Creativity and Innovation
2025, VOL. 9, NO. 6, 115-118
DOI: 10.47297/wspciWSP2516-252725.20250906

A Review on the Applications of Transformer Models in Autonomous Driving Perception Systems

Haoxuan Han

No.58 High School, Qingdao, Shandong, 266000, China

ABSTRACT
With the rapid development of artificial intelligence technologies, convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have played a crucial role in perception and trajectory prediction tasks within intelligent driving systems. However, their limitations in global modeling and long-term dependency handling restrict their applicability in complex and dynamic environments. In recent years, the Transformer model, empowered by the attention mechanism, has achieved breakthroughs in natural language processing and has gradually been introduced into autonomous driving. It is now widely applied in multimodal information fusion, bird's-eye view (BEV) feature generation, and trajectory prediction. This paper reviews the major applications of Transformers in autonomous driving, summarizes their advantages and existing challenges, and discusses future research directions.

KEYWORDS
Transformer; Attention; Autonomous driving; AI training; Attention mechanism; Self-attention; BEV; CNN; RNN

1 Introduction

In recent years, intelligent driving technologies have become a central research focus for many scholars. The core task of artificial intelligence (AI) in this field is to enable vehicles to accurately and efficiently perceive their surrounding environment, thereby supporting reliable decision-making. Traditional deep learning models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have been widely adopted in object recognition and trajectory prediction. Nevertheless, both CNNs and RNNs show limitations in global modeling capabilities, which often lead to errors when handling complex driving scenarios.

The Transformer model, thanks to its self-attention mechanism, is capable of capturing global dependencies and supporting multimodal fusion. As a result, it is increasingly regarded as a promising direction for intelligent driving research. This paper introduces the current status and challenges of Transformer-based approaches in autonomous driving, aiming to provide useful references for future developments in AI modeling and intelligent driving technologies.

2 Related Concepts

2.1 The Role of Artificial Intelligence in Autonomous Driving

The applications of artificial intelligence (AI) in autonomous driving can be summarized in three main aspects: environmental perception, behavior prediction, and support for autonomous driving development.

First, in terms of environmental perception, Long Short-Term Memory (LSTM), a variant of RNN, has achieved prediction accuracy as high as 90.5% in driving behavior forecasting.

Second, regarding behavior and trajectory prediction, Transformer-based models have demonstrated significant improvements by reducing Average Displacement Error (ADE) and Final Displacement Error (FDE) by 30%–50%, thereby providing more reliable support for vehicle decision-making.

Third, from an industry perspective, autonomous driving is progressing from Level 2 (L2) toward Level 5 (L5). It is projected that by 2030–2035, new vehicles will be equipped with highly automated or even fully autonomous driving capabilities. These examples highlight AI's core role in advancing intelligent driving.

2.1 Current Applications of AI Models in Autonomous Driving

In the development of AI models for autonomous driving, traditional perception and prediction approaches mainly rely on CNNs and RNNs. For example, Tesla's AI Day technical reports noted that its early perception stack heavily relied on CNN-based visual networks, combined with recurrent structures to capture temporal features. Similarly, Waymo's research demonstrated the use of CNNs for feature extraction and LSTMs for trajectory prediction. Huawei's white paper on autonomous driving also emphasized CNN and RNN architectures as fundamental to environmental perception and behavior prediction.

These cases indicate that, prior to the emergence of Transformer models in autonomous driving, CNNs and RNNs were the mainstream frameworks adopted by both industry and academia.

2.3 CNN and RNN

Convolutional Neural Networks (CNNs) were first introduced by LeCun et al. in 1989 and were successfully applied to

handwritten digit recognition in LeNet-5 (1998). With the breakthrough performance of AlexNet in the 2012 ImageNet competition, CNNs became the core methodology in computer vision. In autonomous driving, CNNs have been widely used since around 2014 – 2016 for tasks such as lane detection, object recognition, and traffic sign recognition[6], providing effective solutions for environmental perception.

In contrast, research on Recurrent Neural Networks (RNNs) dates back to the 1980s. However, due to issues such as gradient explosion, their development was limited. The introduction of Long Short-Term Memory (LSTM) by Hochreiter and Schmidhuber in 1997 addressed these problems, paving the way for sequence modeling applications. Between 2016 and 2018, LSTM and related architectures were applied in trajectory prediction and driving behavior modeling, enabling predictions of the future positions of vehicles and pedestrians based on historical trajectories.

Traditional perception models in autonomous driving mainly rely on CNNs and RNNs. While CNNs excel at extracting local features from image data, they lack global modeling capacity. RNNs, on the other hand, are effective for sequential modeling but face challenges in training efficiency and in retaining long-term dependencies.

2.4 Development of Transformers

The Transformer model was designed to overcome the limitations of CNNs and RNNs. Since Bahdanau et al. first introduced the attention mechanism, researchers have continuously improved its modeling capacity. In 2015, Luong proposed local and global attention mechanisms, enhancing the Transformer's ability for global representation. Building upon these advances, Vaswani et al. introduced the Transformer architecture in 2017, which abandoned recurrence altogether and relied solely on attention mechanisms. This enabled effective global modeling while supporting parallel computation, fundamentally changing the trajectory of model development.

Subsequent milestones include Radford's GPT and Devlin's BERT in 2018, which established Transformers as mainstream models in natural language processing. In 2020, Dosovitskiy et al. proposed the Vision Transformer, demonstrating the Transformer's potential in computer vision tasks.

Since 2020, Transformers have been increasingly introduced into autonomous driving research, particularly in Bird's-Eye View (BEV) modeling, spatial representation, object detection, and trajectory prediction. With their global modeling and multimodal fusion capabilities, Transformers show strong potential for future applications in intelligent driving.

3 Related Applications

3.1 Transformers and BEV

Bird's-Eye View (BEV) has become one of the major research directions in autonomous driving in recent years. The BEV approach converts sensor data—such as from cameras, LiDAR, and millimeter-wave radar—into a unified top-down 2D representation. This enables tasks such as lane detection, traffic sign recognition, and pedestrian pose estimation. Compared with raw camera images or point clouds, BEV has the advantage of providing a clear spatial structure: road lanes, obstacles, and vehicles are intuitively represented, which facilitates subsequent computational analysis. However, BEV inevitably introduces information loss during projection, particularly with respect to height information, which poses challenges for modeling complex 3D driving environments. Against this backdrop, Transformers have been introduced into BEV modeling and training due to their self-attention mechanism and global modeling capability. Traditional CNN-based BEV models primarily capture local spatial features, making it difficult to model long-range relationships between regions. RNNs, while effective for sequential data, suffer from low computational efficiency and the risk of forgetting long-distance dependencies. By contrast, Transformers can globally model relationships across different spatial positions and modalities, while also supporting parallel computation. This gives them a clear advantage in BEV-based tasks. For example, the recently proposed BEVFormer leverages Transformers to convert multi-camera images across consecutive frames into a unified BEV representation, significantly improving the accuracy of object detection and trajectory prediction in autonomous driving systems. These advances suggest that the combination of BEV and Transformers is becoming a key trend in the development of perception systems for intelligent vehicles.

4 Transformer and Attention

4.1 History of Attention

As the complexity of perception and prediction tasks in autonomous driving has increased, traditional CNNs and RNNs have gradually revealed their limitations: CNNs are weak in capturing global dependencies, while RNNs suffer from gradient issues and low inference efficiency in handling long sequences.

To address these challenges, Bahdanau et al. (2014) first proposed the Attention Mechanism in neural machine translation. By assigning different weights to input sequence elements, the model could selectively "focus" on the most relevant information for the current prediction task, thereby improving global modeling ability while reducing computational cost.

Luong et al. (2015) further refined the mechanism by introducing local and global attention concepts, enhancing

flexibility and representational power. The real breakthrough came with Vaswani et al.' s Transformer architecture (2017), whose core innovation—Multi-Head Self-Attention—not only resolved long-distance dependency issues but also enabled parallel computation, greatly improving both training and inference efficiency. This architecture has since become the mainstream in autonomous driving research.

4.2 The Attention Mechanism

At its core, the idea of the Attention Mechanism is to compute the relevance between three elements—Query (Q), Key (K), and Value (V)—to determine which parts of the input sequence are most important for the current task. By focusing on specific segments of the input sequence, the decoder can generate the target sequence more effectively.

Mathematically, attention is typically expressed as:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Here, Q measures the correlation between queries and keys, the softmax operation converts this into a probability distribution, and a weighted sum produces the output. Multi-Head Attention further extends this by running the process in parallel across multiple subspaces, enabling the model to capture diverse feature relationships simultaneously.

In autonomous driving, attention mechanisms are particularly well-suited to multimodal and complex environments. For instance, in BEV models, attention can establish global relationships across different viewpoints and sensor inputs, leading to more precise understanding of road structures and interactions with dynamic obstacles. This theoretical foundation underpins the widespread application of Transformers in perception and prediction tasks for autonomous driving.

5 Challenges and Open Issues

5.1 Limited Real-Time Performance

In autonomous driving systems, Transformer models face notable shortcomings in inference efficiency. Research shows that the median reaction time of a Transformer model in urban street driving tasks can reach 370 ms—an unacceptably high latency for real-time applications. Another issue lies in computational complexity: the attention mechanism scales with O(n^2) when the input length is n. In autonomous driving, the inputs often involve multimodal high-resolution images or multi-channel LiDAR data, leading to an explosion in the number of tokens and computational load, far exceeding that of CNNs. For example, CNN-based 3D detection models (e.g., PointPillars) can achieve 20–25 FPS, while Transformer-based BEVFormer models run only at 6–10 FPS on the same hardware, falling short of the 10–30 FPS required for real-time driving. This highlights that Transformer models still suffer from heavy computation and high latency in autonomous driving applications.

5.2 Strong Data Dependence

Transformer models require large-scale, high-quality datasets (e. g., nuScenes, Waymo Open Dataset) for training. However, collecting and annotating such datasets is extremely costly, especially in complex traffic scenarios or under extreme weather conditions. Moreover, recognition accuracy often degrades under adverse conditions. For example, BEV models may suffer a 15–20% drop in detection accuracy during harsh weather. Sensor inputs can also be heavily affected —for instance, glare from high-beam headlights can cause BEV-based Transformers to misidentify vehicles on the road.

One study compared Transformer performance across sunny, rainy, and foggy conditions: the Mean Squared Error (MSE) was 0.023 in sunny weather, but rose to 0.028 (+21.7%) in rain and 0.035 (+52.2%) in fog[24]. This demonstrates that Transformers are highly vulnerable to environmental interference.

5.3 Hardware Deployment Challenges

Deploying Transformers in vehicle-mounted systems presents significant hardware barriers. Most automotive-grade AI chips are optimized for CNNs, using INT8 quantization or other low-precision computations. Transformers, by contrast, rely on large-scale floating-point matrix operations, which are inefficient on such platforms.

Additionally, Transformers often contain massive parameter counts and are computationally intensive. Direct deployment in embedded vehicle systems demands much higher compute and memory resources. For example, Tesla revealed in its 2022 Autonomy Day that its latest BEV + Transformer perception model has reached a billion parameters—ten times larger than the previous generation. This means that without aggressive model compression and operator optimization, it is difficult to meet the strict latency and energy constraints of automotive systems.

5.4 Black-Box Nature of Transformers

Another major challenge is the lack of interpretability. While attention mechanisms provide some visualization of "focus regions," they do not directly explain the underlying decision logic. This issue is particularly acute in multimodal

BEV-Transformer models, which process image, radar, and trajectory data simultaneously. Their internal representations are difficult to translate into human-understandable reasoning, making it challenging for developers to trace the causal basis of predictions.

In safety-critical applications like autonomous driving, such "black-box" characteristics create obstacles for accountability after accidents and pose challenges for regulatory approval and large-scale commercial deployment.

6 Future Directions

Future research on the integration of Transformers with autonomous driving can be explored from the following perspectives: First, more efficient attention mechanisms should be designed to alleviate the heavy computational burden of large-scale Transformer models. This would improve their real-time performance and energy efficiency in in-vehicle environments. Second, multimodel fusion strategies hold promise. By combining CNNs, RNNs, and Transformers, researchers can leverage their complementary strengths: CNNs excel at local feature extraction, RNNs at temporal sequence modeling, and Transformers at capturing global dependencies. A hybrid framework where each model contributes to different subtasks could lead to more robust overall performance. Third, data acquisition and efficient utilization remain critical challenges. Transformer training heavily depends on large-scale, high-quality datasets, but collecting such data for autonomous driving is expensive and highly affected by external factors such as weather and lighting conditions. Thus, future work should focus on methods that maximize data efficiency—improving robustness and generalization while reducing data collection costs.

7 Conclusion

This paper reviews the applications and development of Transformer models in autonomous driving. Beginning with the limitations of CNNs and RNNs, we discussed how Transformers, with their self-attention mechanism, enable global modeling and multimodal fusion, thereby showing significant potential in BEV modeling, trajectory prediction, and related tasks. At the same time, we highlighted several challenges faced by Transformers in vehicle deployment: insufficient real-time performance, heavy dependence on large datasets, hardware deployment difficulties, and poor interpretability. Looking ahead, we suggest that research should focus on the design of more efficient attention mechanisms, strategies for multimodel fusion, and improvements in data utilization efficiency. Addressing these directions will be essential to advancing the role of Transformers in the intelligent driving domain.

References

[1] A. Jain, A. Singh, H. S. Koppula, S. Soh, and A. Saxena, "Recurrent Neural Networks for Driver Activity Anticipation via Sensory-Fusion Architecture."

[2] J. Du, Y. Zhao, and H. Cheng, "Target-point Attention Transformer: A novel trajectory prediction network for end-to-end autonomous driving."

[3] K. Shirokinskiy, W. Bernhart, and S. Keese, "Advanced Driver-Assistance Systems: A ubiquitous technology for the future of vehicles."

[4] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems*, 2012.

[6] C. Chen, A. Seff, A. Kornhauser, and J. Xiao, "DeepDriving: Learning affordance for direct perception in autonomous driving," in *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, 2015.

[7] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, 1986.

[8] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[9] N. Deo and M. M. Trivedi, "Convolutional social pooling for vehicle trajectory prediction," in *Proc. IEEE CVPR Workshops*, pp. 1468–1476, 2018.

[10] A. Vaswani, L. Jones, N. Shazeer, A. N. Gomez, N. Parmar, J. Uszkoreit, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[11] M.-T. Luong, H. Pham, and C. D. Manning, "Effective Approaches to Attention-based Neural Machine Translation," in *Proc. EMNLP*, 2015.

[12] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving Language Understanding by Generative Pre-Training," OpenAI Technical Report, 2018.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. NAACL-HLT*, 2019.

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *Proc. Int. Conf. on Learning Representations (ICLR)*, 2021.

[15] F. Tian, "Applications and Challenges of Artificial Intelligence in Autonomous Driving Technologies."

[16] J. Zhong, Z. Liu, and X. Chen, "Transformer-based models and hardware acceleration analysis in autonomous driving: A survey."

[17] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Y. Qiao, and J. Dai, "BEVFormer: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers," 2022.

[18] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," in *Proc. ICLR*, 2015.

[19] M.-T. Luong, H. Pham, and C. D. Manning, "Effective Approaches to Attention-based Neural Machine Translation," in *Proc. EMNLP*, 2015.

[20] A. Vaswani, et al., "Attention Is All You Need," in *Proc. NeurIPS*, 2017.