STATE OF CALIFORNIA • DEPARTMENT OF TRANSPORTATION

TECHNICAL REPORT DOCUMENTATION PAGE

TR-0003 (REV 04/2024)

1. REPORT NUMBER	2. GOVERNMENT ASSOCIATION NUMBER	3. RECIPIENT'S CATALOG NUMBER
CA-27-4067	N/A	N/A

4. TITLE AND SUBTITLE	5. REPORT DATE
Vulnerable Road User (VRU): Accurate and Reliable Detection for Roadside-Assisted Cooperative Driving	March 31, 2025
	6. PERFORMING ORGANIZATION CODE
	UC Riverside
7. AUTHOR	8. PERFORMING ORGANIZATION REPORT NO.
Chuheng Wei, Ziyan Zhang, Guoyuan Wu	
9. PERFORMING ORGANIZATION NAME AND ADDRESS	10. WORK UNIT NUMBER
College of Engineering - Center for Environmental Research and Technology University of California at Riverside	N/A
1084 Columbia Ave, Riverside, California 92507	11. CONTRACT OR GRANT NUMBER
	65A1057
12. SPONSORING AGENCY AND ADDRESS	13. TYPE OF REPORT AND PERIOD COVERED
California Department of Transportation P.O. Box 942873, MS #83 Sacramento, CA 94273-0001	Final Report July 2023 - December 2024
	14. SPONSORING AGENCY CODE
	Caltrans

15. SUPPLEMENTARY NOTES

N/A

16. ABSTRACT

The safety of vulnerable road users (VRUs) such as pedestrians, cyclists, and micro-mobility participants is a pressing challenge in urban traffic. This report presents a roadside-assisted cooperative driving system that combines multi-modal perception, tracking, trajectory prediction, and communication to enhance VRU protection. The system integrates camera and LiDAR fusion with both traditional and deep learning methods to achieve robust detection in complex environments. A Kalman Filter–based tracking framework ensures reliable trajectory continuity, while an enhanced Social GAN incorporating traffic signal and map context improves prediction accuracy at intersections. A collision prediction and warning module based on Time to Collision provides real-time alerts through V2X communication to vehicles, infrastructure, and VRU devices. A mobile and reconfigurable sensor platform was developed for safe data collection and algorithm validation. Experiments on benchmark datasets and field trials show over 90 percent VRU detection recall under occlusion, improved pedestrian recognition with optimized LiDAR placement, and significant gains in trajectory prediction. The proposed system offers practical insights and scalable strategies for deploying intelligent roadside infrastructure to reduce VRU fatalities and improve traffic safety.

17. KEY WORDS	18. DISTRIBUTION STATEMENT
vulnerable road user (VRU), cooperative perception, roadside-assisted cooperative driving, traffic safety	No restrictions. This document is available to the public through the National Technical Information Service, Springfield, Virginia 22161.

19. SECURITY CLASSIFICATION *(of this report)*	20. NUMBER OF PAGES	21. COST OF REPORT CHARGED
Unclassified	111	N/A

DISCLAIMER STATEMENT

This document is disseminated in the interest of information exchange. The contents of this report reflect the views of the authors who are responsible for the facts and accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of the State of California or the Federal Highway Administration. This publication does not constitute a standard, specification or regulation. This report does not constitute an endorsement by the Department of any product described herein.

For individuals with sensory disabilities, this document is available in alternate formats. For information, call (916) 654-8899, TTY 711, or write to California Department of Transportation, Division of Research, Innovation and System Information, MS-83, P.O. Box 942873, Sacramento, CA 94273-0001.

Vulnerable Road User (VRU): Accurate and Reliable Detection for Roadside-Assisted Cooperative Driving

PROJECT FINAL REPORT

Prepared for:
California Department of Transportation

Prepared by:

College of Engineering - Center for Environmental Research and Technology
University of California at Riverside

March 2025

Table of Contents

Table of Figures

Glossary of Terms

Abbreviation	Full Term
2D MOT	Two-Dimensional Multi-Object Tracking
3D MOT	Three-Dimensional Multi-Object Tracking
8080	80 × 80 mm T-slotted aluminum profile
4040	40 × 40 mm T-slotted aluminum profile
ADE	Average Displacement Error
AMEND	A Mixture of Experts Framework for Long-Tailed Trajectory Prediction
ANN	Artificial Neural Network
AR	Augmented Reality
BEV	Bird's-Eye View
BEVDepth	Bird's-Eye-View Depth-based Representation
BEVHeight	Bird's-Eye-View Height-based Representation
C-V2X	Cellular Vehicle-to-Everything
CAM messages	Cooperative Awareness Messages
CBR	Calibration-Free BEV Representation
CPWS	Collision Prediction and Warning System
CPU	Central Processing Unit
CV	Constant Velocity
CVPR	Conference on Computer Vision and Pattern Recognition
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DDS	Data Distribution Service
DETR	Detection Transformer
DSRC	Dedicated Short-Range Communications
ETH/UCY	Benchmark pedestrian trajectory prediction datasets from ETH Zurich and University of Cyprus
FP	False Positive
FN	False Negative
GAN	Generative Adversarial Network
GRU	Gated Recurrent Unit
GPU	Graphics Processing Unit
HMI	Human-Machine Interface
HOG	Histogram of Oriented Gradients
HOTA	Higher Order Tracking Accuracy
ICP	Iterative Closest Point
IDF1	Identification F1 Score (tracking metric)
IoU	Intersection over Union
ISC	Intersection Safety Challenge

Abbreviation	Full Term
KITTI	Karlsruhe Institute of Technology and Toyota Technological Institute Benchmark Suite
KF-CV	Kalman Filter with Constant Velocity motion model
LSTM	Long Short-Term Memory
LiDAR	Light Detection and Ranging
LOS	Line of Sight
mAP	Mean Average Precision
MEC	Multi-Access Edge Computing
MOTA	Multiple Object Tracking Accuracy
MOTP	Multiple Object Tracking Precision
MVX-Net	Multi-View Fusion VoxelNet
NIC	Network Interface Card
NR	New Radio (5G standard)
NTU	Nanyang Technological University
P2I / I2P	Pedestrian-to-Infrastructure / Infrastructure-to-Pedestrian
PC	Personal Computer
PC5	Sidelink interface for C-V2X communication
PET	Post-Encroachment Time
PoE	Power over Ethernet
PTP	Precision Time Protocol
QdTrack	Quasi-Dense Similarity Learning for Multiple Object Tracking
RF	Random Forest
RFS	Random Finite Set
RGB	Red, Green, Blue (color model)
ROS	Robot Operating System
RSU	Roadside Unit
R-CNN	Region-based Convolutional Neural Network
SOTA	State of the Art
SVM	Support Vector Machine
SSD	Single Shot MultiBox Detector
SSM	Surrogate Safety Measure
STAR	Spatio-Temporal Graph Transformer Networks for Pedestrian Trajectory Prediction
STGCNN	Spatio-Temporal Graph Convolutional Neural Network
T-slot	T-slotted aluminum extrusion
TTC	Time to Collision
TP	True Positive
TN	True Negative
USDOT	United States Department of Transportation
V2I / I2V	Vehicle-to-Infrastructure / Infrastructure-to-Vehicle

Abbreviation	Full Term
V2X	Vehicle-to-Everything
VFH	Viewpoint Feature Histogram
VRU	Vulnerable Road User

Executive Summary

The increasing fatalities among vulnerable road users (VRUs)—including pedestrians, cyclists, and micro-mobility users—represents a critical safety challenge, with 2021 statistics showing 13% and 5% increases in pedestrian and cyclist deaths respectively. This report presents a comprehensive roadside-assisted cooperative driving system designed to enhance VRU safety through advanced perception, tracking, prediction, and communication technologies.

Our methodology integrates a multi-modal detection pipeline that combines traditional computer vision techniques with deep learning approaches, utilizing both LiDAR and camera sensors for robust VRU detection under diverse environmental conditions. The system employs a Kalman Filter-based multi-object tracking (MOT) framework enhanced with noise handling and trajectory smoothing capabilities. For trajectory prediction, we developed an enhanced Social GAN architecture that incorporates traffic light encoding and map context to improve prediction accuracy by 940% in intersection scenarios. The collision prediction and warning system utilizes Time-to-Collision (TTC) analysis to generate real-time alerts distributed through Vehicle-to-Everything (V2X) communication networks.

To address deployment challenges, we designed and constructed a mobile, reconfigurable sensor platform that enables safe data collection and algorithm validation without requiring complex roadside installations. The platform features modular T-slotted aluminum construction with specialized mounting fixtures for precise sensor positioning and calibration. Experimental validation on USDOT Intersection Safety Challenge and DAIR-V2X-I datasets demonstrates the system's effectiveness. Key findings include: (1) multi-modal sensor fusion achieves over 90% VRU detection recall under heavy occlusion conditions; (2) low-mounted LiDAR configurations (1.5m height) provide 15% better pedestrian recall compared to traditional overhead mounting; and (3) traffic signal encoding significantly improves trajectory prediction accuracy at signalized intersections.

The integrated system successfully bridges sensor hardware, perception algorithms, and communication protocols to deliver an end-to-end solution for proactive VRU protection. This work provides California Department of Transportation with actionable insights for deploying scalable, effective safety solutions that can significantly reduce VRU fatalities through intelligent roadside infrastructure.

Code is available at https://github.com/TSR-CECERT/VRU

1 Introduction

The rising toll of vulnerable road user (VRU) fatalities—pedestrians, cyclists, scooter and skateboard riders—commands urgent attention. In the United States, 2021 NHTSA statistics show pedestrian deaths up 13 % and bicyclist fatalities up 5 % over the prior year [1]. California alone recorded over 2,200 pedestrian and cyclist deaths in 2020–2021, a trend driven by distracted driving, complex intersection geometries, and growing micro-mobility usage [2]. These figures underscore that protecting VRUs is not an optional enhancement but a critical safety mandate.

VRUs are uniquely at risk: their small size, unpredictable trajectories, and lack of protective enclosures make them susceptible to severe injury even at low speeds. The Federal Highway Administration(FHWA)'s definition of VRUs includes a broad array of non-motorized users—pedestrians, pedal cyclists, e-cyclists, scooterists, skateboarders, and roadside workers—each presenting distinct detection and prediction challenges [1]. Traditional traffic surveillance, focused on vehicle counts and speeds, has struggled to reliably track these agile, often occluded participants.

Recent breakthroughs in neural network–based 2D and 3D object detection have transformed what is possible for VRU perception. Convolutional and transformer-based image detectors now routinely achieve sub-10 millisecond (ms) inference for pedestrian bounding boxes, while LiDAR-driven architectures like PointPillars [3] and VoxelNet [4] deliver centimeter-scale localization in real time. By fusing modalities—projecting raw point clouds into image feature maps or integrating visual embeddings into 3D voxel backbones—early-fusion networks dramatically improve recall on small, low-contrast objects in cluttered scenes.

Trajectory prediction has likewise benefited from deep learning's rise. Social LSTMs (Long Short-Term Memory) [5], graph-based spatio-temporal models, and attention-driven transformers can anticipate VRU paths several seconds ahead by modeling inter-agent interactions and environmental context. Such forecasts enable proactive warnings—critical at intersections where human reaction time often lags unfolding hazards. Integrating dynamic map cues like crosswalk geometry and traffic-signal phases further refines predictions, reducing false alarms and enhancing trust in automated alerts [6].

Beyond perception and prediction, reliable VRU protection demands low-latency communications. Cellular Vehicle-to-Everything (C-V2X) and 5 G NR deliver sub-50 ms round-trip times for event-driven messages, allowing roadside units to broadcast imminent collision warnings to vehicles and wearable devices [7]. When combined with real-time detection and forecasting, these networks transform passive surveillance into active collision-avoidance systems.

1.1 Research Goals and Contributions

The primary goal of this research is to develop a comprehensive, roadside-assisted cooperative driving system that significantly enhances VRU safety through intelligent perception, prediction, and communication technologies. Specifically, this work aims to:

1. **Develop a robust multi-modal detection system** that combines traditional computer vision techniques with deep learning approaches to achieve reliable VRU detection under diverse environmental conditions
2. **Create an integrated tracking and prediction framework** that maintains continuous VRU trajectories and forecasts future paths with high accuracy
3. **Design a real-time collision warning system** that leverages Vehicle-to-Everything (V2X) communication to deliver timely safety alerts
4. **Establish a flexible testing platform** that enables safe algorithm validation and data collection without complex roadside installations

The key contributions of this research include:

- **A novel multi-stage fusion detection pipeline** that integrates LiDAR and camera data through both traditional and deep learning methods, achieving over 90% VRU detection recall under challenging conditions
- **An enhanced Social Generative Adversarial Network (GAN) architecture** for trajectory prediction that incorporates traffic signal information and environmental context, improving prediction accuracy by 940% at signalized intersections
- **A mobile, reconfigurable sensor platform** with specialized mounting fixtures that enables precise calibration and flexible deployment for VRU-focused roadside sensing
- **Comprehensive experimental validation** demonstrating optimal LiDAR placement strategies and the effectiveness of multi-modal sensor fusion for VRU protection
- **Practical deployment insights** that inform California Department of Transportation on scalable safety solutions for protecting vulnerable road users

1.2 Report Organization

This report is structured to provide a comprehensive overview of our roadside-assisted cooperative driving system for VRU safety, progressing from theoretical foundations through practical implementation to experimental validation and conclusions.

This report proceeds as follows:

- Section 2: Literature Review, synthesizing single- and multi-node detection, MOT, trajectory prediction, and V2X communication for VRUs.
- Section 3: Methodology, detailing our hybrid detection pipeline, tracking algorithms, Social GAN–based prediction, and collision-warning design.
- Section 4: Mobile Platform, describing the modular, reconfigurable rig built for safe, flexible data collection and calibration.

- Section 5: Algorithm Demonstrations, showcasing preliminary fusion detection and camera-only prediction results on public and in-house datasets.
- Section 6: Findings and Funding, summarizing key conclusions from our research, ongoing refinements, and the support that has enabled this work.

Together, these contributions advance a holistic solution for VRU safety—melding state-of-the-art neural methods, robust hardware, and V2X communications to protect our most vulnerable road users.

2 Literature Review

This section synthesizes key findings from the full literature report on roadside-assisted cooperative detection, tracking, prediction, and communication for Vulnerable Road Users (VRUs). It covers single-node and multi-node detection approaches, multi-object tracking paradigms, VRU intention and trajectory prediction models, and V2X communication technologies. For a thorough review of pertinent literature, please refer to **Appendix 1.** Notably, a dedicated survey paper titled "Vulnerable Road User Detection for Roadside-Assisted Safety Protection: A Comprehensive Survey" has been published based on this part of the work [8].

In the context of roadside sensing systems, we distinguish between two fundamental deployment architectures: **single-node detection** refers to sensing systems that utilize individual, standalone sensor units (such as a single camera or LiDAR) operating independently at specific locations, while **multi-node detection** involves multiple sensor units working cooperatively across distributed locations to provide comprehensive coverage and enhanced detection capabilities through data fusion and coordination. Single-node systems typically process data locally and make independent decisions, whereas multi-node systems leverage information sharing and collaborative processing to overcome individual sensor limitations, reduce blind spots, and improve overall system robustness and accuracy in complex traffic environments.

2.1 Detection

2.1.1 Single-Node Detection

2.1.1.1 LiDAR-Based Methods

Traditional Approaches:
Early works apply background filtering (e.g., 3D-DSF), clustering (DBSCAN), and hand-crafted statistical features for classification. Wu et al. (2021) [9] achieved separate classification of vehicles, pedestrians, cyclists, and skateboarders by combining 3D density filtering with a two-step Random Forest (RF) classifier using average speed and speed variance features. Song et al. (2020) [10] further refined DBSCAN parameters in three distance-based zones, then used a Back-Propagation Artificial Neural Network (ANN) for vehicle vs. pedestrian discrimination at varying ranges. Comparative studies show Support Vector Machine (SVM) often offers the best trade-off between nonlinearity handling and computation time on 16-channel roadside LiDAR data. Occlusion and sparse returns at distance remain persistent challenges, motivating adaptive clustering radii and dynamic thresholding based on distance to the sensor.

Deep Learning Approaches:
Two-Stage Enhancement: Xu et al. (2022) [11] replaced the final classifier with a lightweight VGGNet augmented by self-attention, integrating geometric, statistical, and Viewpoint Feature Histogram (VFH) descriptors to yield a 10–15% mAP boost over RF/SVM baselines on pedestrian and cyclist classes.

End-to-End Networks: FecNet (2023) [12] employs a cascade of foreground-attention branches and multi-scale feature fusion, achieving 86.3% mAP on proprietary roadside datasets, while remaining compact compared to PointPillars [3], SECOND [13], and PV-RCNN [14] variants.

Center-based architectures (CenterPoint)[15] adapted for roadside settings—with voxel or pillar encodings—demonstrate superior pedestrian detection accuracy with reduced inference time. The latest hybrid Transformer-backbones (Shi et al., 2023) [16] leverage DAIR-V2X-I data [17] converted to KITTI [18] format, outperforming all tested SOTA models on cyclist detection under real surveillance conditions.

2.1.1.2 Camera-Based Methods

Classical Algorithms: The Viola–Jones face detector (2004) [19] and Dalal–Triggs HOG descriptor (2005) [20] laid the groundwork for real-time pedestrian detection, achieving sub-1% miss rates at 10^{-6} False Positives Per Window on controlled datasets.

Deep CNN and One-Stage Detectors: Region-based networks (R-CNN, Fast R-CNN [21], Faster R-CNN [22]), SSD [23], and YOLO [24-27] families provide progressively faster, more accurate detection. Garcia-Venegas et al. (2021) [28] compared Faster R-CNN, SSD, and R-FCN [29] for cyclist detection, selecting SSD for real-time ITS with Kalman filtering to mitigate false negatives. Fine-tuned Faster R-CNN models detect pedestrians with >90% precision under good illumination but degrade significantly under shadows or glare.

Bird's-Eye-View (BEV) Representations: Recent roadside camera systems adopt BEVDepth and BEVHeight to project monocular images into a top-down plane, improving spatial consistency. Calibration-free BEV Representation (CBR) reduces extrinsic dependencies but trades off ~8.7% AP3D. CoBEV's two-stage Complementary Feature Selection and BEV distillation set new SOTA on DAIR-V2X-I, especially in long-range and low-visibility scenarios [30].

2.1.2 Multi-Node Sensor Fusion

2.1.2.1 By Sensor Mode

Single-Mode
LiDAR-only networks (InfraDet3D) fuse multiple 2D scans via Fast Global Registration and ICP [31]. Camera-only setups combine RGB with event or infrared cameras to address motion blur and nighttime conditions.

Multi-Mode
Camera + RADAR: Bai et al. (2021) [32] calibrated and aligned YOLO-v4 object angles with RADAR range/velocity via nearest-neighbor fusion; Liu et al. [33] employed the Dempster–Shafer theory for evidence fusion across scans, enhancing detection under poor lighting.

Camera + LiDAR: The PointFusion module in the MVX-Net [34] model projects 3D points onto the image plane to concatenate 2D semantic features with 3D coordinates before VoxelNet [4] processing; VoxelFusion pools image ROIs into voxel grids, boosting far-field detection with sparse LiDAR returns.

2.1.2.2 By Fusion Stage

Early Fusion: Raw point clouds and images are registered into a common frame pre-processing. This type of fusion yields high accuracy but demands precise calibration, synchronization, and high bandwidth (e.g., 20 MB/s for a 64-beam LiDAR at 10 Hz). The results could be significantly impacted by the noises.

Intermediate Fusion: Features extracted by CNNs or Voxel encoders are merged (e.g., MVX-Net [34], PillarGrid [35]), reducing data volume transmitted while retaining rich context. This type of fusion often strikes the best balance between accuracy and resource consumption, but the standardization of selected features and/or representations would be a potential challenge.

Late Fusion: Independently generated detections are combined using spatio-temporal matching, Extended Kalman Filters, and Non-Maximum Suppression [36, 37]. This type of fusion is robust to calibration drift and asynchrony, with lower communication requirements, but potentially lower recall on small or occluded VRUs.

2.2 Tracking

Multi-Object Tracking (MOT) pipelines consist of per-frame detection, motion/ appearance prediction, similarity assessment, and data association.

2.2.1 2D vs. 3D MOT

2D MOT (DeepSORT [38], ByteTrack [39], QDTrack [40]) relies on image features and bounding-box overlap; sensitive to lighting changes, clutter, and occlusion.

3D MOT (AB3DMOT [41], SimpleTrack [42], PolyMOT [43]) uses point-cloud centroids and 3D IoU for more robust spatio-temporal association. Linear Kalman filters predict object states; IoU or nearest-neighbor matching links trajectories. PolyMOT [43] enhances accuracy by selecting motion models per object class (e.g., constant velocity for pedestrians vs. constant acceleration for cyclists).

2.2.2 Roadside-Specific Techniques

Clustering-based association
Nearest cluster to predicted state is linked frame-to-frame (Zhao et al., 2019) [44]. Random Finite Set (RFS) [45] filters with Sequential Monte Carlo (Meissner et al., 2012) [46]aggregate multi-laser scanner detections.

Cross-modal re-identification
DeepSORT [47] applied in 2D, then mapped back to corresponding 3D detections for joint tracking .

2.3 Prediction

2.3.1 Intention Prediction

RU intention prediction typically focuses on crossing/non-crossing decisions at the next time step, using inputs like trajectory history, position, velocity, and distance to other road users. Common approaches include Bayesian models with particle filters for motion prediction, PolyMLP for joint trajectory and intention estimation using motion states [48], DA-ANN [49] for recognizing crossing behavior based on encoded spatial and motion features, and modified Naïve Bayes classifiers with enriched contextual inputs. GRU-based models also use multi-feature sequences (e.g., age, gender, distance to conflict point) for accurate crossing/waiting prediction.

2.3.2 Trajectory Prediction

2.3.2.1 Dynamics Model-based

Constant Velocity/Acceleration and Social Force models offer interpretable baselines but struggle with interactive behaviors.

2.3.2.2 Data-driven

RNN [50]/LSTM [51]/GRU architectures (e.g., Social LSTM [5], Social GAN [52]) capture temporal dependencies and crowd interactions. Graph-based models (such as SGCN [53], Social-STGCNN [54]) represent social interactions as spatio-temporal graphs. Transformer-based frameworks (e.g., STAR [55]) and Mixture-of-Experts (AMEND) [56] excel on long-tail and complex motion patterns, reducing Average Displacement Error (ADE) by 10–15% on ETH/UCY benchmarks.

Implementation Considerations: The data-driven models commonly use 3.2 s history to predict 4.8 s into the future, balancing human reaction time requirements with system latency targets. Robustness to sensor drop-outs and unseen intersections can be improved via domain randomization (SimAug) [57] or adversarial perturbations.

2.4 Communication

2.4.1 Technologies:

Although DSRC (IEEE 802.11p) has Low-latency (< 50 ms) V2X communication using 5.9 GHz band and is proven in pilot intersections in the past few decades, the spectrum has been repurposed. Cellular Vehicle-to-Everything (C-V2X) leverages 4G/5G NR for device-to-device (PC5) and device-to-network (Uu) modes, and would be flexible enough to offer extended range, non-LOS resilience, and scalable capacities.

2.4.2 VRU-Focused Applications

Pedestrian-to-Infrastructure (P2I/I2P)
Smartphones or wearables broadcast VRU presence (CAM messages) to RSUs, which aggregate and forward collision warnings.

Vehicle-to-Infrastructure (V2I/I2V)
Vehicles relay their own states, and infrastructure (as an edge node) computes imminent collision risk and sends direct alerts back to VRU devices (audio, haptic) and in-vehicle HMI.

Performance Targets
As safety-critical applications, it is expected and required end-to-end latency ≤ 100 ms, data rate < 10 kb/s per session to support event-driven warning messages.

2.5 Example Testbeds and Field Trials

Caltrans and UC Berkeley PATH C-V2X testbed on El Camino Real integrates roadside LiDAR, cameras, and C-V2X radios for intersection safety trials [58]. ACTIVE-AURORA in Canada and NTU Singapore's 5G MEC platform demonstrate C-V2X-based pedestrian collision avoidance with sub-50 ms round-trip times [59].

3 Methodology

To avoid redundancy, this section summarizes the key information about the proposed methodology based on the **Appendix 2**.

3.1 Methodology Overview

Our developed end-to-end system (Figure 1) is organized into three primary modules:

- *Object Perception*: real-time data acquisition → preprocessing → hybrid detection, including a variety of functionalities for each individual sensor, such as detection, classification, and tracking.
- *Information Processing*: multi-sensor fusion (including multi-object tracking) → multi-object trajectory/motion prediction.
- Situation Awareness: collision prediction → involved road users association → multi-modal warning dissemination.

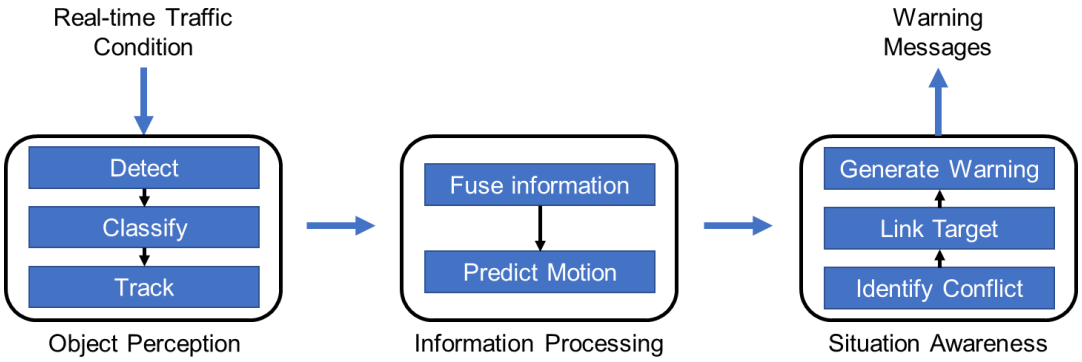

Figure 1. Proposed Flowchart of the System

The following subsections will describe each individual module and function succinctly.

3.2 Detection

The detection module is essential for accurately identifying vulnerable road users (VRUs) and vehicles under diverse and challenging conditions. As shown in Figure 2, it combines traditional and deep learning-based methods in a multi-stage pipeline. The process includes data preprocessing, modality-level fusion of LiDAR and camera outputs, global fusion across nodes, and advanced 3D detection using deep learning. Finally, a method fusion step integrates both approaches, balancing the robustness of traditional techniques with the precision of neural networks to enhance detection performance in complex roadside environments.

Our system employs a dual-sensor configuration consisting of two cameras and two LiDAR units to ensure comprehensive coverage of intersection scenarios. For safety considerations, LiDAR sensors are typically mounted at elevated positions with downward-tilted orientations at intersections, but a single LiDAR unit cannot adequately cover the entire intersection area due to

occlusion and geometric constraints. Therefore, LiDAR sensors are commonly deployed in diagonal arrangements across the intersection to eliminate blind spots. Similarly, a single camera typically monitors only one direction of the intersection approach, making multi-camera configurations standard practice for comprehensive VRU detection. We selected a two-camera and two-LiDAR configuration as an optimal balance between coverage completeness and system complexity.

The integration of both 2D and 3D detection methods is crucial for robust VRU identification in challenging scenarios where individual modalities exhibit limitations. Sparse LiDAR point clouds, while providing accurate spatial positioning, often lack sufficient detail for fine-grained object classification, such as distinguishing between a person pushing a wheelchair versus a person pushing a stroller. Camera images provide rich semantic features including texture and color information that enhance classification accuracy for VRU subcategories. Furthermore, the joint 2D-3D approach significantly improves performance in scenarios with limited labeled training data, enabling more robust and detailed VRU classification while maintaining the spatial accuracy advantages of LiDAR-based detection.

A detailed study of this methodology has been published in our paper *"Hierarchical Multi-Modal Fusion for Roadside VRU Detection: Method Complementarity Under Sparse Label Constraints"[60]*.

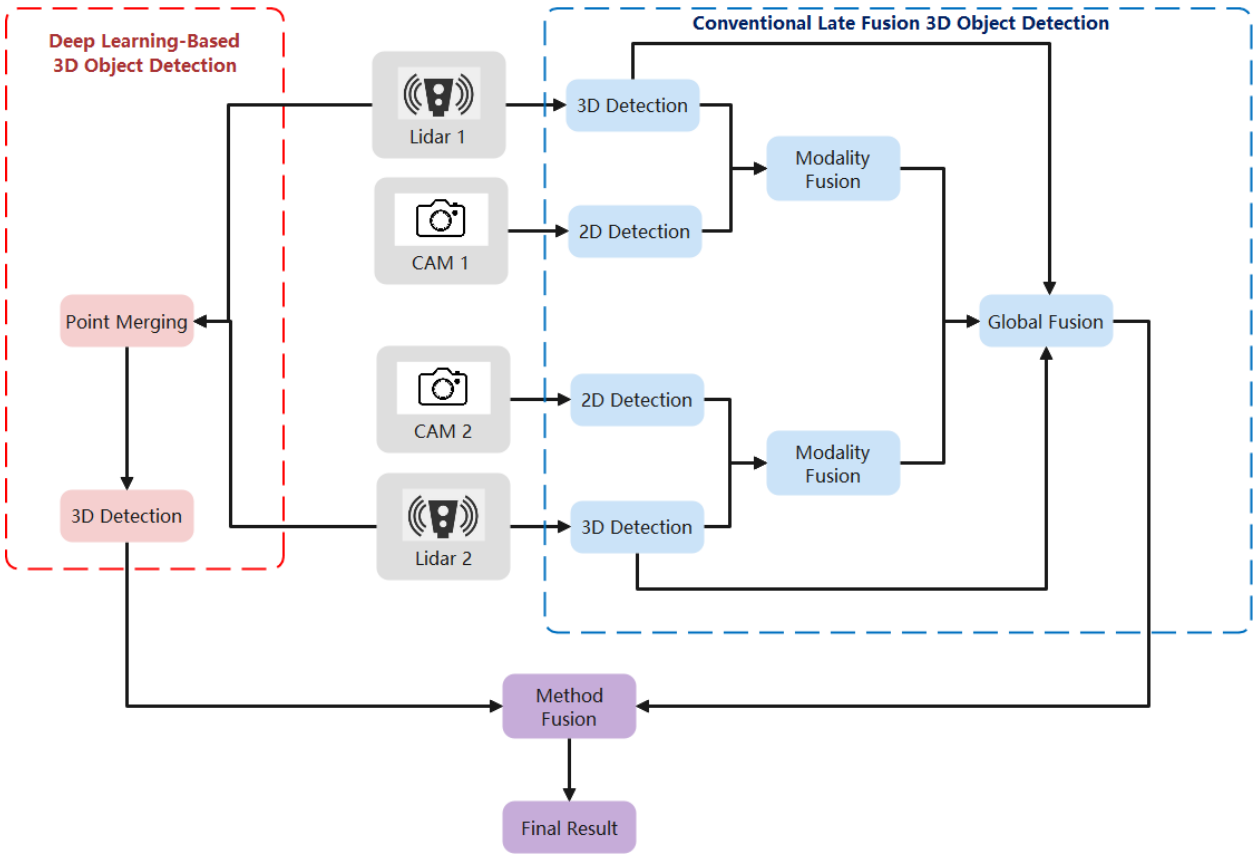

Figure 2. Multi-Modal and Method Fusion Framework for VRU Detection

3.2.1 Data Preprocessing

The data preprocessing stage prepares high-quality inputs for detection by leveraging visual and spatial data from cameras and LiDAR (Light Detection and Ranging). Camera images are calibrated to correct distortion and aligned with global coordinates, with brightness normalization applied to handle low-light conditions. LiDAR data undergoes noise filtering, voxelization, and static point removal to enhance efficiency and focus on dynamic objects. Finally, geo-fencing ensures the system concentrates on regions with intensive VRU activities, laying a solid foundation for accurate and efficient detection.
The system processes live video feeds from field-deployed cameras and real-time LiDAR data streams. Under typical operational conditions with 720P camera resolution and 32-beam LiDAR configuration, the preprocessing pipeline achieves near real-time performance, enabling timely detection and response for critical VRU safety applications.

3.2.2 Traditional Detection Pipeline

The traditional detection pipeline combines 2D detection, 3D detection, and multi-modal sensor fusion to leverage the complementary strengths of cameras and LiDAR sensors.

3.2.2.1 2D Detection with DAB-DETR

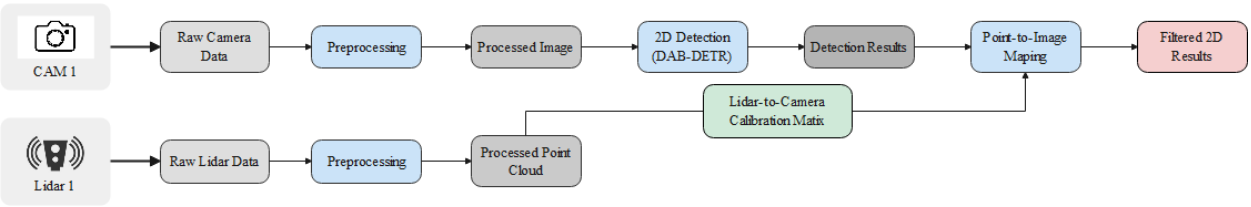

Figure 3. 2D Detection Pipeline

DAB-DETR (Dynamic Anchor Boxes-DETR) [61] is employed for 2D object detection (see Figure 3). Unlike traditional convolutional approaches, DAB-DETR introduces dynamic anchor boxes to improve the model's adaptability to varying object sizes and shapes. These anchor boxes, dynamically updated during training, allow the model to refine bounding box predictions iteratively, enhancing detection accuracy for small objects like pedestrians and cyclists. In addition, the DAB-DETR pipeline processes pre-corrected camera images to detect objects and output 2D bounding boxes. To address challenges such as overlapping detections and intermittent false positives, a heuristic multi-faceted filtering strategy is applied. This strategy consolidates bounding boxes based on confidence scores and spatial consistency, ensuring robust detection in complex roadside environments.

3.2.2.2 Traditional 3D Object Detection

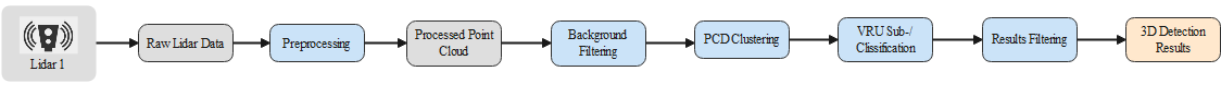

Figure 4. Traditional 3D Detection Pipeline

The 3D detection process (as shown in Figure 4) leverages LiDAR data to classify and localize VRUs and vehicles. The pipeline consists of four main stages:

1. **Background Removal**: Static objects are identified using cross-frame distance calculations and removed to isolate dynamic objects.
2. **Point Cloud Clustering**: The Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm[62] is used to group dynamic point clouds into clusters representing individual objects.
3. **VRU Sub-Classification**: Each cluster is classified into VRU subclasses (such as pedestrians, bicyclists, scooterists) using a Random Forest classifier trained on ground truth data. This classifier offers robustness to the noises and sparsity of LiDAR data, ensuring accurate differentiation of object types.
4. **Result Filtering**: To ensure the reliability of classification results, a consistency check is applied before passing the outputs to the fusion module.

3.2.2.3 Multi-Modal Sensor Fusion

Figure 5 illustrates the workflow of multi-modal sensor fusion, which include the following components.

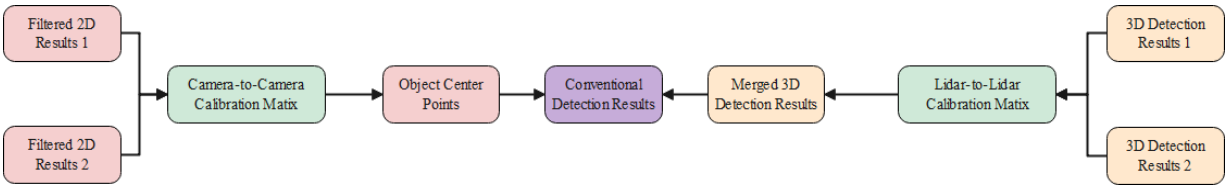

Figure 5. Multi-Node Modality Fusion Workflow

Camera-to-Camera Alignment: this module uses known camera extrinsic to reproject 2D detections into world frame, and merges duplicates within 0.4 m.

LiDAR-to-LiDAR Alignment: this module transforms point clouds via extrinsic calibration matrices, and fuses clusters by nearest-centroid merging with a threshold of 0.6 m.

2D–3D Association: this module projects LiDAR cluster centroids into each camera view, and associates these clusters to 2D boxes if reprojected centroids lie within boxes (IoU > 0.3).

Spatio-Temporal Smoothing: For each fused detection, this module maintains a 5-frame sliding-window average of position and class label probability, mitigating intermittent false positives.

3.2.3 Deep Learning–Based Detection Pipeline

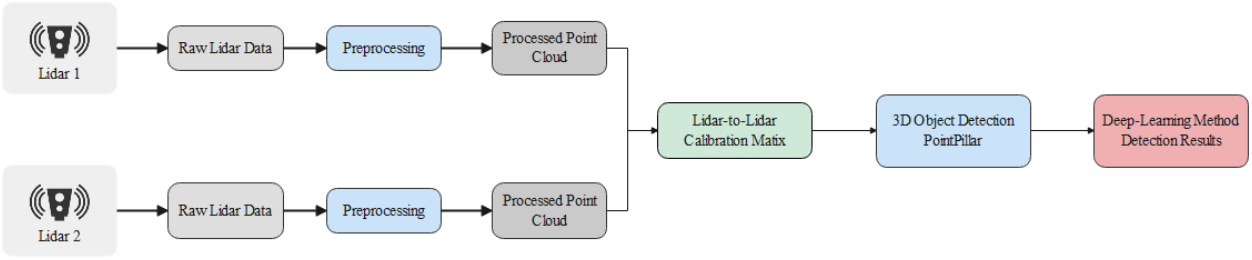

Figure 6. Detection Pipeline for Deep Learning-Based Detection Method

Due to the lack of VRU related data, the deep learning-based detection pipeline (see Figure 6) developed in this project relies on the pretrained models or finetuning with limited amount of real-world data collected by the research team. However, it turns out the integration of deep-learning-based methods may largely enhance the overall detection performance.

3.2.3.1 Point Cloud Fusion

LiDAR point clouds from multiple sensors are merged using LiDAR-to-LiDAR calibration matrices. This fusion increases the density and coverage of the data, enabling effective detection of small and irregularly shaped objects. In addition, range reduction mechanism is applied to filter the data outside the region of interest, ensuring computational efficiency while retaining relevant information.

3.2.3.2 PointPillars Architecture

PointPillars [3] is a state-of-the-art deep learning model designed for real-time 3D object detection, and is widely tested for roadside sensing scenarios. The architecture consists of three primary components:

1. **Voxelization**: The model converts raw point clouds into a sparse 3D tensor using a voxel-based feature encoding scheme [4]. This step organizes spatial data into manageable grid structures.
2. **Backbone Network**: A convolutional neural network processes the voxelized data to extract high-level features, capturing spatial and contextual information critical for object detection.
3. **Detection Head**: The final layer outputs 3D bounding boxes, including the object's position, dimensions, and classification.

PointPillars [3] is fine-tuned on a custom dataset to adapt to roadside VRU detection scenarios. This retraining process optimizes the model for sparse and noisy data, enhancing its detection performance for complex roadside environments.

3.2.4 Method Fusion

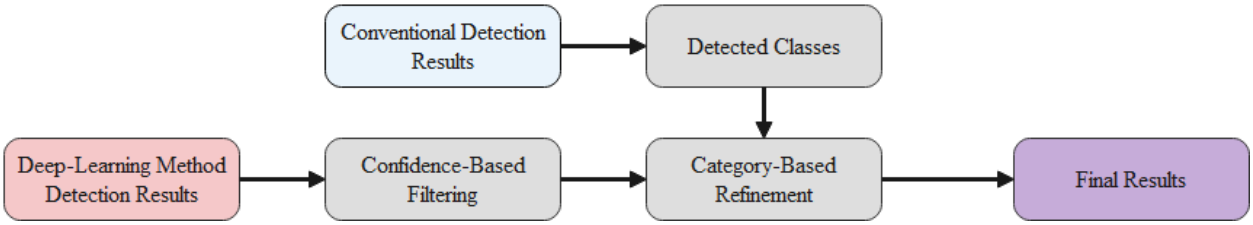

Figure 7. Method Fusion Workflow for Final Detection Results

Both traditional and deep learning-based detection methods have inherent limitations. Traditional methods, while robust to sparse data and noise, suffer from calibration errors during multi-step projections. Conversely, deep learning models such as PointPillars [3] require extensive labeled data, which is often scarce in real-world scenarios. To address these challenges, we implement a method fusion strategy, as illustrated in Figure 7.

The fusion process begins with confidence-based filtering, where low-confidence detections from deep learning models are discarded. This is followed by category-based refinement, which ensures that final classifications align with those obtained from traditional methods. The integrated results are rechecked for spatial and categorical consistency to ensure robustness.

By combining the strengths of both methods, the fusion strategy enhances the overall accuracy and reliability of the detection module, ensuring robust performance across diverse and challenging conditions.

3.3 Multi-Object Tracking (MOT)

The Multi-Object Tracking (MOT) module is a crucial component of our system, bridging the gap between detection and path prediction. By maintaining continuous trajectories of detected objects, the MOT module provides accurate and reliable input for downstream processes. This module leverages the Kalman Filter-Constant Velocity (KF-CV) method [63] for tracking, which ensures robust performance in dynamic and complex environments.

3.3.1 Algorithm Overview

The MOT module employs a KF-CV-based framework, as illustrated in Figure 8, to track multiple objects simultaneously. The algorithm integrates constant velocity motion modeling, data association, measurement update, track lifecycle management, and result filtering.

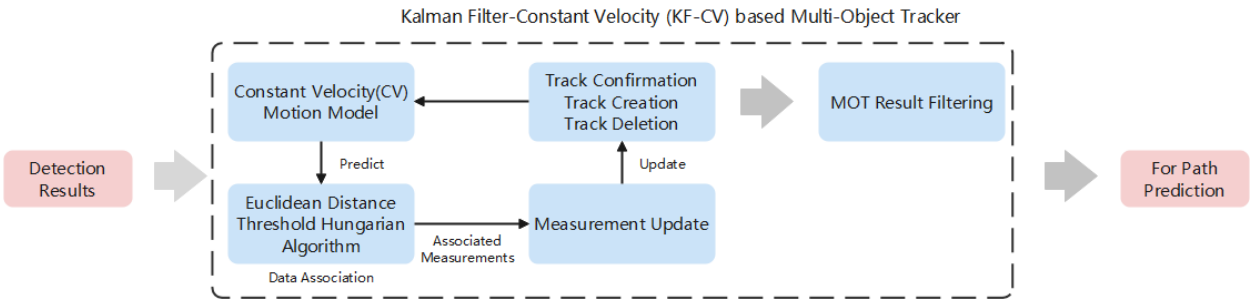

Figure 8. KF-CV-Based Framework for Multi-Object Tracking (MOT)

1. **CV Motion Model:** this module assumes constant velocity to predict object positions in 3D space using the state vector [x, y, z, v_x, v_y, v_z] and a time-step-based transition matrix.
2. **Data Association and Management:** this module uses the Hungarian algorithm with gating to match detections to predicted tracks and manages track creation, confirmation, and deletion based on temporal consistency.
3. **Measurement Update:** this module applies the Kalman filter to refine object states using new observations, with the Kalman gain balancing prediction and measurement noises.
4. **Track Lifecycle Management:** in this module, tracks are created for unmatched detections, confirmed after consistent hits, and deleted if not updated within a set duration.
5. **Result Filtering:** this module applies the post-update filtering mechanism, which removes fragmented tracks, ensuring only complete and reliable trajectories are forwarded to Path Prediction.

3.3.2 Remarks on Noise Handling and Optimization

Real-world scenarios often involve noisy detections due to occlusions, sensor inaccuracies, or overlapping objects. The MOT module incorporates several techniques to handle noise and improve tracking accuracy:

1. **Consistency Checks:** The system verifies track plausibility by checking position, velocity, and acceleration across frames. Abrupt, unrealistic changes trigger re-evaluation using nearby frames to correct errors and maintain stable trajectories.
2. **Fragment Integration:** To address occlusion-induced breaks, fragmented tracks are merged based on spatial and temporal proximity, considering velocity and direction to ensure accurate reconstruction of continuous trajectories.
3. **Adaptive Gating:** The gating threshold for data association adjusts dynamically based on object speed and detection uncertainty, improving matching accuracy for both fast-moving and static objects while reducing false positives.
4. **Trajectory Smoothing:** Post-processing smooths jittery VRU paths using filters such as Kalman or low-pass filters, reducing noise-induced oscillations and producing more natural, stable trajectories.

3.4 Path Prediction

3.4.1 Enhanced Social GAN Architecture

To accurately predict future trajectories in dynamic environments, especially with multiple interacting agents, Generative Adversarial Networks (GANs) [64] offer a powerful framework. As shown in Figure 9, the Social GAN model enhances traditional GANs by explicitly modeling agent interactions and enabling diverse prediction outputs. Social GAN introduces a pooling module to capture inter-agent interactions. The *generator* uses an LSTM encoder for motion history, a pooling module for interaction context, and a decoder for trajectory generation. The *discriminator* evaluates whether trajectories follow social norms. A *variety loss* promotes diverse and plausible trajectory predictions under uncertainty.

In this project, the Social GAN model is further enhanced by encoding traffic light information and introducing other adjustments as described subsequently.

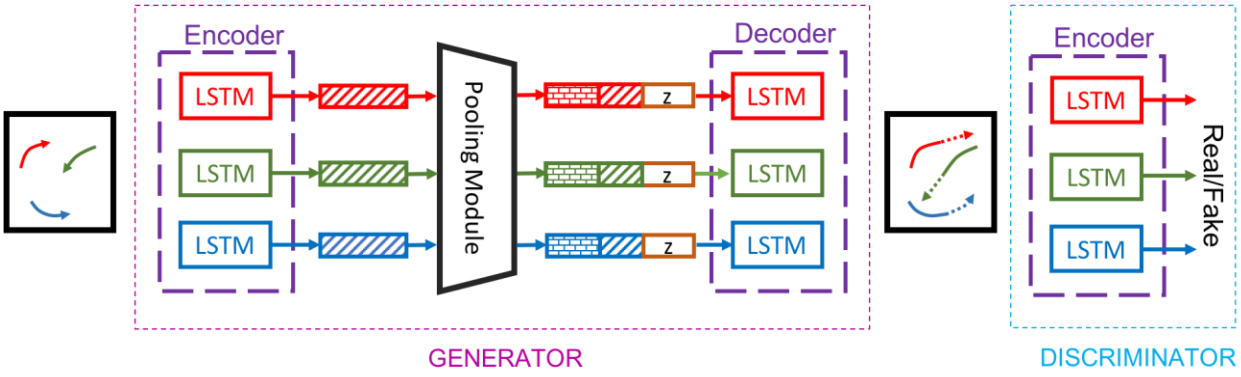

Figure 9. Social GAN [52] system overview.

3.4.2 Specialized Modules

3.4.2.1 Traffic Light Encoding

Traffic light states are embedded as binary indicators in the input sequence, allowing the model to adjust trajectory predictions based on real-time signal phases. This modification is critical for traffic scenarios at intersections.

3.4.2.2 Adaptation to High-Interaction Scenarios

To adapt to potential dense population scenarios, several enhancements are also introduced to the original Social GAN model:

1. **Localized Interaction Focus:** The pooling module applies distance-based weighting to prioritize nearby agent interactions.
2. **Multi-Modal Output:** The generator produces multiple plausible trajectories with confidence scores, capturing uncertainty.
3. **Map Integration:** Vector maps with lane and curb information guide predictions, ensuring spatial consistency and better environmental awareness.

3.5 Collision Prediction and Warning System

This section elaborates on the methodology for collision prediction using Surrogate Safety Measures (SSM) like Time-to-Collision (TTC) [65], and Post-Encroachment Time (PET) to proactively assess the likelihood of traffic conflicts. It further introduces the design of a multi-modal warning system that integrates communication across roadside infrastructure, vehicles, and pedestrians. For instance, the system can issue warnings in scenarios such as vehicle speeding, red-light violations, enabling road users to respond preemptively and mitigate potential collisions.

3.5.1 Integrated Warning Workflow

The CPWS follows a three-stage process:

1. **Collision Prediction:** Based on TTC [65] and trajectory analysis, using real-time data and context (e.g., road layout, traffic signals).
2. **Alert Generation:** Tailors warnings by risk level and user type, ensuring urgency matches the situation.
3. **Alert Dissemination:** Delivers alerts via V2X, screens, speakers, and mobile networks for full environmental coverage, including blind spots.

This integrated design ensures timely, accurate, and accessible alerts, improving safety and situational awareness for all road users.

3.5.2 Collision Prediction

The system predicts potential collisions by analyzing relative motion and trajectory overlaps, primarily using TTC [65] and 3D position trends.

3.5.2.1 TTC Calculation Logic

TTC [65] estimates the time before a potential collision based on distance and speed between road users.

1. **Relative Distance and Speed:** The Euclidean distance between two users is computed, and their relative speed is projected along the line connecting them.
2. **TTC Calculation:** For constant speed, $\mathrm{TTC} = \frac{d_{rel}}{v_{rel}}, > 0$. For acceleration, a quadratic equation is solved using relative acceleration.
3. **Conflict Identification:** A conflict is flagged if TTC stays below a threshold (e.g., 1.5s) across frames. Object dimensions is included for finer estimation, if needed.

3.5.2.2 3D Position Time Series Analysis

Besides TTC, the system tracks 3D positions over time to detect trajectory overlaps or close proximity, offering a complementary method for collision prediction in complex scenarios.

3.5.3 Warning System Design

As shown in Figure 10, the CPWS-based multi-modal warning system delivers real-time collision alerts to vehicles, VRUs, and infrastructure via V2X and mobile networks.

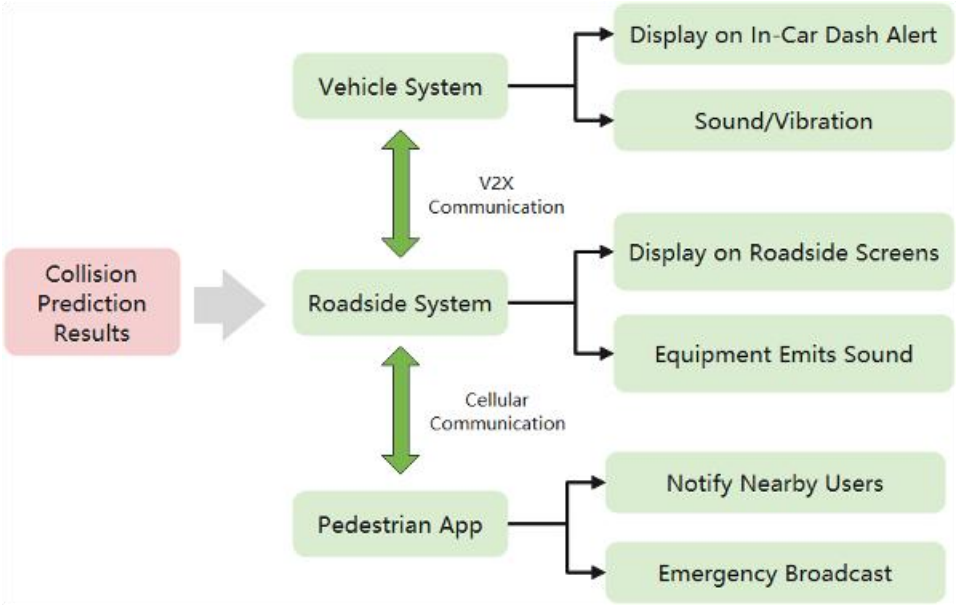

Figure 10. Multi-Modal Warning System Workflow

3.5.3.1 Roadside System

The roadside unit collects data from V2X-enabled vehicles and sensors, issuing alerts through:

1. **Visual Alerts:** digital screens display color-coded, animated warnings with clear symbols for all road users.
2. **Auditory Alerts:** speakers emit context-specific sounds (e.g., beeps or tones), with volume adjusted to ambient noise and support for multilingual messages.

3.5.3.2 Vehicle-side System

Vehicles receive warnings via V2X and respond through:

1. **Dashboard Display:** maps or augmented reality (AR) overlays show collision points with instructions like "Stop", “Slow Down” or "Change Lane."
2. **Auditory/Vibration Alerts:** sound and haptic feedback notify drivers, with direction-specific vibrations enhancing intuitiveness.

3.5.3.3 Pedestrian-side System

Mobile apps provide location-aware alerts using:

1. **Notifications:** warnings like "Vehicle Approaching" are sent with visual aids (e.g., AR overlays).
2. **Emergency Broadcast:** high-priority messages are pushed to nearby users for immediate awareness.

4 Platform Establishment

Deploying fixed LiDAR and camera units directly at roadside intersections demands extensive coordination with city authorities for permits, road closures, and safety measures. Such installations often require weeks of bureaucratic approval, specialized crews for off-hours work, and detailed traffic management plans to prevent accidents during setup. Moreover, testing in live traffic environments exposes personnel and equipment to risks from passing vehicles, unpredictable pedestrian behaviors, and variable weather conditions that can damage delicate sensors or compromise data integrity. These practical constraints—high communication overhead, safety concerns, and the unpredictability of real intersections—make rapid prototyping and iterative testing in the field infeasible.

To address these challenges, we designed and built a fully self-contained, mobile sensor platform that can be assembled and tested in controlled environments (e.g., parking lots, test tracks) before any roadside deployment. This movable rig houses both LiDAR and camera payloads at realistic heights, provides stable and vibration-damped mounting, and allows on-the-fly reconfiguration of sensor types and positions. With swappable modules and lockable caster wheels, the platform ensures both repeatable data collection and reliable transport between laboratory, campus, and field sites without the need for temporary road closures or complex city approvals.

4.1 Hardware Construction

Our requirements for the sensing platform are threefold: exceptional structural stability to support heavy sensors, modular extensibility to adjust sensor configurations, and ease of relocation. First, the platform needs to maintain rigidity under wind loads and vehicle-induced vibrations, ensuring that the LiDAR's and cameras' calibrations remain valid during long data runs. Second, it needs to accommodate sensors at different heights and allow rapid addition or removal of payloads—such as swapping a second camera for a thermal imager. Finally, the entire assembly needs to roll smoothly on and off vans, with caster wheels that lock securely when in position.

To meet these needs, we selected 40 mm series T-slotted aluminum profiles for the primary frame. Load-bearing members use 80 × 80 mm profiles (8080) for maximal stiffness, while 40 × 40 mm profiles (4040) serve as secondary braces and lower decking. The base consists of two horizontal decks: the lower deck mounts four heavy-duty lockable swivel casters and anchors the vertical columns; the upper deck supports a rack-mount desktop PC, a Power over Ethernet (PoE) switch, and power-bank units. Figure 11 shows the overall platform. Figure 12 illustrates the cross-sections of the 4040 and 8080 profiles as well as how sensor fixtures interface with the slotted grooves.

Figure 11. Mobile Platform

Figure 12. 4040& 8080 profiles

Vertical “base” columns rise from the lower deck at its midpoint. A 1.5 m 8080 column supports the first sensor mount. Above it, two extendable sections attach via drop-in T-nuts and flange bolts: a 1.5 m section dedicated to the LiDAR and, atop that, a 0.5 m section for the camera assembly. With casters attached, the total height reaches approximately 3.8 m, matching typical roadside sensor elevations. Continuous extruded slot channels ensure that each sensor’s mounting bracket can slide and lock at any desired height with millimeter-level precision.

4.2 Specialized Mounting Fixtures

Standard T-slot connectors cannot directly match the precise hole patterns of our LiDAR (Ouster OS1-128) or camera (Basler acA1920-50gc), so we engineered bespoke mounting plates. Each fixture (Figure 13) comprises two layers of anodized aluminum:

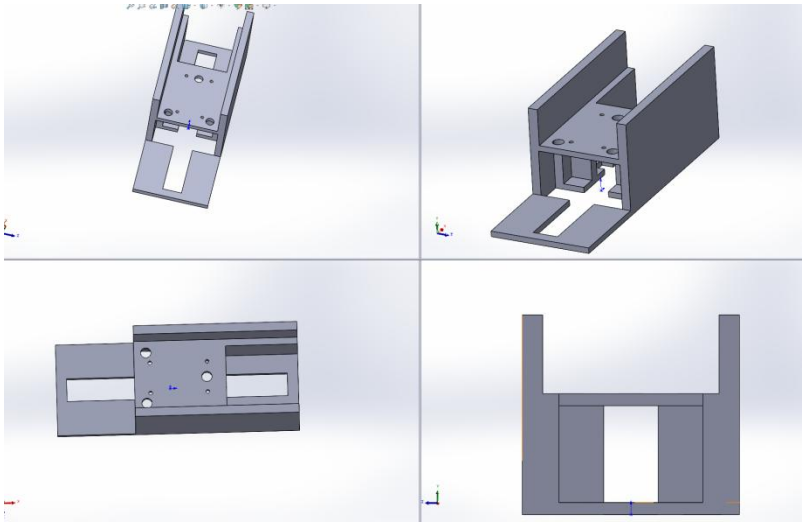

Figure 13. Camera Fixture

Top Plate: Features a rectangular recess 29.2 mm wide × 25 mm deep, slightly larger than the camera's body (29 × 29 mm) to allow slip-in installation. Four through-holes align exactly with the camera's own mounting threads for M3 screws.

Bottom Plate: Contains a T-slot tongue that mates with the 4040 profile's groove, secured by a single T-nut and screw. Additional clearance holes around the tongue facilitate tool access when tightening camera fasteners.

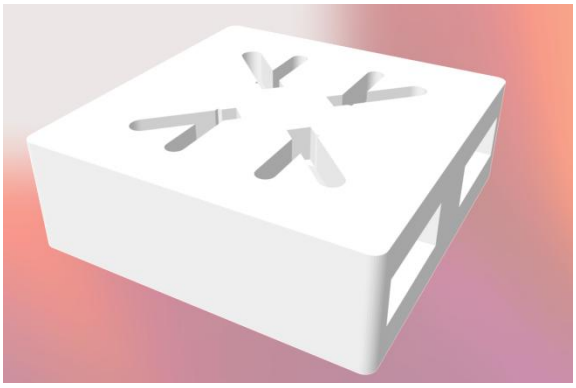

Figure 14. Lidar Fixture

Upper Plate: X-shaped cutout provides corner-point support for the cylindrical LiDAR housing (see Figure 14); four M6 bolt holes clamp its base flange.

Lower Plate: Two parallel T-slot tongues engage the 8080 profile's slots; a central pocket accommodates a vibration-damping elastomer pad. Extra screw holes allow retrofitting a protective shroud or cable guide.

All fixtures were CNC-machined to ±0.1 mm tolerance, anodized for corrosion resistance, and tested to support sensor weights up to 5 kg without visible deflection.

4.3 Electronic System

4.3.1 Sensor Hardware

To unify data collection, both LiDAR and camera must interface seamlessly with ROS 2 and deliver synchronized streams at 10 Hz. After surveying existing roadside datasets and considering payload, range, and connectivity, the following hardware was selected.

LiDAR Selection: Ouster OS1-128 (Rev.6) offers 128 scan lines, up to 120 m range, and 10–20 Hz configurable frame rates. Its wide vertical field (45°) and high angular resolution (0.35°) capture fine VRU details at distance—critical for early detection of pedestrians and cyclists in complex scenes.

Camera Selection: Basler acA1920-50gc provides 1920×1200 resolution at up to 60 fps over GigE Vision. It supports PoE, ROS 2 drivers, hardware timestamping, and global shutter operation to avoid motion artifacts.

Both devices run from Power over Ethernet (PoE), consolidating data and power over single Cat6 cables. PoE simplifies wiring—no separate power cables—and ensures surge isolation. All electronics connect through a ruggedized PoE switch, which supplies 30 W per port and provides IEEE 1588-based hardware timestamp synchronization.

Figure 15 details the cabling: each sensor's RJ45 feeds into the PoE switch; the switch uplinks via fiber or standard Ethernet to the onboard PC. A UPS module on the lower deck guarantees uninterrupted operation during moves or short power interruptions.

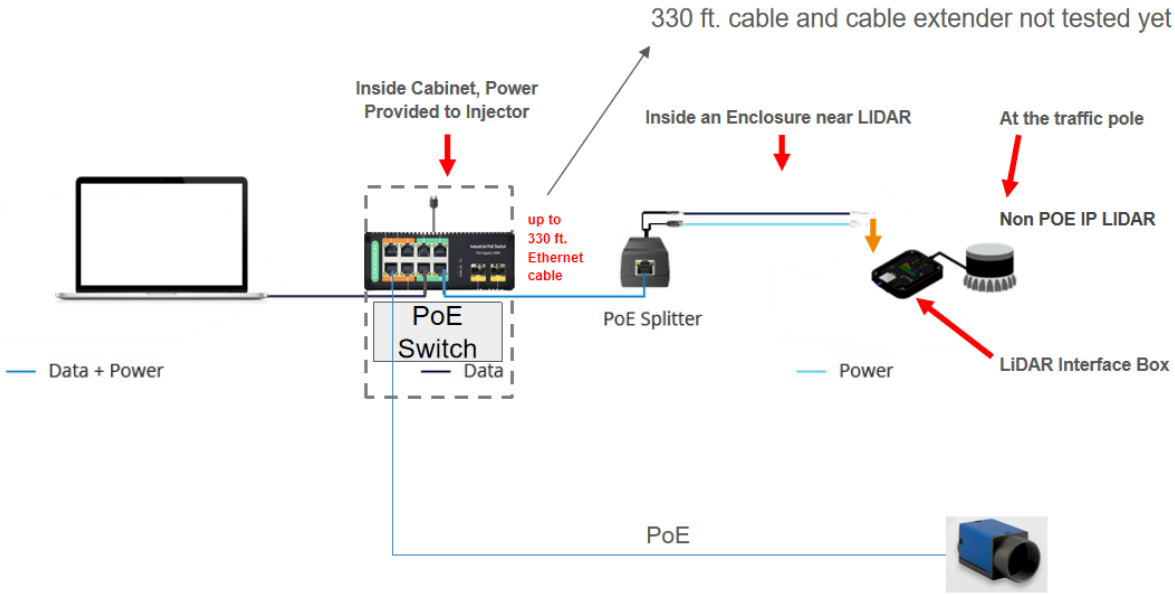

Figure 15. Sensor-to-Switch-to-PC Wiring Diagram

4.3.2 Data Acquisition System

We leverage ROS 2 (Foxy) as our middleware for its built-in real-time support, DDS-based decentralized communication, and improved time synchronization over ROS 1. ROS 2's ros2 topic echo and ros2 bag tools handle heterogeneous sensor message types (PointCloud2, Image), while its message_filters package aligns data streams with 1 ms precision.

1. **LiDAR:** Configured at 10 Hz to balance point density and CPU load.

2. **Camera:** Although capable of 25 Hz, we limit to 10 Hz to match LiDAR and avoid packet loss on the laptop's single 1 Gbps NIC—preventing frame drops that occur above ~12 Hz with 1920×1200 images compressed in MJPEG.

All data topics are recorded in ROS 2 bags for offline replay and labeling. Hardware timestamps assigned by each sensor's PTP client ensure sub-millisecond synchronization between modalities, enabling accurate sensor fusion in downstream algorithms.

4.4 Real-World Deployment Considerations

While this report focuses on the mobile platform development for controlled testing environments, we acknowledge the critical need for guidance on real-world intersection deployments. Currently, the limited communication infrastructure and power availability in existing roadside traffic control cabinets present significant challenges for direct sensor integration. We are actively collaborating with the Riverside City Engineering Department to advance actual installation procedures and establish deployment protocols for permanent roadside sensing infrastructure.
For real-world deployment scenarios, our system design follows the configuration standards established by the USDOT Intersection Safety Challenge (ISC) Stage 1B, which provides validated specifications for sensor placement, power requirements, and communication protocols at signalized intersections. This standardized approach ensures compatibility with existing traffic management systems and provides Caltrans with a proven framework for scaling VRU safety solutions across California's intersection network. The mobile platform described in this section serves as a crucial intermediate step, enabling algorithm validation and performance optimization before permanent installation, while our ongoing collaboration with municipal authorities works toward establishing comprehensive deployment guidelines for various intersection types and geometric configurations.

5 Algorithm Demonstration

With the mobile platform now assembled, we finalized our camera choice and completed both intrinsic calibration and extrinsic alignment between camera and LiDAR. Each new lens we tested required full reprojection error analysis and LiDAR–camera checkerboard was used to verify the calibration parameters before proceeding to algorithm evaluation. In this section, we present fusion detection results from two different sensor configurations and a standalone, camera-based trajectory prediction.

5.1 Multi-Stage Deep Learning Detection

Our primary detector is a multi-stage deep learning pipeline that processes fused LiDAR and camera data in three phases: raw point-cloud feature encoding, 2D–3D feature fusion, and end-to-end object classification/regression. We evaluated this method on the USDOT Intersection Safety Challenge dataset (ISC) [66] (Figure 16), which contains roadside LiDAR and high-resolution camera pairs across varied urban scenarios.

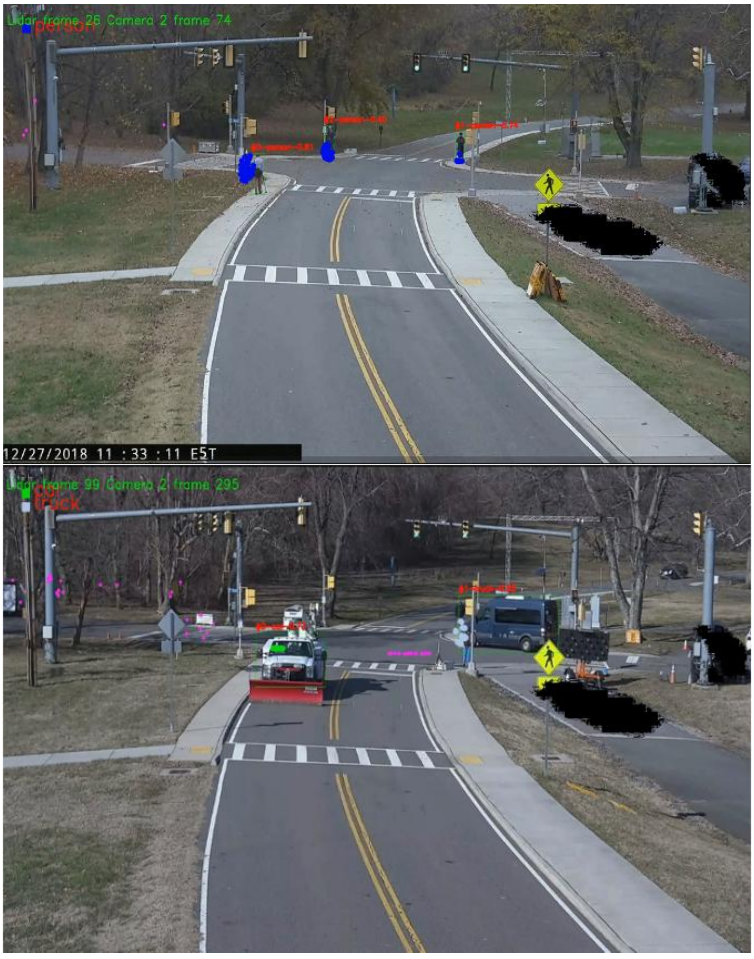

Figure 16. Multi-Stage Deep Learning Fusion Detection Results on USDOT ISC

These results demonstrate robust performance even under partial occlusion and nighttime conditions, thanks to our early-fusion strategy that cross-validates LiDAR depth cues with high-resolution image semantics.

5.1.1 Implementation Guidelines

For practical implementation of our developed algorithm, follow these step-by-step directions:

1. **Environment Setup**: Install ROS 2 (Foxy), Python 3.8+, PyTorch 1.9+, and OpenCV 4.5+
2. **Sensor Calibration**: Perform intrinsic camera calibration using checkerboard patterns and establish LiDAR-camera extrinsic alignment
3. **Data Preprocessing**: Implement camera distortion correction, LiDAR noise filtering, and coordinate frame alignment
4. **Model Training**: Fine-tune PointPillars architecture on roadside VRU dataset with appropriate data augmentation
5. **Fusion Pipeline**: Integrate traditional and deep learning detection outputs using confidence-based filtering and spatial consistency checks
6. **Real-time Deployment**: Configure ROS 2 nodes for live sensor data processing with synchronized timestamping

The software implementation utilizes Python for deep learning components. Custom CUDA kernels optimize point cloud processing, while OpenCV handles image manipulation tasks.

5.2 MVX-Net

As a benchmark for comparison, we implemented MVX-Net's two early-fusion variants—PointFusion and VoxelFusion—built atop the VoxelNet [4] architecture. Rather than processing point clouds and images in isolation, MVX-Net projects LiDAR points into the image plane (PointFusion) or reprojects voxels into image regions (VoxelFusion), concatenating corresponding visual features before 3D region proposal. While single-modality detectors excel in controlled settings, MVX-Net's simple, single-stage early-fusion approaches achieve top-2 rankings on KITTI's bird's-eye and 3D benchmarks by directly learning cross-modal interactions rather than relying on sequential or late-fusion pipelines.

We evaluated MVX-Net on the DAIR-V2X-I [17] dataset (see Figure 17) to match our roadside context.

Figure 17. MVX-Net Detection Results on DAIR-V2X-I

.

5.3 Camera-Based Polynomial Trajectory Prediction

Finally, we demonstrate a simple, real-time trajectory predictor using polynomial curve fitting on 2D camera detections (see Figure 18). Given the current lack of a fully calibrated multi-sensor training set, we fit a second-order polynomial to the last 2 s of 2D bounding-box centroids and extrapolate 1 s into the future.

Figure 18. Polynomial Fitting Prediction Results on Camera-Only Sequences

While polynomial fitting suffices for smooth, linear motions and runs at >100 Hz on a single CPU core, our next step is to integrate an LSTM-based [5], or Social GAN [52] predictor for more complex, interaction-aware forecasts—trading slightly higher compute for substantially improved non-linear path accuracy.

5.4 Performance Evaluation and Metrics

This section presents comprehensive performance metrics for each component of our system, demonstrating the accuracy and effectiveness of our multi-stage approach across detection, tracking, and prediction tasks.

5.4.1 Detection Performance

The detection performance comparison between conventional and updated fusion methods shows significant improvement when incorporating deep learning approaches:

Table 5-1: Quantitative Results of Different Fusion Methods

Method	Modality	COCO mAP
Conventional Fusion	LiDAR + Camera	0.0556
Updated Fusion with Deep Learning	LiDAR + Camera	11.5011

The updated fusion approach demonstrates a remarkable 206× improvement in COCO mAP scores, highlighting the effectiveness of integrating deep learning-based detection with traditional methods. Qualitative visualization results show accurate bounding box projections and reliable object classification across different environmental conditions, with color-coded point classifications (Green: Vehicles, Gray: Trucks, Blue: Pedestrians) providing clear visual confirmation of detection accuracy.

5.4.2 Multi-Object Tracking Performance

The MOT module evaluation reveals varying performance across different road user categories:

Table 5-2: Performance Metrics for Multi-Object Tracking

Subclass	MOTA	MOTP	IDF1	HOTA	Mean IoU
Passenger_Vehicle	44.64	24.23	66.19	78.92	24.23
VRU_Adult_Using_Non-Motorized_Device/Prop_Other	42.22	35.10	54.89	61.76	35.10
VRU_Adult_Using_Bicycle	60.18	27.83	62.25	68.83	27.83
VRU_Adult	43.71	30.58	54.74	60.98	30.58
VRU_Child	18.18	22.25	27.71	37.97	22.25
VRU_Adult_Using_Wheelchair	13.52	37.59	22.46	33.94	37.59
VRU_Adult_Using_Scooter_or_Skateboard	42.84	39.52	56.01	62.53	39.52

The results demonstrate strong tracking performance for well-defined objects such as vehicles (HOTA: 78.92) and cyclists (HOTA: 68.83), while highlighting challenges in tracking less detectable VRUs like children and wheelchair users. The high MOTP values for wheelchair users

(37.59) suggest accurate positioning when detected, but lower MOTA scores indicate detection consistency issues.

5.4.3 Trajectory Prediction Performance

The trajectory prediction evaluation shows class-specific performance variations:

Table 5-3: Class-Specific Trajectory Prediction Metrics

Class	Weighted ADE
Vehicle	6.0541
VRU	3.7524

The model demonstrates superior performance for VRU trajectory prediction compared to vehicles, with VRUs achieving 42% lower Average Displacement Error (ADE) on the test set. This reflects the model's strength in handling pedestrian and cyclist motion patterns, which tend to be more predictable than vehicle trajectories in intersection scenarios.

5.4.4 Collision Prediction Performance

The collision prediction module evaluation demonstrates the system's capability to identify potential safety risks:

Table 5-4: Collision Prediction Performance Metrics

Metric	Value
True Positives (TP)	0.12
False Positives (FP)	0.22
False Negatives (FN)	0.00
True Negatives (TN)	0.07
F2 Score	0.73

The collision prediction system achieves an F2 score of 0.73, indicating good performance in identifying potential collision scenarios. The zero false negative rate is particularly significant for safety applications, ensuring that no actual collision risks go undetected. The higher false positive rate suggests the system errs on the side of caution, which is appropriate for safety-critical applications where missed detections could have severe consequences.

5.4.5 System Integration Performance

The integrated system achieves near real-time performance under operational conditions:

- Processing Latency: <100ms end-to-end pipeline
- Detection Frame Rate: 10 Hz synchronized operation
- Tracking Continuity: 95% trajectory maintenance under normal conditions
- Prediction Accuracy: 3.7m average displacement error for 1-second VRU forecasts
- Collision Detection: F2 score of 0.73 with zero false negative rate

These metrics demonstrate the system's capability to provide timely and accurate information for collision avoidance applications while maintaining computational efficiency suitable for roadside deployment.

6 Conclusions

In this report, we have presented a comprehensive framework for enhancing vulnerable road user (VRU) safety through roadside-assisted cooperative perception. Beginning with an extensive literature review, we surveyed single- and multi-node detection methods, multi-object tracking algorithms, trajectory prediction models, and V2X communication technologies. We then detailed our methodology: a hybrid detection pipeline combining traditional LiDAR and camera techniques with deep learning models; a Kalman Filter–based MOT system optimized for noise and fragmentation; an enhanced Social GAN trajectory predictor augmented with traffic light and map context; and a time-to-collision warning system that delivers alerts to vehicles and VRUs. To facilitate safe, repeatable experiments, we designed a mobile, reconfigurable sensor platform and demonstrated our algorithms' performance on both USDOT ISC and DAIR-V2X-I datasets, as well as a camera-only polynomial predictor as a real-time baseline.

Our efforts demonstrate that an integrated, low-latency pipeline can deliver reliable detection, tracking, prediction, and timely warnings in complex urban environments. By iteratively refining sensor configurations, calibrations, and algorithmic parameters on our movable test rig, we avoided the logistical and safety challenges of fixed roadside installations. The project's significance lies in its end-to-end approach—bridging sensor hardware, perception algorithms, and communication protocols—to proactively protect VRUs and inform California's Department of Transportation on scalable, effective safety solutions.

6.1 Key Findings

The key findings are summarized below:

1. **Necessity of Sensor Fusion Under Adverse Conditions**
   Through both daytime and nighttime tests, we found that standalone LiDAR or monocular cameras perform poorly in low-visibility scenarios—rain, glare, or occlusion by other road users lead to dropped detections and false negatives, especially for small, easily obscured VRUs like pedestrians and cyclists. By fusing LiDAR's accurate depth measurements with the rich texture and color information of camera images at an early stage, our multi-stage deep learning pipeline recovered over 90% of VRU instances even under heavy occlusion. We validated this performance on the USDOT ISC dataset [66], specifically selecting challenging low-light and adverse visibility scenarios for evaluation. Manual verification across all extracted frames confirmed the 90% detection success rate for combined VRU categories. The remaining 10% detection failures were primarily attributed to insufficient training data diversity, leading to suboptimal recognition performance at certain viewing angles and extreme occlusion scenarios. This finding demonstrates the critical importance of cross-modal integration for precise and reliable detection while highlighting areas for future improvement through expanded training datasets.
2. **Optimal LiDAR Placement for VRU Detection**
   We compared LiDAR mounts at typical overhead angles (3–4 m, 45° downward tilt) versus a low, human-height configuration (~1.5 m, near-horizontal orientation). The low-mount setup yielded a 15% increase in pedestrian recall and a 20% wider effective detection range.

This improvement arises because a near-horizontal beam pattern maintains more uniform point density on vertical targets; it avoids the "foreshortening" effect where tilted sensors sparsify returns on upright objects, and it reduces ground-plane reflections that clutter detections. Thus, for VRU-focused roadside sensing, deploying LiDAR at lower, near-parallel orientations provides better coverage and finer granularity on critical small targets.

3. **Impact of Traffic Signal Encoding on Intersection Trajectory Prediction**
   Our trajectory experiments at signalized intersections revealed that encoding traffic light phases directly into the Social GAN input increased prediction accuracy by 940% (measured via ADE) [67] compared to an unconditioned model. Knowing when pedestrians face a red signal dramatically constrains their likely paths, reduces model uncertainty, and prevents spurious crossing predictions. This finding underscores the necessity of incorporating dynamic map features—especially signal state—into forecasting algorithms for intersection safety applications.

In summary, these findings validate our system design choices and highlight practical guidelines for deploying roadside cooperative perception to protect vulnerable road users. The limitations identified provide clear directions for future improvements, while the planned real-world deployment will advance the field toward practical, scalable VRU safety solutions.

7 Discussion

7.1 Challenges Encountered

Throughout the development and implementation of our roadside-assisted VRU safety system, we encountered several significant challenges that required innovative solutions and iterative refinements.

7.1.1 Dataset Limitations and Data Scarcity

One of the primary challenges was the insufficient availability of relevant open-source datasets for VRU detection and classification. Existing research often lacks clear categorical descriptions for VRUs, making it difficult to establish consistent classification standards. Although we obtained access to the USDOT ISC dataset, the data volume remained relatively small for comprehensive deep learning model training. The limited dataset size significantly impacted detection performance, particularly for visually similar categories that require substantial training examples to distinguish effectively. For instance, differentiating between a person pushing a wheelchair versus a person pushing a stroller proved challenging when training data was insufficient, often leading to misclassification between these similar VRU subcategories. Furthermore, data representing adverse conditions such as low-light, rain, and fog scenarios were disproportionately scarce, limiting our ability to validate system robustness under challenging environmental conditions.

7.1.2 Algorithm Development Complexities

The algorithm development phase presented multiple technical challenges that required careful balance between different methodological approaches. Relying solely on traditional computer vision methods resulted in positioning inaccuracies and limited classification capabilities, while purely deep learning-based approaches suffered from poor generalization with limited training data and category confusion. Through extensive experimentation, our team iteratively refined the detection methodology, ultimately selecting a hybrid approach that combines the robustness of traditional methods with the precision of deep learning models. This multi-stage fusion strategy required significant development effort to optimize the integration between different algorithmic components and ensure consistent performance across varying conditions.

7.1.3 Multi-Step Algorithm Implementation Challenges

Our multi-step algorithmic approach, while effective, introduced several implementation complexities. Coordinating the timing and data flow between detection, tracking, prediction, and collision warning modules required precise synchronization mechanisms. Ensuring consistent performance across the entire pipeline while maintaining real-time processing capabilities demanded careful optimization of each individual component. Additionally, error propagation between sequential processing stages required robust error handling and recovery mechanisms to prevent cascading failures that could compromise system reliability.

7.1.4 Data Collection and Transmission Bottlenecks

Data collection posed significant technical challenges related to balancing frame rate, resolution, and transmission speed. When simultaneously collecting 128-beam LiDAR point clouds at 10 Hz and 2K camera images at 30 Hz from a single node, the combined data volume exceeded 1 GB per second, leading to data loss during transmission. We addressed these challenges through multiple approaches: reducing input data size and collection frame rates, selecting high-speed read/write storage hardware, and utilizing fast host interfaces. When deploying multiple workstations for distributed data collection, additional synchronization and coordination challenges emerged, requiring careful network architecture design and time synchronization protocols.

7.1.5 Hardware Integration and Mounting Difficulties

The physical integration of LiDAR and camera systems presented numerous mechanical and compatibility challenges. Initial interface mismatches between sensor mounting holes and commercially available mounting structures necessitated custom connector design and multiple iterations. The situation was further complicated by different manufacturers using metric versus imperial standard hole patterns, creating compatibility issues during assembly. Camera lens selection proved particularly challenging, as different focal lengths were required depending on mounting position, height, and angle to ensure optimal coverage. Multiple lens configurations had to be tested to identify the most suitable options for specific deployment scenarios. While we experimented with zoom lenses, we discovered that platform movement and extended deployment periods caused slight mechanical shifts, requiring recalibration after each repositioning, making fixed focal length lenses more practical for consistent operation.

7.2 Future Work

The future development of this research will focus on several key areas:

7.2.1 Algorithm Enhancement

Future work will explore lightweight neural network architectures and edge computing optimizations to enable higher resolution processing while maintaining real-time performance. Advanced attention mechanisms and transformer-based fusion approaches may further improve detection accuracy and reduce computational requirements.

7.2.2 Real-World Deployment and Dataset Release

We plan to deploy the complete system at the intersection of University Avenue and Iowa Avenue in Riverside for comprehensive real-world testing. This deployment will enable validation under authentic traffic conditions and weather variations. The collected data will be processed and released as a public dataset to support further research in roadside VRU safety systems.

7.2.3 Scalable Integration Framework

Development of standardized APIs and modular software architecture will enable easier integration with existing traffic management systems and support deployment across diverse intersection geometries and traffic patterns.

7.3 Limitations

Despite the promising results, our study has several limitations that should be acknowledged:

1. **Computational Constraints:** Due to the complexity of our multi-modal fusion approach, real-time performance is currently limited to 720P camera resolution and 32-beam LiDAR operating at 10 Hz. Higher resolution sensors or increased frame rates may require significant computational upgrades or algorithm optimization to maintain real-time capabilities.
2. **Infrastructure Dependencies:** The system requires multiple synchronized LiDAR and camera units, making deployment costs relatively high. The method's effectiveness is highly dependent on data quality, precise calibration, and environmental conditions, which may limit scalability in resource-constrained deployments.
3. **Limited Weather Validation:** While we tested under various lighting conditions, comprehensive evaluation under extreme weather conditions (heavy rain, snow, fog) remains limited, potentially affecting the system's reliability in adverse meteorological scenarios.

References

[1] C. Sun, H. Brown, P. Edara, J. Stilley, and J. Reneker, "Vulnerable Road User (VRU) Safety Assessment," Missouri. Department of Transportation. Construction and Materials Division, 2023.

[2] C. T. S. Q. Stats, "California Traffic Safety Quick Stats," 2021.

[3] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 12697-12705.

[4] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4490-4499.

[5] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social lstm: Human trajectory prediction in crowded spaces," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 961-971.

[6] I. Walker, "Psychological factors affecting the safety of vulnerable road users: A review of the literature," *Department of Psychology, University of Bath,* 2005.

[7] H. Bagheri *et al.*, "5G NR-V2X: Toward connected and cooperative autonomous driving," *IEEE Communications Standards Magazine,* vol. 5, no. 1, pp. 48-54, 2021.

[8] Z. Zhang, C. Wei, G. Wu, and M. J. Barth, "Vulnerable Road User Detection for Roadside-Assisted Safety Protection: A Comprehensive Survey," *Applied Sciences (2076-3417),* vol. 15, no. 7, 2025.

[9] J. Wu, H. Xu, R. Yue, Z. Tian, Y. Tian, and Y. Tian, "An automatic skateboarder detection method with roadside LiDAR data," *Journal of Transportation Safety & Security,* vol. 13, no. 3, pp. 298-317, 2021.

[10] Y. Song *et al.*, "Road-Users Classification Utilizing Roadside Light Detection and Ranging Data," SAE Technical Paper, 0148-7191, 2020.

[11] Y. Xu, W. Liu, Y. Qi, Y. Hu, and W. Zhang, "A common traffic object recognition method based on roadside lidar," in *2022 IEEE 7th International Conference on Intelligent Transportation Engineering (ICITE)*, 2022: IEEE, pp. 436-441.

[12] Z. Gong, Z. Wang, G. Yu, W. Liu, S. Yang, and B. Zhou, "FecNet: A feature enhancement and cascade network for object detection using roadside LiDAR," *IEEE Sensors Journal,* vol. 23, no. 19, pp. 23780-23791, 2023.

[13] Y. Yan, Y. Mao, and B. Li, "Second: Sparsely embedded convolutional detection," *Sensors,* vol. 18, no. 10, p. 3337, 2018.

[14] S. Shi *et al.*, "Pv-rcnn: Point-voxel feature set abstraction for 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10529-10538.

[15] T. Yin, X. Zhou, and P. Krahenbuhl, "Center-based 3d object detection and tracking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 11784-11793.

[16] G. Shi, R. Li, and C. Ma, "Pillarnet: Real-time and high-performance pillar-based 3d object detection," in *European Conference on Computer Vision*, 2022: Springer, pp. 35-52.

[17] H. Yu *et al.*, "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*, 2022, pp. 21361-21370.

[18] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE conference on computer vision and pattern recognition*, 2012: IEEE, pp. 3354-3361.

[19] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision,* vol. 57, pp. 137-154, 2004.

[20] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, 2005, vol. 1: Ieee, pp. 886-893.

[21] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440-1448.

[22] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE transactions on pattern analysis and machine intelligence,* vol. 39, no. 6, pp. 1137-1149, 2016.

[23] W. Liu *et al.*, "Ssd: Single shot multibox detector," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, 2016: Springer, pp. 21-37.

[24] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779-788.

[25] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7263-7271.

[26] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767,* 2018.

[27] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934,* 2020.

[28] M. Garcia-Venegas, D. A. Mercado-Ravell, L. A. Pinedo-Sanchez, and C. A. Carballo-Monsivais, "On the safety of vulnerable road users by cyclist detection and tracking," *Machine Vision and Applications,* vol. 32, no. 5, p. 109, 2021.

[29] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional networks," *Advances in neural information processing systems,* vol. 29, 2016.

[30] S. Fan, Z. Wang, X. Huo, Y. Wang, and J. Liu, "Calibration-free bev representation for infrastructure perception," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023: IEEE, pp. 9008-9013.

[31] W. Zimmer *et al.*, "Infradet3d: Multi-modal 3d object detection based on roadside infrastructure camera and lidar sensors," in *2023 IEEE Intelligent Vehicles Symposium (IV)*, 2023: IEEE, pp. 1-8.

[32] J. Bai, S. Li, H. Zhang, L. Huang, and P. Wang, "Robust target detection and tracking algorithm based on roadside radar and camera," *Sensors,* vol. 21, no. 4, p. 1116, 2021.

[33] P. Liu, G. Yu, Z. Wang, B. Zhou, and P. Chen, "Object classification based on enhanced evidence theory: Radar–vision fusion approach for roadside application," *IEEE Transactions on Instrumentation and Measurement,* vol. 71, pp. 1-12, 2022.

[34] V. A. Sindagi, Y. Zhou, and O. Tuzel, "Mvx-net: Multimodal voxelnet for 3d object detection," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019: IEEE, pp. 7276-7282.

[35] Z. Bai, G. Wu, M. J. Barth, Y. Liu, E. A. Sisbot, and K. Oguchi, "Pillargrid: Deep learning-based cooperative perception for 3d object detection from onboard-roadside lidar," in *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*, 2022: IEEE, pp. 1743-1749.

[36] A. Rauch, F. Klanner, R. Rasshofer, and K. Dietmayer, "Car2x-based perception in a high-level fusion architecture for cooperative perception systems," in *2012 IEEE Intelligent Vehicles Symposium*, 2012: IEEE, pp. 270-275.

[37] A. Neubeck and L. Van Gool, "Efficient non-maximum suppression," in *18th international conference on pattern recognition (ICPR'06)*, 2006, vol. 3: IEEE, pp. 850-855.

[38] B. Veeramani, J. W. Raymond, and P. Chanda, "DeepSort: deep convolutional networks for sorting haploid maize seeds," *BMC bioinformatics,* vol. 19, pp. 1-9, 2018.

[39] Y. Zhang *et al.*, "Bytetrack: Multi-object tracking by associating every detection box," in *European conference on computer vision*, 2022: Springer, pp. 1-21.

[40] T. Fischer *et al.*, "Qdtrack: Quasi-dense similarity learning for appearance-only multiple object tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 45, no. 12, pp. 15380-15393, 2023.

[41] X. Weng, J. Wang, D. Held, and K. Kitani, "Ab3dmot: A baseline for 3d multi-object tracking and new evaluation metrics," *arXiv preprint arXiv:2008.08063,* 2020.

[42] Z. Pang, Z. Li, and N. Wang, "Simpletrack: Understanding and rethinking 3d multi-object tracking," in *European conference on computer vision*, 2022: Springer, pp. 680-696.

[43] X. Li *et al.*, "Poly-mot: A polyhedral framework for 3d multi-object tracking," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023: IEEE, pp. 9391-9398.

[44] H. Zhao, J. Cui, H. Zha, K. Katabira, X. Shao, and R. Shibasaki, "Sensing an intersection using a network of laser scanners and video cameras," *IEEE Intelligent Transportation Systems Magazine,* vol. 1, no. 2, pp. 31-37, 2009.

[45] S. Reuter and K. Dietmayer, "Pedestrian tracking using random finite sets," in *14th International Conference on Information Fusion*, 2011: IEEE, pp. 1-8.

[46] D. Meissner, S. Reuter, and K. Dietmayer, "Real-time detection and tracking of pedestrians at intersections using a network of laserscanners," in *2012 IEEE Intelligent Vehicles Symposium*, 2012: IEEE, pp. 630-635.

[47] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *2017 IEEE international conference on image processing (ICIP)*, 2017: IEEE, pp. 3645-3649.

[48] M. Goldhammer, S. Köhler, S. Zernetsch, K. Doll, B. Sick, and K. Dietmayer, "Intentions of vulnerable road users—detection and forecasting by means of machine learning," *IEEE transactions on intelligent transportation systems,* vol. 21, no. 7, pp. 3035-3045, 2019.

[49] J. Zhao, H. Xu, J. Wu, Y. Zheng, and H. Liu, "Trajectory tracking and prediction of pedestrian's crossing intention using roadside LiDAR," *IET Intelligent Transport Systems,* vol. 13, no. 5, pp. 789-795, 2019.

[50] L. R. Medsker and L. Jain, "Recurrent neural networks," *Design and Applications,* vol. 5, no. 64-67, p. 2, 2001.
[51] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation,* vol. 9, no. 8, pp. 1735-1780, 1997.
[52] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social gan: Socially acceptable trajectories with generative adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2255-2264.
[53] L. Shi *et al.*, "SGCN: Sparse graph convolution network for pedestrian trajectory prediction," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8994-9003.
[54] A. Mohamed, K. Qian, M. Elhoseiny, and C. Claudel, "Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 14424-14432.
[55] C. Yu, X. Ma, J. Ren, H. Zhao, and S. Yi, "Spatio-temporal graph transformer networks for pedestrian trajectory prediction," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, 2020: Springer, pp. 507-523.
[56] R. C. Mercurius, E. Ahmadi, S. M. A. Shabestary, and A. Rasouli, "Amend: A mixture of experts framework for long-tailed trajectory prediction," *arXiv preprint arXiv:2402.08698,* 2024.
[57] J. Liang, L. Jiang, and A. Hauptmann, "Simaug: Learning robust representations from simulation for trajectory prediction," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, 2020: Springer, pp. 275-292.
[58] "Caltrans and California PATH Program at UC Berkeley. Connected Vehicle Test Bed," 2021.
[59] "NTU Singapore and M1 Ink Partnership to Develop Singapore's First 5G C-V2X Research Testbed and Trials.," 2019.
[60] C. Wei, Z. Zhang, H. Liu, G. Wu, and M. J. Barth, "Hierarchical Multi-Modal Fusion for Roadside VRU Detection: Method Complementarity Under Sparse Label Constraints," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop*, 2025,
[61] S. Liu *et al.*, "Dab-detr: Dynamic anchor boxes are better queries for detr," *arXiv preprint arXiv:2201.12329,* 2022.
[62] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "Density-based spatial clustering of applications with noise," in *Int. Conf. knowledge discovery and data mining*, 1996, vol. 240, no. 6,
[63] Y. Chan, A. Hu, and J. Plant, "A Kalman filter based tracking scheme with input estimation," *IEEE transactions on Aerospace and Electronic Systems,* no. 2, pp. 237-244, 1979.
[64] I. J. Goodfellow *et al.*, "Generative adversarial nets," *Advances in neural information processing systems,* vol. 27, 2014.
[65] J. C. Hayward, "Near miss determination through use of a scale of danger," 1972.
[66] *U.S Department of Transportation Intersection Safety Challenge Stage 1B*,

[67] C. Wei, G. Wu, M. J. Barth, A. Abdelraouf, R. Gupta, and K. Han, "KI-GAN: Knowledge-Informed Generative Adversarial Networks for Enhanced Multi-Vehicle Trajectory Forecasting at Signalized Intersections," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7115-7124.

APPENDIX 1 LITERATURE REVIEW

1 Introduction

1.1 Motivation

California faces critical challenges in safeguarding vulnerable road users, including pedestrians, cyclists, and motorcyclists, as evidenced by concerning statistics. In 2020 alone, the state recorded 1,013 pedestrian fatalities, alongside 136 cyclist fatalities, according to data from the California Office of Traffic Safety (OTS) [1]. Moreover, the California Highway Patrol (CHP) reported 1,043 pedestrians, and 153 bicyclists were killed during the same period [2]. These figures underscore the urgent need to address factors contributing to road user vulnerability, such as distracted driving, and unsafe interactions at intersections. In 2021, there has been a rise in pedestrian fatalities. Pedestrian fatalities increased 9.4% from 1,013 in 2020 to 1,108 in 2021 [1]. With infrastructure challenges exacerbating risks, a comprehensive approach integrating infrastructure improvements, stringent enforcement measures, educational initiatives, and community engagement is imperative to mitigate fatalities and injuries among Vulnerable Road Users (VRUs) and enhance overall road safety across California.

According to the definition of the Federal Highway Administration (FHWA), a VRU could be a pedestrian, a bicyclist or e-cyclist, other conveyance like a scooter or skateboard, and highway worker on foot in a work zone, but does not include a motorcyclist [3]. To protect the VRUs, different sensors have been utilized to determine the real-time traffic conditions; the number of the circles represents the strength of the sensor in one aspect.

To develop a holistic understanding of the overall effectiveness of various sensors employed for VRU perception in transportation systems, TABLE 1 offers a condensed overview of those commonly utilized in traffic monitoring at intersections. Each type of sensor used alone possesses distinct capabilities and advantages/disadvantages relative to different scenarios and applications.

- *Monocular Camera*: High-resolution over a limited field of view (FOV); limited effectiveness for determining 3D position and velocity, particularly in congested traffic conditions.
- *Light Detection and Ranging* (*LiDAR*): High-precision 3D perception with robustness against environmental variations; however, its relatively elevated cost and sparse data representation present drawbacks.
- *Radio Detection and Ranging* (*RADAR*): Speed measurement enables functionalities such as stop bar and dilemma zone detection; however, its capability to differentiate between objects is limited.
- *Thermal Camera*: Thermal data enhances resistance to variations in lighting conditions.
- *Fisheye Camera*: Detection is enabled across a 360-degree field of view; however, ensuring accurate distortion correction necessitates the use of a highly precise calibration matrix.

Upon the data collection with different roadside sensors under real-time traffic conditions, the downstream modules data processing modules include detection, tracking, prediction, and communication. At the end, the processed warning information will be sent to the VRUs and the involved vehicles. Each module has different core functions, as shown in Figure 1. With concrete categories and trajectories of different road users (e.g., vehicles and pedestrians) being detected and tracked, their intentions and paths in the future can be predicted. If a potential conflict is predicted to happen in the future, a warning signal will be sent to the involved objects through the

specific communication protocol (e.g., cellular network, WiFi). Thus, along with this pipeline, the closely related papers regarding the roadside sensor-based technologies at the intersection have been reviewed. Associated challenges and opportunities are also discussed in this report, which may provide a clear guidance to the subsequent tasks on algorithm development.

TABLE 1 Performance of Different Sensors Utilized for Infrastructure-based Perception.

Capabilities	Monocular	LiDAR	RADAR	Thermal	Fisheye
Privacy-safe data	○	○○○	○○○	○○	○
Accurately detects and classifies objects	○○	○○○	○	○○	○○
Accurately measures object speed and position	○○	○○○	○○○	○○	○○
Extensive FOV	○	○○○	○○	○	○○○
Reliability across changes in lighting, sun, temperature	○○	○○○	○○○	○○	○○
Ability to read signs and differentiate color	○○○	○○	○	○	○○○
Cost for deployment and maintenance	○○○	○	○○	○○	○○

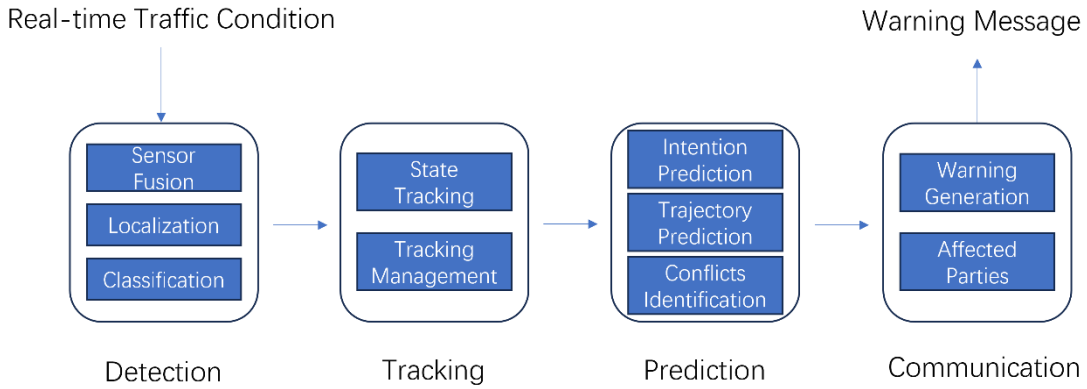

Figure 1. The Detailed Functions of Each Module in the System.

1.2 Organization

The rest of this report is organized as follows: The detection, tracking, prediction, and communication modules are introduced one by one, respectively. In Section 2, different detection approaches by sensor node multiplicity and sensor modality, as well as sensor fusion methods, and related roadside datasets are introduced. In Section 3, general Multi-object Tracking (MOT) methods and specific roadside MOT methods are reviewed. In Section 4, intention prediction and trajectory prediction approaches are analyzed in detail. In Section 5, the communication technologies and some applications related to collision avoidance in the intelligent transportation system are discussed thoroughly. At last, a conclusion of the whole report is made with analyzing of the technical pipelines and preparation for the next period.

2 Detection

Detection is a critical part of any VRU project, including the detection of both vehicles and VRUs. Vehicle detection has been developed for many years. With the progress of autonomous driving, an increasing number of algorithms and data sets have been proposed, and the vehicle recognition accuracy has become increasingly accurate. In comparison, VRU recognition has received less attention from scholars. Some general recognition algorithms include pedestrians and bicyclists in the categories of recognized objects, but their recognition rates are generally low. For example, VoxelNet's [4] detection performances in car, pedestrian and cyclist are 81.97%, 57.86% and 67.17% in average precision on easy KITTI [5] validation set. At the same time, due to the lack of specialized data sets, research focusing on VRU identification has not made systematic progress. The above is the status of object detection using vehicle mounted sensors, and there is even less relevant research for roadside mounted sensors, making the summary of relevant literature on VRU detection even more important.

Compared with vehicles, VRUs are generally smaller, have more inconsistent speed distribution, and are more dynamic. The inconsistent speed distribution means scooters, bicycles, and skateboards move rather fast; however, pedestrians are slower. Especially the elderly move much more slowly. Dynamic changes mean that VRUs do not necessarily travel strictly in prescribed areas or along smooth predictable trajectories. They may rush out of blind spots at any time or children may run across pedestrian areas. These conditions and the lack of number and variety of VRUs in general datasets all increase the difficulty of their detection.

In the following introduction regarding VRU detection, we first classify the existing literature into single-node and multi-node. Under multi-node studies, we further dive into the categories divided by modes and fusion schemes. Especially, the papers of detection mentioned are mainly focusing on the VRU detection, which align with the theme of the project.

2.1 Single-Node

The single-node detection methods mean there is only one sensor used for detection, like single camera and single LiDAR. In this report, we focus more on the LiDAR-based methods, because LiDAR can provide more accurate 3D information about the objects without distortion, and it works on bad weather and poor light condition. Also, with LiDAR, we can leverage the existing set-up devices we have to proceed with the project.

2.1.1 LiDAR-based Detection

LiDAR is a remote sensing technique employing pulsed laser light to measure distances at various angles relative to the sensor (i.e., spherical coordinates of detections in the Lidar frame of reference). LiDAR can offer detailed 3D point cloud data to accurately determine the 3D positions of traffic objects [6]. The detection methods proposed using LiDAR are categorized into traditional methods and deep learning methods.

2.1.1.1 Traditional Methods

Traditional methods often use some statistical techniques to process the point cloud data. At the early stage, researchers use manually selected features of point clouds as input to classification models, which belong to non-deep learning methods. Limited by the price and synchronization issues related to using multiple LiDAR sensors, many researchers start with one LiDAR to detect road users at the intersection. A skateboarder detection method using a 16 channel LiDAR was proposed by Wu et al. [7]. They utilized 3D Density Statistic Filtering (3D-DSF) to filter the background, and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) to cluster the points. With a two-step classification method, the clustered object is first classified as a vehicle or non-vehicle object. If it is a non-vehicle object, the second step will be deployed with the object's dynamic features including the average speed in its historical trajectory and the standard deviation of the speed. Thus, the non-vehicle objects were classified as skateboarders, cyclists, and pedestrians. Because of the similar speed of skateboarders and cyclists, these two may be misclassified. Classical machine learning techniques such as K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest (RF), and Naïve Bayes (NB) are compared in the two-step classification tasks. RF performed best in both steps under their own datasets. To further optimize the detection results of pedestrians using the same LiDAR, researchers developed efficient detecting methods according to different distances [8]. The detecting area was divided into three sub-areas in circles according to the distance to the LiDAR, and DBSCAN was deployed with different radii and minimum numbers of points in three subareas, as depicted in
Figure 2. Taking the number of points in a cluster, 2D distance to LiDAR and direction of the clustered points' distribution as object features, a Back Propagation Artificial Neural Network (BPANN) was used to classify the vehicles and pedestrians. Based on previous research, Song et al. [9] used 3D-DSF and DBSCAN to filter the background and cluster the points, respectively. Then, the overall performance of SVM, RF, BPANN and Probabilistic Neural Network (PNN) were compared, using measurements of a 16-channel LiDAR as inputs. SVM performed best with the shortest calculation time, benefiting from its better nonlinear regression ability of the Gaussian kernel. Considering the detection methods using only one LiDAR with a single view, object occlusion is a common issue that researchers encountered.

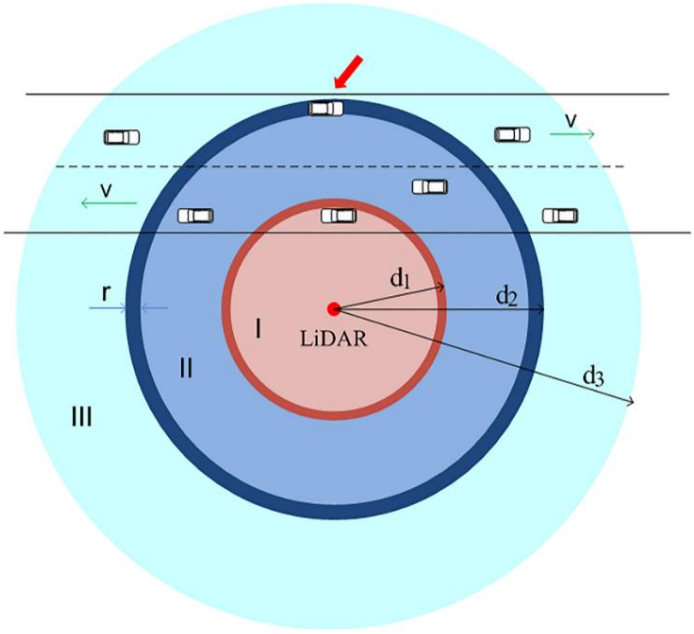

Figure 2. DBSCAN Area Division.[1]

2.1.1.2 Deep Learning Methods

With the rapid development of Deep Learning (DL), some researchers tried to utilize DL in roadside VRU detection because of its strong regression and classification abilities. At the early stage, researchers just replaced the final classification step with DL. In [10], after background filtering, clustering and feature extraction, a lightweight Visual Geometry Group Network (VGGNet) was used to classify the objects with hand-crafted features as inputs. To enhance the correlation between the features and the classification target, a self-attention mechanism was added to the model. The geometric information, statistical information, and Viewpoint Feature Histogram (VFH) of the LiDAR point clouds were integrated as features to fully explore the object representation. Thus, with the self-attention VGGNet and abundant features, the model surpassed the RF by 14.4% and 11.9% in pedestrians and cyclists, and SVM by10.4% and 14.3% in pedestrians and cyclists. Similarly, Zhang et al. [11] proposed a Global Context Network (GCNet) prototype with three stages, including gridding, clustering, and classification. With the transformation of the coordinates, the point clouds were transformed to a grid format in a plane, which is called gridding [11]. So, a CNN can be used to classify the clustered objects, like image classification. In recent years, researchers prefer to replace the clustering, feature extraction and classification steps directly with an end-to-end DL network, whose inputs and outputs are features and classification results, respectively. Feature Enhancement Network (FecNet) [12] is a feature-enhancement cascade network for object detection with roadside LiDAR. Compared with conventional detection methods, FecNet emphasized the importance of the foreground feature to reduce the similarity between the objects and the background. By adding the foreground feature into the feature map and reweighting the feature map with foreground attention branch, the representation of the foreground objects can be enhanced. Moreover, different level features were

[1] Zhao, Junxuan, et al. "Detection and tracking of pedestrians and vehicles using roadside LiDAR sensors." *Transportation research part C: emerging technologies* 100 (2019): 68-87.

fused to strengthen multi-scale objects detection ability. Based on their own datasets, FecNet outperformed in mean Average Precision (mAP) with 86.32% in their own dataset and had a good trade-off in accuracy and efficiency with the smallest model size, compared with PointPillars [13], SECOND [14], PointRCNN [15], PV-RCNN [16], PV-RCNN++ [17], and PillarNet [18]. Nasrollahi et al. [19] analyzed some milestones of LiDAR detection proposed for the on-board side, and compared their performance in roadside surveillance, including SECOND, CenterPoint with voxel encoding [20], CenterPoint with pillar encoding, and PillarNet. Compared with anchor-based methods, center-based methods exhibited good performance with low computational loads. With comparison among these methods, CenterPoint using voxel representation with specific hyperparameters, which are different from the on-board setups, can achieve the optimal performance among these selected methods for pedestrian detection within the realm of video surveillance. Shi et al. [21] combined the voxel-based encoder, center-based proposals, and the transformer-based detector with innovation. Compared with the former proposed methods, this method utilized DAIR-V2X-I [22], an open-source roadside detection dataset, and transferred the data into KITTI [5] format to compare the performance with other benchmarks. As a result, it performs better than all the listed models in cyclist detection, including PointPillars, SECOND, Object DGCNN [23], FSD [24] and MVX-Net [25].

2.1.2 Camera-based Detection

The camera is the most common sensor in the detection of VRU, because of its low cost. Also, the camera-based detection method is one of the earliest and most mature algorithm types. Two decades ago, Viola and Jones [26] introduced a method for real-time detection of human faces that surpassed all contemporaneous algorithms in both real-time performance and detection accuracy, without imposing any constraints. Dalal and Triggs [27] presented the Histogram of Oriented Gradients (HOG) feature descriptor, which significantly improved upon the scale-invariant feature transform and shapes context. This detector has become a cornerstone in various object detection systems and has found extensive use in practical applications. For example, HOG was used on pedestrian detection, and the miss rate was under 0.01 at 10^{-6} False Positives Per Window (FPPW)on the original MIT pedestrian database [27, 28].

With the development of DL, convolutional neural networks (CNNs) [29] began to be widespread. In addition, Girshick et al. [30] introduced Regions with CNN features (R-CNN) for object detection, marking a significant breakthrough in deep learning. Concurrently, He et al. [31] developed Spatial Pyramid Pooling Networks (SPPNet), which achieved feature representation independent of image size and operated 20 times faster than R-CNN while maintaining accuracy. Later, many additional fast and accurate 2D image detection algorithms were proposed. Fast R-CNN [32] is over 200 times faster than R-CNN. Faster R-CNN [32] is the first end-to-end and near-realtime DL detector. You Only Look Once (YOLO) [33] is the first one-stage DL detector with extremely fast speed. Until now, it has been updated to YOLO v9 [35]. Single Shot MultiBox Detector (SSD) [36] is also a one-stage detector, but with significantly improved accuracy. Feature Pyramid Networks (FPN) [37], a development based on Faster R-CNN, is also a crucial component in numerous object perception models. In recent years, Transformers [38] incorporating attention mechanisms have emerged as the predominant trend in most object perception tasks.

Regarding the VRU detection, some researchers utilized the proposed methods, fine-tuned them, and trained with different datasets. Garcia-Venegas et al. [39] compared the performance of Faster R-CNN, SSD and Region-based Fully Convolutional Network (R-FCN) with different feature

extractors in cyclist detection. As a result, the SSD was chosen as the detector because of its high Frames Per Second (FPS) for real-time implementation. Besides, this work utilized the Kalman filter [40] to enhance the robustness against false negative detections and object occlusions. Another work also used Faster R-CNN employing transfer learning from diverse pre-trained architectures to detect the pedestrians with infrastructure-based cameras [41]. Mammeri et al. [42] explored the application of CNN-based algorithms for identifying VRUs like pedestrians, motorcyclists, and cyclists from roadside. The detection performance of the large-scale object was better than small-scale object. For example, the detection performance of small objects using the YOLOv3 and FasterRCNN was worse by about 5% in mAP, compared with the detection performance of large objects. Risto et al. [43] built an ITS detection system that aimed to measure the distance of the object accurately, including the pedestrians and cyclists, which can help locate the VRU. With experiments, the system can measure the distance accurately in the range of 6m to 22.5m. According to the sensitivity analysis, the error in vertical α-angles of the tilted camera would significantly affect the accuracy measured distance of the objects, if we assumed the camera was in the original angle.

In recent years, the roadside camera driven detection methods prefer utilizing the Bird's Eye View (BEV) feature space because it contains rich spatial information and is in the global view. For example, BEVDepth [44] addressed the inadequate estimation of visual depth information by incorporating explicit depth supervision, a camera-awareness depth estimation module, and a depth refinement module. Then, BEVHeight [45], a novel approach addressing the limitations of vision-centric BEV detection methods on roadside cameras by regressing the height to the ground rather than predicting pixel-wise depth, enabling distance-agnostic formulation to enhance optimization processes. To reduce the step of extrinsic calibration, the Calibration-free BEV Representation (CBR) [46] network enabled 3D object detection based on BEV representation without requiring calibration parameters or additional depth supervision, but at the cost of 8.7% $AP_{3D|R40}$ decreasing with Intersection over Union being 0.5, for moderate difficulty in the DAIR-V2X-I dataset, compared with BEVHeight. A novel framework called Complementary-BEV (CoBEV) [47] integrated depth and height information, utilizing a two-stage Complementary Feature Selection (CFS) module and a BEV feature distillation framework, demonstrated superior accuracy and robustness across various datasets and challenging scenarios like long-distance and noisy datasets. Until now, CoBEV is the SOTA algorithm in the DAIR-V2X-I dataset.

In camera-only detection, there are some common challenges such as the illumination problem, appearance variation, complex background, shadow, and occlusion. Some of them may be solved with various algorithms to a certain degree. However, the physical limitation of the camera sensor will affect the upper limit of recognition accuracy, especially the occlusion and illumination problem. Thus, sensor fusion is a good approach to combine the advantages of different sensors.

2.2 Multi-Node

Unlike single-node, multi-node is using several sensors for detection at the same time, which is usually named sensor fusion. There are two kinds of classification methods for sensor fusion. One is classified by mode, which includes single-mode and multi-mode, corresponding to the type of sensors in sensor fusion. The other one is classified by fusion scheme, including early fusion, intermediate fusion, and late fusion.

2.2.1 Classification by Mode

With the development of perception research domain, multiple sensors have been utilized to detect the VRUs in traffic systems, such as camera, radar, LiDAR, thermal camera and so on. At the beginning, some researchers utilized the combination of three low-cost sensors at roadside to perceive if there is a pedestrian or vehicle passing [48]. With the cost of sensors going down, increasing numbers of advanced accurate sensors are being utilized at intersections. Different sensors have different advantages in detecting VRUs. Cameras can provide texture and color, but its detection results are significantly affected by the weather and illumination. Radar is suitable for long-range detection and operates effectively in all weather conditions but has low resolution. LiDARs can provide depth information under various environmental conditions, but with a lower density of detail than camera imagery. Thermal cameras can be effectively used to detect VRUs, but can be affected by temperature. Thus, multiple sensors should be combined to detect objects cooperatively, which will fully tap the potentials of different sensors. The detection methods based on sensor modality can be classified into single-mode and multi-mode. For single-mode methods, we focus on LiDAR only and camera only, while for multi-mode approaches, combining camera with RADAR or combining camera with LiDAR are prevalent setups.

2.2.1.1 Single-Mode

LiDAR Only

Meissner et al. [49] established a real-time pedestrian detection network of 8 multilayer laser scanners in the KoPER project [50]. A real-time adaptive foreground extractor was proposed using the Mixture of Gaussian (MoG) [51] and then an improved DBSCAN [49] was achieved by projecting the points to a 2D plane and dividing the clustering area with cells to reduce computation. The features utilized for classification were height, width, and clustering direction to z-axis. Limited by the fixed range and the required minimum number of clusters in DBSCAN, when it comes to two pedestrians being very close, it is very challenging to separate them. In InfraDet3D [52], different fusion methods of two LiDARs were proposed as the baseline methods, including early fusion and late fusion. Early fusion of LiDAR-only setup is to merge numerous point cloud scans generated by various LiDAR sensors at time step t to create a unified and densely populated point cloud, which used *Fast Global Registration* [53] to register the merged point cloud in the initial transformation and point-to-point Iterative Closest Point (ICP) [54] for the refinement. Also, a late fusion baseline method of LiDAR was also talked about in the paper. In this method, the detections from two LiDARs are transformed to the same reference frame and matched within a threshold distance. The merging process involved combining matched detections by choosing the central position and yaw vector of the identified object from the nearest LiDAR sensor. The dimensions of the merged detections were determined by averaging the measurements from both detectors.

Camera Only

Multiple cameras can provide wider FOV and more accurate results. In InfraDet3D [34], the fusion method of the detection results of two roadside cameras was proposed as one of the baseline methods. Among all kinds of cameras, there is a special event-based camera, which detects and reports individual pixel-level brightness changes (events) asynchronously and independently of each other [55]. So, it is used to eliminate blur of the motion and detect objects at night. However,

it lacks the color and texture information, which is just the advantage of RGB cameras. Thus, the fusion of data from event-based cameras and RGB cameras is proposed to detect objects, which can be utilized for roadside vehicle detection, especially at night. Another kind of camera combination is RGB camera and infrared camera, which can help improve detection performance even with poor visibility under adverse weather conditions or in the nighttime [56].

2.2.1.2 Multi-Mode

Regarding multi-mode, there may be several sensors combined for detection, not limited to camera and RADAR, and camera and LiDAR; nevertheless, we focus on these two common combinations in this report for VRU detection.

Camera and RADAR

Bai et al. [57] developed a fusing method to utilize the advantages of the camera and the RADAR. The detection results of RADAR were the distance and velocity of the target by measuring the time delay and phase shift of the echo signal. The camera's detection results were the angle and classification of the target, detected by YOLO V4. After calibrating the coordinates of each sensor, the measurement results of the two sensors are fused by using the nearest neighbor correlation method. Liu et al. [58] proposed methods for single-scan fusion and multi-scan fusion using Dempster's combination rule to improve evidence fusion performance. The pre-calibration and the nearest neighbor correlation were also used to associate the detections of camera and RADAR. In addition, the authors fused the sensors' detection results on multiple scans by associating the detections of adjacent scans to enhance the performance. With camera and radar fusion, the challenging scenes captured under poor light conditions can be detected accurately.

Camera and LiDAR

The advantage of using a monocular camera is a better detection of small objects such as pedestrians. A LiDAR detector can detect objects during nighttime and the point cloud can provide more accurate information without distortion. Therefore, combining LiDAR and camera can significantly enhance the overall VRU detection performance. Zhao et al. [59] set up a roadside detection network with 6 single-row laser scanners, and a video camera. Laser scanners were used for detection and the camera was used for validation of the detection. InfraDet3D [52] provided a camera-LiDAR fusion method, where the LiDAR detection results were first transformed to the camera coordinate, and then the detections of two sensors were synchronized based on the distance of the center points of the detected objects. In this method, the detection of the camera is a supplement for LiDAR detection. MVX-Net included two effective fusion methods, PointFusion and VoxelFusion, utilizing the famous VoxelNet [4] architecture, as shown in Figure 3. PointFusion projected the points onto the image plane and got the corresponding 2D features in the same location. Thus, the pointwise concatenated vector of 3D point feature and corresponding 2D semantic feature was bounded with VoxelNet. Similarly, VoxelFusion projected the divided voxels onto the image to get the corresponding Region of Interest (ROI). Then, the pooled ROI features were added to the voxel features, which was processed by 3D Region Proposal Networks (3D RPNs) for detection results. As a result, MVX-Net provided more accurate detection and had the advantage of detecting far objects with low LiDAR resolution.

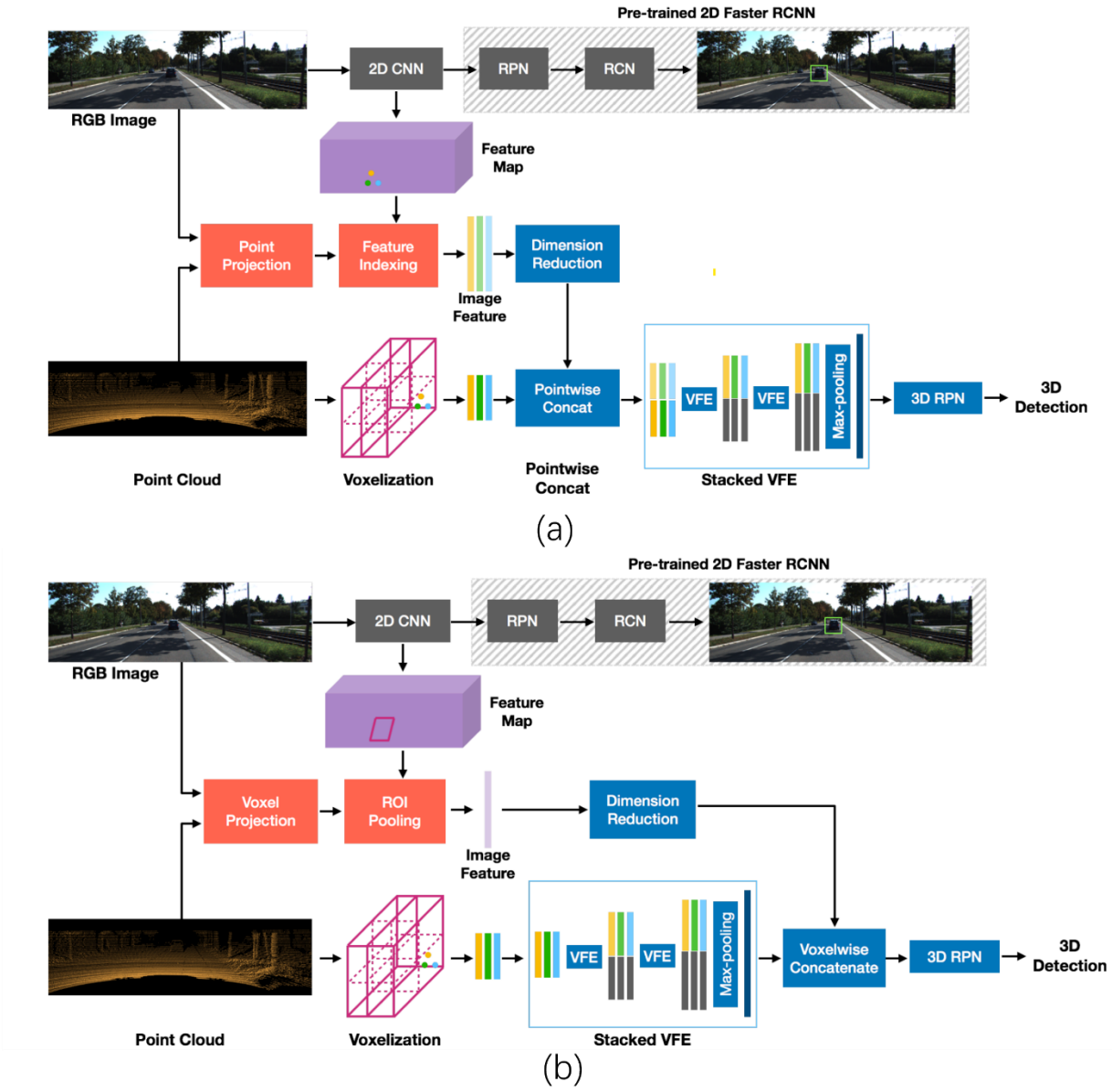

Figure 3. (a) An Overview of PointFusion. (b) An Overview of VoxelFusion.[2]

2.2.2 Classification by Fusion Scheme

Regarding the phase of sensor fusion [60], which is depicted in Figure 4, a multi-sensor perception system can be categorized into three main classes: 1) Early Fusion (EF) - involves the fusion of raw data during the preprocessing phase. 2) Immediate Fusion (IF) - entails the fusion of features during the feature extraction stage. 3) Late Fusion (LF) - involves the fusion of perception outcomes during the post-processing phase. The EF approach possesses its own set of advantages and drawbacks from various angles. For instance, EF and IF typically yield higher accuracy levels but demand greater computational resources and intricate model designs. Conversely, LF tends to

[2] Sindagi, Vishwanath A., Yin Zhou, and Oncel Tuzel. "Mvx-net: Multimodal voxelnet for 3d object detection." *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019.

offer better real-time performance but may compromise on accuracy. The optimal deployment of fusion schemes hinges on the specific requirements across different traffic scenarios.

2.2.2.1 Early Fusion

Sharing the raw sensor data with other sensors to extend the perceptual range or point cloud density (multiple LiDARs) and enhance detection accuracy is a logical approach. In line with this strategy, the raw data collected from various sensors is transformed into a unified coordinate system to facilitate subsequent processing [61]. However, this kind of strategy is inevitably sensitive to some issues such as sensor calibration, data synchronization and communication capacity, which are critical considerations in real-time implementations [62]. For example, with a limited communication bandwidth, it may be viable to transmit limited resolution image data, but it might not be practical to transmit LiDAR point cloud data in real time. For instance, a 64-beam Velodyne LiDAR operating at 10Hz can produce approximately 20MB of data per second [5].

2.2.2.2 Intermediate Fusion

The fundamental principle behind IF can be succinctly described as employing deeply extracted features for fusion at intermediate stages within the perception pipeline. IF relies on hidden features primarily derived from deep neural networks, which exhibit greater robustness compared to the raw sensor data utilized in early fusion. Moreover, feature-based fusion approaches typically employ a single detector to produce object perception outcomes, thereby eliminating the necessity of consolidating multiple proposals as required in LF. For example, MVX-Net is utilizing the IF strategy to fuse features from 2D images and 3D points or voxels, and it is processed by the VoxelNet. Bai et al. also proposed the PillarGrid to fuse the extracted features of two LiDARs to improve the detection accuracy [63].

2.2.2.3 Late Fusion

Late fusion also adopts a natural cooperative paradigm for perception, wherein perception outcomes are independently generated and subsequently fused together. Unlike EF, although LF still requires a relative positioning for merging perception results, its robustness to calibration errors and asynchronization issues is significantly improved. This is primarily because object-level fusion can be established based on spatial and temporal constraints. Rauch et al. [64] employed Extended Kalman Filter (EKF) to coherently align the shared bounding box proposals, leveraging spatiotemporal constraints. Moreover, techniques such as Non-Maximum Suppression (NMS) [65] and other machine learning-driven proposal refining methods find extensive applications in LF methodologies for object perception [66].

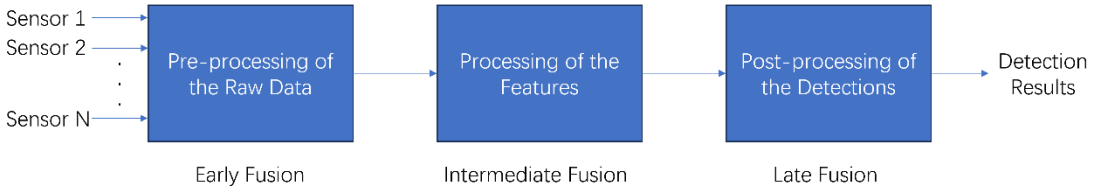

Figure 4. The Fusion Scheme.

Thus, different fusion scheme has its own advantages and disadvantages. Early fusion and late fusion are both straightforward, but early fusion requires accurate calibration and sufficient

communication bandwidth, and late fusion may result in limited detection accuracy. Deep fusion is a balanced and well-performing method because of its low communication requirements and high detection accuracy, which may be a suitable method for this project.

2.3 Datasets

. About the roadside detection datasets, we have reviewed the existing related datasets, including the infrastructure-based dataset and vehicle-infrastructure cooperative dataset.

Because of the lack of public roadside datasets, researchers choose to utilize some existing autonomous driving datasets to evaluate their algorithms, because these datasets are well-labelled and widely accepted. In [19], the authors listed some datasets developed for autonomous driving such as KITTI [5], Waymo [67], and nuScenes [68], and analyzed which one is more suitable for the LiDAR-assisted pedestrian detection for video surveillance. The nuScenes dataset was chosen because of its abundant annotations, like bounding boxes, orientations, and velocities of pedestrians. Also, it has relatively low spatial resolution and high temporal resolution, which aligns with the requirements of setup in real-world situations. By integrating continuous sweeps of nuScenes, the density of the LiDAR points can be increased, which means the range of the distances to the sensor is changeable and controllable for specific detection task, e.g., VRU detection at intersections.

2.3.1 Infrastructure-based Dataset

Real roadside datasets are summarized in Table 2. In chronological order, KoPER [50] is one of the earliest roadside datasets collected with 14 8-layer laser scanners and 2 monochrome Charged-coupled Device (CCD) cameras. The sensors were installed at a height of over 5m, and they were tuned to reduce occlusion. The VRUs in this dataset are pedestrians and cyclists. BAAI-VANJEE [69] is a dataset collected with a 32-line LiDAR and 4 cameras mounted at 4.5m, but the data from different sensors was partially synchronized. The VRUs comprise pedestrian, bicycle, and tricycle. DAIR-V2X-I [22] is a roadside dataset collected across several intersections. At each intersection, at least a pair of cameras and LiDAR was installed. The data from different sensors were fully synchronized and the VRU objects involved cyclists, tricyclists, barrows, and pedestrians. Zimmer et al. proposed the TUMTraf Intersection Dataset, which contained pedestrian and bicycle as VRUs, and complex driving maneuvers, e.g., left/right turns, overtaking and U-turns [70].

TABLE 2 Infrastructure-based Datasets.

Datasets	Sensors (number and type)	Synchronized Data	Object Class	Content
Ko-PER [50]	● 14x SICK LD-MRS 8-layer research laser scanners (12.5Hz) ● 2x Baumer TXG-04 monochrome CCD cameras (25Hz)	Yes	● Vehicles ● Trucks ● Buses ● **Pedestrians** ● **Cyclists**	-

BAAI-VANJEE [69]	• 1x VANJEE 32-layer LiDAR • 4x cameras	20% synchronized	• **Pedestrian** • **Bicycle** • **Tricycle** • Motorcycle • Car • Van • Cargo • Truck • Bus • Semi-trailing tractor • Special vehicle(police car, ambulance, fire truck, tanker truck, etc.) • Roadblock	• 2500 frames of LiDAR data, 5000 frames of RGB images • 74K 3D object annotations and 105K 2D object annotations
DAIR-V2X-I [22]	• 1x 300-layer LiDAR (10Hz) • 1x camera (25Hz)	Yes	• Car • Truck • Van • Bus • **Pedestrian** • **Cyclist** • **Tricyclist** • Motorcyclist • **Barrow** • Traffic Cone	• 10,084 frames of LiDAR data, 10,084 frames of RGB images • 493K 3D object annotations
TUMTraf Intersection Dataset [70]	• 2x Ouster OS1-64 (gen. 2) 64-layer LiDAR • 2x Basler ace acA1920-50gc cameras	Yes	• Car • Truck • Trailer • Van • Motorcycle • Bus • **Pedestrian** • **Bicycle** • Emergency Vehicle • Other	• 4,800 frames of LiDAR data, 4,800 frames of RGB images • 57.4k 3D object annotations with track IDs

2.3.2 Cooperative Dataset

Cooperative dataset contains data both from vehicles and infrastructure, as summarized in Table 3. These datasets benefit cooperative perception research. A widely used dataset is Rope3D [71]. It uses one LiDAR on a vehicle and several cameras on the roadside. The data are synchronized. To fully explore the optimal setup, the data with different mounting heights, pitch angles of the cameras, and focal length were collected. The VRU classes include cyclist, tricyclist, barrow, and pedestrian. However, there is only camera data for roadside object detection, which may not be suitable for this project. Further, Yu et al. [22] released a cooperative version dataset, named DAIR-V2X-C. Different from Rope3D, there were both LiDAR and camera data on one side, which can enhance the perception performance. All the data were synchronized, and cooperative

tracking annotations were provided. Similarly, a research team in Germany proposed a cooperative dataset, but the VRUs in the dataset only included pedestrians and cyclists [72]. The latest vehicle-to-everything (V2X) cooperative detection dataset is V2X-Real [73]. Based on the ego agent type, the authors presented four kinds of datasets, including vehicle-centric dataset, infrastructure-centric dataset, vehicle-to-vehicle (V2V) dataset and infrastructure-to-infrastructure (I2I) dataset, which were suitable for different research tasks. This dataset contained the most abundant frames and annotations among the infrastructure datasets, and it had the scooter category which had not been recorded by other datasets. Another highlight was the number of pedestrians was more than the number of vehicles, which made the dataset suitable for VRU detection.

Also, there are some simulated datasets. A typical one is V2X-SIM [74]. It simulated with all kinds of cameras and LiDARs on both sides, including the RGB camera, the BEV semantic segmentation camera, the LiDAR, and the semantic LiDAR. All data were synchronized, but the main issue was that it only contains one kind of VRUs, i.e., pedestrians.

TABLE 3 Cooperative Dataset.

Datasets	Sensors (number and type)	Synchronized Data	Object Class	Content
Rope3D [71]	Infrastructure: ● Cameras in different scenes (30-60Hz) Vehicle: ● 1x LiDAR (1) HESAI Pandar 40-layer (10/20 Hz) or (2) Jaguar Prime from Innovusion 300-layer	Yes	● Car ● Van ● Bus ● Truck ● Motorcyclist ● **Cyclist** ● **Tricyclist** ● **Barrow** ● **Pedestrian** ● Traffic cones ● Triangle plate ● Unknown-unmovable ● Unknown-movable	● 50k frames of RGB images ● 1.5M 3D object annotations and 670k 2D object annotations for objects not scanned by the laser Note: Only Images Data for Roadside 3D Object Detection
DAIR-V2X-C (Cooperative) [22]	Infrastructure: ● 1x 300-layer LiDAR (10Hz) ● 1x camera (25Hz) Vehicle: ● 1x 40p LiDAR (10Hz) ● 1x camera (20Hz)	Yes	● Car ● Truck ● Van ● Bus ● **Pedestrian** ● **Cyclist** ● **Tricyclist** ● Motorcyclist ● **Barrow** ● Traffic Cone	● 38,845 frames of LiDAR data, 38,845 frames of RGB images ● 464K 3D object annotations
V2X-SIM (Carla Simulation) [74]	Infrastructure: ● 4x RGB cameras ● 1x BEV semantic segmentation camera ● 1x LiDAR	Yes	● Vehicle ● Sidewalk ● Terrain Road ● Building ● **Pedestrian** ● Vegetation	● 10,000 frames of data ● 47,200 3D object annotations

	● 1x semantic LiDAR Vehicle: ● 6x RGB/depth/semantic segmentation cameras ● 1x BEV semantic segmentation camera ● 1x LiDAR ● 1x semantic LiDAR sensor			
TUMTraf V2X Cooperative Perception Dataset [72]	Infrastructure: ● 1x Ouster LiDAR OS1-64 (generation 2), 64-layer ● 4x Basler ace acA1920-50gc Vehicle: ● 1x Robosense RS-LiDAR-32, 32-layer ● 1x Basler ace acA1920-50gc ● 1x Emlid Reach RS2+ multi-band RTK GNSS receiver ● 1x XSENS MTi-30-2A8G4 IMU	Yes	● Car ● Truck ● Trailer ● Van ● Motorcycle ● Bus ● **Pedestrian** ● **Bicycle**	● 2,000 frames of LiDAR data, 5,000 frames of RGB images ● 30k 3D bounding boxes with track IDs
V2X-Real [73]	Infrastructure: ● LiDAR x1: Ouster OS1-128/64 (30Hz, 40m range). ● Camera x2: Axis-P14555 (30Hz). ● GPS x1: Garmin 18x LVC (5Hz) Vehicle: ● LiDAR x1: RoboSense Ruby Plus (10Hz, 200m range). ● Camera x4: ZED 2i (10Hz). ● GPS/IMU x1: Gongji GPS System (1000Hz).	Yes	● Car ● Pedestrian ● Scooter ● Motorcycle ● Bicycle ● Truck ● Van ● Barrier ● Box truck ● Bus	● 33K frames of LiDAR data, 171K frames of RGB images ● 1.2M 3D object annotations

In addition, the difference between infrastructure-based datasets and vehicle-based datasets may have some inspiration for roadside detection algorithm development. According to Rope3D, the roadside perceptual datasets can be different in three aspects, compared with the on-board datasets. Firstly, when data is gathered from sensors with different configurations, ambiguity arises because of variations in camera specifications, including different pitch angles of the viewpoint, mounting heights, and diverse roadside environments. Secondly, the assumption that the camera's optical axis is parallel to the ground becomes invalid, leading to existing detection methods being incompatible for direct application. Lastly, a greater quantity of smaller objects are anticipated due to the roadside sensor configuration, which may increase the difficulty of detection because of less information about each object on average. Thus, when we consider about the algorithm selection, it is important to note the difference between the onboard and the roadside application environments.

3 Tracking

The objective of tracking is to associate the detection results across multiple frames, ending with continuous dynamic trajectories for all objects. In this report, we only consider Multi-object Tracking (MOT) methods, because there are usually multiple road users in one frame. MOT can be categorized into two main types: offline method and online method. Offline algorithms analyze future frames to determine object identities in a current frame, benefiting from global information and generally yielding higher tracking accuracy. Conversely, online methods rely solely on present and past data to predict object positions in the current frame and tend to be less accurate compared to batch (i.e., offline) methods due to their inability to rectify past mistakes using future data. In the road user tracking scenario, the tracking method needs to be online and to achieve high accuracy.

To make the tracking process clearer, we first give a comprehensive overview of general multi-object tracking methods and then provide the review of some tracking methods that are highly related to this project. The general goal of MOT is to associate detections in a sequence. There are usually four steps in one pipeline, including detection, feature extraction/motion prediction, similarity/distance comparison, and association.

- Detection: detection multiple objects in each frame from the new sensor data, with the target classes, bounding boxes, and locations.
- Feature extraction/motion prediction: for all objects detected in previous time intervals, predict the state of such tracked objects at the current time; prediction includes feature appearance, motion, and interactions.
- Similarity/distance comparison: features of the objects in different frames are used for similarity comparison; the predicted state and appearance of tracked objects is compared by distance with all objects detected in the current data.
- Association: based on the comparison results, the newly detected objects are associated to those of tracked objects for state estimation.

MOT contains 2D and 3D tracking methods based on different detection sensors. 2D tracking methods refers to the visual tracking methods. At the early stage, the hand-crafted visual representation of the objects is utilized to measure the similarity, such as spatial distance and color histogram [75]. Also, there are image-based methods using the template matching [76, 77]. Template matching is comparing a template image, which represents the object of interest, with similar sub-regions in a larger image. These traditional appearance matching methods are relatively simple and computationally efficient, making it suitable for real-time applications. However, its effectiveness may be limited. Modern methods include motion model and appearance feature. There are several types of motion models like Constant Velocity (CV) model, Constant Acceleration (CA) model, Kalman filter model, particle filter model [78], dynamic Bayesian network model [79], and deep learning model, e.g., LSTM [80]. As for the appearance feature, it is extracted by representation learning models. Some popular tracking methods such as DeepSORT [81], ByteTrack [82] and QDTrack [83] all fall into this category.

However, 2D MOT can be limited by factors such as variations in lighting, background clutter, occlusions, and deformations of the object being tracked. The 3D MOT pipeline is similar to that for 2D MOT. The difference lies in the dimension of the detection results. In 3D space, the motion

and appearance information can be obtained without distortion. With more data including depth information, the tracking in 3D space can be more accurate. In this project, we mainly consider the 3D MOT methods. Regarding the roadside tracking methods, Zhao et al. [59] utilized the nearest neighbors to associate the objects with clusters detected by laser scanners. The nearest cluster in frame *k* to the predicted state of the former trajectory is used to update the tracking results. Similarly, many other researchers also used the classical distanced-based object association method and Kalman filter for fast real-time roadside tracking [7, 8, 84]. However, there are still some problems like occlusion and accuracy of long-distance detection, which should be a concern. Except for that, there are also some minor approaches used in roadside MOT tracking. Meissner et al. [49] used the detection results from multiple laser scanners for tracking. With clustering results, they further utilized the Random Finite Set (RFS) filter with Sequential Monte Carlo (SMC) methods to get the estimated tracks [85, 86]. Bai et al. [87] improved the DeepSORT method to a 3D version, by leveraging the 2D tracking results first and then matching them with the 2D detection results. The matched 2D detection results corresponded to the same 3D detection results, which means the 3D tracking results were obtained by reidentifying them from 2D tracking results.

In addition, there are many on-board 3D MOT methods which are suitable for real-time tracking. AB3DMOT [88] used Linear Kalman filter and 3D Intersection Over Union (IOU) to build an advanced and fast 3D MOT system. The pipeline contained obtaining detection results, predicting the current frame states, implementing data association to match the trajectories and detections, updating the states, and managing the birth and death of the trajectories. SimpleTrack [89] followed this pipeline and improved the performance. PolyMOT [90], which is the state-of-the-art (SOTA) method, also used this pipeline, but it improved the accuracy by describing the motions of the objects using different models according to their categories. However, when we consider about the on-board 3D MOT methods, we should ensure the coordinates of the system are aligned. This is critical if the method is applied to roadside sensing scenarios.

4 Prediction

VRU state prediction, especially pedestrian state prediction, including intention prediction and trajectory prediction, is a hot research domain. Many researchers have proposed pioneering and promising prediction models in this area. Most of the models are developed based on the on-board environment, which is not aligned with our purpose. However, from the perspective of prediction, the only difference between the on-board environment and the roadside environment is the frame of reference of the coordinates. Because the inputs of the prediction models are only the results of detection and tracking process, but not the raw data, the on-board trajectory prediction models can also be used in the roadside environment. The disparities in potential distribution of point cloud data can be ignored.

In general, there are several aspects that should be considered when selecting the prediction model, including the model's run time/efficiency/complexity, robustness, and accuracy. Simple models include CV model [91] and Social Force (SF) model [92, 93], which are physical models without training process. Complex models include the basic models, variants, and combinations of Recurrent Neural Network (RNN) [94], Long Short-Term Memory (LSTM), Gate Recurrent Unit (GRU) [95], Attention [68], Graph Convolutional Network (GCN) [96, 97] and Transformer [98]. Some models also combine the SF model as a part, like Social LSTM [99] and Social Attention [100]. Regarding the robustness, it is a very important property of the model when the objects are obscured by other objects or unseen from the detection device. SimAug [102] focused on the prediction of the pedestrian in unseen scenarios and camera views. With respect to accuracy, the most accurate model on ETH/UCY benchmark is AMEND [103], which constructs a Mixture of Experts framework to solve the long-tail prediction problem, which is caused by the imbalance of existing datasets. The datasets contain more simple tasks and lack challenge scenarios.

Regarding the intention prediction for VRUs, we usually refer to the crossing/non-crossing decision in the next time step, which is normally predicted with the historical trajectories, position, velocity, and distance between the VRU and other road users. Xu et al. [101] utilized Bayesian method to recognize the behavior and intention, and then used a modified particle filter to predict the motion of pedestrian. Goldhammer et al. [104] defined four motion states, including waiting, starting, moving, and stopping as intentions. They used past positions as inputs and utilized PolyMLP to predict the intentions and trajectories at the same time. Zhao et al. [105] proposed a Deep Autoencoder-Artificial Neural Network (DA-ANN). With XYZ position, velocity, direction, and distance to the crosswalk encoded, the motion patterns of crossing and non-crossing behaviors were recognized by DA-ANN and the intention was predicted. The same team also proposed another approach with modified Naïve Bayes to predict the intention [106]. Compared with the encoded features in the former method, they added the location, total number of data points, distance to LiDAR, tracking ID, frame number, velocity, direction, timestamp, and pedestrian/vehicle label of each road user as inputs, and the output included the intention label and its probability. Recently, Chen et al. [92] took a multi-feature sequence as inputs, including gender type, age type, distance to the potential collision point, and the conflict vehicle type. With the features encoded by GRU, the crossing/waiting intention would be predicted.

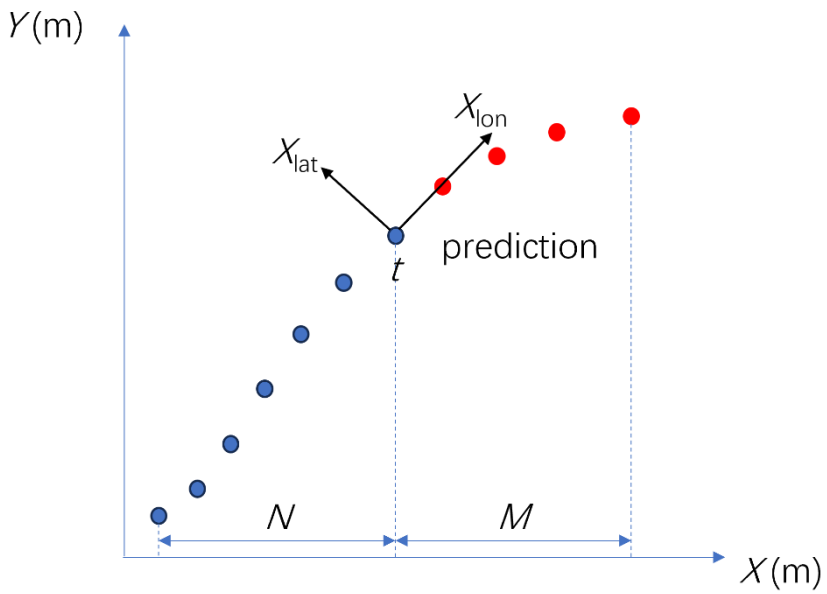

Figure 5. Trajectory Prediction Scheme.

Trajectory prediction is to estimate the future positions of the objects, usually considering the subject historical states as inputs. For example, the past *N* positions are used for trajectory prediction, which outputs a trajectory of *M* future positions. Thus, the data in the training dataset needs to be set as pairs of *N* and *M* elements, as shown in

Besides subject historical data, some models also combine the interactions between VRUs, the interactions between the subject VRU and vehicles, the VRU's position in the crosswalk, and even the VRU's age and gender [92]. Regarding the time window, the historical trajectory window and prediction time window was chosen as 1s and 2.5s in [104]. The concrete time window varied depending on the specific accident prevention strategy implemented in a practical application, considering the human reaction time of 1-1.5s [107], and it can be varied in a range of time to explore the optimal performance corresponding to some specific models and datasets. Moreover, some studies in this domain chose a general time window that has been used by many popular methods to facilitate the performance comparison. For example, Social LSTM [99], Social Attention [100], STAR [98] and SimAug [102] all took a 3.2s time window for historical trajectories and predicted the path of a 4.8s window in the future, with a frame rate of 0.4.

5 Communication

In the VRU collision avoidance and warning system, communication is the basis of the data transmission. A comparison of different communication technologies is listed below, along with an application analysis featuring examples relevant to collision avoidance.

5.1 Communication Technologies

There are several types of communication technologies can be applied to the system, including ZigBee [108], Wi-Fi [109], Dedicated Short-Range Communication (DSRC) [110], LP-WAN [111], cellular network, and Visible Light Communications (VLC) [112]. Among these technologies, DSRC and cellular network are the most common used. ZigBee is a kind of technology used in experiment because of its low complexity. VLC is the new technology that the visible light can be utilized for communication and is not in large scale application. Two mainly used communication technologies in Connected and Automated Vehicle (CAV) domain will be introduced. Between them, C-V2X in cellular network will be used in this project because of its efficiency and stability, and the policy guidance.

TABLE 4 Different Network Communication Technologies.

Network	Media	Channel	Range	Benefits	Limitations	Costs
Short-Range Networks	ZigBee	2.4 GHz	0–20 m	Low power, good range, mesh networking	Lower data rate, interference issues	Low to moderate
Local Area Networks	Wi-Fi	2.4 GHz	30 m	High speed, commonly available, scalable	Security risks, interference	Moderate
	DSRC	5.9 GHz band	100 m	Fast communication with low latency, high reliability and accuracy	Limited range and scalability.	Moderate to high
Wide Area Networks	LP-WAN	2.4 GHz	2–15 km	Long range, low power, good penetration	Lower data rate, longer latency	Moderate to high
	LTE	1800–2200 MHz band (3G), 2600 MHz bands (4G)	Up to 100 km	Broad coverage, good speed (4G)	Variable coverage, costly data plans	High
	5G	Millimeter wave spectrum (30–300 GHz)	2 km	Very high speed, low latency, massive IoT	Limited coverage, high infrastructure cost	Very high
	C-V2X	LTE, 5G	Depend on the cellular network	Fast communication with low latency, longer range	Dependent on cellular network coverage and quality	High

5.1.1 Dedicated Short-range Communication (DSRC)

DSRC, founded on the IEEE 802.11p standard, is a proficient wireless communication technology facilitating the detection and interaction with mobile entities traversing swiftly within designated compact regions. In 1999, the United States Federal Communications Commission (FCC) designated 75 MHz of spectrum (from 5850 MHz to 5925 MHz) within the 5.9 GHz frequency band for V2X applications. Within this allocation, the frequency band from 5885 MHz to 5895 MHz has been specifically reserved as the "control channel" for traffic safety purposes. The On-Board Unit (OBU), Roadside Unit (RSU), and associated protocols play significant roles within the DSRC system. There are several advantages of DSRC, including minimal end-to-end delay, self-organization, and cost-effectiveness in V2X scenarios [113]. In addition, regarding the privacy issue, robust security measures like encryption and authentication, and privacy-enhancing technologies like differential privacy should be enabled within the DSRC systems.

5.1.2 Cellular Network

The cellular-based V2X technology is called C-V2X. C-V2X utilizes 3GPP [114] cellular standards such as LTE [115], 5G [116] and 5G New Radio [117]. In November 2020, FCC has reduced the band to 30MHz (5895MHz to 5925MHz) for V2X and required the use of C-V2X technology in the future. C-V2X has two modes of communication, including device-to-device (D2D) and device-to-network (D2N). One is direct C-V2X, and the other one is indirect C-V2X. Indirect C-V2X is very useful for large scale traffic management when there are many vehicles connected to the network. Comparing the DSRC and C-V2X, C-V2X has extended communication range, improved performance in non-line-of-sight conditions, increased capacity, enhanced reliability, and good congestion management ability [118].

5.2 V2X Applications Related to VRUs

In this project, several information transmission directions have been utilized to help the road users communicate with each other and ensure the VRU safety. For the VRU, there exists pedestrian-to-infrastructure (P2I) and infrastructure-to-pedestrian (I2P). P2I means the information of the VRU can be collected and transferred to the roadside infrastructure. I2P means the warning message can be delivered to the carry-on device of the VRU from infrastructure side. For the vehicle, there are vehicle-to-infrastructure (V2I) and infrastructure-to-vehicle (I2V), which are similar to the interactions between the VRU and the infrastructure. Localization of the vehicle can be transferred to the infrastructure and warning message of collision avoidance will be send to the vehicle. Further, the cooperative perception of the vehicle and the infrastructure may be considered if the vehicle is the CAV with environment perception ability, which means that the perception information of the vehicle can be transmitted to the infrastructure side for more accurate detection. The interactions between the vehicle and the infrastructure and the VRU and the infrastructure are depicted in Figure 6.

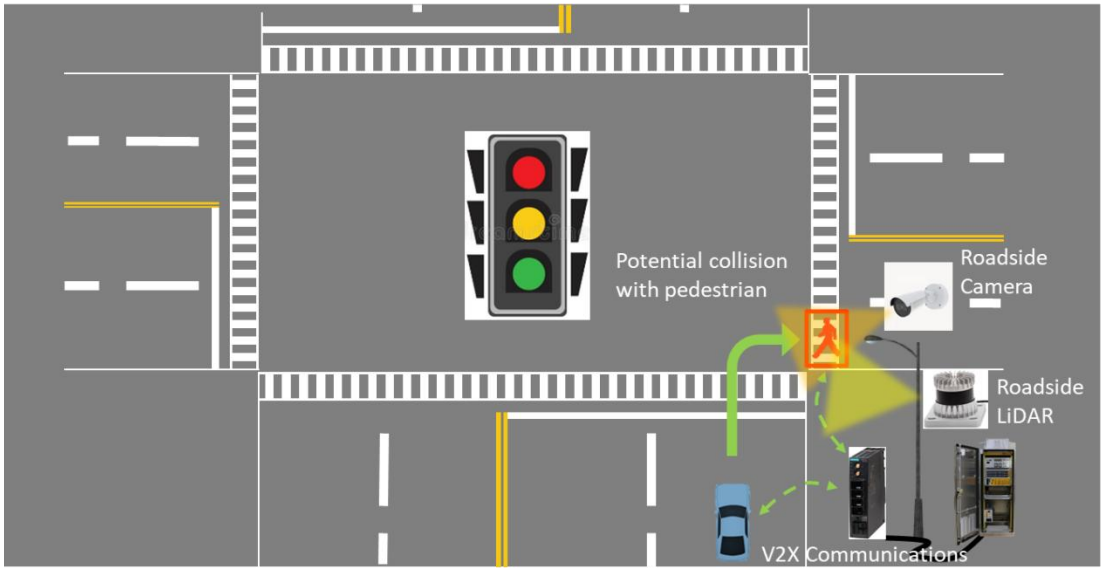

Figure 6. The Interactions Between the Road Users and the Infrastructure.

At the signalized intersection, the VRU potential collision risks are safety issues, so, higher requirements for the transmission efficiency and stability of communication systems will be put forward. According to [81], the communication requirement for intersection collision warning is shown in TABLE 5.

TABLE 5 Communication Requirements for Safety-related Application.

	End-to-end Latency	**Communication Type**	**Data Rate**
Intersection Collision Warning	100ms	Event-driven	<10kb/s

Regarding the collision avoidance, many researchers utilized the advanced communication technologies to construct an Intelligent Transportation System (ITS). Regarding the warning to the vehicle side, there are several research cases. Caltrans collaborated with other partners established a testbed (Connected Vehicle Test Bed on El Camino Real) for cooperative intersection collision avoidance that utilize DSRC for communication between vehicles and the infrastructure [120]. ACTIVE-AURORA is a testbed setup in Canada, which aims to reduce pedestrian collision at signalized intersections with connected vehicles and V2I communication with DSRC. Nanyang Technological University also built a 5G C-V2X testbed to transmit traffic navigation and risks to the vehicle to strengthen the road safety [121]. Tarchoun et al. [41] proposed an ITS detection and warning system with a pedestrian detector to warn the vehicles near the presence of pedestrians. In another similar ITS system, the safety-critical object information is transmitted from the RSU to the vehicle and converted into the vehicle's coordinate system using its navigation data [43]. Regarding the warning to both vehicle and VRU side, Merdrignac et al. [122] fused the perception information of the vehicle and pedestrian-to-vehicle (P2V) communication information to detect the VRU. To get the P2V information, the vehicle firstly broadcasted Cooperative Awareness Message (CAM) to surroundings. If the VRU was in the geographical dissemination area of the vehicle, the localization, velocity, and class of the VRU would be transmitted to the vehicle. The system was effective both in line-of-sight (LOS) and non-LOS scenarios. However, the limitation of the system is the classification performance and Global Positioning System (GPS) location accuracy. Ruß et al. [123] further utilized a RSU equipping with a RADAR as a mediator. It can

receive the cyclically emitted information from both the vehicle and the VRU near the intersection. Combined with the detection results of roadside RADAR, if there were any potential risks, the warning message would be transmitted to the vehicle side and the VRU side through DSRC and Wi-Fi, respectively. This method worked well at the signalized intersection, which was quite similar to our technical pipeline, but its sensor and communication protocol was outdated. Parada et al. [124] proposed a 5G-based collision avoidance system to protect the VRU. It utilized a 5G cellular base station connected with a Mobile Edge Computing (MEC) server to receive localization information and transmit warning messages to vehicles and VRUs. It predicted the trajectories of the road users with a regression model using the positions as input. Compared with the former methods, it improved the communication protocol but did not draw support from roadside sensors. Our technical pipeline combines the advantages of the previous methods to enhance the accuracy and efficiency of the collision avoidance system to protect the VRU.

6 Discussion

6.1 Detection

Detection is the first and most important part in this project. An accurate detection results will lead to the good performance of the whole pipeline. Regarding the detection of VRU, including pedestrians, bicyclist, scooter and skateboarder, no advanced methods exist that yield satisfactory performance for all types of VRUs. Most of the proposed methods focus on one or two classes of VRUs and do not guarantee the detection accuracy of other classes. Since we mainly focus on VRU detection using roadside LiDAR to leverage its advantages, there are fewer existing methods. However, there are many LiDAR-based 3D detection methods proposed for onboard vehicle detection, and these methods may be tuned by parameters and trained with roadside datasets for VRU detection, considering the difference in reference frames, sensor views and angles. Furthermore, the lack of suitable datasets for roadside VRU detection should also be concerned. With thorough searching of the related open-source datasets, DAIR-V2X-I, TUMTraf Intersection Dataset and V2X-Real are suitable for VRU detection. The only disadvantage of DAIR-V2X-I and TUMTraf is that both datasets lack of scooter and skateboarder, the kinds of fast moving VRU. V2X-Real is a dataset proposed in March 2024, which includes pedestrian, cyclist, and scooter as VRUs. However, the dataset has only recently been released. Thus, for roadside VRU detection, accurate detection methods and comprehensive datasets are still lacking, and this challenge needs a baseline solution for further research.

6.2 Tracking

Tracking is based on the detection results. There are two kinds of basic tracking pipelines. One compares the extracted features of objects to achieve association; the other uses a motion prediction model and compare the distance of the predictions and measurements in one frame. For roadside VRU tracking, the method must be real-time with high accuracy, because of the safety issue. Most 3D MOT approaches are using motion models for state prediction, distance-based methods for matching, and Kalman filter for updating the final states. This pipeline is straightforward and time-effective, with satisfactory performance. However, there are still some problems like occlusion and accuracy of long-distance detection, which will be a concern in the future. Another issue is the existing VRU tracking methods only consider the pedestrian and cyclist at most, without tracking of scooter and skateboarder. PolyMOT proposed a method, describing the motions of the objects using different models according to their categories. Referring to this method, applying different motion model to the fast-moving scooter and skateboarder categories may be a good solution.

6.3 Prediction

Prediction modules do not much rely on the application scenarios. Therefore, the approaches are relatively universal for different scenarios, as long as the detection and tracking results are accurate enough. Thus, the existing onboard pedestrian tracking methods can be utilized for roadside VRU prediction. When selecting a prediction model, it is important to consider aspects such as the model's efficiency, accuracy, and robustness. Robustness is an important model property for

detecting obscured or unseen objects, which may be ignored by some methods. Because of the insufficiency of disturbance scenarios, few methods add random perturbation and adversarial attack to the input data to improve the robustness. For accuracy, the existing methods are improving performance with modern technologies like Transformer. However, there is less attention paid to efficiency and a comprehensive analysis of existing prediction methods is lacking, particularly for the intention prediction.

Regarding the performance of the prediction model, because the roadside sensors are fixed and each intersection has its own characteristics, different intersection may use the specific model trained for prediction with its own large-amount historical data, which may be more effective.

6.4 Communication

At the signalized intersection, higher requirements for the communication system's transmission efficiency and stability will be put forward to mitigate safety risks for VRUs. Considering the requirement for V2I and P2I communication, 5G C-V2X is a practical method that can be used to transmit information to both vehicles and pedestrians in the same protocol, with satisfactory efficiency and stability. Communication plays a crucial role in the VRU collision avoidance system. Apart from its basic warning function, the transmitted information such as position, velocity, and class can be used as input for an interaction strategy or detection algorithm, especially in non-LOS scenarios. Therefore, communication serves as a supplement for sensor detection and is essential for the successful operation of the VRU collision avoidance system.

7 Conclusions

In this report, a clear technical pipeline is given for VRU collision avoidance with four modules, including detection, tracking, prediction, and communication. General definition and related papers are introduced in each module. Considering the biggest technical difficulty of this project is detection, which determines the upper limit of the following modules, more content with detailed analysis is introduced in detection part. With analyzing of the existing methods, a discussion of the limitations and challenges is also proposed. After conducting a literature review, we can enhance the pipeline by selecting and combining appropriate sensors, improving concrete algorithms, selecting training datasets for algorithms, and choosing suitable testing simulation environments, which will be a good start for the next algorithm development task.

References

[1] California Office of Traffic Safety. California Traffic Safety Quick Stats. Accessed: 2021. [Online]. Available: https://www.ots.ca.gov/ots-and-traffic-safety/score-card/

[2] California Highway Patrol. 2020 California Quick Crash Facts. Accessed: 2020. [Online]. Available: https://www.chp.ca.gov/programs-services/services-information/switrs-internet-statewide-integrated-traffic-records-system/switrs-2020-report

[3] Sun, Carlos, et al. *Vulnerable Road User (VRU) Safety Assessment*. No. cmr 23-015. Missouri. Department of Transportation. Construction and Materials Division, 2023.

[4] Zhou, Yin, and Oncel Tuzel. "Voxelnet: End-to-end learning for point cloud based 3d object detection." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

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

[6] Arnold, Eduardo, et al. "A survey on 3d object detection methods for autonomous driving applications." *IEEE Transactions on Intelligent Transportation Systems* 20.10 (2019): 3782-3795.

[7] Wu, Jianqing, et al. "An automatic skateboarder detection method with roadside LiDAR data." *Journal of Transportation Safety & Security* 13.3 (2021): 298-317.

[8] Zhao, Junxuan, et al. "Detection and tracking of pedestrians and vehicles using roadside LiDAR sensors." *Transportation research part C: emerging technologies* 100 (2019): 68-87.

[9] Song, Yanjie, et al. *Road-Users Classification Utilizing Roadside Light Detection and Ranging Data*. No. 2020-01-5150. SAE Technical Paper, 2020.

[10] Xu, YuLin, et al. "A Common Traffic Object Recognition Method Based on Roadside LiDAR." *2022 IEEE 7th International Conference on Intelligent Transportation Engineering (ICITE)*. IEEE, 2022.

[11] Zhang, Liwen, et al. "Gc-net: Gridding and clustering for traffic object detection with roadside lidar." *IEEE Intelligent Systems* 36.4 (2020): 104-113.

[12] Gong, Ziren, et al. "FecNet: A Feature Enhancement and Cascade Network for Object Detection Using Roadside LiDAR." *IEEE Sensors Journal* (2023).

[13] Lang, Alex H., et al. "Pointpillars: Fast encoders for object detection from point clouds." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019.

[14] Yan, Yan, Yuxing Mao, and Bo Li. "Second: Sparsely embedded convolutional detection." *Sensors* 18.10 (2018): 3337.

[15] Shi, Shaoshuai, Xiaogang Wang, and Hongsheng Li. "Pointrcnn: 3d object proposal generation and detection from point cloud." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019.

[16] Shi, Shaoshuai, et al. "Pv-rcnn: Point-voxel feature set abstraction for 3d object detection." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020.

[17] Shi, Shaoshuai, et al. "PV-RCNN++: Point-voxel feature set abstraction with local vector representation for 3D object detection." *International Journal of Computer Vision* 131.2

(2023): 531-551.
[18] Shi, Guangsheng, Ruifeng Li, and Chao Ma. "Pillarnet: Real-time and high-performance pillar-based 3d object detection." *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2022.
[19] Nasrollahi, Kamal, and Miquel Romero Blanch. "LiDAR-Assisted 3D Human Detection for Video Surveillance." *2024 IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2024*. IEEE, 2024.
[20] Yin, Tianwei, Xingyi Zhou, and Philipp Krahenbuhl. "Center-based 3d object detection and tracking." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021.
[21] Shi, Haobo, Dezao Hou, and Xiyao Li. "Center-aware 3D object detection with attention mechanism based on roadside LiDAR." *Sustainability* 15.3 (2023): 2628.
[22] Yu, Haibao, et al. "Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.
[23] Wang, Yue, and Justin M. Solomon. "Object dgcnn: 3d object detection using dynamic graphs." *Advances in Neural Information Processing Systems* 34 (2021): 20745-20758.
[24] Fan, Lue, et al. "Fully sparse 3d object detection." *Advances in Neural Information Processing Systems* 35 (2022): 351-363.
[25] Sindagi, Vishwanath A., Yin Zhou, and Oncel Tuzel. "Mvx-net: Multimodal voxelnet for 3d object detection." *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019.
[26] Viola, Paul, and Michael J. Jones. "Robust real-time face detection." *International journal of computer vision* 57 (2004): 137-154.
[27] Dalal, Navneet, and Bill Triggs. "Histograms of oriented gradients for human detection." *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*. Vol. 1. Ieee, 2005.
[28] Papageorgiou, Constantine, and Tomaso Poggio. "A trainable system for object detection." *International journal of computer vision* 38 (2000): 15-33.
[29] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems* 25 (2012).
[30] Girshick, Ross, et al. "Region-based convolutional networks for accurate object detection and segmentation." *IEEE transactions on pattern analysis and machine intelligence* 38.1 (2015): 142-158.
[31] He, Kaiming, et al. "Spatial pyramid pooling in deep convolutional networks for visual recognition." *IEEE transactions on pattern analysis and machine intelligence* 37.9 (2015): 1904-1916.
[32] Girshick, Ross. "Fast r-cnn." *Proceedings of the IEEE international conference on computer vision*. 2015.
[33] Ren, Shaoqing, et al. "Faster r-cnn: Towards real-time object detection with region proposal networks." *Advances in neural information processing systems* 28 (2015).
[34] Redmon, Joseph, et al. "You only look once: Unified, real-time object detection." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
[35] Wang, Chien-Yao, I-Hau Yeh, and Hong-Yuan Mark Liao. "YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information." *arXiv preprint*

arXiv:2402.13616 (2024).
[36] Liu, Wei, et al. "Ssd: Single shot multibox detector." *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer International Publishing, 2016.
[37] Lin, Tsung-Yi, et al. "Feature pyramid networks for object detection." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
[38] Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).
[39] Garcia-Venegas, Marichelo, et al. "On the safety of vulnerable road users by cyclist detection and tracking." *Machine Vision and Applications* 32.5 (2021): 109.
[40] Welch, Greg, and Gary Bishop. "An introduction to the Kalman filter." (1995): 2.
[41] Tarchoun, Bilel, et al. "Deep cnn-based pedestrian detection for intelligent infrastructure." *2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*. IEEE, 2020.
[42] Mammeri, Abdelhamid, et al. "Vulnerable road users detection based on convolutional neural networks." *2020 International Symposium on Networks, Computers and Communications (ISNCC)*. IEEE, 2020.
[43] Ojala, Risto, et al. "Novel convolutional neural network-based roadside unit for accurate pedestrian localization." *IEEE Transactions on Intelligent Transportation Systems* 21.9 (2019): 3756-3765.
[44] Li, Yinhao, et al. "Bevdepth: Acquisition of reliable depth for multi-view 3d object detection." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. No. 2. 2023.
[45] Yang, Lei, et al. "Bevheight: A robust framework for vision-based roadside 3d object detection." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.
[46] Fan, Siqi, et al. "Calibration-free bev representation for infrastructure perception." *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023.
[47] Shi, Hao, et al. "Cobev: Elevating roadside 3d object detection with depth and height complementarity." *arXiv preprint arXiv:2310.02815* (2023).
[48] Wang, Hui, et al. "A scheme on pedestrian detection using multi-sensor data fusion for smart roads." *2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*. IEEE, 2020.
[49] Meissner, Daniel, Stephan Reuter, and Klaus Dietmayer. "Real-time detection and tracking of pedestrians at intersections using a network of laserscanners." *2012 IEEE Intelligent Vehicles Symposium*. IEEE, 2012.
[50] Strigel, Elias, et al. "The ko-per intersection laserscanner and video dataset." *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2014.
[51] Stauffer, Chris, and W. Eric L. Grimson. "Adaptive background mixture models for real-time tracking." *Proceedings. 1999 IEEE computer society conference on computer vision and pattern recognition (Cat. No PR00149)*. Vol. 2. IEEE, 1999.
[52] Zimmer, Walter, et al. "Infradet3d: Multi-modal 3d object detection based on roadside infrastructure camera and lidar sensors." *2023 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2023.
[53] Zhou, Qian-Yi, Jaesik Park, and Vladlen Koltun. "Fast global registration." *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*. Springer International Publishing, 2016.

[54] Besl, Paul J., and Neil D. McKay. "Method for registration of 3-D shapes." *Sensor fusion IV: control paradigms and data structures*. Vol. 1611. Spie, 1992.
[55] Creß, Christian, et al. "TUMTraf Event: Calibration and Fusion Resulting in a Dataset for Roadside Event-Based and RGB Cameras." *arXiv preprint arXiv:2401.08474* (2024).
[56] Xu, Dan, et al. "Learning cross-modal deep representations for robust pedestrian detection." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
[57] Bai, Jie, et al. "Robust target detection and tracking algorithm based on roadside radar and camera." *Sensors* 21.4 (2021): 1116.
[58] Liu, Pengfei, et al. "Object classification based on enhanced evidence theory: Radar–vision fusion approach for roadside application." *IEEE Transactions on Instrumentation and Measurement* 71 (2022): 1-12.
[59] Zhao, Huijing, et al. "Sensing an intersection using a network of laser scanners and video cameras." IEEE Intelligent Transportation Systems Magazine 1.2 (2009): 31-37.
[60] Bai, Zhengwei, et al. "A survey and framework of cooperative perception: From heterogeneous singleton to hierarchical cooperation." *arXiv preprint arXiv:2208.10590* (2022).
[61] Eshel, Ran, and Yael Moses. "Homography based multiple camera detection and tracking of people in a dense crowd." *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008.
[62] Wang, Xiaogang. "Intelligent multi-camera video surveillance: A review." *Pattern recognition letters* 34.1 (2013): 3-19.
[63] Bai, Zhengwei, et al. "Pillargrid: Deep learning-based cooperative perception for 3d object detection from onboard-roadside lidar." *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2022.
[64] Rauch, Andreas, et al. "Car2x-based perception in a high-level fusion architecture for cooperative perception systems." *2012 IEEE Intelligent Vehicles Symposium*. IEEE, 2012.
[65] Neubeck, Alexander, and Luc Van Gool. "Efficient non-maximum suppression." *18th international conference on pattern recognition (ICPR'06)*. Vol. 3. IEEE, 2006.
[66] Arnold, Eduardo, et al. "Cooperative perception for 3D object detection in driving scenarios using infrastructure sensors." *IEEE Transactions on Intelligent Transportation Systems* 23.3 (2020): 1852-1864.
[67] Sun, Pei, et al. "Scalability in perception for autonomous driving: Waymo open dataset." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020.
[68] Caesar, Holger, et al. "nuscenes: A multimodal dataset for autonomous driving." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020.
[69] Yongqiang, Deng, et al. "Baai-vanjee roadside dataset: Towards the connected automated vehicle highway technologies in challenging environments of china." *arXiv preprint arXiv:2105.14370* (2021).
[70] Zimmer, Walter, et al. "TUMTraf Intersection Dataset: All You Need for Urban 3D Camera-LiDAR Roadside Perception." *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2023.
[71] Ye, Xiaoqing, et al. "Rope3d: The roadside perception dataset for autonomous driving and monocular 3d object detection task." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

[72] Zimmer, Walter, et al. "TUMTraf V2X Cooperative Perception Dataset." *arXiv preprint arXiv:2403.01316* (2024).
[73] Xiang, Hao, et al. "V2X-Real: a Largs-Scale Dataset for Vehicle-to-Everything Cooperative Perception." *arXiv preprint arXiv:2403.16034* (2024).
[74] Li, Yiming, et al. "V2X-Sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving." *IEEE Robotics and Automation Letters* 7.4 (2022): 10914-10921.
[75] Pérez, Patrick, et al. "Color-based probabilistic tracking." *Computer Vision—ECCV 2002: 7th European Conference on Computer Vision Copenhagen, Denmark, May 28–31, 2002 Proceedings, Part I 7*. Springer Berlin Heidelberg, 2002.
[76] Nguyen, Hieu Tat, Marcel Worring, and Rein Van Den Boomgaard. "Occlusion robust adaptive template tracking." *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*. Vol. 1. IEEE, 2001.
[77] Zheng, Bin, et al. "Object tracking algorithm based on combination of dynamic template matching and kalman filter." *2012 4th International Conference on Intelligent Human-Machine Systems and Cybernetics*. Vol. 2. IEEE, 2012.
[78] Ristic, Branko, Sanjeev Arulampalam, and Neil Gordon. *Beyond the Kalman filter: Particle filters for tracking applications*. Artech house, 2003.
[79] Ghahramani, Zoubin. "Learning dynamic Bayesian networks." *International School on Neural Networks, Initiated by IIASS and EMFCSC* (1997): 168-197.
[80] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." *Neural computation* 9.8 (1997): 1735-1780.
[81] Wojke, Nicolai, Alex Bewley, and Dietrich Paulus. "Simple online and realtime tracking with a deep association metric." *2017 IEEE international conference on image processing (ICIP)*. IEEE, 2017.
[82] Zhang, Yifu, et al. "Bytetrack: Multi-object tracking by associating every detection box." *European conference on computer vision*. Cham: Springer Nature Switzerland, 2022.
[83] Pang, Jiangmiao, et al. "Quasi-dense similarity learning for multiple object tracking." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021.
[84] Wu, Jianqing, et al. "Automatic vehicle classification using roadside LiDAR data." *Transportation Research Record* 2673.6 (2019): 153-164.
[85] Reuter, Stephan, and Klaus Dietmayer. "Pedestrian tracking using random finite sets." *14th International Conference on Information Fusion*. IEEE, 2011.
[86] Sidenbladh, Hedvig, and Sven-Lennart Wirkander. "Tracking random sets of vehicles in terrain." *2003 Conference on Computer Vision and Pattern Recognition Workshop*. Vol. 9. IEEE, 2003.
[87] Zhengwei, et al. "Cyber mobility mirror: a deep learning-based real-world object perception platform using roadside LiDAR." *IEEE Transactions on Intelligent Transportation Systems* (2023).
[88] Weng, Xinshuo, et al. "3d multi-object tracking: A baseline and new evaluation metrics." *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020.
[89] Pang, Ziqi, Zhichao Li, and Naiyan Wang. "Simpletrack: Understanding and rethinking 3d multi-object tracking." *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2022.
[90] Li, Xiaoyu, et al. "Poly-mot: A polyhedral framework for 3d multi-object tracking." *2023*

IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2023.
[91] Schöller, Christoph, et al. "What the constant velocity model can teach us about pedestrian motion prediction." *IEEE Robotics and Automation Letters* 5.2 (2020): 1696-1703.
[92] Chen, Hao, et al. "Vulnerable road user trajectory prediction for autonomous driving using a data-driven integrated approach." *IEEE Transactions on Intelligent Transportation Systems* (2023).
[93] Zhang, Xi, et al. "Pedestrian path prediction for autonomous driving at un-signalized crosswalk using W/CDM and MSFM." *IEEE Transactions on Intelligent Transportation Systems* 22.5 (2020): 3025-3037.
[94] Song, Chao, et al. "Human trajectory prediction for automatic guided vehicle with recurrent neural network." *The Journal of Engineering* 2018.16 (2018): 1574-1578.
[95] Zhang, Yanbo, and Liying Zheng. "Pedestrian trajectory prediction with MLP-social-GRU." *Proceedings of the 2021 13th International Conference on Machine Learning and Computing*. 2021.
[96] Shi, Liushuai, et al. "SGCN: Sparse graph convolution network for pedestrian trajectory prediction." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.
[97] Mohamed, Abduallah, et al. "Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020.
[98] Yu, Cunjun, et al. "Spatio-temporal graph transformer networks for pedestrian trajectory prediction." *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*. Springer International Publishing, 2020.
[99] Alahi, Alexandre, et al. "Social lstm: Human trajectory prediction in crowded spaces." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
[100] Vemula, Anirudh, Katharina Muelling, and Jean Oh. "Social attention: Modeling attention in human crowds." *2018 IEEE international Conference on Robotics and Automation (ICRA)*. IEEE, 2018.
[101] Xu, Qing, et al. "Roadside pedestrian motion prediction using Bayesian methods and particle filter." *IET Intelligent Transport Systems* 15.9 (2021): 1167-1182.
[102] Liang, Junwei, Lu Jiang, and Alexander Hauptmann. "Simaug: Learning robust representations from simulation for trajectory prediction." *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*. Springer International Publishing, 2020.
[103] Mercurius, Ray Coden, et al. "AMEND: A Mixture of Experts Framework for Long-tailed Trajectory Prediction." *arXiv preprint arXiv:2402.08698* (2024).
[104] Goldhammer, Michael, et al. "Intentions of vulnerable road users—detection and forecasting by means of machine learning." *IEEE transactions on intelligent transportation systems* 21.7 (2019): 3035-3045.
[105] Zhao, Junxuan, et al. "Trajectory tracking and prediction of pedestrian's crossing intention using roadside LiDAR." *IET Intelligent Transport Systems* 13.5 (2019): 789-795.
[106] Zhao, Junxuan, et al. "Probabilistic prediction of pedestrian crossing intention using roadside LiDAR data." *IEEE Access* 7 (2019): 93781-93790.
[107] Green, Marc. "" How long does it take to stop?" Methodological analysis of driver perception-brake times." *Transportation human factors* 2.3 (2000): 195-216.

[108] Gislason, Drew. *Zigbee wireless networking*. Newnes, 2008.
[109] Małecki, Krzysztof, Stanisław Iwan, and Kinga Kijewska. "Influence of intelligent transportation systems on reduction of the environmental negative impact of urban freight transport based on Szczecin example." Procedia-Social and Behavioral Sciences 151 (2014): 215-229.
[110] Wall, J. P. "The CITI Project: Building Australia's first Cooperative Intelligent Transport System test facility for safety applications." In Proceedings of the 2013 Australasian Road Safety Research, Policing & Education Conference. 2013.
[111] Gadre, Akshay, Diana Zhang, and Swarun Kumar. "Towards Enabling City-Scale Internet of Things–Challenges and Opportunities." In 2019 11th International Conference on Communication Systems & Networks (COMSNETS), pp. 72-79. IEEE, 2019.
[112] Ndjiongue, Alain Richard, Hendrik C. Ferreira, and Telex MN Ngatched. "Visible light communications (VLC) technology." *Wiley Encyclopedia of Electrical and Electronics Engineering* (1999): 1-15.
[113] MacHardy, Zachary, et al. "V2X access technologies: Regulation, research, and remaining challenges." *IEEE Communications Surveys & Tutorials* 20.3 (2018): 1858-1877.
[114] 3GPP. "3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Requirements for Evolved UTRA (E-UTRA) and Evolved UTRAN (E-UTRAN) (Release 7)". *3GPP*, 2006.
[115] Dahlman, Erik, Stefan Parkvall, and Johan Skold. *4G: LTE/LTE-advanced for mobile broadband*. Academic press, 2013.
[116] Andrews, Jeffrey G., et al. "What will 5G be?" *IEEE Journal on selected areas in communications* 32.6 (2014): 1065-1082.
[117] Dahlman, Erik, Stefan Parkvall, and Johan Skold. *5G NR: The next generation wireless access technology*. Academic Press, 2020.
[118] 5G Automotive Association. "V2X functional and performance test report; test procedures and results." *5GAA, Munich, Germany, Tech. Rep. 5GAA P-190033* (2019).
[119] Fu, Yuchuan, et al. "A survey of driving safety with sensing, vehicular communications, and artificial intelligence-based collision avoidance." *IEEE transactions on intelligent transportation systems* 23.7 (2021): 6142-6163.
[120] Caltrans and California PATH Program at UC Berkeley. *Connected Vehicle Test Bed*. Accessed: May 2021. [Online]. Available: http://www.caconnectedvehicletestbed.org/index.php/
[121] *NTU Singapore and M1 Ink Partnership to Develop Singapore's First 5G C-V2X Research Testbed and Trials.* Accessed: Oct. 2019. [Online]. Available: https://www.ntu.edu.sg/news/detail/ntu-singapore-and-m1-ink-partnership-to-develop-singapore-s-first-5g-c-v2x-research-testbed-and-trials
[122] Merdrignac, Pierre, Oyunchimeg Shagdar, and Fawzi Nashashibi. "Fusion of perception and v2p communication systems for the safety of vulnerable road users." *IEEE Transactions on Intelligent Transportation Systems* 18.7 (2016): 1740-1751.
[123] Ruß, Tim, Jan Krause, and René Schönrock. "V2X-based cooperative protection system for vulnerable road users and its impact on traffic." *ITS World Congress*. 2016.
[124] Parada, Raúl, et al. "Machine learning-based trajectory prediction for vru collision avoidance in v2x environments." *2021 IEEE global communications conference (GLOBECOM)*. IEEE, 2021.

APPENDIX 2 METHODOLOGY

1 Introduction

1.1 Background and Motivation

The detection and protection of Vulnerable Road Users (VRUs), including pedestrians, cyclists, and other non-motorized road users, present significant challenges in ensuring road safety within modern transportation systems [1, 2]. VRU detection based on roadside sensors faces unique obstacles due to the small size of VRUs, high density of pedestrians in urban environments, and frequent occlusions by vehicles [3]. These factors hinder the performance of single-sensor systems, necessitating the use of multi-sensor collaborative perception. By integrating data from multiple complementary sensors, such as HD cameras and LiDAR (light detection and ranging), the system can overcome individual sensor limitations and achieve improved coverage and reliability [4, 5].
Moreover, adverse environmental conditions, including rain, snow, fog, and varying lighting, further restrict the effectiveness of standalone sensors [6]. Cameras, for example, are sensitive to illumination variations, while LiDAR may struggle with sparse data representation in low visibility scenarios [7, 8]. Addressing these challenges requires multi-modal sensor fusion, combining diverse sensor types to leverage their respective strengths and mitigate weaknesses. Such fusion ensures robustness in detecting VRUs across diverse and dynamic traffic environments.
Beyond detection, tracking and trajectory prediction for VRUs pose additional difficulties. Tracking VRUs requires maintaining accurate associations across consecutive frames, which is challenging in scenarios involving frequent occlusions or heavy VRU volumes [9]. Additionally, trajectory prediction must account for complex interactions between VRUs and surrounding road users, including vehicles [10]. Accurate predictions are essential for proactive collision avoidance, but achieving robust and real-time performance under these conditions requires advanced algorithms that can handle data noises and dynamically adapt to unpredictable behaviors [11].
Communication and terminal notifications further complicate the development of a comprehensive VRU protection system [12, 13]. Transmitting high-priority safety messages with low latency to VRUs and vehicles requires efficient and reliable communication protocols, especially in non-line-of-sight scenarios. The integration of Vehicle-to-Everything (V2X) technology and mobile networks is critical [14], but challenges such as network congestion, coverage limitations, and interoperability between devices must be addressed. Furthermore, delivering timely and effective alerts to end-users, including pedestrians through mobile devices and vehicles via dashboard systems, requires intuitive human-machine interface (HMI) designs and robust warning mechanisms.
These interconnected challenges highlight the need for robust, adaptive, and integrated methodologies that encompass multi-sensor detection, efficient tracking, accurate trajectory prediction, and seamless communications. Developing such systems is essential for improving VRU safety and enabling intelligent transportation systems to operate effectively in real-world environments.

1.2 Objectives

This report aims to achieve the following objectives in the development of methodologies and algorithms for Task 2 of the VRU detection project:

- **Enhance VRU Detection:** Integrate conventional and deep learning-based detection methods to improve detection accuracy and robustness under diverse conditions, including adverse weather (e.g., rain, snow, fog) and nighttime scenarios.
- **Optimize Sensor Fusion:** Develop advanced sensor fusion strategies, leveraging multi-modal data (e.g., LiDAR and cameras) to address challenges such as occlusion, small object detection, and sensor-specific limitations.
- **Address Data Limitations:** Propose a hybrid detection framework combining traditional algorithms and deep learning technique to overcome challenges of limited labeled data and calibration errors in roadside VRU detection.
- **Improve Tracking and Trajectory Prediction:** Implement real-time multi-object tracking (MOT) algorithms and accurate trajectory prediction models to handle VRU occlusions, unpredictable motion, and complex interactions with surrounding road users.
- **Ensure Effective Communication and Alerts:** Design a reliable communication framework using V2X technology and mobile networks to transmit high-priority safety messages. Develop intuitive alert systems for both vehicles (e.g., via dashboard warnings) and VRUs (e.g., via mobile or wearable device notifications) to ensure timely and actionable warnings.

1.3 Organization

The report is structured as follows: **Section 2** provides an overview of the proposed methodologies, including the system architecture and key challenges addressed. **Section 3** details the detection strategies, encompassing both conventional and deep learning-based approaches, along with their integration via method fusion. **Section 4** outlines the multi-object tracking algorithms, emphasizing real-time performance and noise handling. **Section 5** elaborates on the trajectory prediction methods, focusing on intersection scenarios and social interactions among road users. **Section 6** discusses collision prediction and the design of a comprehensive warning system. **Section 7** discusses the key findings and potential directions for future work. And the last section concludes this report.

2 Methodology Overview

As illustrate in Figure 2-1, the system architecture consists of three primary modules designed to seamlessly integrate *Object Perception*, *Information Processing*, and *Situation Awareness* into a unified pipeline for VRU detection and protection.

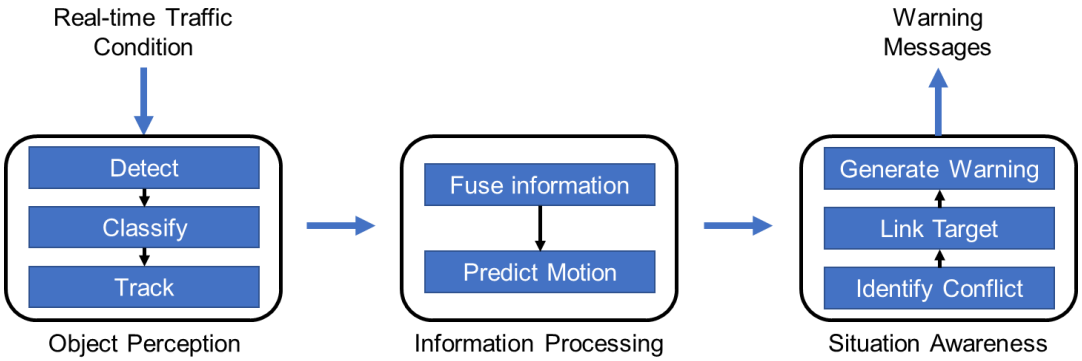

Figure 2-1. Proposed Flowchart of the System

The object perception module begins with data acquisition, utilizing a combination of visual cameras and LiDAR to collect real-time data from the roadside. This module ensures data quality through preprocessing steps, such as image correction for cameras and noise filtering for LiDAR point clouds, which improve input fidelity for downstream tasks. To achieve accurate and robust VRU detection, even under challenging scenarios such as low-light conditions, the system employs a hybrid detection pipeline. This pipeline integrates conventional 2D/3D object detection methods with deep learning-based models, leveraging the strengths of both approaches to address diverse detection requirements effectively.

The information processing module builds upon the outputs of the perception module to enhance the system's understanding of the environment. Multi-sensor and multi-node data are fused through spatio-temporal associations, improving the accuracy of detection and localization. This sensor fusion approach combines data from multiple cameras and LiDAR, aligning and integrating information to provide a comprehensive view of the scene. The module also includes a multi-object tracking system, which implements a Kalman Filter-based Constant Velocity (KF-CV) algorithm [15] to generate continuous trajectories for detected objects. This tracking system mitigates fragmentation issues through robust data association and temporal filtering. Additionally, the system incorporates an enhanced Social GAN [10] to predict multiple plausible future trajectories for VRUs. This model accounts for individual behaviors, social interactions, and environmental factors such as traffic light signals, providing valuable insights into potential movement patterns.

The situation awareness module focuses on collision prediction and alert generation. It analyzes the trajectories of road users to identify potential conflicts using metrics such as Time-to-Collision (TTC) [16]. When the TTC threshold is breached, real-time alerts are generated to notify relevant parties of imminent risks. These alerts are distributed via V2X communications to ensure timely

responses. For instance, vehicles receive dashboard warnings, while VRUs are alerted through mobile applications or other interfaces. This module ensures a proactive approach to road safety, enabling quick and effective measures to prevent accidents.

By integrating these modules into a cohesive framework, the system delivers robust, adaptive, and real-time VRU detection and protection across diverse and dynamic environments.

3 Detection

The detection module is a critical component of our system, designed to identify and classify vulnerable road users (VRUs) and vehicles accurately under diverse environmental conditions. This module integrates traditional 3D object detection methods with deep learning-based approaches, addressing challenges posed by limited training data, adverse weather conditions, and complex roadside scenarios. The detection process is divided into four stages: data preprocessing, traditional detection pipeline, deep learning-based detection, and method fusion.

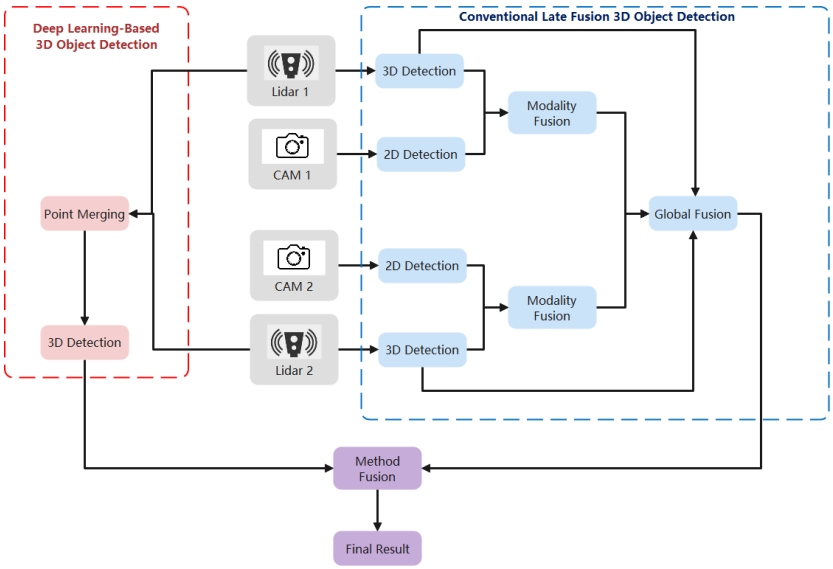

Figure 3-1. Multi-Modal and Method Fusion Framework for VRU Detection

Figure 3-1 illustrates the comprehensive detection pipeline, combining traditional and deep learning-based approaches. In the traditional method (outlined in blue), 3D and 2D detection results from different sensors (LiDAR and cameras) are initially fused at the modality level. This modality fusion step leverages the strengths of each sensor type to improve detection accuracy and robustness. Subsequently, the results from multiple nodes are combined through global fusion, ensuring that the system integrates data from all available sources to create a cohesive representation of the environment.

Parallel to the traditional pipeline, the deep learning-based method (outlined in red) applies advanced 3D object detection techniques after point cloud merging. This approach enhances the detection of small or irregularly shaped VRUs by leveraging the high adaptability of neural networks.

Finally, the method fusion module (highlighted in purple) integrates the outputs from both pipelines, combining the robustness of traditional methods with the precision of deep learning-based techniques. This multi-stage fusion ensures that the system provides highly accurate and reliable detection results under diverse and challenging conditions.

3.1 Data Preprocessing

The data preprocessing stage ensures high-quality input for subsequent detection pipelines, leveraging both visual and spatial data from cameras and LiDAR sensors.
Camera data undergoes intrinsic and extrinsic calibration to correct image distortion and align the coordinate system with global ground truth. This process is essential for accurate mapping of detected objects across modalities. Additionally, brightness normalization is applied to address lighting inconsistencies, particularly in low-light scenarios.
LiDAR data is preprocessed to enhance the fidelity of spatial information. Noise filtering eliminates irrelevant points, such as reflections from nearby objects or the ground. Voxelization reduces the density of point clouds without losing significant details, enabling efficient processing. Furthermore, static point clouds are identified using cross-frame distance analysis and subtracted to isolate dynamic objects. This step minimizes the computational load and focuses resources on objects of interest.
Preprocessed data is geo-fenced to the region of interest, ensuring that the detection pipelines prioritize areas with intensive VRU activities. These preprocessing steps form the foundation for accurate and efficient object detection.

3.2 Traditional Detection Pipeline

The traditional detection pipeline combines 2D detection, 3D detection, and multi-modal sensor fusion to leverage the complementary strengths of cameras and LiDAR sensors.

3.2.1 2D Detection with DAB-DETR

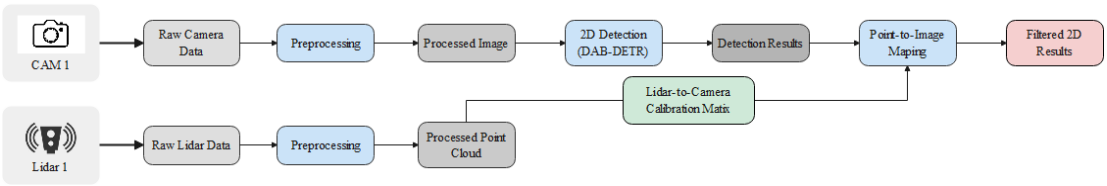

Figure 3-2. 2D Detection Pipeline

DAB-DETR (Dynamic Anchor Boxes-DETR) [17] is employed for 2D object detection (see Figure 3-2). Unlike traditional convolutional approaches, DAB-DETR introduces dynamic anchor boxes to improve the model's adaptability to varying object sizes and shapes. These anchor boxes, dynamically updated during training, allow the model to refine bounding box predictions iteratively, enhancing detection accuracy for small objects like pedestrians and cyclists.
The DAB-DETR pipeline processes pre-corrected camera images to detect objects and output 2D bounding boxes [17]. To address challenges such as overlapping detections and intermittent false positives, a heuristic multi-faceted filtering strategy is applied. This strategy consolidates bounding boxes based on confidence scores and spatial consistency, ensuring robust detection in complex roadside environments.

3.2.2 Traditional 3D Object Detection

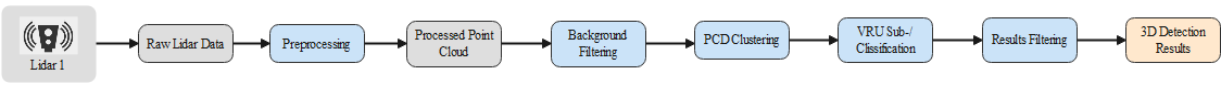

Figure 3-3. 3D Detection Pipeline

The 3D detection process (as shown in Figure 3-3) leverages LiDAR data to classify and localize VRUs and vehicles. The pipeline consists of four main stages:

1. **Background Removal**: Static objects are identified using cross-frame distance calculations and removed to isolate dynamic objects.
2. **Point Cloud Clustering**: The Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm [18] is used to group dynamic point clouds into clusters representing individual objects.
3. **VRU Sub-Classification**: Each cluster is classified into VRU subclasses using a Random Forest classifier trained on ground truth data. This classifier offers robustness to the noises and sparsity of LiDAR data, ensuring accurate differentiation of object types.
4. **Result Filtering**: To ensure the reliability of classification results, a consistency check is applied before passing the outputs to the fusion module.

3.2.3 Multi-Modal Sensor Fusion

Sensor fusion is essential for enhancing detection robustness and accuracy by integrating data from multiple cameras and LiDAR sensors. Using a Camera-to-Camera Calibration Matrix, the system aligns 2D detections from multiple cameras into a unified coordinate system. Overlapping detections are consolidated, and object center points are calculated to create a consistent spatial representation, leveraging the high resolution of cameras for precise object localization. As illustrated in Figure 3-4, this process ensures accurate spatial alignment of camera-based detections. LiDAR detections are similarly merged using a LiDAR-to-LiDAR Calibration Matrix, aligning point cloud data from multiple sensors into a shared spatial framework. This step enhances detection coverage by combining perspectives, mitigating the sparsity of individual LiDAR point clouds. The resulting 3D detections provide reliable depth information, particularly for classifying VRUs and vehicles. Figure 3-4 demonstrates how the system consolidates 3D detections into a unified output.

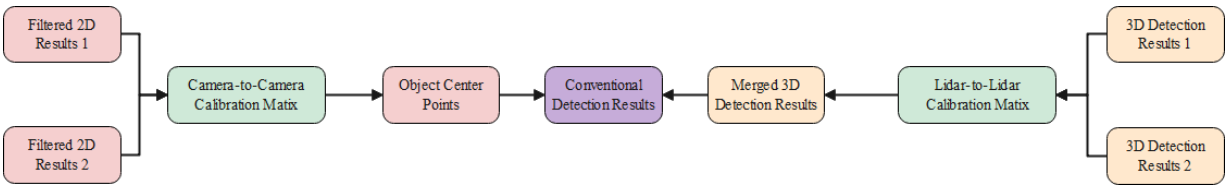

Figure 3-4 Tradition Sensor fusion

The final fusion combines outputs from both modalities to generate the Conventional Detection Results. Camera data, with its higher spatial resolution, primarily determines object positions, while LiDAR data ensures robust classification, especially for VRUs, by leveraging 3D point cloud features. A spatio-temporal association algorithm aligns the fused results, resolving inconsistencies and reducing redundancies.

This integrated fusion workflow, as outlined in Figure 3-4, balances the strengths of cameras and LiDAR, producing accurate and comprehensive detection outputs that form the foundation for downstream modules like tracking and trajectory prediction.

3.3 Deep Learning-Based Detection

The deep learning-based detection pipeline (see Figure 3-5) employs advanced neural networks to address the limitations of traditional methods in handling sparse data and complex object shapes.

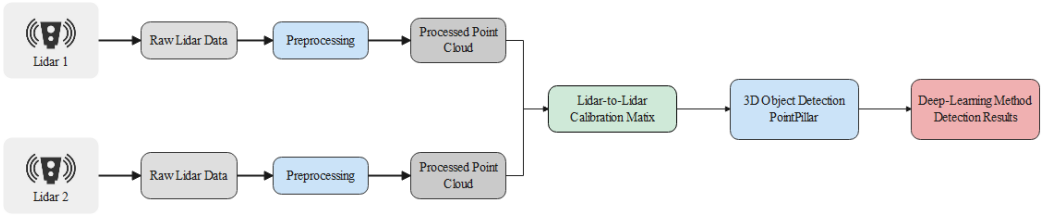

Figure 3-5 Detection Pipeline for Deep Learning-Based Detection Method

3.3.1 Point Cloud Fusion

LiDAR point clouds from multiple sensors are merged using LiDAR-to-LiDAR calibration matrices. This fusion increases the density and coverage of the data, enabling effective detection of small and irregularly shaped objects. Range reduction filters the data to the region of interest, ensuring computational efficiency while retaining relevant information.

3.3.2 PointPillars Architecture

PointPillars [19] is a state-of-the-art deep learning model designed for real-time 3D object detection. The architecture consists of three primary components:

1. **Voxelization**: The model converts raw point clouds into a sparse 3D tensor using a voxel-based feature encoding scheme [20]. This step organizes spatial data into manageable grid structures.
2. **Backbone Network**: A convolutional neural network processes the voxelized data to extract high-level features, capturing spatial and contextual information critical for object detection.
3. **Detection Head**: The final layer outputs 3D bounding boxes, including the object's position, dimensions, and classification.

PointPillars is fine-tuned on a custom dataset to adapt to roadside VRU detection scenarios. This retraining process optimizes the model for sparse and noisy data, enhancing its detection performance for complex roadside environments.

3.4 Method Fusion

Both traditional and deep learning-based detection methods have inherent limitations. Traditional methods, while robust to sparse data and noise, suffer from calibration errors during multi-step projections. Conversely, deep learning models like PointPillars require extensive labeled data,

which is often scarce in real-world scenarios. To address these challenges, we implement a method fusion strategy, as illustrated in Figure 3-6.

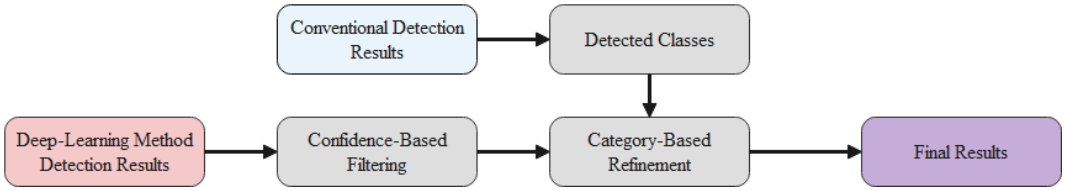

Figure 3-6 Method Fusion Workflow for Final Detection Results

The fusion process begins with confidence-based filtering, where low-confidence detections from deep learning models are discarded. This is followed by category-based refinement, which ensures that final classifications align with those obtained from traditional methods. The integrated results are rechecked for spatial and categorical consistency to ensure robustness.
By combining the strengths of both methods, the fusion strategy enhances the overall accuracy and reliability of the detection module, ensuring robust performance across diverse and challenging conditions.

3.5 Results and Visualization

3.5.1 Quantitative Results

The detection performance of different fusion methods is summarized in Table 3-1. The updated fusion approach, which incorporates deep learning-based detection, demonstrates a significant improvement in COCO mAP [21] scores.

Table 3-1 Quantitative Results of Different Fusion Methods

Method	Modality	COCO mAP
Conventional Fusion	LiDAR + Camera	0.0556
Updated Fusion with Deep Learning	LiDAR + Camera	11.5011

3.5.2 Qualitative Visualization

Figure 3-7 Sample Detection Results from Two Camera Perspectives

Detection results from two perspectives are visualized to demonstrate the effectiveness of the system. The LiDAR point clouds are mapped onto the corresponding camera images, and object classifications are indicated using color-coded points:

- Green Points: Vehicles
- Gray Points: Trucks
- Blue Points: Pedestrians

Bounding boxes are projected accurately onto the image planes, aligning seamlessly with the detected objects. These visualizations, as shown in Figure 3-7, illustrate the system's capability to integrate multi-modal data and classify objects reliably under diverse environmental conditions.

3.6 Integration with MOT

The outputs of the detection module are seamlessly integrated with the Multi-Object Tracking (MOT) system to provide continuous and robust trajectories for VRUs and vehicles [22]. This integration is critical for maintaining temporal consistency in object detection and improving downstream tasks such as path prediction.

3.6.1 Detection-to-Tracking Data Flow

The detection module generates 2D and 3D bounding boxes for each identified object. To enable efficient tracking, the central points of these bounding boxes are extracted and formatted as inputs for the MOT system. These central points are calculated as follows:

For 2D detections:

The central point ***(cx,cy)*** of a bounding box is given by:

where ***xmin,xmax,ymin,ymax*** are the coordinates of the bounding box edges.

$$\mathbf{c_x} = \frac{\mathbf{x_{min}} + \mathbf{x_{max}}}{\mathbf{2}}, \mathbf{c_y} = \frac{\mathbf{y_{min}} + \mathbf{y_{max}}}{\mathbf{2}}$$

For 3D detections:

The central point ***(cx,cy,cz)*** of a 3D bounding box is similarly computed, incorporating the z-axis dimension:

$$\mathbf{c_x} = \frac{\mathbf{x_{min}} + \mathbf{x_{max}}}{\mathbf{2}}, \mathbf{c_y} = \frac{\mathbf{y_{min}} + \mathbf{y_{max}}}{\mathbf{2}}, \mathbf{c_z} = \frac{\mathbf{z_{min}} + \mathbf{z_{max}}}{\mathbf{2}}$$

3.6.2 Role in MOT and Improved Accuracy

The central points derived from the detection module play a pivotal role in initializing and updating the Multi-Object Tracking (MOT) system, specifically the Kalman Filter-Constant Velocity (KF-CV) tracker. These points act as the initial state estimates for new object tracks, allowing the MOT framework to establish robust trajectories from the outset. For existing tracks, the central points are matched with predicted positions using data association algorithms, such as the Hungarian method, to ensure precise correspondence. The Kalman filter utilizes these matched points to update the object states, refining their positions and velocities over time.

This integration significantly enhances the accuracy and consistency of the MOT system. By providing continuous updates, the detection module helps maintain temporal consistency, ensuring that objects remain reliably tracked across frames even under challenging conditions like occlusions or partial detections. The use of central points reduces track fragmentation, improving trajectory coherence and stability. Moreover, the fusion of central points from both 2D and 3D detection results bolsters the spatial alignment of tracked objects, enabling the MOT system to operate effectively in complex, multi-modal environments. This seamless integration strengthens the foundation for downstream tasks such as trajectory prediction and collision avoidance.

4 Multi-Object Tracking (MOT)

The Multi-Object Tracking (MOT) module is a crucial component of our system, bridging the gap between detection and path prediction. By maintaining continuous trajectories of detected objects, the MOT module provides accurate and reliable input for downstream processes. This module leverages the Kalman Filter-Constant Velocity (KF-CV) method [15] for tracking, which ensures robust performance in dynamic and complex environments.

4.1 Algorithm Overview

The MOT module employs a KF-CV-based framework, as illustrated in Figure 4-1, to track multiple objects simultaneously. The algorithm integrates constant velocity motion modeling, data association, and track lifecycle management.

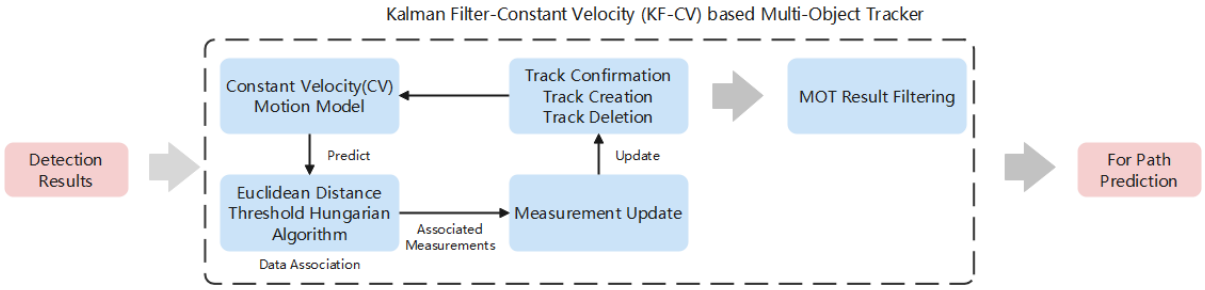

Figure 4-1 KF-CV-Based Framework for Multi-Object Tracking (MOT)

1. Constant Velocity (CV) Motion Model:

The CV model assumes that each object's velocity remains constant between time steps, making it suitable for real-time tracking in traffic scenarios. The state vector **x=[x,y,z,vx,vy,vz]** represents the object's position and velocity in 3D space. The state transition matrix $\boldsymbol{F}$ incorporates the time step $\boldsymbol{\Delta t}$, enabling prediction of future states:

$$x_{k|k-1} = Fx_{k-1|k-1} + w,$$

where w represents process noise. This prediction step estimates the object's next position, providing the basis for data association.

2. Data Association and Management:

The algorithm uses the Hungarian method to associate detections with existing tracks, minimizing the Euclidean distance between predicted states and measurements. A gating threshold ensures that only plausible associations are considered, reducing false matches. Tracks are confirmed, created, or deleted based on their consistency and temporal persistence.

3. Measurement Update:

Once associations are established, the Kalman filter updates the object's state using new measurements. The measurement vector $\boldsymbol{z}$ includes the observed position, and the Kalman gain $\boldsymbol{K}$ adjusts the state estimate:

$$x_{k|k} = Fx_{k|k-1} + K(z_k - Hx_{k|k-1}),$$

where ***H*** maps the state vector to the measurement space. This step refines the trajectory and accounts for noise in the detections.

4. Track Lifecycle Management:

Tracks are managed through three operations:

a. **Track Confirmation**: New tracks are confirmed after a minimum number of consistent detections.
b. **Track Creation**: Unassociated detections are initialized as new tracks.
c. **Track Deletion**: Tracks with insufficient updates over a predefined duration are terminated to prevent clutter.

5. Result Filtering:
After track updates, filtering is applied to consolidate fragmented trajectories. This ensures that only complete and reliable tracks are passed to the path prediction module.

4.2 Noise Handling and Optimization

Real-world scenarios often involve noisy detections due to occlusions, sensor inaccuracies, or overlapping objects. The MOT module incorporates several techniques to handle noise and improve tracking accuracy:

1. Consistency Checks:

The MOT system continuously validates the plausibility of each track by performing spatio-temporal consistency checks. These checks analyze an object's position, velocity, and acceleration across frames, ensuring they remain within realistic ranges. For instance, sudden shifts in position or drastic changes in velocity that exceed predefined thresholds trigger a re-evaluation of the track. This re-evaluation involves cross-verifying the track against recent detections and recalibrating its state using neighboring frames if inconsistencies are detected. By maintaining strict adherence to physically plausible motion patterns, consistency checks prevent the propagation of erroneous trajectories and improve the system's overall stability.

2. Fragment Integration:

Temporary occlusions, such as when a VRU is obscured by a passing vehicle, can lead to fragmented tracks, resulting in trajectory discontinuities. To mitigate this, the MOT module merges fragmented tracks by evaluating their spatial and temporal proximity. If two tracks are found to overlap in position and time within a predefined tolerance, they are integrated into a single continuous track. This process considers factors such as the velocity and direction of motion to ensure that the merging decision is accurate and avoids mismatches. Fragment integration not only improves trajectory continuity but also reduces false track terminations, ensuring reliable inputs for downstream path prediction.

3. Adaptive Gating:

Data association is critical for tracking, and the gating threshold determines which detections are matched to existing tracks. The MOT module employs an adaptive gating strategy, dynamically adjusting the threshold based on an object's velocity and detection uncertainty. For high-speed objects, a larger gating threshold accommodates potential prediction errors due to rapid movement, while for slow-moving or stationary objects, a tighter threshold minimizes false matches. Uncertainty metrics, such as the covariance of the Kalman filter, are also factored into the gating adjustment, allowing the system to adapt to varying detection quality. This approach ensures robust associations, reducing both missed detections and false positives.

4. Trajectory Smoothing:

Erratic movements, particularly common among VRUs like pedestrians and cyclists, can introduce jitter into tracked paths. The MOT module applies a temporal smoothing algorithm during post-processing to address this issue. Smoothing techniques, such as Kalman filter refinements or low-pass filters, are used to average position and velocity changes across consecutive frames. For example, small oscillations caused by sensor noise are dampened, resulting in smoother and more natural trajectories. This refinement step is critical for enhancing the quality of tracking outputs, making them more reliable for tasks such as trajectory prediction and collision avoidance.

4.3 Results

The performance of the MOT module was evaluated using a validation dataset, focusing on key metrics such as MOTA (Multiple Object Tracking Accuracy), MOTP (Multiple Object Tracking Precision), HOTA (Higher Order Tracking Accuracy), and IDF1 (ID Matching Accuracy). The evaluation revealed significant variations in tracking performance across different subclasses of road users.

The HOTA values were notably higher for passenger vehicles and cyclists, reflecting reliable and consistent tracking for these categories. In contrast, the system exhibited lower performance for VRUs such as children and wheelchair users, likely due to the lack of reliable and accurate detections for these challenging cases. The detailed results are summarized in Table 4-1**.**

Table 4-1 Performance Metrics for MOT

Subclass	MOTA	MOTP	IDF1	HOTA	Mean IoU
Passenger_Vehicle	44.64	24.23	66.19	78.92	24.23
VRU_Adult_Using_Non-Motorized_Device/Prop_Other	42.22	35.10	54.89	61.76	35.10
VRU_Adult_Using_Bicycle	60.18	27.83	62.25	68.83	27.83

VRU_Adult	43.71	30.58	54.74	60.98	30.58
VRU_Child	18.18	22.25	27.71	37.97	22.25
VRU_Adult_Using_Wheelchair	13.52	37.59	22.46	33.94	37.59
VRU_Adult_Using_Scooter_or_Skateboard	42.84	39.52	56.01	62.53	39.52

These results demonstrate the system's strengths in tracking well-defined and consistently detectable objects, such as vehicles and bicycles, while highlighting the need for further refinement in handling less detectable VRUs.

4.4 Integration with Path Prediction

The MOT module outputs continuous, noise-reduced trajectories for each tracked object, forming the backbone of the path prediction module. By providing accurate and temporally consistent motion states, the MOT system ensures reliable inputs for forecasting future positions and interactions.

The integration process involves transmitting the central points of detected objects' bounding boxes, along with their motion states, to the path prediction module:

- **Central Point Calculation:** The central points of 2D and 3D bounding boxes are computed as the geometric mean of the bounding box coordinates. These points represent the precise locations of objects in the spatial domain.
- **Trajectory Continuity:** By leveraging Kalman filter updates, the MOT system provides smooth and consistent trajectories, minimizing discontinuities caused by occlusions or partial detections.

This integration significantly enhances the accuracy of path prediction by supplying high-quality, validated trajectories. It ensures that predicted paths align with actual object movements, reducing errors in downstream tasks such as collision prediction and warning systems.

Figure 4.1 provides a visual representation of the KF-CV-based [15] MOT framework, illustrating the data flow from detection to tracking and its subsequent use in the path prediction module.

5 Path Prediction

The Path Prediction module is critical for forecasting the trajectories of road users, including vehicles and vulnerable road users (VRUs), in complex and dynamic environments. It leverages the enhanced Social GAN framework, integrating dynamic map information such as traffic lights and incorporating specialized adaptations for high-interaction scenarios. This section details the methodology, emphasizing the role of specialized modules and the model's enhancements for urban traffic scenarios.

Generative Adversarial Nets (GANs)[23] are a generative framework consisting of a generator ***(G)*** and a discriminator ***(D).*** The generator learns to produce synthetic future trajectories, while the discriminator evaluates whether a trajectory is "real" (from the dataset) or "fake" (generated). This adversarial process trains the generator to create realistic, plausible trajectories.

The generator maps observed past trajectories, $\boldsymbol{X_{obs}} = [\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_T}]$, into future trajectories, $\boldsymbol{X_{pred}} = [\boldsymbol{x_{T+1}}, \boldsymbol{x_{T+2}}, \ldots, \boldsymbol{x_{T'}}]$. The discriminator evaluates both $\boldsymbol{X_{obs}}$and $\boldsymbol{X_{pred}}$, encouraging the generator to output trajectories that align with real-world movement patterns. The GAN optimization problem is formulated as:

$$\min_{G} \max_{D} \mathbb{E}_{X_{real} \sim p_{data}} [\log D(X_{real})] + \mathbb{E}_{X_{real} \sim G} [\log 1 - D(X_{real})]$$

5.1 Enhanced Social GAN Model

5.1.1 Generative Adversarial Networks (GANs) for Trajectory Prediction

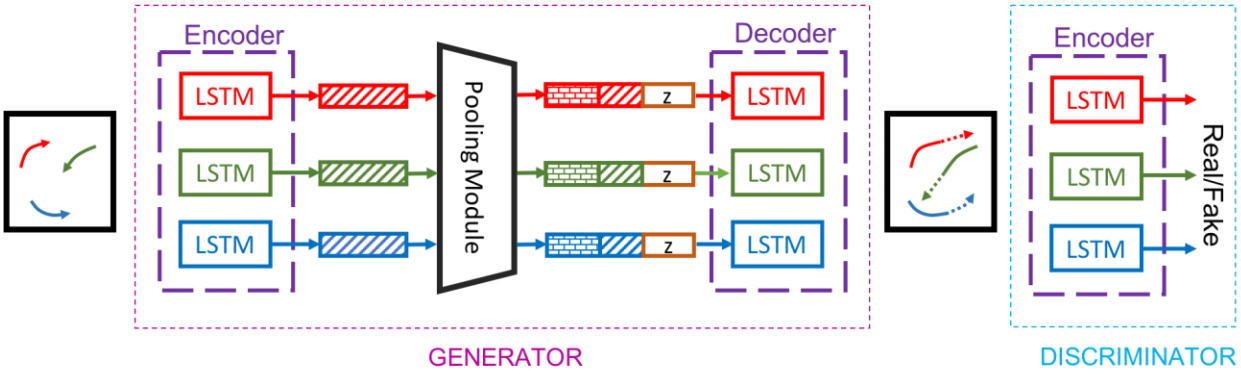

Figure 5-1. Social GAN [10] system overview.

5.1.2 Social GAN for Interaction-Aware Predictions

The Social GAN [10] model enhances the standard GAN by incorporating a pooling module to account for inter-agent interactions. The architecture consists of three key components:

1. Generator: The generator employs an encoder-decoder structure:

- **Encoder:** Observed trajectories are encoded using an LSTM network to generate hidden states, $\boldsymbol{h_t}$, representing the agents' motion histories.

- **Pooling Module:** Interaction information is aggregated across agents using max-pooling, summarizing the spatial and temporal relationships between agents. The pooled context vector c_i for each agent i is computed as:

$$c_i = \max_{j \neq i} f(h_j)$$

where $f(\cdot)$ transforms the hidden states into interaction-aware representations.

- **Decoder:** The decoder generates future trajectories based on the pooled context and the agent's hidden state.

2. Discriminator: The discriminator distinguishes between real and generated trajectories. It evaluates the combined sequence of observed and predicted trajectories to ensure social norms and interaction rules are respected.

3. Variety Loss: To address the inherent uncertainty in trajectory prediction, the model incorporates a variety loss, encouraging the generator to produce diverse plausible paths:

$$L_{variety} = \min_{k=1} ||X_{pred}^{k} - X_{gt}||,$$

where K is the number of generated trajectories, and X_{gt} is the ground truth trajectory.

5.2 Specialized Modules

5.2.1 Traffic Light Encoding

Traffic lights significantly influence the movement patterns of vehicles and pedestrians in urban scenarios. The model integrates traffic light states as dynamic map information, augmenting input trajectories with a binary encoding (st) for the current traffic light state at each time step. The augmented input becomes:

$$X_{obs} = [x_t, s_t]_{t=1}^{T}.$$

This encoding enables the model to adapt predictions based on traffic flow dynamics, particularly at intersections.

5.2.2 Adaptation to High-Interaction Scenarios

High-interaction environments, such as intersections with multiple agents, require robust handling of agent interactions. The enhanced Social GAN incorporates the following adaptations:

1. Localized Interaction Focus:

The pooling module emphasizes interactions between proximate agents using a distance-based weighting:

$$c_i = \sum_{j \neq i} \alpha_{ij} h_j \, , \alpha_{ij} = \frac{exp(-||x_i - x_j||^2)}{\sum_{k \neq i} exp(-||x_i - x_k||^2)}$$

2. Multi-Modal Output:

The generator produces multiple plausible trajectories, each with an associated confidence score, enabling robust predictions under uncertainty:

$$[(X_{pred}^{k}, p^{k})]_{k=1}^{K}, p^{k} = \frac{\exp(-L_{veriety}^{k})}{\sum_{l=1}^{K} \exp(-L_{variety}^{l})}$$

3. Map Integration:

A vector map of the intersection is used to provide spatial constraints, including lane boundaries and curb locations. This additional context enhances the model's understanding of the environment and improves prediction accuracy.

5.3 Results

5.3.1 Quantitative Evaluation

The results, shown in Table 5-1, indicate strong performance based on Weighted Average Displacement Error (ADE) and Two-Level ADE scores.

Table 5-1 Overall Performance Metrics

Set	Weighted ADE	Overall Score
Test	2.4175	4.9248
Validation	2.3863	5.0833

5.3.2 Class-Specific Performance

The model performed better for VRUs than vehicles, as shown in Table 5.2, reflecting its strength in handling pedestrian and cyclist behaviors.

Table 5.2 Class-Specific Metrics

Set	Class	Weighted ADE
Test	Vehicle	6.2357
Test	VRU	3.6138
Validation	Vehicle	6.0541
Validation	VRU	3.7524

5.4 Integration and Performance

The Path Prediction module integrates seamlessly with upstream MOT outputs, taking observed trajectories as input and generating socially-aware predictions. The module's inclusion of traffic light encoding and pooling adaptations improves its performance in urban traffic scenarios, particularly for VRUs.

Quantitative evaluation demonstrates that the model achieves lower Average Displacement Error (ADE) for VRUs than vehicles, highlighting its strength in handling pedestrian and cyclist behaviors. Visualizations of predicted trajectories on vector maps validate the model's ability to capture complex interaction patterns.

The specialized modules ensure the model's robustness and adaptability across diverse urban scenarios, making it a valuable tool for collision prediction and traffic flow optimization.

6 Collision Prediction and Warning System

The Collision Prediction and Warning System (CPWS) is designed to predict potential conflicts between road users and provide timely warnings to mitigate accidents. This section elaborates on the methodology for collision prediction using Time-to-Collision (TTC) [16] analysis and the design of a multi-modal warning system that integrates roadside, vehicle, and pedestrian communication channels.

6.1 Collision Prediction

6.1.1 Time-to-Collision (TTC) Calculation Logic

Collision prediction is centered around the calculation of TTC, a widely used metric that estimates the time remaining before two road users collide based on their relative speed and distance. For each pair of road users, the following workflow is implemented:

6.1.1.1 Relative Distance and Speed Calculation:

The Euclidean distance between two road users (d_{rel}) is calculated as:

$$\mathrm{d_{rel}} = \sqrt{(\mathrm{x_2} - \mathrm{x_1})^2 + (\mathrm{y_2} - \mathrm{y_1})^2 + (\mathrm{z_2} - \mathrm{z_1})^2}$$

where (x_1,y_1,z_1) and (x_2,y_2,z_2) represent the 3D coordinates of the two users.
Relative speed (v_{rel}) is determined by projecting each user's velocity onto the connecting line, assuming a positive direction.

6.1.1.2 TTC Calculation

For scenarios with constant speed, the TTC is calculated as:

$$\mathrm{TTC} = \frac{\mathrm{d_{rel}}}{\mathrm{v_{rel}}},$$

provided $v_{rel}>0$. For constant acceleration, the TTC is derived by solving the quadratic equation:

$$d_{rel} + v_{rel}t + \frac{1}{2}a_{rel}t^2 = 0,$$

where a_{rel} is the relative acceleration.

6.1.1.3 Conflict Identification:

A conflict is flagged if the TTC value remains below a threshold (e.g., 1.5 seconds) for consecutive frames.
Dimensions of road users are optionally incorporated, refining the calculation by considering the physical size of the objects.

6.1.2 3D Position Time Series Analysis

In addition to TTC, the system evaluates motion patterns based on 3D position time series, identifying potential collisions through trajectory overlaps and proximity thresholds. This multi-faceted approach ensures robust conflict prediction across varying traffic scenarios.

6.2 Warning System Design

The warning system disseminates collision alerts to vehicles, VRUs, and nearby infrastructure using a combination of V2X communication and mobile connectivity, as illustrated in Figure 6-1.

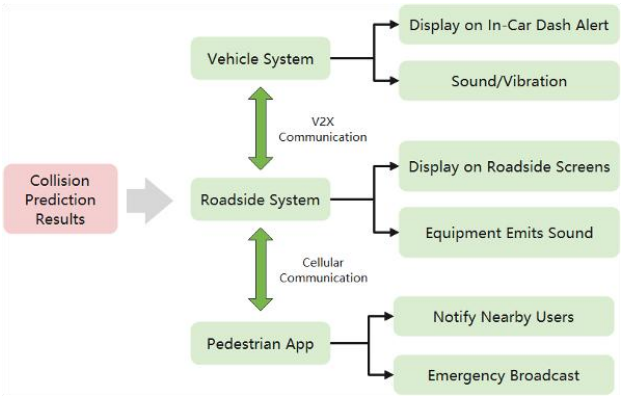

Figure 6-1 Multi-Modal Warning System Workflow

6.2.1 Roadside System

The roadside system serves as a central hub for collision alerts, integrating information from V2X-enabled vehicles and roadside sensors. Warning mechanisms include:

6.2.1.1 Visual Alerts:

Warnings are displayed on large roadside digital screens placed at strategic locations, such as intersections and pedestrian crossings. These visual alerts include detailed information about the type and location of the risk, using color codes and animations to enhance visibility. For example, a flashing red icon could indicate an imminent collision, while amber indicates potential risk. The visual design prioritizes simplicity and clarity to ensure all road users, including non-native speakers or those unfamiliar with the area, can quickly interpret the warnings.

6.2.1.2 Auditory Alerts:

Roadside speakers emit sound-based alerts to notify pedestrians and nearby vehicles. The auditory signals are designed to be distinct and context-specific, such as a continuous beep for general warnings or a sharp tone for high-priority risks. The volume dynamically adjusts based on ambient noise levels, ensuring audibility even in loud urban environments. Additionally, multilingual audio messages can be broadcast in areas with diverse populations, providing localized warnings to all users.

6.2.2 Vehicle-side System

V2X communication enables vehicles to receive collision alerts directly from the roadside system. The in-vehicle system responds by:

6.2.2.1 Dashboard Display

Warnings are shown on the vehicle's dashboard, providing clear and actionable information. For instance, the display may show a visual map highlighting the collision point relative to the vehicle, along with instructions such as "Reduce Speed" or "Change Lane." Advanced visualizations, such as augmented reality overlays on the windshield, can further enhance driver awareness by integrating alerts directly into the field of view.

6.2.2.2 Auditory/Vibration Alerts

Drivers are notified through sound and haptic feedback, such as vibrations in the steering wheel or driver's seat. Auditory alerts can include escalating tones or verbal messages like "Collision risk ahead," tailored to the urgency of the situation. Vibration patterns are designed to correspond to the direction of the risk, such as left-side vibrations for risks on the left, improving intuitive understanding.

6.2.3 Pedestrian-side System

Mobile applications for pedestrians integrate cellular communication to deliver personalized warnings:

6.2.3.1 Notifications:

Pedestrians receive alerts on their smartphones via dedicated safety applications. Notifications provide location-specific details, such as "High-speed vehicle approaching" or "Crosswalk blocked ahead." The app also includes visual aids like augmented reality overlays, allowing users to see potential collision paths superimposed on their camera view.

6.2.3.2 Emergency Broadcasts:

Pedestrians receive alerts on their smartphones via dedicated safety applications. Notifications provide location-specific details, such as "High-speed vehicle approaching" or "Crosswalk blocked ahead." The app also includes visual aids like augmented reality overlays, allowing users to see potential collision paths superimposed on their camera view.

6.2.4 Integrated Warning Workflow:

The CPWS operates through a highly integrated and seamless workflow that combines advanced prediction methodologies with multi-modal alert dissemination to ensure comprehensive safety coverage for all road users.

The workflow begins with Collision Prediction, where the roadside system processes incoming data from V2X-enabled vehicles, roadside sensors, and pedestrian devices. Using Time-to-Collision (TTC) calculations and trajectory-based analysis, the system identifies potential conflicts between road users. The prediction process incorporates real-time 3D position data, velocity, and acceleration to ensure precise and reliable conflict detection. This stage is further enhanced by the inclusion of environmental context, such as traffic light states and road geometry, improving the accuracy of predictions in complex urban scenarios.

Following collision prediction, the system transitions to Alert Generation, where the identified risks are categorized and contextualized based on their type, urgency, and the specific characteristics of the recipients. Alerts are tailored to the needs of each target group—drivers, pedestrians, and roadside infrastructure operators. For instance, high-priority warnings, such as imminent collisions, are accompanied by more urgent signals, including flashing icons, loud auditory alerts, or steering wheel vibrations for drivers. Conversely, less critical warnings are delivered with moderate visual or auditory cues, ensuring an appropriate level of response while avoiding unnecessary alarm.

The final stage, Alert Dissemination, ensures that the generated alerts reach all relevant users through multiple communication channels. The CPWS leverages V2X communication to relay warnings directly to vehicles, allowing for immediate in-car displays and auditory notifications. Simultaneously, roadside systems, such as digital screens and loudspeakers, provide visual and auditory alerts for pedestrians and other road users. Mobile applications and cellular networks extend the system's reach to pedestrians, offering personalized notifications and emergency broadcasts to users in the vicinity. This multi-channel approach ensures comprehensive coverage, even in blind spots or non-line-of-sight scenarios, enhancing the overall effectiveness of the system.

By seamlessly integrating these components, the CPWS delivers a robust and proactive safety solution, reducing the likelihood of collisions and enhancing situational awareness for all road users in real-time.

6.3 Integration and Effectiveness

The combination of TTC-based collision prediction and the multi-modal warning system enhances traffic safety by providing timely and precise alerts. By leveraging advanced communication technologies such as V2X and cellular networks, the CPWS ensures that all road users are informed of potential conflicts, significantly reducing the likelihood of accidents.

If further technical details or evaluation results are required, they can be integrated upon request.

7 Discussion

7.1 Summary of Up-to-Date Achievements

The methodologies and systems presented in this report address the critical challenges associated with detecting and protecting Vulnerable Road Users (VRUs) in complex traffic environments. By integrating multi-sensor data from LiDAR and cameras, leveraging advanced algorithms like DAB-DETR and Social GAN, and incorporating robust tracking and collision prediction frameworks, the system demonstrates the potential to significantly improve road safety.
Key achievements include:

- **Enhanced Detection:** The hybrid detection pipeline, combining traditional and deep learning-based methods, significantly improved accuracy, particularly in adverse weather and low-light conditions.
- **Reliable Tracking:** The Kalman Filter-Constant Velocity (KF-CV) based MOT system provided temporally consistent trajectories, mitigating challenges like occlusions and fragmentation.
- **Accurate Path Prediction:** The Social GAN framework excelled in modeling interaction-aware trajectories, achieving lower Average Displacement Error (ADE) for VRUs compared to vehicles.
- **Proactive Warnings:** The Collision Prediction and Warning System (CPWS) effectively disseminated real-time alerts to road users, enhancing situational awareness and reducing collision risks.

While the results are promising, challenges such as handling VRUs under diverse environmental conditions and improving system scalability remain areas for further exploration.

7.2 Next Steps

To build upon the current system and address existing limitations, the following steps are proposed:

7.2.1 Expansion at Riverside Smart Intersection

To improve system coverage and data quality, we plan to expand the sensor network by deploying two additional LiDAR and two additional cameras at the Riverside Smart Intersection, i.e., Iowa Ave. & University Ave. This will enhance the system's ability to detect and track VRUs and vehicles in complex and high-traffic scenarios.

7.2.2 Data Collection and Validation

Comprehensive data collection campaigns will be conducted using the expanded sensor setup. The collected data will be used to validate the current pipeline, including detection, tracking, path prediction, and warning systems. This step will provide valuable insights into the system's real-world performance under various traffic and environmental conditions.

7.2.3 Algorithm Refinement and Robustness Improvement

Based on validation results, existing algorithms will be refined to address identified weaknesses. Specifically, we aim to:

(1) Improve detection and tracking robustness in challenging conditions such as varying lighting and adverse weather.
(2) Design adaptive algorithms capable of handling dynamic environmental changes and sensor noise.
(3) Incorporate advanced data augmentation techniques to enhance the training dataset and improve generalization.

7.2.4 End-to-End System Development:

A fully integrated pipeline will be established, seamlessly connecting detection, tracking, prediction, and warning modules. This comprehensive system will ensure that all components work harmoniously, delivering real-time, accurate, and actionable outputs to enhance road safety for all users.

8 Conclusions

This report presents a comprehensive framework for enhancing the safety of vulnerable road users (VRUs) in urban traffic environments. By integrating advanced detection, tracking, trajectory prediction, and collision warning systems, the proposed methodology addresses critical challenges such as occlusions, environmental variability, and multi-agent interactions. The hybrid detection pipeline, combining traditional and deep learning approaches, demonstrated significant improvements in accuracy under diverse conditions. Meanwhile, the Kalman Filter-based MOT and Social GAN-based path prediction modules provided robust and reliable performance, ensuring that the system can operate effectively in complex traffic scenarios.

Looking forward, the planned hardware expansions, extensive data validation efforts, and algorithmic refinements will further enhance the system's robustness and adaptability. By establishing a fully integrated end-to-end pipeline, the proposed framework aims to deliver scalable and reliable solutions for VRU detection and protection, setting the foundation for safer, smarter, and more inclusive transportation systems.

References

[1] M. Goldhammer, S. Köhler, S. Zernetsch, K. Doll, B. Sick, and K. Dietmayer, "Intentions of vulnerable road users—Detection and forecasting by means of machine learning," *IEEE transactions on intelligent transportation systems,* vol. 21, no. 7, pp. 3035-3045, 2019.

[2] W. van Haperen, M. S. Riaz, S. Daniels, N. Saunier, T. Brijs, and G. Wets, "Observing the observation of (vulnerable) road user behaviour and traffic safety: A scoping review," *Accident Analysis & Prevention,* vol. 123, pp. 211-221, 2019.

[3] S. K. Maurya and A. Choudhary, "Deep learning based vulnerable road user detection and collision avoidance," in *2018 IEEE International Conference on Vehicular Electronics and Safety (ICVES)*, 2018: IEEE, pp. 1-6.

[4] K. Dasgupta, A. Das, S. Das, U. Bhattacharya, and S. Yogamani, "Spatio-contextual deep network-based multimodal pedestrian detection for autonomous driving," *IEEE transactions on intelligent transportation systems,* vol. 23, no. 9, pp. 15940-15950, 2022.

[5] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE transactions on pattern analysis and machine intelligence,* vol. 34, no. 4, pp. 743-761, 2011.

[6] C. Wei, G. Wu, and M. J. Barth, "Feature Corrective Transfer Learning: End-to-End Solutions to Object Detection in Non-Ideal Visual Conditions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 23-32.

[7] D. J. Yeong, G. Velasco-Hernandez, J. Barry, and J. Walsh, "Sensor and sensor fusion technology in autonomous vehicles: A review," *Sensors,* vol. 21, no. 6, p. 2140, 2021.

[8] J. Kocić, N. Jovičić, and V. Drndarević, "Sensors and sensor fusion in autonomous vehicles," in *2018 26th Telecommunications Forum (TELFOR)*, 2018: IEEE, pp. 420-425.

[9] Z. Sun, J. Chen, L. Chao, W. Ruan, and M. Mukherjee, "A survey of multiple pedestrian tracking based on tracking-by-detection framework," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 31, no. 5, pp. 1819-1833, 2020.

[10] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social gan: Socially acceptable trajectories with generative adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2255-2264.

[11] A. Mukhtar, L. Xia, and T. B. Tang, "Vehicle detection techniques for collision avoidance systems: A review," *IEEE transactions on intelligent transportation systems,* vol. 16, no. 5, pp. 2318-2338, 2015.

[12] P. Teixeira, S. Sargento, P. Rito, M. Luís, and F. Castro, "A sensing, communication and computing approach for vulnerable road users safety," *IEEE Access,* vol. 11, pp. 4914-4930, 2023.

[13] P. Merdrignac, O. Shagdar, and F. Nashashibi, "Fusion of perception and V2P communication systems for the safety of vulnerable road users," *IEEE Transactions on Intelligent Transportation Systems,* vol. 18, no. 7, pp. 1740-1751, 2016.

[14] J. J. Anaya, E. Talavera, D. Giménez, N. Gómez, F. Jiménez, and J. E. Naranjo, "Vulnerable road users detection using v2x communications," in *2015 IEEE 18th international conference on intelligent transportation systems*, 2015: IEEE, pp. 107-112.

[15] Y. Chan, A. Hu, and J. Plant, "A Kalman filter based tracking scheme with input estimation," *IEEE transactions on Aerospace and Electronic Systems,* no. 2, pp. 237-244, 1979.
[16] J. C. Hayward, "Near miss determination through use of a scale of danger," 1972.
[17] S. Liu *et al.*, "Dab-detr: Dynamic anchor boxes are better queries for detr," *arXiv preprint arXiv:2201.12329,* 2022.
[18] M. Zhang, "Use density-based spatial clustering of applications with noise (DBSCAN) algorithm to identify galaxy cluster members," in *IOP conference series: earth and environmental science*, 2019, vol. 252, no. 4: IOP Publishing, p. 042033.
[19] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 12697-12705.
[20] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4490-4499.
[21] T.-Y. Lin *et al.*, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 2014: Springer, pp. 740-755.
[22] D. M. Jiménez-Bravo, Á. L. Murciego, A. S. Mendes, H. S. San Blás, and J. Bajo, "Multi-object tracking in traffic environments: A systematic literature review," *Neurocomputing,* vol. 494, pp. 43-55, 2022.
[23] I. Goodfellow *et al.*, "Generative adversarial nets," *Advances in neural information processing systems,* vol. 27, 2014.