RGB+X Scene Parsing: Five Years of Explorations by the MIAS Group

Rui Fan | Ph.D.

12/08/2024, ACCV/ACML-MIRA

同济大学
TONGJI UNIVERSITY

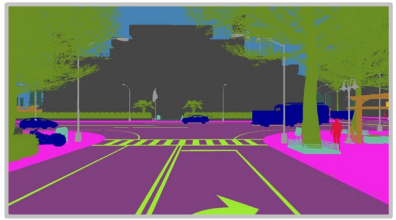
上海自主智能无人系统科学中心
Shanghai Research Institute for Intelligent Autonomous Systems

MIAS
MACHINE INTELLIGENCE & AUTONOMOUS SYSTEMS

College of Electronic & Information Engineering, Tongji University | Email: rui.fan@ieee.org

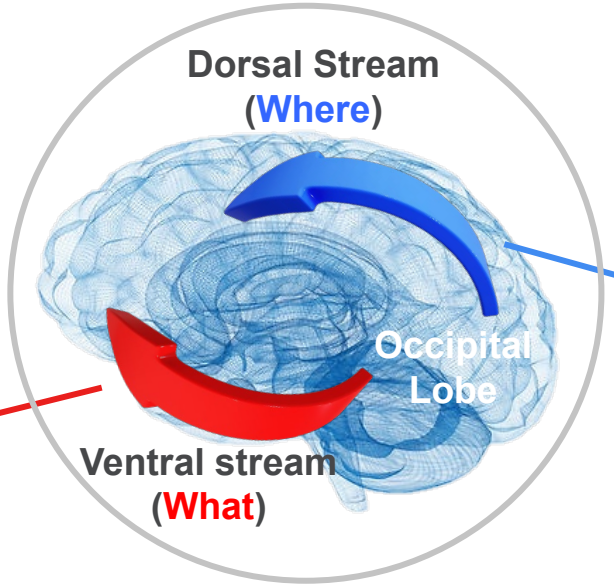
• Research Interests

Input: RGB Image

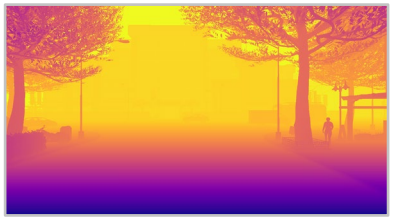


Output: Semantics

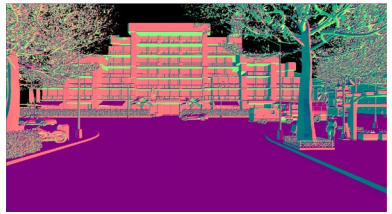
Scene Parsing



Human-like Visual Environment Perception

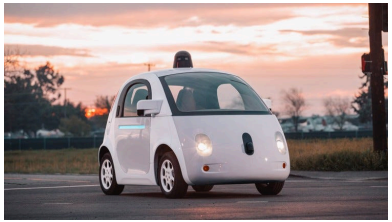


Output: Depth



Output: Surface Normal

Geometry Information Acquisition



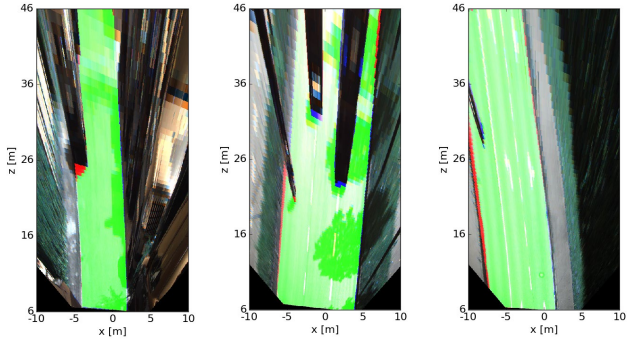
Autonomous Cars

Mobile Robots

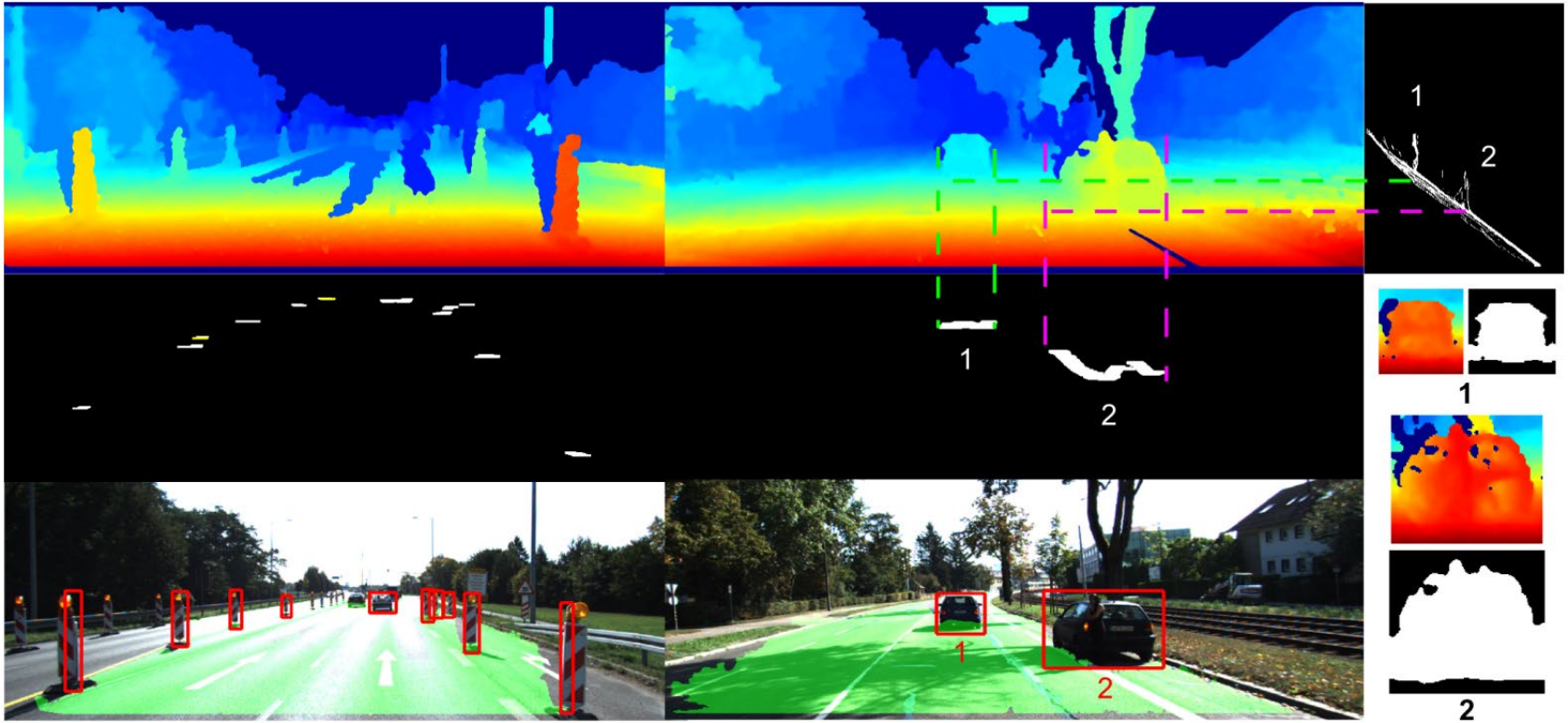
Two-streams Hypothesis (1992)

• Motivation

- Freespace detection is an essential component of autonomous car perception.
- Freespace detection can be formatted as a binary semantic driving scene segmentation problem.
- Freespace detection approaches generally classify each pixel in an RGB or depth/disparity image as drivable or undrivable.
- Such pixel-level classification results are then utilized by other modules in the autonomous system, such as trajectory prediction and path planning, to ensure that the autonomous car can navigate safely in complex environments.

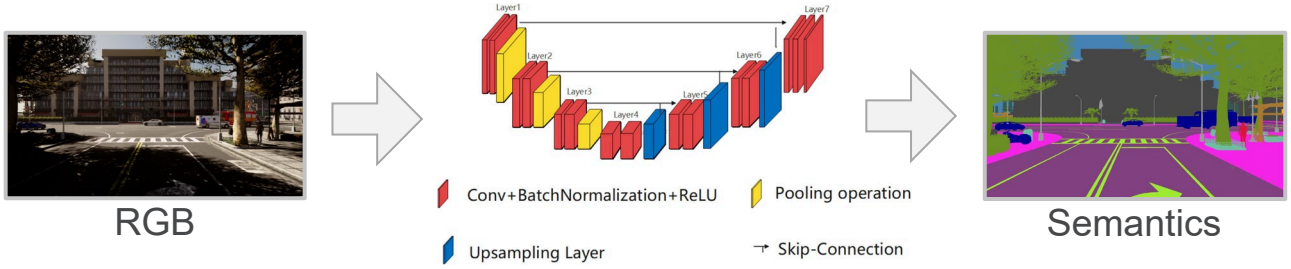


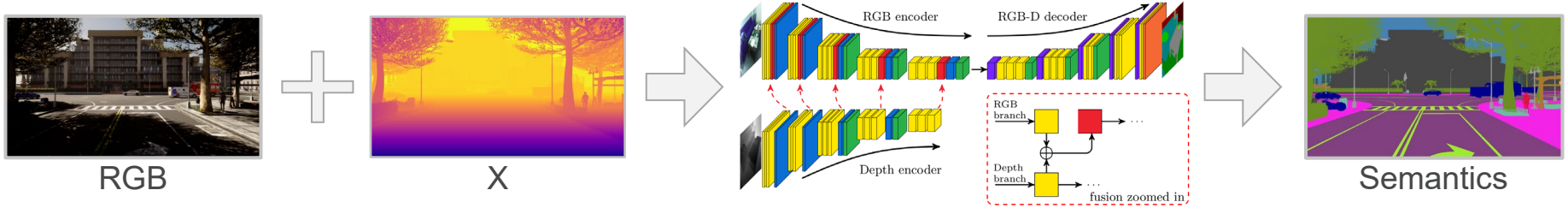
- # Traditional Methods for Freespace Detection



➢ The V-disparity histogram has been commonly utilized to address this problem.

- # RGB-X Scene Parsing

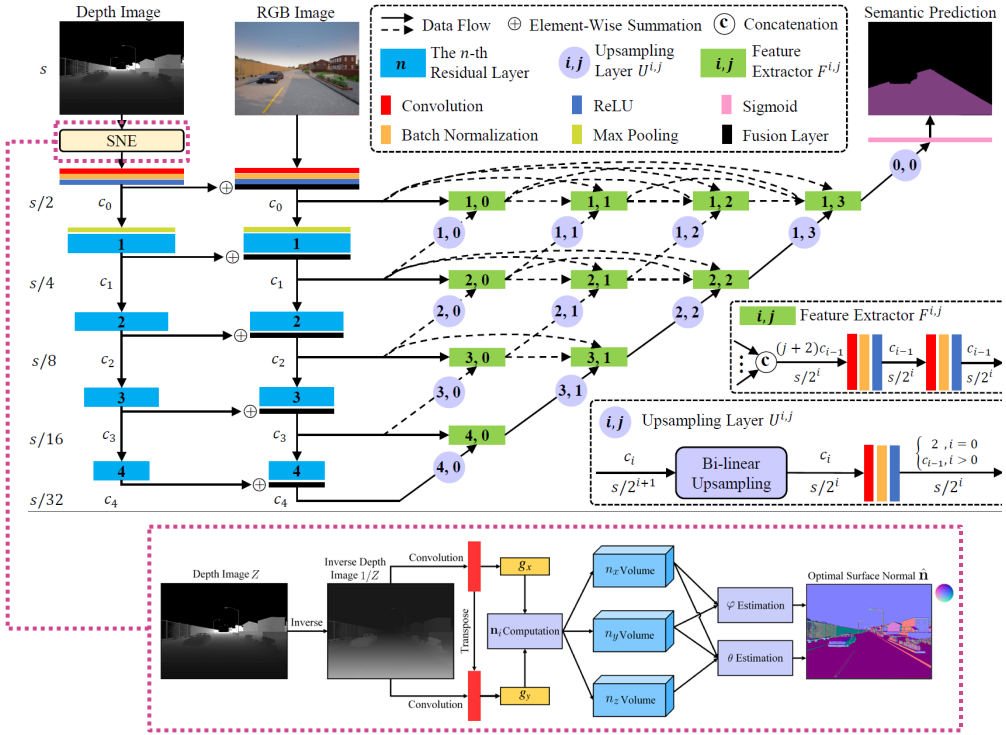




- We extract heterogeneous features from both RGB images and another source or modality of visual information (denoted as "X"), and fuse these features to provide a more comprehensive understanding of the environment.

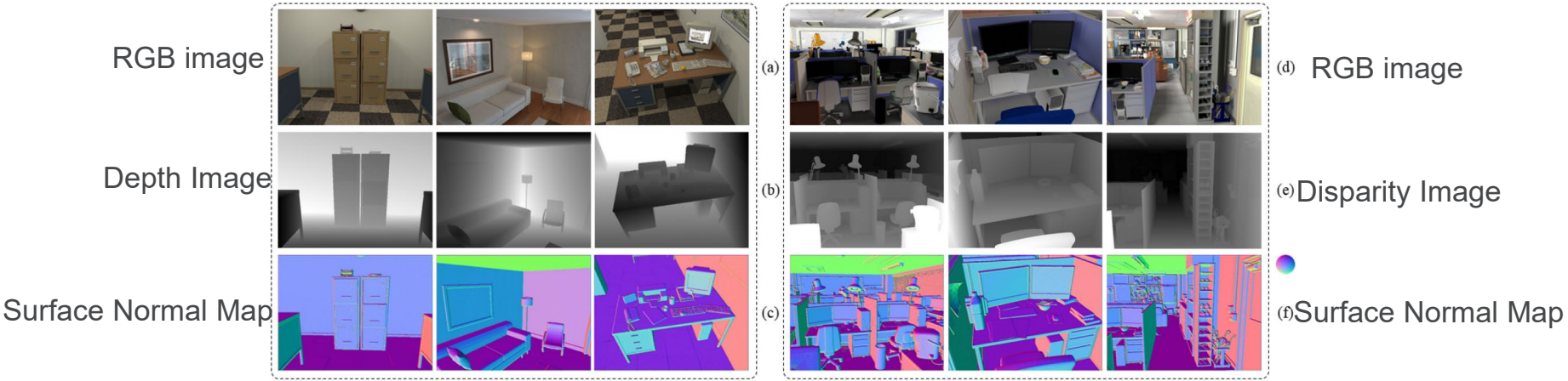
• Data-Fusion Freespace Detection

- We proposed a novel freespace detection approach, referred to as **SNE-RoadSeg**.
- It consist of 1) **a novel module, named surface normal estimator (SNE)**, which can infer surface normal information from dense depth/disparity images with high accuracy and efficiency; and 2) a **data-fusion CNN architecture, referred to as RoadSeg**, which can extract and fuse features from both RGB images and the inferred surface normal information for accurate freespace detection.



- **Fan, R.**, Wang, H., Cai, P. and Liu, M. SNE-RoadSeg: Incorporating surface normal information into semantic segmentation for accurate freespace detection. **ECCV 2020**.

• Three-Filters-to-Normal



- We proposed **Three-Filters-to-Normal (3F2N)**, a geometry-based surface normal estimator (SNE), which is designed for structured range sensor data, such as depth/disparity images.
- 3F2N SNE computes surface normals by simply performing **three filtering operations (two image gradient filters in horizontal and vertical directions, respectively, and a mean/median filter)** on an inverse depth image or a disparity image.

- **Fan, R.**, Wang, H., Xue, B., Huang, H., Wang, Y., Liu, M. and Pitas, I., 2021. Three-filters-to-normal: An accurate and ultrafast surface normal estimator. **RAL + ICRA 2021**

• Three-Filters-to-Normal

Accuracy

Method	t (ms) ↓	e_A (degrees) ↓			π (degrees/kHz) ↓		
		Easy	Medium	Hard	Easy	Medium	Hard
PlaneSVD [14]	393.69	2.07	6.07	17.59	813.87	2389.73	6923.18
PlanePCA [13]	631.88	2.07	6.07	17.59	1306.29	3835.59	11111.92
VectorSVD [4]	563.21	2.13	6.27	18.01	1199.63	3529.11	10142.34
AreaWeighted [4]	1092.24	2.20	6.27	17.03	2407.74	6843.56	18600.68
AngleWeighted [4]	1032.88	1.79	**5.67**	**13.26**	1850.00	5855.62	13693.24
FALS [5]	4.11	2.26	6.14	17.34	9.26	25.20	71.17
SRI [5]	12.18	2.64	6.71	19.61	32.18	81.66	238.78
LINE-MOD [3]	6.43	6.53	9.94	31.45	41.93	63.84	202.08
SNE-RoadSeg [12]	7.92	2.04	6.28	16.37	16.16	49.74	129.65
FD-Mean (ours)	**3.72**	2.14	6.66	15.30	**7.96**	**24.80**	**56.96**
FD-Median (ours)	10.97	**1.66**	5.69	15.31	18.18	62.38	168.03

Comparisons among the geometry-based surface normal estimators w.r.t. the average angular degree .

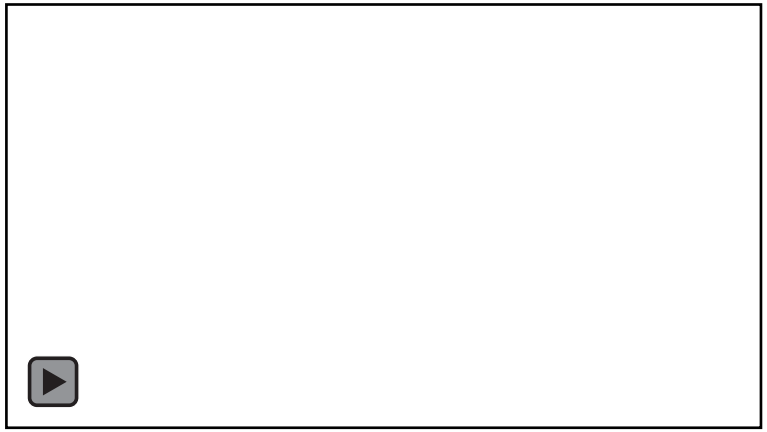
$$e_A = \frac{1}{m}\sum_{k=1}^{m}\psi_k$$

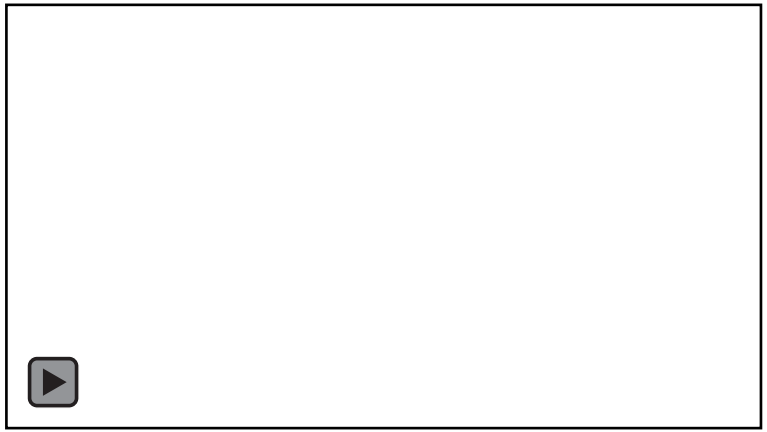
Speed

Method	Jetson TX2	GTX 1080 Ti	RTX 2080 Ti
FD-Mean	**0.823521**	**0.049504**	**0.046944**
Sobel-Mean	0.855843	0.052288	0.051232
Scharr-Mean	0.860319	0.052320	0.051280
Prewitt-Mean	0.857762	0.052256	0.050816
FD-Median	**1.206337**	**0.102368**	**0.065536**
Sobel-Median	1.217023	0.104608	0.067840
Scharr-Median	1.239041	0.105376	0.071008
Prewitt-Median	1.240479	0.105152	0.069024

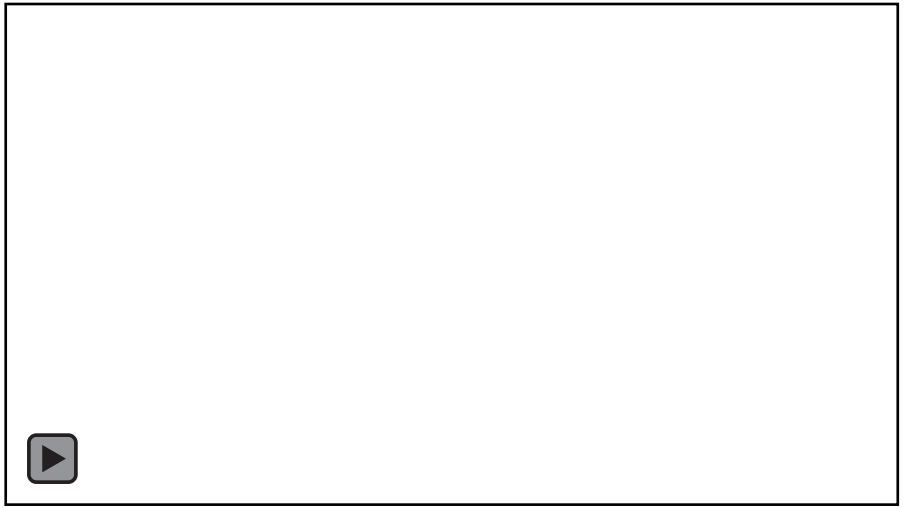
Runtime (ms) on different platforms.

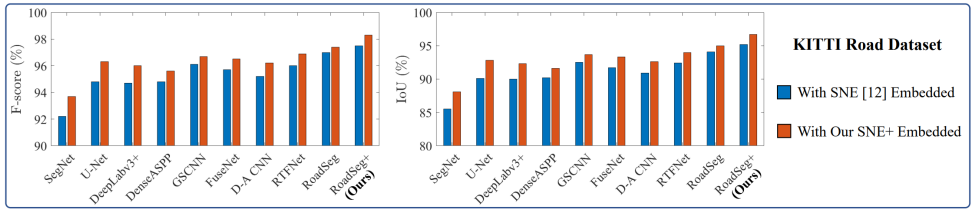
- Our **C++ and CUDA implementations** achieve a processing speed of over **260 Hz and 21 kHz**, respectively.





• Data-Fusion Freespace Detection

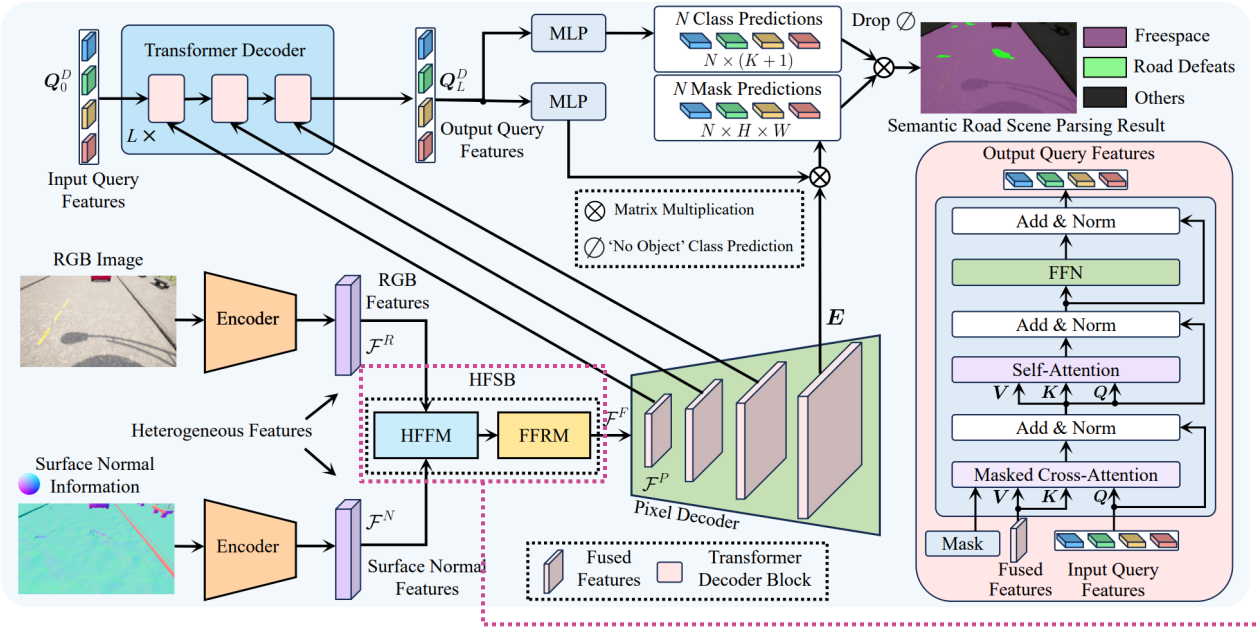


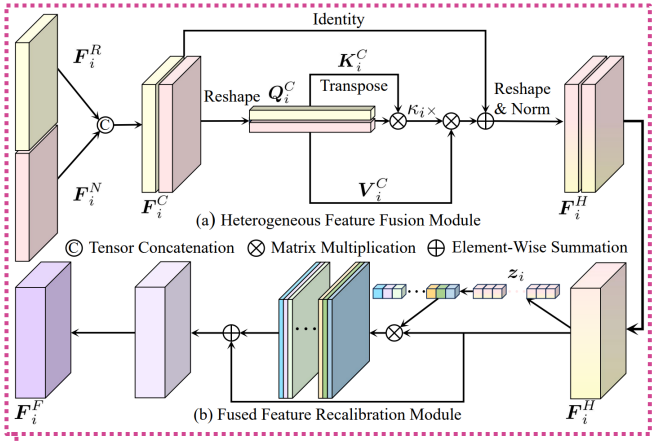


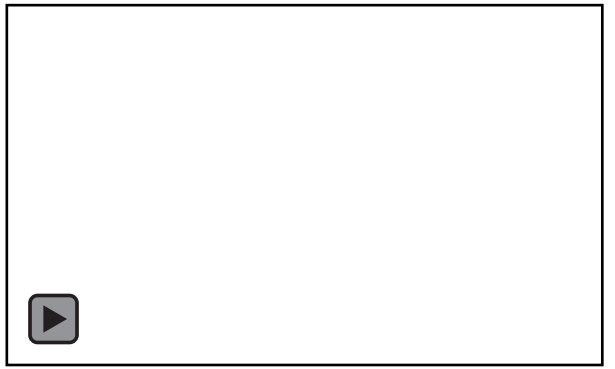
Method	MaxF (%)	AP (%)	PRE (%)	REC (%)	Rank
RBNet [10]	93.21	89.18	92.81	93.60	21
TVFNet [17]	95.34	90.26	95.73	94.94	16
LC-CRF [16]	95.68	88.34	93.62	**97.83**	13
LidCamNet [5]	96.03	93.93	96.23	95.83	7
RBANet [28]	96.30	89.72	95.14	97.50	6
SNE-RoadSeg (Ours)	**96.75**	**94.07**	**96.90**	96.61	**2**

Approach	MaxF (%)	AP (%)	Runtime (s)
RBNet [34]	94.97	91.49	0.18
LC-CRF [35]	95.68	88.34	0.18
LidCamNet [36]	96.03	93.93	0.15
SNE-RoadSeg [12]	96.75	**94.07**	0.10
PLARD [8]	97.03	94.03	0.16
SNE-RoadSeg+ (**Ours**)	**97.50**	93.98	**0.08**

• Freespace Detection







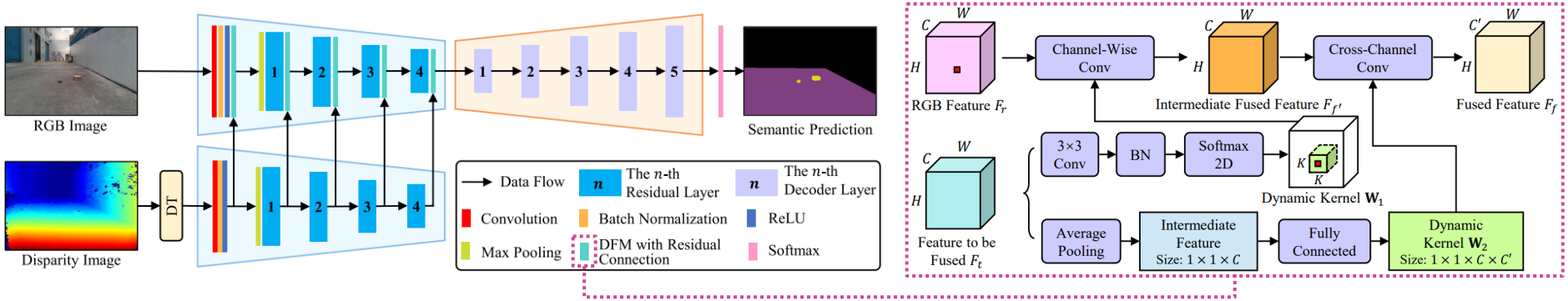
- Li, J., Zhang, Y., Yun, P., Zhou, G., Chen, Q., & **Fan, R.** (2023). **RoadFormer**: Duplex Transformer for RGB-normal semantic road scene parsing. arXiv preprint arXiv:2309.10356. **IEEE T-IV**

• Road Anomaly Detection

- Ground mobile robots, such as robotic wheelchairs and sweeping robots, can significantly improve people's comfort and life quality.
- Among all visual environmental perception tasks for mobile robots, the joint detection of drivable areas and road anomalies at the pixel level is a crucial one. Accurate and efficient drivable area and road anomaly detection can help avoid accidents for such vehicles.
- Some benchmark datasets, such as KITTI and Cityscapes, have been widely used. However, **the existing benchmarks are mostly designed for self-driving cars**. There lacks a benchmark for ground mobile robots, such as robotic wheelchairs.

v.s.

• Road Anomaly Detection



- We propose a novel module, referred to as the **dynamic fusion module (DFM)**, which can be easily deployed in existing data-fusion networks to fuse different types of visual features effectively and efficiently.
- We explored the effectiveness of data-fusion network with **different inputs of vision sensor data**, including RGB images; 2) disparity images; 3) normal images; 4) HHA images; 5) elevation maps; and 6) transformed disparity images.

[6] Wang, H.#, **Fan, R.**#, Sun, Y. and Liu, M. Applying surface normal information in drivable area and road anomaly detection for ground mobile robots. ***IROS 2020***.

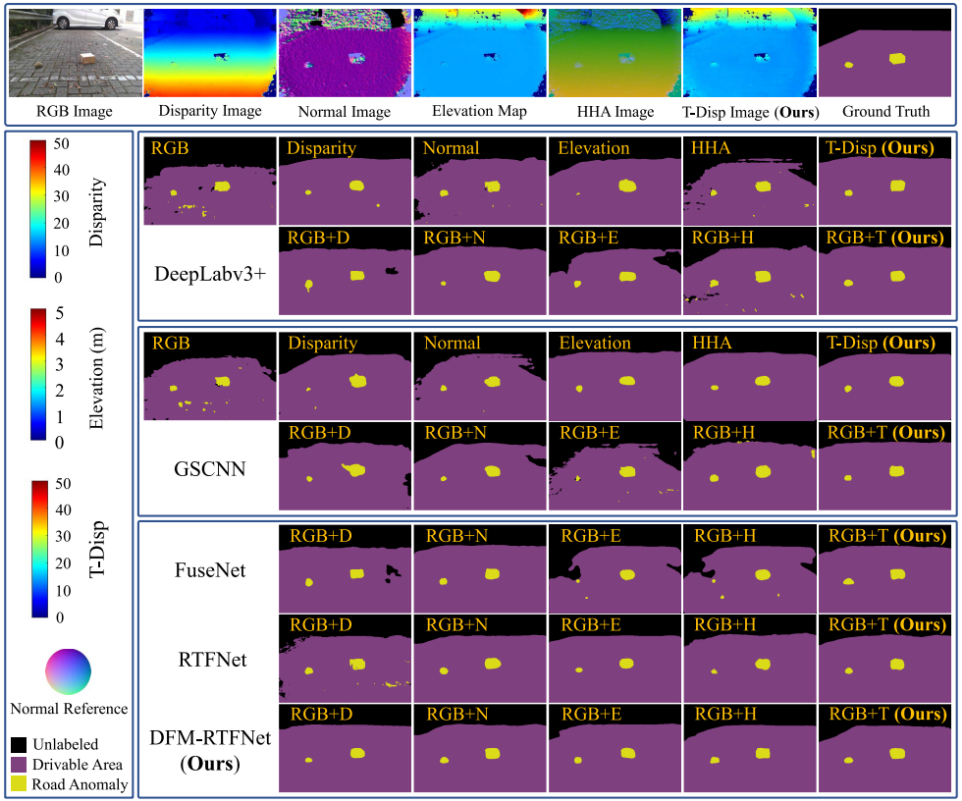
[7] Wang, H.#, **Fan, R.**#, Sun, Y. and Liu, M. Dynamic fusion module evolves drivable area and road anomaly detection: A benchmark and algorithms. ***IEEE T-CYB***.

• Road Anomaly Detection

- Transformed disparity images are the most informative type of vision sensor data in terms of road anomaly detection.
- Data-fusion networks are more robust than single-modal networks.

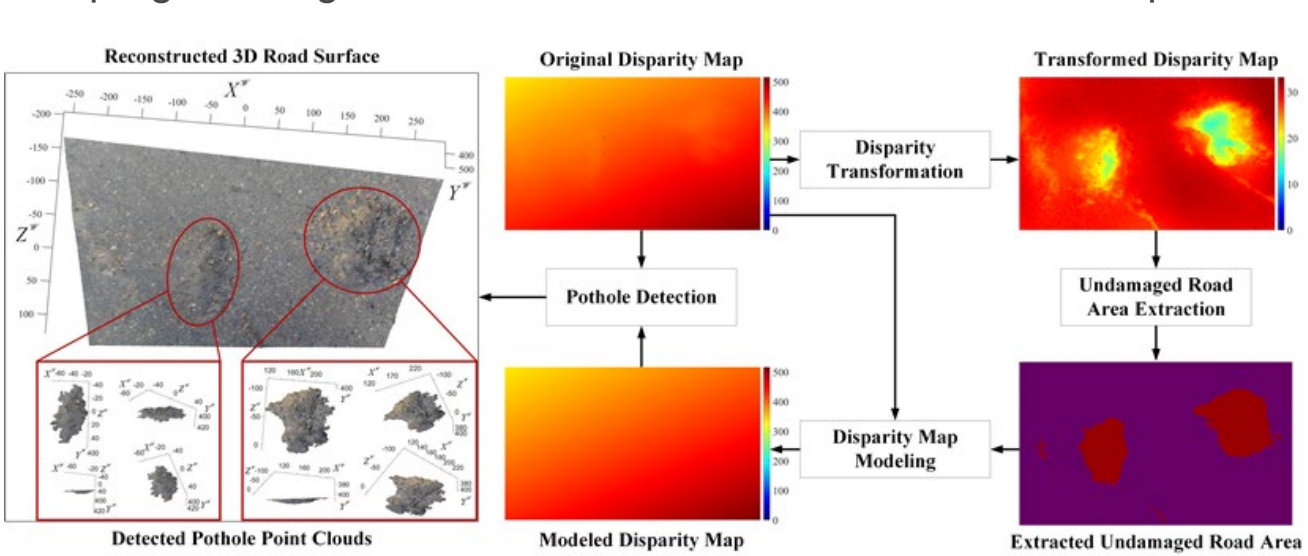
Approach	Setup	AP_D	AP_R	mAP
DeepLabv3+ [10]	T-Disp (**Ours**)	99.71	92.45	96.08
ESPNet [27]	T-Disp (**Ours**)	99.68	91.79	95.74
GSCNN [26]	T-Disp (**Ours**)	99.36	93.61	96.49
FuseNet [28]	RGB+T (**Ours**)	99.25	93.39	96.32
RTFNet [29]	RGB+T (**Ours**)	99.70	96.27	97.99
DFM-RTFNet (**Ours**)	RGB+D	99.72	92.17	95.95
	RGB+N	99.67	97.12	98.40
	RGB+E	99.69	94.83	97.26
	RGB+H	99.61	96.13	97.87
	RGB+T (**Ours**)	**99.85**	**97.61**	**98.73**

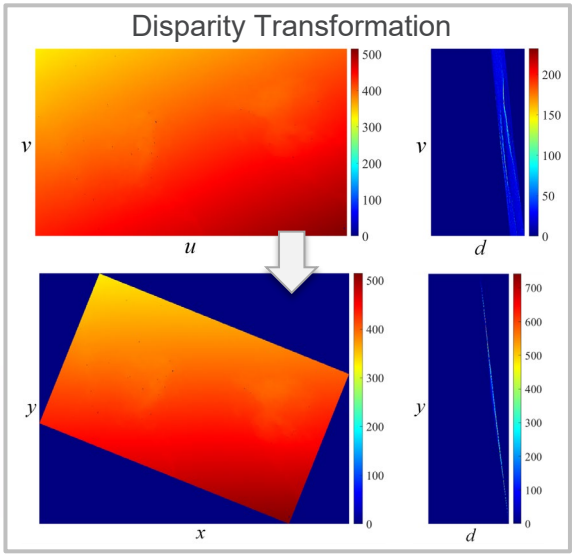
Experimental results (%) of four SOTA data-fusion networks & our DFM-RTFNet w.r.t. different training data setups on the KITTI semantic segmentation dataset. Best results shown in bold type.



• Road Anomaly Detection

- A dense disparity map is first transformed to better distinguish between damaged and undamaged road areas.
- To achieve greater disparity transformation efficiency, golden section search and dynamic programming are utilized to estimate the transformation parameters.



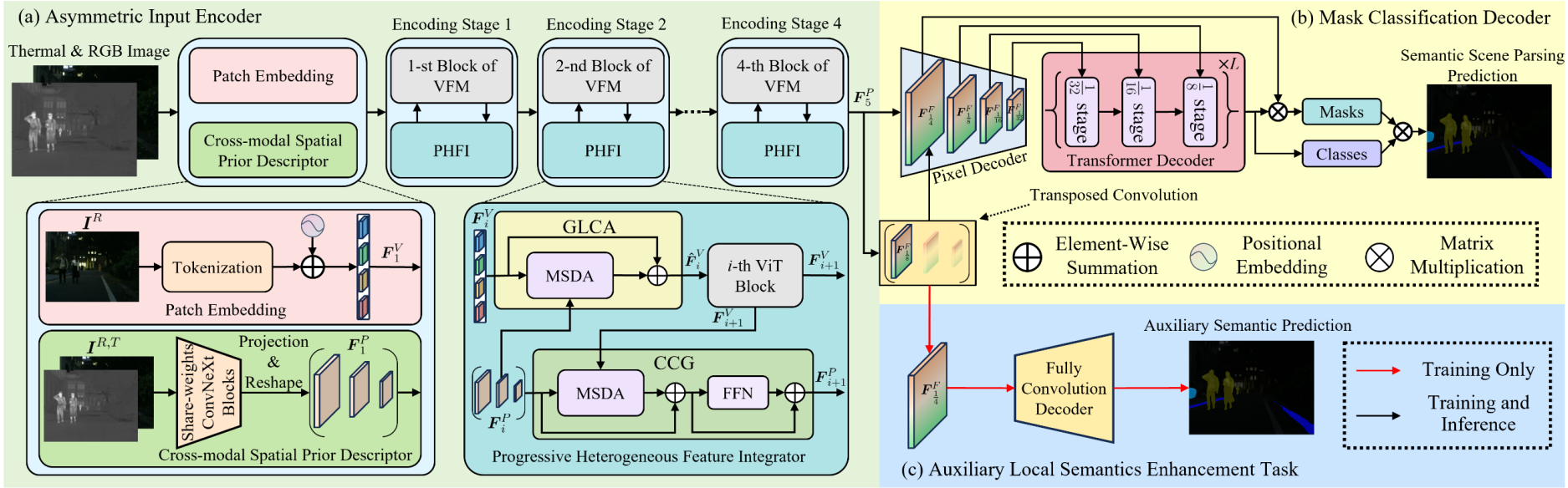


[8] **Fan, R.**, Ozgunalp, U., Hosking, B., Liu, M. and Pitas, I., 2019. Pothole detection based on disparity transformation and road surface modeling. **IEEE T-IP**.

- **Research Trends**

  - **From traditional CNNs to Transformers**, and now to vision foundation models, the field has evolved significantly.
  - Networks have progressed from **single-modal architectures to data-fusion networks**, particularly for multi-modal data fusion.
  - Today, a single vision foundation model can **address multiple tasks**.

• First Attempt – RGB-T scene parsing with VFM

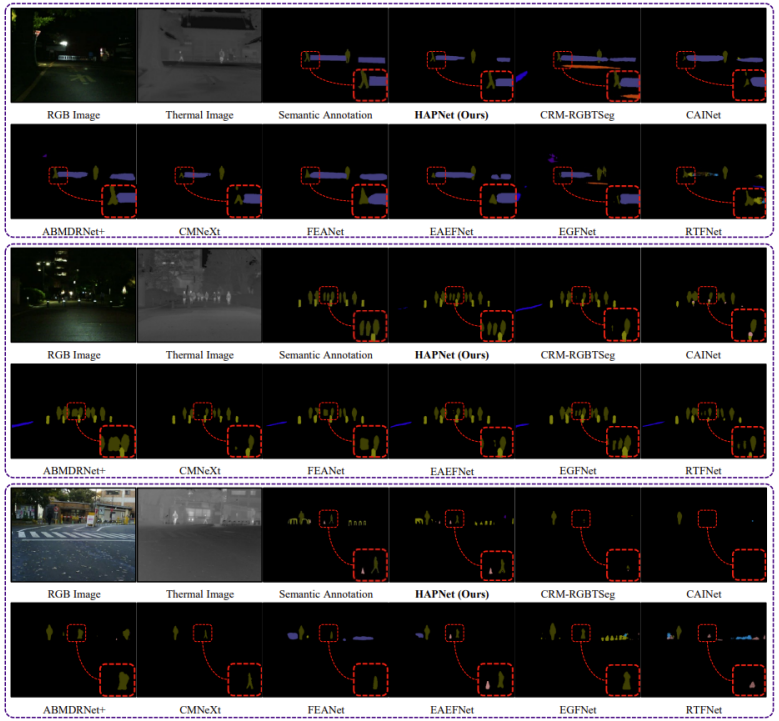


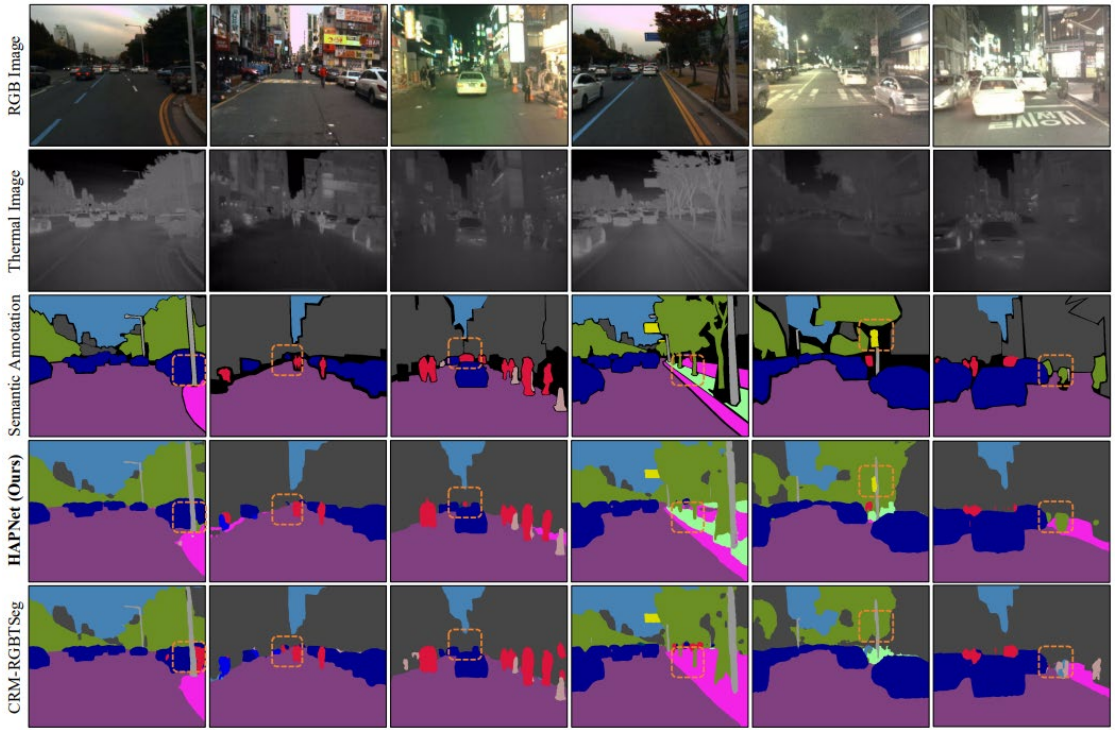
- **HAPNet**, a novel data-fusion network designed for accurate RGB-T scene parsing tasks.
- This **Transformer-CNN hybrid** encoder fully leverages the unique strengths of both modalities to achieve robust heterogeneous feature extraction and accurate parsing results.

- Li, J., Yun, P., Chen, Q. and **Fan, R.**, 2024. HAPNet: Toward Superior RGB-Thermal Scene Parsing via Hybrid, Asymmetric, and Progressive Heterogeneous Feature Fusion. **Accepted with Minor Revisions to IEEE T-ITS**.

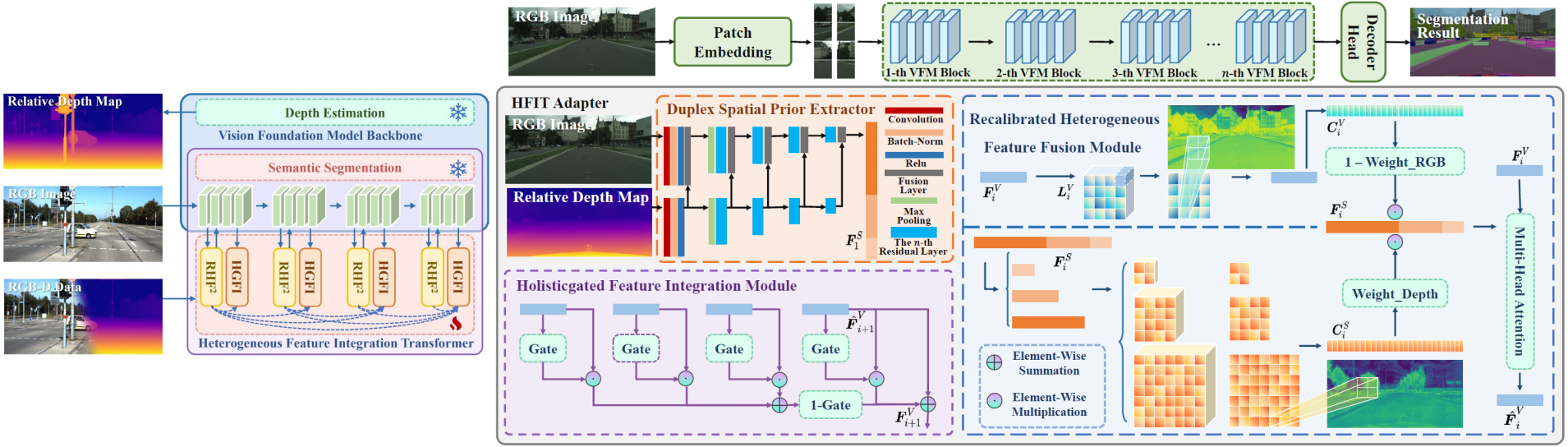
• First Attempt – RGB-T scene parsing with VFM

➢ HAPNet achieves state-of-the-art performance on three widely-utilized RGB-T scene parsing datasets: **MFNet, PST900, and KP Day-Night**.





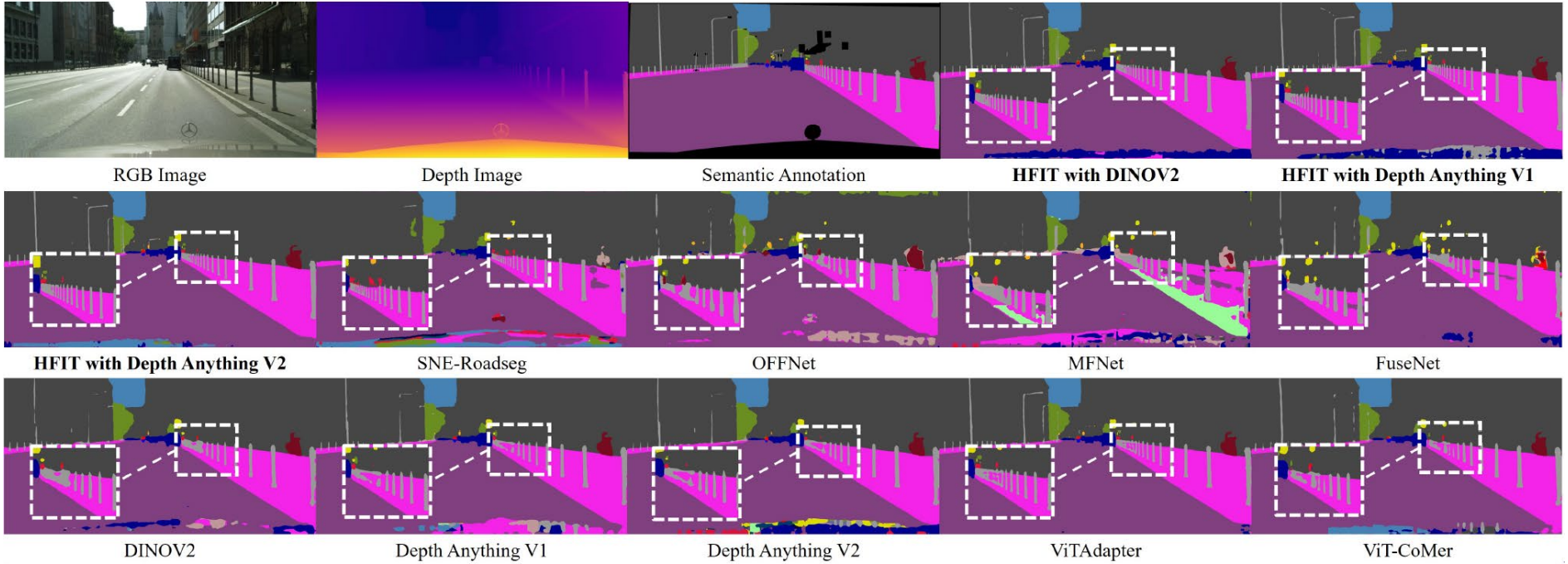
• Second Attempt – RGB-D scene parsing with VFM



- **HFIT**, a novel data-fusion network designed for accurate RGB-D scene parsing tasks.
- This **Transformer-CNN hybrid** encoder fully leverages the unique strengths of both modalities to achieve robust heterogeneous feature extraction and accurate parsing results.

- Guo, S., Wen, T., Liu, CW., Chen, Q., and **Fan, R.**, 2024. Fully Exploiting Vision Foundation Model's Profound Prior Knowledge for Generalizable RGB-Depth Driving Scene Parsing. **Submitted to IEEE T-IV**.

- **Second Attempt – RGB-D scene parsing with VFM**



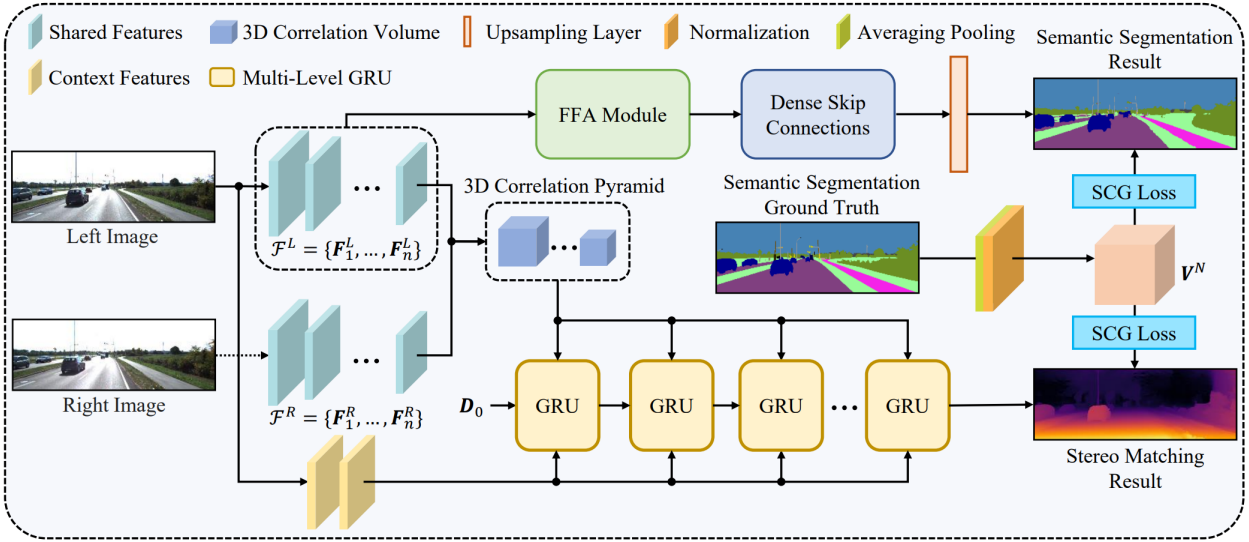
- Our proposed HFIT demonstrates superior performance compared to all other traditional RGB-Depth scene parsing networks, pretrained VFMs, and ViT adapters on the Cityscapes and KITTI Semantics datasets.

- **Existing challenges**

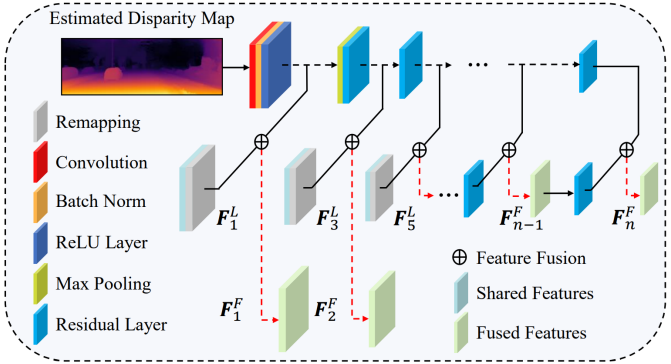
➢ When features extracted from RGB images and the "X" data are considered independently, several questions arise:

- Can additional tasks be introduced to enhance RGB-X scene parsing performance?
- Is it feasible to perform RGB-X scene parsing when the "X" data is unavailable?

• Recent Advances – RGB-X scene parsing + Stereo Matching



Overall Architecture

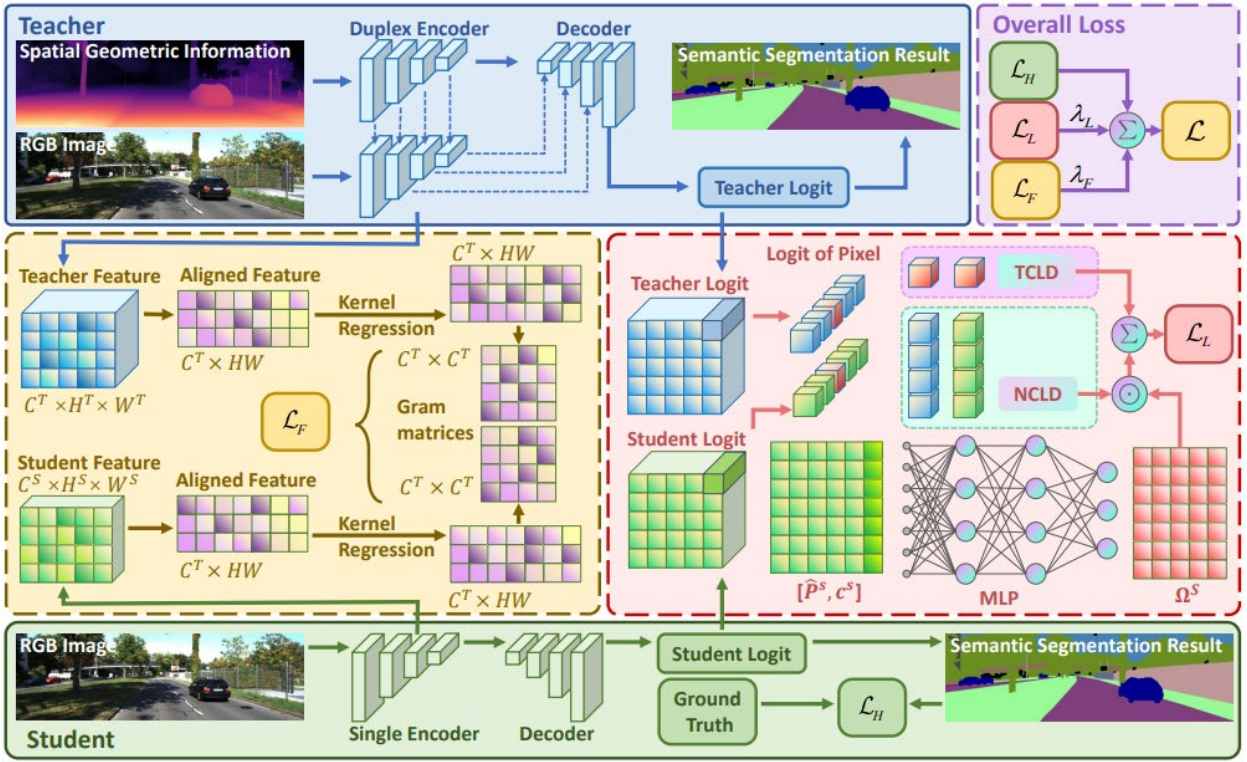


FFA module

➢ **S³M-Net**, a joint learning framework designed to address semantic segmentation and stereo matching simultaneously, where both tasks collaboratively leverage the features extracted from RGB images, enhancing the overall understanding of the driving scenario.

[9] Wu, Z., Feng, Y., Liu, C. W., Yu, F., Chen, Q., & **Fan, R.** (2024). **S³M-Net**: Joint Learning of Semantic Segmentation and Stereo Matching for Autonomous Driving. **IEEE T-IV**.

• Recent Advances – Spatial Geometric Prior Knowledge Infusion



- We implicitly infuse spatial geometric prior knowledge into visual semantic segmentation, distilling a teacher duplex-encoder (RGB-X) DNN architecture into a student DNN operating only RGB images.
- The student network with a single encoder can achieve comparable performance to the teacher network with a duplex encoder.

[10] Guo, S., Wu, Z., Chen, Q., Pitas, I. and **Fan, R.**, 2024. LIX: Implicitly Infusing Spatial Geometric Prior Knowledge into Visual Semantic Segmentation for Autonomous Driving. arXiv preprint arXiv:2403.08215.

• Summary

- RGB-X scene parsing typically provides a more comprehensive understanding of the environment.
- Heterogeneous feature fusion represents a popular research area warranting further attention.
- An asymmetric encoder architecture is often a more suitable choice for RGB-X scene parsing.
- When deploying AI algorithms on resource-limited hardware, it is crucial to consider computational complexity, especially since the discussed applications often require real-time performance.

Thank you very much for your attention!

Q & A