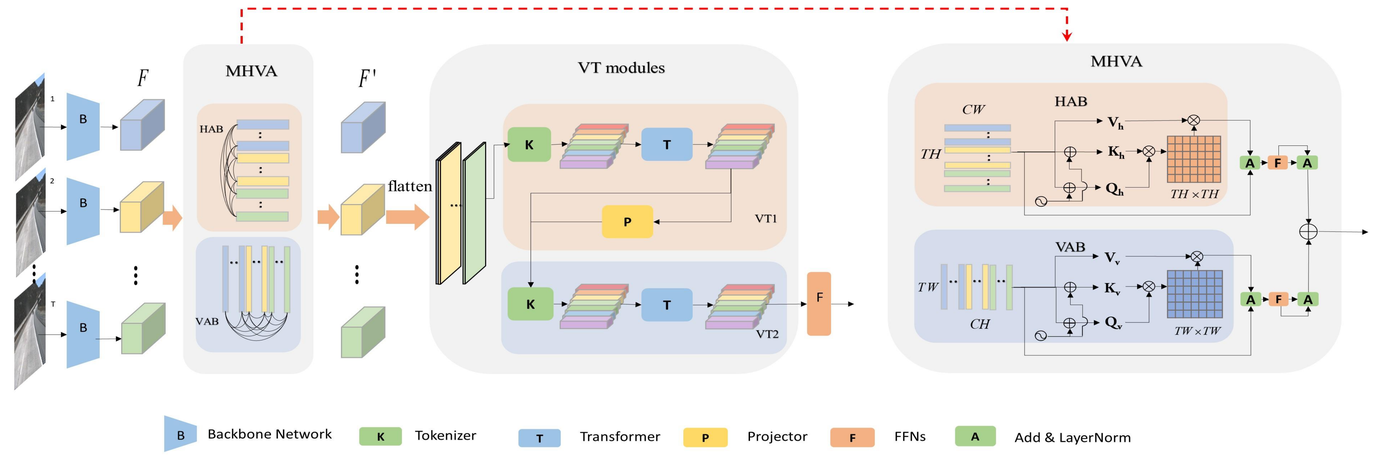
Lane Detection Transformer based on Multi-frame Horizontal and Vertical Attention and Visual Transformer Module.

Han Zhang, Yunchao Gu, Xinliang Wang, Junjun Pan and Minghui Wang
Beihang University, XueYuan Road No.37, HaiDian District, BeiJing, China

Abstract

Lane detection requires adequate global information due to the simplicity of lane line features and changeable road scenes. In this paper, we propose a novel lane detection Transformer based on multi-frame input to regress the parameters of lanes under a lane shape modeling. We design a Multi-frame Horizontal and Vertical Attention (MHVA) module to obtain more global features and use Visual Transformer (VT) module to get “lane tokens” with interaction information of lane instances. Extensive experiments on two public datasets show that our model can achieve state-of-art results on VIL-100 dataset and comparable performance on TuSimple dataset. In addition, our model runs at 46 fps on multi-frame data while using few parameters, indicating the feasibility and practicability in real-time self-driving applications of our proposed method.

Framework



Background

- Many methods [1, 2, 3] that focus on obtaining sufficient global features from the current frame are Inefficient and cumbersome.
- More visual information of previous frames in dynamic driving can infer more complete and accurate detection results than only using individual frames as input.

➢ We propose a novel curve-fitting lane detection method using multi-frame information and achieve instance-level lane detection.

➢ We design a lane detection Transformer based on MHVA and VT modules capturing more global information for parametric regression.

➢ The experiments verify the effectiveness of our method. In addition, we achieve the state-of-the-art results on the VIL-100 dataset and competitive performance on the TuSimple dataset, both with real time speed.

Approach & Experiments

We propose a novel lane detection method based on curve fitting, which is shown in Framework. It receives several time-ordered RGB images taken from a camera mounted in the vehicle as input and outputs the parameters of the predicted lanes. It consists of a backbone network, a MHVA module, several VT modules, and feed-forward networks (FFNs) for parametric regression. Given several continuous images in order, the backbone network first extracts a high level feature map F. Then, the feature map F and positional embedding E are fed into the MHVA module to get the enhanced feature map F’. After received “lane tokens” through VT modules, FFNs will regress the parameters on them. Hungarian fitting loss are used to train our network.

we evaluate the performance of our method on two public datasets TuSimple and VIL-100 [4]. Afterward, we provide a detailed ablation study to prove the effectiveness of the Multi-frame Mechanism and the rationality of our structure.

Conclusion

In conclusion, we propose a novel lane detection Transformer using multiple frames as input. Based on curve fitting, it can detect lanes directly and efficiently. Besides, the customized MHVA can capture more global information in two directions, and the VT modules are very effectual in improving detection result. Our method can achieve real-time results despite the use of multi-frame information, which enables the deployment in practical applications.

References

[1] Hou, Y., Ma, Z., Liu, C., Loy, C.C.: Learning lightweight lane detection cnns by self attention distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1013–1021 (2019)

[2] Pan, X., Shi, J., Luo, P., Wang, X., Tang, X.: Spatial as deep: Spatial cnn for traffic scene understanding. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)

[3] Zheng, T., Fang, H., Zhang, Y., Tang, W., Yang, Z., Liu, H., Cai, D.: Resa: Recurrent feature-shift aggregator for lane detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 3547–3554 (2021)

[4] Zhang, Y., Zhu, L., Feng, W., Fu, H., Wang, M., Li, Q., Li, C., Wang, S.: Vil-100: A new dataset and a baseline model for video instance lane detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15681–15690 (2021)