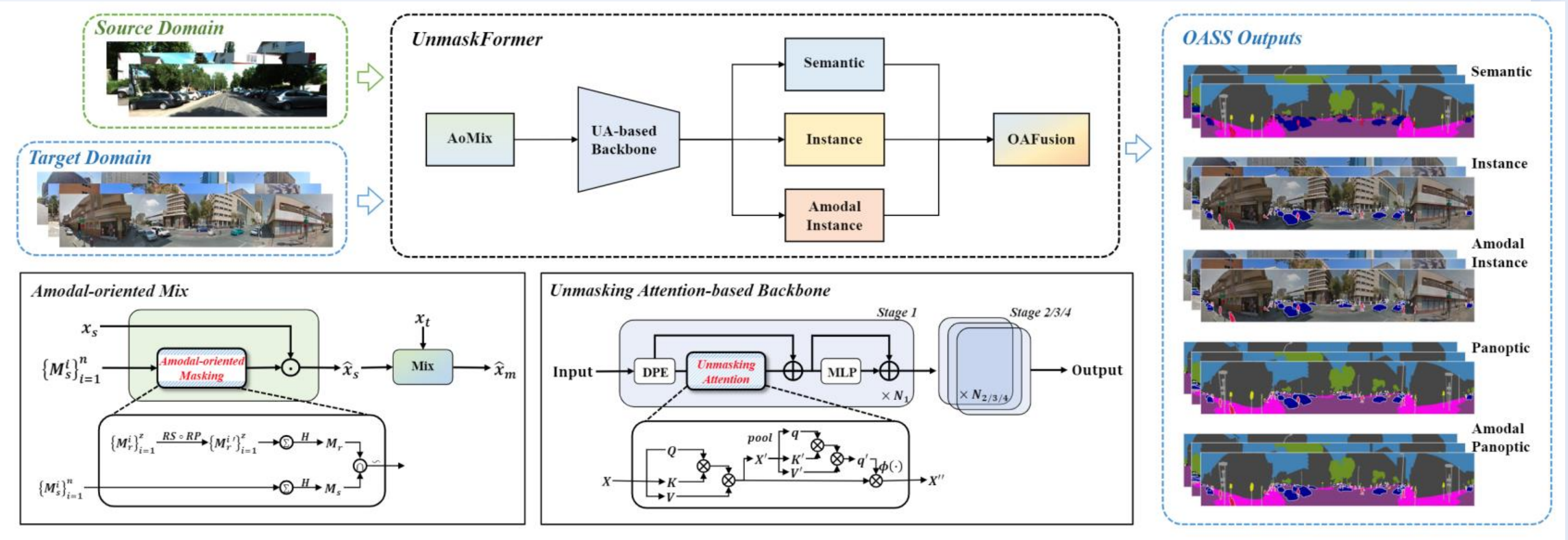


[1] Cao, Yihong, et al. "Occlusion-Aware Seamless Segmentation." European Conference on Computer Vision (ECCV), 2024.

1. Panoramic Scene Segmentation

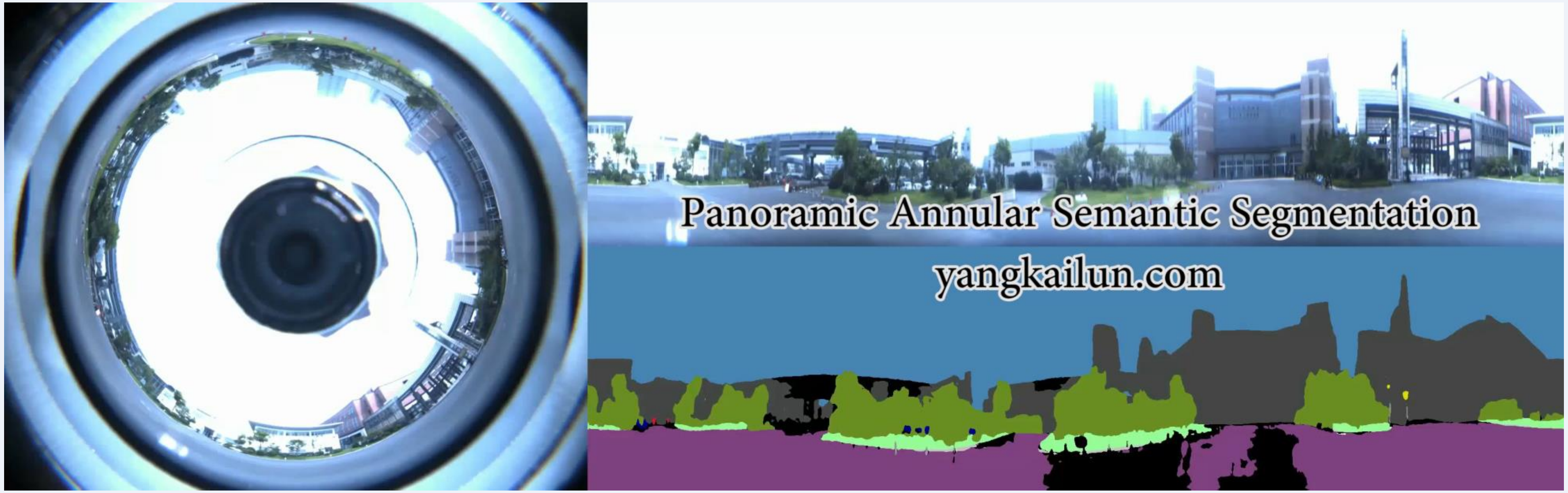
- Occlusion-aware Seamless Segmentation, ECCV 2024 [1]





[1] Cao, Yihong, et al. "Occlusion-Aware Seamless Segmentation." European Conference on Computer Vision (ECCV), 2024.

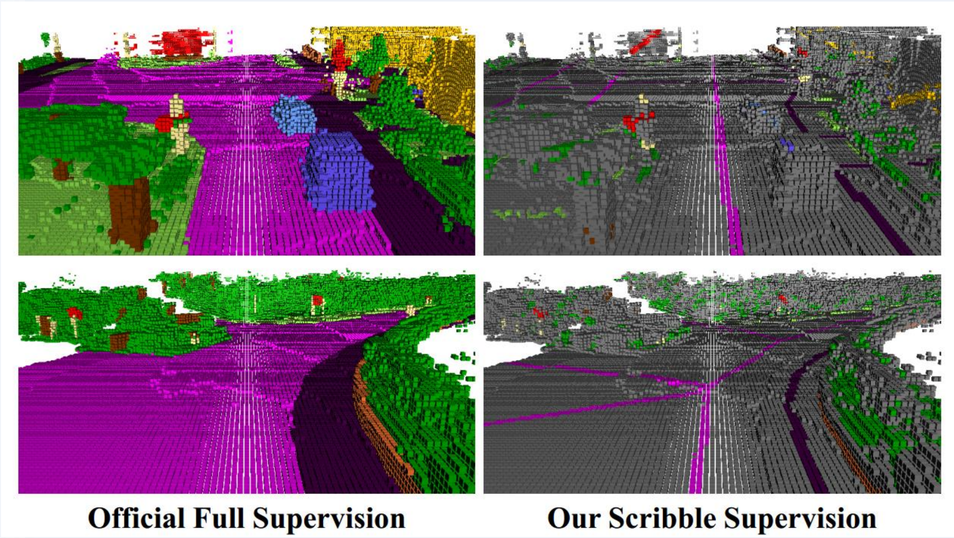
1. Panoramic Scene Segmentation

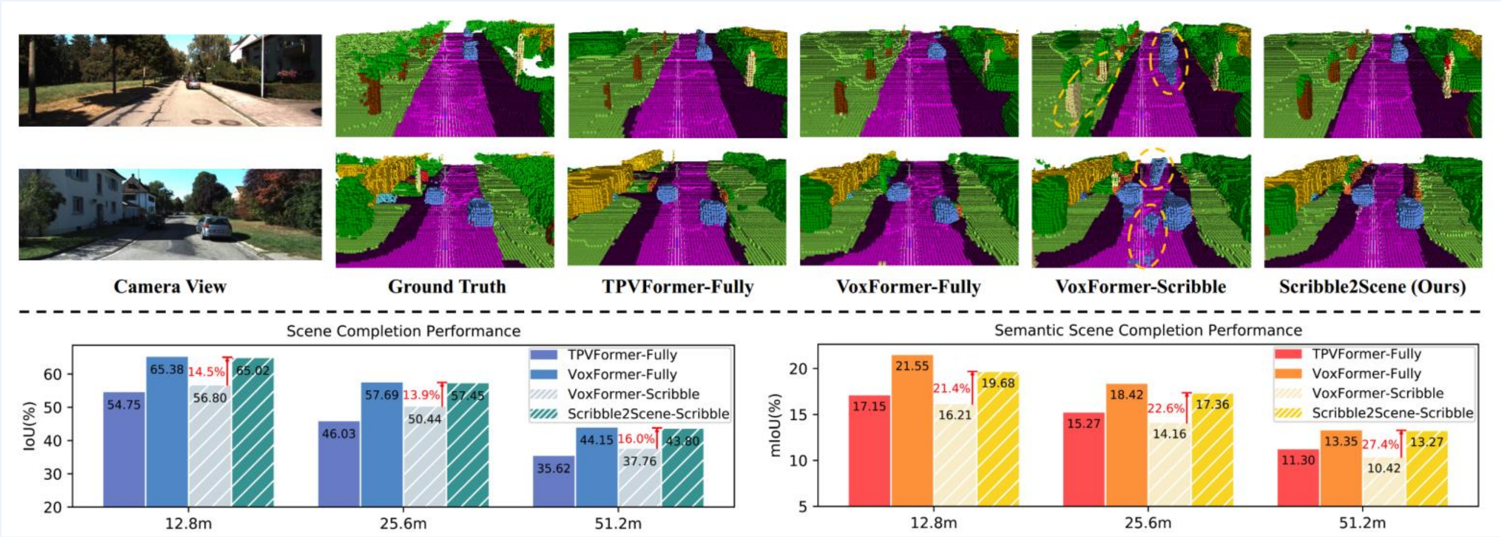


[1] Yang, Kailun, et al. "Pass: Panoramic annular semantic segmentation." IEEE Transactions on Intelligent Transportation Systems 21.10 (2019): 4171-4185.
[2] Yang, Kailun, et al. "Capturing omni-range context for omnidirectional segmentation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

2. Label-efficient Occupancy Prediction

- Weakly scribble-supervised Scribble2Scene, 99% performance
- ScribbleSC benchmark with scribble-based semantic occupancy labels and dense geometric structure [1]

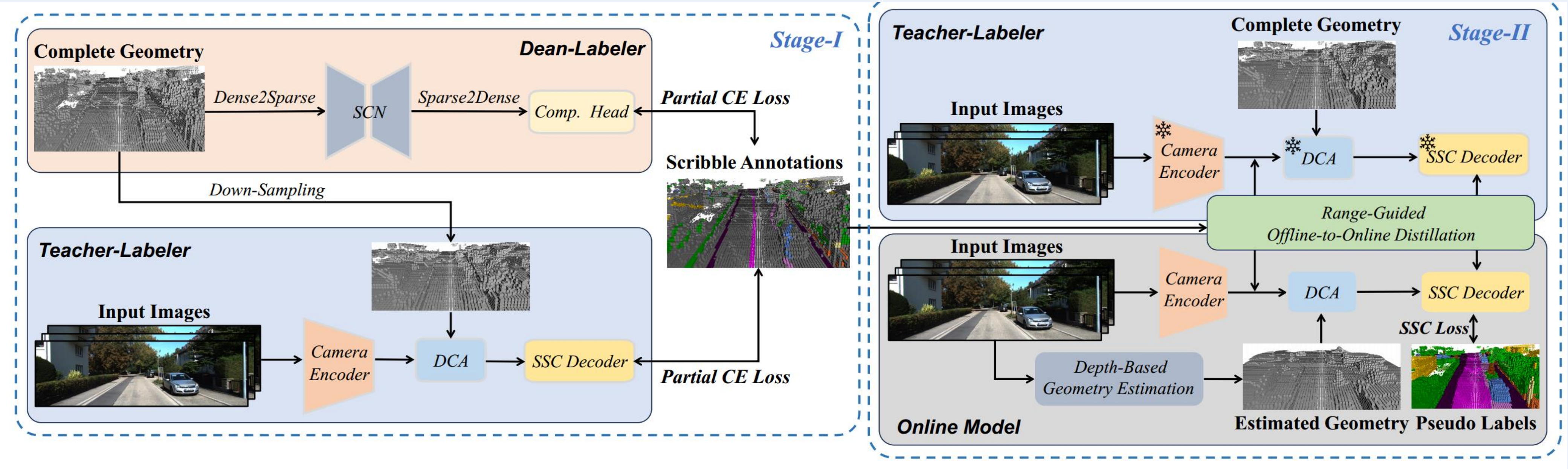




[1] Wang, Song, et al. "Label-efficient Semantic Scene Completion with Scribble Annotations." IJCAI, 2024.

2. Label-efficient Occupancy Prediction

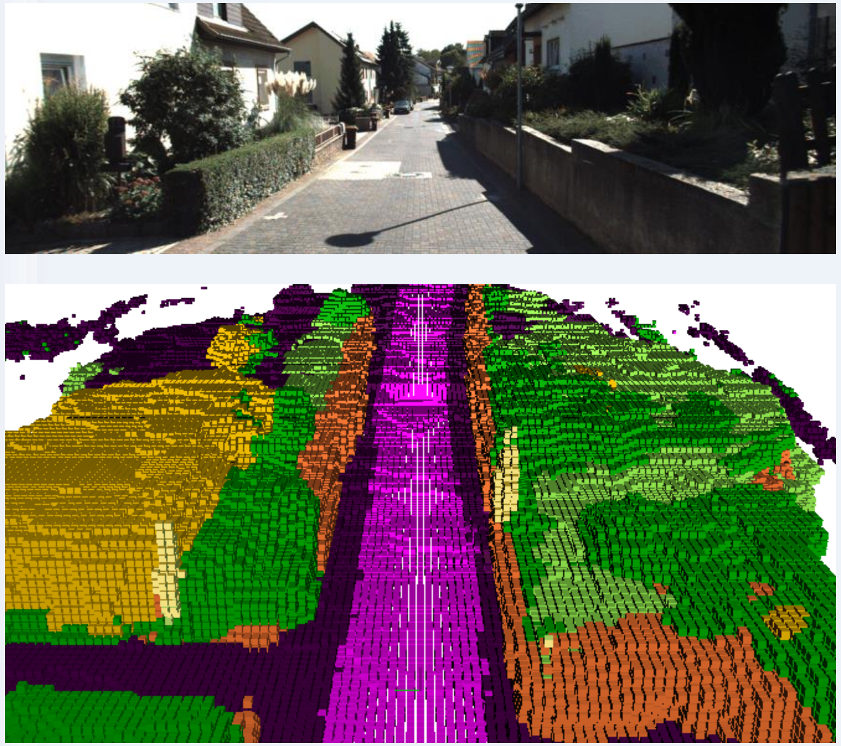
- Dean-Labeler: Treats the complete geometric structure as input, converts into an easier semantic segmentation problem
- Teacher-Labeler: trained in offline mode with input image and complete geometry, the same architecture as the online model



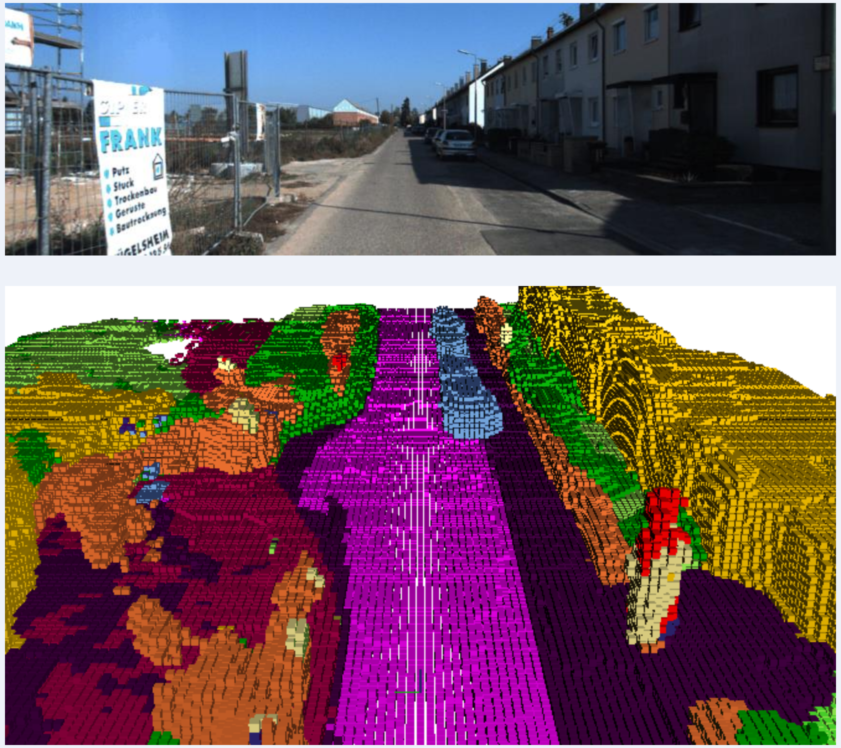
[1] Wang, Song, et al. "Label-efficient Semantic Scene Completion with Scribble Annotations." IJCAI, 2024.

2. Label-efficient Occupancy Prediction

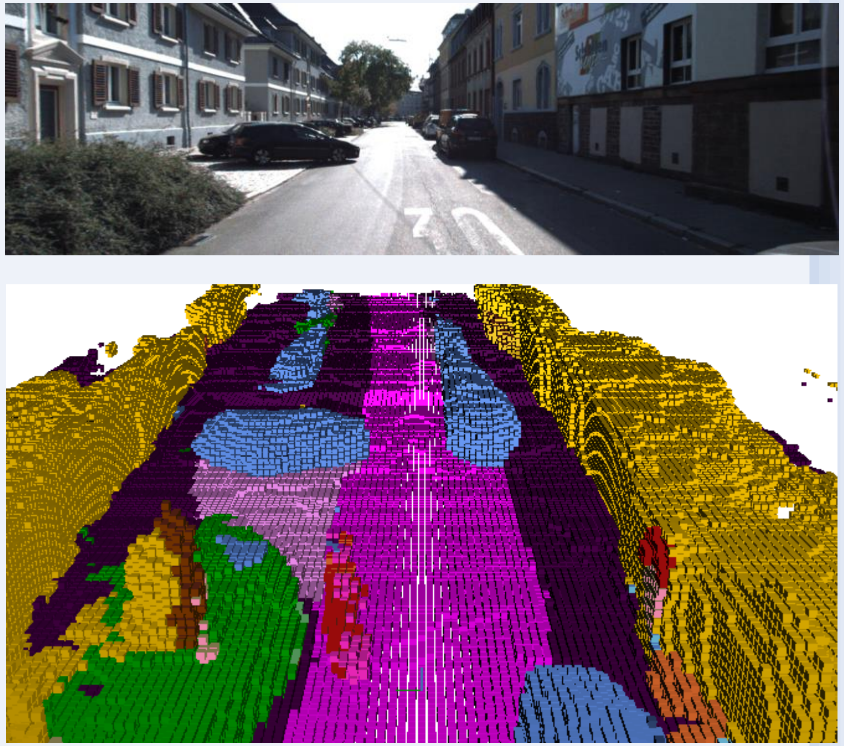
Performance on Sequence 11



Performance on Sequence 15

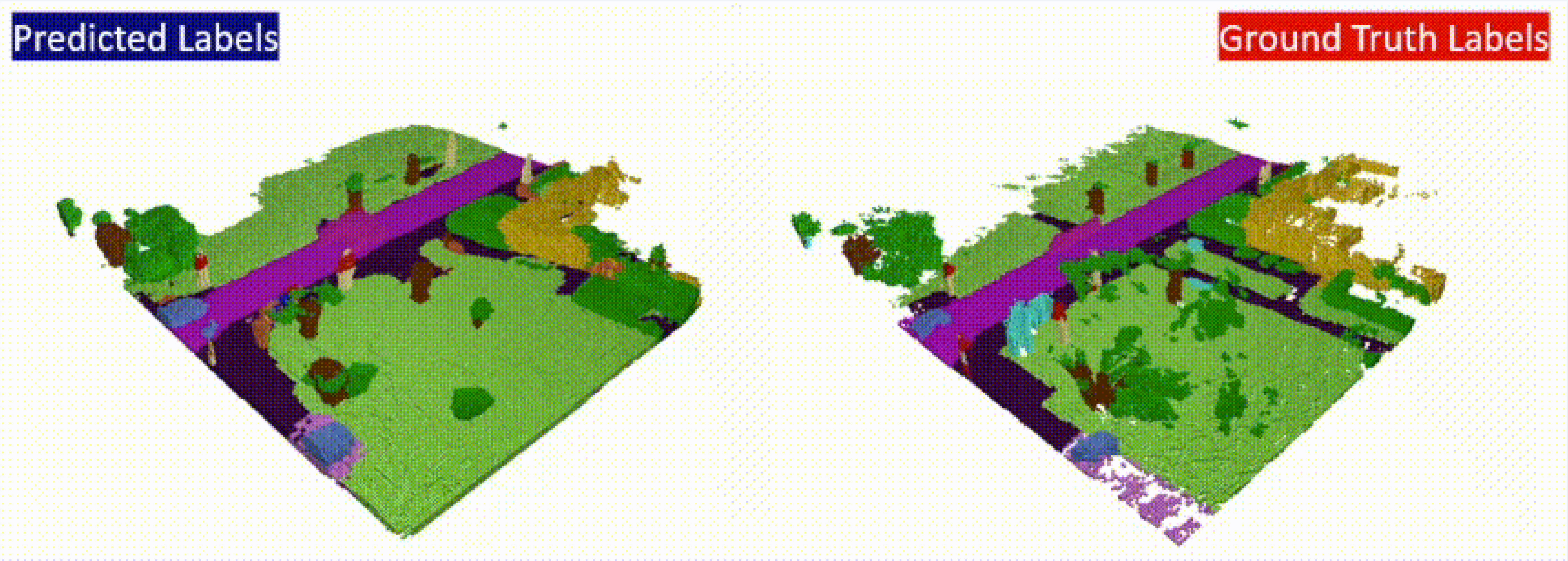


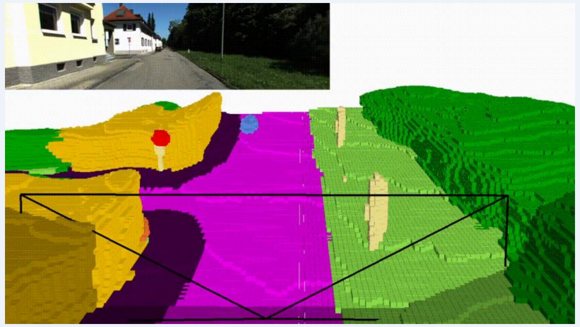
Performance on Sequence 19



car | pole | fence | person | bicycle | parking | vegetation | other-vehicle | bicyclist | motorcyclist

road | truck | terrain | trunk | building | sidewalk | traffic-sign | other-ground | motocycle



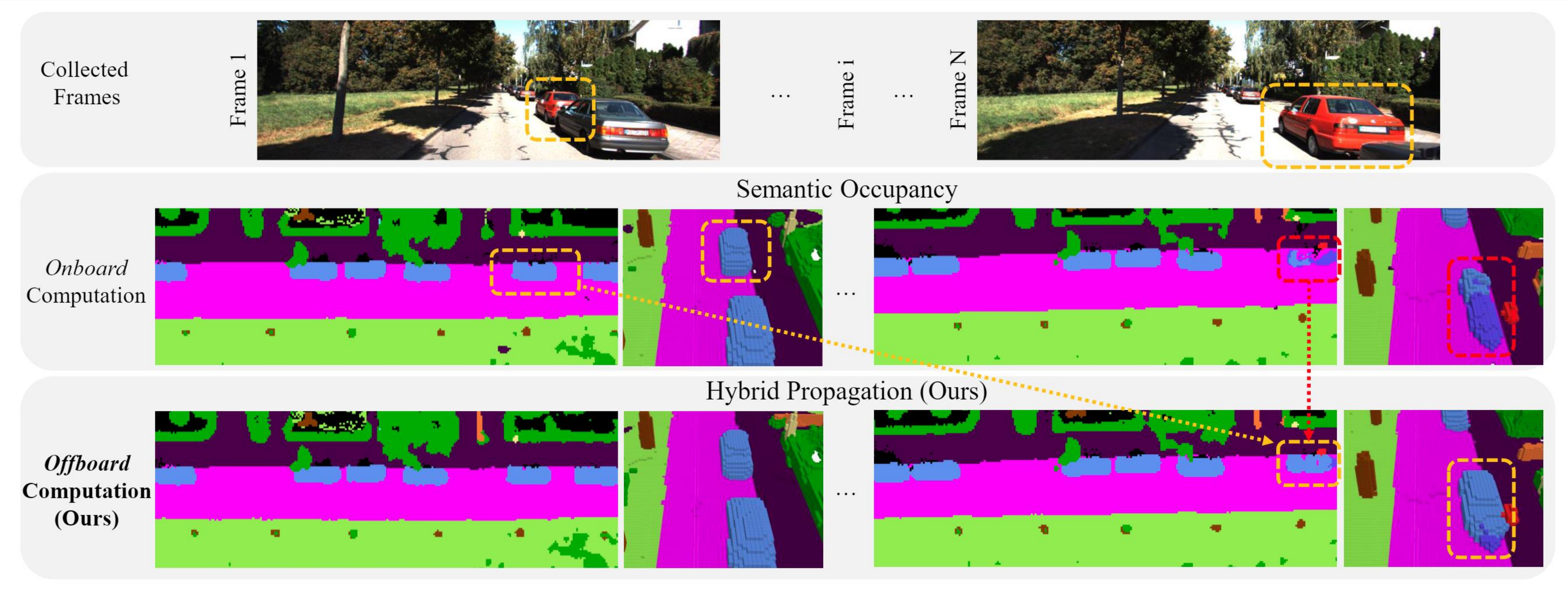


Dataset: SemanticKITTI
Input: Partial Voxel
Output: Complete Voxel Grids with Semantics

Dataset: KITTI360
Input: Front Camera RGB
Output: Complete Voxel Grids with Semantics

2. Label-efficient Occupancy Prediction

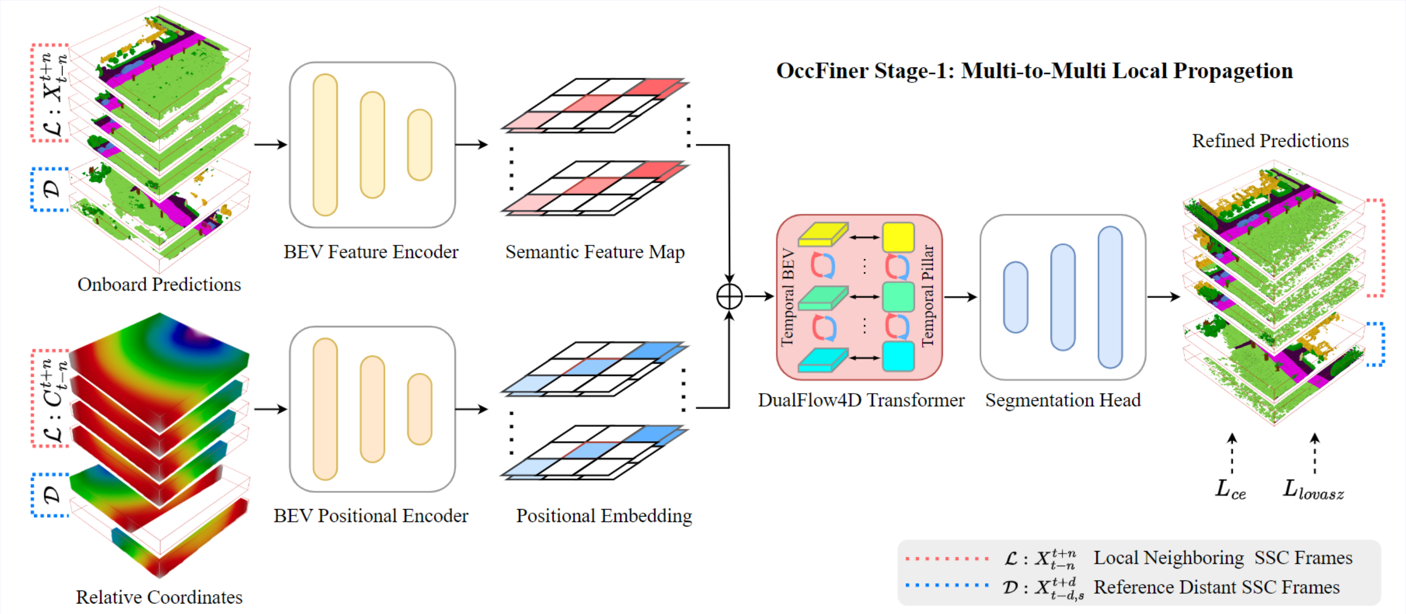
- Offboard OccFiner: Constructs unified and multi-view consistent occupancy maps, with continuity across varying viewpoints [1]

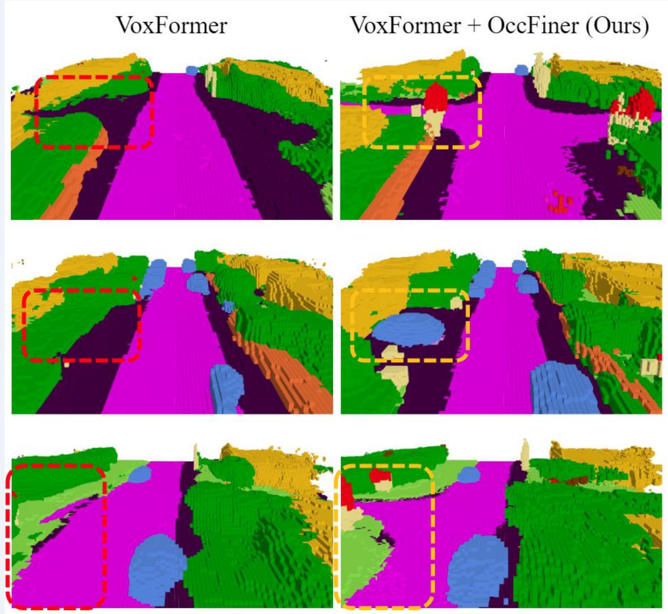


[1] Shi, Hao, et al. "Offboard Occupancy Refinement with Hybrid Propagation for Autonomous Driving." arXiv preprint arXiv:2403.08504 (2024).

2. Label-efficient Occupancy Prediction

- **Multi-to-multi local propagation network:** implicitly aligns and processes multiple local frames for correcting onboard errors
- **Region-centric global propagation:** focuses on refining labels using explicit multi-view geometry and integrating sensor bias

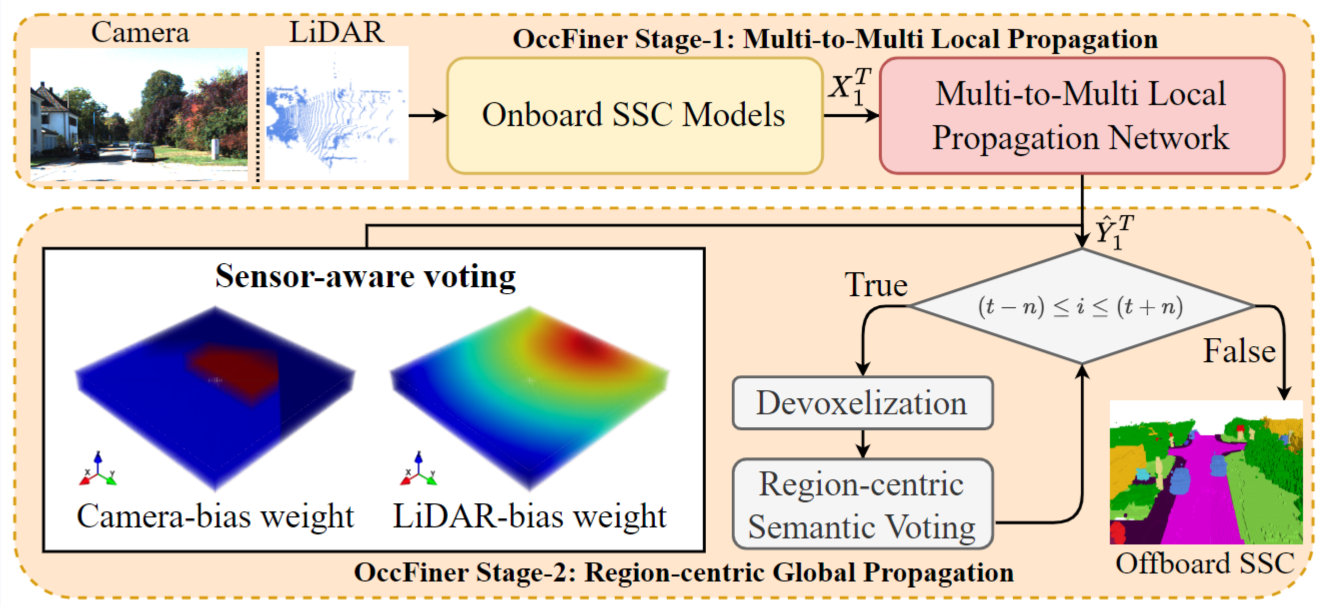


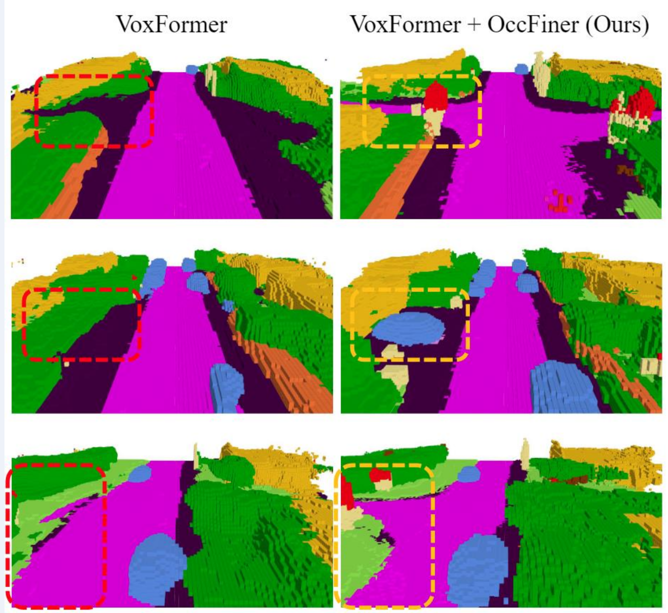


[1] Shi, Hao, et al. "Offboard Occupancy Refinement with Hybrid Propagation for Autonomous Driving." arXiv preprint arXiv:2403.08504 (2024).

2. Label-efficient Occupancy Prediction

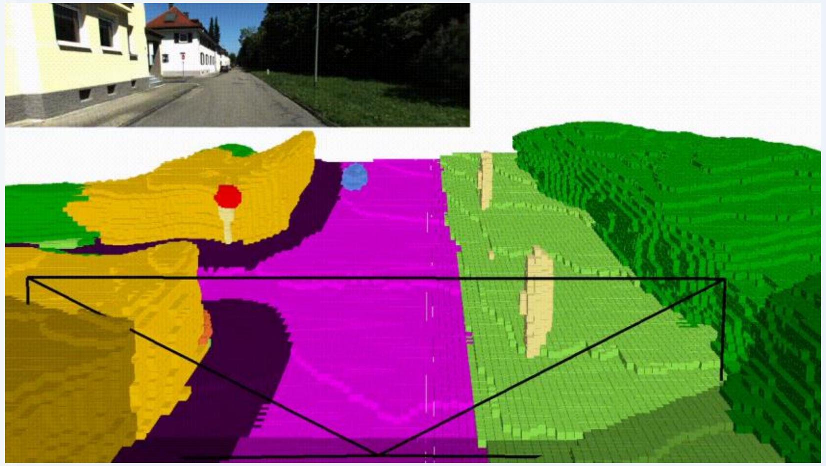
- **Multi-to-multi local propagation network:** implicitly aligns and processes multiple local frames for correcting onboard errors
- **Region-centric global propagation:** focuses on refining labels using explicit multi-view geometry and integrating sensor bias



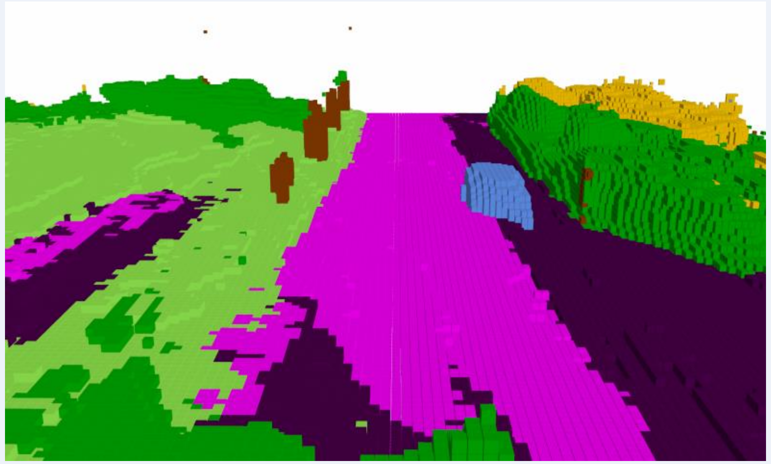


[1] Shi, Hao, et al. "Offboard Occupancy Refinement with Hybrid Propagation for Autonomous Driving." arXiv preprint arXiv:2403.08504 (2024).

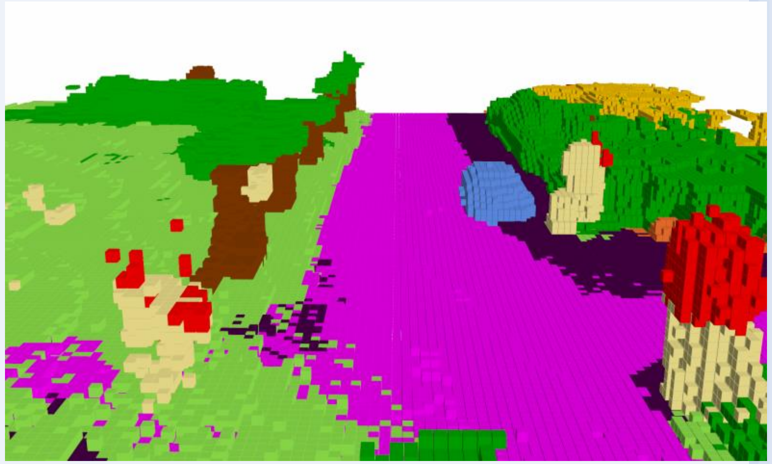
2. Label-efficient Occupancy Prediction



Semantic Scene Completion



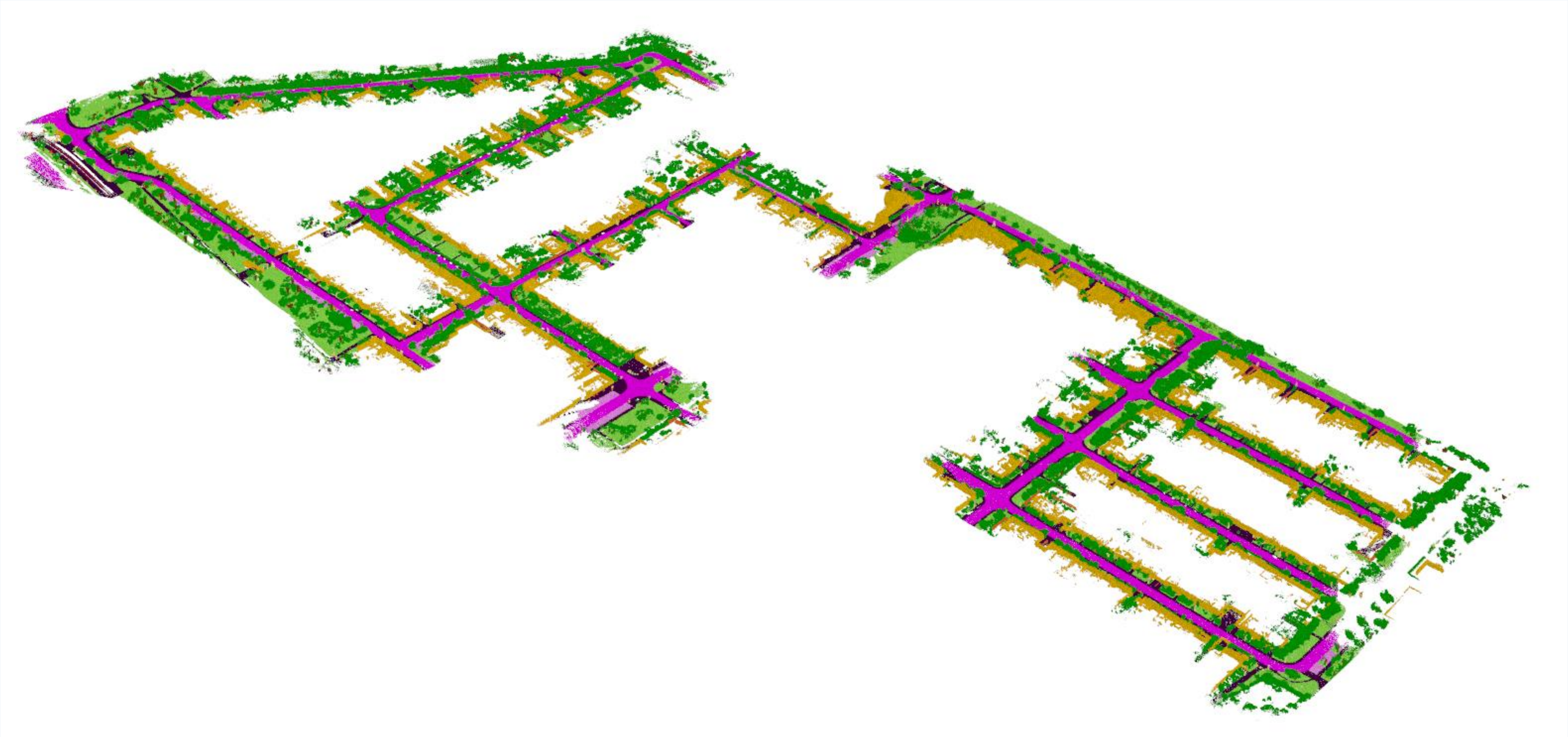
VoxFormer



OccFiner + VoxFormer (Ours)

[1] Shi, Hao, et al. "Offboard Occupancy Refinement with Hybrid Propagation for Autonomous Driving." arXiv preprint arXiv:2403.08504 (2024).

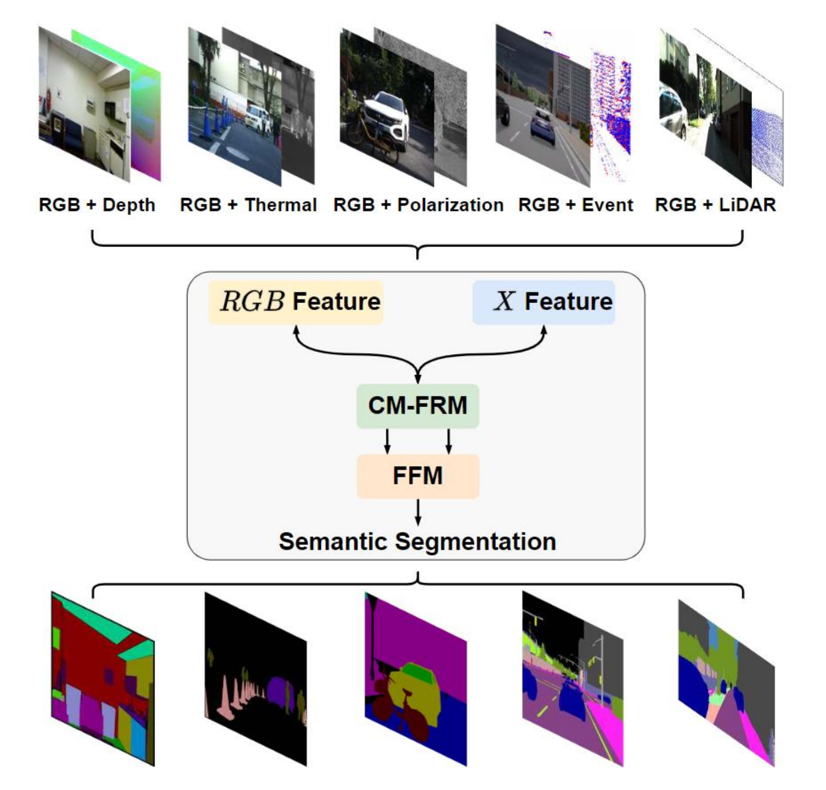
2. Label-efficient Occupancy Prediction



[1] Shi, Hao, et al. "Offboard Occupancy Refinement with Hybrid Propagation for Autonomous Driving." arXiv preprint arXiv:2403.08504 (2024).

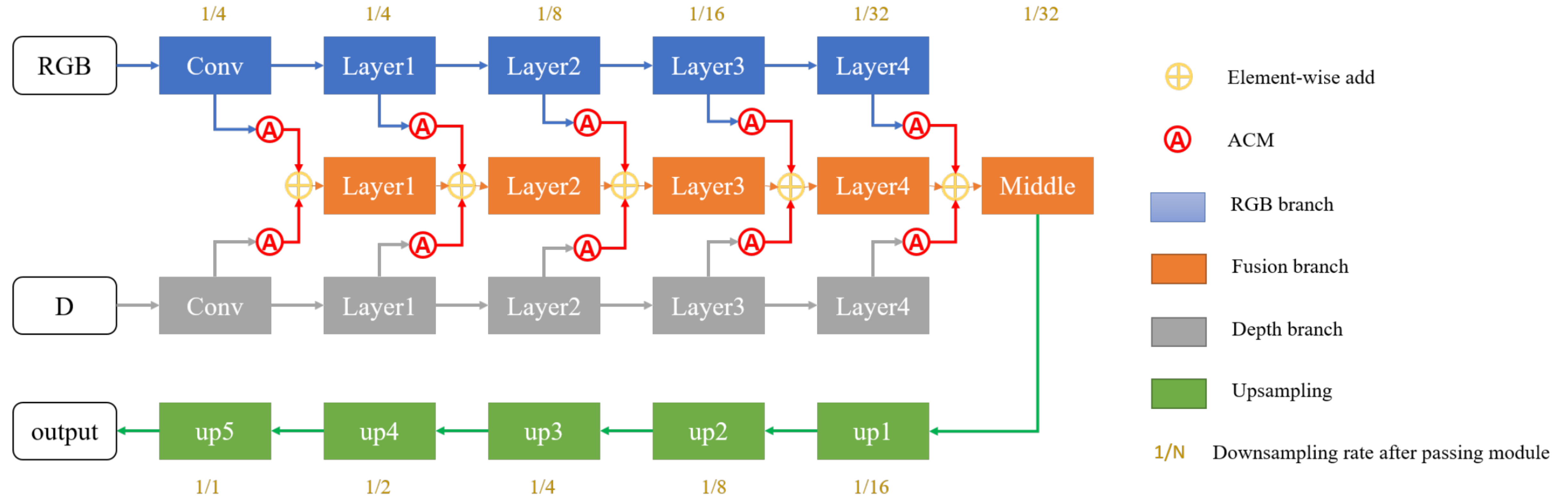
3. Arbitrary-modal Scene Segmentation

- **RGB image semantic segmentation**
  - Great progress on accuracy
  - Difficulties when objects have similar colors/textures
  - E.g., in low-illumination or high-dynamic conditions
- **RGB-X semantic segmentation**
  - Using complementary features from the X-modality
  - Depth: Geometric information
  - Thermal: Infrared information
  - Polarization: Beneficial for specular scenes
  - Event: Beneficial for dynamic scenes
  - LiDAR: Spatial information



3. Arbitrary-modal Scene Segmentation

- **ACNet:** Attention Complementary Network for RGBD Segmentation [1]



[1] Hu, Xinxin, et al. "Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation." 2019 IEEE international conference on image processing (ICIP). IEEE, 2019.

3. Arbitrary-modal Scene Segmentation

- **ACNet:** Attention Complementary Network for RGBD Segmentation [1]

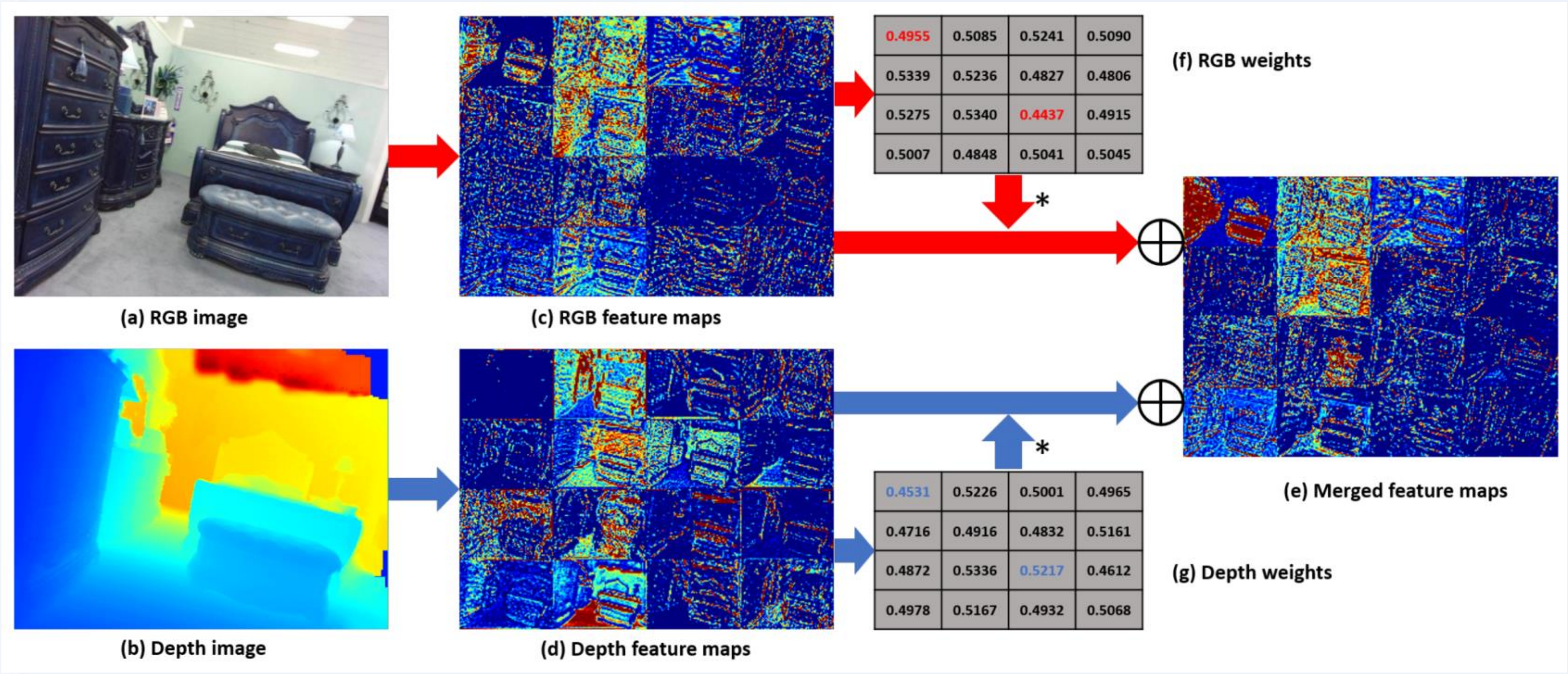


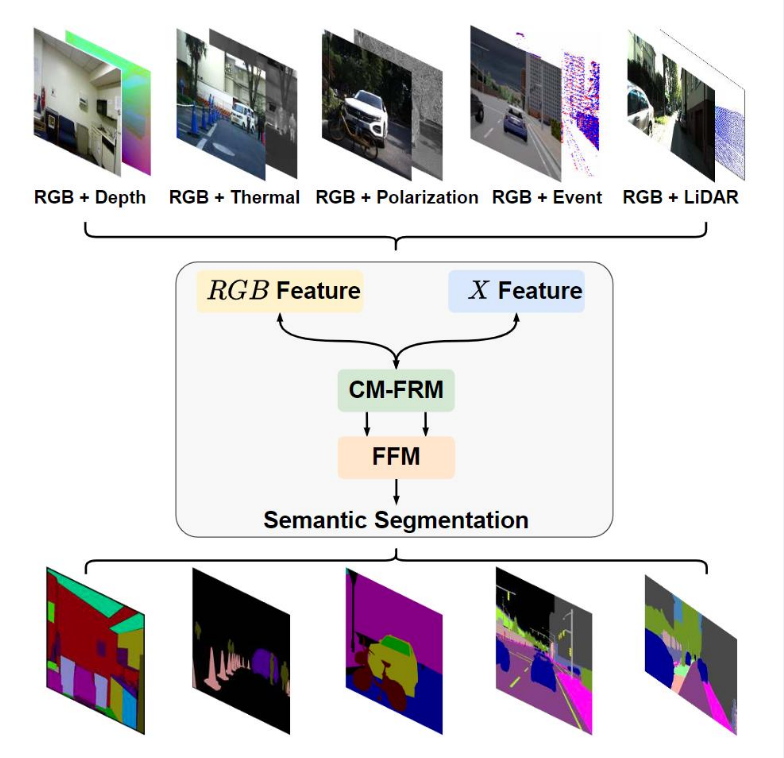
Table 1. Comparison with other state-of-the-art methods on NYUDv2 test set and SUN-RGBD test set.

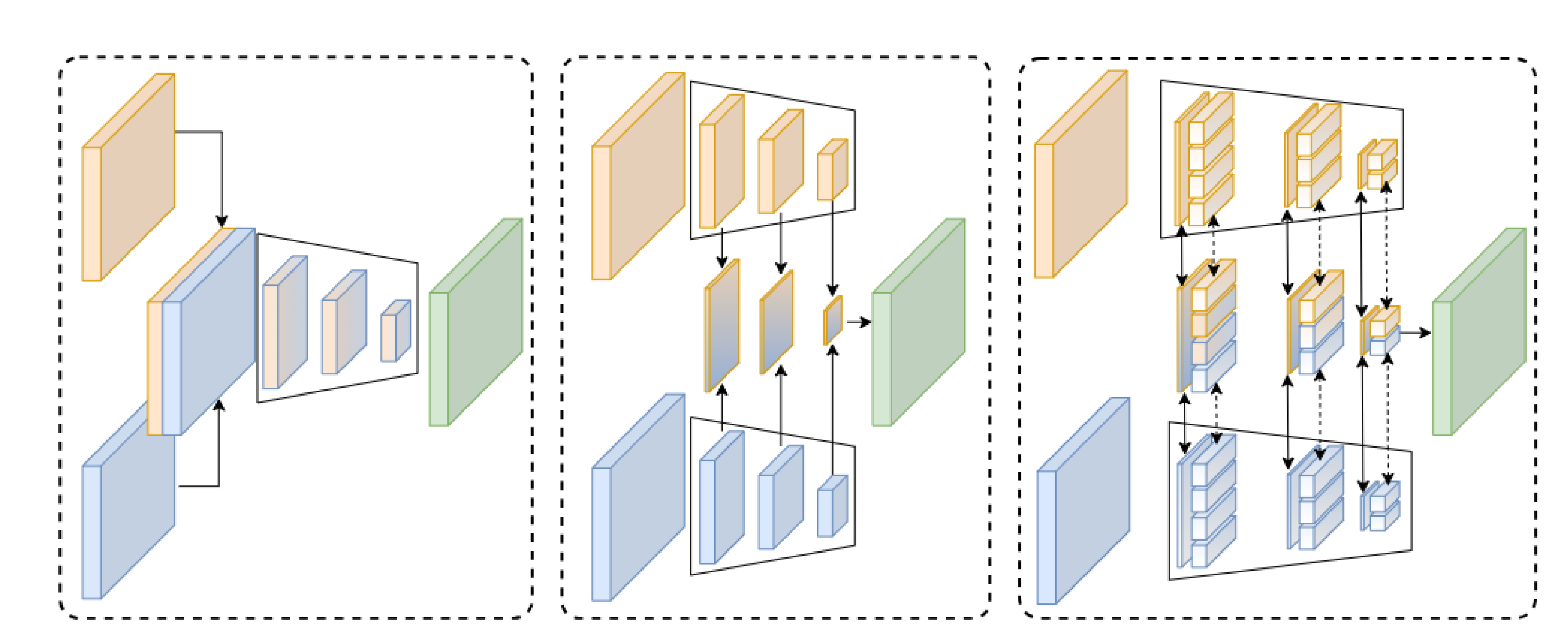
Model	NYUDv2	SUN-RGBD
3DGNN [6]	39.9%	44.1%
RefineNet (ResNet152) [17]	46.5%	45.9%
Depth-aware CNN [7]	43.9%	42.0%
LSD [8]	45.9%	-
CFN (VGG-16) [18]	41.7%	42.5%
CFN (RefineNet-152) [18]	47.7%	**48.1%**
ACNet (ResNet-50)	**48.3%**	**48.1%**

[1] Hu, Xinxin, et al. "Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation." 2019 IEEE international conference on image processing (ICIP). IEEE, 2019.

3. Arbitrary-modal Scene Segmentation

- **CMX:** Cross-Modal Fusion for RGB-X Semantic Segmentation [1]



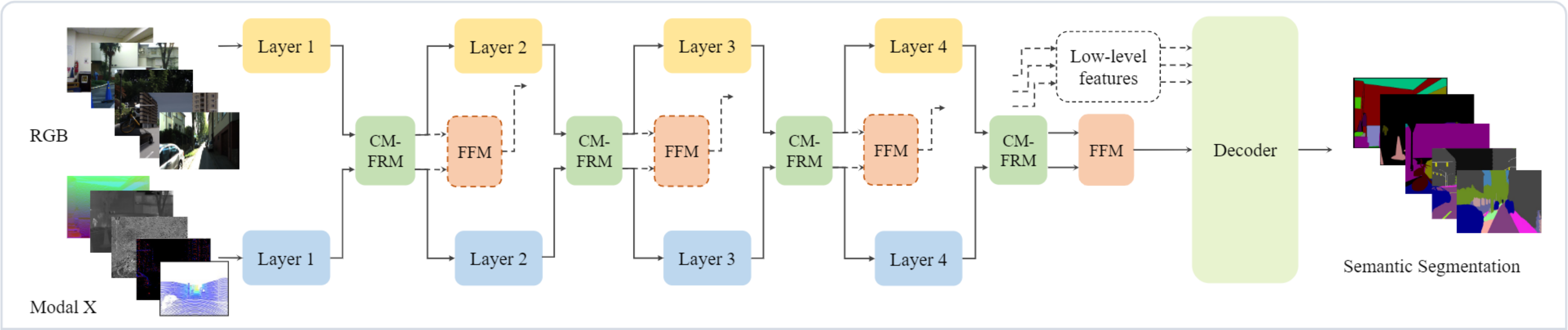


(a) Input fusion (b) Feature fusion (c) Interactive fusion

[1] Zhang, Jiaming, et al. "CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers." IEEE Transactions on intelligent transportation systems (2023).

3. Arbitrary-modal Scene Segmentation

- **CMX:** Cross-Modal Fusion for RGB-X Semantic Segmentation [1]
- Feature-wise rectification using cross-modal information
  - Channel-wise and spatial-wise rectification
- Sequence-wise interaction using cross-modal information
  - Cross-attention and mixed channel embedding

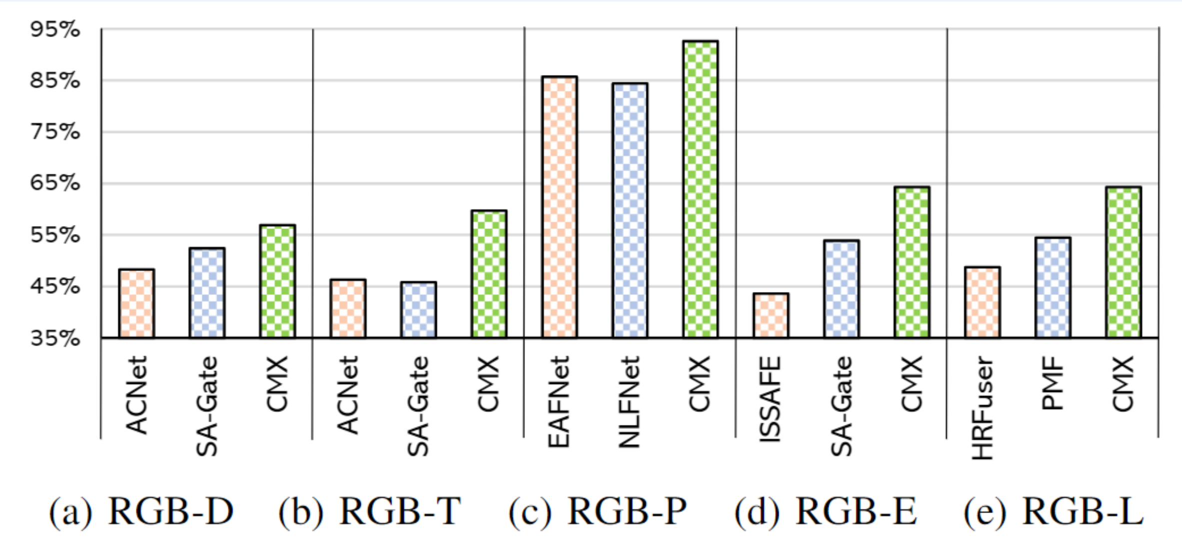


a) Overall Framework

[1] Zhang, Jiaming, et al. "CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers." IEEE Transactions on intelligent transportation systems (2023).

3. Arbitrary-modal Scene Segmentation

- **CMX:** Cross-Modal Fusion for RGB-X Semantic Segmentation [1]

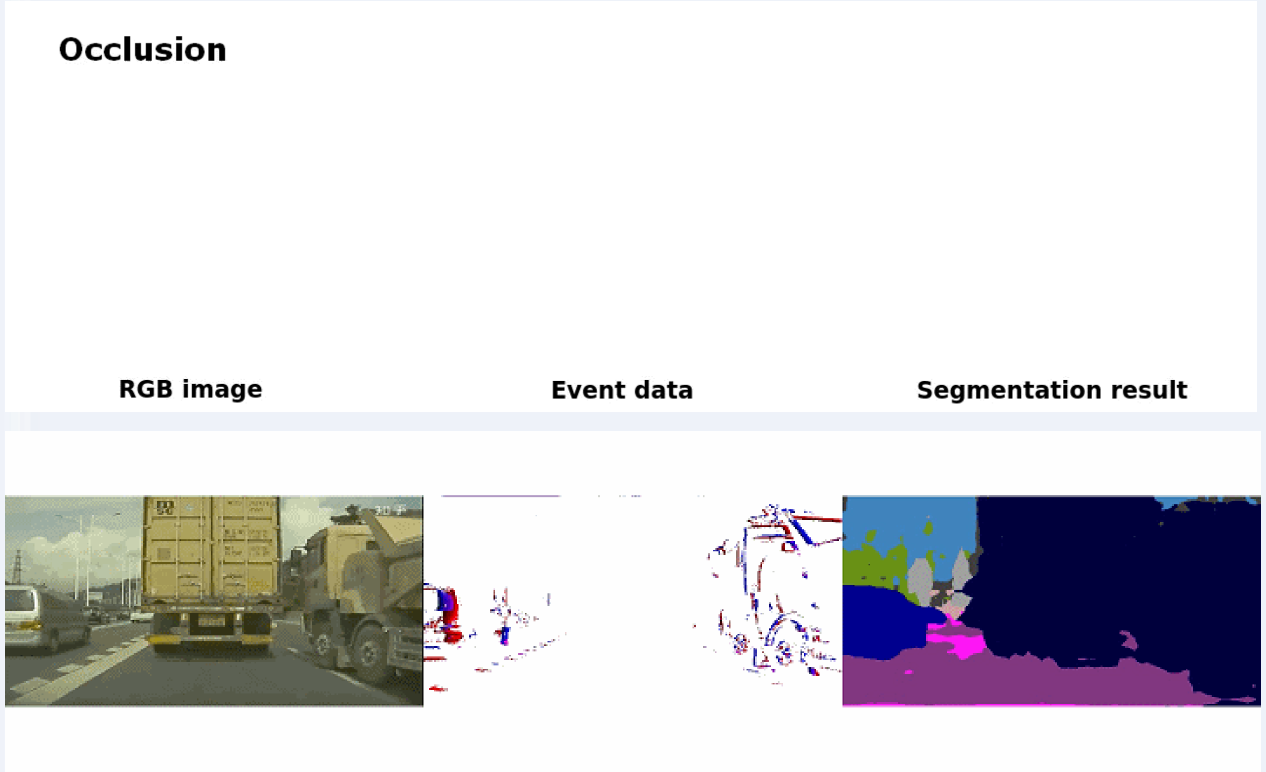


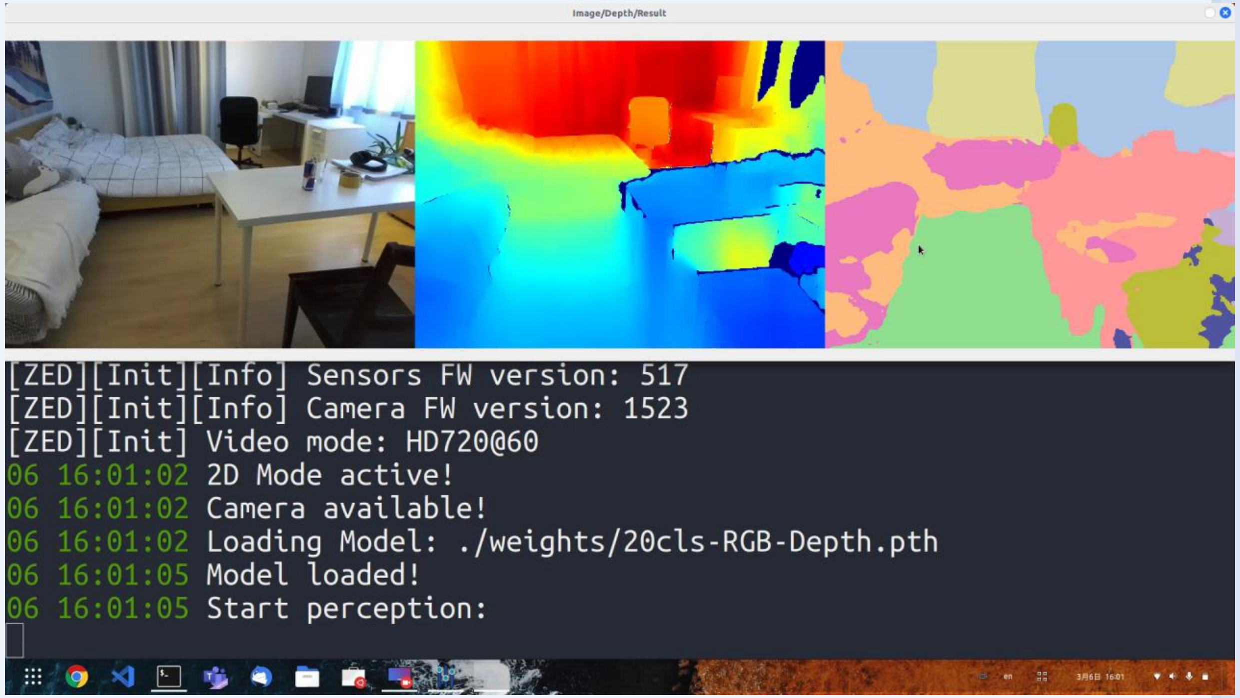
Method	Modal	NYU Depth V2	Cityscapes	MFNet	ZJU-RGB-P	EventScape	KITTI-360
SegFormer-B2 [33]	RGB-only	48.0	81.0	53.2	89.6	58.7	61.3
CMX-B2	Multimodal	54.1 (RGB-D)	81.6 (RGB-D)	58.2 (RGB-T)	92.2 (RGB-P)	61.9 (RGB-E)	64.3 (RGB-L)

[1] Zhang, Jiaming, et al. "CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers." IEEE Transactions on intelligent transportation systems (2023).

3. Arbitrary-modal Scene Segmentation

- **CMX:** Applications in driving and walking assistance [1]

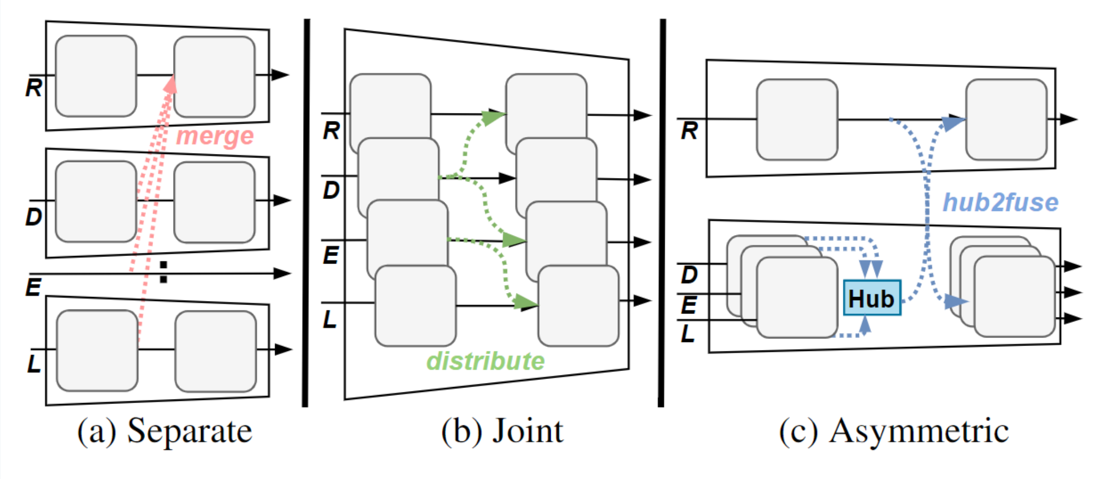


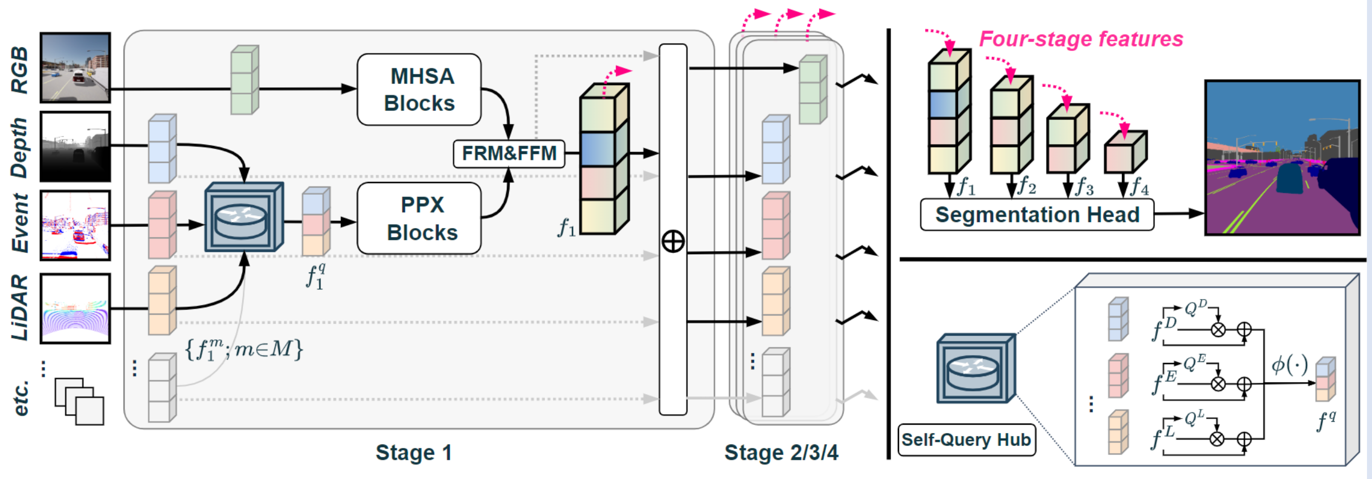


[1] Zhang, Jiaming, et al. "CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers." IEEE Transactions on intelligent transportation systems (2023).

3. Arbitrary-modal Scene Segmentation

- **CMNeXt:** Asymmetric fusion for arbitrary-modal segmentation [1]

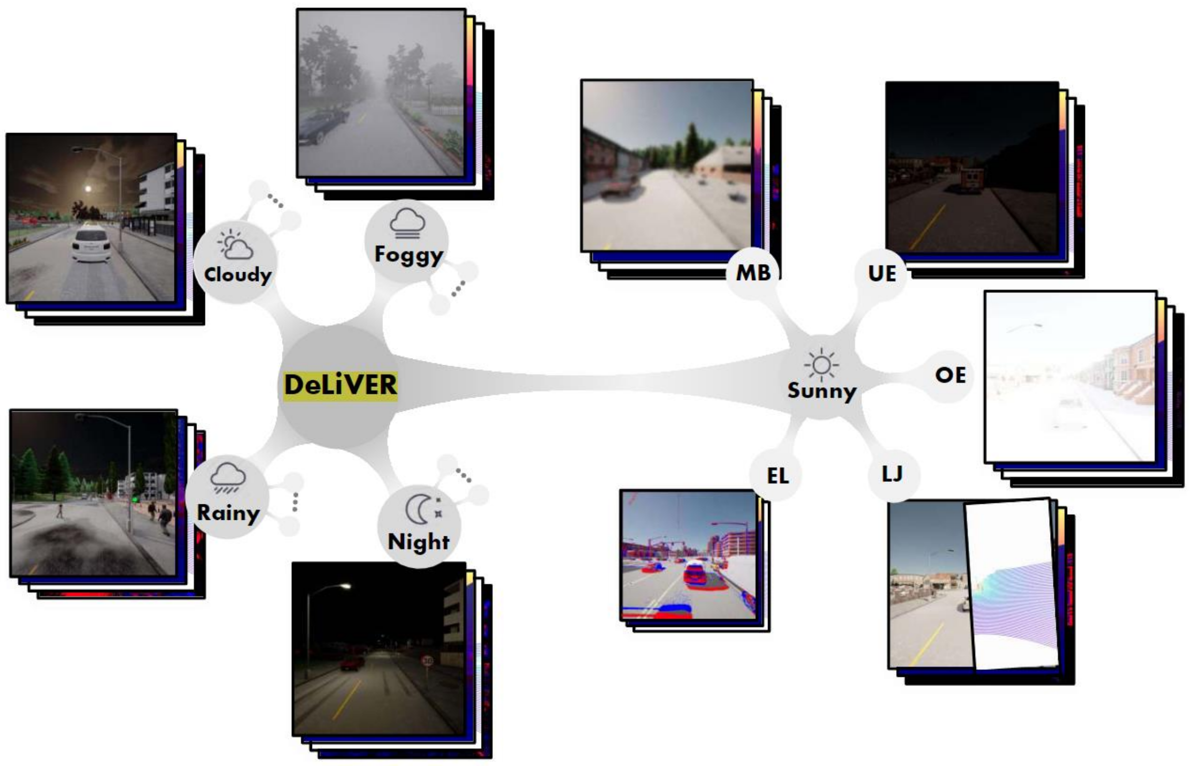




[1] Zhang, Jiaming, et al. "Delivering arbitrary-modal semantic segmentation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

3. Arbitrary-modal Scene Segmentation

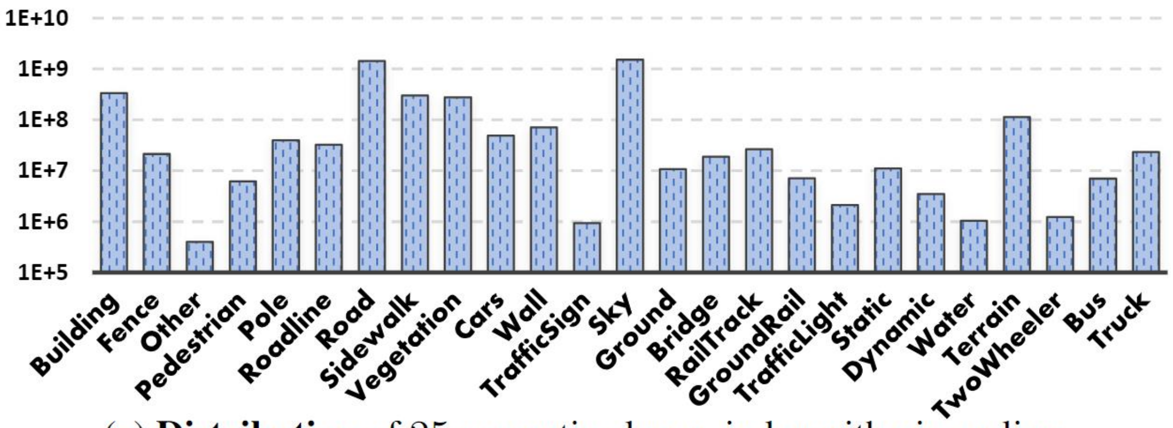
- **DeLiVER:** Arbitrary-modal segmentation benchmark [1]



(a) **Structure and samples** of four adverse conditions and five failure cases.

(b) **Statistic** of different data splits and views.

Split	Cloudy	Foggy	Night	Rainy	Sunny	Normal	Corner	Total
Train	794	795	797	799	798	2585	1398	3983
Val	398	400	410	398	399	1298	707	2005
Test	379	379	379	380	380	1198	699	1897
Front-view	1571	1574	1586	1577	1577	5081	2804	7885
All six views	9426	9444	9516	9462	9462	30486	16824	47310



(c) **Distribution** of 25 semantic classes in logarithmic scaling.

[1] Zhang, Jiaming, et al. "Delivering arbitrary-modal semantic segmentation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

3. Arbitrary-modal Scene Segmentation

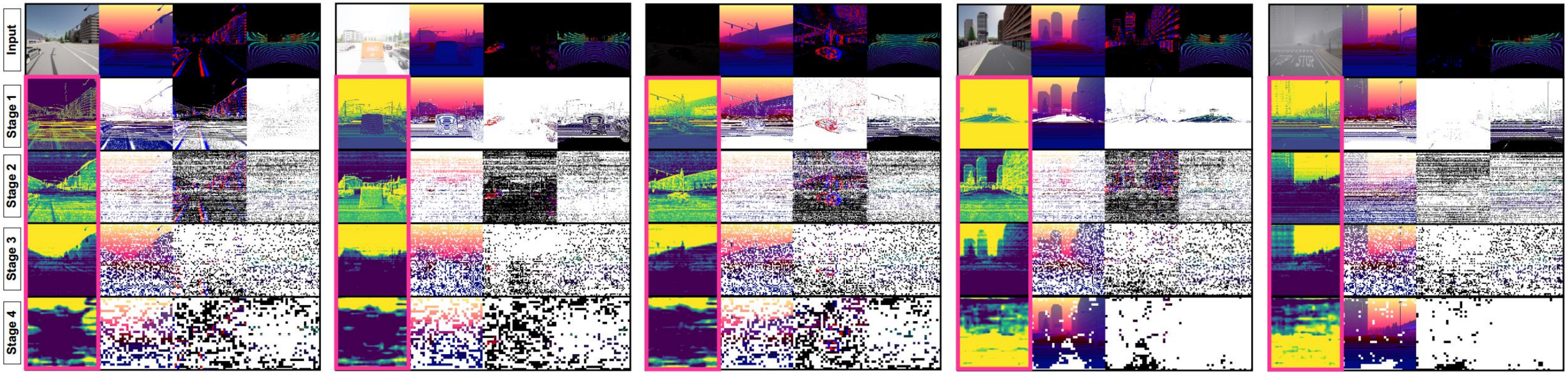
- **DeLiVER:** Arbitrary-modal segmentation benchmark [1]

Model-modality	#Params(M)	GFLOPs	Cloudy	Foggy	Night	Rainy	Sunny	MB	OE	UE	LJ	EL	Mean
HRFuser-RGB	29.89	217.5	49.26	48.64	42.57	50.61	50.47	48.33	35.13	26.86	49.06	49.88	47.95
SegFormer-RGB	25.79	38.93	59.99	57.30	50.45	58.69	60.21	57.28	56.64	37.44	57.17	59.12	57.20
TokenFusion-RGB-D	26.01	54.96	50.92	52.02	43.37	50.70	52.21	49.22	46.22	36.39	49.58	49.17	49.86
CMX-RGB-D	66.57	65.68	63.70	62.77	60.74	62.37	63.14	59.50	60.14	55.84	62.65	63.26	62.66
HRFuser-RGB-D	30.46	223.0	54.80	51.48	49.51	51.55	52.12	50.92	41.51	44.00	54.10	52.52	51.88
HRFuser-RGB-D-E	31.04 **(+0.57)**	229.0 **(+6.00)**	54.04	50.83	50.88	51.13	52.61	49.32	41.75	47.89	54.65	52.33	51.83
HRFuser-RGB-D-E-L	31.61 **(+0.57)**	235.0 **(+6.00)**	56.20	52.39	49.85	52.53	54.02	49.44	46.31	46.92	53.94	52.72	52.97
CMNeXt-RGB-D	58.69	62.94	67.21	62.79	61.64	62.95	65.26	61.00	64.64	58.71	64.32	63.35	63.58
CMNeXt-RGB-D-E	58.72 **(+0.03)**	64.19 **(+1.25)**	68.28	63.28	62.64	63.01	66.06	62.58	64.44	58.73	65.37	65.80	64.44
CMNeXt-RGB-D-E-L	58.73 **(+0.01)**	65.42 **(+1.23)**	68.70	65.67	62.46	67.50	66.57	62.91	64.59	60.00	65.92	65.48	66.30
w.r.t. SegFormer-RGB			**(+8.71)**	**(+8.37)**	**(+12.01)**	**(+8.81)**	**(+6.36)**	**(+5.63)**	**(+7.95)**	**(+22.56)**	**(+8.75)**	**(+6.36)**	**(+9.10)**

[1] Zhang, Jiaming, et al. "Delivering arbitrary-modal semantic segmentation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

3. Arbitrary-modal Scene Segmentation

- **DeLiVER:** Arbitrary-modal segmentation benchmark [1]

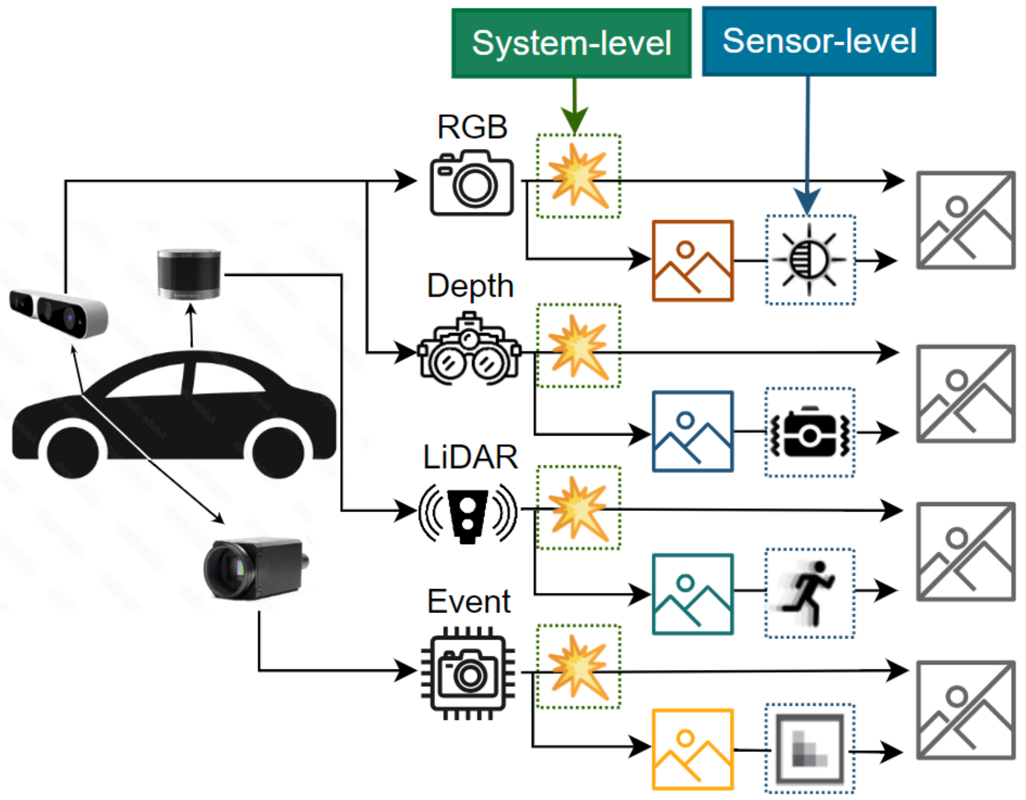


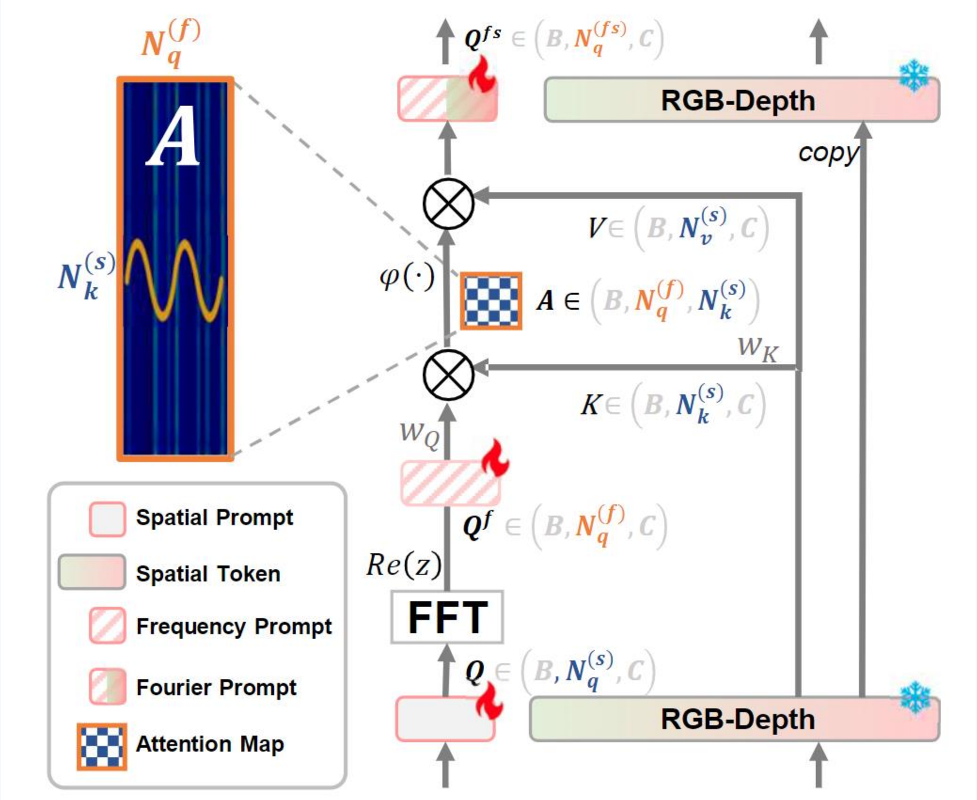
(a) Motion Blur. (b) Over-Exposure. (c) Under-Exposure. (d) LiDAR Jitter. (e) Event Low-resolution.

[1] Zhang, Jiaming, et al. "Delivering arbitrary-modal semantic segmentation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

3. Arbitrary-modal Scene Segmentation

- Fourier Prompt Tuning for **Modality-Incomplete Scene Segmentation** [1]

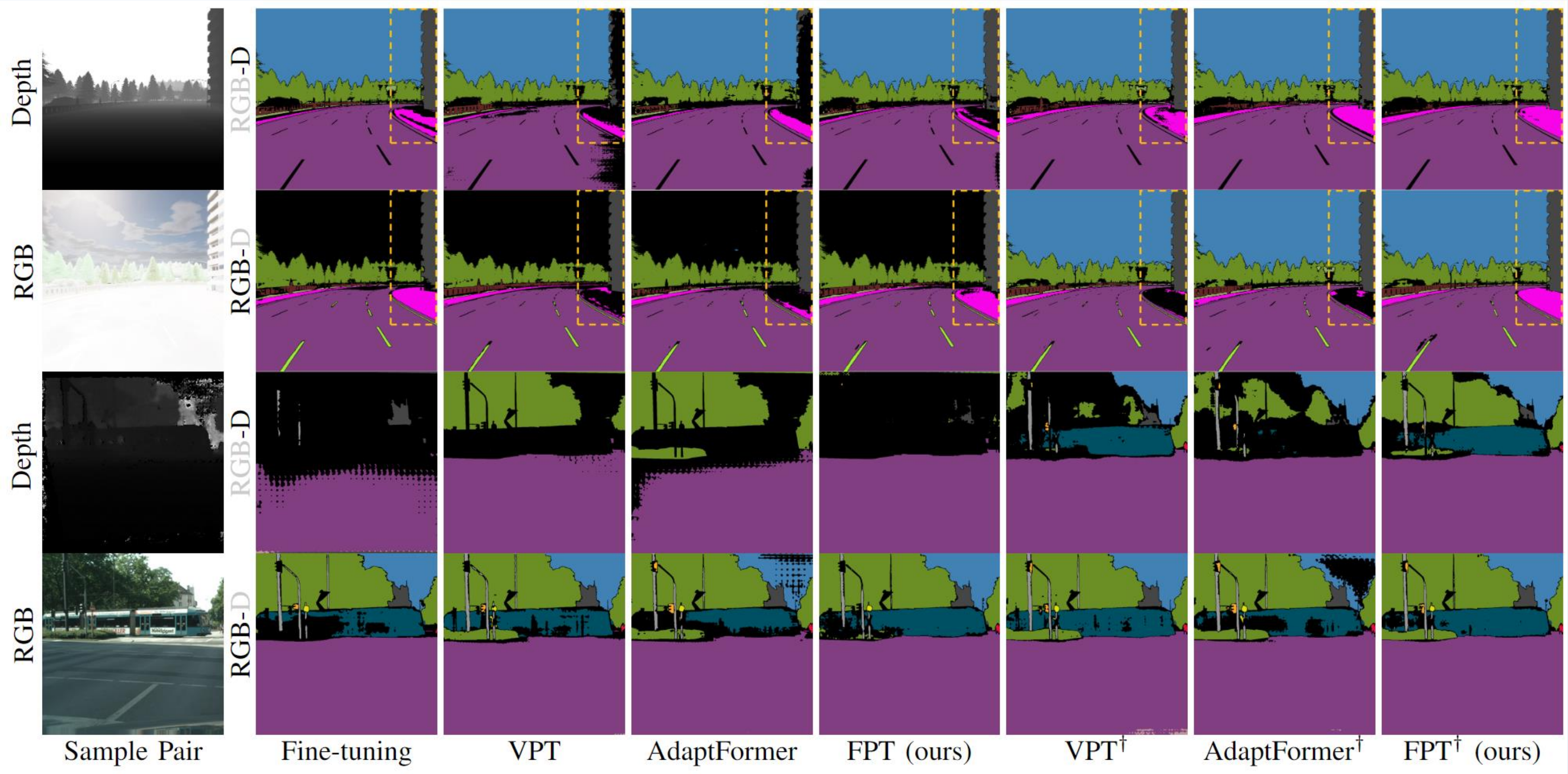




[1] Liu, Ruiping, et al. "Fourier Prompt Tuning for Modality-Incomplete Scene Segmentation." IEEE Intelligent Vehicles Symposium (IV), 2024.

3. Arbitrary-modal Scene Segmentation

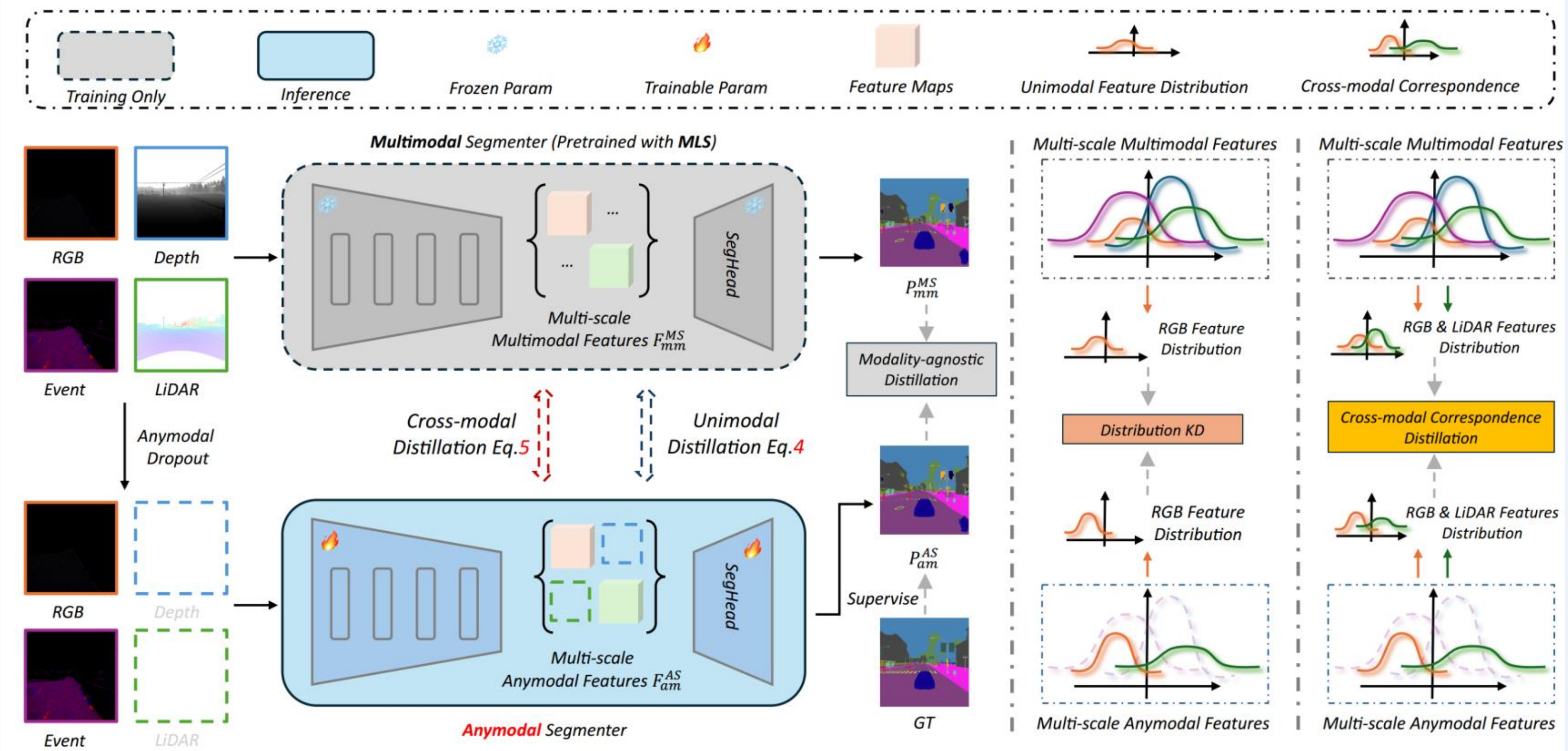
- Fourier Prompt Tuning for **Modality-Incomplete Scene Segmentation** [1]

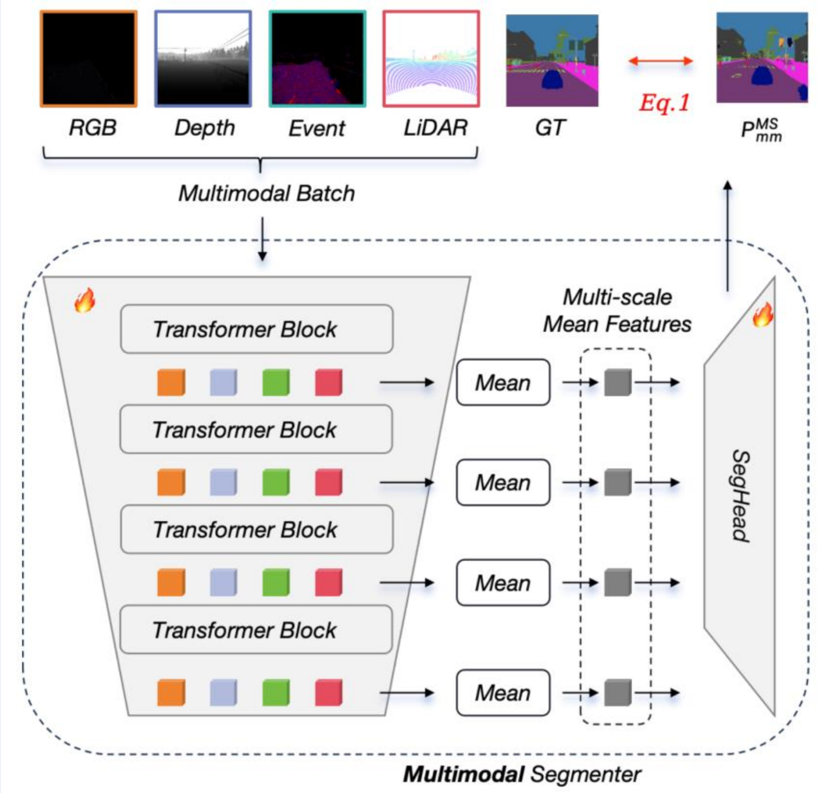


[1] Liu, Ruiping, et al. "Fourier Prompt Tuning for Modality-Incomplete Scene Segmentation." IEEE Intelligent Vehicles Symposium (IV), 2024.

3. Arbitrary-modal Scene Segmentation

- **Anymodal Segmentor:** Learning Robust Anymodal Segmentor with Unimodal and Cross-modal Distillation [1]





[1] Zheng, Xu, et al. "Learning Robust Anymodal Segmentor with Unimodal and Cross-modal Distillation." arXiv preprint arXiv:2411.17141 (2024).

3. Arbitrary-modal Scene Segmentation

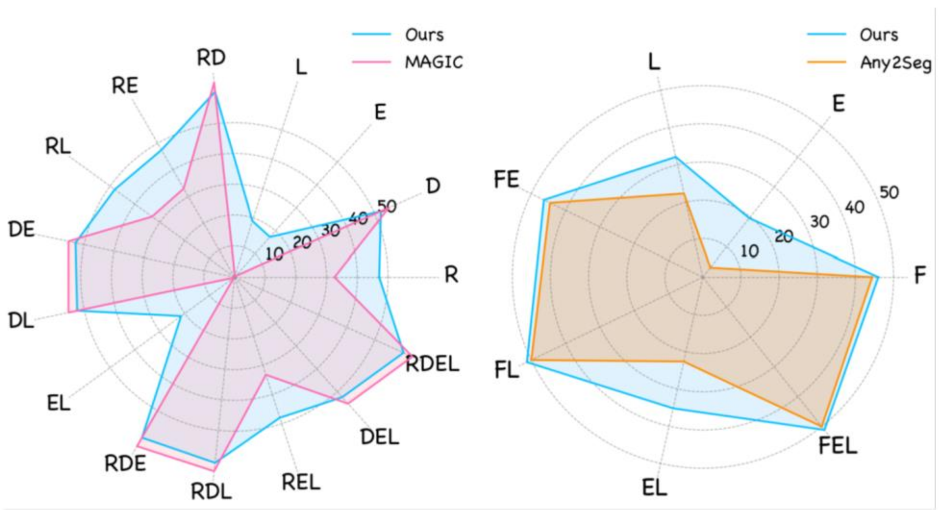
- **Anymodal Segmentor:** Learning Robust Anymodal Segmentor with Unimodal and Cross-modal Distillation [1]

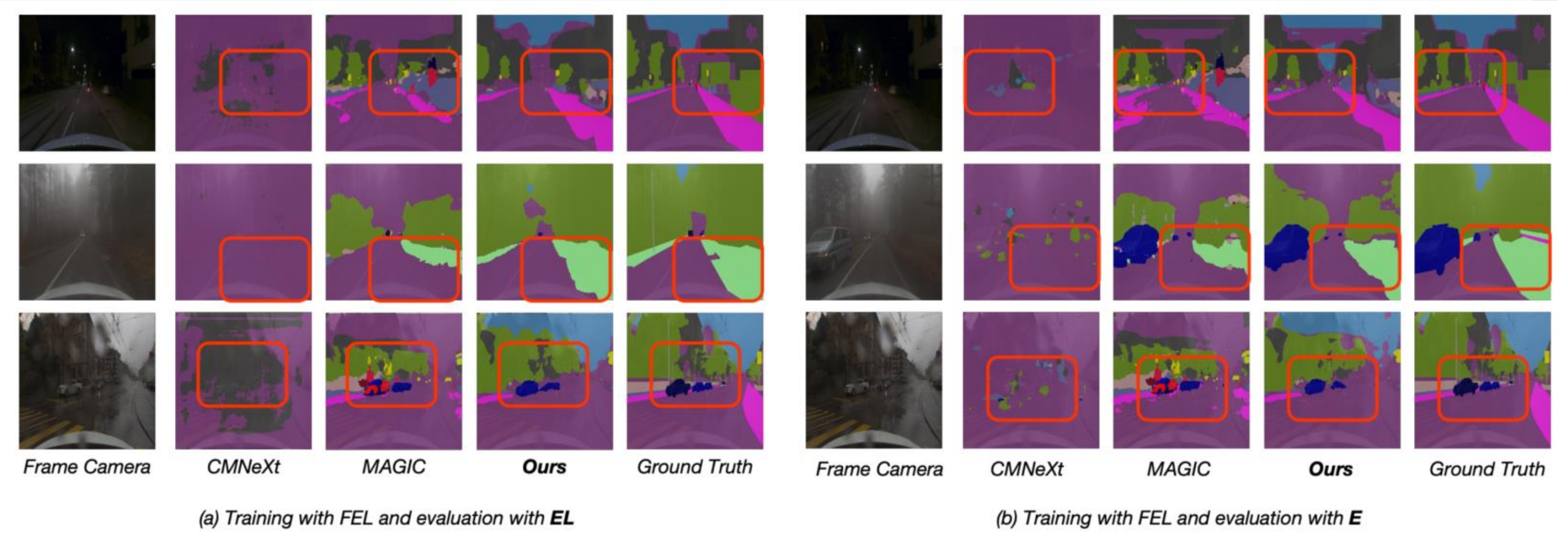
Method	Pub.	Training	Anymodal Evaluation							Mean
			F	E	L	FE	FL	EL	FEL	
CMX (Zhang et al., 2023a)	T-ITS 2023	FEL	2.52	2.35	3.01	41.15	41.25	2.56	42.27	19.30
CMNeXt (Zhang et al., 2023b)	CVPR 2023		3.50	2.77	2.64	6.63	10.28	3.14	46.66	10.80
MAGIC (Zheng et al., 2024b)	ECCV 2024		43.22	2.68	22.95	43.51	49.05	22.98	49.02	33.34
Any2Seg (Zheng et al., 2024a)	ECCV 2024		44.40	3.17	22.33	44.51	49.96	22.63	50.00	33.86
Ours	-		**46.01**	**19.57**	**32.13**	**46.29**	**51.25**	**35.21**	**51.14**	**40.23**
w.r.t SoTA	-	-	**+1.61**	**+16.40**	**+9.80**	**+1.78**	**+1.29**	**+12.58**	**+1.14**	**+6.37**

[1] Zheng, Xu, et al. "Learning Robust Anymodal Segmentor with Unimodal and Cross-modal Distillation." arXiv preprint arXiv:2411.17141 (2024).

3. Arbitrary-modal Scene Segmentation

- **Anymodal Segmentor:** Learning Robust Anymodal Segmentor with Unimodal and Cross-modal Distillation [1]





[1] Zheng, Xu, et al. "Learning Robust Anymodal Segmentor with Unimodal and Cross-modal Distillation." arXiv preprint arXiv:2411.17141 (2024).

Summary

- Towards holistic scene understanding by overcoming the limit in field of view, annotations, and cross-modal fusion

- **All works are with open codes!**

- Future perspectives
  - Leverage SAM's capabilities to address cross-modal inconsistencies
  - Roadside and V2X collaborative perception using complementary representations
  - World models for predicting future semantic occupancy estimation

Thanks!

https://yangkailun.com

kailun.yang@hnu.edu.cn