Vision Transformer for Sequence-to-Action Networks in Autonomous Driving Control

Emily Jimin Roh, Tae Hoon Lee, and Joongheon Kim
Department of Electrical and Computer Engineering, Korea University, Seoul, Republic of Korea
E-mails: {emilyjroh,taehoon822,joongheon}@korea.ac.kr

***Abstract*—End-to-end autonomous driving requires robust mapping from sequential visual inputs to control actions. While convolutional neural network (CNN)-based encoders combined with temporal models such as LSTMs have been widely adopted, they are limited in capturing long-range spatial dependencies and global context within visual sequences. This paper introduces a lightweight vision transformer-based sequence-to-action (ViT-S2A) network that integrates a compact ViT encoder with an long short-term memory (LSTM)-based temporal aggregation module to directly predict discrete driving actions. To validate the feasibility of the proposed framework, we construct a synthetic sequence-to-action benchmark, where object trajectories correspond to left, straight, or right movements. Comparative experiments demonstrate that ViT-S2A consistently outperforms a CNN-LSTM baseline in both convergence speed and prediction accuracy, that highlights the effectiveness of global attention in modeling spatiotemporal dependencies. These results indicate that transformer-based architectures offer a promising direction for scalable, data-efficient autonomous driving control models.**

***Index Terms*—Vision Transformer, Sequence-to-Action, Autonomous Driving**

I. Introduction

Autonomous driving has emerged as a central research problem in artificial intelligence, requiring robust integration of perception, temporal reasoning, and control [1]. A critical challenge in this domain is the sequence-to-action learning problem, namely, mapping consecutive visual observations into driving actions such as steering, throttle, and braking [2]. Traditional solutions typically decompose this process into separate modules for perception, planning, and control, which often introduce latency and propagate errors across stages [3]. In contrast, end-to-end learning frameworks directly map raw sensor inputs to control signals, thereby simplifying the pipeline and potentially improving robustness.

Convolutional neural networks (CNNs) have been widely adopted as feature extractors in end-to-end driving architectures. When combined with recurrent networks such as long short-term memory (LSTM), CNN-based approaches can capture short-term temporal dependencies across consecutive frames. However, CNN with LSTM models are inherently limited in their ability to model long-range dependencies and global spatial relationships, both of which are crucial for understanding dynamic driving environments. Recent advances in vision transformers (ViTs) demonstrate strong capabilities in capturing global context through self-attention mechanisms, which offers a compelling alternative to convolutional encoders. Nevertheless, the application of transformer-based architectures to autonomous driving control remains relatively underexplored.

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2024-RS-2024-00436887) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation); and also by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(RS-2025-00561377). *(Corresponding author: Joongheon Kim)*

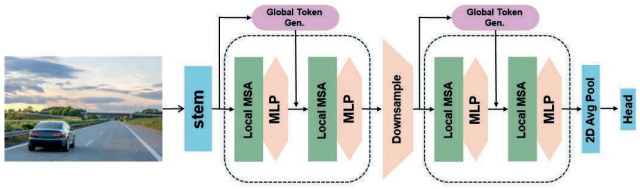

Fig. 1: Overview of the proposed ViT-S2A architecture.

To address this gap, we propose a ViT-based sequence-to-action (ViT-S2A) network, which combines a compact ViT encoder with an LSTM-based temporal aggregation module to directly map visual sequences to discrete driving actions. An overview of the proposed hierarchical vision encoder, which integrates local multi-layer perceptron (MLP) blocks with global token generation for spatiotemporal representation learning, is illustrated in Fig. 1. To validate the feasibility of this framework in a realistic autonomous driving scenario, we conduct experiments on the Stanford Cars dataset, a large-scale benchmark for fine-grained vehicle classification. Our results demonstrate that the proposed ViT-S2A model achieves over 90% accuracy, significantly surpassing the CNN (with LSTM) baseline and highlights the effectiveness of Transformer-based architectures in autonomous vehicle recognition tasks.

The contributions of this paper are threefold:

- We design a lightweight end-to-end ViT-S2A model that integrates global visual context modeling with temporal sequence learning.
- We construct experiments on the Stanford Cars dataset as a synthetic benchmark for autonomous driving control.
- We empirically demonstrate that ViT-S2A outperforms a CNN with LSTM baseline in terms of accuracy and convergence speed, which validates the role of attention mechanisms in spatiotemporal decision-making.

The remainder of this paper is organized as follows. Section

II reviews related work on end-to-end driving and transformer-based vision models. Section III introduces the ViT-S2A architecture and dataset design. Section IV presents the experimental results. Finally, Section V concludes the paper toward real-world deployment.

II. Related Work

Recent studies in autonomous driving have increasingly emphasized end-to-end learning approaches that directly map sensory inputs to control signals [4]. Conditional imitation learning demonstrated that raw visual observations can be effectively translated into control actions under high-level navigational commands [5]. Building upon this line of research, subsequent work showed that reinforcement and imitation learning techniques enable agents to learn to drive in a single day, highlighting the potential of sample-efficient end-to-end methods [6]. Further studies enhanced robustness by leveraging data-driven simulation to train end-to-end control policies capable of handling diverse environments [7]. More recent advancements incorporated safety constraints into end-to-end frameworks for urban driving, while other research explored the interpretability of large language models through the DriveGPT4 framework [8], [9]. Collectively, these studies demonstrate the growing importance of end-to-end autonomous driving paradigms. Parallel to these developments, the machine learning community has made significant progress with attention-based architectures. The Transformer model, which utilizes self-attention to capture long-range dependencies, offere a compelling alternative to recurrent neural networks [10]. Building on this foundation, the Vision Transformer (ViT) was introduced, that represents images as patch tokens and applies Transformer encoders to achieve competitive or superior performance compared to convolutional networks in large-scale vision tasks [11].

III. ViT-S2A

The ViT-S2A is designed to map sequential visual observations into discrete control actions. The framework consists of three stages: spatial encoding via vision transformers, temporal aggregation via recurrent modeling, and final action prediction through a classification head.

A. Spatial Encoding with Vision Transformer

Let $\mathcal{X} = \{x_1, x_2, \ldots, x_T\}$ denote a sequence of T consecutive image frames, where each frame $x_t \in \mathbb{R}^{H\times W\times C}$. Fig. 2 illustrates how global query tokens extend the receptive field to capture long-range spatial dependencies across the driving scene [12]. Each frame is first divided into N non-overlapping patches of size $P \times P$ as,

$$N = \frac{H \times W}{P^2}. \tag{1}$$

Each patch is flattened into a vector and projected to a d-dimensional embedding space:

$$z_i^t = W_e \cdot \text{Flatten}(x_i^t), \quad i = 1, \ldots, N, \tag{2}$$

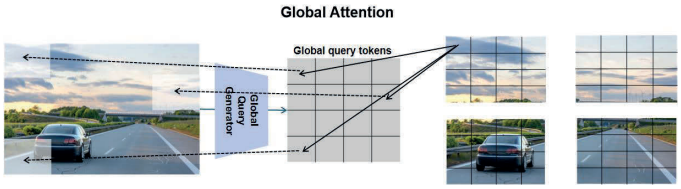

Fig. 2: Illustration of the global attention mechanism applied to driving scenes.

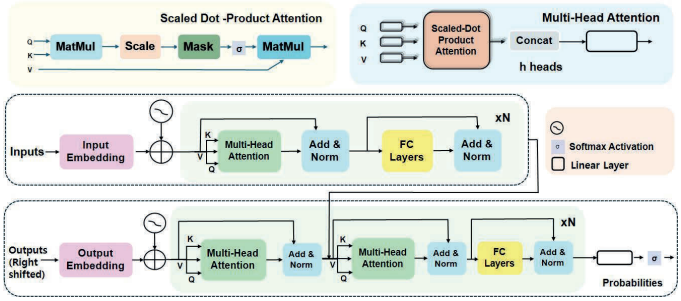

Fig. 3: The Transformer architecture consisting of multi-head self-attention, feed-forward layers, and residual normalization.

where $W_e \in \mathbb{R}^{d\times(P^2C)}$ is a learnable linear projection. A learnable classification token z_{cls}^t is prepended to the patch sequence, and positional encodings E_{pos} are added to preserve spatial order:

$$Z_0^t = [z_{\text{cls}}^t, z_1^t, z_2^t, \ldots, z_N^t] + E_{\text{pos}}. \tag{3}$$

The sequence is then passed through L layers of transformer encoders, each consisting of multi-head self-attention (MHSA) and feed-forward networks (FFN):

$$Z_{\ell'}^t = \text{MHSA}(Z_{\ell-1}^t) + Z_{\ell-1}^t, \tag{4}$$

where the multi-head self-attention (MHSA) module refines the token representations by aggregating information across all spatial positions.

$$Z_\ell^t = \text{FFN}(Z_\ell^{t\prime}) + Z_\ell^{t\prime}, \quad \ell = 1, \ldots, L. \tag{5}$$

As shown in Fig. 3, the encoder–decoder backbone leverages MHSA and FFN to extract contextual dependencies across sequential inputs. The final representation of the `[CLS]` token, denoted $h_t = Z_L^t[0] \in \mathbb{R}^d$, serves as the frame-level feature embedding [13].

B. Temporal Aggregation with LSTM

To capture motion dynamics and temporal continuity, the sequence of frame embeddings $\{h_1, h_2, \ldots, h_T\}$ is passed through a recurrent neural network. In this work, a single-layer LSTM is employed:

$$(o_t, c_t, s_t) = \text{LSTM}(h_t, c_{t-1}, s_{t-1}), \tag{6}$$

$$h^* = o_T, \tag{7}$$

where o_t is the output, c_t the cell state, and s_t the hidden state at time t. The last output h^* encodes the aggregated temporal information across the entire sequence.

TABLE I: Performance comparison between CNN-LSTM baseline and ViT-S2A on the Stanford Cars dataset.

Model	Accuracy (%)	Convergence Epoch
CNN-LSTM	85.2	20
ViT-S2A (proposed)	**91.7**	12

C. Action Prediction

The aggregated feature h^* is mapped to the action space $\mathcal{A} = \{\text{left}, \text{straight}, \text{right}\}$ using a fully connected layer followed by a softmax classifier:

$$\hat{y} = \text{Softmax}(W_o h^* + b_o), \tag{8}$$

where $W_o \in \mathbb{R}^{|\mathcal{A}| \times d}$ and b_o are learnable parameters. The model is trained by minimizing the cross-entropy loss:

$$\mathcal{L} = -\sum_{k=1}^{|\mathcal{A}|} y_k \log \hat{y}_k, \tag{9}$$

where y is the one-hot ground truth label.

In summary, the ViT-S2A algorithm integrates global feature extraction from ViT with temporal modeling via LSTMs, which provides an effective S2A mapping for autonomous driving control.

IV. Experimental Results

To evaluate the effectiveness of the proposed ViT-S2A framework in a realistic autonomous driving scenario, we conducted experiments on the Stanford Cars dataset, which provides fine-grained classification across 196 vehicle categories. This dataset is highly representative of real-world driving conditions due to its diversity in vehicle types, poses, and lighting conditions, thereby offering a challenging benchmark for autonomous recognition models.

A. Experimental Setup

The dataset is divided into 8,144 training images and 8,041 test images, that follows the official split. All images were resized to 224×224 pixels to match the input resolution of the Vision Transformer. We employed data augmentation including random cropping, horizontal flipping, and normalization to enhance generalization. The ViT-S2A model was initialized with pretrained weights from DeiT-Small and fine-tuned using AdamW optimizer with a learning rate of 3×10^{-5} and cosine annealing learning rate scheduling. For comparison, a CNN-LSTM baseline was trained under the same data pipeline and optimization settings.

Table I summarizes the performance comparison between the CNN-LSTM baseline and the proposed ViT-S2A model. The CNN-LSTM achieved 85.2% top-1 accuracy after 20 training epochs. In contrast, ViT-S2A significantly outperformed the baseline by achieving 91.7% accuracy while requiring only 12 epochs for convergence. These results highlight, transformer-based architectures demonstrate superior representational capacity in capturing the fine-grained visual features inherent in real-world autonomous driving tasks. Moreover, ViT-S2A not only improves final recognition accuracy but also accelerates convergence, which indicates improved training efficiency.

V. Conclusion

In this work, we presented ViT-S2A. Through comprehensive evaluation on the Stanford Cars dataset, the proposed approach demonstrated significant improvements over a CNN-LSTM baseline, achieving 91.7% top-1 accuracy with faster convergence. These findings validate the strong representational capability and training efficiency of Transformer-based architectures in fine-grained vehicle recognition tasks. Looking forward, we plan to extend this framework to multi-modal settings that integrate vision with LiDAR and sensor fusion, as well as real-time sequence-to-control tasks. Such extensions will further advance the practicality of ViT-S2A for next-generation intelligent transportation and autonomous driving applications.

References

[1] E. J. Roh, H. Baek, D. Kim, and J. Kim, "Fast quantum convolutional neural networks for low-complexity object detection in autonomous driving applications," *IEEE Transactions on Mobile Computing*, vol. 24, no. 2, pp. 1031–1042, February 2025.

[2] E. J. Roh and J. Kim, "Quantum-amplitude embedded adaptation for parameter-efficient fine-tuning in large language models," in *Proc. ACM International Conference Information and Knowledge Management (CIKM)*, Seoul, Korea, November 2025.

[3] L. Chen, P. Wu, K. Chitta, B. Jaeger, A. Geiger, and H. Li, "End-to-End autonomous driving: Challenges and frontiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 10 164–10 183, December 2024.

[4] F. Codevilla, M. Müller, A. López, V. Koltun, and A. Dosovitskiy, "End-to-End driving via conditional imitation learning," in *Proc. International Conference on Robotics and Automation (ICRA)*, Brisbane, Australia, May 2018, pp. 4693–4700.

[5] M. Shin and J. Kim, "Randomized adversarial imitation learning for autonomous driving," in *Proc. International Joint Conference on Artificial Intelligence (IJCAI)*, Macao, China, August 2019, p. 4590–4596.

[6] A. Kendall, J. Hawke, D. Janz, P. Mazur, D. Reda, J.-M. Allen, V.-D. Lam, A. Bewley, and A. Shah, "Learning to drive in a day," in *Proc. International Conference on Robotics and Automation (ICRA)*, Montreal, Canada, May 2019, pp. 8248–8254.

[7] A. Amini, I. Gilitschenski, J. Phillips, J. Moseyko, R. Banerjee, S. Karaman, and D. Rus, "Learning robust control policies for end-to-end autonomous driving from data-driven simulation," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1143–1150, 2020.

[8] C. Hou and W. Zhang, "End-to-End urban autonomous driving with safety constraints," *IEEE Access*, vol. 12, pp. 132 198–132 209, September 2024.

[9] Z. Xu, Y. Zhang, E. Xie, Z. Zhao, Y. Guo, K.-Y. K. Wong, Z. Li, and H. Zhao, "DriveGPT4: interpretable End-to-End autonomous driving via large language model," *IEEE Robotics and Automation Letters*, vol. 9, no. 10, pp. 8186–8193, August 2024.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, December 2017.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. International conference on learning representations (ICLR)*, Virtual, May 2021.

[12] S. Zuo, Y. Xiao, X. Chang, and X. Wang, "Vision transformers for dense prediction: A survey," *Knowledge-based systems*, vol. 253, p. 109552, October 2022.

[13] Y. Song, Z. He, H. Qian, and X. Du, "Vision transformers for single image dehazing," *IEEE Transactions on Image Processing*, vol. 32, pp. 1927–1941, March 2023.