Lane Change Prediction of Surrounding Vehicles using Video Vision Transformers

[1]Srinath NS
Student, School of Computer Science and Engineering
Vellore Institute of Technology
Chennai, India
srinath.ns2020@vitstudent.ac.in

Maria Anu Vensuslaus
Faculty, School of Computer Science and Engineering
Vellore Institute of Technology
Chennai, India
mariaanu.v@vit.ac.in

Joshua Thomas John Victor
Associate Professor, Computer Science
UOW Malaysia KDU Penang University College, Penang
Malaysia, jjoshua@uow.edu.my

Abstract

Anticipating lane changes of surrounding vehicles is paramount for the safe and efficient operation of autonomous vehicles. Previous works employ the usage of physical variables which do not contain contextual information. Recent methodologies relied on action recognition models such as 3D CNNs and RNNs, thereby dealing with complex architecture. Albeit the advent of transformers into action recognition, there are limited works employing transformer architectures. This research addresses the critical challenge of Lane Change Prediction (LCP) for autonomous vehicles, employing Video Action Prediction with a focus on the integration of ViViT (Video Vision Transformers). Utilizing the PREVENTION dataset, which provides detailed annotations of vehicle trajectories and critical events, the proposed approach outperforms prior methods, achieving over 85% test accuracy in predicting lane changes with a horizon of 1 second. Comparative analyses underscore ViViT's superiority in capturing spatio-temporal dependencies in video data while requiring fewer parameters, enhancing computational efficiency. This research contributes to advancing autonomous driving technology by showcasing ViViT's efficacy in real-world applications and advocating for its further exploration in enhancing vehicle safety and efficiency.

***Keywords**: Lane Change Prediction, Video Vision Transformers, Computer Vision, Tubelet Embeddings, Autonomous Vehicles*

1 Introduction

The evolution of autonomous driving technology has ushered in a new era of innovation, where the convergence of artificial intelligence and computer vision is reshaping the future of transportation. As autonomy levels beyond SAE Levels 2 and 3 become

increasingly critical, the pursuit of Levels 4 and 5 gains significance. These levels signify the pinnacle of driving automation, promising a future where vehicles navigate highways and urban landscapes with unparalleled precision and efficiency.

Lane change prediction stands out as a crucial component in the journey towards fully autonomous vehicles. In the dynamic realm of traffic, the ability to anticipate lane changes of surrounding vehicles is indispensable for ensuring safety and enhancing traffic flow. Lane change prediction empowers autonomous systems to adapt their trajectories proactively, navigate complex scenarios, and maintain safe distances, thereby ushering in a new era of road safety and efficiency.

Traditional approaches to lane change prediction have predominantly relied on physical variables such as speed, acceleration, and distance. While somewhat effective, these do not capture the nuanced intentions of surrounding vehicles. In contrast, human drivers rely on visual cues to anticipate lane changes, leveraging a complex interplay of spatial and temporal information. This human-inspired approach forms the basis for adopting Video Vision Transformers (ViViT) as a groundbreaking solution in the realm of autonomous driving.

ViViT represents a paradigm shift in video analysis, harnessing the power of Transformers to extract spatio-temporal features from video data. Unlike traditional methods that necessitate manual feature engineering or employ complex architectures, ViViT offers a streamlined and efficient means of capturing contextual information and modelling long-range dependencies. By leveraging spatio-temporal attention mechanisms, ViViT empowers autonomous systems to comprehensively analyse video sequences, enabling accurate lane change prediction with unprecedented precision.

The focal point of this research lies in lane change prediction using ViViT, specifically tailored for application in autonomous vehicles. By framing the lane change prediction problem as a Video Action Prediction task, this project aims to demonstrate the efficacy of ViViT in capturing intricate motion patterns and contextual information from video data. Leveraging the detailed annotations provided by the PREVENTION dataset, this research seeks to evaluate the performance of ViViT-based models in predicting lane changes of surrounding vehicles.

The contributions of this project extend beyond mere demonstration, aiming to showcase the superiority of ViViT over traditional methodologies. Through rigorous experimentation and meticulous evaluation, this research seeks to highlight the pivotal role of ViViT in advancing the field of autonomous driving technology. By achieving high prediction accuracy with significantly fewer parameters, ViViT offers a transformative solution that paves the way for safer, more efficient autonomous vehicles.

The primary contributions of this study can be delineated as follows:

1. Development of an end-to-end framework for lane change prediction utilizing front-facing cameras through video action recognition.
2. Exploration of the ViViT model's applicability and optimization to achieve optimal performance.
3. Comparative analysis of the ViViT model against other state-of-the-art approaches in the field.

The subsequent sections of this paper are structured as follows: Section 2 provides an extensive examination of prior studies. Section 3 offers a synopsis of the problem formulation. Section 4 delves into the methodology. Section 5 revolves around the experimental results. Finally, Section 6 discusses the conclusions drawn from the study and outlines avenues for future research.

2 Related Work

The subsequent section entails a thorough examination of related works, categorized into two main segments. Initially, a scrutiny of various datasets utilized in the domain is presented, elucidating their attributes, merits, and demerits. Following this, a comprehensive review of previous studies is conducted, delineating key methodologies, findings, and advancements in the field. This review aims to furnish a comprehensive understanding of the current landscape in lane change prediction research, setting the stage for the proposed methodology.

A. Datasets

The review of lane change datasets serves as a crucial foundation in understanding the landscape of research within the domain of autonomous driving. This section provides a comprehensive analysis of various datasets employed in lane change prediction studies, highlighting their respective characteristics, strengths, and limitations. Such an overview is essential for selecting appropriate datasets for model training and evaluation, ensuring the development of robust and generalizable lane change prediction algorithms.

TABLE I: Comparison of Traffic Datasets

Feature	**[18] PREVENTION**	**NGSIM HW101**	**NGSIM I80**	**[16] HighD**	**PKU**	**[17] ApolloScape**
Sensors Used	LiDAR, radar, cameras	Cameras	Cameras	Aerial images	2D-LiDARs	Cameras, Laser scanners
Coverage	Long-range (up to 80m)	Short-range	Short-range	Short-range	Short-range	Short-range
Lane Markings	Yes	No	No	No	No	No
Redundancy	Yes	No	No	No	No	No
Critical Situations	Yes	No	No	No	No	No

Table I compares the features of the available traffic datasets. The [18] PREVENTION dataset outperforms other datasets in almost all aspects. It utilizes a diverse range of sensors including LiDAR, radar, and cameras, providing redundancy and ensuring fault-tolerant development. In terms of coverage, it offers long-range coverage up to 80 meters around the ego-vehicle, allowing for accurate prediction of trajectories. Additionally, the inclusion of lane markings enhances road scene understanding, while annotations of critical situations such as cut-in, cut-out, and lane changes provide valuable data for prediction tasks. Thus, the study opts to proceed with the PREVENTION dataset.

B. Lane Change Prediction Systems

Lane change prediction in the context of autonomous vehicles is a critical task for ensuring safe and efficient navigation in complex traffic scenarios. A variety of input variables and methodologies have been explored in the literature to address this challenge. [1] introduces a methodology that considers lateral position, heading error, and lateral speed extracted from vehicle motion data to predict future vehicle trajectories, utilizing neural network models and SVM classifiers. However, the study's reliance on kinematic data alone may limit its ability to capture complex driving scenarios and environmental factors. [2] presents a sophisticated approach that extracts input variables from sensor fusion of radar and camera data, employing CNN-based models to predict lane change intentions of neighboring vehicles. Despite promising results, limitations exist due to small-scale datasets and simplifications in the driving scene reconstruction process. [3] adapts an LSTM-based methodology for trajectory prediction, utilizing historical trajectory data and lateral/longitudinal maneuver classifications. While the model predicts multi-modal trajectories effectively, challenges remain in accurately predicting lane changes, especially in unconventional traffic scenarios. [4] employs track histories of vehicles and convolutional social pooling layers to predict future vehicle positions, but its reliance on vehicle track data may overlook additional cues from visual and map-based information [5]. Another approach utilizes previous states of vehicles and composes them into a composite lane-based SRNN model, although it relies on simplified assumptions about vehicle behavior and may struggle to generalize to diverse driving environments. [6] explores input variables extracted from visual data, aiming to predict lane-change intentions based on motion history and context information encoded in images, but faces challenges in accurately differentiating between left and right lane changes. Additionally, studies such as [7], [8], [9], and [10] introduce various methodologies for lane-change prediction, utilizing input variables ranging from enriched RGB images to sequences of images obtained from front-view cameras on vehicles. While these studies demonstrate promising results, they also highlight the importance of addressing limitations such as computational complexity and reliance on visual cues alone. Specifically, [9] and [10] use video action recognition models [11], [12], [13], [14] and [15]. There is a need for further research to develop robust and generalized models capable of accurately predicting lane changes in diverse real-world driving scenarios. Consequently, Vision Transformer based architectures have not been explored in this context. As we embark on this journey towards fully autonomous vehicles, it is essential to address the challenges and limitations that lie ahead. Traditional approaches to lane change prediction have predominantly relied on physical variables such as speed, acceleration, and distance. While somewhat effective, these methods often struggle to capture the nuanced intentions of surrounding vehicles. In contrast, human drivers rely on visual cues to anticipate lane changes, leveraging a complex interplay of spatial and temporal information. This human-inspired approach forms the basis for adopting Video Vision Transformers (ViViT) as a groundbreaking solution in the realm of autonomous driving. While ViViT offers significant advancements in lane change prediction, there are still hurdles to overcome, including real-world deployment, regulatory considerations, and public acceptance. By acknowledging these challenges, we can work towards developing comprehensive solutions that address the needs of all stakeholders.

3 Problem Formulation

The problem formulation revolves around predicting lane change events as a multi-classification task, where the objective is to classify whether a surrounding vehicle will execute a left lane change, a right lane change, or maintain its current lane (no lane change) within a specified time horizon N. Illustrated in Figure 1, a lane change event occurs when the midpoint of the rear bumper aligns with the lane markings. The observation horizon, represented by a window of N images stacked frame by frame, captures the contextual information leading up to the prediction moment. This problem is approached as a prediction task, focusing on the Time To Lane Change (TTLC). Setting N to 40 frames (equivalent to 4 seconds at 10 FPS) and TTLC to 10 frames (1 second) defines the length of each sample as 50 frames, facilitating evaluation of the model's predictive capacity one second into the future.

Formally, this problem is framed as a video action prediction challenge, where an input video clip sample x with dimensions $\mathbb{R}^{F\times W\times H\times C}$ denoting frames, width, height, and channels respectively, is analyzed to forecast the probability of a lane change event denoted by y from the set $\{0, 1, 2\}$, representing Left Lane Change (LLCE), Right Lane Change (RLCE), or No Lane Change Event (NLCE). Preprocessing steps, detailed further, are applied to the input video. Following the final fully connected layer, a softmax activation function is employed for prediction, resulting in $\hat{y}$. The loss function utilized is sparse categorical cross-entropy (1).

$$Loss(y, \hat{y}) = -\sum_{i}^{N} y_i \log(\hat{y}_i) \tag{1}$$

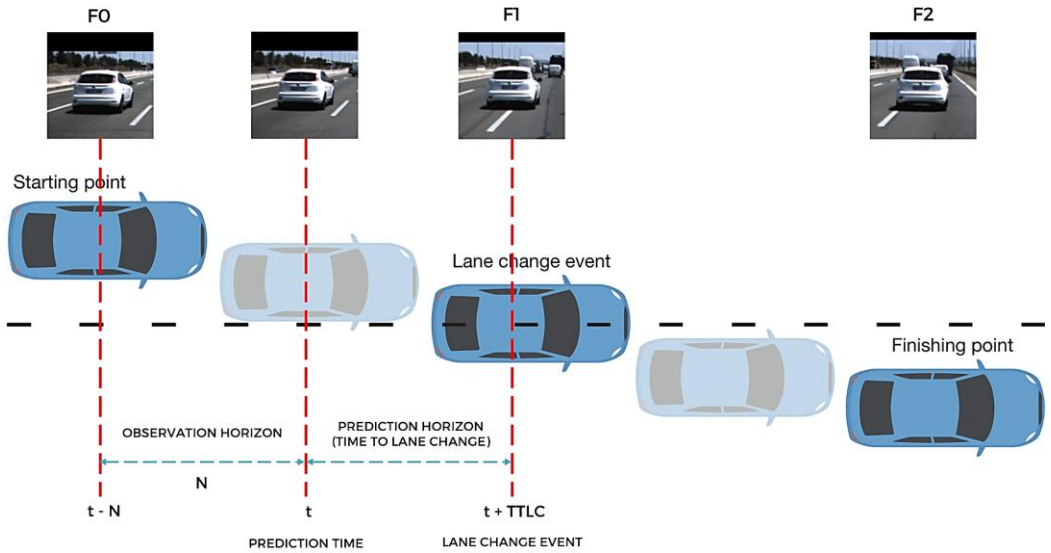


Fig 1. Problem Formulation: A "Lane Change Event" occurs when, in frame (F2), the midpoint of the vehicle's rear bumper surpasses the lane marking.

4 Methodology

The methodology section encompasses three key subsections: preprocessing, the architecture of the Video Vision Transformer, and the implementation of the Video Vision

Transformer. In the preprocessing stage, the raw data undergoes initial cleaning and transformation to prepare it for further analysis. Subsequently, the architecture of the Video Vision Transformer is delineated, outlining its components and underlying mechanisms. Finally, the section delves into the practical implementation of the Video Vision Transformer, detailing the steps involved in applying the model to the dataset for analysis and inference.

A. Preprocessing

The PREVENTION dataset includes original raw videos along with associated text files detailing lane change events and other contextual data. This dataset comprises five records, each containing multiple drives, with each drive containing raw video footage and corresponding text files. The videos are captured at a resolution of 1920 x 600 pixels, recorded at 10 frames per second (FPS), and consist of three colour channels. Lane change information is provided within the respective drives, with a total of 381 left lane change events and 468 right lane change events identified from the analysis. To balance event distribution, 420 instances of no lane change events were randomly sampled from all records, resulting in a total of 1269 training samples.

To aid contextual understanding, bounding boxes are generated for each vehicle in the green channel of the frame, referenced from the "detections_filtered.txt" file. Following this, the videos undergo preprocessing steps. Initially, the videos are centre cropped to dimensions of 1600 x 600 pixels to remove irrelevant context. Subsequently, spatial downsampling to 400 x 400 pixels is applied to reduce computational complexity. Additionally, the frame rate is halved from 10 to 5 FPS to further alleviate computational demands. This preprocessing results in each sample being represented as a tensor with dimensions of $\mathbb{R}^{25\times400\times400\times3}$. Finally, the preprocessed videos are divided into training, testing, and validation sets in an 80-10-10 ratio, yielding 1015 training, 127 testing, and 127 validation samples, respectively. The impact of the preprocessing steps is depicted in Fig 2.

(a) (b)

Fig 2. Preprocessing: (a) Sample Frame of a video before preprocessing (b) Sample Frame of a video after preprocessing.

B. Video Vision Transformers for Lane Change Prediction

After completing data preprocessing, an effective deep learning framework becomes essential for executing the lane change prediction task. In the current approach, Video Vision Transformers [21] are employed for this purpose.

1) *Tubelet Embedding:* Tubelet Embedding technique is utilized to convert a given video clip sample $x \in \mathbb{R}^{F\times W\times H\times C}$ into a sequence of tokens $\tilde{\boldsymbol{k}} \in \mathbb{R}^{n_f\times n_w\times n_h\times d}$, akin to [21]. Unlike patch embedding methods for images [20], Tubelet Embedding, proposed by [21], suggests extracting non-overlapping, spatio-temporal tubes from the input sequences of

video frames. These tubes encompass both temporal and frame-specific patches. The resulting patches are then linearly projected to generate multiple video-encoded tokens. For a tubelet with dimensions $f \times w \times h$, the number of encoded tokens is determined by the temporal, width, and height dimensions, denoted by (2).

$$n_f = \left\lfloor \frac{F}{f} \right\rfloor , n_w = \left\lfloor \frac{W}{w} \right\rfloor \text{ and } n_h = \left\lfloor \frac{H}{h} \right\rfloor \quad (2)$$

Moreover, [21] recommends incorporating an additional CLS token into the set of embedded tokens, inspired by [22]. This CLS token is tasked with aggregating global video frame information to facilitate final prediction. Additionally, a learned positional embedding, as proposed by [19], denoted as $P \in \mathbb{R}^{N \times d}$, is appended to the tokens to retain positional information, ensuring that self-attention remains permutation invariant.

2) *Spatio-Temporal Attention:* The research opts for the **spatio-temporal attention** (Fig 3) variant of ViViT, introduced by Arnab et al. In this approach, all video tokens undergo tubelet embedding and are subsequently forwarded directly to a standard transformer encoder [19]. The sequence of input tokens is

$$\boldsymbol{m} = [m_{cls}, \boldsymbol{E}x_1, \boldsymbol{E}x_2, \dots, \boldsymbol{E}x_N] + P \quad (3)$$

where $\boldsymbol{E}$ is patch embedding, resulting in equivalent representation such as a 3D convolution.

3) *Transformer Encoder:* The encoder comprises multiple stacks of L identical blocks, each consisting of two components: Multi-Head Self Attention ($MHSA$) and Multi-Layer Perceptron (MLP). Both components include Layer Normalization ($LayerNorm$) and a residual skip connection. At each layer, the model outputs embeddings of dimension D. The previous m vector is passed through the transformer encoder to generate the output vector o.

$$M_l = m_{l-1} + MHSA\big(LayerNorm(m_{l-1})\big) \qquad l = 1, \dots, L \quad (4)$$

$$O_{l+1} = M_l + MLP\big(LayerNorm(M_l)\big) \qquad l = 1, \dots, L \quad (5)$$

The Self Attention (SAT) is a crucial element of $MHSA$, responsible for identifying significant connections among all input tokens. To achieve this, the input vector m undergoes projection into three distinct matrices for each SAT component: Q (Query), K (Key), and V (Value). This projection is accomplished through multiplication with trainable weights W_Q, W_K, and W_V, respectively.

$$Q = m \times W_Q \quad (6)$$

$$K = m \times W_K \quad (7)$$

$$V = m \times W_V \tag{8}$$

The Queries Q are subjected to multiplication by the transpose of Keys K^T. The resulting vector is then divided by the square root of the embedding dimension D to mitigate the impact of peaky affinities. Subsequently, a SoftMax activation is applied, and the resulting output is multiplied by the Values V to generate the final output referred to as Head H.

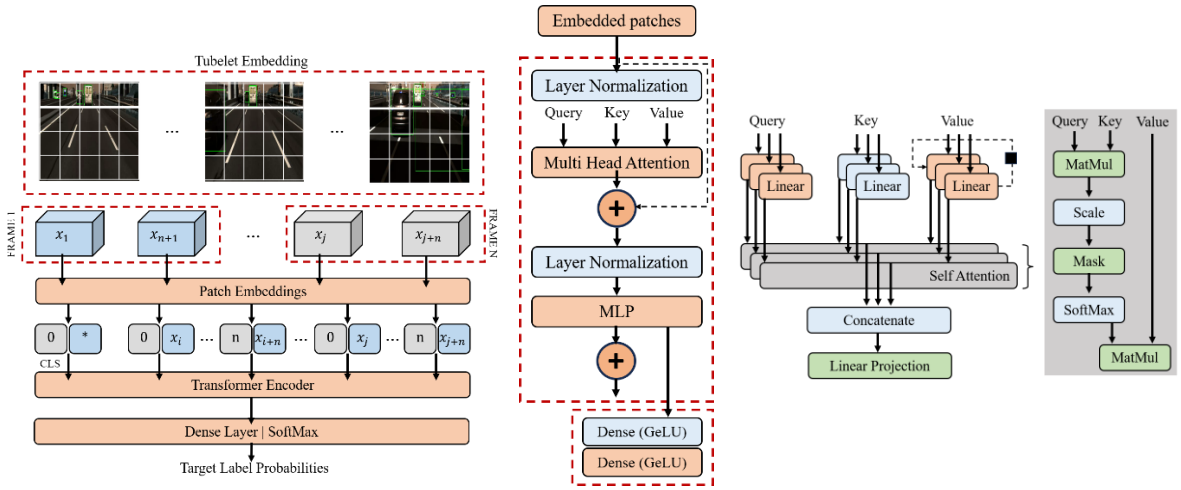
$$H = SoftMax\left(\frac{Q \times K^T}{\sqrt{D}}\right) \times V \tag{9}$$

SAT is employed h times to yield h attention heads. The outcomes of each attention head are concatenated and then processed through a feedforward layer equipped with learnable weights W^0.

$$MHSA = concat(SAT_1, SAT_2, SAT_3, \dots, SAT_N) \times W^0 \tag{10}$$

$$\boldsymbol{O} = [\boldsymbol{o_0}, \boldsymbol{o_1}, \boldsymbol{o_2}, \boldsymbol{o_3}, \dots, \boldsymbol{o_N}] \tag{11}$$

The MLP section consists of fully connected dense layers employing GeLU activation. The initial token of the output vector, denoted as $\boldsymbol{O}^0$, represents the CLS token, essential for classification purposes. This token undergoes processing through a dense layer with SoftMax activation to generate a probability distribution for the video clip's target label. Depending on this distribution, the clip is categorized into RLCE, NLCE, or LLCE.



(a) (b) (c)

Fig 3. Overview of (a) ViViT Model Architecture (b) Transformer Encoder (c) Self Attention

5 Experimental Results

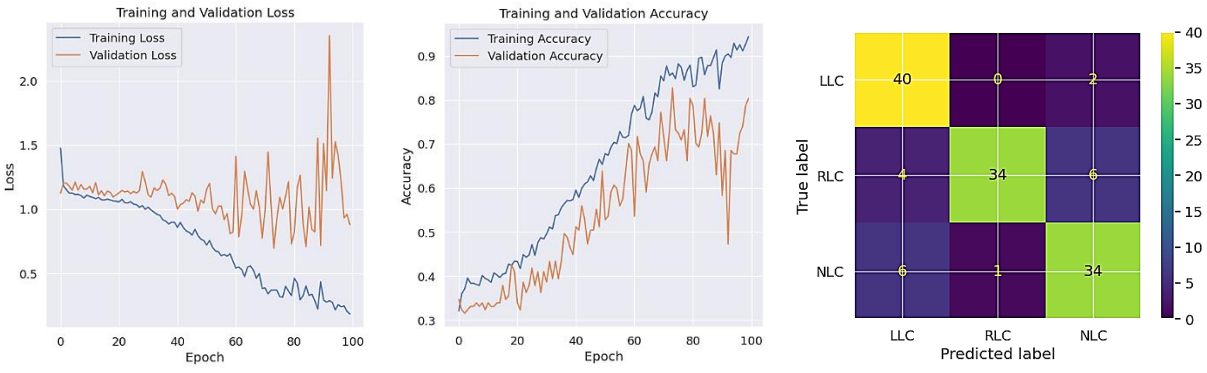
The Experimental Results section presents a comprehensive evaluation of the ViViT model, encompassing the determination of optimal hyperparameter combinations and the model's performance during inference. Additionally, a comparative study is conducted to assess the efficacy of the ViViT model against existing methodologies.

A) Evaluation of ViViT Model for Lane Change Prediction

This section entails the assessment of the ViViT model's efficacy in predicting lane changes. The ViViT model, implemented according to the methodology outlined in [21], underwent minor adjustments. Notably, a 3D convolutional layer was utilized to execute the learnable tubelet embedding, drawing inspiration from [20]. Furthermore, a refinement was introduced where, instead of solely considering the CLS token, all final token representations underwent layer normalization anew. Subsequently, global average pooling was employed to render them into a 1D format, enabling the aggregation of information across the entire sequence. This pooling operation effectively reduced the spatial dimensionality of the token representations while preserving crucial features, thereby priming them for predictive analysis.

Various combinations of hyperparameters were tested to identify the optimal set yielding the highest accuracy. The specific combination of hyperparameters resulting in the greatest accuracy following multiple tuning iterations is detailed in Table II.

Following extensive hyperparameter tuning, the model achieved its peak training and validation accuracies of 94.38% and 80.04%, respectively, after 100 epochs. Notably, convergence was observed around the 74th epoch, where the model attained training and validation accuracies of 86.11% and 82.68%, respectively. The learning curves depicting the performance on the training and validation sets are illustrated in Fig 4. These curves demonstrate a decent learning algorithm, as both reach a stable point with minimal discrepancy.



(a) (b) (c)

Fig. 4: Learning Curves of Training and Validation (a) Loss and (b) Accuracy and (c) Test Set Confusion Matrix

TABLE II. Tuned Hyper Parameters

Hyperparameter	Value/ Attribute
Batch Size	4
Input Dimension	(25, 400, 400, 3)
Number of Classes	3
Learning Rate	0.0001
Weight Decay	0.001

Epochs	100
Patch Size	(4, 32, 32)
Layer Normalization	0.0001
Projection Dimension	1024
Number of Attention Heads	8
Number of Transformer Layers	8
Optimizer	Adam

TABLE III. Evaluation Metrics

	Precision	Recall	F1-Score	Support
LLCE	0.80	0.95	0.87	42
RLCE	0.97	0.77	0.86	44
NLCE	0.81	0.83	0.82	41
Accuracy			0.85	127
Macro Average	0.86	0.85	0.85	127
Weighted Average	0.86	0.85	0.85	127

The evaluation metrics of the test set is detailed in Table III. The model demonstrates good performance in predicting lane change events, with high precision, recall, and F1-score values across all classes. Specifically, the model achieves precision values of 0.80, 0.97, and 0.81 for LLCE, RLCE, and NLCE classes respectively, indicating its ability to make accurate positive predictions for each class. The recall values of 0.95 for LLCE and 0.83 for NLCE suggest that the model effectively captures most of the actual instances of lane changes. The F1-score values, which consider both precision and recall, are also high, indicating a balanced performance. Additionally, the overall accuracy of 0.85 demonstrates the model's ability to correctly classify lane change events. Overall, the results suggest that the model performs well in predicting lane change events across different classes.

B) Comparative Analysis

The following section provides a comparative analysis of the proposed Video Vision Transformer (ViViT) methodology against existing approaches for lane change prediction. The superior performance of the Video Vision Transformer (ViViT) over other models in lane change prediction can be attributed to several factors. Firstly, ViViT leverages the strengths of transformer architectures, which have lower inductive biases compared to traditional CNN or RNN models, leading to better generalization and improved performance on unseen data. Additionally, transformers excel at capturing global information and long-range dependencies in large datasets, which is crucial for tasks such as lane change prediction where contextual information over extended periods is vital.

TABLE IV: Comparison of the proposed methodology with other approaches

Method	Accuracy
Lane SRNN [5]	48.70%
CNN LSTM [6]	74.41%
Vision Transformer [8]	81.10%
I3D [9][13]	83.28%
Two Stream [9][12]	84.54%
X3D [10]	84.79%
Video Vision Transformer	**85.04%**
Spatio-Temporal Multiplier [9][14]	85.69%
Slow Fast [9][15]	88.64%

While ViViT outperforms several models, it falls slightly behind the last two entries in Table IV. This may be attributed to transformers' reliance on large amounts of data for optimal performance. However, [19][21] Transformers typically require fewer computational resources in terms of Floating-Point Operations (FLOPs) and have a smaller number of parameters compared to models like Slow Fast [14] and Spatio-Temporal Multiplier [15]. This computational efficiency makes ViViT a practical choice for real-world applications where efficiency is paramount.

6 Conclusion

This study has presented ViViT for lane change prediction in autonomous vehicles, demonstrating notable performance improvements over existing methodologies. ViViT excels in capturing spatio-temporal dependencies in video data while maintaining computational efficiency. However, challenges remain in accurately distinguishing between left and right lane changes, particularly in complex traffic scenarios. Further exploration into ViViT variants and the incorporation of techniques such as GANs for data augmentation could enhance model robustness and generalization. Investigating the impact of inductive biases and the acquisition of larger datasets will be crucial for improving ViViT's performance. Research efforts should focus on refining ViViT's ability to handle complex traffic scenarios, potentially by incorporating additional contextual information or refining the model architecture. By addressing these challenges and continuing to innovate, ViViT holds the potential to revolutionize lane change prediction and contribute significantly to the advancement of autonomous driving technology.

References (MLA)

[1] Izquierdo, Ruben, et al. "Vehicle trajectory and lane change prediction using ANN and SVM classifiers." *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2017.

[2] Lee, Donghan, et al. "Convolution neural network-based lane change intention prediction of surrounding vehicles for ACC." *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2017.

[3] Deo, Nachiket, and Mohan M. Trivedi. "Multi-modal trajectory prediction of surrounding vehicles with maneuver based lstms." *2018 IEEE intelligent vehicles symposium (IV)*. IEEE, 2018.

[4] Deo, Nachiket, and Mohan M. Trivedi. "Convolutional social pooling for vehicle trajectory prediction." *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2018.

[5] Patel, Sajan, et al. "Predicting future lane changes of other highway vehicles using RNN-based deep models." *arXiv preprint arXiv:1801.04340* (2018).

[6] Izquierdo, Rubén, et al. "Experimental validation of lane-change intention prediction methodologies based on CNN and LSTM." *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2019.

[7] Izquierdo, Rubén, et al. "Vehicle lane change prediction on highways using efficient environment representation and deep learning." *IEEE Access* 9 (2021): 119454-119465.

[8] Konakalla, Neha, Avrum Noor, and Josh Singh. "Cnn, cnn encoder-rnn decoder, and pretrained vision transformers for surrounding vehicle lane change classification at future time steps." (2022).

[9] Biparva, Mahdi, et al. "Video action recognition for lane-change classification and prediction of surrounding vehicles." *IEEE Transactions on Intelligent Vehicles* 7.3 (2022): 569-578.

[10] Liang, Kai, Jun Wang, and Abhir Bhalerao. "Lane Change Classification and Prediction with Action Recognition Networks." *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2022.

[11] Karpathy, Andrej, et al. "Large-scale video classification with convolutional neural networks." *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2014.

[12] Simonyan, Karen, and Andrew Zisserman. "Two-stream convolutional networks for action recognition in videos." *Advances in neural information processing systems* 27 (2014).

[13] Carreira, Joao, and Andrew Zisserman. "Quo vadis, action recognition? a new model and the kinetics dataset." *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.

[14] Feichtenhofer, Christoph, Axel Pinz, and Richard P. Wildes. "Spatiotemporal multiplier networks for video action recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

[15] Feichtenhofer, Christoph, et al. "Slowfast networks for video recognition." *Proceedings of the IEEE/CVF international conference on computer vision*. 2019.

[16] Krajewski, Robert, et al. "The highd dataset: A drone dataset of naturalistic vehicle trajectories on german highways for validation of highly automated driving systems." *2018 21st international conference on intelligent transportation systems (ITSC)*. IEEE, 2018.

[17] Huang, Xinyu, et al. "The apolloscape dataset for autonomous driving." *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2018.

[18] Izquierdo, Rubén, et al. "The prevention dataset: a novel benchmark for prediction of vehicles intentions." *2019 IEEE intelligent transportation systems conference (ITSC)*. IEEE, 2019.

[19] Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

[20] Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv preprint arXiv:2010.11929* (2020).

[21] Arnab, Anurag, et al. "Vivit: A video vision transformer." *Proceedings of the IEEE/CVF international conference on computer vision*. 2021.

[22] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

References (APA)

[1] Izquierdo, R., Parra, I., Muñoz-Bulnes, J., Fernández-Llorca, D., & Sotelo, M. A. (2017, October). Vehicle trajectory and lane change prediction using ANN and SVM classifiers. In *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)* (pp. 1-6). IEEE.

[2] Lee, D., Kwon, Y. P., McMains, S., & Hedrick, J. K. (2017, October). Convolution neural network-based lane change intention prediction of surrounding vehicles for ACC. In *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)* (pp. 1-6). IEEE.

[3] Deo, N., & Trivedi, M. M. (2018, June). Multi-modal trajectory prediction of surrounding vehicles with maneuver based lstms. In *2018 IEEE intelligent vehicles symposium (IV)* (pp. 1179-1184). IEEE.

[4] Deo, N., & Trivedi, M. M. (2018). Convolutional social pooling for vehicle trajectory prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 1468-1476).

[5] Patel, S., Griffin, B., Kusano, K., & Corso, J. J. (2018). Predicting future lane changes of other highway vehicles using RNN-based deep models. *arXiv preprint arXiv:1801.04340*.

[6] Izquierdo, R., Quintanar, A., Parra, I., Fernández-Llorca, D., & Sotelo, M. A. (2019, October). Experimental validation of lane-change intention prediction methodologies based on CNN and LSTM. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)* (pp. 3657-3662). IEEE.

[7] Izquierdo, R., Quintanar, Á., Lorenzo, J., García-Daza, I., Parra, I., Fernández-Llorca, D., & Sotelo, M. Á. (2021). Vehicle lane change prediction on highways using efficient environment representation and deep learning. *IEEE Access*, *9*, 119454-119465.

[8] Konakalla, N., Noor, A., & Singh, J. (2022). Cnn, cnn encoder-rnn decoder, and pretrained vision transformers for surrounding vehicle lane change classification at future time steps.

[9] Biparva, M., Fernández-Llorca, D., Gonzalo, R. I., & Tsotsos, J. K. (2022). Video action recognition for lane-change classification and prediction of surrounding vehicles. *IEEE Transactions on Intelligent Vehicles*, *7*(3), 569-578.

[10] Liang, K., Wang, J., & Bhalerao, A. (2022, October). Lane Change Classification and Prediction with Action Recognition Networks. In *European Conference on Computer Vision* (pp. 617-632). Cham: Springer Nature Switzerland.

[11] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (pp. 1725-1732).

[12] Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, *27*.

[13] Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 6299-6308).

[14] Feichtenhofer, C., Pinz, A., & Wildes, R. P. (2017). Spatiotemporal multiplier networks for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4768-4777).

[15] Feichtenhofer, C., Fan, H., Malik, J., & He, K. (2019). Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 6202-6211).

[16] Krajewski, R., Bock, J., Kloeker, L., & Eckstein, L. (2018, November). The highd dataset: A drone dataset of naturalistic vehicle trajectories on german highways for validation of highly automated driving systems. In *2018 21st international conference on intelligent transportation systems (ITSC)* (pp. 2118-2125). IEEE.

[17] Huang, X., Cheng, X., Geng, Q., Cao, B., Zhou, D., Wang, P., ... & Yang, R. (2018). The apolloscape dataset for autonomous driving. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 954-960).

[18] Izquierdo, R., Quintanar, A., Parra, I., Fernández-Llorca, D., & Sotelo, M. A. (2019, October). The prevention dataset: a novel benchmark for prediction of vehicles intentions. In *2019 IEEE intelligent transportation systems conference (ITSC)* (pp. 3114-3121). IEEE.

[19] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

[20] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

[21] Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., & Schmid, C. (2021). Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 6836-6846).

[22] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.