IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 1 September 2023, accepted 16 September 2023, date of publication 20 September 2023,
date of current version 28 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3317699

RESEARCH ARTICLE

Non-Autoregressive Transformer Based Ego-Motion Independent Pedestrian Trajectory Prediction on Egocentric View

YUJIN KIM[1], EUNBIN SEO[2], CHIYUN NOH[1], AND KYONGSU YI[1], (Member, IEEE)
[1]Department of Mechanical Engineering, Seoul National University, Seoul 08826, South Korea
[2]Interdisciplinary Program in Artificial Intelligence, Seoul National University, Seoul 08826, South Korea

Corresponding author: Kyongsu Yi (kyi@snu.ac.kr)

This work was supported by Korea Evaluation Institute of Industrial Technology (KEIT) grant funded by the Korea government (MOTIE) (No. 1415181716, Development of super high difficulty autonomous driving mobility cognitive prediction sensor technology). This research was also supported by Seoul National University Institute of Advanced Machines and Design (SNU-IAMD) and Seoul National University Future Mobility Technology Center (SNU-FMTC). The Institute of Engineering Research at Seoul National University provided research facilities for this work.

ABSTRACT Predicting the future trajectories of surrounding pedestrians is undoubtedly one of the most essential but challenging tasks for safe urban autonomous driving. Despite this importance, there has been limited research conducted on the egocentric view from easy-to-access vehicle-mounted cameras for autonomous driving applications. This paper presents a non-autoregressive transformer based trajectory prediction methodology for pedestrian on egocentric view. Furthermore, our proposed model predicts ego-motion independent future trajectories for utilization in downstream tasks such as motion planning in autonomous vehicles. This approach differs from previous researches as it focuses on predicting the future position of pedestrians based on the current observed image context, rather than their future positions in future observed images. The proposed model, referred to as the TransPred network in this paper, is composed of three main modules: vehicle motion compensation, non-autoregressive transformer, and conditional variational autoencoder(CVAE). The transformer structure is employed to effectively handle raw images and the historical trajectory of the target pedestrian, enabling the generation of advanced future predictions. Additionally, the CVAE module is utilized in the final part of the overall model to predict plausible multiple future trajectories. It contributes to generating diverse and realistic future trajectory predictions. The performance of our model has been evaluated on Nuscenes and In-house dataset obtained from our vehicle equipped with sensors. We achieves the state-of-the-art performance for prioritized trajectories on both datasets. Moreover, the usability of the proposed ego-motion independent trajectories for autonomous driving is demonstrated through risk assessment experiments.

INDEX TERMS Autonomous driving, autonomous vehicle, attention mechanism, egocentric view, non-autoregressive transformer, pedestrian trajectory prediction.

I. INTRODUCTION

Autonomous driving technology has made significant advancements over the past few decades, expanding its scope from simple highway environments to complex urban settings. In line with this progress, ensuring pedestrian safety in the context of sharing urban driving environments has emerged as a critical focal point. To enhance safety, it is crucial to understand pedestrians' underlying intentions and predict their future actions, as this plays a vital role in preventing potential collisions and disruptions to traffic flow. However, despite this importance, there have been insufficient researches conducted to adequately apply these studies to autonomous vehicles. Existing studies [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18] predicting pedestrian trajectories have

The associate editor coordinating the review of this manuscript and approving it for publication was Hui Ma.

predominantly been conducted with fixed cameras on bird's eye view or building cctv view, rather than cameras mounted on the autonomous vehicle. This distinction in domains arises from the fact that the former measures the position on the top view, not the bounding box of the pedestrian's body, and is not affected by ego-motion. On the other hand, research utilizing easily accessible vehicle-mounted cameras with an egocentric perspective for prediction purposes remains largely underexplored in the field of autonomous driving. Moreover, even a limited number of existing studies [19], [20], [21], [22], [23], [24], [25] in the egocentric view predict pedestrian's future trajectory that encompass both ego-motion and the movement of the target object. In other words, the trajectory predicted in these previous studies aims to track the ego-motion-dependent trajectory, which is the cumulative sequence of pedestrian bounding boxes observed at each future time step from the vehicle-mounted camera during driving. These ego-motion dependent predictive outcomes, as evident in Fig. 1, demonstrate challenges in grasping pedestrians' intention and assessing potential risks due to their inconsistency with the surrounding context on the current captured images.

This paper proposes ego-motion independent trajectory prediction considering visual context and historical trajectory based on a non-autoregressive transformer in egocentric view. Fig. 1 illustrates an example of the proposed model's prediction result on Nuscenes dataset, along with the ego-motion-dependent future trajectory pursued in previous researches. On the current observed image, the ego-motion-dependent trajectory appears to be pushed to the left due to the future movements of the vehicle. On the other hand, the trajectory we aim to track represents the future positions of the pedestrian in the current image coordinates, independent of the future ego-motion. As shown in the example image, our model generates predictions that adhere to an ego-motion independent trajectory. The output trajectory is in harmony with the surrounding image context at the measured current image, allowing for various applications in autonomous vehicles. For instance, it can be utilized to determine when pedestrians will cross or move away from my vehicle's driving lane. Moreover, our focus has also been on extracting the necessary contextual information from the raw image to predict the future trajectory of pedestrians. During driving, human drivers predict pedestrian movements by comprehensively considering the visual information, road structures such as nearby crosswalks, sidewalks, and lanes, as well as the past movements of pedestrians in the vicinity. Following this aspects, our study actively utilizes both the visual contextual information and historical trajectories of target based on an cross-attention mechanism of transformer. The transformer model [26], which has made significant advancements in the field of natural language processing (NLP), has also been extended to the domain of computer vision, including tasks such as image classification, object detection, and segmentation. The transformer model is also a suitable framework for prediction tasks, as it effectively utilizes the key image context and handles time sequential data. In this study, cross-attention within the visual transformer architecture is employed to focus on crucial image pixels that have a significant impact on the future trajectory. Moreover, in reference to [12], we employ a non-autoregressive transformer with a learnable query, departing from the time-consuming autoregressive structure of the original transformer.

The proposed model was evaluated on two datasets: the Nuscenes dataset, which contains a large-scale collection of urban driving videos from an egocentric perspective, and an in-house dataset obtained from our vehicle equipped with sensors. The trajectory prediction performance of the proposed model demonstrated superiority over other prediction models in the egocentric perspective. Moreover, the usability of ego-motion independent future trajectories was demonstrated through risk evaluation experiments with nearby pedestrians. The primary contributions of this work can be summarized in three parts.

1) We are the first to apply the transformer's cross-attention mechanism to actively incorporate not only the historical trajectory but also dense information from raw RGB images in the future trajectory prediction task, achieving state-of-the-art performance for prioritized trajectories.
2) To the best of our knowledge, this study represents the first attempt to predict ego-motion independent future trajectories of pedestrians from an egocentric view, specifically focusing on their applicability in autonomous driving.
3) We perform a risk assessment with surrounding pedestrians to showcase the utilization of our predicted ego-motion independent future trajectories in autonomous driving.

The remaining sections of the paper are as follows: Section II describes related studies focusing on image-based prediction approaches. In Section III, we outline the comprehensive model architecture proposed in this research and provide detailed explanations of the methodology. Section IV presents various experiments and ablation studies conducted to evaluate the prediction performance and provide analysis. Lastly, Section V concludes this research.

II. RELATED WORK

A. TRAJECTORY PREDICTION ON BIRD'S EYE VIEW

Most of the pedestrian trajectory prediction research has been conducted using bird's eye views or building views captured in crowded indoor or outdoor squares. Accordingly, a significant amount of research has focused on the interactions with surrounding agents to predict pedestrian movement patterns in such congested spaces. Alahi et al. [1] proposed a Social-LSTM model, which utilizes social pooling techniques to learn the interactions between nearby pedestrians in crowded environments. A Social-GAN [2], which leverages the concept of Generative Adversarial Networks (GAN), was proposed as an extension of the work presented in [1]. This approach incorporates global pooling

to effectively capture interactions among all individuals depicted in an image. Xue [3]'s work focused on predicting multi-destination conditional trajectories using bi-directional LSTM path classification method. Yue et al. [27] integrates dynamics for crowd modeling into a deep learning approach. Multiple research studies [4], [5], [6], [7], [8], [9] have utilized graph-based networks to understand and model interactions involving multiple objects.

Recently, the attention mechanism [26] in transformers has brought significant improvements in long-term prediction performance in the field of NLP, leading to various attempts to apply transformers in trajectory prediction. Yu et al. [10] encoded the interactions with nearby pedestrians within a specific range using a spatial transformer and applied temporal transformers to encode the temporal information of each agent's historical trajectory. GAT [11] constructed interactions between agents and between agents and infrastructure using a sparse graph structure, and captured the most noteworthy interactions based on attention mechanism. Additionally, they enhanced prediction performance by incorporating additional inputs such as satellite maps, semantic maps, traffic signals, and agent position maps. Agentformer [12] proposed a methodology where a single transformer module handles both temporal and spatial information, enabling consideration of the impact of one agent at a specific time on the future state of other agents. This was achieved by flattening the agent information and temporal sequence to apply attention simultaneously. Achaji et al. [13] pointed out the time-consuming nature of using merged attention in Agentformer [12] and proposed a method based on temporal-spatial divided attention. They also utilized a non-autoregressive model based on learnable queries to enable parallel application of Transformers. Liu et al. [14] also addressed the time-consuming nature and error accumulation in auto-regressive models, proposing a non-autoregressive transformer-based prediction model. These transformer-based prediction methodologies have demonstrated superior performance compared to various RNN-based approaches.

On the other hand, Lee et al. [15] demonstrated the first attempt to utilize Conditional Variational Autoencoders (CVAE) for plausible multiple trajectory prediction. Nowadays, multi-modal forecasting [16], [17], [18], [24], [25], [28], [29], [30] has become the dominant approach in the field of trajectory prediction, as it offers more realistic prediction results compared to single-path forecasting. In accordance with this research trend, our research also applies CVAE to derive diverse trajectories, offering applicability for downstream tasks that require probabilistic predictions rather than deterministic trajectories.

B. TRAJECTORY PREDICTION ON EGOCENTRIC VIEW

There have been several studies focused on predicting the future bounding box sequence of pedestrians' bodies in urban driving scenarios using data captured from vehicle-mounted cameras, rather than their future positions in the top view. Bhattacharyya et al. [19] pioneered pedestrian trajectory prediction on images from onboard camera. By incorporating vehicle motion and pedestrians' past trajectories as inputs, Bayesian-LSTM was employed to estimate future trajectories and associated uncertainties. Yagi et al. [20] introduced a convolution-deconvolution framework that considers the vehicle's motion, pedestrian's past trajectory, bounding box scales, and poses to predict pedestrian future trajectories. Quan et al. [23] introduced a modified version of the LSTM model that effectively integrates various inputs, including pedestrian intention, vehicle motion, and global scene information. Makansi et al. [22] predicted reachable locations based on semantic segmentation images and ego-motion. Subsequently, many researchers attempted to use prior destination with historical trajectories to predict pedestrians' future trajectories. Two recent studies [24], [25] have proposed trajectory prediction methods that consider both top view and egocentric view perspectives. Both studies shared similarities in having the prediction models based on CVAE modules and goal-conditioned decoders. Yao et al. [24] specifically proposed a bi-directional decoder with dual forward and backward GRUs based on the ultimate destination. Wang et al. [25] suggested estimating the goal position sequence during the prediction time horizon rather than the final destination.

The studies on urban driving data with an egocentric view have primarily prioritized advanced interpretation of vehicle movements and pedestrians' past trajectories, rather than focusing on sparse interactions among pedestrians in urban driving situation. However, all the aforementioned studies in the egocentric domain focus on predicting future bounding boxes in the future image coordinate system that will be observed at each time step, relying on the future motion of the vehicle. In the context of autonomous vehicles, it is considerably challenging to utilize these predicted results

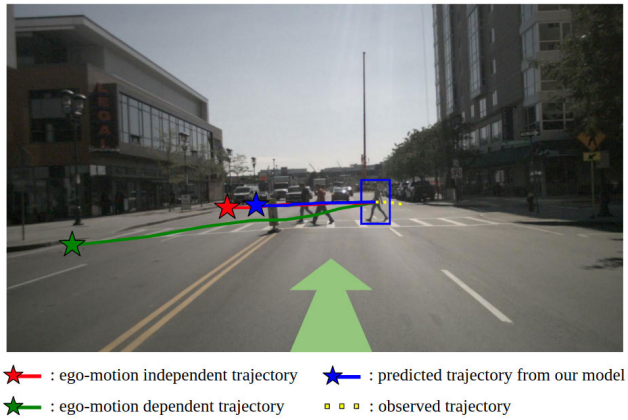

FIGURE 1. Illustration of ego-motion independent trajectory prediction. The past and future trajectories from the dataset are represented by the yellow and green colors, respectively. These trajectories depict the cumulative positions perceived in the image at each time step. The ground truth of the ego-motion independent future trajectory, which we aim to follow, is highlighted in red. The blue color represents the predicted trajectory generated by our proposed model. The final positions are represented by star-shaped markers and the bounding box is used to indicate the current position of the target pedestrian.

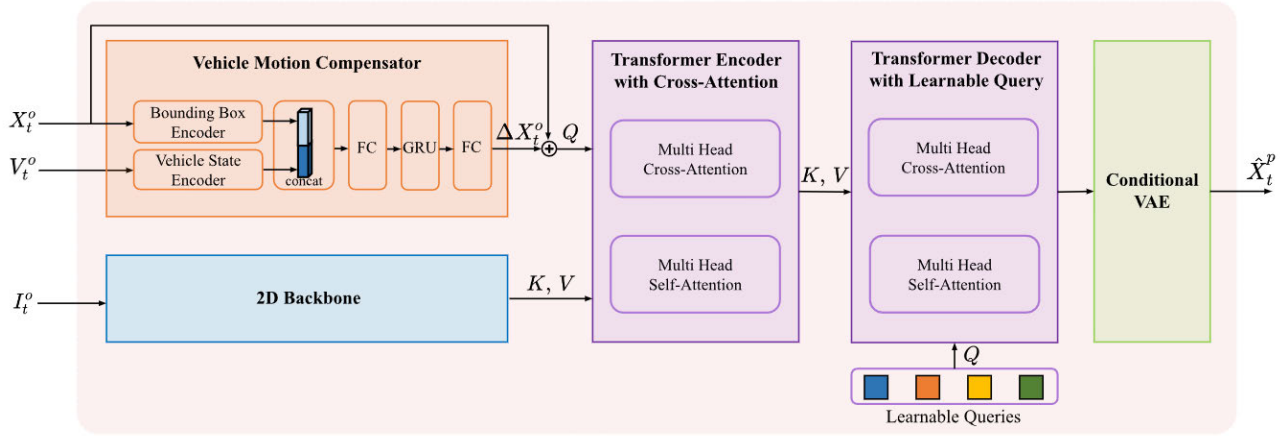

FIGURE 2. Architecture of the proposed model based on a non-autoregressive transformer with learnable queries for pedestrian trajectory prediction.

by comparing them with available information such as lane markings or pedestrian crosswalk in the current observed image. In this study, the future positions of pedestrians on the current image coordinate system are predicted independently of the future vehicle movements. Moreover, the majority of these prior studies have either disregarded the valuable information present in images or relied on processed information derived from images, such as semantic segmentation images, which required additional algorithms for interpretation. In contrast, our proposed approach aims to leverage the rich and unprocessed information directly extracted from the raw images, enabling a more comprehensive and holistic understanding of the scene, facilitating higher-level anticipatory capabilities. To accomplish this, we introduce a non-autoregressive transformer-based predictor that effectively incorporates image scene information and historical trajectory into the prediction process using a cross-attention mechanism.

III. PROPOSED METHOD

The aim of this study is to derive the future trajectories of the target pedestrians within an egocentric view for the N_p prediction horizon, using the past trajectories of target pedestrians X_t^o, subject vehicle states V_t^o, and the last observed raw image I_t^o. Notably, the desired future trajectories in this study, unlike existing works, indicate the anticipated positions of pedestrian within the context of the currently observed image at time t. In other words, these trajectories capture the pedestrian movements independent of the future motions of vehicles, thereby facilitating comprehension of pedestrian behaviors from the perspective of autonomous vehicles. We denote $X_t^o = [x_{t-N_o+1}, x_{t-N_o+2}, \ldots, x_t]$ as the observed past trajectory of the target pedestrian at time t, where $x_t \in \mathbb{R}^4$ represents the center position, height and width of the bounding box in pixel units. The predicted trajectory at time t is denoted as $X_t^p = [x_{t+1}, x_{t+2}, \ldots, x_{t+N_p}]$, and the ground truth of the future trajectory is denoted as $Y_t^p = [y_{t+1}, y_{t+2}, \ldots, y_{t+N_p}]$. The desired ego-motion independent trajectories are denoted with hat symbols as $\hat{X}_t^o$, $\hat{X}_t^p$, $\hat{Y}_t^o$, and $\hat{Y}_t^p$. The ground truth $\hat{Y}_t^o$ and $\hat{Y}_t^p$ for past and future trajectory at time step t are obtained by processing the X_t^o and Y_t^p provided in the dataset. This procedure is addressed in Section III-F.

A. OVERVIEW

The proposed model for obtaining the final output $\hat{X}_t^p$ consists of three key modules: the vehicle motion compensator, the transformer-based trajectory predictor, and the conditional variational autoencoder (CVAE). Initially, the past trajectory X_t^o is transformed into a compensated trajectory $\hat{X}_t^o$ using the trained vehicle motion compensation module. For instance, in the case of a stationary pedestrian, the observed historical trajectory X_t^o may not align with the same coordinate positions throughout due to the past motions of the subject vehicle. However, if the model is well trained, ideally, the trajectory X_t^o will be transformed into a trajectory $\hat{X}_t^o$ that aligns with the current bounding box in the current image coordinate system, represented by x_t, after passing through the vehicle motion compensation module. Subsequently, $\hat{X}_t^o$ is utilized as input for the trajectory prediction module. The trajectory predictor is based on a transformer structure, which is composed of an encoder and a decoder. In the encoder, a cross-attention mechanism is employed to effectively integrate the encoded image features from the 2D backbone and the historical trajectory. The decoder then generates hidden features for the future trajectory of the target pedestrian. Lastly, the CVAE module incorporates latent variables to learn the distribution of the target trajectory and uses this distribution to predict K multiple trajectories. The overall architecture of the proposed model is illustrated in Fig. 2.

B. VEHICLE MOTION COMPENSATION

The displacement $\triangle X_t^o$ of 2D bounding boxes on the image, caused by the vehicle's motion, is estimated through the vehicle motion compensator. There are limitations in finding

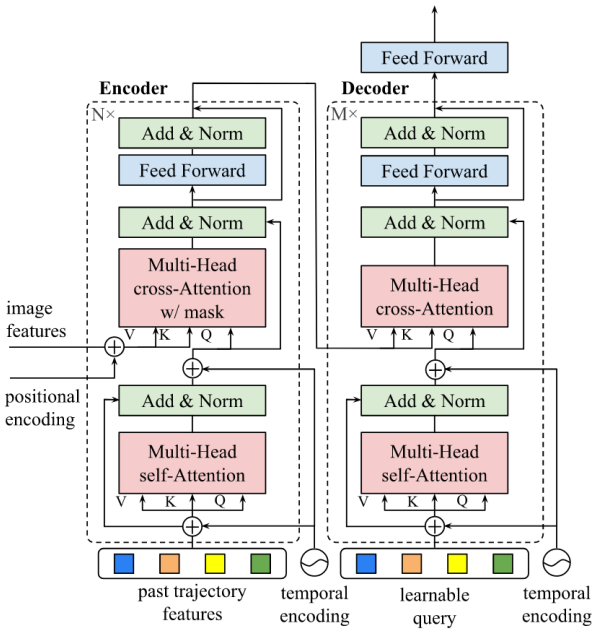

FIGURE 3. Architecture of the modified transformer module with cross-attention mechanism to effectively fuse the historical trajectory and image feature.

a unique solution for the displacement of dynamic target objects in the dynamic image coordinate system captured by a moving camera through purely algebraic methods. Therefore, the proposed vehicle motion compensation module is employed to estimate displacements based on a learning methodology that allows for the incorporation of more intricate relationships.

The observed past trajectory X_t^o and the vehicle state sequence V_t^o, which contains information on the vehicle's speed and yaw rate during the observation time horizon N_o, serve as inputs to the vehicle motion compensation module. They are individually encoded by the bounding box encoder and the vehicle state encoder. The compensation module is composed of a combination of fully connected layers and a GRU, forming its overall architecture. The derived $\triangle X_t^o$ through this module is added to the observed trajectory X_t^o, resulting in the transformation of the observed trajectory X_t^o into the compensated observed trajectory $\hat{X}_t^o$. In other words, the target pedestrian bounding boxes at each image coordinate during the time steps $[t - N_o + 1, \ldots, t - 1]$ are transformed to their corresponding positions in the current image coordinates at time t. The compensated observed trajectory $\hat{X}_t^o$ can be defined by the following equation.

$$\hat{X}_t^o = X_t^o + \triangle X_t^o \tag{1}$$

The computed $\hat{X}_t^o$ is then used as input to the transformer-based trajectory predictor with image features.

C. TRANSFORMER-BASED TRAJECTORY PREDICTION

To construct a trajectory prediction that comprehensively incorporates both the compensated past trajectory and

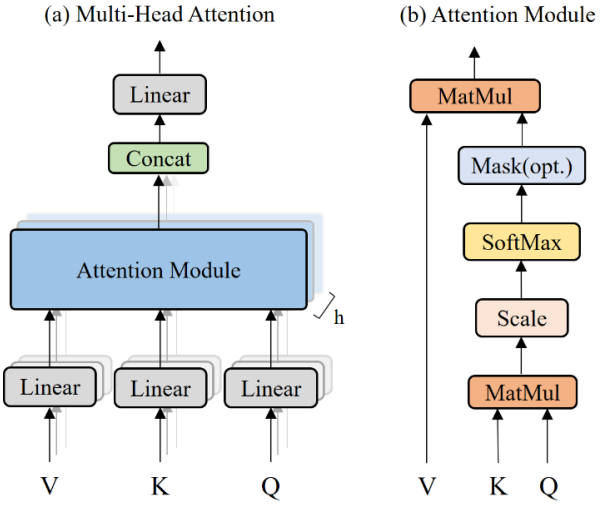

FIGURE 4. Architecture of the multi-head attention module.

the image pixels containing crucial contextual information, we present a modified transformer architecture. This modification involve adapting the transformer structure to effectively integrate the compensated past trajectory and the relevant image pixels, which are vital for predicting future paths. The proposed structure of modified transformer is shown in Fig. 3. In the transformer encoder, the encoded past trajectory $\hat{X}_t^o$ is treated as query $Q \in \mathbb{R}^{N_o \times d_m}$, where N_o each query element represents the encoded bounding box at each time step within the observation horizon. To introduce temporal information to the query sequence, each query element is added with an encoded temporal value. The temporal encoder follows a sinusoidal design, similar to the original transformer [26], to extract the timestamp feature. The timestamp feature $\tau^t \in \mathbb{R}^{d_m}$ for each time step t is computed as follows:

$$\tau_j^t = \begin{cases} sin(t/10000^{j/d_m}) & \text{for j even} \\ cos(t/10000^{(j-1)/d_m}) & \text{for j odd} \end{cases} \tag{2}$$

where the notation τ_j^t refers to the j-th feature of τ^t with a feature dimension of d_m. On the other hands, the encoded image feature serves as the key-value $K, V \in \mathbb{R}^{H \cdot W \times d_m}$ for the transformer encoder, incorporating learned positional encoding, following the approach proposed in [31]. Additionally, taking inspiration from [31] and [32], a 2D Gaussian weight mask is employed to efficiently identify influential key pixels near the past trajectory within the observed image space, encompassing a broader range. The mask $M \in \mathbb{R}^{H \cdot W \times N_o}$ is represented by the expression below.

$$M_{uv,t} = exp\left(-\frac{(u - c_{x,t})^2 + (v - c_{y,t})^2}{\sigma r^2}\right) \tag{3}$$

where (u, v) represents the indices of image pixels, and $(c_{x,t}, c_{y,t})$ denotes the center point of the bounding box at t time step of compensated past trajectory $\hat{X}_t^o$. r represents the distance from the center point of the bounding box to one of

its corner points, and σ is a hyper-parameter used to control the extent of the masked region. The σ has been configured as shown below.

$$\sigma = 4(2r + 1) \tag{4}$$

The multi-head cross-attention module, including the weight mask, can be expressed mathematically as follows:

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$$
$$Attention(Q, K, V) = (softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \odot M)V \tag{5}$$

Each query, key, and value is projected using $W_i^Q \in \mathbb{R}^{d_m \times d_k}$, $W_i^K \in \mathbb{R}^{d_m \times d_k}$, and $W_i^V \in \mathbb{R}^{d_m \times d_k}$, respectively, and then applied to the attention mechanism. The masked attention weights are obtained by element-wise multiplication of the attention weights and the 2D weight mask, and then these masked attention weights are multiplied by the V.

$$MultiHead = Concat(head_1, \ldots, head_h)W^O \tag{6}$$

The results from h heads are concatenated, and a projection matrix $W^O \in \mathbb{R}^{h \cdot d_k \times d_m}$ is applied to ensure that the output has the same size as the input query. Equations (5) and (6) are visualized in Fig. 4.

The derived output of transformer encoder is utilized as the key-value input for the transformer decoder component. In the transformer decoder, the ultimate objective is to derive the hidden features of future trajectory over the prediction time horizon N_p based on the hidden features of length N_o, which encapsulate the past trajectory and image context information. The original transformer [26] operates in an autoregressive manner, which not only results in significant time consumption on inference but also introduces exposure bias due to the disparity between the training process that employs ground truth and the inference process that relies on previous predictions instead. Furthermore, if both training and inference operate autoregressive manner applying previous predictions, it hampers the utilization of parallelization power during training, resulting in inefficient learning time. Therefore, taking inspiration from [13], we adopt a learnable query with a temporal dimension of N_p. By incorporating the learnable query, parallel decoding can be performed during both training and inference, effectively addressing the aforementioned issues. The learnable query is randomly initialized and serves as the query input to the multi-head cross-attention module after passing through self-attention. The multi-head cross-attention is conducted in a similar manner to the encoder, with the exception that no mask is applied. The final hidden feature derived from the decoder undergoes a CVAE process to facilitate diverse multi-trajectory prediction.

D. CONDITIONAL VARIATIONAL AUTOENCODER (CVAE)

In this study, the CVAE is employed as an additional module to derive plausible multi-trajectories for safe urban driving. The CVAE module is designed to learn a target distribution by introducing latent variables for one-to-many modeling. In our CVAE module, the hidden feature of the future trajectory $h_{\hat{x}_t^p}$ obtained from the transformer-based trajectory predictor is used to conditionally model the distribution of the ground truth trajectory $\hat{Y}_t^p$. In accordance with references [15], the CVAE module is composed of three parts: the recognition network $Q_\phi(z|h_{\hat{x}_t^p}, \hat{Y}_t^p)$, the prior network $P_\psi(z|h_{\hat{x}_t^p})$, and the generation network $P_\theta(\hat{X}_t^p|h_{\hat{x}_t^p}, z)$. Here, ϕ, ψ, and θ represent the parameters of the three respective networks. The recognition network is composed of a combination of GRU and fully-connected layers. On the other hand, the other networks are solely comprised of fully-connected layers. All distributions are assumed to be Gaussian, and during training, samples are drawn from the distribution $N(\mu_z^q, \sigma_z^q)$ obtained from the recognition network. During inference, samples are drawn from the prior network. To reconstruct the distribution from recognition network, the recognition network, and the prior network are trained to minimize the difference between the Gaussian distributions derived from each network. Ultimately, K diverse future trajectories $\hat{X}_t^p$ can be obtained by passing through the generation network, where K represents the number of samples.

E. LOSS FUNCTION

The overall loss function consists of three components: trajectory prediction loss, compensation loss, and KLD loss. The trajectory prediction loss measures the error between the predicted trajectories $\hat{X}_t^p$ from the proposed model and the target trajectories $\hat{Y}_t^p$. The trajectory prediction loss is formulated using the Best-of-Many (BoM) approach, as referenced in [19], [24] and [25], where the best prediction among k multiple trajectories derived from the stochastic prediction model is selected. The loss is calculated using the L2 norm. The formulation of the trajectory prediction loss can be expressed as follows:

$$L_{pred} = \min_{\forall k \in K} \|\hat{X}_t^{p,k} - \hat{Y}_t^p\|_2 \tag{7}$$

The compensation loss is derived by comparing the compensated observed trajectory $\hat{X}_t^o$ with the target trajectory $\hat{Y}_t^o$ using the L2 norm. Accordingly, the compensation loss can be formulated simply as shown below.

$$L_{comp} = \|\hat{X}_t^o - \hat{Y}_t^o\|_2 \tag{8}$$

The KLD loss captures the discrepancy between the target Gaussian distribution $N(\mu_z^q, \sigma_z^q)$, represented by Z_q, derived from the recognition network of the CVAE module and the distribution $N(\mu_z^p, \sigma_z^p)$, represented by Z_p, derived from the prior network. The overall loss, including the KLD loss, is given as follows.

$$L_{total} = L_{pred} + L_{comp} + D_{KL}(Z_q||Z_p) \tag{9}$$

F. GROUND TRUTH TRAJECTORY GENERATION

Since this research is based on urban driving data from a camera-equipped vehicle, the camera coordinates change

at each time step with respect to the world coordinate system. According to the research objective, the ground truth trajectories required for training are the ego-motion compensated past and future trajectories of the target pedestrian, represented as $\hat{Y}_t^o = [\hat{y}_{t-N_o+1}, \hat{y}_{t-N_o+2}, \ldots, \hat{y}_t]$ and $\hat{Y}_t^p = [\hat{y}_{t+1}, \hat{y}_{t+2}, \ldots, \hat{y}_{t+N_p}]$ on t image coordinate. The $\hat{y}_{t+i} \in \mathbb{R}^4$ for $i \in [-N_o + 1, N_p]$ represents the compensated center position and size of the bounding box of the target pedestrian observed at time $t+i$, with respect to the image coordinate system observed at time t. The conventional approach, feature descriptor [33], [34], [35] and matching [36], [37], can be used to transform images captured at different times and locations within the same spatial context into a unified coordinate system. However, this method is both time-consuming and prone to inaccuracies, particularly in the presence of dynamic objects in the images. In order to achieve a higher degree of precision in the transformation process, it is necessary to utilize the positional information of the target pedestrian in the physical 3d space.

In this research, the model was trained and evaluated using both the NuScenes dataset and our own dataset. Due to the differences in the information provided by these two datasets, the ground truth was generated in slightly different ways for each dataset. In the case of the NuScenes dataset, precise information regarding the 3d bounding boxes of the target pedestrians in the world coordinate system, as well as the global location of the ego-vehicle, is provided. Therefore, the center of the global 3d bounding box p_{ped}^g within the range of $[t - N_o + 1, t + N_p]$ is transformed into a local coordinate system based on the global coordinates of the ego-vehicle at time t, represent as $p_{veh,t}^g$.

$$p_{ped,t+i}^{O_v(t)} = p_{ped,t+i}^g - p_{veh,t}^g \quad i \in [-N_o + 1, N_p] \tag{10}$$

where $p_{ped,t+i}$ represents the center position of the pedestrian's bounding box observed at time $t+i$ in the 3d coordinate system, while $O_v(t)$ denotes a local coordinate system with the ego-vehicle's coordinates at time t as the reference point. Subsequently, by utilizing the extrinsic and intrinsic matrix of the camera sensor in the ego-vehicle coordinate system, the bounding box is projected onto the t image plane, allowing us to obtain the desired ground truth, $\hat{Y}_t^o$ and $\hat{Y}_t^p$. In mathematical notation, it can be represented as follows.

$$\hat{y}_{t+i} = f(K[R|t]b_{ped,t+i}^{O_v(t)}) \quad i \in [-N_o + 1, N_p] \tag{11}$$

where K and $[R|t]$ are the camera intrinsic and extrinsic matrix provided in Nuscenes dataset, respectively. The $b_{ped,t+i}$ represents the corner position of the bounding box observed at time $t + i$ in the 3d coordinate system. $f(\cdot)$ represents the post-processing step that derives a bounding box vector consisting of the 2D center position, height, and width in the image from the projected corner positions.

On the other hand, for our self-collected dataset, the global coordinates of the ego-vehicle and the target pedestrian are not available. Instead, we can utilize the local position of the target pedestrian with respect to the vehicle's coordinate system, obtained from the LIDAR sensor mounted on the vehicle, along with the subject vehicle motion states from the vehicle's chassis sensor. In order to obtain the ground truth $\hat{Y}_t^o$ and $\hat{Y}_t^p$, vehicle motion compensation in 3D space is required using vehicle motion states, specifically the velocity and yaw rate. The compensation process entails transforming the 3d bounding box sequence of the target pedestrian in the ego-vehicle coordinates system within the range of $[t-N_o+1, t+N_p]$, based on the motion information of the vehicle, to the positions in the ego-vehicle coordinate system at time t. The process can be expressed mathematically as follows. Here, the position z value of the bounding box is not considered.

$$p_{ped,t+i}^{O_v(t)} = \begin{cases} p_{ped,t}^{O_v(t)} & \text{for i = 0} \\ (T_{t+1}^t T_{t+2}^{t+1} \cdots T_{t+i}^{t+(i-1)}) p_{ped,t+i}^{O_v(t+i)} & \text{for i} \in [1, N_p] \\ (T_{t-1}^t T_{t-2}^{t-1} \cdots T_{t+i}^{t+(i+1)}) p_{ped,t+i}^{O_v(t+i)} & \text{for i} \in [-N_o + 1, -1] \end{cases} \tag{12}$$

$$T_{t+i}^{t+(i-1)} \left(p_{ped,t+i}^{O_v(t+i)}\right) = \begin{bmatrix} cos(\gamma^{t+i}dt) & sin(\gamma^{t+i}dt) \\ -sin(\gamma^{t+i}dt) & cos(\gamma^{t+i}dt) \end{bmatrix} p_{ped,t+i}^{O_v(t+i)} + \begin{bmatrix} -\frac{v^{t+i}}{\gamma^{t+i}} sin(\gamma^{t+i}dt) \\ \frac{v^{t+i}}{\gamma^{t+i}} - \frac{v^{t+i}}{\gamma^{t+i}} cos(\gamma^{t+i}dt) \end{bmatrix} \tag{13}$$

$$T_{t+i}^{t+(i+1)} \left(p_{ped,t+i}^{O_v(t+i)}\right) = \begin{bmatrix} cos(\gamma^{t+i}dt) & -sin(\gamma^{t+i}dt) \\ sin(\gamma^{t+i}dt) & cos(\gamma^{t+i}dt) \end{bmatrix} p_{ped,t+i}^{O_v(t+i)} + \begin{bmatrix} \frac{v^{t+i}}{\gamma^{t+i}} sin(\gamma^{t+i}dt) \\ \frac{v^{t+i}}{\gamma^{t+i}} - \frac{v^{t+i}}{\gamma^{t+i}} cos(\gamma^{t+i}dt) \end{bmatrix} \tag{14}$$

where $p_{ped,t+n}^{O_v(t+m)}$ represents the center x, y position of the pedestrian's bounding box observed at the time $t + n$ in the vehicle coordinate system at time $t + m$. T_{t+n}^{t+m} represents the transformation from the coordinate system at time $t + n$ to the coordinate system at time $t + m$. Additionally, γ and v denote the yaw rate and velocity of the subject vehicle, respectively. Once the ego-motion is compensated, the same process is applied to project the compensated 3D bounding box of the pedestrian onto the image plane in the vehicle coordinate system using the camera sensor's extrinsic matrix and intrinsic matrix. The resulting ground truth $\hat{Y}_t^o$ and $\hat{Y}_t^p$ are then utilized to train model and evaluate performances.

TABLE 1. Trajectory prediction results of multi-modal models on nuscenes and in-house datasets. the performance of the proposed model is shown at the bottom, and the best results are highlighted in bold.

k	Methods	Nuscenes			In-house Dataset		
		ADE (1.0s/2.0s/3.0s/4.0s)	C_{ADE} (4.0s)	C_{FDE} (4.0s)	*ADE* (1.0s/2.0s/3.0s/4.0s)	C_{ADE} (4.0s)	C_{FDE} (4.0s)
1	LSTM-CVAE	423/1154/2243/3684	3671	9030	223/509/998/1789	1762	4863
	BiTraP-NP [24]	375/1090/2144/3507	3491	8418	201/480/917/1446	1423	3177
	SGNet-ED [25]	275/838/1739/2935	2888	7222	170/424/811/1349	1325	3303
	TransPred	**146/403/820/1419**	**1404**	**3851**	**52/71/97/183**	**168**	**620**
2	LSTM-CVAE	361/1083/2161/3567	3554	8223	178/455/925/1623	1596	3597
	BiTraP-NP [24]	187/528/1015/1641	1627	3883	110/245/476/796	775	1846
	SGNet-ED [25]	172/501/1001/1675	1648	4097	85/199/369/600	578	1365
	TransPred	**136/368/731/1203**	**1188**	**2067**	**44/62/85/149**	**134**	**356**
20	LSTM-CVAE	178/808/1802/3157	3144	6489	64/285/690/1273	1246	1774
	BiTraP-NP [24]	50/123/210/317	302	625	27/50/85/158	139	310
	SGNet-ED [25]	**42/108/192/296**	**283**	**614**	21/**33/49/79**	**60**	**99**
	TransPred	86/244/447/802	786	1209	**20**/37/57/94	80	142

TABLE 2. Trajectory prediction results of deterministic models on nuscenes and in-house datasets. the performance of the proposed model is shown at the bottom, and the best results are highlighted in bold.

Methods	Nuscenes			In-house Dataset		
	ADE (1.0s/2.0s/3.0s/4.0s)	C_{ADE} (4.0s)	C_{FDE} (4.0s)	*ADE* (1.0s/2.0s/3.0s/4.0s)	C_{ADE} (4.0s)	C_{FDE} (4.0s)
LSTM	231/661/1294/2169	2149	5421	140/369/810/2078	2048	7691
BiTraP-D [24]	241/680/1284/2047	2036	4838	87/218/463/997	971	3250
SGNet-ED [25]	227/633/1207/1946	1934	4654	74/171/331/611	593	1729
TransPred-D	**196/539/1038/1698**	**1686**	**4129**	**53/97/169/270**	**253**	**641**

IV. EXPERIMENTS

A. DATASETS

In this study, the model's performance was evaluated using both the Nuscenes dataset and an in-house dataset. Both datasets consist of urban driving data, providing ego-centric view video data with 1600 × 900 resolution. The Nuscenes dataset is a large-scale dataset that provides diverse scene data with a duration of 20 seconds. It offers annotations for the object class, location, and size information of surrounding targets, as well as the states of the ego-vehicle, including its position, velocity, acceleration, and yaw rate, at a frequency of 2Hz. However, since the Software Development Kit (SDK) provided by Nuscenes is primarily focused on vehicle prediction tasks, additional preprocessing was performed to extract scenes of pedestrian appearance for pedestrian prediction. As for the in-house dataset, consecutive scenes were captured at a rate of 20Hz and included front camera video, chassis data, and LiDAR point clouds for generating ground turth. During the training and inference processes of the in-house dataset, the data was downsampled to 2Hz, consistent with the Nuscenes dataset. The trajectory data from the in-house dataset and the NuScenes dataset are preprocessed by shifting them one timestep at a time to create overlapping data, in order to increase the dataset size. The total number of samples for each dataset was 23,916 and 4,425, respectively, with approximately 20% allocated as test data. Due to the limited amount of in-house dataset, a transfer learning method was employed using a model pre-trained on the large-scale Nuscenes dataset for training on the in-house dataset.

B. IMPLEMENTATION DETAILS

All experiments were conducted using a single GPU setup with a NVIDIA GeForce RTX 3090 Ti graphics card. The proposed model has an observation length, denoted as N_o, set to 4, allowing for a 2-second time horizon in 2Hz data. The prediction length N_p is set to 8, allowing for a 4-second time horizon. Considering the usability in the downstream task, we present a longer term prediction than previous studies. For the transformer-based predictor, all feature dimensions, including d_m and d_k, are set to 128 and the number of multi-heads, h, is set to 8. The CVAE module utilized a sample count of 20 (K=20). Resized images to 640 (W) × 360 (H) were used as inputs to the model in order to handle the computational load. On the other hands, the prediction results were computed at the original resolution of 1600 (W) × 900 (H). To encode image features, we utilized pre-trained ResNet50 and FPN models provided by MMDetection3D as 2D backbones. The provided model was pre-trained on instance segmentation tasks using the nuImages dataset. Additionally, the bounding box encoder and vehicle state encoder were composed of a combination of GRU and fully connected (FC) layers. The transformer encoder and decoder in our model were constructed with a depth of

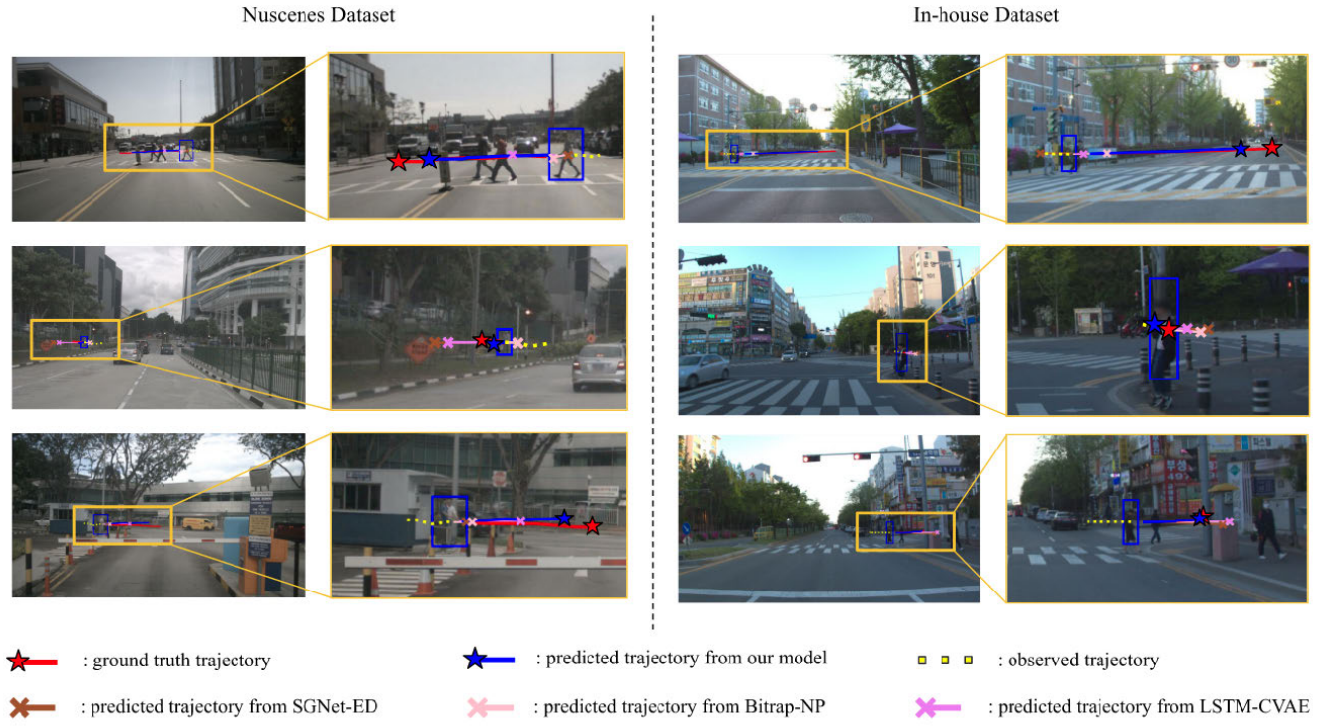

FIGURE 5. Qualitative comparisons of trajectory prediction models. The past and future trajectories from the dataset are represented by the yellow and green colors, respectively. These trajectories depict the cumulative positions perceived in the image at each time step. The ground truth of the ego-motion independent future trajectory, which we aim to follow, is highlighted in red. The blue color represents the predicted trajectory generated by our proposed model. The final positions are represented by star-shaped markers and the bounding box is used to indicate the current position of the target pedestrian. Additionally, the predicted trajectories of the comparison models are displayed as lines with 'X' marks at the final predicted positions.

2 and 1, respectively. For our in-house dataset, transfer learning was performed by unfreezing the final prediction head, input encodings and vehicle motion compensation from a pre-trained model on the NuScenes dataset. The proposed model underwent training with batch size 20 for 50 epochs, employing an initial learning rate of 5×10^{-4}. A reduction of 0.8 in the learning rate was applied when there was no improvement observed in the test loss during 5 epochs.

C. EXPERIMENTS FOR TRAJECTORY PREDICTION

1) OUTLINE

In this section, the performance of the proposed prediction network is compared to other recent methods on both the Nuscenes dataset and an in-house dataset. The comparison models include multi-modal predictors SGNet [25] and BiTrap [24], which have demonstrated state-of-the-art performance in pedestrian path prediction using vehicle-mounted camera data from prominent datasets such as PIE and JAAD. Additionally, a simplified model consisting of a single LSTM and FC layer, augmented with a CVAE for multi-modal trajectory generation, is used as the baseline comparison model. As a side note, the JAAD and PIE datasets do not provide annotations for positions of objects in the physical 3d space. Therefore, it is not possible to obtain the ground truth of the ego-motion independent future trajectory proposed in this paper, rendering them inapplicable for this study.

The objective of this paper is to derive future trajectories that are independent of ego-motion, distinguishing it from previous studies. Therefore, we reproduce the performance results of two state-of-the-art models to align with our specific task. The model we propose is mainly composed of the ego-motion compensation, transformer-based predictor, and CVAE modules, where the ego-motion compensation module is responsible for generating ego-motion independent trajectories. This module can operate independently and be attached to other models for utilization. Therefore, we integrated this module into the front end of the SGNet [25] and BiTrap [24] models to derive ego-motion independent trajectories and evaluated their performance through re-training on the NuScenes dataset. By doing so, we were able to compare the performance of predictor with that of the state-of-the-art models under equivalent settings and task.

To facilitate downstream tasks such as motion planning, it is essential to prioritize and find the most probable future trajectory among the diverse predicted paths. Given the consideration of all diverse predictive trajectories, vehicle might exhibit overly cautious movements. Hence, numerous motion planning methodologies [38], [39], [40], [41], [42], [43], [44], [45], [46] adopt a prioritized prediction outcome. Given our study's emphasis on applications in autonomous driving, we concentrate on presenting results centered around the performance of the selected highest-priority or second-highest priority trajectories among the $K = 20$ predicted

trajectories, instead of the performance based on the lowest error compared to the ground truth among the K trajectories as in other studies [24], [25]. The selection process uses a K-means clustering methodology referring to previous studies [47], [48], [49]. Given that the CVAE module generates K prediction trajectories from the gaussian distribution, the largest trajectory cluster can be considered the dominant cluster having a higher probability. We cluster the predicted trajectories into five groups, and then the average trajectories of the largest or second largest clusters were used to derive the results. Moreover, in order to clearly demonstrate the superiority of our approach in the unimodal aspect, we have also incorporated a comparison of the performance of deterministic versions, which inherently generate a single trajectory, with other state-of-the-art models. Following the approach of SGNet and BiTrap, the deterministic version is obtained by removing the CVAE module.

2) METRICS

To assess the prediction performance, we followed the evaluation metrics utilized in previous studies [22], [24], [25], [50], specifically the Average Displacement Error (*ADE*) and Final Displacement Error (*FDE*). The *ADE* metric provides the average error between the predicted and ground truth trajectories over a specified prediction time horizon. On the other hand, the *FDE* metric quantifies the positional error of the final point of the predicted trajectory. We computed the *ADE* values for different time steps, including 1.0, 2.0, 3.0, and 4.0 seconds into the future from the current time t. The bounding box's upper-left and lower-right pixel coordinates were used to calculate the errors. Additionally, we also computed the *ADE* and *FDE* based on the center pixel coordinate of the bounding boxes, denoted as C_{ADE} and C_{FPE}. The errors were calculated using the Mean Square Error (MSE) between the target trajectory and the predicted future trajectory. The smallest error value among the selected k paths from the K generated predictive paths through sampling is recorded as the performance.

3) RESULTS

The experimental results for $k = 1$, 2 and 20 are summarized in Table 1. Although it slightly lags behind in the prediction performance for $k = 20$ without considering the priority of trajectories, the proposed model, TransPred, outperforms the compared models in terms of prediction accuracy for the selected highest-priority or second-highest priority trajectories on both datasets. Furthermore, as indicated in Table 2, our model exhibited superior performance in the deterministic prediction of deriving a single future trajectory. The results indicate that while existing state-of-the-art studies hold a stronger position from a diversity perspective, the proposed TransPred model exhibits superior predictive performance for a prioritized trajectory compared to other models. This establishes its proficiency not only in predicting performance but also its potential strengths in downstream tasks.

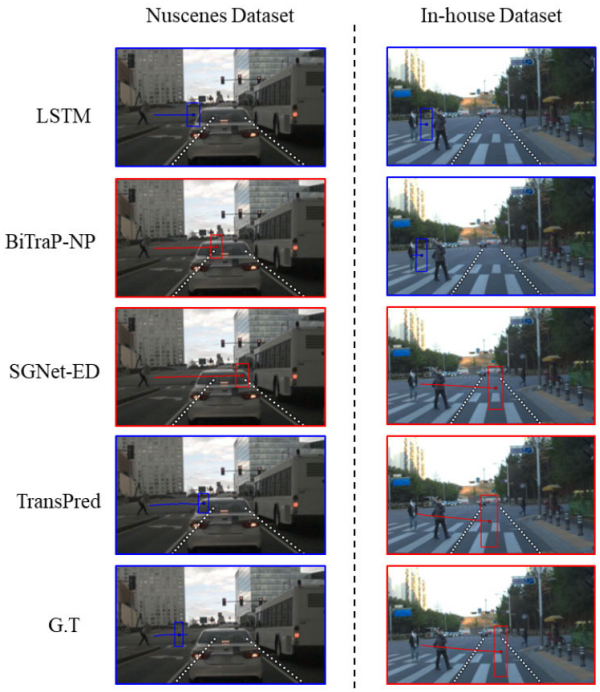

FIGURE 6. Qualitative comparisons of trajectory prediction models on risk assessment experiment. The predicted trajectories are depicted as lines, and the pedestrian's bounding box at the final prediction time is indicated. Blue indicates cases classified as non-risky situations, while red represents the results classified as risky situations.

Additionally, the results also demonstrate the effectiveness of our modified transformer-based prediction model, which leverages the cross-attention module to appropriately fuse information from the raw image and historical trajectories for advanced prediction. Further analysis into the contributions of each component to the prediction performance is discussed in detail through the ablation study. The qualitative results are presented in Fig. 5, illustrating a prioritized trajectory on multi-modal prediction in example scenes. As depicted in the Fig. 5, the proposed model successfully predicts ego-motion independent future trajectories from the ego-motion dependent historical trajectory as input. Moreover, the presented figures demonstrate that our proposed model consistently adheres better to the ground truth trajectory compared to other comparative models.

D. EXPERIMENTS FOR RISK ASSESSMENT

1) OUTLINE

One of the notable aspects of this study is its practicality in applying pedestrian path prediction in autonomous driving. Therefore, in this section, we demonstrate the practicality by conducting experiments that assess the risk from surrounding pedestrians using the predicted ego-motion independent trajectories in urban driving scenarios. As mentioned above, it is challenging to use ego-motion-dependent future trajectories derived from existing prediction studies to assess pedestrian encroachment in hazardous areas, such as within the ego-lane, on the current observed image. However, it becomes feasible

TABLE 3. Risk prediction results on nuscenes and In-House dataset.

Datasets	Methods	Balanced Accuracy	Recall	F2-score
Nuscenes	LSTM-CVAE	0.82	0.68	0.69
	BiTraP-NP [24]	0.80	0.63	0.65
	SGNet-ED [25]	0.82	0.65	0.67
	TransPred	**0.89**	**0.81**	**0.77**
In-house Dataset	LSTM-CVAE	0.86	0.74	0.79
	BiTraP-NP [24]	0.87	0.79	0.81
	SGNet-ED [25]	0.90	0.82	0.83
	TransPred	**0.93**	**0.85**	**0.87**

by utilizing the ego-motion independent future trajectories obtained from our proposed model. We evaluated the risk by determining whether the target pedestrians enter the ego-vehicle's driving lane based on their predicted trajectories during prediction time horizon. The entrance status is determined by evaluating whether the ratio of the encroached width to the predicted bounding box width exceeds a pre-defined threshold. The ground truth for risk assessment is also obtained by comparing the ground truth of future trajectory $\hat{Y}_t^p$ with the lanes obtained on the last observed image at t time step.

This experiment is conducted on the nuscenes dataset and an in-house dataset. For the nuscenes dataset, the lane information is provided through HD map and is projected onto the image for use. On the other hand, for the in-house dataset, lane information is obtained using CondLaneNet [51], one of the various approaches [51], [52], [53], [54], [55], [56] for lane detection in images, and for scenes with ambiguous lanes, we performed manual labeling. The risk prediction results of the comparison models were computed based on a prioritized trajectory of multi-modal results.

2) METRICS

The outcome of the risk assessment is provided as binary results indicating whether the situation is risky or not. However, it appears that there is an imbalance in the data labels, with a much larger number of non-risky situations compared to risky situations. Accordingly, we use balanced accuracy as a main performance metric. Furthermore, from the perspective of autonomous driving, misclassifying risky pedestrians as non-risky is a much more dangerous situation than the opposite. Therefore, recall and F2 score are used as additional metrics. Recall represents the number of positive instances correctly predicted among the actual positive data, while the F2 score is a comprehensive performance metric widely used in tasks where recall holds more weight. It takes into account both precision and recall, with a higher emphasis on recall, making it suitable for evaluating the performance in situations where correctly identifying positive instances is crucial.

TABLE 4. Ablation results on image input and attention weight mask.

Datasets	Raw Image	Weight mask	C_{ADE} (4.0s)	C_{FDE} (4.0s)
Nuscenes	×	×	2481	6012
	√	×	2215	5995
	√	√	1404	3851
In-house Dataset	×	×	893	1933
	√	×	476	1385
	√	√	168	620

3) RESULTS

The results are summarized in Table 3. The proposed model, TransPred, exhibits the highest risk evaluation performance compared to other models in terms of all metrics for both datasets. Based on intuitive recall metrics, we observe a high positive prediction performance of over 80% for both datasets. In the case of F2 score, all models show relatively low performance on the nuscenes dataset, which can be attributed to the imbalance in the data leading to generally low precision. However, even in such a situation, our model demonstrates the highest performance. These results demonstrate the applicability of ego-motion-independent predicted trajectories in autonomous driving and showcase the superiority of our proposed model. The qualitative results of risk assessment, along with comparisons to other models, can be observed in Fig. 6.

E. ABLATIONS AND ANALYSIS

1) IMAGE INPUT EXCLUSION

In this research, the raw image was used as input without any additional pre-processing such as segmentation. The transformer architecture was employed to encode the historical trajectory and raw image information, allowing for effective feature encoding and accurate prediction of future trajectories. This section focuses on assessing the contribution of image features to the model's performance. Additionally, the experiment investigates the role of the attention weight mask applied near the target pedestrian's location. The results are shown in Table 4. Firstly, when no image information was used at all, the model exhibited a significant performance decrease, with the error increasing by approximately 1.8 times for the nuScenes dataset and 5.3 times for the in-house dataset compared to the complete model based on the C_{ADE} metric. This result demonstrates that the proposed transformer architecture effectively encodes key points from raw images, which are essential for predicting future trajectories. Furthermore, without the weight mask alone, the error increased by approximately 1.6 times for the nuScenes dataset and 2.8 times for the in-house dataset compared to the complete model based on the C_{ADE} metric. In the case of the nuScenes dataset, there was little difference in performance between not using images at all and excluding only the weight mask. This indicates that in the Nuscenes

TABLE 5. Comparison of inference time with prediction performance.

Methods	C_{ADE} (4.0s)	C_{FDE} (4.0s)	Inference Time (ms)
LSTM-CVAE	3671	9030	3.2
BiTrap-NP [24]	3491	8414	4.9
SGNet-ED [25]	2888	7222	27.1
TransPred w/o image	2481	6012	10.1
TransPred	1404	3851	59.8

TABLE 6. Comparison of performance and inference time based on the input image resolutions.

Image Resolution	C_{ADE} (4.0s)	C_{FDE} (4.0s)	Inference Time (ms)
480 × 270	1693	4434	38.0
640 × 360	1404	3851	59.8
960 × 540	1408	3607	119.2
1600 × 900	1306	3556	318.4

dataset, which provides video data with a relatively wide field of view (FOV), the attention weight mask plays a crucial role in effectively encoding relevant image information by focusing attention on the areas near the target pedestrian within the wide coverage area.

2) INFERENCE TIME

In the context of autonomous vehicles, not only the accuracy of predictions but also real-time performance is crucial. Table 5 presents the measured inference times for 10 samples. We selected a sample size of 10 based on the assumption that in our dataset, there were no more than 10 pedestrians present in a single image frame, and it is generally expected that the maximum number of pedestrians would be around 10. In the results table, the LSTM-CVAE and BiTrap-NP [24] models demonstrated very fast inference times below 10ms, but they exhibited lower performance. On the other hand, SGNet-ED [25] demonstrated slightly higher performance compared to the previous two models. Nevertheless, it should be noted that SGNet-ED [25] demonstrated inferior performance in comparison to our model, even including the version without image inputs. Additionally, SGNet-ED [25] reported an inference time roughly 2.7 times slower when compared to our image-less model. Our complete model, which utilizes dense image information, comes with a relatively high computational load. However, we have mitigated the temporal overhead through techniques such as image resizing and the non-autoregressive structure of the transformer. As a result, an inference time of approximately 17Hz was achieved, which is considerably practical for application in autonomous driving. Furthermore, our complete model demonstrated significantly notable performance, recording error values nearly half that of SGNet. This demonstrates that our proposed prediction model presents a competitive solution even considering the trade-off of computational efficiency.

TABLE 7. Comparison of performance based on the depth of transformer encoder and decoder.

Datasets	# Encoder Layers	# Decoder Layers	C_{ADE} (4.0s)	C_{FDE} (4.0s)
Nuscenes	1	1	1875	4582
	2	1	1404	3851
	3	1	1488	3865
	3	3	1573	3997
	5	5	1597	3862

3) IMAGE RESOLUTIONS

The proposed approach utilizes image inputs to actively incorporate key contextual information around the target pedestrians for prediction. In this regard, the resolution of input images can influence both the model's performance and inference time. Therefore, we analyze the impact of various input image resolutions, which we have documented in Table 6. The image resolutions encompass four settings while maintaining a 16:9 aspect ratio and including the full resolution provided by the dataset. Naturally, higher image resolutions lead to increased inference time and improved performance. However, when comparing image resolution 960 × 540 to 1600 × 900, there is a very slight performance degradation of approximately 1.08 times based on C_{ADE}, while the inference time is reduced by around 37%. On the other hand, for 640 × 360, the performance is similar to 960 × 540, but the inference time is significantly reduced by approximately 18% compared to 1600 × 900. As observed from the results, it can be deduced that the advantages in terms of inference time outweigh the performance degradation caused by reducing the resolution. Nevertheless, at the very low resolution of 480 × 270, a relatively significant performance degradation was observed. Therefore, considering the trade-off between performance differences and computational load, we adopt input images of size 640 × 360 for our model.

4) TRANSFORMER LAYERS DEPTH

The performance of the proposed model was compared on the Nuscenes dataset by adjusting the number of encoder and decoder layers in the transformer. The results are shown in Table 7. It was found that there was not a significant performance difference based on the depth of the layers. However, when only one layer was used for the encoder, a slight decrease in performance was observed. On the other hand, when the encoder had two or more layers, the performance was quite similar. The highest performance was achieved when the encoder had two layers and the decoder had one layer. Thus, it can be concluded that the selected layer depth is sufficient for capturing the relevant information and generating accurate future predictions.

V. CONCLUSION

We propose a multi-modal future trajectory prediction model for pedestrians that effectively integrates historical trajectory

and raw image using a transformer with an attention mechanism in an egocentric view. Particularly, unlike previous research, we derive future trajectories that are independent of ego-motion and demonstrate the utility of ego-motion independent future trajectory in autonomous driving through risk assessment experiments. Furthermore, our proposed prediction model exhibited superior performance not only in terms of application aspects but also in terms of prediction accuracy for prioritized trajectories compared to previous studies. We also present a light version model with slightly lower performance but much faster design without using images that can be useful depending on the specifications of autonomous driving equipment. In the future, we plan to consider interactions among agents, with a particular focus on integrating the influence of the ego vehicle's movement on pedestrians' future trajectories. Moreover, we have plans to further develop the model to ensure robustness. This includes enhancing the model's ability to handle slightly unstable input trajectories resulting from real-time 2D object detection performance, as well as variations in image quality caused by surrounding environmental factors or weather conditions. These efforts will contribute to the overall reliability and performance of our proposed model, enabling its deployment in real-world autonomous driving scenarios.

ACKNOWLEDGMENT

The Institute of Engineering Research at Seoul National University provided research facilities for this work.

REFERENCES

[1] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social LSTM: Human trajectory prediction in crowded spaces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 961–971.

[2] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social GAN: Socially acceptable trajectories with generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2255–2264.

[3] H. Xue, D. Q. Huynh, and M. Reynolds, "Bi-prediction: Pedestrian trajectory prediction based on bidirectional LSTM classification," in *Proc. Int. Conf. Digit. Image Comput., Techn. Appl. (DICTA)*, Nov. 2017, pp. 1–8.

[4] V. Kosaraju, A. Sadeghian, R. Martín-Martín, I. Reid, H. Rezatofighi, and S. Savarese, "Social-BiGAT: Multimodal trajectory forecasting using bicycle-GAN and graph attention networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–10.

[5] L. Li, M. Pagnucco, and Y. Song, "Graph-based spatial transformer with memory replay for multi-future pedestrian trajectory prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 2231–2241.

[6] L. Shi, L. Wang, C. Long, S. Zhou, M. Zhou, Z. Niu, and G. Hua, "SGCN: Sparse graph convolution network for pedestrian trajectory prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8994–9003.

[7] H. Zhou, D. Ren, H. Xia, M. Fan, X. Yang, and H. Huang, "AST-GNN: An attention-based spatio-temporal graph neural network for interaction-aware pedestrian trajectory prediction," *Neurocomputing*, vol. 445, pp. 298–308, Jul. 2021.

[8] P. Dendorfer, S. Elflein, and L. Leal-Taixé, "MG-GAN: A multi-generator model preventing out-of-distribution samples in pedestrian trajectory prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 13158–13167.

[9] L. Huang, J. Zhuang, X. Cheng, R. Xu, and H. Ma, "STI-GAN: Multimodal pedestrian trajectory prediction using spatiotemporal interactions and a generative adversarial network," *IEEE Access*, vol. 9, pp. 50846–50856, 2021.

[10] B. Yu, H. Yin, and Z. Zhu, "Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting," 2017, *arXiv:1709.04875*.

[11] Y. Huang, H. Bi, Z. Li, T. Mao, and Z. Wang, "STGAT: Modeling spatial–temporal interactions for human trajectory prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6272–6281.

[12] Y. Yuan, X. Weng, Y. Ou, and K. Kitani, "AgentFormer: Agent-aware transformers for socio-temporal multi-agent forecasting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9813–9823.

[13] L. Achaji, T. Barry, T. Fouqueray, J. Moreau, F. Aioun, and F. Charpillet, "PreTR: Spatio-temporal non-autoregressive trajectory prediction transformer," in *Proc. IEEE 25th Int. Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2022, pp. 2457–2464.

[14] D. Liu, Q. Li, S. Li, J. Kong, and M. Qi, "Non-autoregressive sparse transformer networks for pedestrian trajectory prediction," *Appl. Sci.*, vol. 13, no. 5, p. 3296, Mar. 2023.

[15] N. Lee, W. Choi, P. Vernaza, C. B. Choy, P. H. S. Torr, and M. Chandraker, "DESIRE: Distant future prediction in dynamic scenes with interacting agents," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 336–345.

[16] C. Xu, W. Mao, W. Zhang, and S. Chen, "Remember intentions: Retrospective-memory-based trajectory prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 6488–6497.

[17] I. Bae, J.-H. Park, and H.-G. Jeon, "Non-probability sampling network for stochastic human trajectory prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 6477–6487.

[18] L. Zhou, D. Yang, X. Zhai, S. Wu, Z. Hu, and J. Liu, "GA-STT: Human trajectory prediction with group aware spatial–temporal transformer," *IEEE Robot. Autom. Lett.*, vol. 7, no. 3, pp. 7660–7667, Jul. 2022.

[19] A. Bhattacharyya, M. Fritz, and B. Schiele, "Long-term on-board prediction of people in traffic scenes under uncertainty," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4194–4202.

[20] T. Yagi, K. Mangalam, R. Yonetani, and Y. Sato, "Future person localization in first-person videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7593–7602.

[21] R. Chandra, U. Bhattacharya, A. Bera, and D. Manocha, "TraPHic: Trajectory prediction in dense and heterogeneous traffic using weighted interactions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8483–8492.

[22] O. Makansi, Ö. Çiçek, K. Buchicchio, and T. Brox, "Multimodal future localization and emergence prediction for objects in egocentric view with a reachability prior," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4354–4363.

[23] R. Quan, L. Zhu, Y. Wu, and Y. Yang, "Holistic LSTM for pedestrian trajectory prediction," *IEEE Trans. Image Process.*, vol. 30, pp. 3229–3239, 2021.

[24] Y. Yao, E. Atkins, M. Johnson-Roberson, R. Vasudevan, and X. Du, "BiTraP: Bi-directional pedestrian trajectory prediction with multi-modal goal estimation," *IEEE Robot. Autom. Lett.*, vol. 6, no. 2, pp. 1463–1470, Apr. 2021.

[25] C. Wang, Y. Wang, M. Xu, and D. J. Crandall, "Stepwise goal-driven networks for trajectory prediction," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 2716–2723, Apr. 2022.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[27] J. Yue, D. Manocha, and H. Wang, "Human trajectory prediction via neural social physics," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2022, pp. 376–394.

[28] B. Ivanovic and M. Pavone, "The trajectron: Probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2375–2384.

[29] B. Ivanovic, E. Schmerling, K. Leung, and M. Pavone, "Generative modeling of multimodal multi-human behavior," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2018, pp. 3088–3095.

[30] C. Choi, A. Patil, and S. Malla, "DROGON: A causal reasoning framework for future trajectory forecast," *CoRR*, vol. abs/1908.00024, pp. 1–14, Jul. 2019.

[31] X. Bai, Z. Hu, X. Zhu, Q. Huang, Y. Chen, H. Fu, and C.-L. Tai, "TransFusion: Robust LiDAR-camera fusion for 3D object detection with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1090–1099.
[32] P. Gao, M. Zheng, X. Wang, J. Dai, and H. Li, "Fast convergence of DETR with spatially modulated co-attention," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3621–3630.
[33] P. C. Ng and S. Henikoff, "SIFT: Predicting amino acid changes that affect protein function," *Nucleic Acids Res.*, vol. 31, no. 13, pp. 3812–3814, Jul. 2003.
[34] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Comput. Vis. Image Understand.*, vol. 110, no. 3, pp. 346–359, Jun. 2008.
[35] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2564–2571.
[36] H. Y. Kim and S. A. De Araújo, "Grayscale template-matching invariant to rotation, scale, translation, brightness and contrast," in *Advances in Image and Video Technology*. Santiago, Chile: Springer, Dec. 2007, pp. 100–113.
[37] M. Muja and D. G. Lowe, "Fast approximate nearest neighbors with automatic algorithm configuration," in *Proc. VISAPP*, 2009, vol. 2, nos. 331–340, p. 2.
[38] Y. Ding, W. Zhuang, L. Wang, J. Liu, L. Guvenc, and Z. Li, "Safe and optimal lane-change path planning for automated driving," *Proc. Inst. Mech. Eng. D, J. Automobile Eng.*, vol. 235, no. 4, pp. 1070–1083, Mar. 2021.
[39] Y. Liang, Y. Li, A. Khajepour, Y. Huang, Y. Qin, and L. Zheng, "A novel combined decision and control scheme for autonomous vehicle in structured road based on adaptive model predictive control," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 16083–16097, Sep. 2022.
[40] W. Lim, S. Lee, J. Yang, M. Sunwoo, Y. Na, and K. Jo, "Automatic weight determination in model predictive control for personalized car-following control," *IEEE Access*, vol. 10, pp. 19812–19824, 2022.
[41] S. Bae, D. Isele, A. Nakhaei, P. Xu, A. M. Añon, C. Choi, K. Fujimura, and S. Moura, "Lane-change in dense traffic with model predictive control and neural networks," *IEEE Trans. Control Syst. Technol.*, vol. 31, no. 2, pp. 646–659, Mar. 2023.
[42] T. Brüdigam, M. Olbrich, D. Wollherr, and M. Leibold, "Stochastic model predictive control with a safety guarantee for automated driving," *IEEE Trans. Intell. Vehicles*, vol. 8, no. 1, pp. 22–36, Jan. 2023.
[43] D. Xu, Z. Ding, X. He, H. Zhao, M. Moze, F. Aioun, and F. Guillemard, "Learning from naturalistic driving data for human-like autonomous highway driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 12, pp. 7341–7354, Dec. 2021.
[44] S. Xu, R. Zidek, Z. Cao, P. Lu, X. Wang, B. Li, and H. Peng, "System and experiments of model-driven motion planning and control for autonomous vehicles," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 52, no. 9, pp. 5975–5988, Sep. 2022.
[45] T. Zhang, W. Song, M. Fu, Y. Yang, X. Tian, and M. Wang, "A unified framework integrating decision making and trajectory planning based on spatio-temporal voxels for highway autonomous driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 10365–10379, Aug. 2022.
[46] W. Lim, S. Lee, M. Sunwoo, and K. Jo, "Hybrid trajectory planning for autonomous driving in on-road dynamic scenarios," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 1, pp. 341–355, Jan. 2021.
[47] N. Deo, E. Wolff, and O. Beijbom, "Multimodal trajectory prediction conditioned on lane-graph traversals," in *Proc. Conf. Robot Learn.*, 2022, pp. 203–212.
[48] Y. Chai, B. Sapp, M. Bansal, and D. Anguelov, "MultiPath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction," 2019, *arXiv:1910.05449*.
[49] N. Deo and M. M. Trivedi, "Trajectory forecasts in unknown environments conditioned on grid-based plans," 2020, *arXiv:2001.00735*.
[50] A. Rasouli, I. Kotseruba, T. Kunic, and J. Tsotsos, "PIE: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6262–6271.
[51] L. Liu, X. Chen, S. Zhu, and P. Tan, "CondLaneNet: A top-to-down lane detection framework based on conditional convolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3773–3782.
[52] B. Janakiraman, S. Shanmugam, R. P. D. Prado, and M. Wozniak, "3D road lane classification with improved texture patterns and optimized deep classifier," *Sensors*, vol. 23, no. 11, p. 5358, Jun. 2023.
[53] X. Pan, J. Shi, P. Luo, X. Wang, and X. Tang, "Spatial as deep: Spatial CNN for traffic scene understanding," in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, no. 1, pp. 1–8.
[54] Q. Zou, H. Jiang, Q. Dai, Y. Yue, L. Chen, and Q. Wang, "Robust lane detection from continuous driving scenes using deep neural networks," *IEEE Trans. Veh. Technol.*, vol. 69, no. 1, pp. 41–54, Jan. 2020.
[55] D. Neven, B. D. Brabandere, S. Georgoulis, M. Proesmans, and L. Van Gool, "Towards end-to-end lane detection: An instance segmentation approach," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2018, pp. 286–291.
[56] W. Van Gansbeke, B. De Brabandere, D. Neven, M. Proesmans, and L. Van Gool, "End-to-end lane detection through differentiable least-squares fitting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 905–913.

YUJIN KIM received the B.S. degree in mechanical and advanced material engineering from Kyunghee University, South Korea, in 2018, and the M.S. degree in mechanical engineering from Seoul National University, South Korea, in 2020, where she is currently pursuing the Ph.D. degree in mechanical engineering. Her research interests include detection and prediction of pedestrians for autonomous driving.

EUNBIN SEO received the B.S. degree in computer science and technology from the Daegu Gyeongbuk Institute of Science and Technology (DGIST), South Korea, in 2023. She is currently pursuing the M.S. degree with the Interdisciplinary Program in Artificial Intelligence, Seoul National University, South Korea. Her research interests include computer vision and computer system for automated driving.

CHIYUN NOH received the B.S. degree in mechanical engineering from Seoul National University, South Korea, in 2023, where he is currently pursuing the Ph.D. degree in mechanical engineering. His research interest includes multi object tracking via sensor fusion for autonomous vehicles.

KYONGSU YI (Member, IEEE) received the B.S. and M.S. degrees in mechanical engineering from Seoul National University, South Korea, in 1985 and 1987, respectively, and the Ph.D. degree in mechanical engineering from the University of California at Berkeley, Berkeley, CA, USA, in 1992. He is currently a Professor with the School of Mechanical Engineering, Seoul National University. His research interests include control systems, driver assistant systems active safety systems, and automated driving of ground vehicles.