Traffic & Transportation

Intent Recognition and Trajectory Prediction for Multiple Types of Traffic Participants at an Unsignalized Intersection Based on Bidirectional Spatiotemporal Attention Network

Yanan HOU[1], Mingbao PANG[2], Huamin LIANG[3]

Original Scientific Paper
Submitted: 30 Apr 2025
Accepted: 15 Sep 2025
Published: 27 May 2026

[1] ynanhou@163.com, School of Civil and Transportation Engineering, Hebei University of Technology, Tianjin, China
[2] Corresponding author, pmbpgy@hebut.edu.cn, School of Civil and Transportation Engineering, Hebei University of Technology, Tianjin, China
[3] 3075992742@qq.com, School of Civil Engineering, Tianjin Chengjian University, Tianjin, China

Publisher:
Faculty of Transport and Traffic Sciences, University of Zagreb

ABSTRACT
This work investigates intent recognition and trajectory prediction for multiple types of traffic participants at an unsignalized intersection within connected intelligence environments based on bidirectional spatiotemporal attention network (Bi-STANet). An unsignalized intersection is used as the studied object, where the participants include Connected and Automated Vehicles (CAVs), Human Vehicles (HVs), bicyclists, and pedestrians. A novel method is proposed based on Bi-STANet. First, a multimodal spatiotemporal feature extraction model is constructed based on (2+1)-dimensional CNN ((2+1)D CNN), where grid encoding method is used to unify the spatial structure and 2D convolution is used to extract spatial features for capturing the disordered characteristics of participants. Temporal dynamics are modelled via 1D convolution along the time axis, enabling spatiotemporal decoupling. Second, a bidirectional dynamic interaction model is developed by integrating LSTM-based temporal feature extraction with (2+1)D CNN layers, where heterogeneous modality fusion is implemented through a Bidirectional Contextual Block (BiCoBlock). Finally, a model integrating dynamic interaction, intent recognition, and trajectory prediction is developed. The proposed method is validated through the inD dataset innovatively. The results show that the average accuracy of intent recognition can reach to 95.4%. Within a 3-second horizon, Average Displacement Error (ADE) and Final Displacement Error (FDE) can be reduced to 0.51 m and 0.64 m, respectively, compared with the best baseline model. In Ablation studies, intent recognition F1-score can be enhanced by 7.2%, and ADE and FDE of trajectory prediction can be enhanced by 41.4% and 39.0%, respectively.

KEYWORDS
unsignalized intersection; bidirectional spatiotemporal attention network (Bi-STANet); (2+1)D CNN; multiple types of traffic participants; intent recognition; trajectory prediction.

1. INTRODUCTION

The safety decision-making and control of Connected and Automated Vehicles (CAVs) rely on precise situational awareness of each different driving situation [1], where the scenario of unsignalized road intersection with multiple types of traffic participants is more complex due to the uncertainty behaviours of participants and the possible interactions among CAVs, Human Vehicles (HVs), bicyclists, and pedestrians [2–4]. The traveling direction and the starting time of participants are not limited by signal light. The probability of interaction is still very large although the traffic flow density is very low. Moreover, the intersection may be in a disordered state when the complex interweaving among participants occurs. By using the conventional methods [5], CAVs have some difficulty in identifying the intents of participants and

predicting their trajectories, especially bicyclists, and pedestrians. Low recognition rate and prediction accuracy lead to inability to meet the needs of engineering applications. It has become one of the urgent problems that should be solved in this field.

In recent years, many scholars have focused on the studies of decreasing the errors of trajectory prediction and enhancing the precision of intent recognition by constructing interaction frameworks among participants in these complex scenarios. For example, Azadani et al. proposed the Spatio-Temporal Attention Graph (STAG) for vehicle path prediction, where asymmetric and dynamic inter-agent interactions are modelled to address complex spatio-temporal dependencies [6]. Byeon et al. proposed a multi-model fusion framework for lane-changing intent recognition, where an attention-enhanced BiLSTM is combined with a reinforcement learning-based CRF to model temporal dependencies and vehicle interactions [7]. Qiao et al. proposed the Social-Attention LSTM model for long-term vehicle trajectory prediction, where interaction relationships captured by the Social-Pooling layer and temporal dependencies extracted via a self-attention mechanism are integrated into the LSTM architecture to enhance prediction accuracy [8]. Evidently, only HVs were discussed, which limits the application of these methods in complex scenarios with bicyclists and pedestrians, e.g., an unsignalized intersection.

In response to the above issues, some scholars have attempted to study dynamic interaction, trajectory prediction, and intent recognition problems of four types of participants by incorporating bicyclists and pedestrians. For example, Hosford et al. investigated pedestrian-bicyclist interactions through observational studies at urban intersections, where crossing behaviours and environmental factors were analysed to identify conditions associated with increased interaction frequency [9]. Benhelal et al. proposed an edge-assisted clustering framework for vehicle trajectory prediction, where multi-agent predictions are aggregated via DBSCAN on an edge server to enhance accuracy and robustness [10]. Long et al. proposed a Dual-LSTM network, in which dynamic interaction information between autonomous vehicles and other participants is processed to reduce long-term trajectory prediction errors [11]. Those studies show the effectiveness of the methods in mapping the nonlinear relation of complex scenarios. But the impact of Vulnerable Road Users (VRUs) with highly random behaviour on vehicle movement, e.g., pedestrians and bicyclists, has not yet been addressed. The bidirectional interaction relationships among the four types of participants have not been effectively mapped in the models. Moreover, the joint interaction modelling of pedestrian and rider spatiotemporal trajectories has not been discussed yet. For an unsignalized intersection with four types of participants, the spatial structure should be refined due to the uncertainty of interactive sub-scenario. The spatial features in response to the disordered characteristics of participants should be exacted and the temporal features in response to the multi-directional characteristics of participants' intents should be extracted. If we use the above methods for modelling nonlinear behaviour, it cannot be achieved to fully capture the dynamic spatiotemporal dependencies among participants. The accuracies of intent recognition and trajectory prediction cannot be enhanced.

The Bi-STANet model, with multiple types of traffic participants as the studied subject, can be used to provide a more accurate solution of trajectory prediction in complex scenarios, by which the participants' intents and trajectories can be discussed deeply, the interaction characteristics between VRUs and vehicles can be thoroughly explored, and the behavioural patterns of participants can be effectively characterized. At the same time, (2+1)D CNN [12,13] employed previously for human activity recognition, can be used to enhance the modelling of traffic participants, by which the spatiotemporal features [14,15] of participants can be captured and represented based on 2D [16] CNN, the dynamic changes in space can be monitored based on 1D CNN, and the spatial relationships between participants and surrounding participants can be handled. By innovatively introducing a bidirectional cross-attention mechanism [17–20], the BiCoBlock module can be used to enhance the accuracy of subsequent intent recognition and trajectory prediction. The temporal and spatiotemporal features of participants can be updated, and the continuity of the temporal dimension and the correlation of the spatial dimension in the model are preserved. Based on these, an unsignalized intersection is used as the studied object, where participants include bicyclists, pedestrians, CAVs and HVs. A Bi-STANet-based method is proposed for intent recognition and trajectory prediction of participants under intelligent connected environments. The grid encoding method [21,22]is used to unify spatial structure and a multimodal spatiotemporal feature extraction model is established based on (2+1)D convolution. The BiCoBlock is constructed based on (2+1)D CNN and Bi-STANet. The effectiveness of the proposed approach is validated through the comprehensive experimental evaluation.

The main contributions of this paper are as follows. The first one is that a method for intent recognition and trajectory prediction of multiple types of participants at an unsignalized intersection is proposed based on Bi-

STANet, and its model is established. The second one is that a multimodal spatiotemporal feature extraction model of participants is constructed based on (2+1)D CNN. The third one is that a bidirectional dynamic interaction model is constructed by fusing (2+1)D convolutional and LSTM-based features.

The remainder of this paper is structured as follows: Section 2 presents the problem formulation. In Section 3, the intent recognition and trajectory prediction modules based on Bi-STANet are established. The experimental analysis is discussed in Section 4. Finally, Section 5 summarizes the conclusions.

2. PROBLEM FORMULATION

An unsignalized intersection shown in *Figure 1* is used as the studied object, where multiple types of traffic participants include CAVs, HVs, bicyclists, and pedestrians. There is relatively low traffic demand during the time period according to the rules set for the signal lights [5]. Here, two complex sub-scenarios are displayed in order illustrate the issue. In *Figure 1a*, bicyclist A is observed to interact with going-straight HV D, left-turning HV B, and straight-moving CAVS E and C during a wide-radius left-turn manoeuvre. In *Figure 1b*, interaction scenarios depicted at the four corner regions show typical behaviours observed in real-world traffic conditions. Most pedestrians cross the intersection through zebra crossings, but there are still a few who do not follow the rules and walk randomly.

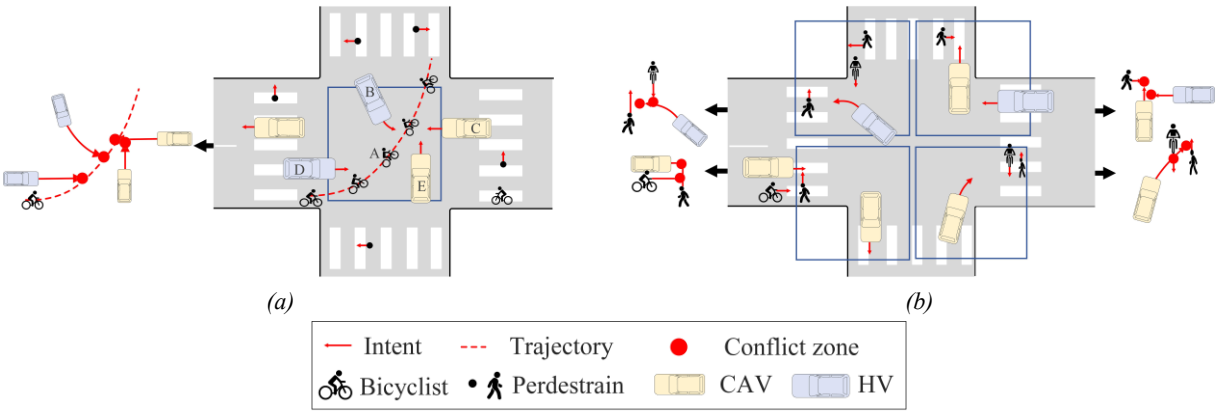


Figure 1 – Studied object: a) Central interaction scenario; b) Corner interaction scenario

Table 1 shows the main behavioural characteristics and common intents of participants. These behavioural patterns highlight the need for models that can effectively capture the diverse and dynamic interactions among these participants, which is crucial for accurate trajectory prediction and intent recognition in autonomous driving systems. By incorporating these characteristics, a more comprehensive understanding of traffic dynamics can be achieved.

Table 1 – Behavioural characteristics and intents of traffic participants

Participant	Behavioural characteristics	Conventional intents
Bicyclist	Moderate speed Frequent wrong-way riding Prone to large-radius left turns	Straight/right turn Wrong-way riding large-radius left turns
Pedestrian	Low speed High randomness Weak rule adherence	Sudden appearance Wrong-way walking Multi-directional walking
HV/CAV	High speed Strong rule adherence Usually yield to VRUs	Straight movement Left/right turns Yielding/stopping

The objective of this work is to simultaneously recognize the motion intent and predict the future trajectory of the target participants. Here $x_t^i \in \mathbb{R}^2$ represents the 2D position of the i-th participant at time t, and the observed trajectory over the past T_{obs} frames is denoted by $x_{1:T_{\text{obs}}}^i = \{x_1^i, x_2^i, \ldots, x_{T_{\text{obs}}}^i\}$. The goal is to predict both

the intent $I^i \in \mathcal{I}$ and the future trajectory $\hat{y}^i_{T_{\text{obs}}+1:T_{\text{pred}}} = \{\hat{y}^i_{T_{\text{obs}}+1}, \ldots, \hat{y}^i_{T_{\text{pred}}}\}$. The surrounding traffic participants denoted by $\mathcal{N}_i = \{V_1, V_2, \ldots, V_N\}$ are considered in interaction model. The state variables for each neighbour $j \in \mathcal{N}_i$, denoted by $s^j_t = (x^j_t, v^j_t, a^j_t, \theta^j_t)$, are the first encoded into grid-based spatiotemporal features via the CNN module and then incorporated into the attention-based fusion process to model inter-agent interactions. The problem can be formulated by a learning function.

$$(I^i, \hat{y}^i) = \Phi\left(x^i_{1:T_{\text{obs}}}, \{x^j_{1:T_{\text{obs}}}\}_{j \in N_i}, \mathcal{M}\right) \tag{1}$$

where $\Phi(\cdot)$ is modeled using a multimodal spatiotemporal attention network integrating (2+1)D CNN and LSTM, capable of jointly performing intent recognition and trajectory prediction.

3. METHODOLOGY

The proposed framework for intent recognition and trajectory prediction is systematically presented in this section. First, the overall model architecture based on Bi-STANet is introduced. Subsequently, the multimodal spatiotemporal feature extraction process using (2+1)D CNN is described. Finally, the bidirectional dynamic interaction mechanism realized through the BiCoBlock module is detailed. Each component is designed to enhance the capture of complex spatiotemporal dependencies and improve overall predictive performance.

3.1 Model structure based on Bi-STANet

Figure 2 shows the structure of the proposed intent recognition and trajectory prediction model based on Bi-STANet, where LSTM, (2+1)D CNN, and the BiCoBlock module are comprehensively utilized to accurately capture the time series dynamics of the observed agent and the spatial interaction characteristics with the surrounding environment. The parameters and variables of the model are listed in *Table 2*.

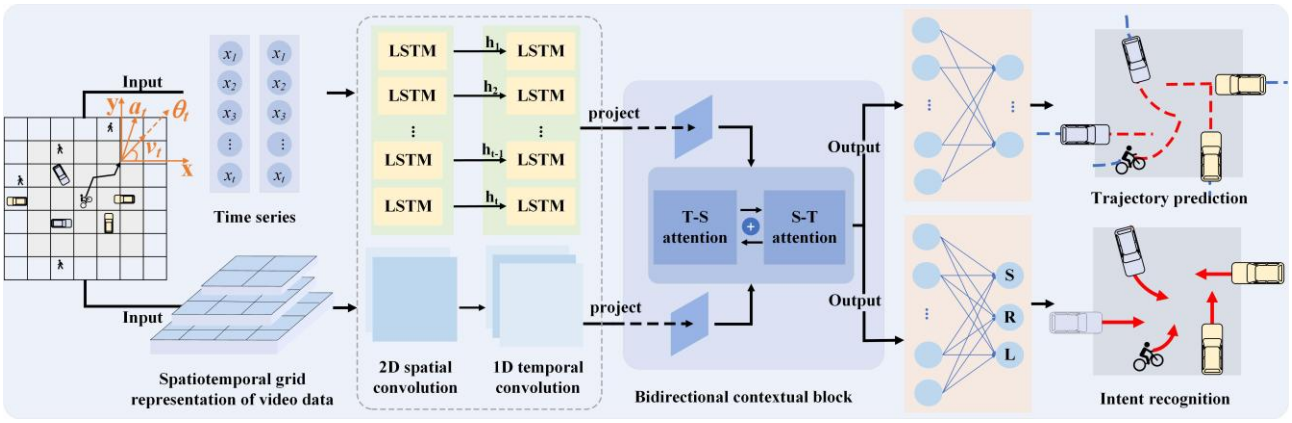


Figure 2 – Structure of intent recognition and trajectory prediction model based on Bi-STANet

The specific working processes of the model are as follows.

1) Input the historical trajectory data of participants into the LSTM network, e.g., position, speed, acceleration, and heading, and then extract trajectory features denoted by $H^{(\text{LSTM})}$. Simultaneously, input the keyframe images of participants at the intersection into the CNN network as video tensors denoted by $V \in \mathbb{R}^{B \cdot C \cdot T \cdot H \cdot W}$.
2) Construct a spatiotemporal occupancy grid [22], encoding both occupancy and motion direction in each grid cell. Transform positional features into a local coordinate system aligned with the agent's heading and normalize continuous variables using statistics from the training set for numerical stability. Subsequently, compute the spatial grid representation of the traffic participants. Obtain spatiotemporal features denoted by $H^{(\text{CNN})}$.
3) Project $H^{(\text{LSTM})}$ and $H^{(\text{CNN})}$ into the feature space denoted by d_{model} via linear projection, by which the condition that temporal features and spatial features are in the same dimension can be ensured.
4) Update the temporal features, where $H^{(\text{LSTM})}$ is used as the query, and the continuous frame grid features denoted by $H^{(\text{CNN})}$ are input into the multi-head attention mechanism as the key and value.

$$H^{(\mathrm{LSTM})}=\mathrm{Softmax}\left(\frac{H^{(\mathrm{LSTM})}\left(H^{(\mathrm{CNN})}\right)^{T}}{\sqrt{d}}\right)H^{(\mathrm{CNN})} \tag{2}$$

5) Update the spatiotemporal features, where $H^{(\mathrm{CNN})}$ is used as the query, and the updated $H^{(\mathrm{LSTM})}$ is input into the multi-head attention mechanism as the key and value.

$$H^{(\mathrm{CNN})}=\mathrm{Softmax}\left(\frac{H^{(\mathrm{CNN})}\left(H^{(\mathrm{LSTM})}\right)^{T}}{\sqrt{d}}\right)H^{(\mathrm{LSTM})} \tag{3}$$

6) Fuse the updated features and construct a joint probability model $P(c_i,Y_i) = P(c_i|X_i,V_i) \cdot P(Y_i|c_i,X_i)$. Extract multi-scale motion features $h_p^{(t)} = \mathrm{LSTM}(\mathrm{CNN}(X_i^{(t-\Delta t:t)}))$. Introduce a learnable modality alignment matrix $Q_i^{(t)} = \frac{\sigma(h_p^{(t)}\mathbf{M})(h_v^{(t)})^T}{\sqrt{d_v}}$, $\mathbf{M} \in \mathbb{R}^{d_p \cdot d_v}$, and achieve semantic fusion of trajectories and video-scene features.

$$F_{\mathrm{fused}}=\mathrm{LayerNorm}\left(h_P+\mathrm{Softmax}\left(\frac{h_P Q h_v^{T}}{\sqrt{d}}\right)h_v\right) \tag{4}$$

7) Visualize the predicted trajectories of traffic participants and output the intent recognition probabilities to intuitively present the experimental results.

Table 2 – Notation of key model parameters and variables

Notation	Definition
$H^{(\mathrm{LSTM})}$	Trajectory features extracted by LSTM.
$H^{(\mathrm{CNN})}$	Trajectory features extracted by CNN.
V	Tensor of key-frame images for intersection participants.
c_i	Discrete intent classes of intersection participants.
X_i	Spatiotemporal tensor encoded from continuous video frames.
Y_i	Probability distribution of future trajectory prediction.
F_{fused}	Fused feature of trajectory and video scene features.
$\tilde{H}_L$	Updated temporal features.
$\tilde{H}_C$	Updated spatiotemporal features.
LayerNorm	Layer Normalization.
MLP	Multilayer perceptron.
Softmax	Softmax function. Normalize logits to a probability distribution.
W_Q^C	Query weight matrix for the h-th attention head.
W_K^C	Key weight matrix for the h-th attention head.
W_V^C	Value weight matrix for the h-th attention head.

3.2 Multimodal spatiotemporal feature extraction based on (2+1)D CNN

Figure 3 shows the architecture of multimodal spatiotemporal feature extraction based on (2+1)D CNN. A grid-based environmental encoding method is employed, in which the non-motorised target agent is designated as the reference centre. A fixed-size spatial region is discretized into uniform grid cells. The relative position of each surrounding participant is computed by measuring the distance and angle to the target participant's centre, thus determining their placement in the grid. This enables frame level occupancy representations to be constructed using the model, capturing both spatial distributions and motion patterns. The spatiotemporal grid

representation enables the simultaneous capture of spatial features within each frame and temporal dependencies across frames, essential for modelling dynamic interactions and accurately predicting trajectories. By using conventional 3D convolutions [23], while capable of directly modelling spatiotemporal relationships by operating jointly in temporal and spatial dimensions, large-scale video data can't be processed in real-time due to high computational complexity and parameter scale. To address this issue, a (2+1)D CNN structure is designed, where 3D convolution is decomposed into a combination of 2D spatial convolution and 1D temporal convolution. According to existing literature on spatiotemporal convolutions, the R(2+1)D architecture achieves a 3-3.8% improvement in accuracy over R3D on both 8-frame and 16-frame inputs, with similar computational complexity, which significantly reduces computational complexity while preserving the effective extraction capability of spatiotemporal features, improving efficiency and accuracy on video-based recognition tasks [24].

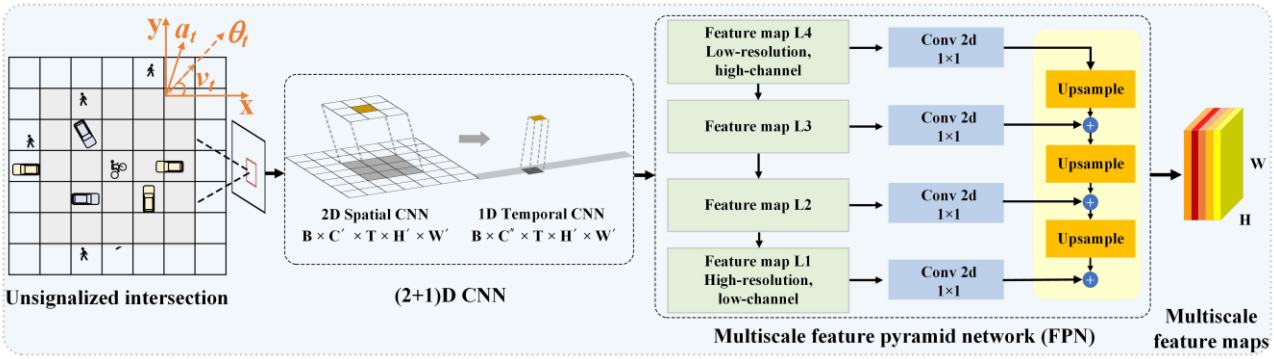


Figure 3 – Structure of the (2+1)D CNN

The specific steps are as follows.

1) (2+1)D CNN. The input video tensor is defined as follow.

$$X \in \mathbb{R}^{B \cdot C \cdot T \cdot H \cdot W} \tag{5}$$

where B, C, and T represent the batch size, the number of channels, and the number of temporal frames, respectively. H and W represent the height and width for each frame, respectively.

2D spatial convolution is performed on the input consecutive frame grids at each time step t to extract local static features. The operation of the 2D spatial convolution is expressed by *Equation 6.*

$$Y_{b,c',t,i,j}^{(2D)} = \sum_{c=1}^{C} \sum_{m=0}^{k_h-1} \sum_{n=0}^{k_w-1} X_{b,c,t,i+m,j+n} \cdot K_{2D}^{(c',c,m,n)} + b_{c'} \tag{6}$$

where $K_{2D} \in \mathbb{R}^{C_{mid} \cdot C \cdot k_h \cdot k_w}$ represents the two-dimensional convolutional kernel, C_{mid} represents the number of intermediate channels, k_h and k_w represent the spatial kernel sizes. $b_{c'}$ is the bias term. The dimension of the output feature tensor is expressed by *Equation 7*.

$$Y^{(2D)} \in \mathbb{R}^{B \cdot C_{mid} \cdot T \cdot H' \cdot W'} \tag{7}$$

One-dimensional convolution is implemented along the temporal dimension on the aforementioned result to capture dynamic associations between frames. The operation of the 1D temporal convolution is formulated by *Equation 8*.

$$Y_{b,c'',t,i,j}^{(2D)} = \sum_{c'=1}^{C_{mid}} \sum_{\tau=0}^{t_k-1} Y_{b,c',t+\tau,i,j}^{(2D)} \cdot K_{1D}^{(c'',c',\tau)} + b_{c''} \tag{8}$$

where $K_{1D} \in \mathbb{R}^{C_{out} \cdot C_{mid} \cdot t_k}$ represents the one-dimensional temporal convolution kernel, C_{out} represents the number of output channels. t_k represents the temporal convolution kernel size. $b_{c''}$ is the bias term. The dimension of the final output feature tensor is expressed by *Equation 9*.

$$Y \in \mathbb{R}^{B \cdot C_{out} \cdot T^{'} \cdot H^{'} \cdot W^{'}} \tag{9}$$

Through the aforementioned design, the complexity of computation can be significantly decreased, and the training efficiency can be enhanced compared to 3D convolution.

2) Multi-scale feature pyramid network (FPN). In the discussion of consecutive frame spatio-temporal grid matrices, multi-scale features are usually exhibited by different objects and dynamic behaviours, e.g., local rapid movements and global slow variations. So, it is difficult to comprehensively describe complex video content when relying solely on single-resolution features. To address this issue, a multi-scale Feature Pyramid Network (FPN)[25] is proposed to achieve multi-level feature fusion and multi-scale modelling. After the grid matrix of the input consecutive-frame is processed by multiple layers of (2+1)D convolutional modules, different levels of features are extracted layer by layer.

$$F^{l} = (2+1)\text{D-Conv}\left(F^{l-1}\right), l = 1,2,...,L \tag{10}$$

where $F^{l} \in \mathbb{R}^{B \cdot C_l \cdot T_l H_l \cdot W_l}$ represents the feature map of the L-th layer. L represents the total number of layers. The resolution of each layer's feature map gradually decreases, while the number of channels gradually increases.

Once features from all levels have been extracted, upsampling operations are performed sequentially, beginning with the highest-level feature F^{l}, and lateral concatenation is conducted with lower-level features.

Step 1. Channel alignment. The number of channels in the lower-level features is adjusted to match that of the higher-level features by employing a 1·1 convolution operation.

$$\widehat{F}^{l} = \text{Conv}_{1\cdot1}\left(F^{l}\right) \tag{11}$$

where $\hat{F}^{1} \in \mathbb{R}^{B \cdot C_{l+1}^{'} \cdot T_l \cdot H_l \cdot W_l}$.

Step 2. Upsample. The features from the previous layer are upsampled so that their spatial resolution is aligned with that of the current layer.

$$\widetilde{P}^{l+1} = \text{Upsample}\left(P^{l+1}\right) \tag{12}$$

where $\tilde{P}^{l+1} \in \mathbb{R}^{B \cdot C_{l+1}^{'} \cdot T_l \cdot H_l \cdot W_l}$.

Step 3. Feature fusion. The upsampled features are added to the adjusted ones of the current layer.

$$P^{l} = \widetilde{P}^{l+1} + \widehat{F}^{l} \tag{13}$$

After performing the above operations, a set of multi-scale feature maps denoted by $\{P^1, P^2, ..., P^L\}$ is output by the FPN, where Semantic information from higher levels and local details corresponding to each resolution are simultaneously preserved within the features of each layer, which can effectively support downstream tasks such as trajectory prediction and intent recognition.

3.3 Dynamic feature interaction based on BiCoBlock

The core issue of intent recognition and trajectory prediction is how to achieve multimodal feature fusion. In the scenario of the unsignalized intersection, a significant complementarity exists between the sequential motion features of participants and the spatial interaction features of their surrounding environment. In order to address the aforementioned issues, a BiCoBlock shown in Figure 4 is proposed, which enables bidirectional feature interaction and captures contextual features effectively [26]. Unlike conventional Transformer-based fusion mechanisms that apply unidirectional or modality-agnostic attention, BiCoBlock enables bidirectional information flow by allowing temporal and spatiotemporal features to serve as queries for each other, tailored for dynamic and asymmetric traffic interactions. This design allows task-relevant information to be selected from the counterpart's context, thereby achieving efficient feature fusion [27]. The problem of feature redundancy can be alleviated by adaptively filtering important information through attention mechanisms, and it is easy to integrate into multimodal networks through modular design, compared to the conventional

methods. Moreover, the incorporation of a lightweight gating mechanism improves the model's adaptability and interpretability, making BiCoBlock particularly suitable for heterogeneous agent modelling in complex, dynamic traffic scenarios.

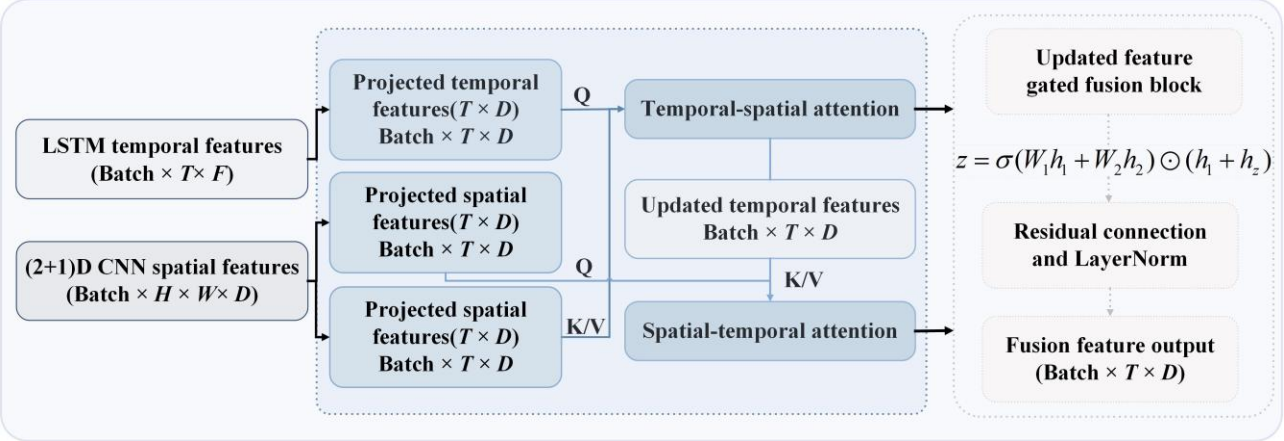


Figure 4 – Structure of the BiCoBlock

The two input feature sets are processed by BiCoBlock, which respectively represent the temporal dynamics of the observed subject and the spatial features of the surrounding environment. Here LSTM is used to extract the temporal features denoted by $H^{(\mathrm{LSTM})} \in \mathbb{R}^{B\cdot T\cdot F}$, where B, T, and F represent the batch size, the number of time steps, and the feature dimension. (2+1)D CNN is used to extract the spatiotemporal features denoted by $H^{(\mathrm{CNN})} \in \mathbb{R}^{B\cdot T\cdot D}$, where D represents the feature dimension of the CNN output. Since the dimensions of the two feature streams may be different, a linear transformation is first applied to project into a unified feature space d_{model}.

$$H_L^{(0)} = H^{(\mathrm{LSTM})} W_L,\ H_C^{(0)} = H^{(\mathrm{CNN})} W_C \tag{14}$$

where $W_L \in \mathbb{R}^{F\cdot d_{\mathrm{model}}}$ and $W_C \in \mathbb{R}^{D\cdot d_{\mathrm{model}}}$ are the learnable parameters, $H_L^{(0)}, H_C^{(0)} \in \mathbb{R}^{B\cdot T\cdot d_{\mathrm{model}}}$.

After the linear transformation, bidirectional information interaction is achieved via two multi-head attention computations. The specific steps are as follows.

Step 1: Update temporal features. The historical trajectory features by the output of LSTM are used as queries, while the consecutive-frame grid features by the output of CNN are employed as keys and values.

$$\tilde{H}_L = \mathrm{MHA}\left(Q = H_L^{(0)} W_Q^L,\ K = H_C^{(0)} W_K^L,\ V = H_C^{(0)} W_V^L\right) \tag{15}$$

where MHA represents the multi-head attention mechanism to compute the attention scores based on the query-key relationships.

$$\mathrm{Attention}(Q,K,V) = \mathrm{Softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right) V \tag{16}$$

where $d_k = d_{\mathrm{model}} / h$ represents the feature dimension of a single head. h represents the number of attention heads. The updated temporal features are obtained through residual connections and layer normalization.

$$H_L^{(1)} = \mathrm{LayerNorm}\left(\tilde{H}_L + H_L^{(0)}\right) \tag{17}$$

Step 2: Update spatiotemporal features. The temporal features updated in Step 1 are now used as keys and values for the LSTM.

$$\tilde{H}_C = \mathrm{MHA}\left(Q = \mathrm{H}_{\mathrm{C}}^{(0)} W_Q^C,\ K = \mathrm{H}_{\mathrm{L}}^{(1)} W_K^C,\ V = \mathrm{H}_{\mathrm{L}}^{(1)} W_V^C\right) \tag{18}$$

The updated spatiotemporal features are further refined through a second multi-head attention operation, followed by residual connections and layer normalization to enhance feature integration.

$$H_C^{(1)} = \text{LayerNorm}\left(\tilde{H}_C + H_C^{(0)}\right) \tag{19}$$

Step 3: Fuse the updated temporal and spatial features by designing a gating mechanism that enables the model to adaptively learn the integration of the two feature streams.

$$Z = g \cdot H_L^{(1)} + (1-g) H_C^{(1)},\ g = \sigma(W_g [H_L^{(1)}, H_C^{(1)}]) \tag{20}$$

where σ and W_g represent the sigmoid activation function and the learnable weight, respectively.

The time complexity of BiCoBlock module is $O(B \cdot T^2 \cdot d_{\text{model}})$ for its main calculation comes from two multi head attention operations. Since the two sets of features share the temporal dimension T, the module exhibits a linear relationship with respect to the time steps and batch size, making it suitable for multimodal learning tasks in actual scenarios. By using the module, temporal and spatial features interact bidirectionally, allowing each to access the contextual information of the other. This enhances feature complementarity while reducing the interference of redundant information through task-relevant adaptive filtering. As a result, the model's robustness and generalization capability are improved.

To enhance the representation capability of BiCoBlock, intent-guided information is incorporated during decoding to promote trajectory prediction accuracy. A joint modelling strategy fuses the intent embedding vector with spatiotemporal features, enabling conditional generation of future trajectories. Kullback-Leibler(KL) divergence is introduced to regularize the latent space distribution, while a weighted loss function is designed to jointly optimize intent recognition and trajectory prediction. This design captures both fine-grained motion dynamics and high-level semantic intent, thereby improving its overall predictive robustness and interpretability.

4. EXPERIMENTAL ANALYSIS

The experimental results are discussed to validate the effectiveness of the proposed method. Subsequently, the ablation experiments and the sensitivity analysis are conducted to further investigate the contributions of different modules and parameter configurations.

4.1 Experimental design

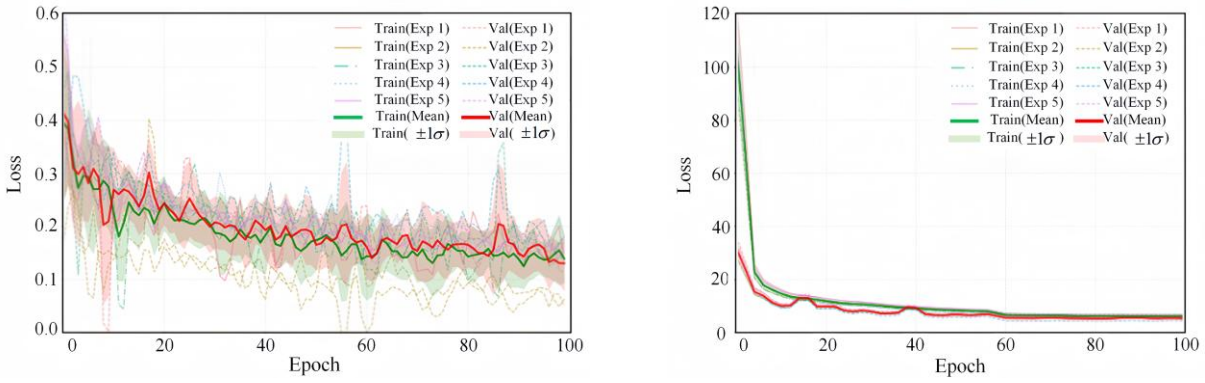
The inD dataset [28] developed by the Automotive Engineering Institute of RWTH Aachen University, was used as the experimental object, which covers natural driving scenarios at four typical German intersections. 80% of the dataset is used for model training, 10% for validation, and 10% for experimental prediction. Specifically, the selected dataset includes four recording sites that represent three distinct types of unsignalized intersections: a standard unsignalized intersection, an asymmetric unsignalized intersection, and an asymmetric T-intersection with channelized medians on both mainlines and side roads.

4.2 Experimental discussion

Figure 5 shows the training and validation loss trends of intent recognition and trajectory prediction using the proposed Bi-STANet model. *Figure 6* presents a comparison of the predicted trajectories, historical ones, and the actual ones for four types of participants. *Table 3* presents the performance comparison of participants.

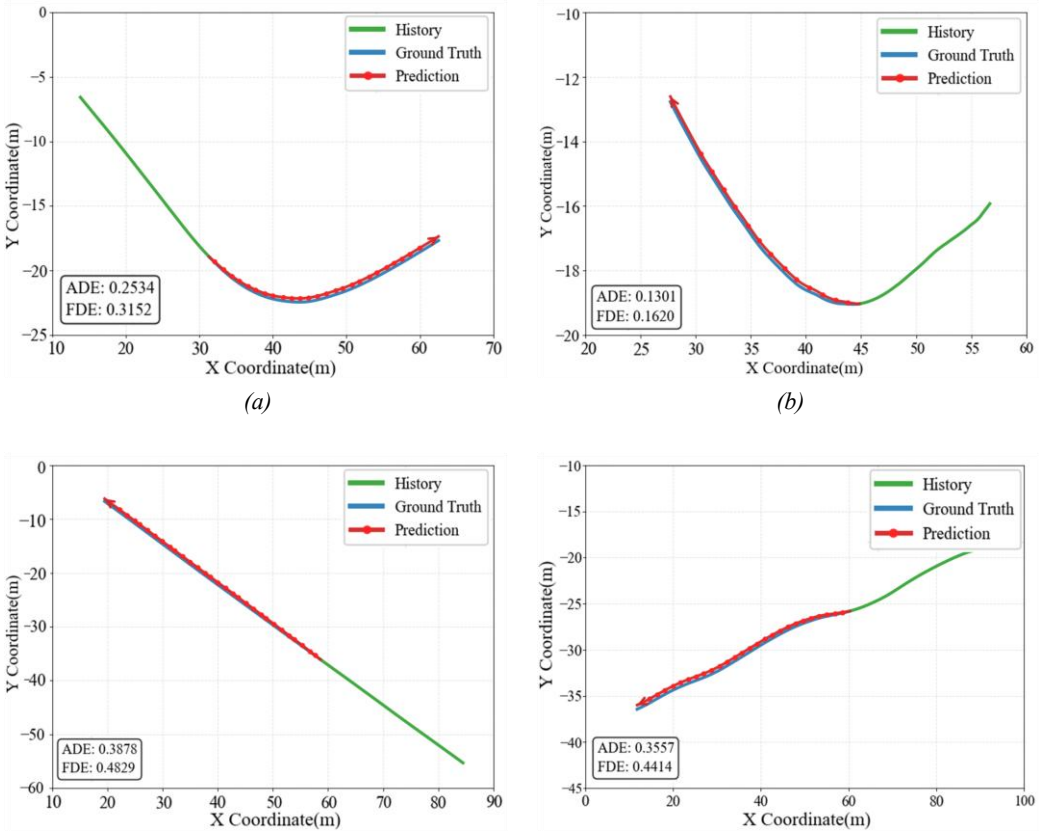
1) As shown in *Figure 5a*, the average training loss decreases significantly after the 20th epoch and stabilizes after the 50th epoch, with no signs of overfitting. Notably, the standard deviation of the validation loss is narrower than that of the training loss, and the intra-group correlation across five independent experiments reaches 0.93, indicating the strong robustness. Rapid convergence is observed in early training stages, while later fluctuations, attributed to distribution differences in the validation set, have minimal impact on performance.
2) As shown in *Figure 5b*, the training loss decreases monotonically, which indicates the effective capture of underlying data patterns. The average training loss and standard deviation interval show consistent convergence behaviour across five independent experiments. The validation loss follows the training loss, stabilizing after 30 epochs without significant increase, confirming the model's generalization capability. The standard deviation of the validation loss is notably narrower than that of the training loss, which shows excellent experimental repeatability. Rapid initial convergence is observed within the first 10 epochs, with

both training and validation losses stabilizing after 30 epochs, narrowing the gap between them, which suggests a robust generalization boundary has been established in the feature space.



(a) *(b)*

Figure 5 – Training loss curves: a) Intent recognition task; b) Trajectory prediction task



(a) *(b)*

(c) *(d)*

Figure 6 – Visualization of predicted trajectories: a) HV 1; b) HV 2; c) Pedestrian; d) Bicyclist

3) The results in *Figure 6* show that the predicted trajectory closely matches the ground truth in both direction and magnitude, which confirms the model's ability to capture global motion patterns. Prediction deviations occur mainly observed in turning areas. This is attributed to the limitations of scene dynamics or nonlinear interaction modelling. Trajectory adjustments of turning areas are strongly correlated with interaction parameters, e.g., relative speed and distance to surrounding participants. The effectiveness of the proposed multimodal interaction feature fusion is proved. The robustness of the proposed model at an unsignalized intersection is achieved by the multi-task joint optimization strategy, where intent recognition is trained with cross-entropy loss and trajectory prediction is optimized using mean squared error loss.

4) The results in *Table 3* that HV intent is recognized with the highest accuracy and lowest prediction errors. In contrast, bicyclist behaviour shows lower intent recognition performance and the highest trajectory errors, likely due to more dynamic and less rule-constrained motion. As a result of higher speeds and larger spatial displacement, even small directional deviations in vehicle trajectories can lead to significantly larger errors in ADE and FDE. In contrast, pedestrians move more slowly and along constrained paths, which leads to lower absolute displacement errors in these metrics. The effectiveness and the robustness of the proposed model with scenario-sensitive prediction accuracy are again proved.

Table 3 – Performance comparison of participants

Participant	Intent recognition				Trajectory prediction	
	Accuracy	Precision	Recall	F1-score	ADE@3sec(m)	FDE@3sec(m)
HV	0.974	0.959	0.974	0.971	0.48	0.66
Bicyclist	0.937	0.917	0.937	0.931	0.61	0.72
Pedestrian	0.951	0.922	0.951	0.948	0.45	0.55
Average	0.954	0.933	0.954	0.950	0.51	0.64

4.3 Ablation experiment

Ablation studies are conducted to evaluate individual module contributions and synergy effects, revealing each component's impact under controlled experimental conditions. The selected baseline methods include representative classical and interaction-aware models that have been widely adopted in intent recognition and trajectory prediction tasks.

Ablation experiment

Table 4 shows the ablation experiments results by sequentially removing key components-LSTM, (2+1)D CNN, and BiCoBlock from the model. The main findings are summarized as follows.

Table 4 –Ablation experiment results

Module	ADE	FDE	MSE	Accuracy	Precision	Recall	F1-score
Bi-STANet	0.51	0.64	0.26	0.954	0.933	0.954	0.950
LSTM + BiCoBlock	0.71	0.93	0.48	0.904	0.887	0.892	0.889
(2+1)D CNN + LSTM	0.59	0.79	0.34	0.933	0.920	0.929	0.924
(2+1)D CNN + BiCoBlock	0.65	0.85	0.40	0.921	0.902	0.915	0.908

1) Removing the (2+1)D CNN module results in a surge in trajectory prediction errors, with ADE increased by 39.2% and FDE increased by 45.3%, which indicates the irreplaceability of spatial interaction features for scene understanding. Furthermore, the F1-score in the recognition task is decreased by 6.4%, which underscores the (2+1)D CNN's key contribution to spatial information encoding.
2) Removing the BiCoBlock module results in a relatively smaller increase in trajectory prediction errors, with ADE rising by 15.7% and FDE increasing by 23.4%. However, the F1-score in the recognition task decreases by 2.7%, reflecting its auxiliary role in enhancing model robustness through spatiotemporal interaction fusion.
3) Removing the LSTM module results in a significant decline in trajectory prediction performance, with ADE increased by 27.5% and FDE increased by 32.8%, which indicates the necessity of temporal dynamic modelling for capturing traffic participants' behaviour evolution. In the intent recognition task, accuracy and F1-score is decreased by 3.5% and 4.4%, respectively.

In all, the complete Bi-STANet model achieves optimal performance in both trajectory prediction and intent recognition. The effectiveness of the multi-module collaborative mechanism is validated, where LSTM captures long-term temporal dependencies, (2+1)D CNN effectively extracts local spatiotemporal features, and BiCoBlock enhances adaptability to unsignalized intersections through a bidirectional cross-attention mechanism.

Module synergy effect

Four models are used to evaluate the module synergy, including the conventional LSTM+CNN model [29], the CNN+Self-Attention model [30], the LSTM+Self-Attention model [31], and the proposed Bi-STANet. The LSTM+CNN model represents a conventional spatiotemporal modelling approach that performs feature-level fusion without interaction modelling. The CNN+Self-Attention model combines CNN with self-attention for capturing global spatial relationships but lacks the capability to model temporal dependencies. The LSTM+Self-Attention model integrates LSTM with self-attention to enhance temporal modelling but does not incorporate spatial feature extraction. *Table 5* shows the comparison of performances by four methods. The main findings are summarized as follows.

Table 5 – Module synergy effect

Module	ADE	FDE	MSE	Accuracy	Precision	Recall	F1-score
Bi-STANet	0.51	0.64	0.26	0.954	0.933	0.954	0.950
LSTM+CNN	0.87	1.05	0.54	0.905	0.882	0.891	0.886
CNN+Self-Attention	0.93	1.12	0.59	0.892	0.871	0.879	0.875
LSTM+Self-Attention	0.89	1.07	0.56	0.913	0.890	0.898	0.894

1) Beyond individual temporal and spatial modelling, bidirectional interaction between temporal and spatiotemporal feature streams is achieved by introducing the BiCoBlock module. By using Bi-STANet model, ADE and FDE are decreased by 41.4% and 39.0%, respectively, and F1-score is increased by 7.2%, compared with the LSTM+CNN baseline model, which performs only feature-level fusion without interaction modelling. These improvements validate the effectiveness of the proposed module synergy, confirming that joint spatiotemporal interaction modelling significantly enhances overall performance.
2) Temporal dependency modelling is enhanced by integrating into Bi-STANet through LSTM. Bi-STANet achieves a 45.2% reduction in ADE, a 42.9% reduction in FDE, and an 8.6% improvement in F1-score, respectively, compared with the CNN+Self-Attention baseline model, which lacks sequence modelling capability. Furthermore, recall is improved by 8.5%, indicating that temporal dynamics are effectively captured. These results confirm the necessity of temporal feature extraction in intent recognition and trajectory prediction tasks.
3) (2+1)D CNN is employed for local spatiotemporal feature extraction to address the limitations in spatial representation. Bi-STANet achieves a 53.6% reduction in MSE, a 4.8% improvement in precision, compared with the LSTM+Self-Attention model, which lacks explicit spatial modelling. The precision of trajectory prediction is enhanced. This highlights the importance of spatial context in modelling complex motion patterns, especially in turning or interactive regions.

4.4 Sensitivity analysis

Sensitivity analyses are performed to investigate the effects of varying channel numbers, hidden layer sizes, and attention configurations on model performance. The trade-offs between prediction accuracy and computational complexity are systematically evaluated to guide optimal parameter selection.

Impact of channel and layer configurations on model performance

Figure 7 visually illustrates the variations in model performance and computational complexity and presents the core performance metrics and the corresponding parameter counts by varying the number of convolutional channels and LSTM hidden layers varies. The main findings are summarized as follows.

1) The performances are improved with the increase of the number of convolutional channels. Accuracy reaches 0.954 and F1-score reaches 0.947 with the optimal configuration achieved at 32 channels. However, beyond this point, a sharp rise in the parameter count is caused by further increases in the number of channels, from 358 K to 952 K, which significantly increases computational complexity. The performance improvements become marginal relative to the computational cost, which indicates diminishing returns beyond 32 channels.
2) Similarly, model performances are affected by the number of LSTM hidden layers. An optimal balance between performance and complexity is achieved by the model with 32 hidden layers, maintaining a

parameter count of 358K. While consistent performance improvements are led by increasing the number of layers from 16 to 32, diminishing returns are yielded by further increases in the number of layers. For example, a minimal improvement in F1-score from 0.947 to 0.935 is achieved by increasing the layers from 32 to 48, while the parameter counts increases from 358K to 625K.

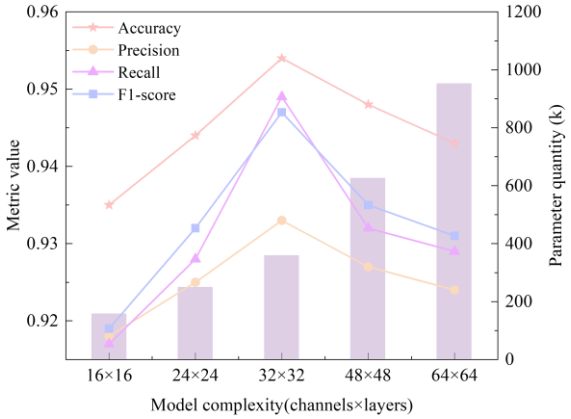


Figure 7 – Analysis of model architecture parameter combinations

Impact of attention configuration on model performance

Figure 8 illustrates the variations in model performance and computational complexity and presents the model's recognition performance metrics and parameter count by varying the number of attention heads and the dimension of each head. The main findings are summarized as follows.

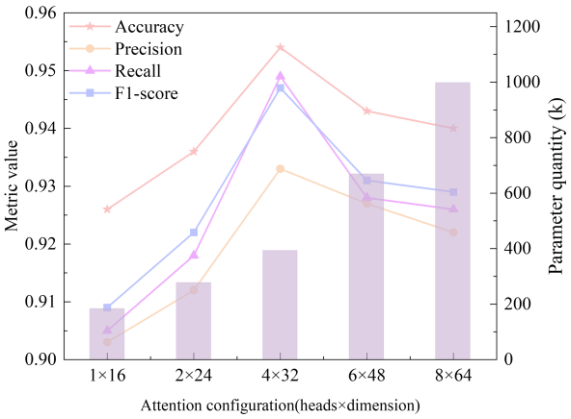


Figure 8 – Analysis of bidirectional attention mechanism parameter combinations

1) The model's accuracy and F1-score are improved with the increase of the number of attention heads. The optimal configuration achieved at 4 heads. Beyond this point, a sharp increase is observed in the parameter count, which leads to a notable rise in computational complexity. This results in diminishing returns, where performance improvement becomes marginal relative to the increase in parameter count, which highlights the trade-off between model accuracy and computational cost.
2) The optimal performance is achieved by the model with 4 attention heads and a head dimension of 32. While performance is improved with the increase of the dimension from 16 to 32, diminishing returns are yielded by further increases in dimension. Additionally, a significant increase in the parameter count is observed as the dimension grows, which underscores the balance between enhancing model performance and controlling computational complexity. An optimal balance is struck at 4 heads and 32-dimensional heads, maintaining a parameter count of 392K, which ensures both high accuracy and computational efficiency.

Additionally, the computational complexity of the proposed model is determined mainly by the feature extraction with cost $O(T \cdot d)$, the BiCoBlock attention mechanism with cost $O(N^2 \cdot h \cdot d)$, and trajectory decoding with cost $O(T \cdot d^2)$. By constraining the attention configuration to h=4 and d=32, a favourable trade-off between representational capacity and computational efficiency was attained. The overall per-sequence complexity is

$O(T \cdot N \cdot d^2)$, which scales linearly with sequence length and quadratically with agent count. This ensures practical applicability in typical traffic scenarios with moderate agent density.

Moreover, the empirical evaluation demonstrates real-time inference capability across various deployment environments. The inference time by the model is 21.6 ms within real-time requirements, where threshold is 50 ms. For edge deployment in autonomous vehicles, a 30% speed improvement is obtained via INT8 quantization, and latency is reduced to approximately 15 ms. Memory-efficient implementations enable deployment on resource-constrained mobile platforms, achieving approximately 40 ms inference time while maintaining competitive accuracy.

5. CONCLUSION

An unsignalized intersection with four types of participants is used as the studied object. A unified modelling framework named Bi-STANet is proposed, which integrates temporal modelling, spatiotemporal feature extraction, and interaction modelling. A grid-based spatiotemporal feature representation of participants is constructed using a (2+1)D CNN, which factorizes 3D convolution into sequential 2D spatial and 1D temporal operations to reduce computational cost while preserving modelling capacity, and a novel bidirectional interaction module, BiCoBlock, is introduced to enable deep fusion of temporal and spatiotemporal features. Ablation studies confirm the individual contributions of each core module to overall performance. Furthermore, parameter sensitivity analysis demonstrates that appropriate configurations of channel number, hidden layer dimension, and attention settings can achieve a favourable balance between prediction accuracy and computational efficiency. The advantages of the proposed method are as follows.

1) The multimodal spatiotemporal modelling is enhanced to improve intent recognition accuracy and robustness through the proposed Bi-STANet model. LSTM and (2+1)D CNN are integrated to effectively encode the historical motion sequences of traffic participants and the dynamic environments of intersections. The incorporation of the BiCoBlock module facilitates deep bidirectional interactions between temporal and spatiotemporal features. Furthermore, strong generalization performance is demonstrated by the model under class imbalance conditions, which highlights its superior robustness and adaptability in complex, real-world traffic scenarios.
2) Trajectory prediction precision is improved through the modelling of complex interactions. Experimental results demonstrate that Bi-STANet significantly outperforms baseline models in terms of ADE and FDE. Compared to the traditional LSTM+CNN architecture, ADE is reduced by approximately 41.4% and FDE by 39.0%. The predicted trajectories closely align with the ground truth, particularly in complex turning scenarios, validating the model's effectiveness in capturing dynamic interactive behaviours.
3) The performance-efficiency trade-off is optimized through module synergy and parameter tuning. Ablation studies and parameter sensitivity analysis reveal the complementary roles of LSTM, (2+1)D CNN, and BiCoBlock in spatiotemporal feature extraction and interaction modelling. The synergistic integration of these modules substantially enhances overall model performance. Moreover, when the number of CNN channels and LSTM hidden units is set to 32, and the number of attention heads, dimension to 4, 32, respectively, Additionally, the proposed framework exhibits favourable computational efficiency and real-time inference performance, enabling practical deployment across server, edge, and mobile platforms.

There are some limitations that need to be improved. For example, the current model may experience performance degradation when the traffic demand is high. The model's robustness against malicious or abnormal behaviours is limited, and such irregular behavioural patterns were observed to exacerbate spatiotemporal decoupling, resulting in significant prediction errors and reduced system reliability. The spatiotemporal decoupling strategy in (2+1)D CNN, although computationally efficient, is prone to overlooking critical spatiotemporal interaction patterns and strong temporal dependencies in specific traffic scenarios, e.g., deceleration decisions at intersection conflict zones. To address these challenges, the Bi-STANet framework should be extended to accommodate multi-agent interactions in more diverse and unstructured environments, incorporating multi-scale spatiotemporal fusion to capture both local and global dependencies, applying positional conditioning to retain context-specific temporal dynamics, and diversifying training scenarios to encompass higher-interaction and more complex traffic conditions. Moreover, the model compression strategies e.g., pruning and attention-head reduction should be explored to reduce computational load and support real-time deployment on edge devices.

ACKNOWLEDGMENT

The work described in this paper was supported by the National Natural Science Foundation of China (50478088), and the Cooperation Special Project of Beijing-Tianjin-Hebei Basic Research in China (F2024202106). The authors gratefully acknowledge the editor's comments and the reviewers of the paper who helped to clarify and improve the presentation.

REFERENCES

[1] Zhu D, Khan Q, Cremers D. Multi-vehicle trajectory prediction and control at intersections using state and intention information. *Neurocomputing*. 2024;574:127220. DOI: 10.1016/j.neucom.2023.127220.

[2] Li Z, Gong J, Lu C, Yi Y. Interactive Behavior Prediction for Heterogeneous Traffic Participants in the Urban Road: A Graph-Neural-Network-Based Multitask Learning Framework. *IEEE/ASME Transactions on Mechatronics*. 2021;26(3):1339-49. DOI: 10.1109/TMECH.2021.3073736.

[3] Mo X, Huang Z, Xing Y, Lv C. Multi-Agent Trajectory Prediction With Heterogeneous Edge-Enhanced Graph Attention Network. *IEEE Transactions on Intelligent Transportation Systems*. 2022;23(7):9554-67. DOI: 10.1109/TITS.2022.3146300.

[4] Li Z, Lu C, Yi Y, Gong J. A Hierarchical Framework for Interactive Behaviour Prediction of Heterogeneous Traffic Participants Based on Graph Neural Network. *IEEE Transactions on Intelligent Transportation Systems*. 2022;23(7):9102-14. DOI: 10.1109/TITS.2021.3090851.

[5] Zyner A, Worrall S, Nebot E. A Recurrent Neural Network Solution for Predicting Driver Intention at Unsignalized Intersections. *IEEE Robotics and Automation Letters*. 2018;3(3):1759-64. DOI: 10.1109/LRA.2018.2805314.

[6] Azadani MN, Boukerche A. STAG: A novel interaction-aware path prediction method based on Spatio-Temporal Attention Graphs for connected automated vehicles. *Ad Hoc Networks*. 2023;138:103021. DOI: 10.1016/j.adhoc.2022.103021.

[7] Byeon H, et al. Reinforcement Learning for Dynamic Optimization of Lane Change Intention Recognition for Transportation Networks. *IEEE Transactions on Intelligent Transportation Systems*. 2025;1-11. DOI: 10.1109/TITS.2025.3529299.

[8] Qiao S, Gao F, Wu J, Zhao R. An Enhanced Vehicle Trajectory Prediction Model Leveraging LSTM and Social-Attention Mechanisms. *IEEE Access*. 2024;12:1718-26. DOI: 10.1109/ACCESS.2023.3345643.

[9] Hosford K, Cloutier M-S, Winters M. Observational Study of Pedestrian and Cyclist Interactions at Intersections in Vancouver, BC and Montréal, QC. *Transportation research record*. 2020;2674(6):410-9. DOI: 10.1177/0361198120919407.

[10] Benhelal MS, Jouaber B, Afifi H, Moungla H. Towards Edge-Assisted Trajectory Prediction for Connected Autonomous Vehicles. *In GLOBECOM 2023-2023 IEEE Global Communications Conference*. 2023. p. 3741-3746. DOI: 10.1109/GLOBECOM54140.2023.10437498.

[11] Xin L, et al. Intention-aware Long Horizon Trajectory Prediction of Surrounding Vehicles using Dual LSTM Networks. *In 2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. 2018. p. 1441-1446. DOI: 10.1109/ITSC.2018.8569595.

[12] Huang M, Qian H, Han Y, Xiang W. R(2+1)D-based Two-stream CNN for Human Activities Recognition in Videos. *In 2021 40th Chinese Control Conference (CCC)*. 2021. p. 7932–7937. DOI: 10.23919/CCC52363.2021.9549432.

[13] Zou Y, et al. Two-Stream (2+1)D CNN Based on Frame Difference Attention for Driver Behavior Recognition. *In2023 10th International Conference on Dependable Systems and Their Applications (DSA)*. 2023. p. 782–788. DOI: 10.1109/DSA59317.2023.00110.

[14] Chen L, et al. A short-term traffic flow prediction model for road networks using inverse isochrones to determine dynamic spatiotemporal correlation ranges. *Physica A: Statistical Mechanics and Its Applications*. 2025;657:130244. DOI: 10.1016/j.physa.2024.130244.

[15] Jiang Y, et al. A spatiotemporal optimization method for connected and autonomous vehicle operations in long tunnel constructions. *Physica A: Statistical Mechanics and Its Applications*. 2024;651:130041. DOI: 10.1016/j.physa.2024.130041.

[16] Wu K, et al. Graph-Based Interaction-Aware Multimodal 2D Vehicle Trajectory Prediction Using Diffusion Graph Convolutional Networks. *IEEE Transactions on Intelligent Vehicles*. 2024;9(2):3630-43. DOI: 10.1109/TIV.2023.3341071.

[17] Meng Z, Zhao H, Tan W, Wang D. A Novel Approach for Stratifying Pulmonary Edema Severity on Chest X-ray via Dual-Mechanic Self-Learning and Bidirectional Multi-Modal Cross-Attention Algorithms. *In Journal of Physics: Conference Series*. 2024;2829:012019. DOI: 10.1088/1742-6596/2829/1/012019.

[18] Chen Y, et al. Bidirectional feature fusion via cross-attention transformer for chrysanthemum classification. *Pattern Analysis and Applications*. 2025;28(2):41. DOI: 10.1007/s10044-025-01419-8.

[19] Liu D, Mao Q, Gao L, Wang G. Leveraging Contrastive Language–Image Pre-Training and Bidirectional Cross-attention for Multimodal Keyword Spotting. *Engineering Applications of Artificial Intelligence*. 2024;138:109403. DOI: 10.1016/j.engappai.2024.109403.

[20] Li X, et al. Multimodal temperature prediction for lithium-ion battery thermal runaway using multi-scale gated fusion and bidirectional cross-attention mechanisms. *Journal of Energy Storage*. 2025;116:116098. DOI: 10.1016/j.est.2025.116098.

[21] Ren B, et al. A data-driven approach to traffic vehicle intent recognition and trajectory prediction. *IEEE Transactions on Intelligent Vehicles*. 2024:1-10. DOI: 10.1109/TIV.2024.3484494.

[22] Xu H, et al. Behavior recognition of non-motorized transport at intersections using dual-channel grid model based on disordered trajectory point data. *Physica A: Statistical Mechanics and Its Applications.* 2024;650:129994. DOI: 10.1016/j.physa.2024.129994.

[23] Lin M, et al. A 3D Convolution-Incorporated Dimension Preserved Decomposition Model for Traffic Data Prediction. *IEEE Transactions on Intelligent Transportation Systems*. 2025;26:673-90. DOI: 10.1109/TITS.2024.3486963.

[24] Tran D, et al. A Closer Look at Spatiotemporal Convolutions for Action Recognition. *InProceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2018. p. 6450-6459. DOI: 10.1109/cvpr.2018.00675.

[25] Yang L, et al. Lite-FPN for keypoint-based monocular 3D object detection. *Knowledge-Based Systems*. 2023;271:110517. DOI: 10.1016/j.knosys.2023.110517.

[26] Zhang Y, He Y, Zhang L. Recognition method of abnormal driving behavior using the bidirectional gated recurrent unit and convolutional neural network. *Physica A: Statistical Mechanics and Its Applications*. 2023;609:128317. DOI: 10.1016/j.physa.2022.128317.

[27] Wang Y, et al. Multi-Vehicle Collaborative Learning for Trajectory Prediction With Spatio-Temporal Tensor Fusion. *IEEE Transactions on Intelligent Transportation Systems*. 2022;23(1):236-48. DOI: 10.1109/TITS.2020.3009762.

[28] Bock J, et al. The inD Dataset: A Drone Dataset of Naturalistic Road User Trajectories at German Intersections. *In2020 IEEE Intelligent Vehicles Symposium (IV)*. 2020. p. 1929-1934. DOI: 10.48550/arXiv.1911.07602.

[29] Xie G, et al. Motion trajectory prediction based on a CNN-LSTM sequential model. *Science China Information Sciences*. 2020;63(11):212207. DOI: 10.1007/s11432-019-2761-y.

[30] Yang W, et al. Method of Predicting Braking Intention Using LSTM-CNN-Attention With Hyperparameters Optimized by Genetic Algorithm. *International Journal of Control, Automation and Systems*. 2024;22(7):2301-12. DOI: 10.1007/s12555-021-1113-x.

[31] Min H, Xiong X, Wang P, Zhang Z. A Hierarchical LSTM-Based Vehicle Trajectory Prediction Method Considering Interaction Information. *Automotive Innovation*. 2024;7(1):71-81. DOI: 10.1007/s42154-023-00261-0.