A Multimodal Switch Transformer for Pedestrian Trajectory Prediction

Yu-Jou Chen [1] , Pao-Kai Wang [1] , Wen-Te Hsiao [1] , Yu-Chen Lin [1*], Kuan-Chu Hou [2]

[1] Department of Automatic Control Engineering, Feng Chia University, Taichung 40724, Taiwan, R.O.C

[2] IoT Solution Development at AdasEco Inc.

[*] E-mail(s): yuchlin@fcu.edu.tw (Y. C. Lin).

***Abstract*—In Taiwan, the high rate of accidents between pedestrians and vehicles remains a pressing issue. Pedestrian trajectory prediction poses a significant challenge for autonomous systems, primarily due to the complex dynamics of social interactions. This paper develops a pedestrian trajectory prediction model based on the multimodal switch Transformer with temporal attention mechanism. In addition, the Gating Mechanism is used to select the best trajectory from multimodal candidates, while the correlated sampling module learns true data distributions through sampling, improving accuracy, stability, and robustness for real-world safety applications. The proposed model was evaluated on the TITAN dataset and showed improved performance over previous models, achieving a 10.24% reduction in minimum average distance error (minADE) and a 23.04% reduction in minimum final displacement error (minFDE).**

Keywords—Pedestrian trajectory prediction, Multimodal, switch Transformer, temporal attention, correlated sampling, TITAN dataset

I. INTRODUCTION

Pedestrian trajectory prediction has always played a crucial role in autonomous vehicle and intelligent mobile robot navigation tasks. By analyzing the intentions of pedestrians and predicting their trajectories, intelligent systems not only significantly enhances its understanding of the surrounding environment but also provides valuable support for path planning and decision-making. For advanced driver assistance systems (ADAS) or autonomous vehicles, predicting the trajectories of pedestrians serves as an early warning mechanism for potential accidents, enabling better safety measures, such as achieving collision prevention or mitigating collision impacts.

Traditionally, researchers have attempted to model and understand human behavior using simple rules and mechanisms, it's called the physics-based models such as the Kalman filter [1] and Extended Kalman filter [2] for prediction tasks, which can be employed to predict future trajectories of pedestrians over short-term horizons (i.e., whether the pedestrian will continue walking or stop to cross the road). However, the inherent randomness of human walking behavior introduces uncertainty to pedestrian trajectories and intertwines with various aspects of social dynamics, such as the significant influence of social interactions and the surrounding environment on human movement patterns. Therefore, traditional prediction methods struggle to handle such highly complex scenarios and long-range inferences effectively. In recent years, data-driven pedestrian behavior modeling methods such as social interaction-based deep learning models have become increasingly popular due to their outstanding trajectory prediction performance. One of the most influential neural network architectures for pedestrian trajectory prediction is the Social LSTM [3] proposed by Alahi *et al*. Since then, numerous researchers have proposed various deep learning architectures, such as the use of recurrent neural networks (RNNs) [4-6], conditional variational autoencoders (CVAE) [7, 8], and the spatial-temporal attention network [9-10].

A common approach in studies focuses on static, top-down perspectives, which excel at capturing intricate spatial-temporal relationships. These models leverage the predictability of structured environments to effectively understand agent interactions and future movements. Techniques like the Edge-enhanced heterogeneous graph transformer (EPHGT) [9] and the Gate-Calibrated Double disentangled distribution matching network (CD3MN) [10] showcase the strengths of this perspective in modeling complex spatial dependencies and temporal dynamics. However, these models are primarily tailored to static, structured settings and struggle to adapt to ego-centric visual scenarios, such as those captured by front-facing vehicle cameras. Their limitations arise from an inability to adapt to the dynamic and ever-changing nature of these environments. As a result, creating trajectory prediction models specifically designed for ego-centric remains a considerable challenge.

This paper focuses on developing a novel Transformer-based model with a temporal attention mechanism for predicting future pedestrian trajectories in egocentric vision, which includes three main contributions: (i) **Improved Training Efficiency:** The model uses teacher-forcing in early training to accelerate convergence and improve learning. A mixture-of-experts framework selects top-k experts, reducing active parameters to boost inference speed while maintaining performance. (ii) **Enhanced Prediction Capabilities:** The gating mechanism selects the best trajectory from 20 multimodal candidates, while the correlated sampling module learns true data distributions through sampling, improving accuracy, stability, and robustness for real-world safety applications. (iii) **Real-Time State-of-the-Art Performance:** On the Honda TITAN dataset [11], the model achieves outstanding performance in minimum average distance error (minADE) and minimum final displacement error (minFDE). With only 1.153 GFLOPs and 0.259M parameters, it meets real-time requirements for automotive embedded platforms.

II. Related Works

Since pedestrians lack any protective measures, they have always been the most vulnerable group in road traffic. This has resulted in a consistently high proportion of pedestrian-related traffic accidents, with road fatality rates continuing to rise. As a result, pedestrian trajectory prediction has become a widely researched topic in the field of autonomous driving. Accurately predicting pedestrians' future movements can help autonomous vehicles respond promptly, effectively avoiding potential hazards and improving overall safety for both pedestrians and drivers. The accuracy of this task is influenced by various environmental factors, including the dynamics of nearby pedestrians and vehicles, vehicle speed, and external conditions such as weather and road conditions. To address these challenges, the academic community has proposed numerous pedestrian trajectory prediction models. With the rapid development of trajectory prediction technologies, significant progress has been made in this field.

A. Research on Transformer architecture

In pedestrian trajectory prediction, long short-term memory (LSTM) [12] networks and gated recurrent units (GRU) [13] have been widely used, achieving good results by processing time-series data. However, with the advancement of research, Transformer architectures have gained attention for their exceptional performance in temporal representation learning, particularly in natural language processing and multimodal prediction. Compared to traditional models, Transformers excel in capturing long-range spatiotemporal dependencies and handling complex motion trajectories. However, they require large datasets and are prone to overfitting, necessitating fine-tuning in practical applications.

Studies have shown that effectively integrating contextual information is crucial for improving prediction accuracy. For example, combining visual information (e.g., images and optical flow) with non-visual data (e.g., vehicle speed) provides richer background details, enhancing prediction performance. According to literature [14], Transformers demonstrate strong capabilities in learning such contextual information. Research by Franco *et al*. [4] further confirms this by comparing transformers and bidirectional transformers (BERT) [15] with traditional models like LSTM and RNN, as well as state-of-the-art methods such as Social-GAN [5] and Trajectron++ [6]. The results show that Transformers and BERT outperform their counterparts in multimodal prediction capabilities, accuracy, and flexibility, solidifying their advantage in capturing complex spatiotemporal dependencies. Transformers and BERT excel in pedestrian trajectory prediction by modeling long-term dependencies, spatiotemporal relationships, and multimodal futures, while LSTM and RNN are better suited for short-term predictions but struggle with long-term and complex tasks.

B. Autoregressive vs non-autoregressive

Damirchi *et al*. [16] proposed a multimodal Transformer-based encoder-decoder architecture that integrates vehicle steering angle and speed as contextual information with observed pedestrian trajectory data. This fusion allows the model to better perceive and adapt to dynamic pedestrian behavior, particularly when the vehicle is in motion. By combining trajectory and vehicle speed data through coordinated position encoding and feature extraction, the model enhances its ability to make accurate and informed predictions of future pedestrian trajectories. Unlike traditional sequential models that predict future trajectories step by step, the proposed Transformer generates the entire trajectory in a single prediction pass. This single-shot prediction significantly improves computational efficiency, making it particularly advantageous in real-time applications such as autonomous driving systems. One of the key highlights of this architecture is its exceptional inference speed. Compared to state-of-the-art models like SGNet [7] and BiTraP [8], it reduces inference time to as little as 2 milliseconds, making it ideal for real-time applications where low latency and high performance are critical, such as in mobile devices and autonomous driving systems. This efficiency not only improves prediction accuracy but also facilitates deployment in real-world scenarios, demonstrating its strong potential for practical applications.

C. Research on ego-view trajectory prediction

Unlike the common use of a third-person perspective primarily for crowd and trajectory monitoring, pedestrian trajectory prediction from a first-person perspective (ego-view) greater emphasis on integrating contextual information, physical constraints, and advancements in multimodal learning. This approach can provide significant support for vehicle warning systems and serves as a foundation for subsequent vehicle safety decision-making, such as proactive emergency braking system control.

In the ego-view pedestrian trajectory prediction, Sang *et al*. [17] proposed a novel method combining physics-based constraints with a probabilistic framework. Their model integrates a differential constraint module during trajectory generation to ensure the predictions align with physical motion laws, improving both accuracy and realism. This approach addresses the limitations of purely data-driven deep learning models, which often produce unrealistic and difficult-to-interpret trajectories. The differential constraint module ensures predictions adhere to real-world dynamics, enhancing interpretability and reliability.

Similarly, PedFormer [18] introduces a multimodal Transformer-based architecture to predict pedestrian behavior from an egocentric perspective. It employs cross-modal attention to capture dependencies between various data modalities, such as pedestrian trajectories, ego-vehicle dynamics, and environmental interactions. The Semantic Attention Interaction Module further encodes interactions by leveraging semantic scene information and visual attention. PedFormer's hybrid gated decoder improves the integration of shared and individual task representations, enabling simultaneous predictions of trajectories and crossing actions.

III. Methodology

The primary objective of this study is to accurately predict pedestrian trajectories using front-view images combined with vehicle state information obtained from IMU sensors as features

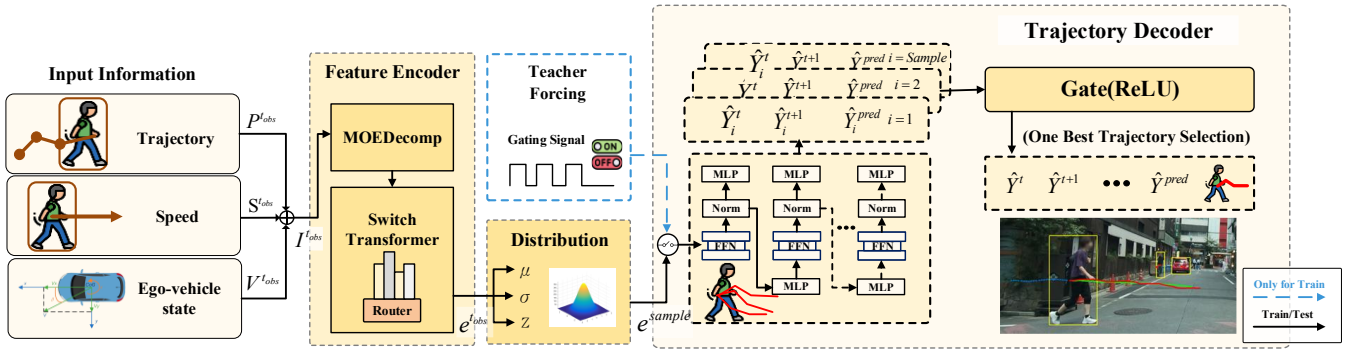


Fig. 1. Overall Architecture Diagram of the Transformer-Based Model.

for trajectory prediction. The research framework, as shown in Fig. 1, is developed based on ego-view, focusing on processing the dynamic trajectories of objects, including pedestrians and vehicles, to enhance the accuracy and practicality of trajectory prediction.

A. Temporal Feature Extraction Encoder

This study utilizes bounding box positions $P^{t_{obs}}$, position speed $S^{t_{obs}}$, and inertial measurement unit (IMU) $V^{t_{obs}}$ data as input features for trajectory prediction. After feature extraction, a mixture-of-experts (MOE) module is employed to further process the features.

$$I^{t_{obs}} = \text{concat}(P^{t_{obs}}, S^{t_{obs}}, V^{t_{obs}}) \tag{1}$$

The MOE module uses a gating mechanism to dynamically assign inputs to specialized experts, each being a simple linear transformation network. The gating network determines expert weights based on the input, enabling adaptive output combination for efficient parameter use and flexibility. SwitchFFN [19], as illustrated in Fig. 2, extends expert selection in Transformers by replacing traditional FFN layers with multiple expert networks. A routing mechanism selects the best expert for each token, reducing active parameters and computational load while maintaining performance.

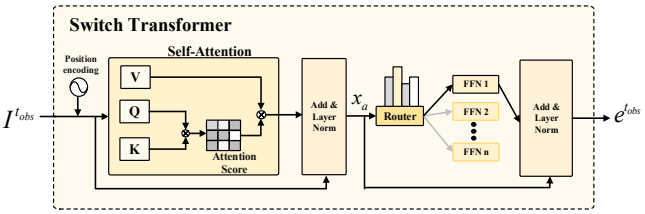


Fig. 2. Switch Transformer Architecture. (FFN with Switch Mechanism)

$$x_a = Att\ \ Q(I^{t_{obs}}), K(I^{t_{obs}}), V(I^{t_{obs}}) = soft\max(\frac{QK^T}{\sqrt{d_k}})V \tag{2}$$

$$Expert(i) = TopK(x_a, k),\ k = 1 \tag{3}$$

$$e^{t_{obs}} = FFN_{Expert(i)}(x_a) \tag{4}$$

where x_a is the feature extraction result computed by the attention mechanism. Subsequently, there are i feed-forward network (FFN) experts available for selection. The expert selection is performed using the TopK method, where $k=1$ in this case, ensuring that the highest-ranked expert is chosen for processing. The final output is $e^{t_{obs}}$.

Switch Transformer scales by replacing the conventional feedforward layers of a standard Transformer with a Sparse Mixture-of-Experts (Sparse MoE) mechanism, where a gating network dynamically routes each token to a set of specialized expert networks. Unlike the standard Transformer, where each feedforward layer (FFN) is a single network, Switch Transformer activates only the most relevant experts for each token, thereby reducing computational overhead while maintaining expressive power. This architecture enhances temporal feature extraction by efficiently capturing complex dependencies, thereby improving the accuracy of trajectory prediction. With adaptive expert selection, Switch Transformer excels in complex and dynamic environments.

B. Correlated Sampling

In real-world scenarios, pedestrian trajectories are highly uncertain. Many studies use Gaussian distributions to model this uncertainty, but this research introduces Correlated Sampling, as illustrated in Fig. 3, to capture interdependencies between samples. This approach better reflects real-world dynamics, where movement direction and speed are interrelated. By incorporating a learnable correlation matrix, the model effectively captures these latent connections, improving prediction accuracy and handling multivariate influences. Linear layers learn distribution features, while the correlation matrix models dependencies among predicted trajectories. This enables more realistic trajectory distributions and enhances overall model performance.

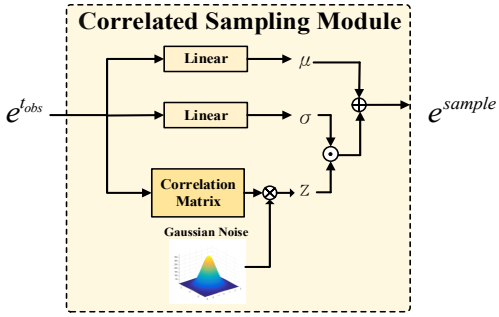


Fig. 3. Correlated Sampling Module.(Sample=20)

$$e^{sample} = \mu + \sigma \odot (L \cdot z),\ z \sim N\ (0, I) \tag{5}$$

where L is a learnable correlation matrix that introduces dependencies between sampled latent variables, ensuring that the generated samples capture underlying correlations in the data. The term $z \sim N\ 0, I$ represents a standard normal distribution, and $\sigma \odot (L \cdot z)$ incorporates the learned correlations into the sampled latent space. Finally, the output e^{sample} is obtained by decoding the latent variables into the trajectory prediction space.

C. Autoregressive Training Strategy with Teacher-Forcing

In the field of trajectory prediction decoder research, model strategies can be broadly categorized into two types: non-autoregressive strategies and autoregressive strategies.

The non-autoregressive strategy generates predictions for all time steps simultaneously, offering efficiency advantages in large-scale data processing and real-time applications. However, its accuracy is often lower. In contrast, the autoregressive strategy predicts step by step, using previous outputs as inputs, allowing it to capture deeper temporal dependencies for more precise predictions. While non-autoregressive models prioritize speed, autoregressive models are better suited for high-accuracy tasks. Therefore, this study adopts the autoregressive approach to improve trajectory prediction accuracy in complex scenarios.

To further improve the learning performance of the autoregressive model, this study introduces the Teacher Forcing technique, which stabilizes sequence prediction by guiding the model with ground-truth inputs during early training. Without Teacher Forcing, the model may learn suboptimal features due to error accumulation in self-predictions, negatively impacting future steps. To mitigate this, a gradually decreasing Teacher Forcing Rate is employed: initially, the model is heavily guided by ground-truth labels, ensuring accurate temporal dependency learning. As training progresses, reliance on ground-truth labels is progressively reduced, allowing the model to adapt to self-prediction. This controlled transition enhances robustness in sequence prediction tasks, ultimately improving real-world trajectory prediction accuracy.

$$r_{step} = \left(1 - {step}/{T_{step}}\right) \tag{6}$$

$$\text{Teacher Forcing}: \begin{cases} 1, \ if\ r_{step} > \text{rand}(1) \\ 0, \ otherwise \end{cases} \tag{7}$$

where r_t represents the Teacher Forcing rate at step T_{step} is the total number of steps over which the Teacher Forcing rate decays The Teacher Forcing decision follows a binary rule: if the condition is met, 1 indicates that Teacher Forcing is applied; otherwise, 0 means the model relies on its own predictions.

D. Loss Function

The loss function used in this study is based on the minimum Euclidean distance trajectory error. The primary reason for selecting this loss function lies in its calculation method. It begins by squaring the differences between the model-generated multimodal trajectory predictions and the ground truth trajectories. The errors across multiple future time steps are then summed and square-rooted to obtain the final distance metric. From the multiple trajectories generated by the model, the one with the smallest error is selected as the final output trajectory.

$$L_{traj} = \min_{i \in q} \left(\sqrt{\sum_{k=obs+1}^{obs+u} \left(m_i^k - \hat{m}_i^k \right)^2} \right) \tag{8}$$

where m_i^k represents the ground truth trajectory, $\hat{m}_i^k$ represents the predicted multimodal trajectory, k represents the total prediction time steps, i denotes the multimodal trajectory sequences generated by the model, $obs+u$ is the final prediction time step, q is the maximum number of multimodal trajectories generated by the model. By calculating the error values for all generated trajectories, the one with the smallest error is selected and used as the loss for the current prediction task.

IV. Experimental and Results

This study proposes a new trajectory prediction model for road users based on a Transformer architecture with a temporal attention mechanism, specifically designed for ego-view vehicle perspectives. On the Honda TITAN dataset, the model achieves state-of-the-art results with minADE of less than 10.16 pixels and minFDE of less than 15.03 pixels. Notably, the model demonstrates exceptional computational efficiency, achieving a throughput of 96.93 FPS with a batch size of 32 on an NVIDIA GeForce RTX 3080 Ti GPU, requiring only 1.153 GFLOPs and a parameter count of just 0.259M. These characteristics highlight the model's suitability for automotive embedded platforms, meeting real-time operational requirements and offering a promising solution for trajectory prediction in complex road scenarios.

A. Implementation details

Our pedestrian trajectory prediction model is developed with an egocentric perspective. The model is trained, validated, and tested using the TITAN dataset. The training parameters include the Adam optimizer, a batch size of 128, and a learning rate of 0.0005. Training was conducted on an NVIDIA GeForce RTX 3080 Ti GPU for 150 epochs.

B. Quantitative Comparison Result

Following previous benchmarks, this study adopts the approach of using 10 observed time steps to predict 20 future trajectory points. Evaluations conducted on the TITAN dataset show that our model outperforms other baseline models. As detailed in Table I., our model achieves a 10.24% improvement in minADE and a 23.04% improvement in minFDE compared to the state-of-the-art TITAN model. These improvements highlight the significant advantage of our model in accurately predicting future pedestrian trajectories.

TABLE I. QUANTITATIVE EVALUATION FOR FUTURE OBJECT LOCALIZATION.ADE ARE FDE IN PIXELS ON THE ORIGINAL SIZE 1920X1200.

	Trajectory Prediction (TITAN)					
	Year	**K=20**			**RTX 3080Ti GPU Batch size 32**	
		ADE↓	**FDE↓**	$FIOU_{Max}$ ↑	**Params**	**FPS**
Social-GAN [5]	2018	35.41	69.41	-	-	-
TITAN [11]	2020	11.32	19.53	0.6559	-	-
GPRAR [20]	2021	12.56	20.36	-	-	-
ABC + [21]	2022	30.52	46.84	-	-	-
PTINet [22]	2024	16.97	28.79	-	-	-
Ours	-	**10.16**	**15.03**	**0.7218**	0.259M	96.93

C. *Ablation Study*

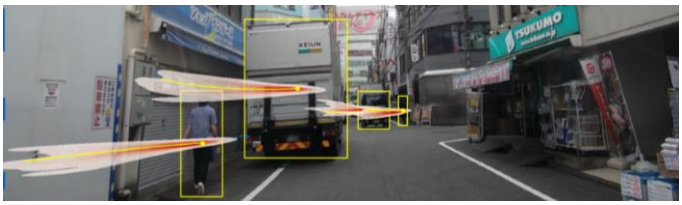
In Table II., the Switch Transformer outperforms the traditional Transformer through an efficient expert selection mechanism. By leveraging this mechanism, the Switch Transformer significantly enhances the accuracy of trajectory prediction.

TABLE II. EVALUATION OF DIFFERENT ENCODERS FOR TRAJECTORY PREDICTION ON TITAN.

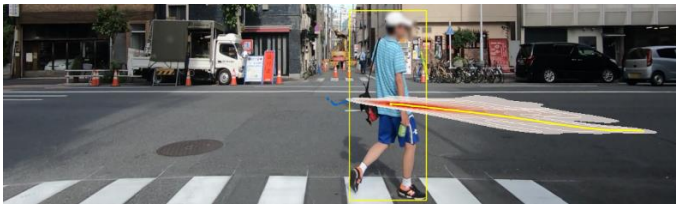
	Trajectory Prediction (TITAN)		
	ADE↓	FDE↓	$FIOU_{Max}$ ↑
LSTM	11.01	16.66	0.7063
GRU	10.84	16.46	0.7130
Standard Transformer	10.42	15.51	0.7180
Switch Transformer	**10.16**	**15.03**	**0.7218**

D. *Visualization Result*

Fig.4 visualizes the model's performance in pedestrian trajectory prediction. The observed 10 time steps are shown in blue, predicted future 20 steps in yellow, and distribution in red. The close alignment between prediction and ground truth highlights the model's accuracy and reliability. In Fig. 4(a), stationary roadside objects appear to move backward as a result of the ego-vehicle's motion. Fig. 4(b) presents the model's prediction in a pedestrian crossing scenario.



(a)



(b)

Fig. 4. Visualization Result. (Blue: Observation; Yellow: Prediction in minADE; Red: Prediction distribution.)

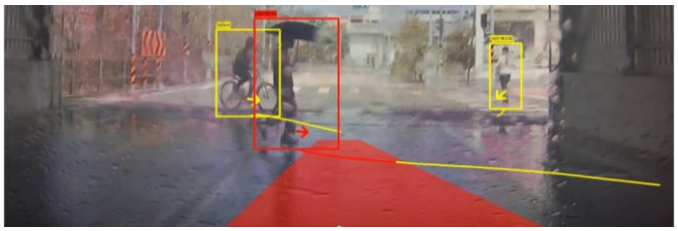
E. *Dangerous Intent of Trajectory*

Considering that ground truth is unavailable in real-world applications, this study has developed a gating mechanism to address this limitation. The gating mechanism is designed to select the most suitable trajectory from 20 multimodal predicted candidate trajectories. By utilizing a threshold-based gating process, the mechanism identifies and outputs the optimal trajectory that best aligns with real-world conditions. This enhancement significantly improves the model's predictive capabilities, ensuring that the generated trajectories are more aligned with practical safety alert requirements sin real-world scenarios. As shown in Fig. 5, based on the selected trajectories, it is determined whether they have entered the vehicle's hazardous zone and pose a high risk of a critical event. Red bounding boxes indicate instances classified as having dangerous intent. In contrast, yellow bounding boxes represent detected pedestrians and objects in the scene, which are assessed to have a lower immediate risk.

Fig. 5. Trajectory intersecting the ego-vehicle's path, indicating a potential collision risk.

(a)

(b)



(c)

Fig. 6. Dangerous intent of trajectory in TAIWAN dataset. (a) Dangerous pedestrian scene. (b) Dangerous motorcycle scene. (c) Adverse weather scene.

F. *TAIWAN Pedestrian Dataset*

Data collections and annotations. Our laboratory is independently recording and compiling the TAIWAN dataset, developed by the Taiwan Car Laboratory in Tainan, Taiwan. It includes data from front-view cameras and vehicle IMU sensors, covering diverse and challenging scenarios. Ground truth annotations are enhanced using state-of-the-art tracking-by-detection algorithms, providing critical inputs for the Transformer-based pedestrian trajectory prediction model to improve reliability, robustness, and accuracy. The results are shown in Fig. 6.

V. CONCLUSION

This paper develops a novel pedestrian trajectory prediction model designed for real-world safety-critical applications. By leveraging onboard front-view camera and IMU sensor data, the proposed Transformer-based architecture with a temporal attention mechanism achieves precise and robust trajectory predictions. The model's lightweight design ensures efficiency, meeting the real-time processing requirements of automotive embedded systems. Evaluations on the Honda TITAN dataset highlight the model's state-of-the-art performance in trajectory prediction, achieving superior accuracy while maintaining computational efficiency. These results demonstrate the model's potential as a reliable and effective solution for enhancing safety in dynamic environments.

ACKNOWLEDGMENT

This Research was supported by National Science and Technology Council, R.O.C., under the Grant NSTC 113-2218-E-035-001.

REFERENCES

[1] F. A. Hermawati *et al*., "Trajectory Prediction Using Kalman Filter Method As Collision Risk Assessment On Autonomous Tram," *International Conference on Information & Communication Technology and System (ICTS)*, pp. 249-254, 2023.

[2] C. Y. Lin, L. J. Kau, and C. Y. Chan, "Bimodal Extended Kalman Filter-Based Pedestrian Trajectory Prediction," *Sensors*, vol. 22, no. 21, pp. 1-18, 2022.

[3] A. Alahi, K. Goel, V. Ramanathan, and N. Correll. "Social LSTM: Social Long Short-Term Memory for Human Trajectory Prediction," *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 623-631, 2016.

[4] L. Franco, *et al*., "Under the Hood of Transformer Networks for Trajectory Forecasting," *Pattern Recognition*, vol. 138, pp. 109372, 2023.

[5] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social GAN: Socially Acceptable trajectories with generative adversarial networks," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2255-2264, 2018.

[6] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone, "Trajectron++: Dynamically-Feasible Trajectory Forecasting with Heterogeneous Data," *European Conference on Computer Vision* , pp. 683-700, 2020.

[7] C. Wang, Y. Wang, M. Xu, and D. J. Crandall, "Stepwise Goal-Driven Networks for Trajectory Prediction," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 2716-2723, 2022.

[8] Y. Yao, E. Atkins, M. Johnson-Roberson, R. Vasudevan, and X. Du, "BiTraP: Bi-Directional Pedestrian Trajectory Prediction with Multi-Modal Goal Estimation," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 1463-1470, 2021.

[9] X. Zhou, X. Chen, and J. Yang, "Edge-Enhanced Heterogeneous Graph Transformer With Priority-Based Feature Aggregation for Multi-Agent Trajectory Prediction," *IEEE Transactions on Intelligent Transportation Systems*, 2024.

[10] Z. Liu, Y. Wu, D. Zeng, S. Du, and B. Peng, "Gate-Calibrated Double Disentangled Distribution Matching Network for Cross-Domain Pedestrian Trajectory Prediction," *IEEE Signal Processing Letters*, 2024.

[11] S. Malla, B. Dariush, and C. Choi, "TITAN: Future forecast using action priors," *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11183-11193, 2020.

[12] S. Hochreiter, and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 1997.

[13] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling," *Neural Information Processing Systems Workshop on Deep Learning*, 2014.

[14] Y. Wang, Z. Guo, C. Xu, and J. Lin, "A Multimodal Stepwise-Coordinating Framework for Pedestrian Trajectory Prediction," *Knowledge-Based Systems*, vol. 299, pp. 112038, 2024.

[15] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding," *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 4171-4186, 2019.

[16] H. Damirchi, M. Greenspan, and A. Etemad, "Context-Aware Pedestrian Trajectory Prediction with Multimodal Transformer," *IEEE International Conference on Image Processing (ICIP)*, pp. 2535-2539, 2023.

[17] H. Sang, J. Wang, Q. Liu, W. Chen, and Z. Zhao, "Physics Constrained Pedestrian Trajectory Prediction with Probability Quantification," *Expert Systems with Applications*, vol. 255, pp. 124743, 2024.

[18] A. Rasouli and I. Kotseruba, "PedFormer: Pedestrian Behavior Prediction via Cross-Modal Attention Modulation and Gated Multitask Learning," *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9844-9851, 2023.

[19] W. Fedus, B. Zoph, and N. Shazeer, "Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity," *arXiv preprint*, 2021.

[20] M. Huynh and G. Alaghband, "GPRAR: Graph Convolutional Network Based Pose Reconstruction and Action Recognition for Human Trajectory Prediction," *British Machine Vision Conference (BMVC)*, pp. 401-410, 2021.

[21] M. Halawa, O. Hellwich, and P. Bideau, "Action-Based Contrastive Learning for Trajectory Prediction," *European Conference on Computer Vision (ECCV)*, pp. 143-159, 2022.

[22] F. Munir and T. Kucner, "Context-aware Multi-Task Learning for Pedestrian Intent and Trajectory Prediction," *arXiv preprint*, 2024.