ViTraj: Learning Dual-Side Representations for Vehicle-Infrastructure Cooperative Trajectory Prediction

Shengzhe You
Zhejiang University of Technology
Hangzhou, Zhejiang, China
ysz980705@163.com

Libo Weng
Zhejiang University of Technology
Hangzhou, Zhejiang, China
wenglibo@zjut.edu.cn

Fei Gao*
Zhejiang University of Technology
Hangzhou, Zhejiang, China
feig@zjut.edu.cn

Abstract

While autonomous driving has made substantial progress, accurately predicting the trajectories of surrounding traffic agents remains a fundamental challenge for ensuring safety. Integrating both infrastructure-side and vehicle-side information has the potential to enhance perception and prediction capabilities. However, existing methods overlook the challenges in Vehicle-Infrastructure Cooperative Trajectory Prediction. To bridge this gap, we propose ViTraj, a model-agnostic framework for VIC-TP that leverages infrastructure-side trajectories to mitigate the inherent limitations of vehicle-side forecasting. ViTraj introduces a Feature-Side Selection and a Cooperative Interaction to aggregate complementary features from both sides, effectively expanding the perceptual horizon of prediction models. In addition, we present a Vehicle-Infrastructure Knowledge Distillation strategy to enforce consistency between multi-side predictions, which efficient global-local feature alignment through a single backward pass. Extensive experiments on large-scale public datasets demonstrate that ViTraj consistently improves advanced trajectory prediction models, achieving the state-of-the-art performance compared to existing vehicle-infrastructure cooperative methods. We believe this work provides a promising step toward the practical deployment of V2X-based autonomous driving systems.

CCS Concepts

• **Computing methodologies** → **Computer vision representations**; **Knowledge representation and reasoning**; • **Applied computing** → **Transportation**.

Keywords

Vehicle-Infrastructure Cooperation, Autonomous Driving, Trajectory Prediction

ACM Reference Format:

Shengzhe You, Libo Weng, and Fei Gao. 2025. ViTraj: Learning Dual-Side Representations for Vehicle-Infrastructure Cooperative Trajectory Prediction. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), Oct. 27–31, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3746027.3755295

*Corresponding author.

MM '25, Dublin, Ireland

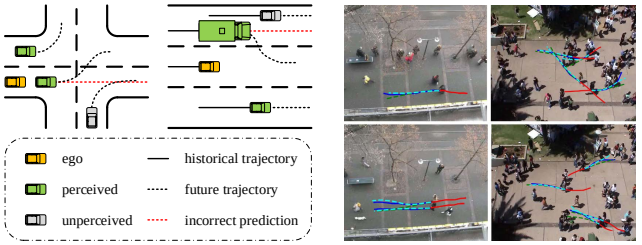
ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3755295

1 Introduction

Trajectory prediction remains one of the most critical and challenging problems in autonomous driving [15, 18]. To safely navigate complex traffic scenarios, autonomous vehicles must accurately perceive their surroundings and forecast the future trajectories of agents. The autonomous driving pipeline typically consists of three key stages: (1) Perception, which involves detecting traffic participants and estimating their future states [27]; (2) Decision-making, which plans the vehicle's next actions [21]; and (3) Control, which executes decisions based on control theory [24]. Trajectory prediction lies at the core of this pipeline, aims to forecast future movements of surrounding agents based on historical trajectories, sensor signals, and spatial map information. Accurate trajectory prediction is essential for robust path planning and decision-making, enabling vehicles to avoid potential hazards. In recent years, deep learning-based methods [26] have shown great promise and have been widely adopted for trajectory prediction. Nevertheless, achieving consistently reliable performance remains a significant challenge. This raises two important questions: Why is trajectory prediction so inherently difficult, and what are the limitations of existing methods? In this paper, we explore these issues from two perspectives: vehicle-side trajectory prediction (VS-TP) and infrastructure-side trajectory prediction (IS-TP).

(1) VS-TP has made notable progress in autonomous driving. However, its performance remains limited in complex scenarios due to the restricted perception range. As illustrated in Fig. 1(a), autonomous vehicles rely on onboard sensors to perceive the environment, capturing data such as high-definition vector maps and agents motion states. With the advancement of graph neural networks (GNNs) and encoder-decoder architectures [22, 25], deep learning models have gradually replaced traditional approaches [36]. In particular, attention-based architectures have improved interaction modeling among agents, significantly enhancing prediction accuracy [28]. However, VS-TP methods are constrained by vehicle's limited field of view, which restricts scene representations and interaction reasoning [20]. As shown in Fig. 1(a), occlusions from large obstacles (e.g., trucks or buses) hinder the vehicle's ability to perceive distant agents, making future trajectory prediction unreliable or even infeasible. Since minor errors in prediction and decision-making can result in critical failures, the limited perception scope remains a major bottleneck for the deployment of VS-TP methods.

(2) Traffic agents exhibit both long-term and short-term interactions, their complexity increase as scene dynamics evolve. Compared to vehicle-side perception, infrastructure-mounted sensors offer a broader field of view, enabling more effective modeling of global interactions, as shown in Fig. 1(b). Existing IS-TP methods



(a) vehicle-side trajectory prediction (VS-TP) (b) infrastructure-side trajectory prediction (IS-TP)

Figure 1: Two widely studied trajectory prediction.

typically rely on fixed-position 2D cameras and employ deep neural networks to extract motion states and predict future trajectories [10]. However, due to the inherent limitations of 2D images and low spatial accuracy, these methods struggle to meet the high-precision requirements of autonomous driving. Recent research has explored the potential of infrastructure-side data to enhance perception. Some works define cooperative 3D object detection and highlight the challenges of collaborative perception [44], while others investigate feature-level transmission from infrastructure to vehicles [45] or apply infrastructure-side data for vehicle control [41]. Nonetheless, most existing efforts focus on perception tasks, leaving effective solutions for VIC-TP largely unexplored.

This paper revisits the limitations of existing methods in vehicle-infrastructure cooperative scenarios, discussing how infrastructure-side information can mitigate performance bottlenecks caused by the constrained vehicle-side perspective. To address this challenge, we propose ViTraj, a model-agnostic framework for vehicle-infrastructure cooperative trajectory prediction (VIC-TP). ViTraj introduces a cooperative interaction module to effectively fuse vehicle-side and infrastructure-side information, thereby expanding perception range. Furthermore, we enforce consistency across predictions from vehicle-side, infrastructure-side, and cooperative features, encouraging the model to learn robust representations from a single side. The key contributions of this work are summarized as follows:

- We investigate the practical yet challenging problem of vehicle-infrastructure cooperative trajectory prediction and propose ViTraj, a model-agnostic VIC-TP framework. By incorporating Feature Side Selection (FeSS) and a COoperative Interaction (COIN) module, ViTraj effectively fuses vehicle-side and infrastructure-side information, significantly enhancing the perception range.
- We propose Vehicle-Infrastructure Knowledge Distillation (VIKD), which enforces consistency across multi-side predictions by measuring singular value deviations among vehicle-side, infrastructure-side, and cooperative features. This strategy aims to effectively learn local representations from vehicle-side data, global representations from infrastructure-side data, and complementary cooperative features.
- ViTraj is evaluated on the real-world public dataset V2X-Seq[46], demonstrating strong effectiveness and generalizability across a range of advanced prediction models.

2 Related Works

2.1 Vehicle-Side Trajectory Prediction

Predicting the behavior of dynamic obstacles is essential for autonomous driving systems. Traditional methods rely on manual feature engineering and rule-based approaches. These methods struggle to capture nonlinear behaviors and complex interactions in dynamic traffic scenarios [19]. With the rise of deep learning, Recurrent Neural Networks (RNNs) have facilitated the modeling nonlinear dependencies of trajectories [30, 32]. Given the pervasive agent interactions in traffic scenarios, Graph Neural Networks (GNNs) are increasingly used to capture spatiotemporal dependencies, enhancing multi-agent prediction in dense environments [34]. Inspired by the success of Transformers and Mamba [16, 31, 50], recent studies have incorporated space state models and self-attention mechanisms to model spatiotemporal relationships and interactions between agents and map elements [14, 52, 53]. Current models rely on searching relevant traffic agents, some select the k-nearest neighbors of the target agent [9, 13], while others consider agents within a fixed distance in the adjacent lanes [8]. Static selections are not applicable to all traffic scenarios. Moreover, as shown in Fig. 1(a), limited perception range and occlusions prevent vehicle-mounted sensors from capturing global scene information. Therefore, integrating infrastructure-side data is crucial to expand perception and enhance prediction accuracy.

2.2 Infrastructure-Side Trajectory Prediction

The rapid advancement of computer vision and the widespread deployment of surveillance cameras have supported the development of IS-TP methods [5, 48, 49]. The key distinction is that the sensor placement of IS-TP is at elevated positions, offering a wider perception for modeling global agent interactions, as shown in Fig. 1(b). Existing IS-TP methods primarily rely on 2D image inputs [51]. Research in IS-TP has progressed from traditional methods to deep learning. The introduction of RNN-based models, such as Social-LSTM, marked a milestone by modeling social interactions using LSTM units [42]. Additionally, numerous studies combines motion information with infrastructure-side images [38]. GNNs construct agent-interaction graphs and embed these into trajectory prediction frameworks [43]. Despite advances in IS-TP, its fixed sensing range and the low localization accuracy of 2D images constrain its application in autonomous driving. Although numerous algorithms have been proposed over the past decade, methods leveraging vehicle-infrastructure cooperation remain scarce [2]. Recent studies, such as V2X-Graph, attempt to explore this direction [33]. However, due to the diversity of vehicle-side and infrastructure-side prediction paradigms, a generalized framework for cooperative trajectory prediction is still lacking. This gap between trajectory prediction research and real-world deployment highlights the need for further exploration in this area.

2.3 Vehicle-Infrastructure Cooperative Autonomous Driving

Recent progress in autonomous driving has been driven by large-scale datasets such as nuScenes, Waymo, Argoverse, and KITTI [3, 4, 11, 35]. However, vehicle-mounted sensors are limited by visibility

and occlusions, resulting in reduced performance in long-range or obstructed areas. A promising solution is vehicle-infrastructure cooperative autonomous driving, which leverages infrastructure-side data to enhance perception. Most existing work focuses on cooperative perception. DAIR-V2X introduces vehicle-infrastructure cooperative object detection and analyzing the benefits of infrastructure data [44]. Some methods embed cross-side extraction techniques and are plug-and-play to any models [6, 37, 47]. Subsequent studies have explored feature encoding for data sharing, balancing detection performance and transmission latency [7], as well as delay compensation modules and feature prediction to mitigate latency-induced degradation [29, 45]. More recently, research has extended to cooperative segmentation [12]. Existing studies on vehicle-infrastructure cooperative autonomous driving overlook motion prediction. The V2X-Seq dataset firstly supports VIC-TP and has demonstrated the effectiveness of incorporating infrastructure-side trajectory [46]. V2X-Graph represents the first dedicated approach for cooperative trajectory prediction, leveraging graph networks to model agent motion and interactions [33]. We further propose a model-agnostic cooperative framework ViTraj for extending to integrate various vehicle-side trajectory prediction models.

3 Preliminary

Given the historical cooperative trajectory $\mathcal{M} = \{\mathcal{M}_{\text{veh}}, M_{\text{inf}}\}$. $\mathcal{M}_{\text{veh}}$ denotes the local historical trajectory data perceived by the vehicle-side, and $\mathcal{M}_{\text{inf}}$ represents the global historical trajectory data captured by the infrastructure-side. Notably, vehicle-side and infrastructure-side trajectories often overlap, as a vehicle's motion can be simultaneously observed from both perspectives. Thus, the total number of cooperative trajectories is given by $N_{\text{vic}} = N_{\text{veh}} + N_{\text{inf}} - N_{\text{veh}\cap\text{inf}}$. N_{veh} and N_{inf} denote the number of trajectories observed from the vehicle-side and infrastructure-side perspectives, respectively. And $N_{\text{veh}\cap\text{inf}}$ represents the number of overlapping trajectories. The cooperative trajectory set is defined as $\mathcal{M} = \{m_0, m_1, ..., m_N\} \in \mathbb{R}^{N\times T\times C}$, where each trajectory has length T, and includes attributes C such as position, heading angle, and agent type. In most prediction settings, the motion state of each trajectory is given by $m_i = \{P_t^i, R_t^i\}$, where $P_t^i = \{x, y\}$ represents the position of agent i at time t, and R_t^i denotes its heading angle. Trajectory prediction typically use a vectorized representation for map elements, often leveraging lane centerline points to encode spatial information. This work follow the setting of V2X-Graph [33] and represent the set of map elements as $G \in \mathbb{R}^{N_l\times 2\times C_l}$, where N_l is the number of lanes and C_l denotes lane attributes such as position and road type. Finally, the VIC-TP task is defined as $\mathcal{Y} = \rho(\mathcal{M}, G)$. Here, $\mathcal{Y} = \{\mathcal{M}_{\text{target}}\}$ denotes future cooperative trajectory data. The prediction model is denoted as ρ. The objective is to exploit key information from cooperative trajectories and model interactions between vehicle-side and infrastructure-side trajectories to enhance prediction.

4 Method

This section introduces ViTraj, a model-agnostic framework for VIC-TP. As illustrated in Fig. 2, ViTraj can be seamlessly integrated with various prediction models: (1) Feature Side Selection (FeSS): Identifies and selects relevant cooperative trajectory data; (2) Cooperative Interaction (COIN): Fuses vehicle-side and infrastructure-side features; (3) Vehicle-Infrastructure Knowledge Distillation (VIKD): Enforces consistency across multi-side inputs to enhance model robustness; (4) Trajectory Prediction Model: Utilizes an encoder-decoder architecture for autonomous driving.

4.1 Feature Side Selection

Feature selection is an optional component in ViTraj. In practice, pointwise multiplication between cooperative data and a vehicle-infrastructure mask provides an intuitive feature selection approach [46]:

$$\mathcal{M}_* = \mathcal{M} \cdot msk_*, * = \text{veh}, \text{inf} \tag{1}$$

where $\mathcal{M} \in \mathbb{R}^{N\times T\times C}$ denotes the cooperative trajectory data, and $msk_* \in \mathbb{R}^{N\times T}$ represents the mask for a specific side (vehicle-side or infrastructure-side), which indexes the associated trajectory within the cooperative data. Rather than applying rigid feature selection, we introduce Feature Side Selection (FeSS) based on a gating mechanism to enable flexible vehicle-infrastructure feature selection. FeSS captures the spatiotemporal correlations in cooperative trajectories and generates differentiated inputs to support subsequent feature fusion. The process begins with positional encoding of the cooperative trajectory:

$$f_{\text{vic}} = \mathcal{M} + \text{F}_{\text{embed}}(\mathcal{M}) \in \mathbb{R}^{N\times T\times C} \tag{2}$$

where $f_{\text{vic}} = \{f_1, f_2, ..., f_N\}$ denotes the encoded trajectory features, and $f_1 \in \mathbb{R}^{T\times C}$; F_{embed} represents the positional encoding function that generates learnable embeddings based on the sequence $\mathcal{M}$. Positional encoding assigns unique identifiers to sequence positions, allowing the model to capture temporal order. Feature side selection is performed using a masked gating layer:

$$g_{\text{veh}} = \text{MLP}_{\text{G}}(msk_{\text{veh}}), g_{\text{inf}} = \text{MLP}_{\text{G}}(msk_{\text{inf}}) \tag{3}$$

$$f_{\text{veh}} = \sigma(g_{\text{veh}}) \cdot f_{\text{vic}}, f_{\text{inf}} = \sigma(g_{\text{inf}}) \cdot f_{\text{vic}} \tag{4}$$

where MLP_{G} denotes a multilayer perceptron that generate gating weights. The cooperative mask msk_* is processed by MLP_{G} to produce gating value g_*, which are then scaled to the range [0,1] using a sigmoid function. Given the encoded cooperative feature f_{vic}, the vehicle-side and infrastructure-side trajectory features f_{veh} and f_{inf} are obtained by element-wise multiplication. This module extracts side-specific features and captures the spatiotemporal correlations within cooperative trajectories.

4.2 Cooperative Interaction

COoperative INteraction (COIN) learns the relationships between vehicle-side and infrastructure-side trajectories, integrating them into cooperative features. After position encoding and feature selection, both vehicle-side and infrastructure-side data are input into the encoder of the prediction model. To efficiently capture the spatiotemporal knowledge of both sides, the encoders share weights:

$$H_{\text{veh}} = \text{Enc}(f_{\text{veh}}), H_{\text{inf}} = \text{Enc}(f_{\text{inf}}) \tag{5}$$

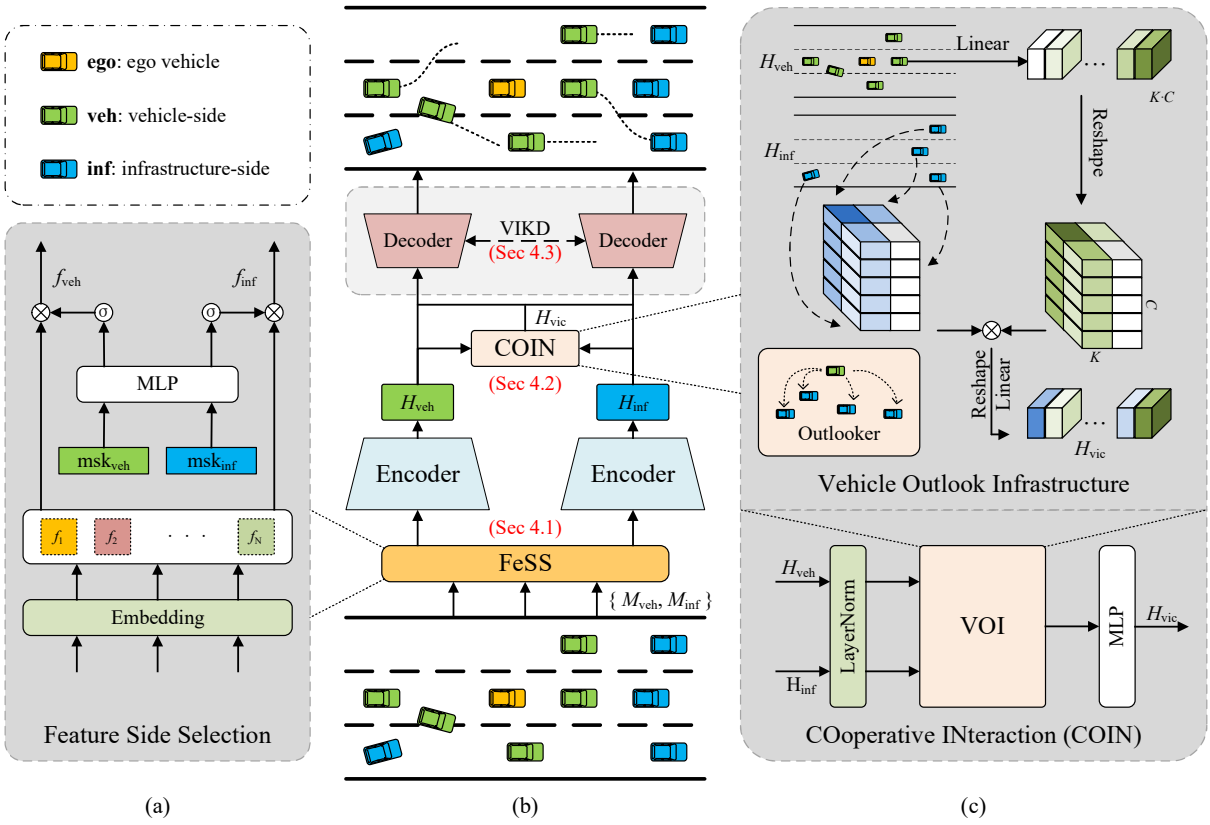


Figure 2: (a) FeSS (b) Overview of the proposed ViTraj. (c) COIN based on vehicle outlook infrastructure(VOI).

where Enc denotes the encoder of any trajectory prediction model. After serializing and modeling single-side features, the encoded features are denoted as $\{H_{\text{veh}}, H_{\text{inf}}\}$. COIN consists of a Vehicle Outlook Infrastructure for spatial relationship interaction and a multilayer perceptron for capturing temporal dependencies. Given the vehicle-side representation $H_{\text{veh}} \in \mathbb{R}^{N_{\text{veh}} \times T \times C}$ and infrastructure-side representation $H_{\text{inf}} \in \mathbb{R}^{N_{\text{inf}} \times T \times C}$, the COIN can be expressed as:

$$Z_{\text{vic}} = \text{VOI}(\text{LN}(H_{\text{veh}}), \text{LN}(H_{\text{inf}})) \tag{6}$$

$$H_{\text{vic}} = \text{MLP}_{\text{vic}}(\text{LN}(Z_{\text{vic}}) + Z_{\text{vic}} \tag{7}$$

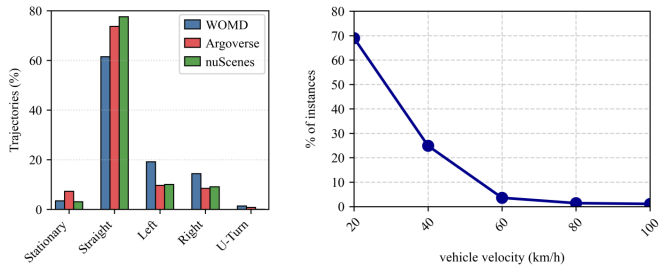
where LN denotes layer normalization. VOI refers to Vehicle Outlook Infrastructure, and MLP_{vic} represents the multilayer perceptron.

Vehicle Outlook Infrastructure: Unlike traditional trajectory prediction models, VIC-TP requires integrating dual-side features to enhance its perception range. V2VNet [40] and FFNet [45] are two advanced vehicle-infrastructure cooperative methods, which are extensively compared with ViTraj in the experimental section. V2VNet introduces a cross-agent aggregation module using convolutional neural networks and updates agent states via ConvGRU. FFNet adopt a unified features flow mechanism to align cross-side features efficiently. However, large-scale infrastructure-side trajectory data often contain redundant or irrelevant agents, which may not benefit prediction and can incur high computational costs. Existing methods aggregate all infrastructure-side information, limiting their effectiveness for VIC-TP. To address this, we propose the Vehicle Outlook Infrastructure (VOI), a selective aggregation that focuses on key infrastructure-side features, as shown in the right part of Fig. 2. The key idea is: (1) Each vehicle-side agent can generate a query weight for locally aggregating features from nearby agents. (2) Dense and localized aggregation reduces redundancy and extract fine-grained spatial cues for VIC-TP.

For each agent feature in H_{veh}, VOI computes its correlation with the K nearest neighbors. Formally, given an agent feature $h^i_{\text{veh}} \in \mathbb{R}^{N_{\text{inf}} \times T \times C}$, both are projected through layer normalization and linear layers $W^i_{\text{veh}} \in \mathbb{R}^{C \times K}$ and $W^i_{\text{inf}} \in \mathbb{R}^{C}$ to to obtain the outlook weight $h'_{\text{veh}} \in \mathbb{R}^{K \times T \times C}$ and $H'_{\text{inf}} \in \mathbb{R}^{N_{\text{inf}} \times T \times C}$, respectively. Let $H'_{\text{near}} \in \mathbb{R}^{K \times T \times C}$ denotes the K nearest features to this agent:

$$H_{\text{near}} = \underset{H'_{\text{near}} \subseteq H'_{\text{inf}}, |H'_{\text{near}}|=K}{\text{argmin}} \sum_{h \in H'_{\text{near}}} \text{dis}\big(\text{xy}(h), \text{xy}(h'_{\text{veh}})\big) \tag{8}$$

where dis denotes the Euclidean distance, and xy refers to the coordinates associated with trajectory features. The infrastructure-side weights are applied directly to outlook weights, and the cooperative



(a) Trajectory distribution (b) Speed distribution

Figure 3: Statistical analysis of trajectory datasets

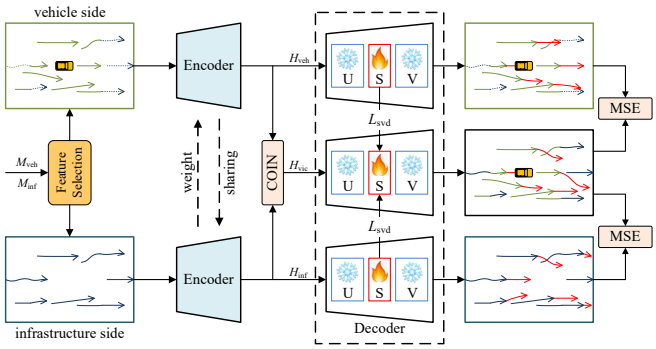


Figure 4: Vehicle-Infrastructure Knowledge Distillation

representation is obtained through the Softmax operation followed by a dot product:

$$H'_{\text{vic}} = \text{Matmul}(\text{Softmax}(\text{H}'_{\text{veh}}), \text{H}_{\text{near}}) \tag{9}$$

VOI performs dense aggregation of cooperative representations. The final output is computed by a weighted sum of the cooperative features associated with each vehicle-side feature:

$$z_{\text{vic}} = \sum_{t=0}^{K-1} H'_{\text{vic}}(i) \in \mathbb{R}^{T\times C} \tag{10}$$

$Z_{\text{vic}} = \{z_{\text{vic}}\}$ is obtained by VOI, and finally, a multilayer perceptron is used to obtain the vehicle-road collaborative features H_{vic} , as shown in Equation 7.

4.3 Vehicle-Infrastructure Knowledge Distillation

As shown in Fig. 3(a), statistical analysis of the WOMD, Argoverse, and nuScenes datasets [3, 4, 35] reveals that vehicle trajectories are predominantly linear. Additionally, Fig. 3(b) shows that over half of the vehicle instances maintain speeds below 20 km/h, further highlighting the sparsity of dynamic behavior in real-world trajectories [39]. These observations indicate the many linear trajectory features are redundant, and models should instead focus on learning features from non-linear movements. However, identifying such informative features remains challenging. Simply computing similarities among all features is expensive and may introduce redundancy to overfitting. Additionally, existing methods are typically designed for either vehicle-side or infrastructure-side data, they struggle to disentangle vehicle-side and infrastructure-side representations within fused features and extract representations beneficial to target agents. To address these limitations, this paper proposes a concise Vehicle-Infrastructure Knowledge Self-Distillation strategy (VIKD), as depicted in Fig. 4. The proposed VIKD incorporates two complementary loss components: SVD contrastive loss $\mathcal{L}_{\text{con}}$ and logit distillation loss $\mathcal{L}_{\text{pre}}$.

4.3.1 Singular Value Contrastive Distillation. $\mathcal{L}_{\text{con}}$ aims to encourage the model learn critical spatiotemporal knowledge from either vehicle-side or infrastructure-side data independently, thereby mitigating feature coupling. Specifically, we apply singular value decomposition to decompose the trajectory features into three sequential matrices and compute similarity only on singular values. Formally, given vehicle-infrastructure feature $H_* \in \mathbb{R}^{B\times N\times D} (* = \text{veh}, \text{inf}, \text{vic})$, where B is the batch size, N is the number of trajectories, and D is the feature dimension. We preserve the top-R singular values with $R = \min(N, D)$ to identify key trajectory features. The decomposition is represented as:

$$H_* = U_* S_* V_*^{\text{T}}, U_* \in \mathbb{R}^{B\times N\times R}, S_* \in \mathbb{R}^{B\times R\times R}, V_* \in \mathbb{R}^{B\times R\times D} \tag{11}$$

The matrices U_* and V_* represent two new trajectory features, while S_* is a diagonal matrix containing singular values. The trajectory feature f is decomposed into three sub-features: (1) projection into a rank-R subspace; (2) identification of key features via comparison of singular values; (3) reconstruction of the self-distilled feature back to the original space f. Based on this decomposition, we define the contrastive loss over the singular value matrix S as follows:

$$\mathcal{L}_{\text{svd}}(S_*) = \frac{\text{diag}(S_{\text{vic}}) \times [\text{diag}(\text{S}_*)]^{\text{T}}}{|\text{diag}(S_{\text{vic}})||[\text{diag}(\text{S}_*)]|}, * \in \{\text{veh}, \text{inf}\} \tag{12}$$

where diag() denotes extracting the diagonal elements, where $*$=veh, inf. The complete singular value contrastive loss $\mathcal{L}_{\text{con}}$ is formulated as:

$$\mathcal{L}_{\text{con}} = \mathcal{L}_{\text{svd}}(S_{\text{veh}}) + \mathcal{L}_{\text{svd}}(S_{\text{inf}}) \tag{13}$$

4.3.2 Logit Distillation. While Singular Value Contrastive Distillation facilitates the learning of nonlinear key features from high-dimensional, sparse, dual-side representations, it lacks the ability to capture global dependencies in dual-side trajectories. To address this, we introduce a trajectory-level logit distillation loss $\mathcal{L}_{\text{pre}}$, which enforces prediction consistency between vehicle-side and infrastructure-side outputs. $\mathcal{L}_{\text{pre}}$ compels the model to generate identical trajectory predictions from either single-side input. The proposed logit distillation loss is defined as follows:

$$\mathcal{L}_{\text{pre}} = \sum \mathcal{L}_2(\rho([\mathcal{M}_{\text{veh}}, \mathcal{M}_{\text{inf}}], G), \rho(\mathcal{M}_*, G)), * \in \{\text{inf}, \text{veh}\} \tag{14}$$

Table 1: Quantitative Comparison

Prediction Model	Cooperative Method	K=1			K=3			K=6		
		↓ ADE	↓ FDE	↓ MR	↓ ADE	↓ FDE	↓ MR	↓ ADE	↓ FDE	↓ MR
HiVT [53]	w/o fusion	3.08	6.32	0.79	1.93	4.02	0.61	1.44	2.52	0.36
	PP-VIC [46]	2.47	5.90	0.77	1.77	3.86	0.58	1.26	2.34	0.35
	V2VNet [40]	2.55	5.98	0.77	1.82	3.92	0.60	1.27	2.30	0.35
	FFNet [45]	2.44	5.63	0.76	1.76	3.77	0.55	1.30	2.25	0.34
	ViTraj (Ours)	**2.20**	**5.19**	**0.73**	**1.59**	**3.37**	**0.52**	**1.15**	**2.06**	**0.31**
ADAPT [1]	w/o fusion	2.92	6.05	0.76	1.75	3.70	0.56	1.49	2.61	0.37
	PP-VIC [46]	2.42	5.80	0.75	1.62	3.48	0.52	1.29	2.40	0.35
	V2VNet [40]	2.78	6.08	0.78	1.80	3.66	0.56	1.54	2.69	0.38
	FFNet [45]	2.57	5.92	0.74	1.69	3.66	0.53	1.43	2.46	0.35
	ViTraj (Ours)	**2.17**	**5.15**	**0.70**	**1.49**	**3.10**	**0.47**	**1.17**	**2.09**	**0.30**
LAFormer [23]	w/o fusion	3.03	6.22	0.78	1.91	3.97	0.57	1.52	2.68	0.38
	PP-VIC [46]	2.46	5.87	0.76	1.77	3.85	0.54	1.31	2.43	0.36
	V2VNet [40]	2.91	6.12	0.78	1.87	3.81	0.57	1.45	2.66	0.38
	FFNet [45]	2.60	6.29	0.77	1.71	3.62	0.55	1.32	2.50	0.36
	ViTraj (Ours)	**2.28**	**5.38**	**0.75**	**1.52**	**3.14**	**0.49**	**1.19**	**2.11**	**0.31**

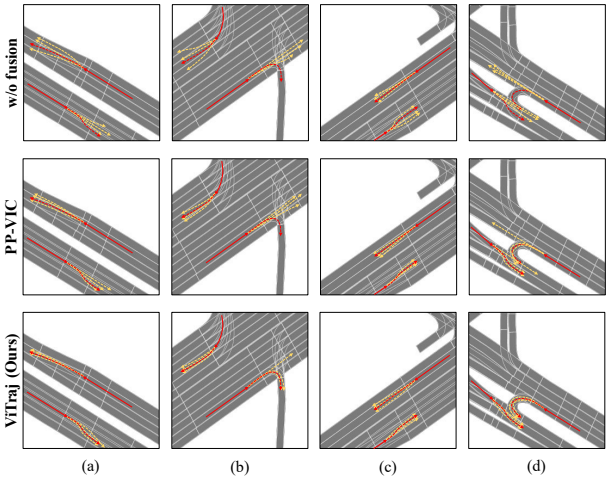


Figure 5: Qualitative results on the V2X-Seq dataset. Historical trajectories are shown as red solid lines, ground-truth future trajectories are depicted with red dashed lines, and predicted trajectories are illustrated using yellow dashed lines.

where ρ denotes the prediction model, $\mathcal{M}$ represents trajectories, G denotes the set of map elements. $\mathcal{L}_2$ is the mean squared error (MSE) function.

4.3.3 Joint Loss. Following prior work [53], we minimize the negative log-likelihood (NLL) between predicted and ground-truth trajectories of the target vehicle as the regression loss, formally defined as:

$$\mathcal{L}_{\text{traj}} = \sum \mathcal{L}_{\text{reg}}(y, \mathcal{M}) \tag{15}$$

where $\mathcal{L}_{\text{reg}}$ employs the probability density function of a Laplace distribution as the regression loss to jointly optimize the position and confidence of the optimal trajectory. The final loss function integrates trajectory regression loss $\mathcal{L}_{\text{traj}}$, logit distillation loss $\mathcal{L}_{\text{pre}}$, and singular value contrastive loss $\mathcal{L}_{\text{con}}$, with equal weights assigned to each component:

$$\mathcal{L} = \mathcal{L}_{\text{traj}} + \mathcal{L}_{\text{pre}} + \mathcal{L}_{\text{con}} \tag{16}$$

4.4 Trajectory Prediction Networks

Existing trajectory prediction networks typically utilize an encoder-decoder architecture. The encoder, responsible for serving as the backbone to extract trajectory features, commonly use networks such as recurrent neural networks (RNNs), graph neural networks (GNNs), or Transformers. The decoder usually consists of a multi-layer perceptron (MLP) cascading several fully connected layers to generate predictions.

Since ViTraj is a model-agnostic VIC-TP framework, the proposed feature side selection, cooperative interaction, and vehicle-infrastructure knowledge distillation are all independent of model architectures. Consequently, the proposed framework can be seamlessly integrated with various state-of-the-art trajectory prediction networks, enhancing prediction performance in cooperative vehicle-infrastructure scenarios. In this study, we evaluate ViTraj using three advanced methods, HiVT [53], ADAPT [1], and LAformer [23]. These models represent recent advancements in trajectory forecasting, and integrating ViTraj enhances their performance in vehicle-infrastructure cooperative settings.

5 Experiments

In this section, we conduct extensive experiments to evaluate the effectiveness of the proposed ViTraj. The experiment are designed to answer the following research questions:

RQ1: How does ViTraj compare with existing V2X trajectory prediction baselines? Can the proposed method improve the performance in cooperative driving scenarios?

RQ2: What are the contributions of individual components within ViTraj to the overall prediction accuracy?

Table 2: Ablation study on VIKD

Model	$\mathcal{L}_{traj}$	$\mathcal{L}_{pre}$	$\mathcal{L}_{svd}$	minADE	minFDE
HiVT [53]	✓			1.26	2.34
	✓	✓		1.18	2.29
	✓		✓	1.22	2.18
	✓	✓	✓	**1.15**	**2.06**
LAFormer [23]	✓			1.31	2.43
	✓	✓		1.23	2.36
	✓		✓	1.28	2.24
	✓	✓	✓	**1.19**	**2.11**

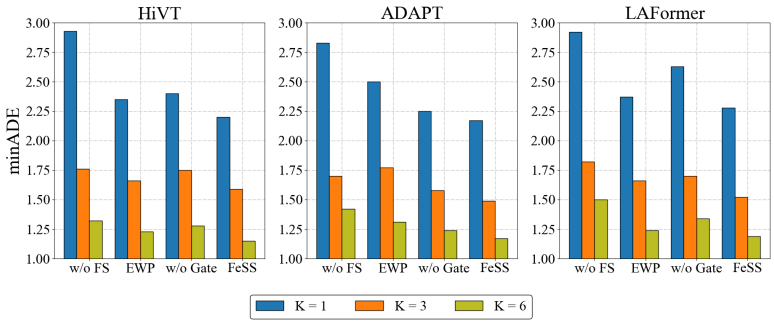


Figure 6: Ablation study on FeSS

RQ3: Does the Cooperative Interaction effectively capture the correlation between vehicle-side and infrastructure-side trajectories?

5.1 Settings

Datasets & **Baselines**: We evaluate ViTraj on V2X-Seq [46], the first large-scale dataset for vehicle-infrastructure cooperative trajectory prediction. This dataset contains over 210,000 real-world driving scenarios, including 51,146 V2X-cooperative scenes collected from 28 different regions, spanning 672 hours of driving data. Each scenario lasts for 10 seconds at a frame rate of 10 Hz. Given the first 5 seconds of historical trajectories from both vehicle and infrastructure sides, the task is to predict the target agent's motion over the next 5 seconds. Following prior work, we split the dataset into 50% for training, 20% for validation, and 30% for testing to ensure fair and comprehensive evaluation.

We evaluated the proposed VIC-TP framework against three state-of-the-art trajectory prediction networks (HiVT[53], ADAPT[1], and LAFormer[23]) and three representative cooperation methods (w/o fusion, PP-VIC[46], V2VNet[40], and FFNet[45]).

Metrics & **Implementation Details**: The proposed VIC-TP framework is evaluated using several widely adopted metrics, including minimum Average Displacement Error (minADE), minimum Final Displacement Error (minFDE), and Miss Rate (MR). Specifically, minADE measures the average displacement error between the best predicted trajectory and the ground truth across all future time steps, while minFDE measures the displacement error at the final prediction time step. MR is defined as the proportion of predicted trajectories whose endpoints deviate from the ground-truth endpoint by more than 2.0 meters. These metrics allow the model to output up to K predicted trajectories per agent. Following prior work, we evaluate performance with K = 1, 3, and 6 to provide a comprehensive assessment.

The proposed model is trained for 64 epochs using the Adam optimizer [17]. Following prior work [53], the batch size, learning rate, weight decay, and dropout rate are set to 32, 3e-4, 1e-4, and 0.1, respectively. The learning rate is scheduled using cosine annealing. All prediction models and experiments are implemented in PyTorch and run on a Linux server equipped with NVIDIA RTX A6000 GPUs.

5.2 RQ1: Performance Comparison

Quantitative Results. The proposed ViTraj is trained and evaluated on the V2X-Seq dataset [46]. Tab. 1 presents a comprehensive comparison of ViTraj with various fusion baselines across three trajectory prediction networks. As observed, all models yield unsatisfactory performance when no cooperative information is used (w/o fusion). This highlights that in complex traffic scenarios, single vehicle-side inputs are insufficient for accurate future trajectory prediction. While PP-VIC, as a straightforward fusion strategy, does offer marginal improvements over w/o fusion, the gains remain limited—suggesting that naive fusion fails to fully exploit environmental cues. V2VNet [40] and FFNet [45] represent classical and advanced V2X approaches, respectively. Although both attempt to utilize surrounding information through more sophisticated mechanisms, their performance is not consistently better than that of PP-VIC. Notably, in the single-trajectory prediction setting (K=1), PP-VIC even slightly outperforms them. This indicates that VIC-TP requires more than general cooperative perception methods—it demands an explicit modeling of the spatiotemporal correlations among agents. In contrast, ViTraj improves the performance of all baseline models, which achieves the best results in terms of minADE, minFDE, and MR across all values of K (1, 3, and 6). In summary, ViTraj consistently enhances a wide range of trajectory prediction models. These results demonstrate its strong generalization ability and practical applicability for autonomous driving and intelligent transportation systems.

Qualitative Results. To comprehensively evaluate the performance of the proposed method across diverse traffic scenarios, we provide a visual comparison of different fusion strategies on the V2X-Seq dataset in Fig. 5. For clarity and consistency, we employ HiVT [53] as the base trajectory prediction model, set the number of predicted trajectories to K=3, and compare ViTraj against both the w/o fusion and PP-VIC. In each test scene, we visualize predictions for two agents to facilitate intuitive analysis of differences between methods. In scenarios (a) and (c), incorporating infrastructure-side trajectory information effectively mitigates occlusion issues. Specifically, cooperative methods (PP-VIC and ViTraj) exhibit enhanced perception coverage, enabling them to anticipate lane-change maneuvers based on broader contextual cues. In scenes (b) and (d), the cooperative integration of both global and local perception enables accurate prediction of abrupt directional changes. Compared to

Table 3: Quantitative Comparison

Model	Cooperation			Param.(M)	FLOPs(G)	FPS
	VIKD	FeSS	COIN			
HiVT [53]				0.66	1.80	23.4
	✓			0.66	1.80	23.4
	✓	✓		0.68	1.82	22.7
	✓		✓	0.70	1.84	21.9
	✓	✓	✓	0.72	1.86	21.4

PP-VIC, ViTraj captures finer-grained correlations between the target agent and its neighboring agents, allowing it to produce more accurate and contextually reasonable predictions, particularly in complex traffic scenarios involving multiple interacting agents.

5.3 RQ2: Ablation Study

Effectiveness of Vehicle-Infrastructure Knowledge Distillation (VIKD). To validate the proposed VIKD, we compare the contributions of singular value contrastive distillation ($\mathcal{L}_{\text{svd}}$) and logit distillation ($\mathcal{L}_{\text{pre}}$). Tab. 2 presents the performance of several VIKD variants: (1) Introducing $\mathcal{L}_{\text{svd}}$ significantly reduces the prediction error at the trajectory endpoint. This proves the effect of aligning the singular value distributions, guiding the model to focus on salient features, such as sharp speed changes and turning points; (2) Combining $\mathcal{L}_{\text{pre}}$ further enhances the overall prediction accuracy. This observation indicates that logit distillation captures the global semantics and suppresses outlier predictions via backward.

Effectiveness of Feature-Side Selection (FeSS). To evaluate the proposed FeSS module, we construct three ablation variants: (1) w/o FS: removing FeSS entirely; (2) EWP: replacing the selection mechanism with element-wise product; (3) w/o Gate: removing the gating mechanism and applying a static vehicle-infrastructure mask. Fig. 6 presents minADE comparisons under: HiVT [53], ADAPT [1], and LAFormer [23]. Across all models and and predicted trajectories (K = 1/3/6), w/o FS performs the worst, confirming that noise and redundancy in raw features severely degrade prediction accuracy. Replacing EWP with FeSS further boosts the performance, indicating that side selection adapts to the varying feature correlations across complex scenes. Moreover, removing the learnable gating mechanism means the non-application of soft feature filtering. This absence limits the ability to decouple and select key features in VIC-TP.

Parameter Size and Inference Cost Comparison. We conduct inference experiments on a single NVIDIA RTX 3090 GPU and compare the computational costs. As a model-agnostic framework, ViTraj can be seamlessly integrated with various encoder-decoder-based trajectory prediction networks. VIKD is exclusively applied during the training phase, thus introducing zero additional inference overhead. For FeSS and COIN, we have quantified their computational requirements and parameter size, and compared them with the representative baseline HiVT [53].The results are summarized in the tab. 3. We observe that while ViTraj's implementation leads to a modest increase in model parameters (approximately 10%) and a slight reduction in inference speed (around 4 ms/sample), the overall inference latency remains well within practical limits for real-time applications.

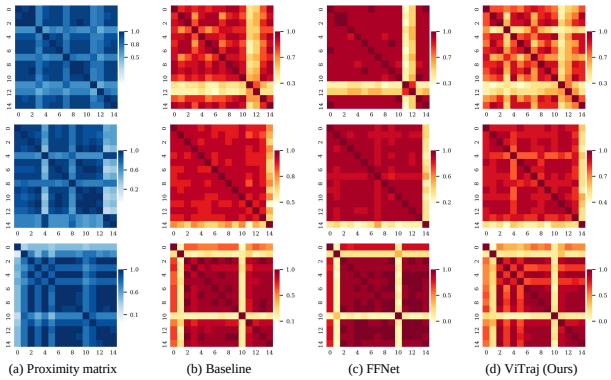


Figure 7: Visualization Comparison of the COIN. The baseline represent a variant of ViTraj by removing COIN

5.4 RQ3: Visualization of COIN

Fig. 7 shows comparisons for agent association capability. Fig. 7(a) shows the agent proximity matrix, while Fig. 7(b) to (d) depict the self-similarity matrices of agent vectors for Baseline, FFNet, and ViTraj, respectively. ViTraj exhibits a more consistent self-similarity matrix with the spatial topology of the proximity matrix. This indicates that COIN can accurately identify strongly correlated local regions, avoiding indiscriminate aggregation. Moreover, self-similarity distribution of ViTraj is concentrated in local banded regions, suggesting that features generated by COIN focus on high-value spatial areas. In contrast, FFNet performs global aggregation across agents, resulting in a dispersed similarity matrix with high redundancy among features. This demonstrates that COIN achieves dense local spatial aggregation, effectively filtering out irrelevant or weakly related agent features and preserving only fine-grained critical information. The comparison of spatial patterns, intensity distributions, and redundancy levels validates the core design of COIN, vehicle-centric dynamic selection of infrastructure features, and reveals the underlying reason for its superior performance over traditional methods.

6 Conclusion

This paper proposes ViTraj, a model-agnostic framework to address key challenges in VIC-TP, such as limited perception range and redundant multi-agent features. First, to generate disentangled representations for dual-side trajectories, ViTraj introduces a Feature-end Selection based on a gating mechanism, enabling flexible selection between local vehicle-side features and global infrastructure-side features. Second, the Vehicle Outlook Interaction utilizes vehicle-side agent features as query vectors to efficiently aggregate critical information from neighboring agents, thereby overcoming perception blind spots inherent to vehicle-only viewpoints. Lastly, a Vehicle-Infrastructure Knowledge Distillation enforces trajectory consistency via singular value contrastive distillation and captures essential nonlinear patterns. Experimental results demonstrate that ViTraj achieves state-of-the-art performance on the V2X-Seq benchmark and is compatible with various prediction models.

Acknowledgments

This work is being supported by Zhejiang Provincial Science and Technology Planning Key Project of China under Grant No.2023C03187 and Hangzhou Science and Technology Planning Key Project of China under Grant No.2023SZD0016.

References

[1] Görkay Aydemir, Adil Kaan Akan, and Fatma Güney. 2023. Adapt: Efficient multi-agent trajectory prediction with adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8295–8305.

[2] Zhengwei Bai, Guoyuan Wu, Xuewei Qi, Yongkang Liu, Kentaro Oguchi, and Matthew J Barth. 2022. Infrastructure-based object detection and tracking for cooperative driving automation: A survey. In *2022 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 1366–1373.

[3] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. 2020. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11621–11631.

[4] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, et al. 2019. Argoverse: 3d tracking and forecasting with rich maps. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8748–8757.

[5] Wei Chen, Yuxuan Liang, Yuanshao Zhu, Yanchuan Chang, Kang Luo, Haomin Wen, Lei Li, Yanwei Yu, Qingsong Wen, Chao Chen, et al. 2024. Deep learning for trajectory data management and mining: A survey and beyond. *arXiv preprint arXiv:2403.14151* (2024).

[6] Fangyuan Chi, Yixiao Wang, Panos Nasiopoulos, and Victor CM Leung. 2025. Parameter-efficient federated cooperative learning for 3D object detection in autonomous driving. *IEEE Internet of Things Journal* (2025).

[7] Fangyuan Chi, Yixiao Wang, Panos Nasiopoulos, and Victor CM Leung. 2025. Parameter-Efficient Federated Cooperative Learning for 3D Object Detection in Autonomous Driving. *IEEE Internet of Things Journal* (2025).

[8] Nachiket Deo and Mohan M Trivedi. 2018. Convolutional social pooling for vehicle trajectory prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 1468–1476.

[9] Wenchao Ding, Jing Chen, and Shaojie Shen. 2019. Predicting vehicle behaviors over an extended horizon using behavior interaction network. In *2019 international conference on robotics and automation (ICRA)*. IEEE, 8634–8640.

[10] Fei Gao, Wanjun Huang, Libo Weng, and Yuanming Zhang. 2024. SIF-TF: A Scene-Interaction fusion Transformer for trajectory prediction. *Knowledge-Based Systems* 294 (2024), 111744.

[11] Andreas Geiger, Philip Lenz, and Raquel Urtasun. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 3354–3361.

[12] Roshan George, Joseph Clancy, Tim Brophy, Ganesh Sistu, William O'Grady, Sunil Chandra, Fiachra Collins, Darragh Mullins, Edward Jones, Brian Deegan, et al. 2025. Infrastructure Assisted Autonomous Driving: Research, Challenges, and Opportunities. *IEEE Open Journal of Vehicular Technology* (2025).

[13] Yeping Hu, Wei Zhan, and Masayoshi Tomizuka. 2018. Probabilistic prediction of vehicle semantic intention and motion. In *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 307–313.

[14] Yizhou Huang, Yihua Cheng, and Kezhi Wang. 2025. Trajectory mamba: Efficient attention-mamba forecasting model based on selective ssm. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 12058–12067.

[15] Yanjun Huang, Jiatong Du, Ziru Yang, Zewei Zhou, Lin Zhang, and Hong Chen. 2022. A survey on trajectory-prediction methods for autonomous driving. *IEEE Transactions on Intelligent Vehicles* 7, 3 (2022), 652–674.

[16] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. 2022. Transformers in vision: A survey. *ACM computing surveys (CSUR)* 54, 10s (2022), 1–41.

[17] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[18] Raphael Korbmacher and Antoine Tordeux. 2022. Review of pedestrian trajectory prediction methods: Comparing deep learning and knowledge-based approaches. *IEEE Transactions on Intelligent Transportation Systems* 23, 12 (2022), 24126–24144.

[19] Raphael Korbmacher and Antoine Tordeux. 2022. Review of pedestrian trajectory prediction methods: Comparing deep learning and knowledge-based approaches. *IEEE Transactions on Intelligent Transportation Systems* 23, 12 (2022), 24126–24144.

[20] Wei-Cheng Lai, Zi-Xiang Xia, Hao-Siang Lin, Lien-Feng Hsu, Hong-Han Shuai, I-Hong Jhuo, and Wen-Huang Cheng. 2020. Trajectory prediction in heterogeneous environment via attended ecology embedding. In *Proceedings of the 28th acm international conference on multimedia*. 202–210.

[21] Shen Li, Keqi Shu, Chaoyi Chen, and Dongpu Cao. 2021. Planning and decision-making for connected autonomous vehicles at road intersections: A review. *Chinese Journal of Mechanical Engineering* 34 (2021), 1–18.

[22] Ming Liang, Bin Yang, Rui Hu, Yun Chen, Renjie Liao, Song Feng, and Raquel Urtasun. 2020. Learning lane graph representations for motion forecasting. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer, 541–556.

[23] Mengmeng Liu, Hao Cheng, Lin Chen, Hellward Broszio, Jiangtao Li, Runjiang Zhao, Monika Sester, and Michael Ying Yang. 2024. Laformer: Trajectory prediction for autonomous driving with lane-aware scene constraints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2039–2049.

[24] Wei Liu, Min Hua, Zhiyun Deng, Zonglin Meng, Yanjun Huang, Chuan Hu, Shunhui Song, Letian Gao, Changsheng Liu, Bin Shuai, et al. 2023. A systematic survey of control techniques and applications in connected and automated vehicles. *IEEE Internet of Things Journal* 10, 24 (2023), 21892–21916.

[25] Antonio Longa, Steve Azzolin, Gabriele Santin, Giulia Cencetti, Pietro Liò, Bruno Lepri, and Andrea Passerini. 2025. Explaining the explainers in graph neural networks: a comparative study. *Comput. Surveys* 57, 5 (2025), 1–37.

[26] Yuhuan Lu, Wei Wang, Rufan Bai, Shengwei Zhou, Lalit Garg, Ali Kashif Bashir, Weiwei Jiang, and Xiping Hu. 2025. Hyper-relational interaction modeling in multi-modal trajectory prediction for intelligent connected vehicles in smart cites. *Information Fusion* 114 (2025), 102682.

[27] Jiageng Mao, Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. 2023. 3D object detection for autonomous driving: A comprehensive survey. *International Journal of Computer Vision* 131, 8 (2023), 1909–1963.

[28] Kaouther Messaoud, Itheri Yahiaoui, Anne Verroust-Blondet, and Fawzi Nashashibi. 2020. Attention based vehicle trajectory prediction. *IEEE Transactions on Intelligent Vehicles* 6, 1 (2020), 175–185.

[29] Meriem Mhedhbi, Salah Elayoubi, and Galaad Leconte. 2022. AI-based prediction for Ultra Reliable Low Latency service performance in industrial environments. In *2022 18th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*. IEEE, 130–135.

[30] Xiaoyu Mo, Yang Xing, and Chen Lv. 2021. Graph and recurrent neural network-based vehicle trajectory prediction for highway driving. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*. IEEE, 1934–1939.

[31] Haohao Qu, Liangbo Ning, Rui An, Wenqi Fan, Tyler Derr, Hui Liu, Xin Xu, and Qing Li. 2024. A survey of mamba. *arXiv preprint arXiv:2408.01129* (2024).

[32] Ruijie Quan, Linchao Zhu, Yu Wu, and Yi Yang. 2021. Holistic LSTM for pedestrian trajectory prediction. *IEEE transactions on image processing* 30 (2021), 3229–3239.

[33] Hongzhi Ruan, Haibao Yu, Wenxian Yang, Siqi Fan, and Zaiqing Nie. 2024. Learning cooperative trajectory representations for motion forecasting. *Advances in Neural Information Processing Systems* 37 (2024), 13430–13457.

[34] Liushuai Shi, Le Wang, Chengjiang Long, Sanping Zhou, Mo Zhou, Zhenxing Niu, and Gang Hua. 2021. SGCN: Sparse graph convolution network for pedestrian trajectory prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8994–9003.

[35] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. 2020. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2446–2454.

[36] Chujie Wang, Lin Ma, Rongpeng Li, Tariq S Durrani, and Honggang Zhang. 2019. Exploring trajectory prediction through machine learning methods. *IEEE Access* 7 (2019), 101441–101452.

[37] Jianda Wang, Zhendong Wang, Bo Yu, Jie Tang, Shuaiwen Leon Song, Cong Liu, and Yang Hu. 2023. Data fusion in infrastructure-augmented autonomous driving system: Why? where? and how? *IEEE Internet of Things Journal* 10, 18 (2023), 15857–15871.

[38] Jizhao Wang, Zhizhou Wu, Yunyi Liang, Jinjun Tang, and Huimiao Chen. 2024. Perception methods for adverse weather based on vehicle infrastructure cooperation system: A review. *Sensors* 24, 2 (2024), 374.

[39] Tianqi Wang, Sukmin Kim, Ji Wenxuan, Enze Xie, Chongjian Ge, Junsong Chen, Zhenguo Li, and Ping Luo. 2024. Deepaccident: A motion and accident prediction benchmark for v2x autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 5599–5606.

[40] Tsun-Hsuan Wang, Sivabalan Manivasagam, Ming Liang, Bin Yang, Wenyuan Zeng, and Raquel Urtasun. 2020. V2vnet: Vehicle-to-vehicle communication for joint perception and prediction. In *Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, part II 16*. Springer, 605–621.

[41] Xianing Wang, Linjun Lu, Zhan Zhang, Ying Wang, and Haoming Li. 2025. Introducing the vehicle-infrastructure cooperative control system by quantifying the benefits for the scenario of signalized intersections. *Transportation Research Part A: Policy and Practice* 192 (2025), 104378.

[42] Jing Yang, Yuehai Chen, Shaoyi Du, Badong Chen, and Jose C Principe. 2024. IA-LSTM: interaction-aware LSTM for pedestrian trajectory prediction. *IEEE transactions on cybernetics* (2024).

[43] Mingxia Yang, Boliang Zhang, Tingting Wang, Jijing Cai, Xiang Weng, Hailin Feng, and Kai Fang. 2024. Vehicle interactive dynamic graph neural network-based trajectory prediction for Internet of Vehicles. *IEEE Internet of Things Journal* 11, 22 (2024), 35777–35790.

[44] Haibao Yu, Yizhen Luo, Mao Shu, Yiyi Huo, Zebang Yang, Yifeng Shi, Zhenglong Guo, Hanyu Li, Xing Hu, Jirui Yuan, et al. 2022. Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21361–21370.

[45] Haibao Yu, Yingjuan Tang, Enze Xie, Jilei Mao, Ping Luo, and Zaiqing Nie. 2023. Flow-based feature fusion for vehicle-infrastructure cooperative 3d object detection. *Advances in Neural Information Processing Systems* 36 (2023), 34493–34503.

[46] Haibao Yu, Wenxian Yang, Hongzhi Ruan, Zhenwei Yang, Yingjuan Tang, Xu Gao, Xin Hao, Yifeng Shi, Yifeng Pan, Ning Sun, et al. 2023. V2x-seq: A large-scale sequential dataset for vehicle-infrastructure cooperative perception and forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5486–5495.

[47] Huihuang Zhang, Haigen Hu, Bin Cao, and Xiaoqin Zhang. 2025. AutoStyleMixer: A universal adaptive N-to-One framework for cross-domain data augmentation. *Knowledge-Based Systems* 323 (2025), 113616.

[48] Xingchen Zhang, Yuxiang Feng, Panagiotis Angeloudis, and Yiannis Demiris. 2022. Monocular visual traffic surveillance: A review. *IEEE Transactions on Intelligent Transportation Systems* 23, 9 (2022), 14148–14165.

[49] Xia Zhao, Limin Wang, Yufei Zhang, Xuming Han, Muhammet Deveci, and Milan Parmar. 2024. A review of convolutional neural networks in computer vision. *Artificial Intelligence Review* 57, 4 (2024), 99.

[50] Jianwei Zheng, Wei Li, Ni Xu, Junwei Zhu, and Xiaoqin Zhang. 2024. Alias-free mamba neural operator. *Advances in Neural Information Processing Systems* 37 (2024), 52962–52995.

[51] Chen Zhou, Ghassan AlRegib, Armin Parchami, and Kunjan Singh. 2024. TrajPRed: Trajectory prediction with region-based relation learning. *IEEE Transactions on Intelligent Transportation Systems* (2024).

[52] Zikang Zhou, Jianping Wang, Yung-Hui Li, and Yu-Kai Huang. 2023. Query-centric trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17863–17873.

[53] Zikang Zhou, Luyao Ye, Jianping Wang, Kui Wu, and Kejie Lu. 2022. Hivt: Hierarchical vector transformer for multi-agent motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8823–8833.