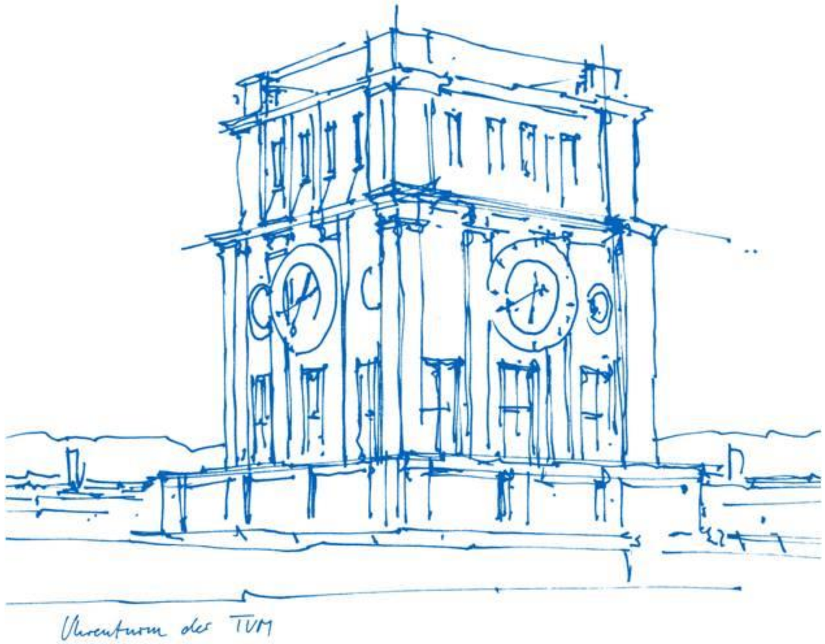
Learning-based Multi-modal Perception

Aleksandar Jevtić

Seminar: Robot Perception and Intelligence

Advisor: Dr. Jaehyung Jung

Munich, 16th of January 2024



Introduction

Learning-based Multi-modal ***Perception***

“The process of perception involves making useful ***models of the environment*** *from a confusion mass of sensory input data”.*

- Semantic segmentation
- Object detection / tracking
- Pose estimation
- *(…)*

Source: R. Nevatia, *Machine perception*. Prentice Hall, 1982.

Introduction

Where do we need Machine Perception?

Many use relevant use cases, e.g.,

- Robotics
- Autonomous Vehicles
- Healthcare

Source: https://www.cnet.com/home/this-robot-isnt-going-to-replace-your-in-home-nurse-yet/
https://venturebeat.com/ai/waymos-autonomous-cars-have-driven-20-million-miles-on-public-roads/

Introduction

Learning-based Multi-modal Perception

*“The term **multimodality** refers to an individual’s use of **different modes** (i.e. channels of communication) for the purpose of conveying meaning.”.*

- RGB images
- Depth
  - dense
  - sparse (e.g. LiDAR)
- Thermal imaging

(non-structural)

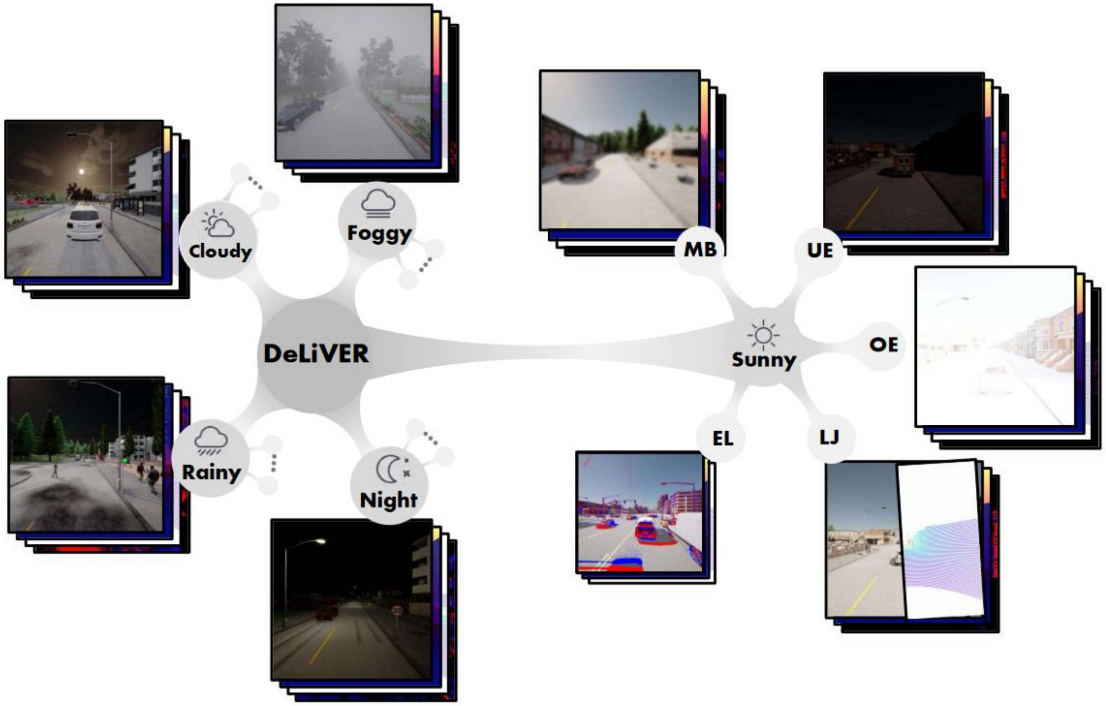
- IMU
- Audio
- Language

Source: M. J. Kessler, “Multimodality,” *ELT Journal*, vol. 76, no. 4, pp. 551–554, Jul. 2022.

Introduction

How does multimodality help?

- Better accuracy
- Robustness
  - Adverse conditions
  - Failure cases



Source: J. Zhang et al., "Delivering Arbitrary-Modal Semantic Segmentation." CVPR, 2023.

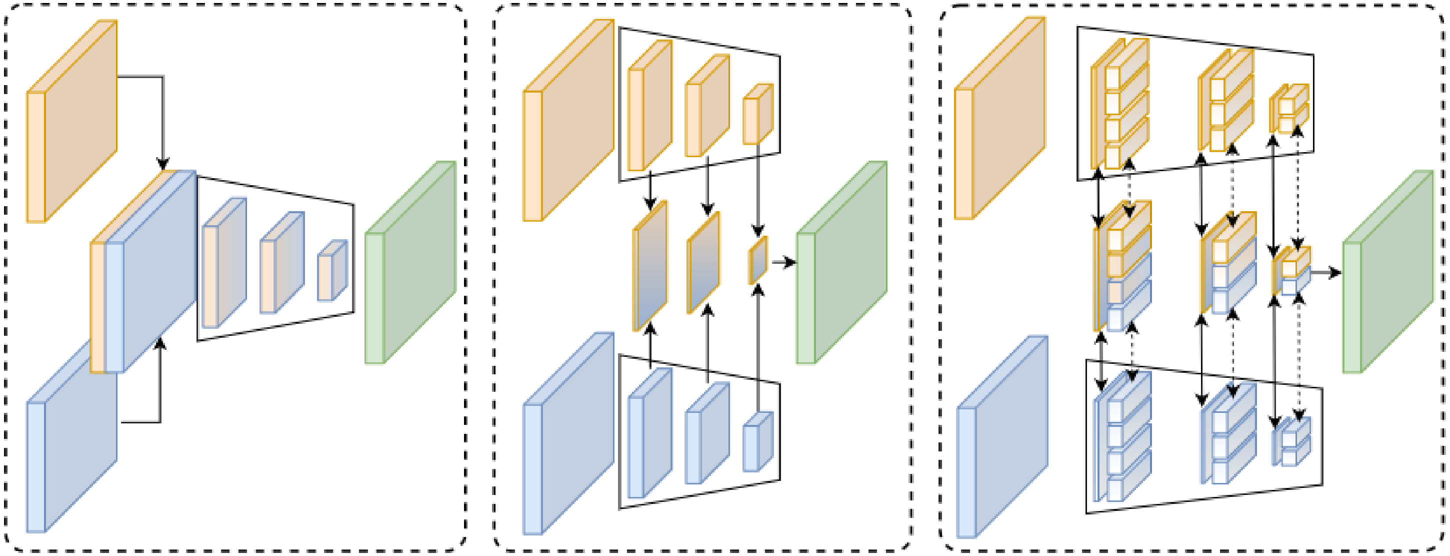
Overview

- Introduction and Overview
- Related Work
- Method Descriptions and Results
  - Multi-modal curb detection
  - CMX and CMNeXt
  - Multi-modal knowledge expansion
- Personal Comments
- Future Work

Related Work

Cross-Modal Fusion

(e.g. ShapeConv) (e.g. SA-Gate)



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

Source: J. Cao, H. Leng, D. Lischinski, D. Cohen-Or, C. Tu, and Y. Li, “ShapeConv: Shape-Aware Convolutional Layer for Indoor RGB-D Semantic Segmentation,” ICCV, 2021.
X. Chen et al., “Bi-directional Cross-Modality Feature Propagation with Separation-and-Aggregation Gate for RGB-D Semantic Segmentation,” ECCV, 2020.

Related Work

High-level: Leveraging RGB data and models

Related concept: Semi-supervised Learning

- Consistency regularization

  Small input and model perturbations → small output changes

  Additional loss term

- Pseudo-labeling

  Teacher – Student Architecture

  Generation of labels for unlabeled data

Source: E. Arazo, D. Ortego, P. Albert, N. E. O'Connor and K. McGuinness, "Pseudo-Labeling and Confirmation Bias in Deep Semi-Supervised Learning," IJCNN, 2020.

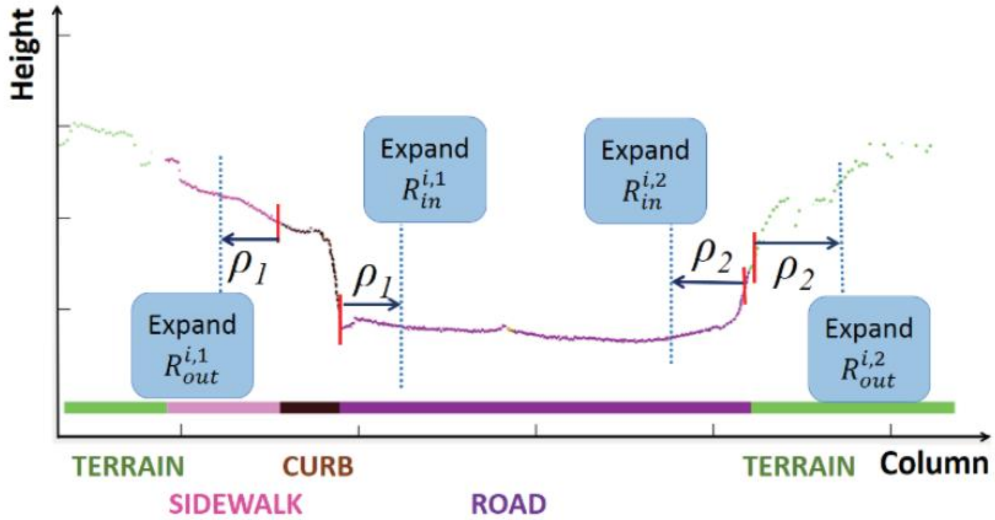
Related Work

Curb Detection Methods with LiDAR

Segmentation in RGB to find Regions of Interest

in these ROIs:
use engineered spatial features

Segmentation is learning-based, but not fusion!



Source: S. E. C. Goga and S. Nedevschi, "Fusing semantic labeled camera images and 3D LiDAR data for the detection of urban curbs," ICCP, 2018.

Multi-modal curb detection and filtering

Sandipan Das[1,2], Navid Mahabadi[2], Saikat Chatterjee[1], Maurice Fallon[3]

[1] KTH EECS, Sweden. {sandipan,sach}@kth.se
[2] Scania, Sweden. {sandipan.das,navid.mahabadi}@scania.com
[3] Oxford Robotics Institute, UK. mfallon@robots.ox.ac.uk

Multi-modal curb detection and filtering

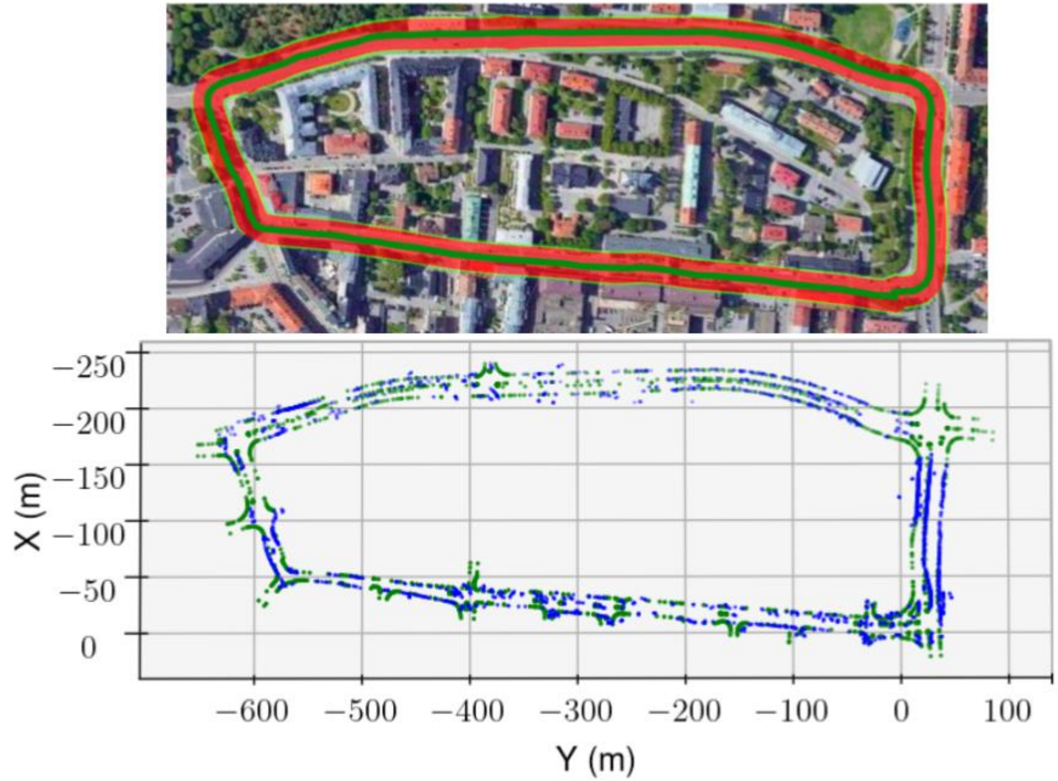
Detection of curb points by unsupervised clustering

Multi-modal fusion of

- RGB
- LiDAR

Data collection vehicle

- 4 sensors
- varying FoVs



Detected curb features (blue) and ground truth (green)

Source: S. Das, N. Mahabadi, S. Chatterjee, and M. Fallon, “Multi-modal curb detection and filtering,” *CoRR*, vol. abs/2205.07096, 2022.

Multi-modal curb detection and filtering

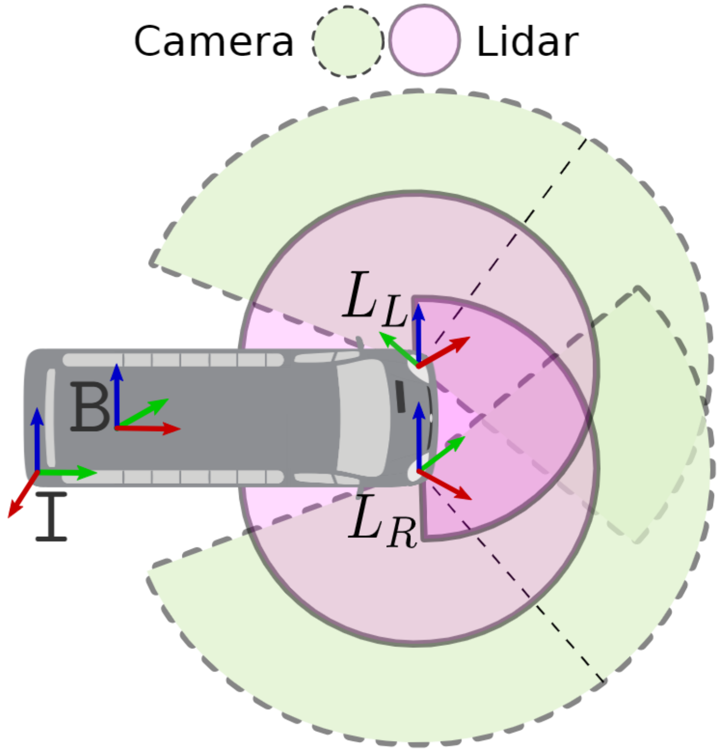
Detection of curb points by unsupervised clustering

Multi-modal fusion of

- RGB
- LiDAR

Data collection vehicle

- 4 sensors
- varying FoVs



Source: S. Das, N. Mahabadi, S. Chatterjee, and M. Fallon, "Multi-modal curb detection and filtering," *CoRR*, vol. abs/2205.07096, 2022.

Multi-modal Curb detection - Method

Curb segmentation on RGB

- EfficientNet

Association with LiDAR

Unsupervised clustering

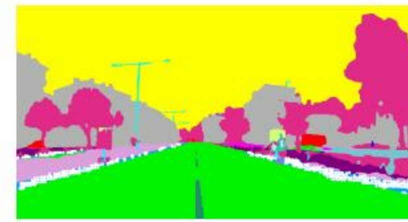
- DBSCAN (density-based)

Filtering

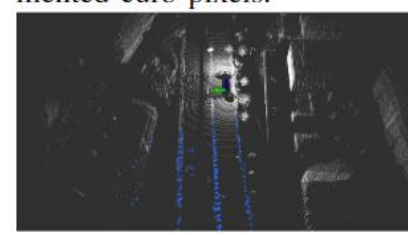
- RANSAC filtering
- Delaunay filtering

(a) Semantic segmentation results using our modified EfficientNet [18].

(b) Fused lidar point clouds from lidar sensors.



(c) Lidar point clouds (white points) overlaid on the segmented curb pixels.



(d) Curb semantics (blue points) with the fused point cloud.

Source: S. Das, N. Mahabadi, S. Chatterjee, and M. Fallon, “Multi-modal curb detection and filtering,” *CoRR*, vol. abs/2205.07096, 2022.

Multi-modal Curb detection - Method

Curb segmentation on RGB

- EfficientNet

Association with LiDAR

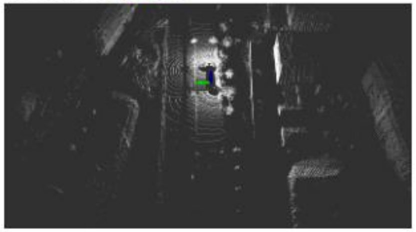
Unsupervised clustering

- DBSCAN (density-based)

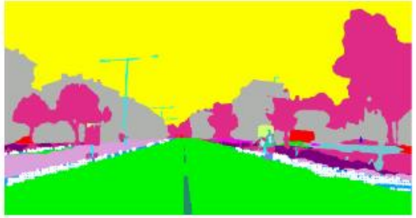
Filtering

- RANSAC filtering
- Delaunay filtering

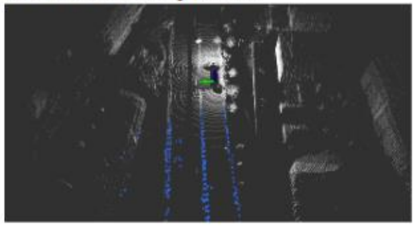
(a) Semantic segmentation results using our modified EfficientNet [18].



(b) Fused lidar point clouds from lidar sensors.



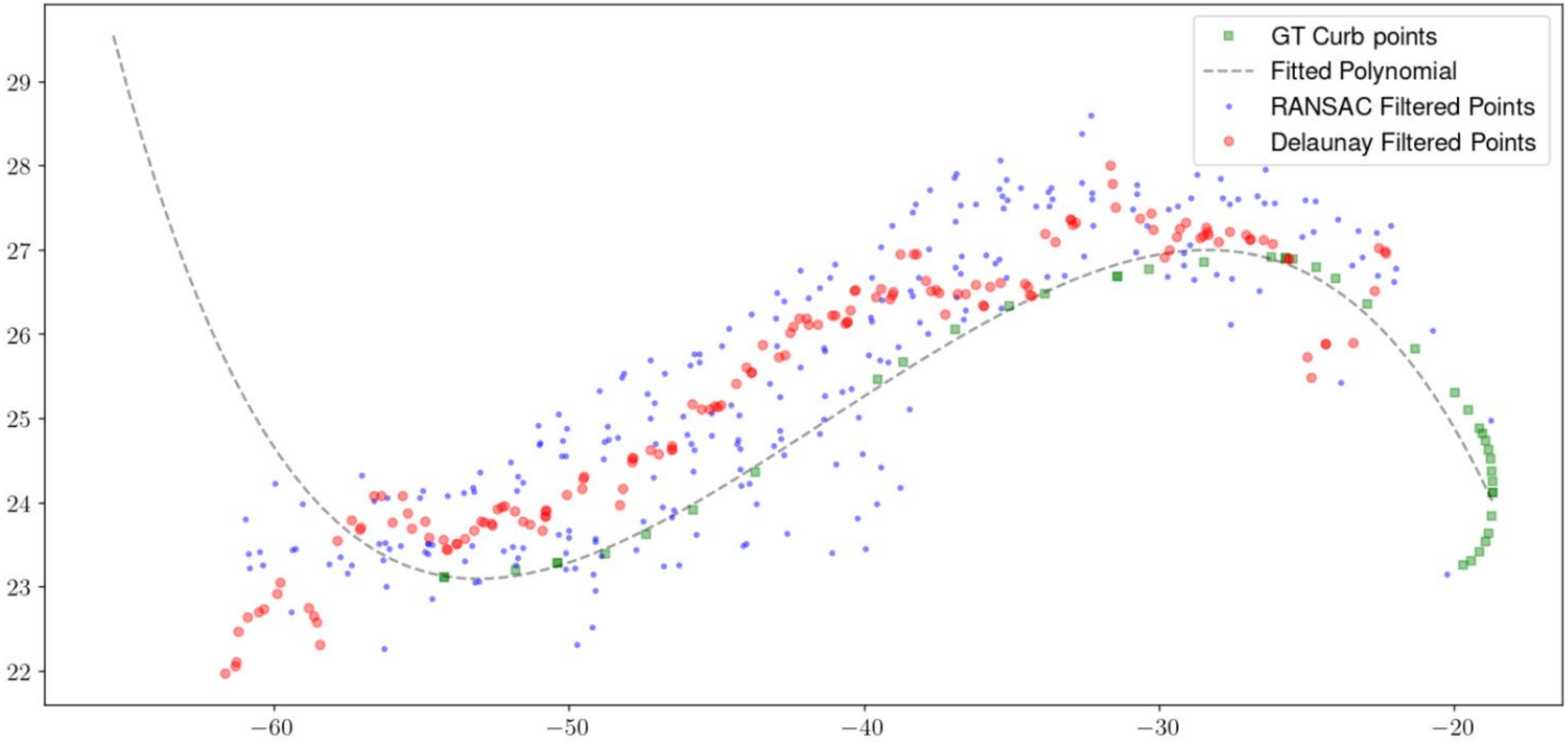
(c) Lidar point clouds (white points) overlaid on the segmented curb pixels.



(d) Curb semantics (blue points) with the fused point cloud.

Source: S. Das, N. Mahabadi, S. Chatterjee, and M. Fallon, “Multi-modal curb detection and filtering,” *CoRR*, vol. abs/2205.07096, 2022.

Multi-modal Curb detection - Results



Source: S. Das, N. Mahabadi, S. Chatterjee, and M. Fallon, "Multi-modal curb detection and filtering," *CoRR*, vol. abs/2205.07096, 2022.

Multi-modal Curb detection - Results

Manual segment-wise association		
No Clustering	**Normalized L_2-Norm**	**# Detected Points**
RANSAC Filtering	27.659	9578
Delaunay Filtering	**19.947**	6904
Automatic segment-wise association		
Outlier Removal (RANSAC)	**Chamfer Distance**	**# Detected Points**
Agglomerative Clustering	17.427	3489
BIRCH	19.596	1351
DBSCAN	**17.220**	5314
OPTICS	18.370	7446
Outlier Removal (Delaunay)	**Chamfer Distance**	**# Detected Points**
Agglomerative Clustering	15.418	3924
BIRCH	16.165	3492
DBSCAN	**14.753**	6678
OPTICS	15.870	4415

Source: S. Das, N. Mahabadi, S. Chatterjee, and M. Fallon, “Multi-modal curb detection and filtering,” *CoRR*, vol. abs/2205.07096, 2022.

CMX: Cross-Modal Fusion for RGB-X Semantic Segmentation with Transformers

Jiaming Zhang*, Huayao Liu*, Kailun Yang*†, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen

J. Zhang, R. Liu, and R. Stiefelhagen are with Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany.

K. Yang is with Hunan University, Changsha 410082, China.

H. Liu is with NIO, Shanghai 201804, China.

X. Hu is with ByteDance Inc., Hangzhou 310000, China.

*indicates equal contribution.

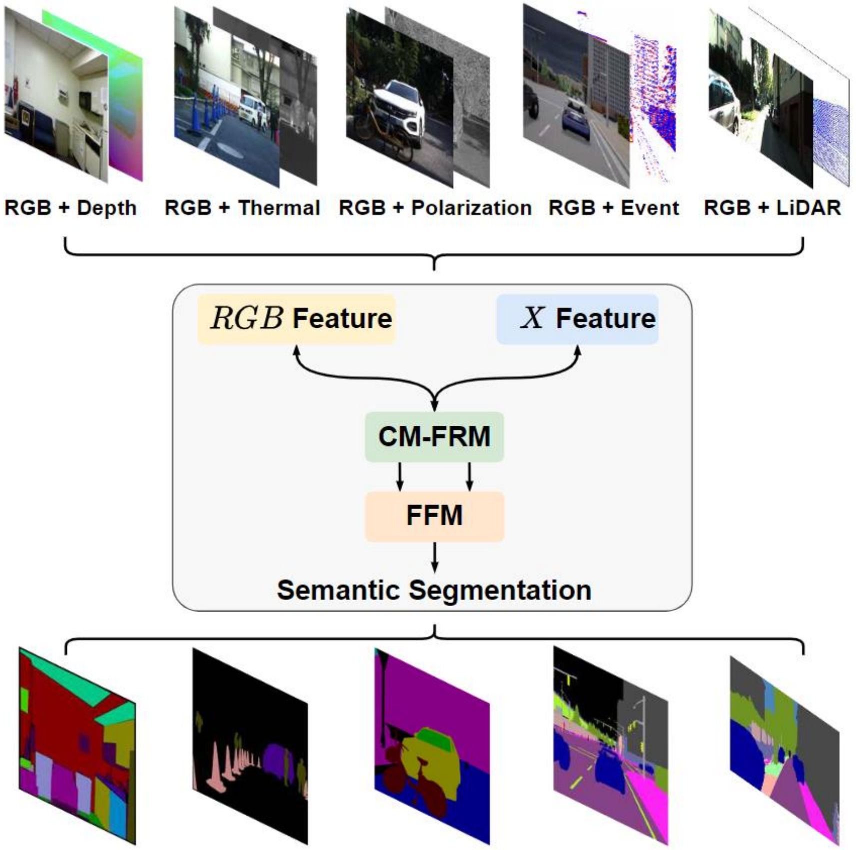
†corresponding author. (E-Mail: kailun.yang@hnu.edu.cn.)

CMX: Cross-Modal Fusion for RGB-X

Unified fusion framework

RGB-X semantic segmentation

Attention mechanisms enable efficient fusion

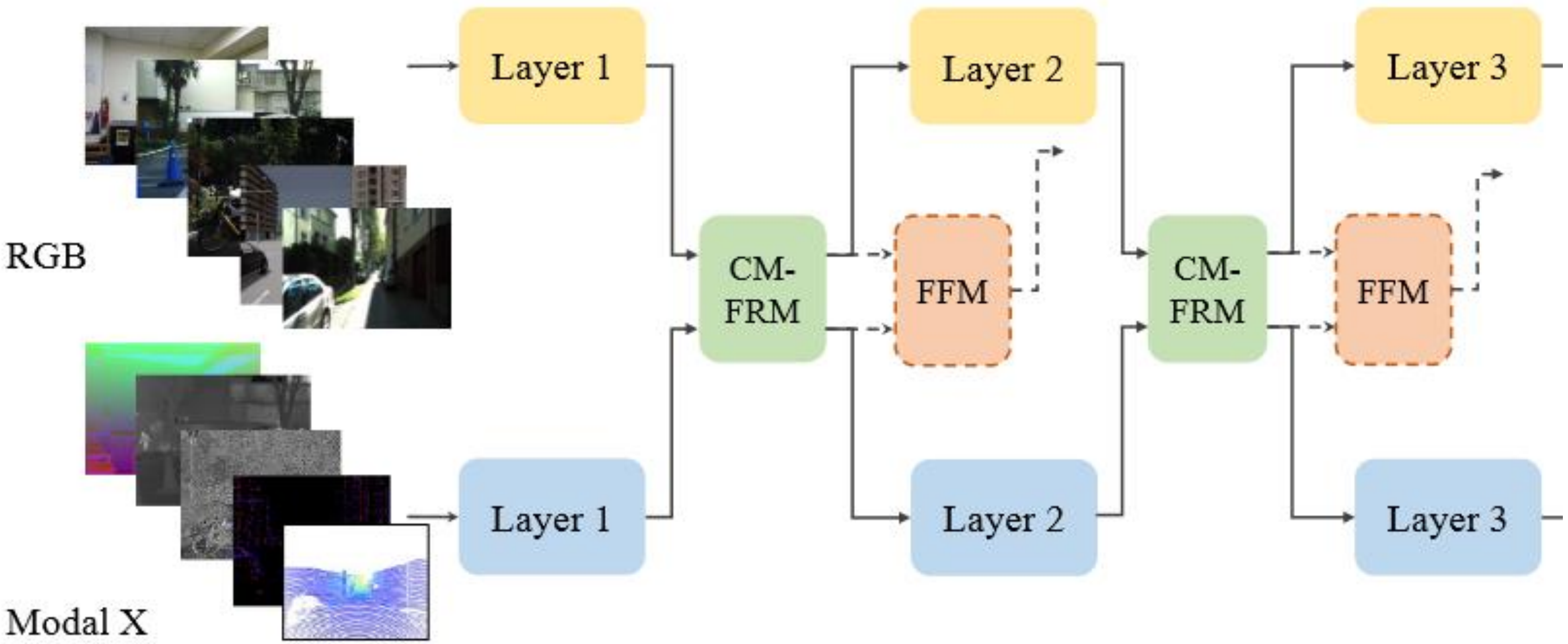


Source: J. Zhang, H. Liu, K. Yang, X. Hu, R. Liu, and Rainer Stiefelhagen, "CMX: Cross-Modal Fusion for RGB-X Semantic Segmentation With Transformers," IEEE Transactions on Intelligent Transportation Systems, vol. 24, no. 12, pp. 14679–14694, 2022.

CMX - Method

Overall Framework

Layer	Mix Transformer (MiT)
CM-FRM	Cross-modal feature rectification
FFM	Feature fusion

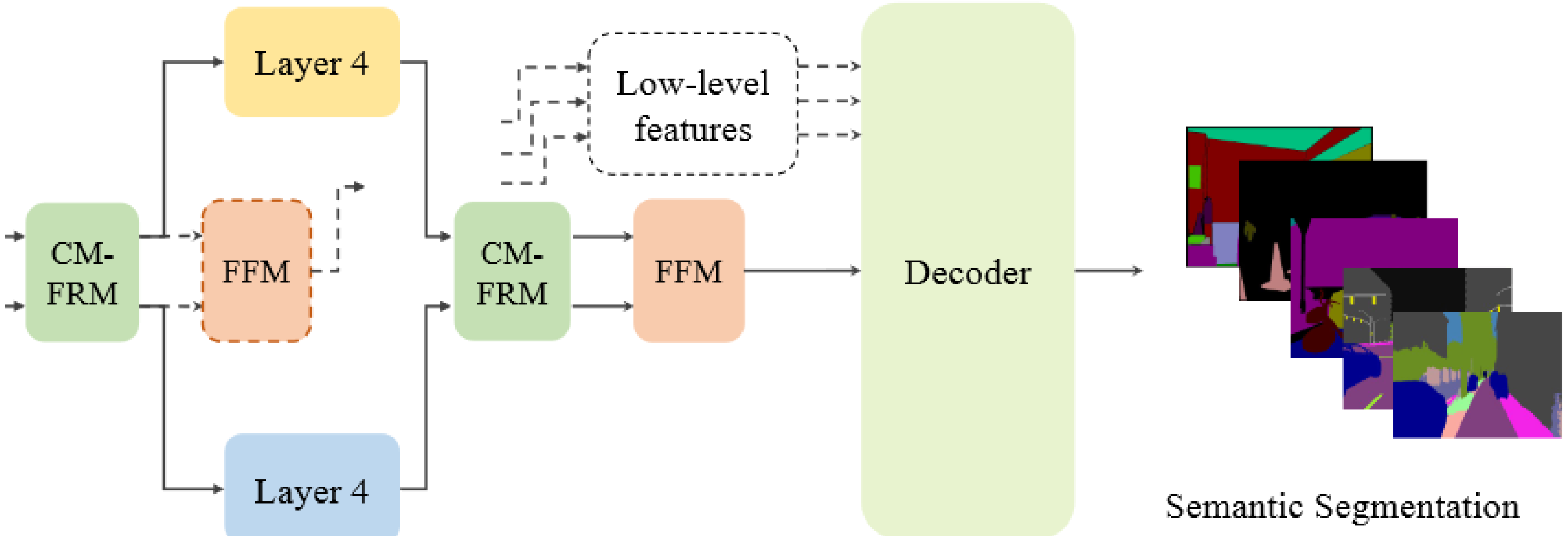


Source: E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, “SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers,” NeurIPS, 2021.
J. Zhang, H. Liu, K. Yang, X. Hu, R. Liu, and Rainer Stiefelhagen, “CMX: Cross-Modal Fusion for RGB-X Semantic Segmentation With Transformers,” IEEE Transactions on Intelligent Transportation Systems, 2022.

CMX - Method

Overall Framework

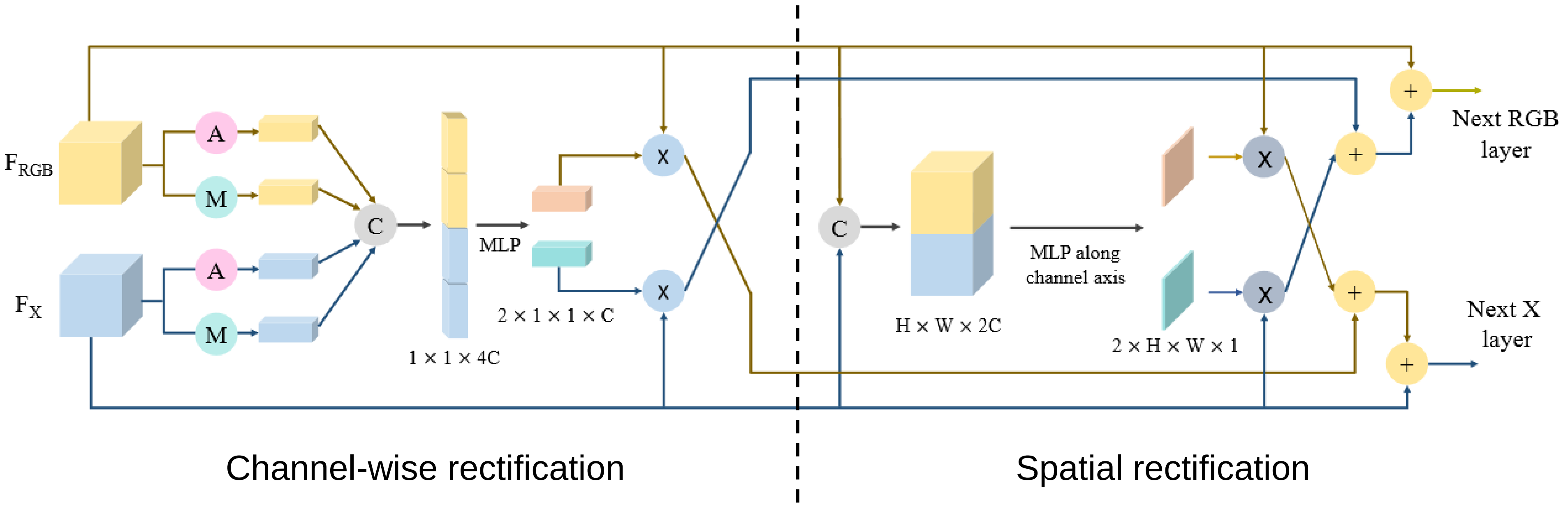
Layer	Mix Transformer (MiT)
CM-FRM	Cross-modal feature rectification
FFM	Feature fusion



Source: E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, “SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers,” NeurIPS, 2021.
J. Zhang, H. Liu, K. Yang, X. Hu, R. Liu, and Rainer Stiefelhagen, “CMX: Cross-Modal Fusion for RGB-X Semantic Segmentation With Transformers,” IEEE Transactions on Intelligent Transportation Systems, 2022.

CMX - Method

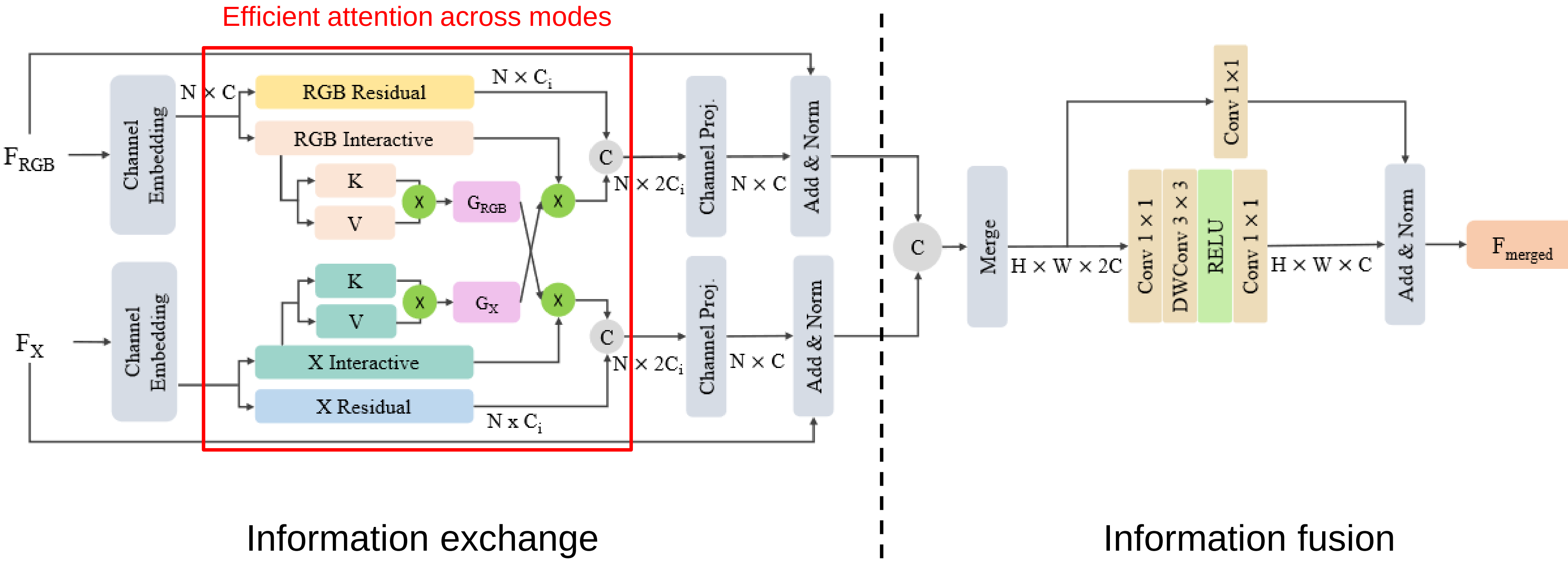
Cross-modal feature rectification module (CM-FRM)



Source: J. Zhang, H. Liu, K. Yang, X. Hu, R. Liu, and Rainer Stiefelhagen, "CMX: Cross-Modal Fusion for RGB-X Semantic Segmentation With Transformers," IEEE Transactions on Intelligent Transportation Systems, vol. 24, no. 12, pp. 14679–14694, 2022.

CMX - Method

Feature fusion module (FFM)



Source: J. Zhang, H. Liu, K. Yang, X. Hu, R. Liu, and Rainer Stiefelhagen, “CMX: Cross-Modal Fusion for RGB-X Semantic Segmentation With Transformers,” IEEE Transactions on Intelligent Transportation Systems, vol. 24, no. 12, pp. 14679–14694, 2022.

Delivering Arbitrary-Modal Semantic Segmentation

Jiaming Zhang[1,*], Ruiping Liu[1,*], Hao Shi[3], Kailun Yang[2,†], Simon Reiß[1], Kunyu Peng[1], Haodong Fu[4], Kaiwei Wang[3], Rainer Stiefelhagen[1]
[1]Karlsruhe Institute of Technology, [2]Hunan University, [3]Zhejiang University, [4]Beihang University

*Equal contribution.
†Corresponding author (e-mail: kailun.yang@hnu.edu.cn).

[1]The DELIVER dataset and our code will be made publicly available at: https://jamycheung.github.io/DELIVER.html.

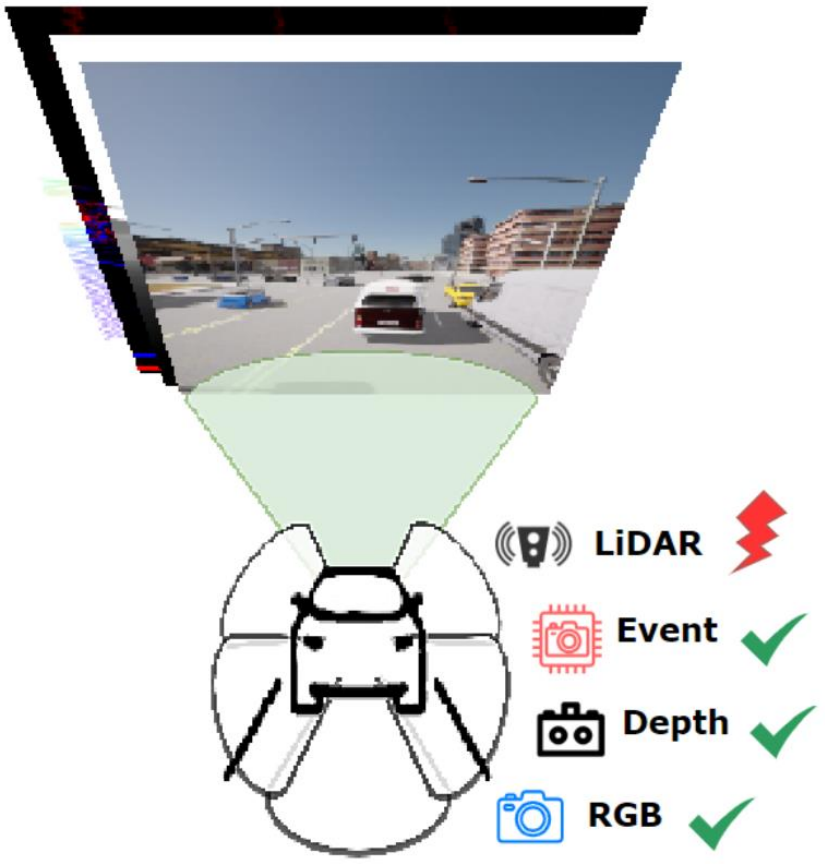
CMNeXt: Arbitrary-Modal Fusion

Extending CMX

- Multiple additional modalities
- Retains two-stream architecture

Synthetic dataset DeLiVER

- Depth
- LiDAR
- Multiple Views
- Event



Source: J. Zhang et al., “Delivering Arbitrary-Modal Semantic Segmentation,” CVPR, 2023.

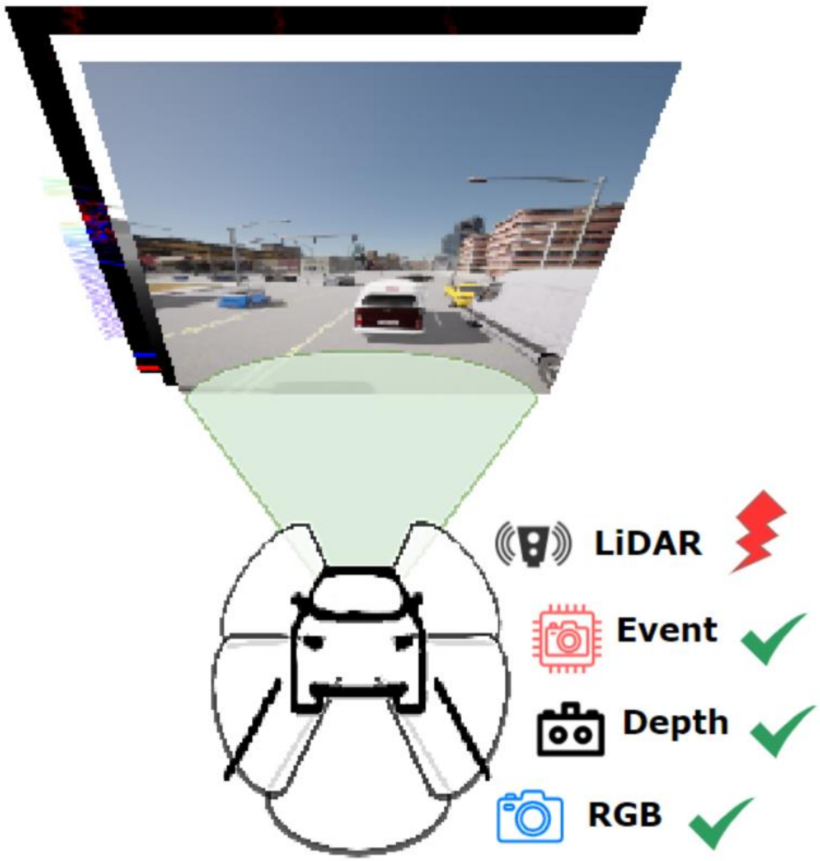
CMNeXt: Arbitrary-Modal Fusion

Extending CMX

- Multiple additional modalities
- Retains two-stream architecture

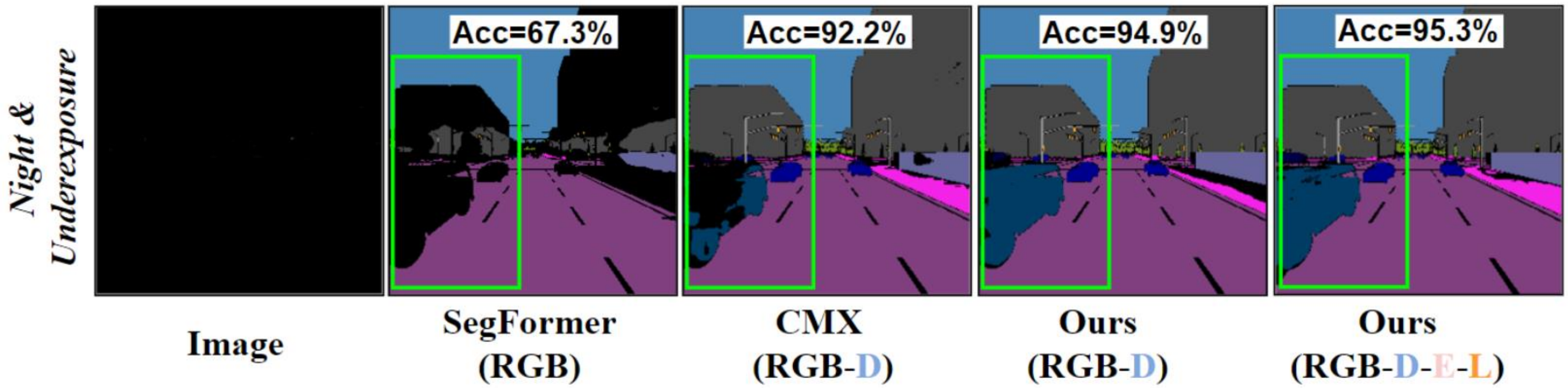
Synthetic dataset DeLiVER

- Depth
- LiDAR
- Multiple Views
- Event



Source: J. Zhang et al., "Delivering Arbitrary-Modal Semantic Segmentation," CVPR, 2023.

CMX and CMNeXt - Results



Source: J. Zhang et al., "Delivering Arbitrary-Modal Semantic Segmentation," CVPR, 2023.

CMX and CMNeXt - Results

(b) Results on MFNet.

Method	Modal	mIoU
SwinT [50]	RGB	49.0
SegFormer [80]	RGB	52.0
ACNet [35]	RGB-T	46.3
FuseSeg [66]	RGB-T	54.5
ABMDRNet [96]	RGB-T	54.8
LASNet [41]	RGB-T	54.9
FEANet [15]	RGB-T	55.3
MFTNet [101]	RGB-T	57.3
GMNet [103]	RGB-T	57.3
DooDLeNet [20]	RGB-T	57.3
CMX (MiT-B2) [49]	RGB-T	58.2
CMX (MiT-B4) [49]	RGB-T	59.7
CMNeXt (MiT-B4)	RGB-T	**59.9**

(c) Results on NYU Depth V2.

Method	mIoU
ACNet [35]	48.3
SGNet [9]	51.1
ShapeConv [5]	51.3
NANet [92]	52.3
SA-Gate [11]	52.4
PGDENet [104]	53.7
TokenFusion [72]	54.2
TransD-Fusion [78]	55.5
MultiMAE [2]	56.0
Omnivore [25]	56.8
CMX (MiT-B4) [49]	56.3
CMX (MiT-B5) [49]	**56.9**
CMNeXt (MiT-B4)	**56.9**

Source: J. Zhang et al., “Delivering Arbitrary-Modal Semantic Segmentation,” CVPR, 2023.

Multimodal Knowledge Expansion

Zihui Xue[1,2], Sucheng Ren[1,3], Zhengqi Gao[1,4], and Hang Zhao [*5,1]

[1]Shanghai Qi Zhi Institute, [2]UT Austin
[3]South China University of Technology
[4]MIT, [5]Tsinghua University

Multimodal Knowledge Expansion - MKE

Models need to be trained on data!

RGB

- Big field of research
- Many datasets
- Well-trained backbones

Multi-modal

- Some labeled datasets,
  lots of unlabeled data
- Not many pre-trained backbones

Source: Z. Xue, S. Ren, Z. Gao, and H. Zhao, “Multimodal Knowledge Expansion,” ICCV, 2021.

Multimodal Knowledge Expansion - MKE

Models need to be trained on data!

RGB

- Big field of research
- Many datasets
- Well-trained backbones

Multi-modal

- Some labeled datasets,
  lots of unlabeled data
- Not many pre-trained backbones

Transfer knowledge to different modes?

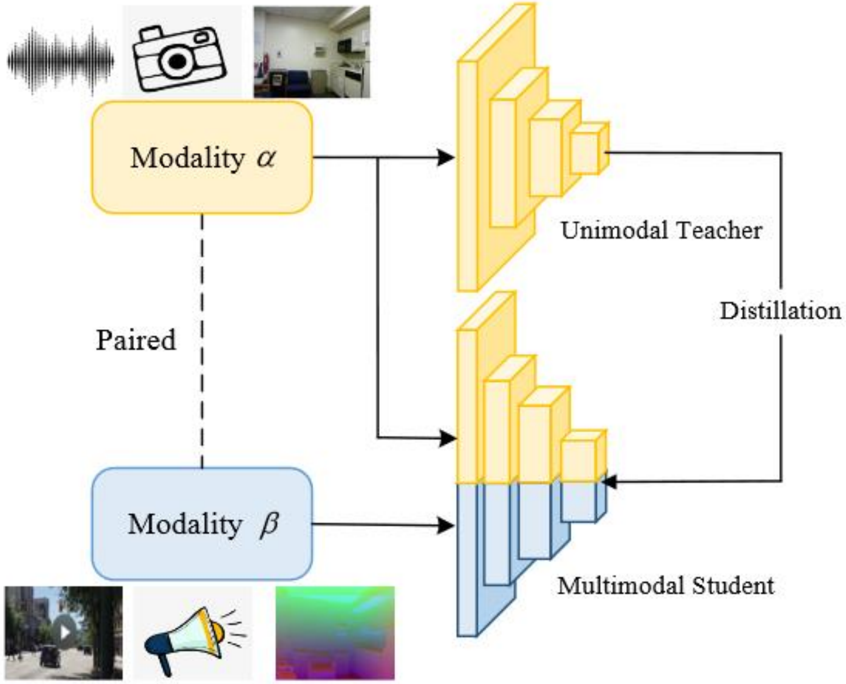
Source: Z. Xue, S. Ren, Z. Gao, and H. Zhao, “Multimodal Knowledge Expansion,” ICCV, 2021.

MKE - Method

Based on knowledge distillation

Teacher-Student architecture

- Teacher
  - unimodal
  - generates pseudo-labels
- Student
  - multi-modal
  - learns on pseudo-labels



Source: Z. Xue, S. Ren, Z. Gao, and H. Zhao, "Multimodal Knowledge Expansion," ICCV, 2021.

MKE - Method

Confirmation Bias

Student should not strictly confirm to Teacher's pseudo-labels!

Solution → Loss term

(like consistency regularization in SSL)

$$\theta_s^\star = \underset{\theta_s}{\operatorname{argmin}}(\mathcal{L}_{pl} + \gamma \mathcal{L}_{reg})$$

$$\mathcal{L}_{pl} = \frac{1}{M} \sum_{i=1}^{M} l_{cls}(\tilde{\mathbf{y}}_i, \mathbf{f}_s(\mathbf{x}_i^\alpha, \mathbf{x}_i^\beta; \theta_s))$$

$$\mathcal{L}_{reg} = \sum_{i=1}^{M} l_{reg}[\mathbf{f}_s(\mathbf{x}_i^\alpha, \mathbf{x}_i^\beta; \theta_s), \mathcal{T}(\mathbf{f}_s(\mathbf{x}_i^\alpha, \mathbf{x}_i^\beta; \theta_s))]$$

$l_{cls}(\cdot)$	Cross-entropy loss
$l_{reg}(\cdot)$	Distance metric (L2)
$f_s(\cdot)$	Student model
$T(\cdot)$	Transformation on student model (i.e. input or model perturbation)

Source: Z. Xue, S. Ren, Z. Gao, and H. Zhao, “Multimodal Knowledge Expansion,” ICCV, 2021.

MKE - Results

Method	Train data			Test mIoU
	mod	D_l	$\tilde{D}_u$	(%)
UM teacher	rgb	✓		44.15
Naive student [10]	rgb		✓	46.13
NOISY student [44]	rgb	✓	✓	47.68
Gupta *et al.* [15]	rgb, d		✓	45.65
CMKD [49]	rgb, d		✓	45.25
MM student (no reg)	rgb, d		✓	46.14
MM student (ours)	rgb, d		✓	**48.88**

Table 4: Results of semantic segmentation on NYU Depth V2. rgb and d denote RGB images and depth images.

Source: Z. Xue, S. Ren, Z. Gao, and H. Zhao, “Multimodal Knowledge Expansion,” ICCV, 2021.

MKE - Results

Methods	Train data			Accuracy (%)	
	mod	D_l	$\tilde{D}_u$	val	test
UM teacher	i	✓		79.67	80.33
UM student	i		✓	79.01	77.79
NOISY student [44]	i	✓	✓	82.54	83.09
MM student (no reg)	i, a		✓	88.73	89.28
MM student (ours)	i, a		✓	90.61	**91.38**
MM student (sup)	i, a		$\star$	97.46	97.35

Table 3: Results of emotion recognition on RAVDESS. *mod*, i and a denote modality, images and audios, respectively. Data used for training each method is listed. $\star$ means that the MM student (sup) is trained on true labels instead of pseudo labels in $\tilde{D}_u$.

Source: Z. Xue, S. Ren, Z. Gao, and H. Zhao, “Multimodal Knowledge Expansion,” ICCV, 2021.

Personal Comments

Multi-modal curb detection

- Unimodal Segmentation
- Simple unsupervised fusion in pipeline
- No interaction / end-to-end learning!

CMX and **CMNeXt**

- Unified fusion framework
- *(close to)* SOTA, even comparing to specialized models
- only image-like formats - no sparse data (LiDAR!)

Personal Comments

Multi-modal knowledge expansion

- Exciting (and surprising) results
- however very theoretical
- still a young field of research

Future Work

Leverage RGB knowledge in SOTA

- Starting point: Multi-modal Knowledge Expansion
- Train SOTA multi-modal model
- Student-teacher architecture
- Improvements?

Future Work

Support of different representations for modalities

- for fully unified multi-modal framework
- Image-like data, point cloud, non-structural (Audio, Language, …)
- no conversion / loss of structure necessary

even further:

Shared representation between modalities

- Recent advance → ImageBind

Source: R. Girdhar et al., “ImageBind: One Embedding Space To Bind Them All,” CVPR, 2023.

Thank you for listening!

Any Questions?