Fully Exploiting Vision Foundation Model's Profound Prior Knowledge for Generalizable RGB-Depth Driving Scene Parsing -Supplementary Material-

Sicen Guo, Tianyou Wen, Chuang-Wei Liu, Qijun Chen *Senior Member, IEEE*, Rui Fan *Senior Member, IEEE*

I. Datasets

We utilize two public datasets to evaluate the performance of our proposed methods: Cityscapes [1] and KITTI Semantics [2] datasets. Both the Cityscapes and KITTI Semantics datasets are highly regarded in autonomous driving research due to their comprehensive annotations, diversity in urban environments, and support for advancing scene parsing in challenging driving scenarios. Their details are as follows:

- The **Cityscapes dataset** offers real-world stereo images with a resolution of 2,048×1,024 pixels. It captures a variety of road conditions, traffic scenarios, diverse weather, and lighting conditions. This diversity makes it an ideal choice for models aimed at generalizable scene parsing across different urban environments. It includes pixel-level annotations for 19 semantic classes relevant to driving, including pedestrians, vehicles, and road markings, which are crucial for detailed and comprehensive scene understanding. Moreover, Cityscapes is commonly used as a benchmark in autonomous driving research, providing a solid basis for comparing the performance of different algorithms in scene parsing. We follow the official division for the training and validation sets, comprising 2,975 images for training and 500 for validation.
- The **KITTI Semantics dataset** consists of 200 real-world images taken from diverse driving scenarios. It includes semantic annotations for 19 different classes, consistent with the Cityscapes [1] dataset. The images are randomly split into training and validation sets at a ratio of 3:1. KITTI Semantics dataset provides synchronized sparse depth data alongside RGB images using a Velodyne HDL-64E LiDAR system, which is invaluable for applications requiring spatial awareness and precise depth perception. This depth information supports advanced understanding of 3D structure in driving scenes. Captured using a sensor-equipped vehicle on real streets, KITTI Semantics includes challenging conditions, such as occlusions, dynamic objects, and varying perspectives, making it highly representative of real-world driving scenes. KITTI is widely recognized in autonomous driving research as a rigorous testbed, especially for algorithms that need to perform well in challenging, real-world conditions.

II. Implementation Details

Our model was trained for 20,000 iterations on an NVIDIA RTX 3090 GPU, employing the AdamW optimizer [3]. The initial learning rate is 1×10^{-3} with a weight decay of 5×10^{-2}. To preserve the original aspect ratio of the dataset images, the Cityscapes dataset images were resized to a resolution of 1,792 × 896, while the KITTI Semantics dataset images were resized to 2,968 × 896. This resizing approach ensures consistency in aspect ratio across the datasets, allowing for accurate and comparable experimental results. Then, images were randomly cropped to a size of 448 × 448 pixels during the training process. The batch size is set to three to balance computational efficiency with memory constraints, ensuring stable model training. To enhance the model's robustness, we employed standard data augmentation techniques, such as random color adjustments, photometric distortion, rescaling, and flipping.

III. Comparisons of VFMs with Different Decoders

In this subsection, we conduct quantitative comparisons of VFMs with different decoders on the Cityscapes [1] and KITTI Semantics [2] datasets, respectively. As shown in Table I, Depth Anything V1 [9] achieves superior performance when the decoder is UperNet [6]. This observation may be attributed to the extensive receptive fields, which are enhanced by the pyramid pooling module at the final stage. Given the excellent performance of UperNet, our experiments will uniformly adopt it as the decoder. Moreover, we observe that Segmenter [7] consistently outperforms other Transformer-based methods. Conversely, the results of the more impressive Mask2Former [8] are unsatisfactory. We speculate that these unexpected results may be due to the smaller crop size, which restricts the range of spatial features that deformable convolutions can access, further reducing the model's ability to deform its kernel over meaningful parts of the image. Therefore, for resource-constrained driving scene parsing, CNN-based methods are

TABLE I: Quantitative comparisons of VFMs with different decoders on the Cityscapes and KITTI Semantics datasets.

Dataset	VFM Backbone	Decoder Type	Decoder	mFsc (%) ↑	mIoU (%) ↑	aAcc (%) ↑	mPre (%) ↑	mRec (%) ↑
Cityscapes	DINOv2 [4]	CNN-Based	DANet [5]	88.19	79.75	95.93	88.69	87.90
			UperNet [6]	**88.94**	**80.88**	96.21	89.12	**88.89**
		Transformer-Based	Segmenter [7]	87.65	78.98	95.65	88.45	86.99
			Mask2Former [8]	81.63	67.56	95.43	77.27	78.02
	Depth Anything V1 [9]	CNN-Based	DANet [5]	88.49	80.19	96.16	89.09	88.11
			UperNet [6]	89.20	81.24	**96.41**	**90.12**	88.41
		Transformer-Based	Segmenter [7]	88.38	80.08	96.05	89.95	87.08
			Mask2Former [8]	79.88	65.83	95.43	81.06	74.96
	Depth Anything V2 [10]	CNN-Based	DANet [5]	88.35	79.98	96.02	88.52	88.30
			UperNet [6]	88.93	80.85	96.27	89.54	88.44
		Transformer-Based	Segmenter [7]	87.88	79.33	95.86	89.12	86.84
			Mask2Former [8]	81.53	63.86	95.21	77.34	73.51
KITTI Semantics	DINOv2 [4]	CNN-Based	DANet [5]	86.59	77.79	95.32	88.25	85.63
			UperNet [6]	86.90	78.28	95.37	87.84	**86.52**
		Transformer-Based	Segmenter [7]	85.41	76.92	95.27	89.63	83.56
			Mask2Former [8]	76.71	53.43	94.27	81.93	61.22
	Depth Anything V1 [9]	CNN-Based	DANet [5]	86.61	78.07	95.42	89.77	84.83
			UperNet [6]	**86.28**	77.65	95.58	89.50	84.55
		Transformer-Based	Segmenter [7]	85.16	76.60	**95.62**	**89.86**	82.90
			Mask2Former [8]	80.56	52.35	94.46	81.51	59.88
	Depth Anything V2 [10]	CNN-Based	DANet [5]	86.51	77.63	95.17	88.51	85.48
			UperNet [6]	87.36	**79.14**	95.61	89.41	86.13
		Transformer-Based	Segmenter [7]	85.80	77.23	95.38	89.27	84.09
			Mask2Former [8]	78.26	54.07	94.30	76.09	61.97

TABLE II: The amounts of parameters, Flops and FPS of HFIT and other methods on the Cityscapes dataset.

Methods	Params(M)	Flops (GFLOPs)	FPS (img/s)
SNE-RoadSeg [11]	201.33	410.11	1.67
OFFNet [12]	25.22	21.54	1.22
MFNet [13]	0.74	6.12	4.78
FuseNet [14]	44.18	186.50	3.10
OCRNet [15]	55.52	176.83	2.03
KNet [16]	60.41	157.77	1.94
EMANet [17]	39.99	129.81	2.99
Single-Modal VFMs	308.11	313.74	1.32
ViTAdapter [18]	329.57	437.37	0.48
ViT-CoMer [19]	390.81	610.86	0.32
HFIT(ours)	412.52	471.56	0.20

more appropriate, as they prioritize local feature extraction and progressively construct a comprehensive understanding, making them more robust when dealing with limited input sizes. Whereas for transformer-based methods that highly rely on the global context, scenarios with diverse datasets and training spaces are more suitable for them, allowing them to capture long-range dependencies among features at different spatial locations.

IV. Parameters, Flops and FPS Comparisons

As shown in Table II, HFIT has the highest number of parameters and FLOPs among all the methods, resulting in the lowest Frames per second (FPS) of 0.20 img/s. Compared to previous models like SNE-RoadSeg [11] and FuseNet [14], HFIT's substantial parameter count and computational cost reflect its complex architecture, likely due to an enhanced feature extraction capability and deeper layers aimed at capturing intricate semantic information from driving scenes. This complexity is beneficial for tasks demanding high accuracy but results in a trade-off with inference speed, especially on large datasets like Cityscapes. In contrast, lightweight models like MFNet [13] maintain higher FPS at the cost of reduced representational power. To address this trade-off, a potential future strategy could involve optimizing HFIT's architecture by applying parameter reduction techniques, such as pruning or quantization, or by incorporating knowledge distillation strategy to retain performance while reducing model size.

V. Discussion

The advancements in multi-modal models for autonomous driving have opened new pathways for robust scene understanding and decision-making. In this section, we discuss how HFIT can contribute to advancing recent studies on large models in autonomous driving.

- **3D Spatial Understanding:** LiDAR-LLM [20] leverages large language models (LLMs) for understanding sparse

outdoor LiDAR data, offering a powerful approach for 3D scene comprehension. If combined with LiDAR-LLM, HFIT could better handle LiDAR-based data and gain deeper insights into 3D spatial structures within driving scenes. This would allow HFIT to process LiDAR data and depth-based visual cues simultaneously, improving its overall scene parsing and 3D spatial understanding abilities of driving environments.

- **Multi-modal Contextual Understanding:** The integration of LLMs into the HFIT framework would enable the model to interpret ambiguous or complex visual inputs through a language-informed lens. This capability could significantly improve semantic alignment across multiple sensor modalities (*e.g.*, RGB, depth, and LiDAR), allowing HFIT to understand context more effectively. By leveraging LLMs, HFIT could contextualize visual features in the driving environment and make better decisions based on high-level semantic understanding.
- **Graph-Based Reasoning for Sequential Decision-Making:** One promising direction for the future evolution of HFIT lies in incorporating graph-based reasoning for multi-step decision-making. HFIT currently excels at parsing RGB-D driving scenes, but it operates primarily as a static model that processes inputs in a single step. In contrast, DriveLM [21] introduces the concept of Graph VQA, where reasoning is structured through a series of perception, prediction, and planning question-answer pairs. By integrating this multi-step reasoning process into HFIT, we could enable it to handle more dynamic and evolving driving environments, where decisions need to be continuously updated based on new inputs. For example, HFIT could first localize key objects in the scene, then predict possible interactions between those objects (e.g., vehicle trajectories or pedestrian movements), and finally plan actions based on these predictions. This would bring HFIT closer to mimicking human-like reasoning in driving situations, making it more adaptable and responsive to real-time challenges.

In conclusion, the future evolution of HFIT lies in integrating advancements from recent multi-modal models in autonomous driving. These advancements would make HFIT not only more accurate and robust but also more flexible and human-like in its decision-making process, paving the way for smarter and more interactive autonomous driving systems.

REFERENCES

[1] M. Cordts *et al.*, "The CityScapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3213–3223. 1

[2] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3061–3070. 1

[3] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations (ICLR)*, 2019. 1

[4] M. Oquab *et al.*, "DINOv2: Learning robust visual features without supervision," *Computing Research Repository (CoRR)*, vol. abs/2304.07193, 2023. [Online]. Available: https://arxiv.org/abs/2304.07193 2

[5] H. Xue *et al.*, "DANet: Divergent activation for weakly supervised object localization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 6589–6598. 2

[6] T. Xiao *et al.*, "Unified perceptual parsing for scene understanding," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 418–434. 1, 2

[7] R. Strudel *et al.*, "Segmenter: Transformer for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 7262–7272. 1, 2

[8] B. Cheng *et al.*, "Masked-attention mask Transformer for universal image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 1290–1299. 1, 2

[9] L. Yang *et al.*, "Depth Anything: Unleashing the power of large-scale unlabeled data," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 10 371–10 381. 1, 2

[10] Yang, Lihe *et al.*, "Depth Anything V2," *Computing Research Repository (CoRR)*, vol. abs/2406.09414, 2024. [Online]. Available: https://arxiv.org/abs/2406.09414 2

[11] R. Fan *et al.*, "SNE-RoadSeg: Incorporating surface normal information into semantic segmentation for accurate freespace detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 340–356. 2

[12] C. Min *et al.*, "ORFD: A dataset and benchmark for OFF-Road freespace detection," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 2532–2538. 2

[13] Q. Ha *et al.*, "MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 5108–5115. 2

[14] Hazirbas *et al.*, "FuseNet: Incorporating depth into semantic segmentation via fusion-based cnn architecture," in *Proceedings of the Asian Conference on Computer Vision (ACCV)*. Springer, 2017, pp. 213–228. 2

[15] Y. Yuan *et al.*, "Object-contextual representations for semantic segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 173–190. 2

[16] W. Zhang *et al.*, "K-Net: Towards unified image segmentation," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, pp. 10 326–10 338, 2021. 2

[17] X. Li *et al.*, "Expectation-maximization attention networks for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 9167–9176. 2

[18] Z. Chen *et al.*, "Vision Transformer adapter for dense predictions," in *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. 2

[19] C. Xia *et al.*, "ViT-CoMer: Vision Transformer with convolutional multi-scale feature interaction for dense predictions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 5493–5502. 2

[20] S. Yang *et al.*, "Lidar-LLM: Exploring the potential of large language models for 3D lidar understanding," *Computing Research Repository (CoRR)*, vol. abs/2312.14074, 2023. [Online]. Available: https://arxiv.org/abs/2312.14074 2

[21] C. Sima *et al.*, "DriveLM: Driving with graph visual question answering," *Computing Research Repository (CoRR)*, vol. abs/2312.14150, 2023. [Online]. Available: https://arxiv.org/abs/2312.14150 3