A Cognitive-Based Trajectory Prediction Approach for Autonomous Driving

Haicheng Liao*, Yongkang Li*, Zhenning Li†, Chengyue Wang, Zhiyong Cui, Shengbo Eben Li, *Senior Member, IEEE*, and Chengzhong Xu†, *Fellow, IEEE*

***Abstract*—In autonomous vehicle (AV) technology, the ability to accurately predict the movements of surrounding vehicles is paramount for ensuring safety and operational efficiency. Incorporating human decision-making insights enables AVs to more effectively anticipate the potential actions of other vehicles, significantly improving prediction accuracy and responsiveness in dynamic environments. This paper introduces the Human-Like Trajectory Prediction (HLTP) model, which adopts a teacher-student knowledge distillation framework inspired by human cognitive processes. The HLTP model incorporates a sophisticated teacher-student knowledge distillation framework. The "teacher" model, equipped with an adaptive visual sector, mimics the visual processing of the human brain, particularly the functions of the occipital and temporal lobes. The "student" model focuses on real-time interaction and decision-making, drawing parallels to prefrontal and parietal cortex functions. This approach allows for dynamic adaptation to changing driving scenarios, capturing essential perceptual cues for accurate prediction. Evaluated using the Macao Connected and Autonomous Driving (MoCAD) dataset, along with the NGSIM and HighD benchmarks, HLTP demonstrates superior performance compared to existing models, particularly in challenging environments with incomplete data. The project page is available at Github.**

***Index Terms*—Autonomous Driving, Trajectory Prediction, Cognitive Modeling, Knowledge Distillation, Interaction Understanding**

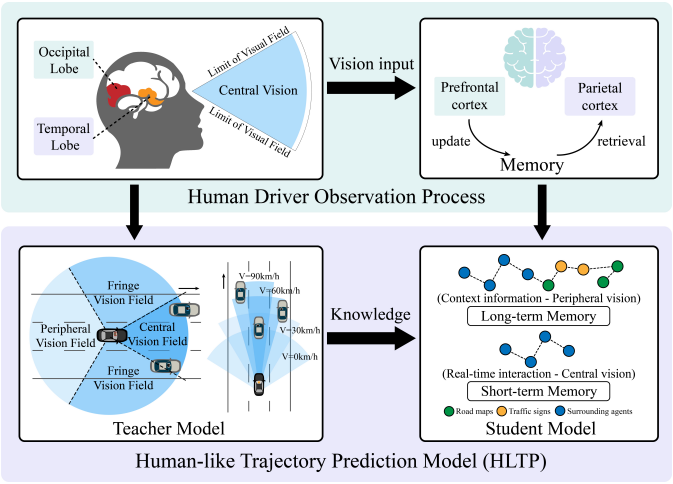


Fig. 1: llustration of the HLTP Model. The "teacher" model integrates an adaptive visual sector and surround-aware encoder, mirroring the occipital and temporal lobes' roles in visual processing. The "student" model emphasizes real-time interaction and decision-making, akin to the prefrontal and parietal cortex functions. These components collectively enable HLTP to replicate the intricate visual and cognitive tasks of human drivers, thereby enhancing trajectory prediction.

I. Introduction

Accurate trajectory prediction is central to advancing autonomous vehicle (AV) technology, requiring more than just an analysis of vehicular dynamics and environmental factors. It demands a deep understanding of decision-making processes, which traditional deep-learning models have yet to fully capture [1]. Despite the leaps made in vehicular pattern recognition and environmental data processing through advanced deep-learning frameworks, a notable gap persists in the emulation of complex, human-like cognitive functions. Current models, although proficient in handling structured data, often falter in scenarios demanding the adaptability and foresight characteristic of human drivers. Bridging this gap calls for an integrative approach that blends the precision of machine learning algorithms with deep insights derived from cognitive and behavioral studies of human drivers.

As depicted in Figure 1, human driving expertise stems from a sophisticated network of neural processes. The brain's capacity to integrate sensory information, especially visual cues, and to make instantaneous decisions based on this integration, represents a remarkable cognitive feat. This involves various brain regions, each playing a specific, interconnected role [2]. The occipital and temporal lobes, responsible for visual processing, work in concert with the prefrontal and parietal cortex, which are central to decision-making and spatial reasoning [3]. It is this harmonious interplay that allows drivers to anticipate potential hazards, adjust to sudden changes in the environment, and make informed decisions based on a blend of current sensory input and past experience [4].

Drawing inspiration from these cognitive mechanisms, our Human-Like Trajectory Prediction (HLTP) model aims to replicate the essence of human decision-making. The model does not merely seek to forecast trajectories based on historical

* Authors contributed equally; † Corresponding author.

Haicheng Liao, Yongkang Li, Zhenning Li, and Chengzhong Xu are with the State Key Laboratory of Internet of Things for Smart City and the Department of Computer and Information Science, University of Macau, Macau. Zhiyong Cui is with the School of Transportation Science and Engineering, Beihang University, Beijing. Shengbo Eben Li is with the School of Vehicle and Mobility, Tsinghua University, Beijing, China. Zhenning Li is also with the Department of Civil and Environmental Engineering, University of Macau, Macau. E-mails: zhenningli, czxu@um.edu.mo

This research is supported by the Science and Technology Development Fund of Macau SAR (File no. 0021/2022/ITP, 0081/2022/A2), and University of Macau (SRG2023-00037-IOTSC).

Manuscript received December 19, 2023; revised February 08, 2024.

data but strives to incorporate the dynamism and context-awareness inherent in human cognition. Our study initially focuses on the dynamic allocation of visual attention by human drivers, who focus primarily on the central vision field and pay less attention to the peripheral vision fields, especially during maneuvers such as lane changes, merging, or speed adjustments. The driver's "vision sector" adapts dynamically with speed: it narrows at high speed for focused attention ahead, and widens at lower speeds to capture more peripheral information, which is critical for slower driving. Therefore, HLTP introduces an adaptive visual sector that mimics this speed-dependent adaptation with a novel pooling mechanism. This mechanism enhances the model's ability to process and interpret visual information, similar to the human visual system, making it adept at recognizing and responding to important details in different driving environments.

Importantly, we propose a heterogeneous knowledge distillation architecture. The "teacher" component of HLTP is designed to mimic the brain's visual processing faculties, employing a neural network architecture that dynamically prioritizes and interprets incoming visual data. This aspect of the model is analogous to the way the occipital and temporal lobes filter, process, and relay visual information to other parts of the brain. The "student" component of HLTP, in turn, represents the decision-making and reasoning aspects of the brain, particularly the functions of the prefrontal and parietal cortex. This part of the model synthesizes the processed visual information, integrating it with spatial awareness and contextual data to make real-time predictive decisions. The goal is to create a model that not only processes information as the human brain does but also adapts and responds to new information with similar flexibility and accuracy. The HLTP model, therefore, is not just an exercise in technical innovation; it is an exploration into the cognitive realms of the human brain, an attempt to harness the complex interplay of neural processes for enhancing AVs' decision-making capabilities. Overall, the contributions of HLTP are multifaceted:

- HLTP introduces a novel vision-aware pooling mechanism that incorporates an adaptive visual sector to mimic the dynamics of human visual attention. Enhanced by a surround-aware encoder and a shift-window attention block, HLTP is able to adeptly capture the critical perceptual cues from diverse scenes.
- HTLP employs a cognitive-inspired teacher-student knowledge distillation framework, improving prediction accuracy and robustness, especially in complex scenarios, by efficiently assimilating intricate patterns of visual attention and spatial awareness.
- The introduction of the Macao Connected and Autonomous Driving (MoCAD) dataset, featuring a right-hand-drive system, offers a novel urban driving context for trajectory prediction studies.

This paper is organized as follows: In Section II, we review the relevant literature in the field. Section III delves into the intricacies of trajectory prediction and presents our proposed model. Our training methods for this model are detailed in Section IV. Section V evaluates the performance of our model on the NGSIM, HighD, and MoCAD datasets. Finally, in section VI, we offer concluding thoughts.

II. Related Work

A. Trajectory Prediction

In the realm of AD, deep learning has catalyzed the adoption of various neural network architectures. Prior studies have explored Recurrent Neural Networks (RNNs) [5], [6], Graph Neural Networks (GNNs) [7], [8], attention mechanisms [9], [10], generative models [11], [12] and Transformers framework [13], [14] to enhance spatial and temporal feature extraction from trajectory data. In recent years, some research has focused on solving trajectory prediction tasks from a human cognitive perspective. Notable research, such as BAT [15], applies graph theory to capture human driving behavior. However, it lacks consideration for human driver observation process. Furthermore, Li et al. [13] include visual regions in their model to represent driver perception, but they fail to account for the dynamic adaptation of the vision sector with changing driving speeds and the human memory storage mechanism. Concurrently, some studies [16]–[18] focus on improving safety in trajectory predictions, while others [19] develop end-to-end models incorporating multimodality. In addition, some recent studies have ventured into the realm of large language models (LLMs) [19], [20] such as GPT and BLIP, and exploited their ability to output future position distributions. However, the computational intensity of existing models and the lack of human-centered driving cognition pose challenges for real-time application and safety. To address this, we introduce a lightweight yet robust model that mimics human driver observation process and memory storage mechanism, enhancing real-time trajectory prediction capabilities.

B. Knowledge Distillation

Knowledge distillation, originally proposed by Hinton et al. [21], involves a complex "teacher" model transferring knowledge to a simpler "student" model. Initially aimed at model compression, its applications have expanded to initialize pre-trained models [22], improve generalization [23], and increase accuracy [24], while reducing computational and time requirements [25]. However, its applications in trajectory prediction, especially in the field of autonomous driving, are not yet widespread. Novel strategies such as Li et al. [26] and Monti et al. [27] allow student models to use fewer observations, while Wsip [28] introduces a wave superposition-inspired social pooling mechanism in this field. Our study integrates human cognitive processes into this paradigm, allowing the teacher model to emulate human reasoning and transfer this knowledge to the student model. This allows for a lightweight model for trajectory prediction that mimics human driving behavior.

III. METHODOLOGY

A. Problem Formulation

The primary aim of this study is to predict the trajectories of all the surrounding vehicles of the AV in a mixed autonomy environment, with each vehicle being referred to

TABLE I: Notation of the main vectors, variables and parameters involved in this paper and their meanings.

Notation	Meaning
$\boldsymbol{X}$	2D position coordinates of the target vehicle and its surrounding vehicles in a traffic scene
$\boldsymbol{Y}$	Predicted future trajectory of target vehicle
$\boldsymbol{W}^Q$	Query weight matrix of the shift-window attention mechanism
$\boldsymbol{W}^K$	Key weight matrix of the shift-window attention mechanism
$\boldsymbol{W}^V$	Value weight matrix of the shift-window attention mechanism
q^i_{mn}	Attention subgraph matrices of the shift window A
k^i_{mn}	Attention subgraph matrices of the shift window B
$\boldsymbol{S}$	Visual vectors of scene representation
M	Context matrices of scene representation
$\boldsymbol{S}_{\Delta p}$	Relative position of the target vehicle to its neighboring vehicle
$\boldsymbol{S}_{\Delta s}$	Relative velocity of the target vehicle to its neighboring vehicle
$\boldsymbol{S}_{\Delta a}$	Relative acceleration of the target vehicle to its neighboring vehicle
$M_{\Delta s}$	Velocity differences among the surrounding vehicles around adjacent times
$M_{\Delta\theta}$	Direction angle differences among the surrounding vehicles around adjacent times
$\tilde{\boldsymbol{S}}$	Output of the vision-aware pooling mechanism
$\boldsymbol{C}$	Probability of different maneuvers

Notation	Meaning
H_{vision}	Visual weight matrix
$\boldsymbol{O}^i_{SWA}$	Output of the ith shift-window attention block
$\boldsymbol{O}_{vis}$	Output of the shift-window attention mechanism
$\mathbf{O}_{sur}$	Outputs of the surround-aware encoder
$\mathbf{O}_{vis}$	Output of the teacher encoder
$\mathbf{O}_{tea}$	Output of the "teacher" model
$\mathbf{O}_{stu}$	Output of the "student" model
$stride_x, stride_y$	Shift strides of the shift window
α	Ratio hyperparameter of knowledge distillation
T	Temperature hyperparameter of knowledge distillation
$\mathcal{L}^N$	Negative Log-Likelihood loss function
$\mathcal{L}^M$	Mean Squared Error loss function
Y^{tea}_t	Output of the "teacher" model
Y^{stu}_t	Output of the "student" model
$\hat{p}^0_t$	Ground-truth 2D position coordinates of the target vehicle
p^0_t	2D position coordinates of the target vehicle predicted by the "teacher" model
$\tilde{p}^0_t$	2D position coordinates of the target vehicle predicted by the "student" model
$\sigma_1, \sigma_2, \sigma_t, \sigma_d$	Learnable uncertainty variances for knowledge distillation modulation

as a *target vehicle* during prediction. At any time t, the state of the ith vehicle can be denoted as p^i_t, where $p^i_t = (x^i_t, y^i_t)$ represents the 2D trajectory coordinate. Given the trajectory coordinate data of the target vehicle (superscript 0) and all its surrounding traffic vehicles (superscripts from 1 to n) in a traffic scene in the interval $[1, T_{obs}]$, denoted as $\boldsymbol{X} = \{p^{0:n}_t\}^{T_{obs}}_{t=1} \in \mathbb{R}^{(n+1)\times T_{obs}\times 2}$, the model aims to predict a probabilistic multi-modal distribution over the future trajectory of the target vehicle, expressed as $P(\boldsymbol{Y}|\boldsymbol{X})$. Here, $\boldsymbol{Y} = \{\boldsymbol{y}^0_t\}^{T_f}_{t=T_{obs}+1} \in \mathbb{R}^{T_f\times 2}$ is the predictive future trajectory coordinates of the target vehicle over a time horizon T_f, and $\boldsymbol{y}^0_t = \{(\tilde{p}^{0,1}_t; \tilde{c}^{0,1}_t), (\tilde{p}^{0,2}_t; \tilde{c}^{0,2}_t), \cdots, (\tilde{p}^{0,C}_t; \tilde{c}^{0,C}_t)\}$ encompasses both the potential trajectory and its associated maneuver likelihood ($\Sigma^1_{\boldsymbol{C}} c^{0,i}_t = 1$), with $\boldsymbol{C}$ denoting the total number of potential trajectories predicted. For the sake of clarity, the main variables and parameters covered in this paper are listed in Table I.

B. Model Architecture

Figure 2 illustrates the architecture of HLTP, which features a novel pooling mechanism with an adaptive visual sector. This sector dynamically adapts, similar to human attention allocation, to capture important cues in different traffic situations. Additionally, the model also employs a teacher-student heterogeneous network distillation approach for human-like trajectory prediction. i) The "teacher" model emulates the human visual observation process by integrating a shift-window attention block (SWA) to mimic central visual focus, along with a surround-aware encoder to capture peripheral vision aspects during maneuvers. ii) The "student" model incorporates a human-like vision-aware pooling mechanism and uses a lightweight framework to efficiently represent the spatio-temporal dynamics among traffic participants. Under the guidance of the "teacher", the "student" strives to bridge the knowledge gap in understanding and emulating human driving behavior, ultimately aiming to improve the accuracy of its predictions, while requiring fewer input observations.

C. Scene Representation

HLTP describes scenes by focusing on the relative positions of traffic vehicles, aligning with human spatial understanding. It preprocesses historical data into two key spatial forms:
1) Visual vectors $\boldsymbol{S}$: These vectors are constructed to encapsulate the relative position, velocity, and acceleration of the target vehicle to its neighboring vehicles. Represented as $\boldsymbol{S} = \{\boldsymbol{S}_{\Delta p}, \boldsymbol{S}_{\Delta s}, \boldsymbol{S}_{\Delta a}\}$, they fall within the dimensional space of $\mathbb{R}^{(n+1)\times T_{obs}\times 4}$; **2) Context matrices M:** These matrices provide a detailed description of speed and direction angle differences among the surrounding vehicles. Defined as $M = \{M_{\Delta s}, M_{\Delta\theta}\} \in \mathbb{R}^{(n+1)\times T_{obs}\times 2}$. They provide a comprehensive view of the relative motion dynamics within the traffic scene. These representations $\boldsymbol{S}$ and M, capturing the geometric nuances of the traffic environment, are then fed into the vision-aware pooling mechanism and teacher-student knowledge distillation framework of HLTP for further feature extraction.

D. Vision-aware Pooling Mechanism

Research [29] indicates that human drivers, limited by their brain's working memory, effectively process information from only a few external vehicles, prioritizing the central visual field, particularly in high-risk frontal collision scenarios. Furthermore, a driver's visual sector, defined by its radius and angle, dynamically adjusts with speed: narrowing at higher speeds for focused attention ahead and widening at lower speeds for broader peripheral awareness.

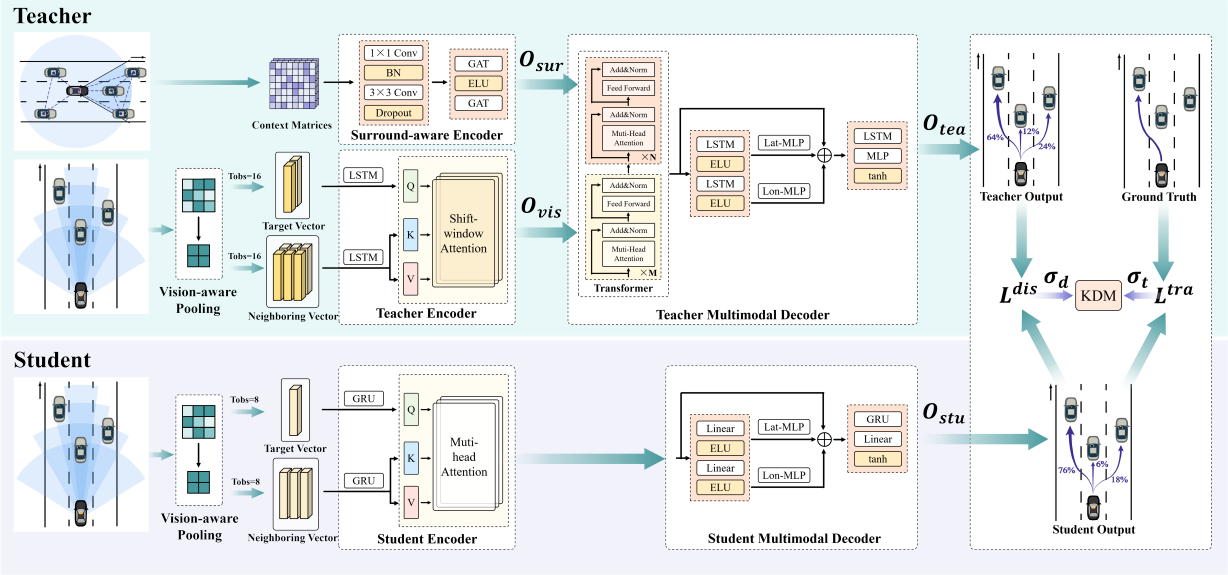


Fig. 2: Overall "teacher-student" architecture of the HLTP. The Surround-aware Encoder and the Teacher Encoder within the "teacher" model process visual vectors and context matrices to produce surround-aware and visual-aware vectors, respectively. These vectors are then fed into the Teacher Multimodal Decoder, which enables the prediction of different potential maneuvers for the target vehicle, each with associated probabilities. The "student" model acquires knowledge from the "teacher" model using a Knowledge Distillation Modulation (KDM) training strategy.

Inspired by this, HLTP introduces a vision-aware pooling mechanism that emulates the adaptive visual attention of human drivers. It features an innovative adaptive visual sector that adjusts the field of view based on vehicle speed, enabling intuitive and context-sensitive trajectory prediction. This approach departs from models that assume a uniform distribution of attention. Instead, we introduce a visual weight matrix H that adapts to represent the driver's shifting focus across the central visual field at different speeds. We establish speed thresholds at $0km/h$, $30km/h$, $60km/h$, and $90km/h$, each corresponding to different visual sectors. Specifically, different speed thresholds correspond to different central angles of the field of view, as shown in Figure 1. We assign different learnable weights, denoted w_c and w_n, to vehicles inside and outside the central field of view, with initial values set to $w_c = 1$ and $w_n = 0.2$. Vehicles within the central field of view are assigned higher weights to indicate their greater importance, which is adaptively adjusted in real-time according to different traffic scenarios. These thresholds allow different weights to be applied to the model based on the speed of the vehicle, resulting in a more refined understanding of attention allocation during driving. The integration of the visual weight matrix H_{vision} with the input visual vectors $\boldsymbol{S}$ by element-wise multiplication. Formally,

$$\tilde{\boldsymbol{S}} = H_{vision} \odot \boldsymbol{S}, \tag{1}$$

This pooling mechanism culminates in the generation of vision-aware vectors $\tilde{\boldsymbol{S}} \in \mathbb{R}^{(n+1)\times T_{obs}\times d_s}$ that encapsulate the varying focus and attention patterns of human drivers.

E. Teacher Model

The "teacher" model integrates Teacher Encoder and Surround-aware Encoder to closely emulate human visual perception, mirroring the retinal processing of the human drivers. As an enhancement, it further employs a transformer framework in the decoder to effectively extract spatio-temporal interactions and contextual features. This hierarchical learning process equips the model with advanced knowledge, which is subsequently used to guide the "student" model to make human-like trajectory predictions with improved accuracy.

1) Teacher Encoder: The encoder processes central field-of-vision information into vision-aware vectors $\tilde{\boldsymbol{S}}$ that are critical for real-time road conditions and immediate obstacles, significantly influencing driving decisions and reactions. In real-world driving, the human brain, with its limited processing capacity, prioritizes information to facilitate efficient decision-making. As a result, the trajectory of a human-driven vehicle is mostly determined by recent traffic scenarios, as the human brain vividly remembers short-term events that influence driving behaviors. In addition, the receptive fields (RFs) of the human visual cortex are limited and typically focus on the most immediate traffic scenes based on neural stimulus intensity. This further implies that human drivers are more inclined to focus on and be mostly influenced by the movements of recent traffic scenarios, as these are perceived as immediate and relevant to their driving decisions.

Furthermore, we introduce a novel shift-window attention block inspired by [30] that involves extracting vectors for the target vehicle $\boldsymbol{F}_{tar}$ and its neighbors $\boldsymbol{F}_{nbr}$ from $\tilde{\boldsymbol{S}}$ and embedding them via the LSTM-ELU-Linear structure for high-level vector production. Specifically, SWA is built on the multi-head attention mechanism that computes the attention

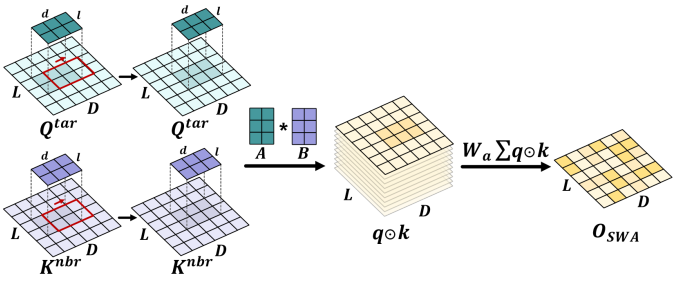


Fig. 3: Visualization of the shift-window function in SWA. A window shifting technique is used to capture features by moving windows across key and query vectors, similar to convolution. Overlapping shifted windows, highlighted in red, connect with prior layers, promoting interaction and information sharing between query and key vectors.

weight within the local windows, while the skip connections are applied for training stability and efficiency. It projects the embedded target vehicle $\tilde{\boldsymbol{F}}_{tar}$, neighboring vehicles $\tilde{\boldsymbol{F}}_{nbr}$, and the vision-aware $\tilde{\boldsymbol{S}}$ vectors into the query $\mathbf{Q}$, key $\mathbf{K}$, value $\mathbf{V}$ representations using Multilayer Perceptrons (MLP) with different embedding dimensions d_q,d_k,d_v, respectively:

$$\mathbf{Q} = \mathbf{W}^Q \tilde{\boldsymbol{F}}_{tar}, \quad \mathbf{K} = \mathbf{W}^K \tilde{\boldsymbol{F}}_{nbr}, \quad \mathbf{V} = \mathbf{W}^V \tilde{\boldsymbol{S}}, \tag{2}$$

where $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V$ are the learnable weight matrices. Then, $\mathbf{Q} \in \mathbb{R}^{(n+1)\times d_q}, \mathbf{K} \in \mathbb{R}^{(n+1)\times d_k}$ and $\mathbf{V} \in \mathbb{R}^{(n+1)\times d_v}$ are divided into H multiple heads, denoted $\mathbf{Q}^i \in \mathbb{R}^{(n+1)\times \frac{d_q}{H}}$, $\mathbf{K}^i \in \mathbb{R}^{(n+1)\times \frac{d_k}{H}}$, and $\mathbf{V}^i \in \mathbb{R}^{(n+1)\times \frac{d_v}{H}}, i \in [1, H]$. The ith attention head $\boldsymbol{head}^i$ and the output $\boldsymbol{O}_{vis}$ of the shift-window attention mechanism can be defined mathematically as:

$$\boldsymbol{O}_{vis} = \sum_{i=1}^{H} \boldsymbol{head}^i = \sum_{i=1}^{H} \phi_{softmax}\left(\frac{\phi_{shift}(\mathbf{Q}^i, \mathbf{K}^i)}{\sqrt{d_k}}\right) \odot \tanh(\mathbf{V}^i), \tag{3}$$

where ϕ_{shift} represents the shift-window function. Figure 3 illustrates our approach of treating local windows as the human visual system's finite receptive fields, akin to convolutional kernels. We employ a shift-window method to compute attention weights within these windows, linking them and using their slight overlap for efficient feature aggregation and updating. For $\mathbf{Q}^i$ and $\mathbf{K}^i$ with dimension $L \times D$, we establish unique shift windows A and B, each with dimension $l \times d$, considerably smaller than $\mathbf{Q}^i$ and $\mathbf{K}^i$. Both windows have fixed shift steps $stride_x$ and $stride_y$ for the dimensions x and y, respectively. Local attention weights are calculated within these windows, which are then sequentially shifted based on their strides. We define q^i_{mn} and k^i_{mn} as attention subgraph matrices for windows A and B, with m and n indicating their respective displacements in the x and y directions. This process produces the output $\boldsymbol{O}^i_{SWA}$:

$$\boldsymbol{O}^i_{SWA} = \phi_{shift}(\mathbf{Q}^i, \mathbf{K}^i) = \mathbf{W}_a \sum_{m=0}^{z_x} \sum_{n=0}^{z_y} q^i_{mn} \odot k^i_{mn}, \tag{4}$$

where the weighted product of these maps is computed using the weight matrix $\mathbf{W}_a$. To ensure synchronized traversal of these windows, it is essential that $z_x = z_y = \frac{L-l}{stride_x} + 1 = \frac{D-d}{stride_y} + 1$. This SWA allows for precise modeling and processing of attention, closely mimicking the selective and dynamic focusing of human vision in complex driving environments.

2) Surround-aware Encoder: Human drivers focus on their central vision while continuously monitoring their peripheral vision through the side and rear-view mirrors to understand their surroundings, including nearby vehicles, pedestrians, and road conditions. To replicate this peripheral monitoring, especially during maneuvers, we introduce the surround-aware encoder. It processes a quarter of the time-segmented matrices $M \in \mathbb{R}^{(n+1)\times \frac{T_{obs}}{4} \times 2}$ with a 1×1 convolutional layer for channel expansion, followed by a 3×3 layer for specific feature extraction, incorporating batch normalization and dropout for robustness. Enhanced with Graph Attention Networks (GAT) and ELU activation, it produces surround-aware vectors $\mathbf{O}_{sur}$.

3) Teacher Multimodal Decoder: The combined outputs of the surround-aware $\mathbf{O}_{sur}$ and the teacher $\mathbf{O}_{vis}$ encoders are fed into a transformer for advanced spatio-temporal interaction analysis, generating hidden states $\mathbf{I}$. The decoder of the "teacher" model, based on a Gaussian Mixture Model (GMM), accounts for uncertainty in trajectory prediction by evaluating multiple possible maneuvers and their probabilities.

To provide a formal description, the future trajectories can be predicted hierarchically in a Bayesian framework on two hierarchical levels: (1) At each time step, the probability of different maneuvers $\boldsymbol{C}$ of the target vehicle is determined; (2) Subsequently, the detailed trajectories of the vehicle conditional on each maneuver are generated within a predefined distributional form. In accordance with the characteristics of the driver's actions during driving, the possible maneuvers of the vehicles are decomposed into a combination of two distinctive sub-maneuvers, comprising the position-wise sub-maneuver c_p and the velocity-wise sub-maneuver c_v. Additionally, the position-wise sub-maneuver includes three discrete driver choices regarding position changes, namely left lane change, right lane change, and lane keeping. Meanwhile, the speed-wise sub-maneuver includes three distinct decisions, including accelerating, braking, and maintaining speed, which serve as branch categories.

Conditional on the estimated maneuvers $\boldsymbol{C}$ in the first layer, the probability distribution of the multimodal trajectory predictions is assumed to follow a Gaussian distribution:

$$P_{\boldsymbol{\Omega}}(\boldsymbol{Y}|\boldsymbol{C}, \boldsymbol{X}) = N(\boldsymbol{Y}|\mu(\boldsymbol{X}), \Sigma(\boldsymbol{X})), \tag{5}$$

where $\boldsymbol{X}$ denotes the input of the proposed model. Moreover, $\boldsymbol{\Omega} = \left[\Omega^{t+1}, \ldots, \Omega^{t+t_f}\right]$ are the estimable parameters of the distribution, and $\Omega^t = [\mu^t, \Sigma^t]$ are the mean and variance of the distribution of predicted trajectory point at t. Correspondingly, in the second layer, the multimodal predictions are formulated as a Gaussian Mixture Model, i.e.,

$$P(\boldsymbol{Y}|\boldsymbol{X}) = \sum_{\forall i} P\left(c_i|\boldsymbol{X}\right) P_{\boldsymbol{\Omega}}\left(\boldsymbol{Y}|c_i, \boldsymbol{X}\right), \tag{6}$$

where c_i is the i-th element in $\boldsymbol{C}$.

This multimodal structure not only provides different predictions, but also quantifies their confidence levels, which

supports decision-making amidst unpredictability in prediction. The hidden states $\mathbf{I}$ are then processed through softmax activation and an MLP, forming a probability distribution $\mathbf{O}_{tea}$ over potential future trajectories. This approach balances precision and adaptability, which is critical for dynamic and uncertain driving environments.

F. Student Model

The "student" model mimics human memory storage strategies by prioritizing real-time interactions and central visual field data, similar to a driver's focus during navigation. Information from peripheral vision, such as road signs and maps, provides broader context but receives less attention, reflecting the division between short-term and long-term memory in human cognition. This design ensures that the "student" model closely mirrors the behavior of the "teacher" model, balancing immediate detail with overarching context in decision-making. It focuses on short-term input, processing a limited number of short-term, vision-aware vectors $\tilde{\boldsymbol{S}}$ ($T_{obs} = 8$), using a lightweight architectural design for efficient learning. In a departure from the more complex Transformer framework of the "teacher" model, the "student" model employs the GRU for its decoder to produce a multimodal prediction distribution $\mathbf{O}_{stu}$. By distilling knowledge from the "teacher" model, the "student" model can make effective predictions even when constrained by limited observational data.

IV. Training

A. Teacher Training

For the "teacher" model, we follow a standard protocol, using 3 seconds of observed trajectory for input ($T_{obs} = 16$) and predicting a 5-second future trajectory ($T_f = 25$) with a sampling rate of 2. To extract complex knowledge from the dataset, we allow slight over-fitting during training. To address the potentially detrimental effects of misclassified maneuver types on trajectory prediction accuracy and robustness, we adopt a multitask learning approach for teacher loss, which can be decoupled as Mean Squared Error (MSE) loss $\mathcal{L}^M$ and Negative Log-Likelihood (NLL) loss $\mathcal{L}^N$.

The NLL Loss function is commonly used in probabilistic models. For a specific data point, given the model's predicted probability distribution $P(\boldsymbol{Y}|\boldsymbol{X})$, where $\boldsymbol{Y}$ represents the output future trajectory and $\boldsymbol{X}$ denotes the input features, the NLL Loss $\mathcal{L}^N$ is defined: $\mathcal{L}^N = -log(P(\boldsymbol{Y}|\boldsymbol{X}))$.

In our trajectory prediction process, we treat the output trajectory as a bivariate Gaussian distribution consisting of the horizontal coordinate x and the vertical coordinate y. Given the ground-truth 2D position coordinate $\hat{p}_t^{gt} = (\hat{x}_t^{gt}, \hat{y}_t^{gt})$, and the final predicted trajectory set by the teacher model is $p_t^{tea} = (x_t^{tea}, y_t^{tea})$, the possibility $P(\hat{p}_t^{gt}|p_t^{tea})$ is defined as:

$$P(\hat{p}_t^{gt}|p_t^{tea}) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{\hat{x}_t^{gt}-\mu_x}{\sigma_x}\right)^2 - 2\rho\frac{\hat{x}_t^{gt}-\mu_x}{\sigma_x}\frac{\hat{y}_t^{gt}-\mu_y}{\sigma_y} + \left(\frac{\hat{y}_t^{gt}-\mu_y}{\sigma_y}\right)^2\right]\right\}, \tag{7}$$

where $P(\hat{p}_t^{gt}|p_t^{tea}) = (\tilde{c}_t^{0,1}, \tilde{c}_t^{0,2}, \tilde{c}_t^{0,3}, \cdots, \tilde{c}_t^{0,C})$, μ_x, μ_y, σ_x, σ_y are the means and variances of $\hat{x}_t^{gt}$, $\hat{y}_t^{gt}$, ρ is the correlation coefficient between $\hat{x}_t^{gt}$ and $\hat{y}_t^{gt}$. Moreover, we calculate the teacher loss using the trajectory prediction set of "teacher" model $\boldsymbol{Y}_t^{tea} = \{(p_t^{0,1}; c_t^{0,1}), (p_t^{0,2}; c_t^{0,2}), \cdots, (p_t^{0,C}; c_t^{0,C})\}$ and the ground-truth set $\hat{\boldsymbol{Y}}_t^{gt} = \{(\hat{p}_t^0; \hat{c}_t^0)\}$ with their corresponding maneuver probability distributions, $t \in [T_{obs}+1, T_f]$. This loss function, denoted as $\mathcal{L}^{tea}$, is derived by a one-to-one calculation between the predicted and actual trajectories:

$$\mathcal{L}^{tea} = \sum_t^{T_f}\sum_i^C \left[\mathcal{L}^M\left(p_t^{0,i}, \hat{p}_t^0\right) + \mathcal{L}^N\left(\boldsymbol{Y}_t^{tea}, \hat{\boldsymbol{Y}}_t^{gt}\right)\right] \tag{8}$$

B. Student Training

The "student" model is trained to predict 5-second trajectories with fewer input observations. To improve its predictive performance with limited observations, we decouple the student loss $\mathcal{L}^{stu}$ as *track* loss $\mathcal{L}^{tra}$ and *distillation* loss $\mathcal{L}^{dis}$. The track loss, similar to that of the "teacher" model, quantifies the discrepancy between the model's predicted trajectories and their ground truths. Mathematically,

$$\begin{aligned} &\mathcal{L}^{tra} = \sum_t^{T_f}\sum_i^C \left[\mathcal{L}^M\left(\tilde{p}_t^{0,i}, \hat{p}_t^0\right) + \mathcal{L}^N\left(\boldsymbol{Y}_t^{stu}, \hat{\boldsymbol{Y}}_t^{gt}\right)\right], \\ &\textit{such that } \boldsymbol{Y}_t^{stu} = \{(\tilde{p}_t^{0,1}; \tilde{c}_t^{0,1}), (\tilde{p}_t^{0,2}; \tilde{c}_t^{0,2}), \cdots, (\tilde{p}_t^{0,C}; \tilde{c}_t^{0,C})\}, \end{aligned} \tag{9}$$

Here, $\boldsymbol{Y}_t^{stu}$ represents the output of "student" model. In contrast, the MSE loss used in *distillation* focuses on guiding "student" model to mimic the behavior of its "teacher":

$$\mathcal{L}^{dis} = 2\alpha T^2 \sum_t^{T_f}\sum_i^C \left[\mathcal{L}^N\left(\tilde{P}_t^{0,i}, P_t^{0,i}\right) + \mathcal{L}^N\left(\tilde{c}_t^{0,i}, c_t^{0,i}\right)\right], \tag{10}$$

Here, T is the distillation temperature, used to modulate the smoothness of the output probability distribution, and α is a hyperparameter to measure the proportion between the two loss functions. Correspondingly, $\tilde{P}_t^{0,i} = \phi_{softmax}(\frac{\tilde{p}_t^{0,i}}{T})$ denotes the output obtained post-softmax application. Eventually, the student loss $\mathcal{L}^{stu}$ is the summation of *track* loss $\mathcal{L}^{tra}$ and *distillation* loss $\mathcal{L}^{dis} = \mathcal{L}^{tra} + \mathcal{L}^{dis}$.

C. Knowledge Distillation Modulation

Given that L^{stu} is composed of sub-loss functions from multiple tasks, determining the proportionality relationships between them poses a challenging problem. Both $\mathcal{L}^{tra}$ and $\mathcal{L}^{dis}$ are further decomposed into maneuver loss function $\mathcal{L}_{man}$ and trajectory coordinate loss function $\mathcal{L}_{coor}$:

$$\mathcal{L}^{tra} = \mathcal{L}_{man}^{tra} + \mathcal{L}_{coor}^{tra},\ \mathcal{L}^{dis} = \mathcal{L}_{man}^{dis} + \mathcal{L}_{coor}^{dis}, \tag{11}$$

where each part is formulated as follows:

$$\begin{aligned} \mathcal{L}_{coor}^{tra} &= \sum_t^{T_f}\sum_i^C \mathcal{L}^M\left(\tilde{p}_t^{0,i}, \hat{p}_t^0\right), \\ \mathcal{L}_{man}^{tra} &= \sum_t^{T_f}\sum_i^C \mathcal{L}^N\left(\boldsymbol{Y}_t^{stu}, \hat{\boldsymbol{Y}}_t^{gt}\right), \\ \mathcal{L}_{coor}^{dis} &= 2\alpha T^2 \sum_t^{T_f}\sum_i^C \mathcal{L}^N\left(\tilde{P}_t^{0,i}, P_t^{0,i}\right), \\ \mathcal{L}_{man}^{dis} &= 2\alpha T^2 \sum_t^{T_f}\sum_i^C \mathcal{L}^N\left(\tilde{c}_t^{0,i}, c_t^{0,i}\right), \end{aligned} \tag{12}$$

TABLE II: Comparison of MoCAD with other AV datasets. The datasets highlighted in **bold** are those that we evaluated specifically for this study. Within the vehicle column, "Veh." represents a range of vehicles, including cars, trucks, trailers, vans, and buses. Meanwhile, "Ped." and "Bic." are shorthand for pedestrians and bicyclists respectively. In the context of Data Type, "Traj." refers to the trajectory coordinates, while "HD maps" denotes the high-definition (HD) maps. Furthermore, "PT" and "SG" in the Location column stand for Pittsburgh and Singapore, correspondingly.

Dataset	Year	vehicle	Scene	Data Type	Tracking Quantity	Sensor	Sampling Rate	Location
NGSIM [32]	2006	Veh.	Highway	Traj.	1.5 hours	Cemara	10 Hz	USA
KITTI [33]	2013	Bic., Ped.	Urban, Highway	Image, Point cloud	50 Sequences	Lidar, Camera	10 Hz	Karlsruhe
HighD [34]	2018	Veh.	Highway	Traj., Lane	447 hours, 110500 vehicles	Drone	25 Hz	German
Apolloscape [35]	2018	Veh., Bic., Ped.	Urban	Traj.	1000 Km Trajectories	Lidar, Camera	10 Hz	China
Argovrse [36]	2019	Vehicles	Urban	Traj., HD maps	320h	Lidar, Camera	30 Hz	Miami, PT
RounD [37]	2020	Veh., Bic., Ped.	Roundabouts	Traj., Lanes	13746 Tracks	Drone	25 Hz	German
nuScenes [38]	2020	Veh., Bic., Ped.	Urban	Traj., HD maps	5.5 hours	Lidar, Camera	12 Hz	Boston, SG
MoCAD (Ours)	2023	Veh., Bic., Ped.	Urban, Campus	Traj., HD maps, Point cloud, BEV	300 hours	Lidar, Camera, Drone	10 Hz	Macao

Notably, inspired by previous work [31], we incorporate a Knowledge Distillation Modulation to weight the *track* loss and the *distillation* loss, with homoscedastic uncertainty. To the best of our knowledge, we are the first to propose a multi-level, multi-task hyperparameter tuning approach to autonomously adjust knowledge distillation hyperparameters during training in this field. A multi-level task is defined where the overarching training loss function is composed of an ensemble of sub-loss functions. Each of these sub-loss functions is further composed of additional sub-loss functions, all of which have some degree of similarity to each other.

Theoretically, there are two options for multi-level, multi-task hyperparameter tuning in our context: Option (1): The first level involves multi-task tuning of $\mathcal{L}^{tra}$ and $\mathcal{L}^{dis}$, while the second level involves multi-task tuning of $\mathcal{L}_{man}$ and $\mathcal{L}_{coor}$. Option (2): The first level deals with multi-task tuning of $\mathcal{L}_{man}$ and $\mathcal{L}_{coor}$, while the second level focuses on $\mathcal{L}^{tra}$ and $\mathcal{L}^{dis}$. We first adopt option (1) using the multi-task tuning approach outlined in [31] for the first level of the loss function, we derive the following equation:

$$\mathcal{L}_1^{stu} = \frac{1}{2\sigma_t^2}\mathcal{L}^{tra}(\mathbf{W}) + \frac{1}{2\sigma_d^2}\mathcal{L}^{dis}(\mathbf{W}) + log\sigma_t\sigma_d, \quad (13)$$

Then, employing multi-task tuning separately for the second level of the loss functions results in the following formulation:

$$\begin{aligned}\mathcal{L}^{tra} &= \frac{1}{2\sigma_1^2}\mathcal{L}_{man}^{tra}(\mathbf{W}) + \frac{1}{2\sigma_2^2}\mathcal{L}_{coor}^{tra}(\mathbf{W}) + log\sigma_1\sigma_2, \\ \mathcal{L}^{dis} &= \frac{1}{2\sigma_1^2}\mathcal{L}_{man}^{dis}(\mathbf{W}) + \frac{1}{2\sigma_2^2}\mathcal{L}_{coor}^{dis}(\mathbf{W}) + log\sigma_1\sigma_2,\end{aligned} \quad (14)$$

where $\sigma_1, \sigma_2, \sigma_t, \sigma_d$ are the learnable uncertainty variances.

Considering the analogous properties of $\mathcal{L}_{man}^{tra}$ with $\mathcal{L}_{man}^{dis}$, $\mathcal{L}_{coor}^{tra}$ with $\mathcal{L}_{coor}^{dis}$, we apply the same variance adjustments, separately. By substituting the above expressions, we derive the following formulation:

$$\begin{aligned}\mathcal{L}_1^{stu}(W, \sigma_1, \sigma_2, \sigma_t, \sigma_d) = \frac{1}{2\sigma_t^2}(\frac{1}{2\sigma_1^2}\mathcal{L}_{man}^{tra} + \frac{1}{2\sigma_2^2}\mathcal{L}_{coor}^{tra}) \\ + \frac{1}{2\sigma_d^2}(\frac{1}{2\sigma_1^2}\mathcal{L}_{man}^{dis} + \frac{1}{2\sigma_2^2}\mathcal{L}_{coor}^{dis}) + log\sigma_1\sigma_2(\frac{1}{2\sigma_t^2} + \frac{1}{2\sigma_d^2}) + log\sigma_t\sigma_d,\end{aligned} \quad (15)$$

Similarly, by adopting option (2), we obtain the following:

$$\begin{aligned}\mathcal{L}_2^{stu}(W, \sigma_1, \sigma_2, \sigma_t, \sigma_d) = \frac{1}{2\sigma_1^2}(\frac{1}{2\sigma_t^2}\mathcal{L}_{man}^{tra} + \frac{1}{2\sigma_d^2}\mathcal{L}_{coor}^{dis}) \\ + \frac{1}{2\sigma_2^2}(\frac{1}{2\sigma_t^2}\mathcal{L}_{man}^{dis} + \frac{1}{2\sigma_d^2}\mathcal{L}_{coor}^{dis}) + log\sigma_t\sigma_d(\frac{1}{2\sigma_1^2} + \frac{1}{2\sigma_2^2}) + log\sigma_1\sigma_2,\end{aligned} \quad (16)$$

The difference between Eq. 15 and Eq. 16 merely lies in their respective constant terms. Specifically, the expression $log\sigma_1\sigma_2(\frac{1}{2\sigma_t^2}+\frac{1}{2\sigma_d^2})+log\sigma_t\sigma_d$ is not equal to to $log\sigma_t\sigma_d(\frac{1}{2\sigma_1^2}+\frac{1}{2\sigma_2^2}) + log\sigma_1\sigma_2$. This demonstrates the non-equivalence of the two multi-task tuning approaches. Therefore, we optimize the formula by omitting the constant term and replacing it with $\log\sigma_1 + \log\sigma_2 + \log\sigma_t + \log\sigma_d$, or alternatively as $\log\sigma_1\sigma_2\sigma_t\sigma_d$. This approach not only introduces regularization by the uncertainty variances $\sigma_1, \sigma_2, \sigma_t, \sigma_d$ and makes different options share the same expression, but also simplifies the formula, making the training process easier.



Fig. 4: Autonomous driving testing platform of University of Macau.

V. Experiment

A. Datasets

We evaluate the effectiveness of our model using the three datasets: Next Generation Simulation (NGSIM) [32], Highway Drone (HighD) [34], and our proposed MoCAD. The MoCAD dataset [15] was collected from the first Level 5 autonomous bus in Macau (as shown in Figure 4) and has undergone extensive testing and data collection since its deployment in 2020. The data collection period spans over 300 hours and covers various scenarios, including a 5-kilometer campus road dataset, a 25-kilometer dataset covering city and urban roads, and complex open traffic environments captured under different weather conditions, time periods, and traffic densities. As shown in Table II, we present a comprehensive comparison of our proposed MoCAD dataset with other datasets for AVS. It is noteworthy that MoCAD stands out with its extensive tracking hours and unique Macao origin, featuring a right-hand-drive system and mandatory left-hand driving, providing

TABLE III: Evaluation results for our proposed model and the other SOTA baselines without using HD maps in the NGSIM, HighD and MoCAD datasets over a different horizon. Note: RMSE (m) is the evaluation metric, where lower values indicate better performance, with some not specifying ("-"). **Bold** and underlined values represent the best and second-best performance in each category. "AVG" indicates the average value of the RMSE.

Dataset	Model	Prediction Horizon (s)					
		1	2	3	4	5	AVG
NGSIM	S-GAN [39]	0.57	1.32	2.22	3.26	4.40	2.35
	CS-LSTM [32]	0.61	1.27	2.09	3.10	4.37	2.29
	MATF-GAN [40]	0.66	1.34	2.08	2.97	4.13	2.22
	IMM-KF [41]	0.58	1.36	2.28	3.37	4.55	2.43
	MFP [42]	0.54	1.16	1.89	2.75	3.78	2.02
	DRBP [43]	1.18	2.83	4.22	5.82	-	3.51
	WSiP [28]	0.56	1.23	2.05	3.08	4.34	2.25
	CF-LSTM [44]	0.55	1.10	1.78	2.73	3.82	1.99
	MHA-LSTM [45]	0.41	1.01	1.74	2.67	3.83	1.91
	HMNet [46]	0.50	1.13	1.89	2.85	4.04	2.08
	TS-GAN [47]	0.60	1.24	1.95	2.78	3.72	2.06
	STDAN [10]	**0.39**	0.96	1.61	2.56	3.67	1.84
	iNATran (M) [48]	0.41	1.00	1.70	2.57	3.66	1.87
	iNATran [48]	**0.39**	0.96	1.61	2.42	3.43	1.76
	DACR-AMTP [49]	0.57	1.07	1.68	2.53	3.40	1.85
	FHIF [50]	0.40	0.98	1.66	2.52	3.63	1.84
	HLTP	0.41	**0.91**	**1.45**	**2.09**	**2.93**	**1.56**
	HLTP (s)	0.48	0.99	1.54	2.22	3.09	1.66
HighD	S-GAN [39]	0.30	0.78	1.46	2.34	3.41	1.69
	WSiP [28]	0.20	0.60	1.21	2.07	3.14	1.44
	CS-LSTM(M) [32]	0.23	0.65	1.29	2.18	3.37	1.54
	CS-LSTM [32]	0.22	0.61	1.24	2.10	3.27	1.48
	MHA-LSTM [45]	0.19	0.55	1.10	1.84	2.78	1.29
	MHA-LSTM(+f) [45]	0.06	0.09	0.24	0.59	1.18	0.43
	DRBP [43]	0.41	0.79	1.11	1.40	-	0.92
	EA-Net [51]	0.15	0.26	0.43	0.78	1.32	0.59
	CF-LSTM [44]	0.18	0.42	1.07	1.72	2.44	1.17
	STDAN [10]	0.19	0.27	0.48	0.91	1.66	0.70
	DACR-AMTP [49]	0.10	0.17	0.31	0.54	1.01	0.42
	HLTP	**0.09**	**0.16**	**0.29**	**0.38**	**0.59**	**0.30**
	HLTP (s)	0.11	0.23	0.32	0.42	0.68	0.35
MoCAD	S-GAN [39]	1.69	2.25	3.30	3.89	4.69	3.16
	CS-LSTM(M) [32]	1.49	2.07	3.02	3.62	4.53	2.95
	CS-LSTM [32]	1.45	1.98	2.94	3.56	4.49	2.88
	MHA-LSTM [45]	1.25	1.48	2.57	3.22	4.20	2.54
	MHA-LSTM(+f) [45]	1.05	1.39	2.48	3.11	4.12	2.43
	WSiP [28]	0.70	0.87	1.70	2.56	3.47	1.86
	CF-LSTM [44]	0.72	0.91	1.73	2.59	3.44	1.87
	STDAN [10]	0.62	0.85	1.62	2.51	3.32	1.78
	HLTP	**0.55**	**0.76**	**1.44**	**2.39**	**3.21**	**1.67**
	HLTP (s)	0.64	0.86	1.55	2.48	3.27	1.76

rare insights into complex driving patterns for trajectory prediction research.

B. Experimental Setup

1) Dataset Segmentation: In this study, all three datasets are treated with a consistent segmentation framework, i.e., each trajectory is divided into 8-second segments, with the first 3 seconds as historical data and the following 5 seconds for evaluation.

2) Metric: Root Mean Square Error (RMSE) and average RMSE are used as our primary evaluation metrics.

3) Implementation Details: Our model is developed using PyTorch and trained to converge on an A40 48G GPU. We use the Adam optimizer along with CosineAnnealingWarmRestarts for scheduling, with a training batch size of 128 and learning rates ranging from 10^{-3} to 10^{-5}. Unless specified, all evaluation results are based on the "student" model of HLTP. In addition, we lighten the architecture of the "student" model by halving the hidden state dimensions of the GRU and reducing the number of attention heads, termed HLTP (s), for subsequent experiments.

C. Experiment Results

1) Comparison with the State-of-the-art Baseline: Our comprehensive evaluation demonstrates HLTP's superior performance compared to SOTA baselines, as detailed in Table III. HLTP's "student" model, trained on only 1.5 seconds ($T_{obs} = 8$) of recent trajectory data, surpasses the standard 3-second data reliance of baseline models. It notably achieves gains of 5.2% for short-term (2s) and 13.8% for long-term (5s) predictions on the NGSIM dataset. On the HighD dataset, all models, including HLTP, display fewer inaccuracies due to HighD's precise trajectories and larger data size. While short-term predictions are comparable across models, HLTP excels in long-term forecasting with a 41.6% RMSE improvement for up to 5 seconds. In complex right-hand-drive environments like urban streets and unstructured roads (MoCAD dataset), HLTP's accuracy gains range from 3.3% to 11.3%, underscoring its robustness and adaptability in diverse traffic scenarios.

2) Comparing Model Performance and Complexity: In our benchmark against SOTA baselines, as detailed in Table IV, HLTP and HLTP (s) models demonstrate superior performance across all metrics, while maintaining a minimal parameter

count. Despite limited access to many models' source codes, our analysis focuses on open-source options. Remarkably, HLTP (s) achieves high performance with significantly less complexity, reducing parameters by 71.41% and 55.89% compared to WSiP and CS-LSTM, respectively. Furthermore, HLTP (s) efficiently outperforms transformer-based STDAN and CF-LSTM, using 82.34% and 77.79% fewer parameters, respectively. This highlights the efficiency and adaptability of our lightweight "teacher-student" knowledge distillation framework, offering practitioners a balance between accuracy and computational resource requirements.

TABLE IV: Comparative evaluation of our model with selected baselines. Emphasizing model complexity via parameter count (Param.). **Bold** and underlined values represent the best and second-best performance in each category.

Model	Param. (K)	Average RMSE (m)		
		NGSIM	HighD	MoCAD
CS-LSTM	194.92	2.29	1.49	2.88
CF-LSTM	387.10	1.99	1.17	1.88
WSiP	300.76	2.25	1.44	1.86
STDAN	486.82	1.87	0.70	1.78
HLTP	149.50	**1.56**	**0.30**	**1.67**
HLTP (s)	**85.97**	1.66	0.35	1.76

3) Qualitative Results: Figure 5 showcases the multimodal probabilistic prediction performance of HLTP on the NGSIM dataset, while showing the velocities of the target vehicle and its surrounding vehicles. The heat maps shown represent the Gaussian Mixture Model of predictions in challenging scenes. These visualizations show that the highest probability predictions of our model are very close to the ground truth, indicating its impressive performance. Additionally, it also visually demonstrates our model's ability to accurately predict complex scenarios such as merging and lane changing, confirming its effectiveness in various traffic situations. Interestingly, in certain complex scenarios, the trajectory predictions of the "student" model exceed the accuracy of the "teacher" model. This result illustrates the ability of the "student" model to selectively assimilate and refine the knowledge acquired from the "teacher" model, effectively "extracting the essence and discarding the dross".

D. Ablation Studies

1) Ablation Study for Core Components: Table V(a) shows that our ablation study evaluates the performance of HLTP using six model variations, each omitting different components. Method A excludes the vision-aware pooling mechanism, using data from all surround vehicles; Method B excludes the surround-aware module; Method C omits the transformer module; Method D replaces SWA with classical attention; Method E omits multimodal probabilistic maneuver prediction; Method F avoids a multitask learning strategy; Method G, for comparison, includes all these elements, i.e., the complete proposed model HLTP.

The results, presented in Table V(b), show that all removal methods underperform compared to Method G. The inclusion

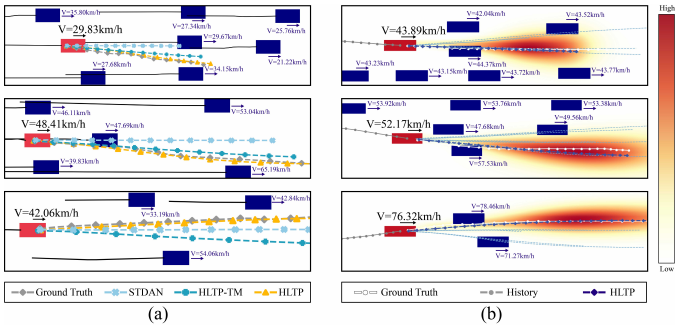


Fig. 5: Multimodal probabilistic prediction (a) and visualizations (b) for the target vehicle on NGSIM. Heat maps illustrate the GMM of predictions: brighter areas denote higher probabilities. The target vehicle is marked in red and its surrounding vehicles in blue.

of the transformer module and a multi-task learning approach in the model significantly improves performance, illustrating their critical role in capturing high-level interactions and evaluating different sources of knowledge for accurate trajectory prediction. Furthermore, Method E omits the multimodal probabilistic maneuver prediction, which also results in a decrease in performance. In addition, the inclusion of the vision-aware pooling mechanism significantly improved performance by adaptively adjusting the attention weights for the surrounding vehicles. By incorporating the human driver's observation process, HLTP predicts the target vehicle's trajectory more insightfully. Similarly, employing the Surround-aware Encoder in Method E outperformed Method B without this encoder, highlighting the relevance of contextual information for traffic scenario relationships. Furthermore, removing the Shift-window Attention block in Method D results in considerable RMSE increases, confirming its vital role in HLTP.

2) Ablation Study for Different Input and Distribution: To explore the impact of varying observation lengths on the HLTP student model, we experimented with different $T_{obs} = 4, 6, 10, 12, 16$, as shown in Table V(c). Furthermore, with $T_{obs} = 6$, HLTP exceeds all baselines in prediction accuracy. As the amount of observation data increases ($T_{obs} > 8$), we observe a slight reduction in prediction error for HLTP; however, the overall performance improvement is marginal. These findings align with research on human attention and decision-making, which suggests that humans tend to prioritize information that is relevant to their current goals and that is within their immediate visual field [52]. To elucidate the effect of different distributions of historical time points on the model's performance, we experimented with different input time distributions. For instance, in the case of $T_{obs} = 4$ we compared the impact of using data from the first to the fourth frames with that of the most recent time points, data from the $(T_{obs} - 3)$th to (T_{obs})th frames. The results clearly indicate that using the initial moments versus the most recent data significantly affects the model's performance, further validating the premise that recent historical data has the most significant influence on future trajectory predictions.

TABLE V: Ablation studies for core component, different observation input, and missing data.

(a) Different methods and components of ablation study.

Components	Ablation Methods						
	A	B	C	D	E	F	G
Vision-aware Pooling Mechanism	✘	✔	✔	✔	✔	✔	✔
Surround-aware Encoder	✔	✘	✔	✔	✔	✔	✔
Transformer Module	✔	✔	✘	✔	✔	✔	✔
Shift-window Attention Block	✔	✔	✔	✘	✔	✔	✔
Multimodal Decoder	✔	✔	✔	✔	✘	✔	✔
Knowledge Distillation Modulation	✔	✔	✔	✔	✔	✘	✔

(b) Ablation results for different models. (NGSIM/HighD)

Time (s)	Ablation Methods					
	A	B	C	D	E	F
1	0.39/0.09	0.52/0.18	0.51/0.10	0.40/0.12	0.46/0.21	0.50/0.10
2	0.93/0.17	1.09/0.25	1.10/0.15	0.95/0.21	1.06/0.37	1.01/0.19
3	1.57/0.36	1.77/0.37	1.80/0.28	1.58/0.33	1.76/0.58	1.56/0.34
4	2.34/0.51	2.56/0.57	2.68/0.46	2.31/0.58	2.66/0.75	2.22/0.47
5	3.28/0.70	3.45/0.72	3.69/0.69	3.23/0.83	3.78/1.02	3.09/0.68
AVG	1.70/0.37	1.88/0.42	1.96/0.34	1.69/0.41	1.94/0.57	1.68/0.36

(c) Ablation results for varying observation frames (recent/initial).

Time (s)	T_{obs}=4	T_{obs}=6	T_{obs}=10	T_{obs}=12	T_{obs}=16
1	0.42/0.68	0.42/0.67	0.41/0.56	0.42/0.53	0.41/0.41
2	0.95/1.33	0.93/1.28	0.92/1.15	0.92/1.12	0.91/0.91
3	1.50/2.04	1.47/1.96	1.45/1.76	1.44/1.69	1.43/1.43
4	2.17/2.85	2.12/2.76	2.08/2.50	2.07/2.39	2.05/2.05
5	3.05/3.86	2.98/3.75	2.91/3.44	2.90/3.29	2.87/2.87

(d) Ablation results for *missing* data.

Time (s)	3-10	4-11	5-12	6-13	7-14
1	0.43	0.44	0.45	0.45	0.48
2	0.95	0.96	0.98	0.99	1.04
3	1.50	1.52	1.55	1.58	1.65
4	2.16	2.19	2.24	2.29	2.37
5	3.02	3.06	3.12	3.20	3.30

TABLE VI: Ablation studies for the hyperparameters of the SWA. Top-performing models are **bolded** in each category.

(a) Ablation study on stride of the shift window in SWA.

Time (s)	$stride_x = 8$ $stride_y = 4$	$stride_x = 8$ $stride_y = 6$	$stride_x = 8$ $stride_y = 8$	$stride_x = 8$ $stride_y = 12$	$stride_x = 16$ $stride_y = 12$
1	0.43	**0.41**	**0.41**	0.42	0.43
2	0.93	**0.91**	**0.91**	0.92	0.94
3	1.47	**1.43**	**1.43**	1.45	1.48
4	2.09	**2.05**	2.06	2.08	2.09
5	2.92	**2.87**	2.88	2.89	2.92

(b) Ablation study on size of the shift window in SWA.

Time (s)	x=32, y=12	x=32, y=18	x=32, y=24	x=32, y=30	x=32, y=36
1	**0.40**	**0.40**	0.41	0.42	0.42
2	0.93	**0.91**	**0.91**	0.92	0.93
3	1.46	1.44	**1.43**	1.45	1.46
4	2.10	2.08	**2.05**	2.07	2.07
5	2.94	2.91	**2.87**	2.89	2.90

3) Ablation Study for Missing Data: We introduced a *missing* test set on the NGSIM dataset, focusing on scenarios where half of the historical data is missing. This set is divided into five subsets based on different distributions of missing data. For example, the $3-10$ subset excludes data from frames 3 through 10. We used linear interpolation to fill these gaps. The results in Table V(d) show that HLTP outperforms all baselines even with 50% missing data, highlighting its adaptability and deep understanding of traffic dynamics. Moreover, we observed that the closer the missing data is to the prediction window ($6-13$, $7-14$ subsets), the more significant the impact on model performance, reinforcing the importance of recent historical data.

4) Ablation Study for the SWA: We investigate the effect of the size and stride of the shifting window in SWA on the predictive performance of the HLTP. As shown in Table VI (a) and (b), in the proposed SWA block, unintuitively, the best experimental results for ablation studies on window stride and window size are achieved with intermediate parameter values. According to ablation studies, the optimal window size is $x = 32, y = 24$ with strides of $stride_x = 8, stride_y = 6$, rather than using a larger sliding window and strides. We attribute this result to two primary factors: the effect of the number of windows on the complexity of the training, and the effect of the number of windows on the amount of information captured. A smaller $stride_y$ results in a larger number of windows, making it more difficult to train the attention weights W_a. Conversely, a larger $stride_y$ reduces the number of windows, thereby limiting the variety of information scales. However, as long as the stride values are within a reasonable range, the union of all windows will still cover the entire dataset, resulting in a small performance penalty. The same principle applies to decreasing the window size. Decreasing the dimension y leads to an increase in the number of windows, thus increasing the training difficulty. Conversely, increasing the dimension y results in fewer windows, thus reducing the amount of information. Overall, our empirical findings resonate with the observations on human attention and decision-making. This research indicates that people typically prioritize information relevant to their current objectives and within their immediate visual field, as suggested by Kahneman [52]. In addition, our findings support the idea that an overly broad range of observations can actually impair predictive performance. This may underscore the importance of focusing on the targeted and relevant information data in AD systems.

VI. Conclusion

In this paper, we propose a novel trajectory prediction model HLTP for AVs that addresses the limitations of previous models in terms of parameter heaviness and limited applicability. It is based on a heterogeneous knowledge distillation network, providing a lightweight yet efficient framework that maintains prediction accuracy. Importantly, HLTP adapts to scenarios with missing data and reduced inputs, using layered learning to simulate human observation and make human-like predictions based on fewer input observations. Our empirical results show that HLTP excels in complex traffic scenarios and delivers SOTA performance. Additionally, we present the right-hand-drive system dataset MoCAD as a new benchmark to refine trajectory prediction methodologies in complex traffic scenes.

REFERENCES

[1] Y. Huang, J. Du, Z. Yang, Z. Zhou, L. Zhang, and H. Chen, "A survey on trajectory-prediction methods for autonomous driving," *IEEE Transactions on Intelligent Vehicles*, vol. 7, no. 3, pp. 652–674, 2022.

[2] S. W. Loke, "Cooperative automated vehicles: A review of opportunities and challenges in socially intelligent vehicles beyond networking," *IEEE Transactions on Intelligent Vehicles*, vol. 4, no. 4, pp. 509–518, 2019.

[3] H. Fukuda, N. Ma, S. Suzuki, N. Harasawa, K. Ueno, J. L. Gardner, N. Ichinohe, M. Haruno, K. Cheng, and H. Nakahara, "Computing social value conversion in the human brain," *Journal of Neuroscience*, vol. 39, no. 26, pp. 5153–5172, 2019.

[4] M. Gao, L. Jin, Y. Jiang, and J. Bie, "Multiple object tracking using a dual-attention network for autonomous driving," *IET Intelligent Transport Systems*, vol. 14, no. 8, pp. 842–848, 2020.

[5] L. Hou, S. E. Li, B. Yang, Z. Wang, and K. Nakano, "Integrated graphical representation of highway scenarios to improve trajectory prediction of surrounding vehicles," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 2, pp. 1638–1651, 2023.

[6] R. Quan, L. Zhu, Y. Wu, and Y. Yang, "Holistic lstm for pedestrian trajectory prediction," *IEEE transactions on image processing*, vol. 30, pp. 3229–3239, 2021.

[7] D. Xu, X. Shang, Y. Liu, H. Peng, and H. Li, "Group vehicle trajectory prediction with global spatio-temporal graph," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 2, pp. 1219–1229, 2023.

[8] X. Li, X. Ying, and M. C. Chuah, "Grip: Graph-based interaction-aware trajectory prediction," in *ITSC*. IEEE, 2019.

[9] C. Chang, J. Zhang, K. Zhang, W. Zhong, X. Peng, S. Li, and L. Li, "Bev-v2x: Cooperative birds-eye-view fusion and grid occupancy prediction via v2x-based data sharing," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 11, pp. 4498–4514, 2023.

[10] X. Chen, H. Zhang, F. Zhao, Y. Hu, C. Tan, and J. Yang, "Intention-aware vehicle trajectory prediction based on spatial-temporal dynamic attention network for internet of vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 10, pp. 19 471–19 483, 2022.

[11] K. Wu, Y. Zhou, H. Shi, X. Li, and B. Ran, "Graph-based interaction-aware multimodal 2d vehicle trajectory prediction using diffusion graph convolutional networks," *IEEE Transactions on Intelligent Vehicles*, pp. 1–13, 2023.

[12] D. Yang, H. Zhang, E. Yurtsever, K. A. Redmill, and Ü. Özgüner, "Predicting pedestrian crossing intention with feature fusion and spatio-temporal attention," *IEEE Transactions on Intelligent Vehicles*, 2022.

[13] Z. Li, Z. Chen, Y. Li, and C. Xu, "Context-aware trajectory prediction for autonomous driving in heterogeneous environments," *Computer-Aided Civil and Infrastructure Engineering*, 2023.

[14] Z. Zhou, J. Wang, Y.-H. Li, and Y.-K. Huang, "Query-centric trajectory prediction," in *Proceedings of the IEEE/CVF CVPR*, 2023.

[15] H. Liao, Z. Li, H. Shen, W. Zeng, D. Liao, G. Li, S. E. Li, and C. Xu, "BAT: behavior-aware human-like trajectory prediction for autonomous driving," *CoRR*, vol. abs/2312.06371, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2312.06371

[16] Y. Yan, L. Peng, T. Shen, J. Wang, D. Pi, D. Cao, and G. Yin, "A multi-vehicle game-theoretic framework for decision making and planning of autonomous vehicles in mixed traffic," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 11, pp. 4572–4587, 2023.

[17] A. Kamenev, L. Wang, O. B. Bohan, I. Kulkarni, B. Kartal, A. Molchanov, S. Birchfield, D. Nistér, and N. Smolyanskiy, "Predictionnet: Real-time joint probabilistic traffic prediction for planning, control, and simulation," 2022.

[18] X. Chen, W. Zhang, H. Bai, C. Xu, H. Ding, and W. Huang, "Two-dimensional following lane-changing (2df-lc): A framework for dynamic decision-making and rapid behavior planning," *IEEE Transactions on Intelligent Vehicles*, 2023.

[19] Y. Hu, J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, S. Du, T. Lin, W. Wang, L. Lu, X. Jia, Q. Liu, J. Dai, Y. Qiao, and H. Li, "Planning-oriented autonomous driving," 2023.

[20] H. Liao, H. Shen, Z. Li, C. Wang, G. Li, Y. Bie, and C. Xu, "Gpt-4 enhanced multimodal grounding for autonomous driving: Leveraging cross-modal attention with large language models," *arXiv preprint arXiv:2312.03543*, 2023.

[21] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *CoRR*, vol. abs/1503.02531, 2015. [Online]. Available: http://arxiv.org/abs/1503.02531

[22] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba, "Eye tracking for everyone," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2176–2184.

[23] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," *CoRR*, vol. abs/1612.03928, 2016. [Online]. Available: http://arxiv.org/abs/1612.03928

[24] T. Furlanello, Z. Lipton, M. Tschannen, L. Itti, and A. Anandkumar, "Born again neural networks," in *International Conference on Machine Learning*. PMLR, 2018, pp. 1607–1616.

[25] S. Bhardwaj, M. Srinivasan, and M. M. Khapra, "Efficient video classification using fewer frames," in *Proceedings of the IEEE/CVF CVPR*, 2019, pp. 354–363.

[26] G. Li, Z. Ji, S. Li, X. Luo, and X. Qu, "Driver behavioral cloning for route following in autonomous vehicles using task knowledge distillation," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 2, pp. 1025–1033, 2023.

[27] A. Monti, A. Porrello, S. Calderara, P. Coscia, L. Ballan, and R. Cucchiara, "How many observations are enough? knowledge distillation for trajectory forecasting," in *Proceedings of the IEEE/CVF CVPR*, 2022.

[28] R. Wang, S. Wang, H. Yan, and X. Wang, "Wsip: wave superposition inspired pooling for dynamic interactions-aware trajectory prediction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 4, 2023, pp. 4685–4692.

[29] A. Tucker and K. Marsh, "Speeding through the pandemic: Perceptual and psychological factors associated with speeding during the covid-19 stay-at-home period," *Accident Analysis & Prevention*, vol. 159, p. 106225, 2021.

[30] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[31] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," 2018.

[32] N. Deo and M. M. Trivedi, "Convolutional social pooling for vehicle trajectory prediction," in *Proceedings of the IEEE CVPR workshops*, 2018, pp. 1468–1476.

[33] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[34] 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 *ITSC*, 2018.

[35] X. Huang, X. Cheng, Q. Geng, B. Cao, D. Zhou, P. Wang, Y. Lin, and R. Yang, "The apolloscape dataset for autonomous driving," in *Proceedings of the IEEE CVPR workshops*, 2018, pp. 954–960.

[36] M.-F. Chang, J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan *et al.*, "Argoverse: 3d tracking and forecasting with rich maps," in *Proceedings of the IEEE/CVF CVPR*, 2019, pp. 8748–8757.

[37] R. Krajewski, T. Moers, J. Bock, L. Vater, and L. Eckstein, "The round dataset: A drone dataset of road user trajectories at roundabouts in germany," in *ITSC*, 2020.

[38] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *the IEEE/CVF CVPR*, 2020.

[39] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social gan: Socially acceptable trajectories with generative adversarial networks," in *Proceedings of the IEEE CVPR*, 2018, pp. 2255–2264.

[40] T. Zhao, Y. Xu, M. Monfort, W. Choi, C. Baker, Y. Zhao, Y. Wang, and Y. N. Wu, "Multi-agent tensor fusion for contextual trajectory prediction," in *Proceedings of the IEEE/CVF CVPR*, 2019.

[41] V. Lefkopoulos, M. Menner, A. Domahidi, and M. N. Zeilinger, "Interaction-aware motion prediction for autonomous driving: A multiple model kalman filtering scheme," *IEEE Robotics and Automation Letters*, vol. 6, no. 1, pp. 80–87, 2020.

[42] C. Tang and R. R. Salakhutdinov, "Multiple futures prediction," *Advances in neural information processing systems*, vol. 32, 2019.

[43] K. Gao, X. Li, B. Chen, L. Hu, J. Liu, R. Du, and Y. Li, "Dual transformer based prediction for lane change intentions and trajectories in mixed traffic environment," *IEEE Transactions on Intelligent Transportation Systems*, 2023.

[44] X. Xie, C. Zhang, Y. Zhu, Y. N. Wu, and S.-C. Zhu, "Congestion-aware multi-agent trajectory prediction for collision avoidance," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 13 693–13 700.

[45] K. Messaoud, I. Yahiaoui, A. Verroust-Blondet, and F. Nashashibi, "Attention based vehicle trajectory prediction," *IEEE Transactions on Intelligent Vehicles*, vol. 6, no. 1, pp. 175–185, 2021.

[46] Q. Xue, S. Li, X. Li, J. Zhao, and W. Zhang, "Hierarchical motion encoder-decoder network for trajectory forecasting," *CoRR*, vol. abs/2111.13324, 2021. [Online]. Available: https://arxiv.org/abs/2111.13324
[47] Y. Wang, S. Zhao, R. Zhang, X. Cheng, and L. Yang, "Multi-vehicle collaborative learning for trajectory prediction with spatio-temporal tensor fusion," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 1, pp. 236–248, 2022.
[48] X. Chen, H. Zhang, F. Zhao, Y. Cai, H. Wang, and Q. Ye, "Vehicle trajectory prediction based on intention-aware non-autoregressive transformer with multi-attention learning for internet of vehicles," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–12, 2022.
[49] P. Cong, Y. Xiao, X. Wan, M. Deng, J. Li, and X. Zhang, "Dacr-amtp: Adaptive multi-modal vehicle trajectory prediction for dynamic drivable areas based on collision risk," *IEEE Transactions on Intelligent Vehicles*, pp. 1–22, 2023.
[50] Z. Zuo, X. Wang, S. Guo, Z. Liu, Z. Li, and Y. Wang, "Trajectory prediction network of autonomous vehicles with fusion of historical interactive features," *IEEE Transactions on Intelligent Vehicles*, 2023.
[51] Y. Cai, Z. Wang, H. Wang, L. Chen, Y. Li, M. A. Sotelo, and Z. Li, "Environment-attention network for vehicle trajectory prediction," *IEEE Transactions on Vehicular Technology*, 2021.
[52] D. Kahneman, *Attention and effort*. Citeseer, 1973, vol. 1063.



Haicheng Liao (Student Member, IEEE) received the B.S. degree in software engineering from the University of Electronic Science and Technology of China (UESTC) in 2022. He is currently pursuing the Ph.D. degree at the State Key Laboratory of Internet of Things for Smart City and the Department of Computer and Information Science, University of Macau. His research interests include connected autonomous vehicles and the application of deep reinforcement learning to autonomous driving.



Yongkang Li (Student Member, IEEE) is an undergraduate student in software engineering from the University of Electronic Science and Technology of China. He is currently serving as a Research Assistant at the State Key Laboratory of Internet of Things for Smart City and the Department of Computer and Information Science, University of Macau. His research interests include computer vision and trajectory predictions for autonomous driving.



Zhenning Li (Member, IEEE) received his Ph.D. in Civil Engineering from the University of Hawaii at Manoa, Honolulu, Hawaii, USA, in 2019. Currently, he holds the position of Assistant Professor at the State Key Laboratory of Internet of Things for Smart City, as well as the Department of Computer and Information Science at the University of Macau, Macau. Over his academic career, he has published over 40 papers. His main areas of research focus on the intersection of connected autonomous vehicles and Big Data applications in urban transportation systems. He has been honored with several awards, including the TRB Best Young Researcher award and the CICTP Best Paper Award, amongst others.



Chengyue Wang is currently a Research Assistant at the State Key Laboratory of Internet of Things for Smart City and the Department of Civil Engineering at the University of Macau. He holds an MS degree in Civil Engineering from the University of Illinois Urbana-Champaign (2022) and a B.E in Transportation Engineering from Chang'an University (2021). His research primarily focuses on the innovative integration of artificial intelligence with autonomous driving technologies and the development of intelligent transportation systems.



Zhiyong Cui received the B.S. degree in software engineering from Beihang University, Beijing, China, in 2012, the M.S. degree in software engineering and microelectronics from Peking University, Beijing, in 2015, and the Ph.D. degree in civil engineering from the University of Washington, Seattle, WA, USA, in 2021. He is currently a Professor at the School of Transportation Science and Engineering, at Beihang University. He was a University of Washington Data Science Postdoctoral Fellow with the eScience Institute, Seattle, WA, USA. His primary research interests include deep learning, machine learning, urban computing, traffic forecasting, connected vehicles, and transportation data science. He was the recipient of the IEEE ITSS Best Dissertation Award in 2021 and Best Paper Award at the 2020 IEEE International Smart Cities Conference.



Shengbo Eben Li (Senior Member, IEEE) received his M.S. and Ph.D. degrees from Tsinghua University in 2006 and 2009, respectively. Before joining Tsinghua University, he had worked at Stanford University, University of Michigan, and UC Berkeley. His active research interests include intelligent vehicles and driver assistance, deep reinforcement learning, optimal control and estimation, etc. He is the author of over 130 peer-reviewed journal/conference papers and the co-inventor of over 30 patents. He is the recipient of the best (student) paper awards of IEEE ITSC, ICCAS, IEEE ICUS, CCCC, etc. His important awards include the National Award for Technological Invention of China (2013), the Excellent Young Scholar of NSF China (2016), the Young Professor of ChangJiang Scholar Program (2016), the National Award for Progress in Sci & Tech of China (2018), Distinguished Young Scholar of Beijing NSF (2018), Youth Sci & Tech Innovation Leader from MOST (2020), etc. He also serves as the Board of Governor of the IEEE ITS Society, Senior AE of IEEE OJ ITS, and AEs of IEEE ITSM, IEEE Trans ITS, Automotive Innovation, etc.



Chengzhong Xu (Fellow, IEEE) received the Ph.D. degree from The University of Hong Kong, in 1993. He is currently the chair professor of computer science and the dean with the Faculty of Science and Technology, University of Macau. Prior to this, he was with the faculty at Wayne State University, USA, and the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China. He has published more than 400 papers and more than 100 patents. His research interests include cloud computing and data-driv intelligent applications. He was the Best Paper awardee or the Nominee of ICPP2005, HPCA2013, HPDC2013, Cluster2015, GPC2018, UIC2018, and AIMS2019. He also won the Best Paper award of SoCC2021. He was the Chair of the IEEE Technical Committee on Distributed Processing from 2015 to 2019.