End-to-End Spatio-Temporal Attention-Based Lane-Change Intention Prediction from Multi-Perspective Cameras

Zhouqiao Zhao, Zhensong Wei, Danyang Tian, Bryan Reimer, Pnina Gershon, and Ehsan Moradi-Pari

2023 IEEE Intelligent Vehicles Symposium (IV) | 979-8-3503-4691-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/IV55152.2023.10186602

Abstract— **Advanced Driver Assistance Systems (ADAS) with proactive alerts have been used to increase driving safety. Such systems' performance greatly depends on how accurately and quickly the risky situations and maneuvers are detected. Existing ADAS provide warnings based on the vehicle's operational status, detection of environments, and the drivers' overt actions (e.g., using turn signals or steering wheels), which may not give drivers as much as optimal time to react. In this paper, we proposed a spatio-temporal attention-based neural network to predict drivers' lane-change intention by fusing the videos from both in-cabin and forward perspectives. The Convolutional Neural Network (CNN)-Recursive Neural Network (RNN) network architecture was leveraged to extract both the spatial and temporal information. On top of this network backbone structure, the feature maps from different time steps and perspectives were fused using multi-head self-attention at each resolution of the CNN. The proposed model was trained and evaluated using a processed subset of the MIT Advanced Vehicle Technology (MIT-AVT) dataset which contains synchronized CAN data, 11058-second videos from 3 different views, 548 lane-change events, and 274 non-lane-change events performed by 83 drivers. The results demonstrate that the model achieves 87% F1-score within the 1-second validation window and 70% F1-score within the 5-second validation window with real-time performance.**

I. Introduction

Improper lane changing is one of the top causes of accidents in the transportation system. Common examples of improper lane changing include changing lanes without utilizing the blinker or thoroughly inspecting the surrounding traffic, particularly the blind spots for vehicles, bicycles, or pedestrians in an adjacent lane. Such actions leave the surrounding road users little time to respond, and accidents were more likely to happen. According to the National Highway Traffic Safety Administration's (NHTSA) General Estimates System (GES) collision database [1], 539,000 crashes (9% of all crashes) involving 1,078,000 cars were caused by lane changes in the US in 1999. Among them, 14% of the lane change crashes led to some form of injury, and major injuries occurred much more frequently on dark unlit portions of interstates [2]. To improve lane-change safety, numerous ADAS have been deployed to alert drivers when a potential conflict is detected during or even before changing lanes. This helps drivers better understand the surrounding environment and thus make better decisions [3]. However, ADAS equipped in existing commercial vehicles only provide warnings based on vehicle's operational status. Systems lack of environmental context, and drivers' overt actions (e.g., using turn signals or steering wheels) can lead to late warnings and may not provide drivers sufficient time to react.

With the development of Connected and Autonomous Vehicle (CAV) technologies, vehicles equipped with sensors and communication devices are capable of perceiving, understanding, and sharing driving scenarios, which makes the early prediction of lane-change maneuvers possible [4]. The most common input for maneuver prediction is the past trajectory of the host vehicle. From this, speed, acceleration, and position can be extracted and fed into the prediction system. In addition to the host vehicle's trajectory, its relation with surrounding vehicles is also commonly used as supplementary input to enhance prediction accuracy. Such methods usually omit drivers' behaviors that are hypothesized to enhance predictions, such as mirror checking, and could potentially delay the lane-change maneuver prediction time. Therefore, in this paper, we present an innovative method to model drivers' lane-change intention from multi-perspective cameras that include both drivers' views and forward view.

In summary, the contributions of this paper are listed below:

- We implemented a semi-automatic method to annotate the start and finish time of 548 lane-change events and 274 non-lane-change events performed by 83 drivers from the MIT-AVT Dataset. The processed lane-change dataset is drawn from high-definition video streams from forward-facing view, driver's face view, and driver's body view, which were synchronized with the vehicles' Controller Area Network (CAN) messages at 30Hz.
- We proposed a CNN-RNN-based neural network with self-attention-based multi-modal fusion architecture to extract spatio-temporal features from multiple videos and CAN data.
- Empirical analysis and evaluations show that the proposed model has reached state-of-the-art performance. The present drawbacks of end-to-end lane change prediction models were investigated and the relation be-

Z. Zhao and Z. Wei are with the Department of Electrical and Computer Engineering, and the Center for Environmental Research and Technology, University of California, Riverside, CA 92507. (e-mail: zzhao084@ucr.edu; zwei030@ucr.edu)

D. Tian (corresponding author) and E. Moradi-Pari are with Honda Research Institute US, Ann Arbor, MI 48103. (e-mail:danyang_tian@honda-ri.com; emoradipari@hondari.com)

B. Reimer and P. Gershon are with the Massachusetts Institute of Technology, Cambridge, MA 02142. (e-mail:reimer@mit.edu; pgershon@mit.edu)

tween pre-lane-change data and lane-change intention of drivers was analyzed.

The remainder of this paper is as follows: Section II introduces the related existing work on lane-change prediction. Section III elaborates on the used lane change dataset preprocessing and summary, followed by the problem formulation and the proposed methodology for driver's lane-change intention prediction. Section V shows the sensitivity evaluation results of the proposed model in different scenarios. Finally, conclusion and future directions are presented in Section VI.

II. Related Work

The research on lane-change prediction has been explored for decades [5]. With a large number of open-source trajectory datasets [6], [7], numerous studies have been conducted using trajectory prediction for lane-change intention recognition [8]–[11]. Kou et al. and Girasc et al. [8], [9] used only the past trajectory to predict the vehicle's future trajectory and maneuvers. Zhang et al. [10] considered the speed and position relationship between the target vehicle and surrounding vehicles in addition to the trajectory data and used the Dempster–Shafer (D-S) evidence theory to calculate the lane-change probability. Liao et al. [11] applied inverse reinforcement learning (IRL) to calculate the probability of all possible lane-change trajectories.

With the development of computer vision and machine learning technologies, more research has been turning to the use of vision sensors to forecast lane changes. For example, Wei et al. [12] used only single-frame forward-view images to predict lane changes. However, because temporal information was not utilized, the prediction performance was unsatisfactory. Video action recognition is one of the representative tasks for video understanding, which has been used for behavior prediction. The state-of-the-art approaches for video-based action recognition include CNN-RNN networks [13], two-stream networks [14], temporal segment networks [15], I3D [16], Non-local [17] and SlowFast [18]. Recently, Biparva et al. [19] applied four different video action recognition methods for lane-change classification and prediction of surrounding vehicles. The implemented models achieved good prediction performance in terms of both accuracy and Time-To-Maneuver (TTM) with only the forward-view camera in the PREVENTION dataset [20]. To further improve the performance, in-cabin videos have been applied to leverage the monitoring of drivers' behaviors. As a representative example of in-cabin driving monitoring dataset, Brain4Cars [21] has been well explored by different researchers [22]–[24]. For example, Jain et al. argued that, by using the in-cabin camera, they can anticipate driving maneuvers up to 3.5 seconds before they occurred [21], which outperforms trajectory-based prediction methods (e.g., 1.1 seconds in advance for lane-change maneuvers [25]). However, research on maneuver/intention prediction using in-cabin video steams is still at an early stage. The existing lane-change datasets with in-cabin videos have a limited number of lane-change events and inadequate annotating quality. In addition, the current studies about vision-based lane-change anticipation heavily rely on the existing video action recognition approaches, while cutting-edge neural network architectures such as self-attention mechanisms have not been well investigated.

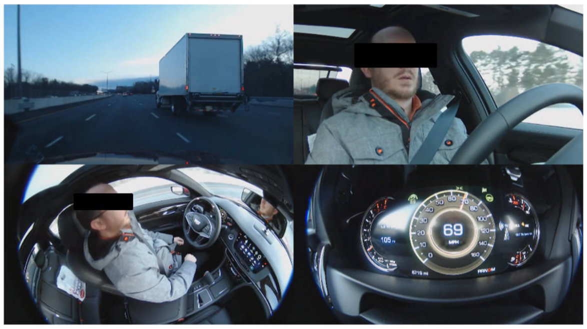

Fig. 1. Different video views of MIT-AVT Dataset.

III. Dataset Preparation

A. MIT Advanced Vehicle Technology (MIT-AVT) Dataset

A subset of the MIT-AVT datasets [26] were used to search lane-change events and prepared for model training that is described in Section IV. The MIT-AVT dataset includes driving data collected through various sensors from different models vehicles in both long-term (over a year per driver) and medium-term (one month per driver) studies. The recorded data streams include Inertial Measurement Unit (IMU), Global Positioning System (GPS), CAN messages, and high-definition video streams of driver's face, driver's cabin, forward roadway, and instrument cluster (on selected vehicles).

The MIT-AVT dataset has been used to explore driver-automation interaction [27]–[30], etc., at the same time, the MIT-AVT dataset contains a large number of pure human-driving or longitudinal automation-only driving periods, which are suitable for understanding drivers' lane-change behavior. In this paper, we used high-definition video streams and CAN messages for lane-change intention prediction. As shown in Fig. 1, the top left view is referred as the front view, the top right view is referred as the face view, and the bottom left view is referred as the body view. The face view depicts the driver's facial expressions, head attitude, and head angle, while the body view captures the body positions and gestures. The front view provides context information about the area in front of the ego vehicle. The information of steering angles, speeds, and automation levels is coming from the CAN messages.

B. Lane Mark Detection

To use the MIT-AVT dataset for lane-change intention prediction, we selected lane-change events based on the front-view video streams. 200 human-driving or longitudinal-automation-only driving trips with highway driving were selected from the dataset, with an average of six high-speed

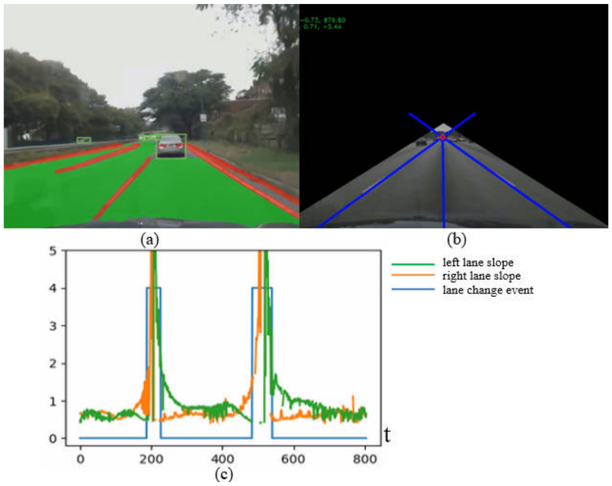

Fig. 2. Lane change detection process.

lane-change events in every 30-minute trip. This amount of data makes manual annotation an extremely time-consuming task. To solve this issue, a machine-learning-based semi-automatic method was implemented to identify the coarse start and finish time of the lane-change events. The precise lane-change event time was then determined manually. The network implemented for selecting the lane-change events is called *You Only Look Once for Panoptic* (YOLOP) [31], which is a state-of-the-art perception method with high accuracy and robustness to perform traffic object detection, drivable area segmentation, and lane detection simultaneously (see Fig. 2 (a)). For the task of recognizing lane-change events, only the output of lane detection is needed. Our experiments show that the lane detection of YOLOP works effectively even when lighting is low or lane markings are barely visible.

C. Lane-Change Events Annotation and CAN Data Equalization

The detected lanes from the YOLOP network are represented in the form of a binary mask, and the tentative lane mark pixels have the value of one. We used polynomial approximation to calculate the left and right lanes, as shown in Fig. 2 (b). During a lane-change event, the slope of one side of the lane will initially increase to infinity, then be miss-detected, and finally be re-detected with a low slope value. The slope of the opposite side of the lane will initially decrease, then be miss-detected, and finally be re-detected going from a high slope value to a low slope value. Based on this feature, the lane-change events and their coarse start and finish time can be labeled automatically, as shown in Fig. 2 (c). The parameters were tuned based on manually labeled sample lane-change events. It should be noted that the parameters were tuned to be sensitive to slope change and missed detection of lane marks so that all the suspicious events could be selected as candidate events. As a result, the false positive rate is also high, therefore, it's necessary to delete those false positive events through manual video review. This can be done efficiently using VidStep [32] with high playback speed. VidStep is a frame-by-frame in-browser video player and annotator, which was originally designed to label driver glances. It offers the ability to play videos at various playback speeds and keyboard shortcuts for frame control and annotation.

The definition of a lane-change start and finish time is contestant across published studies. For example, Scheel et al. [25] used a 3-second criterion, before the surrounding target vehicle's lane assignment changes in simulation, to label a lane change. Wu et al. [33] defined the start-point and end-point by calculating the heading angle θ of the vehicle. As a result, the performances of TTM were not comparable. In this paper, we defined the beginning and end of lateral displacement (i.e., pixel movement perpendicular to the road direction) as the start and finish points of lane-change events. The lane deviation should not be used as the criteria because some drivers do not keep in the center of the lane long before and after the lane change.

The two CAN messages used for lane-change intention prediction are steering angle and vehicle speed. However, it's worth mentioning that the scale of steering angle data varies with vehicle models. Additionally, even for the same vehicle, the scale of steering angle data can vary due to tire wear. Scalars were used to equalize the steering angle distributions for trip data collected by different vehicles.

D. Lane-Change Dataset Summary

Since the drivers may behave differently when conducting consecutive lane changes, we eliminated the lane-change events that happened close in time (i.e., within 10 seconds). The non-lane-change events were randomly selected within the period that was far away (15 seconds) from the selected lane-change events. To make sure that the selected non-lane-change samples were not affected by the lane-change maneuvers, we selected those that were 5 seconds before or after the lane-change events. The statistics of the lane-change dataset are shown in Fig. 3. The average lane-change duration is 6 seconds, and the lane-change speed ranges from 37 mph to 85 mph. Each sample event was trimmed starting from 5 seconds before to 2 seconds after the lane change for data balancing purpose. The lane-change dataset of 822 selected events from 83 drivers includes 261 left lane changes, 287 right lane changes, and 274 staying in the same lane. This processed dataset is a large lane-change-event dataset with a variety of drivers comprising both in-cabin and forward-facing video streams, and CAN messages.

IV. Driver Intention Modeling

Although drivers' lane-change intentions are implicit, we hypothesize they would give hints by making expressions that would reveal their intentions. For instance, a driver might shift his/her body toward the direction of a potential lane change and check mirrors and blind spots. Alternatively, if a driver tries to pass a slow car ahead, such contextual information can

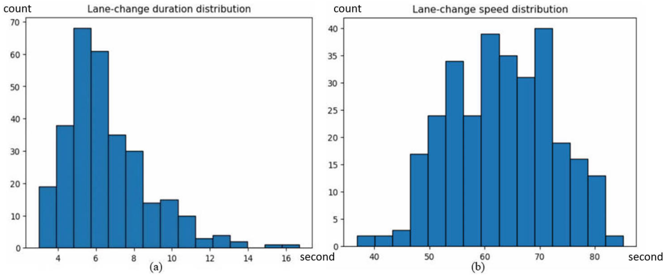

Fig. 3. Duration and speed distribution of lane change events.

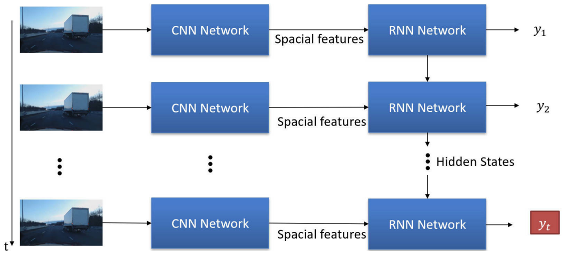

Fig. 4. CNN-RNN backbone.

be an indicator of incoming lane-change maneuvers. As such, the spatio-temporal features of the pre-lane-change behaviors may contain information about drivers' lane-change intentions and aid in the early prediction of lane-change maneuvers. In this paper, we proposed and trained a model to predict the driver's lane-change intentions by using the pre-lane-change videos, synchronized CAN data, and the corresponding future maneuver tags. The problem can be then formulated as a multi-classification problem, with the three classes being left lane change, right lane change, and staying in the same lane (or going straight), represented by L, R, and S.

A. Backbone Network

As reviewed in Section II, CNN-RNN architecture is a simple yet efficient approach to extracting spatio-temporal features from videos, which is often used for video classification. The goal of CNN models is to extract the high-level spatial information of each frame, and the goal of RNN models is to extract the temporal correlation between the images by keeping a memory of past images. The features are then fed into fully connected layers, also known as multilayer perceptron (MLP), to get the classification output. As shown in Fig. 4, the backbone of the proposed model follows the CNN-RNN scheme, where we used ResNet18 [34] as the CNN model and GRU [35] as the RNN model. The classification output head was used to predict the future maneuver which falls in one of the three classes that belong to $\mathbb{R}^3$.

B. Attention Mechanism

Inspired by [36], the self-attention mechanism of transformers [37] was implemented in addition to the original CNN-RNN architecture, to capture and incorporate the context information from different video perspectives and time steps. Following the prior studies on the image-recognition application of transformers [38], the input $x \in \mathbb{R}^{(C \times d_x)}$ to the transformer module is the embedding of the grid-structured feature maps. C represents the number of tokens (i.e., the encoded sequence of integers) and d_x is the dimension of the feature vectors in each token. In this study, this embedded token is concatenated from the CNN features of different perspectives and time steps. Then, the transformer blocks apply linear projections for calculating a set of queries, keys, and values, denoted as Q, K, and V, as shown below:

$$Q = xW^Q, K = xW^K, V = xW^V \tag{1}$$

where $W^Q \in \mathbb{R}^{(d_x \times d_Q)}$,$W^K \in \mathbb{R}^{(d_x \times d_K)}$, and $W^V \in \mathbb{R}^{(d_x \times d_V)}$ are parameter matrices. Then the attention is calculated using the dot products between Q, K, and V, as shown below:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2}$$

Finally, the calculated attention will go through MLP to form the output y, which has the same size as input x.

$$y = MLP(Attention(Q, K, V)) + x \tag{3}$$

C. Proposed Network

The architecture of the proposed network is shown in Fig. 5. The inputs are videos from the three different views and one-dimensional CAN messages (i.e., vehicle speed and steering angle). The frequency of the collected and synchronized videos and CAN messages is 10Hz. The frames from different views and different time steps were resized into 256 by 256, to save computational time and power, and sent into the ResNet separately. The one-dimensional CAN messages were fused in the Transformer blocks of this network.

As shown in Fig. 5, the feature maps from different views and different time steps are fused multiple times at different scales (i.e., after each ResNet layer). The self-attention transformers require the input to be a two-dimensional token structure. Therefore, the feature maps from the ResNet are first reshaped and concatenated to the required dimension before fed into the transformers. More specifically, let a feature map of a single view and single time step be a 3D tensor of dimension $H \times W \times C$, where H and W are the height and width and C is the number of channels. All the features are reshaped and stacked together to form a sequence of dimension $(V \times T \times H \times W) \times C$ (illustrated by the green blocks), where V is the number of views and T is the number of frames. Then, a learnable positional embedding feature (illustrated by the light blue block) reshaped to the same size is added to the input feature map so that the network can infer spatial-temporal dependencies between different tokens while training. Also, the CAN messages are broadcasted and copied to the size of positional embedding denoting velocity embedding and steering embedding (illustrated by the orange and blue blocks). Finally, the transformer output is reshaped back to the size of its input and is fed into the following

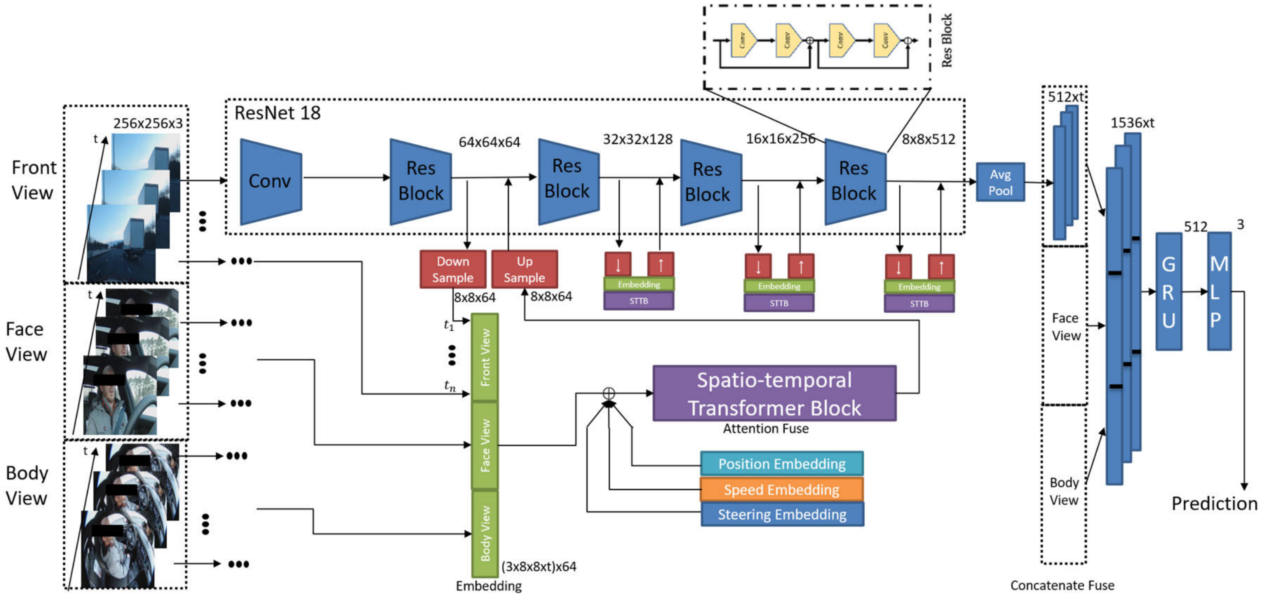

Fig. 5. Proposed spatio-temporal attention-based network.

Resnet residual blocks. Since the same structure is applied multiple times at each down-sampling level along every residual block, it is computationally expensive to operate all the above calculations on the original-sized feature maps. Therefore, the transformer input is first down-sampled to $H = W = 8$, then the output is resized back and added to the input to form a residual connection (illustrated by the red blocks).

After the CNN feature extraction and dense feature fusion via spatio-temporal transformers, the output feature map becomes $8 \times 8 \times 512$ for each view and each time step. We reduced their dimension to 512 by average pooling and fused them by concatenation. The size of the fused feature is $(512 \times V) \times T$. This fused feature vector is a condensed environment representation that captures the overall spatio-temporal context of the pre-lane-change and can be used for lane-change intention prediction. The fused feature vector is then sent to the GRU network to further extract the temporal relationship between frames. The GRU has one layer and the number of features in the hidden state is 512. As a result, the network reduces the size of the feature from $(512 \times V) \times T$ to 512×1. Finally, the MLP network (i.e, output head) is a series of fully connected layers to reduce the number of channels to 3 at the last layer, which consists of 3 hidden layers with 256, 128, and 64 units.

V. Experiments and Results

A. Assumptions and Specifications

The implemented CNN backbone is a pre-trained ResNet18. The transformer block at each resolution has eight layers and each layer has four attention heads. As suggested by [36], we used the AdamW optimizer, which is a variant of Adam. Weight decay is set to 0.01, and Adam beta values to the PyTorch defaults of 0.9 and 0.999. In addition, the loss function is the classic cross-entropy loss for the classification task. We trained the models with 2 RTX A6000 GPUs. We randomly split the entire dataset into a training subset of 70% and a validation subset of 30%.For model training and evaluation, the following periods of lane-change events were defined:

- Pre-Lane-Change Period: The Pre-Lane-Change Period refers to the period that starts five seconds before lane changes.
- Observation Window: The observation window was defined as a sliding window that contains the time-series data before the predicted time point. The wider the observation window is, the more historical information is utilized for prediction.
- Ground Truth (GT) Time: As the drivers' intentions are implicit, it was assumed that the pre-lane-change periods contain information about drivers' lane change intentions. The Lane Change Intention "GT" is a subset of the pre-lane-change period, in which they were labeled as L and R for training and validation. Regarding the non-lane-change events, the whole periods were labeled as S. The Lane Change Intention "GT" period ends at the start time of the lane-change events and its start time was defined as GT time. GT time is a tunable hyperparameter.
- Validation Time: The Validation Period is also a subset of the pre-lane-change period. The observation windows slide along it while validating the model's performance. Regarding non-lane-change events, the whole periods were used for validation. The Validation Period ends at the start time of the lane-change events and its start time was defined as Validation Time. Validation Time is a parameter to be decided for different experiments.

TABLE I

Results of Experiment 1: STACNN-GRU with single video stream

	Face	Body	Front
Acc	**72.7%**	69.0%	58.9%

B. Metrics

For the prediction measurement, Accuracy (Acc), Precision (Pr), Recall (Re), and F1 score were used to evaluate the performance of the intention prediction model. It should be noted that the maneuvers herein only refer to the class L and R. Following [21], the definitions are shown in Equation (4)-(7).

$$Acc = \frac{tp + tsp}{tp + fp + fpp + tsp + mp} \tag{4}$$

$$Pr = \frac{tp}{tp + fp + fpp} \tag{5}$$

$$Re = \frac{tp}{tp + fp + mp} \tag{6}$$

$$F1 = 2 \times \frac{Pr \times Re}{Pr + Re} \tag{7}$$

where true predictions (tp) represent correct maneuver predictions, true straight predictions (tsp) represent correct predictions when there's no lane change, false predictions (fp) represent false maneuver predictions, false positive predictions (fpp) represent the predictions of maneuvers, but actually driving straight, and missed predictions (mp) represent the prediction of driving straight but driver performs a maneuver.

C. Lane-Change Intention Prediction Results

This section presents the experiments that were conducted to validate the proposed network, which was named as Spatio-Temporal Attention-based CNN-GRU (STACNN-GRU).

1) Experiment 1: STACNN-GRU with single video stream: To investigate the effects of different views on lane-change intention prediction, the network was trained using only a single video streaming in the first experiment. Additionally, the CAN messages were not integrated into the network. The observation window, the GT time, and the validation time were set up as 1 second, 1 second, and 3 seconds, respectively. The results are shown in TABLE I. As can be seen from the results, the prediction using the face view reaches the highest accuracy, 72.7%, which is 3.7% higher than the body view, and 13.8% higher than the front view. This result indicates that the face view has the most spatio-temporal information, such as the head movement for mirror-checking, representing the lane-change intention at the early stage; the spatio-temporal information from the body view, such as the arm movement for using turn signals, can also help to infer the lane-change intention at a lower level of accuracy; same as the expectation, the front view showing the forward roadway plays the least important role for the early stage lane-change prediction.

TABLE II

Results of Experiment 2: STACNN-GRU VS CNN-GRU

Acc	GT Time 1s	GT Time 2s	GT Time 3s
CNN-GRU	79.0%	80.7%	75.4%
STACNN-GRU (proposed)	**85.5%**	**83.1%**	**78.4%**

TABLE III

Results of Experiment 3: STACNN-GRU with different sizes of observation window

Acc	GT Time 2s	GT Time 3s	GT Time 4s
Observation Window: 1s	79.3%	72.1%	64.9%
Observation Window: 1.5s	**83.1%**	**78.4%**	**70.2%**

2) Experiment 2: STACNN-GRU VS CNN-GRU: To prove the effectiveness of the self-attention mechanism, the performances between the proposed network and the CNN-GRU network were compared in the second experiment. The only difference between those two is that the baseline CNN-GRU network does not have attention-based fusion blocks at different scales. The observation window was set up as 1.5 seconds, and the validation time was set up as 3 seconds. The GT time varies from 1 second to 3 seconds. The results are shown in TABLE II. For three different GT times, the proposed STACNN-GRU network achieves better accuracy compared to the baseline by 6.5%, 3.6%, and 3.0%. It should also be noted that the STACNN-GRU network can run on a single A6000 GPU at 8.38 frame-per-second (fps), which is suitable for real-time application.

3) Experiment 3: STACNN-GRU with different sizes of observation window: As the observation window presents historical information to the network for prediction, intuitively, the wider observation window results in better performance. This hypothesis was validated by the third experiment as shown in TABLE III. The model was trained and validated with a fixed validation time, i.e., 3 seconds, the observation window was varied from 1 second to 1.5 seconds, and the GT time was varied from 2 seconds to 4 seconds. The reasons why we can not further extend the observation window are threefold. First, for early-stage prediction, larger observation windows may also include unrelated information, which could reduce the model's accuracy. Second, a wider observation time corresponds to a larger transformer network and more feature tokens in the time dimension. Because the attention network is data-hungry, the current size of the lane-change dataset can not ensure that the larger network will learn useful features. Third, the larger attention network increases the computing load, which makes it challenging for real-world implementation.

4) Experiment 4: STACNN-GRU with different GT Times and Validation Times: In the last experiment, the observation window was fixed to 1 second while the network was trained with various GT times, ranging from 1 second to 4 seconds.

TABLE IV

Results of Experiment 4: STACNN-GRU with different GT Times and Validation Times

	Validation Time 1s	Validation Time 2s	Validation Time 3s	Validation Time 4s	Validation Time 5s
GT Time 1s	**Pr: 0.86** **Re: 0.88** **F1:0.87**	**Pr: 0.81** **Re: 0.80** **F1:0.80**	**Pr: 0.81** **Re: 0.80** **F1:0.80**	Pr: 0.76 Re: 0.67 F1:0.71	Pr: 0.73 Re: 0.61 F1:0.66
GT Time 2s	Pr: 0.74 Re: 0.76 F1:0.75	Pr: 0.74 Re: 0.76 F1:0.74	Pr: 0.77 Re: 0.82 F1:0.79	**Pr: 0.74** **Re: 0.76** **F1:0.75**	**Pr: 0.71** **Re: 0.70** **F1:0.70**
GT Time 3s	Pr: 0.74 Re: 0.84 F1:0.79	Pr: 0.74 Re: 0.83 F1:0.78	Pr: 0.71 Re: 0.76 F1:0.73	Pr: 0.69 Re: 0.70 F1:0.70	Pr: 0.68 Re: 0.64 F1:0.66
GT Time 4s	Pr: 0.73 Re: 0.87 F1:0.80	Pr: 0.70 Re: 0.83 F1:0.76	Pr: 0.77 Re: 0.79 F1:0.78	Pr: 0.65 Re: 0.79 F1:0.70	Pr: 0.63 Re: 0.74 F1:0.68

The models were validated by different validation times ranging from 1 second to 5 seconds. Note that the GT intention was not labeled based on the actual drivers' intention. As a substitute, we assumed that the explicit expression of the implicit intention lies in the pre-lane-change behaviors, as such this experiment was designed to explore how the explicit expression is distributed along the pre-lane-change data. The results are shown in TABLE IV. We observed that the model performs best for validation times ranging from 1 second to 3 seconds, even if we only labeled the closest 1 second to the lane change as the GT intention. For VT of 4 or 5s, the model performs best when trained with a GT of 2s. This mismatch of the best-model corresponding GT time and validation time indicates that model quality relies on the density of the lane-change spatio-temporal information in the pre-lane-change period. Because we labeled the whole GT time with lane change intention, the higher density of the lane-change spatio-temporal information translates into better GT labeling quality. The lane-change-intention-related maneuvers were more concentrated right before (i.e., last two seconds) lane changes and more sparse when relatively far away from lane changes in time.

VI. Conclusion and Future Work

In this paper, a semi-automatic technique was first implemented to create a lane-change dataset based on the MIT-AVT full-trip data. The processed dataset includes synchronized CAN data, 11058-second videos from 3 different views, 548 lane-change events, and 274 non-lane-change events. Second, a CNN-RNN network was implemented for drivers' real-time lane-change intention prediction. The CNN network is the well-established ResNet18 with pre-trained parameters, and the RNN network is a Gated Recurrent Unit. More importantly, the spatio-temporal information was fused at multiple scales of the CNN network using the transformers with a spatio-temporal self-attention mechanism.

The results show that the proposed model can efficiently learn and predict drivers' lane-change intention achieving 87% F1-score within the 1-second validation window and 70% F1-score within the 5-second validation window. Also, it was demonstrated that the face view plays the most important role in representing the driver's lane-change intention. The ablation study results show that the attention mechanism helps to improve prediction accuracy. Larger observation windows led to better model performance, since more spatio-temporal features can be captured by the transformer blocks, while increasing the computing load due to the considerably growing size of the transformer network. Interestingly, we observed that the spatio-temporal information, representing the lane-change intention, does not spread evenly along the pre-lane-change period. The results indicate that the model achieves the best prediction performance by training it using the closest 1 second of the pre-lane-change period as the GT intention. This can be explained by the distribution of pre-lane-change behaviors, which were more concentrated when closer in time before the lane changes.

Future work includes more evaluations of the model on other transformer architectures and state-of-the-art video detection networks such as I3D and SlowFast. Moreover, instead of using end-to-end networks, it is worth modeling the driver's behaviors in a higher resolution. For example, Markov Decision Process (MDP) can be used to depict and model the movement of the driver. To realize this, the annotation of detailed driver behaviors is necessary, such as mirror-checking behaviors. The actual drivers' intention data for annotation as the ground truth is helpful for model training accuracy as well. Third, it is essential to annotate the lane-change events with more detailed contextual information, which may have impacts on the driver's lane-change behaviors and thus is useful for behavior anticipation.

VII. Acknowledgment

We are grateful to the Advanced Vehicle Technology Consortium, Massachusetts Institute of Technology, for providing the dataset, and thanks to Sue Bai and Doug Longhitano for the coordination that made this research activity happen.

References

[1] B. Sen, J. D. Smith, W. G. Najm *et al.*, "Analysis of lane change crashes," United States. National Highway Traffic Safety Administration, Tech. Rep., 2003.

[2] J. C. McCall, D. P. Wipf, M. M. Trivedi, and B. D. Rao, “Lane change intent analysis using robust operators and sparse bayesian learning,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 8, no. 3, pp. 431–440, 2007.

[3] V. A. Butakov and P. Ioannou, “Personalized driver/vehicle lane change models for adas,” *IEEE Transactions on Vehicular Technology*, vol. 64, no. 10, pp. 4422–4431, 2014.

[4] Z. Zhao, G. Wu, Z. Wang, and M. J. Barth, “Optimal control-based eco-ramp merging system for connected and automated vehicles,” in *2020 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2020, pp. 540–546.

[5] P. Kumar, M. Perrollaz, S. Lefevre, and C. Laugier, “Learning-based approach for online lane change intention prediction,” in *2013 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2013, pp. 797–802.

[6] R. Krajewski, J. Bock, L. Kloeker, and L. Eckstein, “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)*. IEEE, 2018, pp. 2118–2125.

[7] J. Colyar and J. Halkias, “Ngsim - us highway 101 dataset,” https://www.fhwa.dot.gov/publications/research/operations/07030/, 2007.

[8] S. Kou, K. Jiang, W. Yu, R. Yan, W. Zhou, M. Yang, and D. Yang, “Lane change intention recognition for intelligent connected vehicle using trajectory prediction,” in *20th COTA International Conference of Transportation ProfessionalsChinese Overseas Transportation Association (COTA) American Society of Civil Engineers*, 2020.

[9] H. Girase, H. Gang, S. Malla, J. Li, A. Kanehara, K. Mangalam, and C. Choi, “Loki: Long term and key intentions for trajectory prediction,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9803–9812.

[10] B. Zhang, Z. Ding, and M. Zhou, “A lane change prediction algorithm based on probabilistic modeling,” in *Proceedings of the 2020 International Conference on Computers, Information Processing and Advanced Education*, 2020, pp. 211–216.

[11] X. Liao, Z. Wang, X. Zhao, Z. Zhao, K. Han, P. Tiwari, M. Barth, and G. Wu, “Online prediction of lane change with a hierarchical learning-based approach,” in *Proceedings 2022 IEEE International Conference on Robotics and Automation. IEEE*, 2022.

[12] Z. Wei, C. Wang, P. Hao, and M. J. Barth, “Vision-based lane-changing behavior detection using deep residual neural network,” in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2019, pp. 3108–3113.

[13] Y. Fan, X. Lu, D. Li, and Y. Liu, “Video-based emotion recognition using cnn-rnn and c3d hybrid networks,” in *Proceedings of the 18th ACM international conference on multimodal interaction*, 2016, pp. 445–450.

[14] K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” *Advances in neural information processing systems*, vol. 27, 2014.

[15] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, “Temporal segment networks: Towards good practices for deep action recognition,” in *European conference on computer vision*. Springer, 2016, pp. 20–36.

[16] J. Carreira and A. Zisserman, “Quo vadis, action recognition? a new model and the kinetics dataset,” in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[17] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[18] C. Feichtenhofer, H. Fan, J. Malik, and K. He, “Slowfast networks for video recognition,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6202–6211.

[19] M. Biparva, D. Fernández-Llorca, R. I. Gonzalo, and J. Tsotsos, “Video action recognition for lane-change classification and prediction of surrounding vehicles,” *IEEE Transactions on Intelligent Vehicles*, 2022.

[20] R. Izquierdo, A. Quintanar, I. Parra, D. Fernández-Llorca, and M. Sotelo, “The prevention dataset: a novel benchmark for prediction of vehicles intentions,” in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2019, pp. 3114–3121.

[21] A. Jain, H. S. Koppula, B. Raghavan, S. Soh, and A. Saxena, “Car that knows before you do: Anticipating maneuvers via learning temporal driving models,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3182–3190.

[22] Y. Rong, Z. Akata, and E. Kasneci, “Driver intention anticipation based on in-cabin and driving scene monitoring,” in *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2020, pp. 1–8.

[23] A. Jain, A. Singh, H. S. Koppula, S. Soh, and A. Saxena, “Recurrent neural networks for driver activity anticipation via sensory-fusion architecture,” in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 3118–3125.

[24] P. Gebert, A. Roitberg, M. Haurilet, and R. Stiefelhagen, “End-to-end prediction of driver intention using 3d convolutional neural networks,” in *2019 IEEE Intelligent vehicles symposium (IV)*. IEEE, 2019, pp. 969–974.

[25] O. Scheel, N. S. Nagaraja, L. Schwarz, N. Navab, and F. Tombari, “Attention-based lane change prediction,” in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 8655–8661.

[26] L. Fridman, D. E. Brown, M. Glazer, W. Angell, S. Dodd, B. Jenik, J. Terwilliger, A. Patsekin, J. Kindelsberger, L. Ding *et al.*, “Mit advanced vehicle technology study: Large-scale naturalistic driving study of driver behavior and interaction with automation,” *IEEE Access*, vol. 7, pp. 102 021–102 038, 2019.

[27] M. Q. Khan and S. Lee, “Gaze and eye tracking: Techniques and applications in adas,” *Sensors*, vol. 19, no. 24, p. 5540, 2019.

[28] J. Orlovska, F. Novakazi, B. Lars-Ola, M. Karlsson, C. Wickman, and R. Söderberg, “Effects of the driving context on the usage of automated driver assistance systems (adas)-naturalistic driving study for adas evaluation,” *Transportation research interdisciplinary perspectives*, vol. 4, p. 100093, 2020.

[29] P. Gershon, S. Seaman, B. Mehler, B. Reimer, and J. Coughlin, “Driver behavior and the use of automation in real-world driving,” *Accident Analysis & Prevention*, vol. 158, p. 106217, 2021.

[30] A. Morando, P. Gershon, B. Mehler, and B. Reimer, “A model for naturalistic glance behavior around tesla autopilot disengagements,” *Accident Analysis & Prevention*, vol. 161, p. 106348, 2021.

[31] D. Wu, M.-W. Liao, W.-T. Zhang, X.-G. Wang, X. Bai, W.-Q. Cheng, and W.-Y. Liu, “Yolop: You only look once for panoptic driving perception,” *Machine Intelligence Research*, vol. 19, no. 6, pp. 550–562, 2022.

[32] “Vidstep: Frame-by-frame in-browser video player and annotator,” https://vidstep.com/, accessed: 2023-01-19.

[33] Z. Wu, K. Liang, D. Liu, and Z. Zhao, “Driver lane change intention recognition based on attention enhanced residual-mbi-lstm network,” *IEEE Access*, 2022.

[34] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[35] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” *arXiv preprint arXiv:1412.3555*, 2014.

[36] A. Prakash, K. Chitta, and A. Geiger, “Multi-modal fusion transformer for end-to-end autonomous driving,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7077–7087.

[37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.

[38] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.