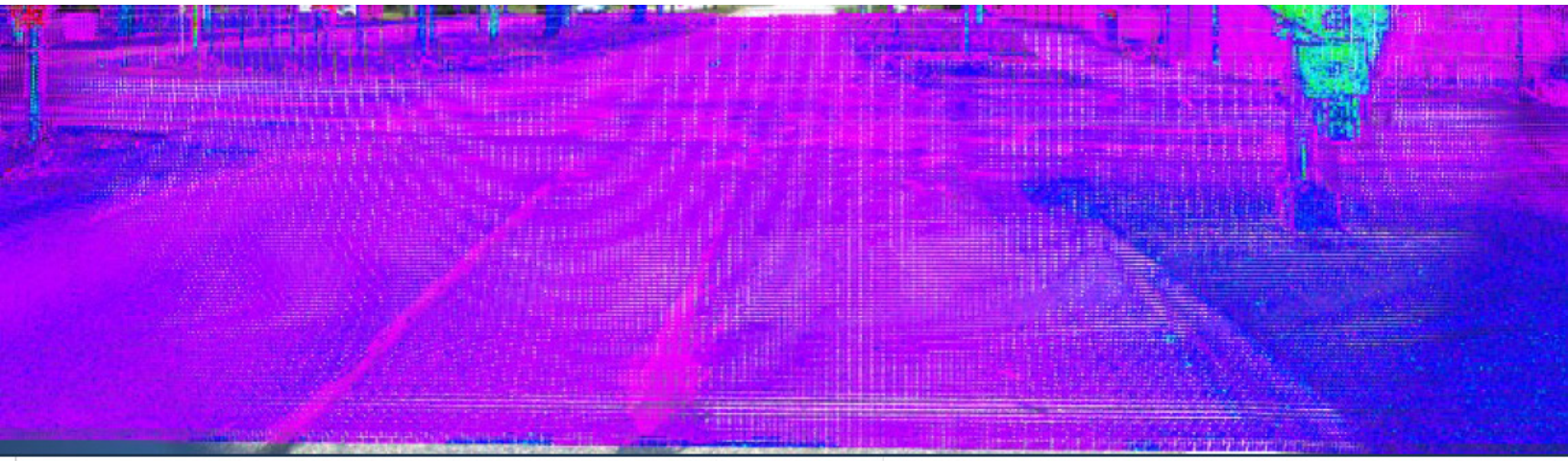
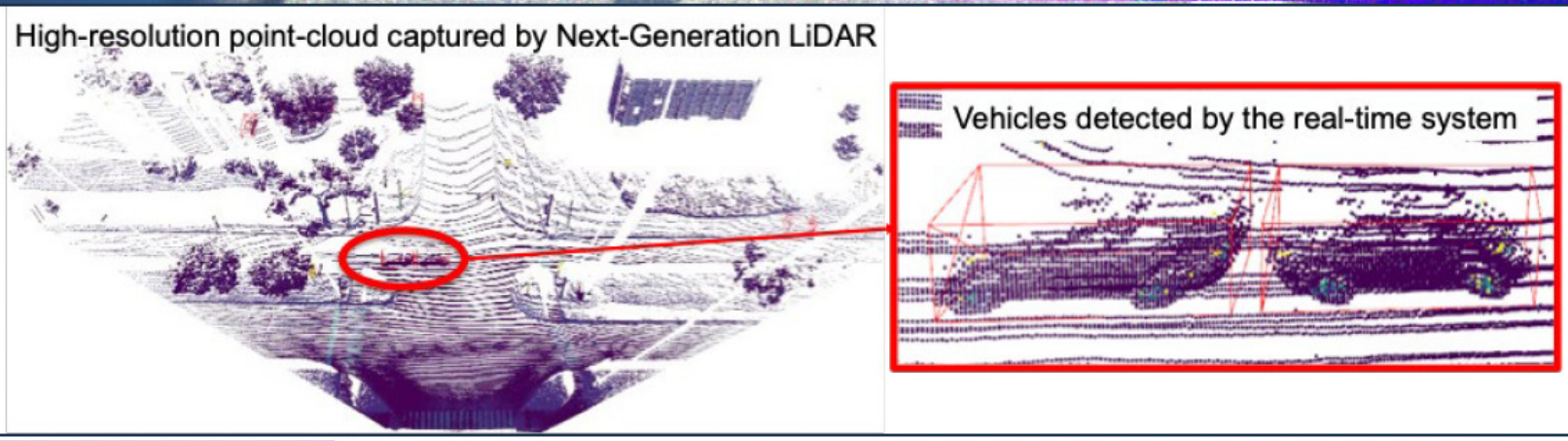
SJSU SAN JOSÉ STATE UNIVERSITY

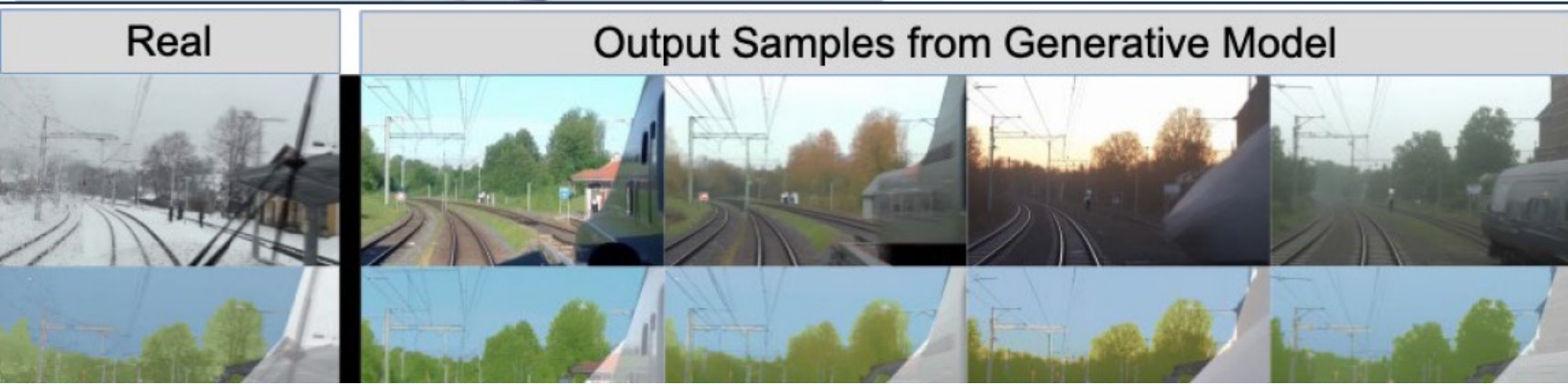
MTI MINETA TRANSPORTATION INSTITUTE

Multimodal Solutions for the Detection of Safety-Related Railroad Objects under Adverse Weather: Architectures, Proof-of-Concepts, and Performance Results

Hayder Radha, PhD







MICHIGAN STATE UNIVERSITY

MINETA TRANSPORTATION INSTITUTE

Founded in 1991, the Mineta Transportation Institute (MTI), an organized research and training unit in partnership with the Lucas College and Graduate School of Business at San José State University (SJSU), increases mobility for all by improving the safety, efficiency, accessibility, and convenience of our nation's transportation system. Through research, education, workforce development, and technology transfer, we help create a connected world. MTI leads the California State University Transportation Consortium (CSUTC) funded by the State of California through Senate Bill 1 and the Rail Resilience to Severe Weather: Training and Research Program (RR2SW) funded by the Federal Railroad Administration. MTI focuses on three primary responsibilities:

Research

MTI conducts multi-disciplinary research focused on surface transportation that contributes to effective decision making. Research areas include: active transportation; planning and policy; security and counterterrorism; sustainable transportation and land use; transit and passenger rail; transportation engineering; transportation finance; transportation technology; and workforce and labor. MTI research publications undergo expert peer review to ensure the quality of the research.

Education and Workforce Development

To ensure the efficient movement of people and goods, we must prepare the next generation of skilled transportation professionals who can lead a thriving, forward-thinking transportation industry for a more connected world. To help achieve this, MTI sponsors a suite of workforce development and education opportunities. The Institute supports educational programs offered by the Lucas Graduate School of Business: a Master of Science in Transportation Management, plus graduate certificates that include High-Speed and Intercity Rail Management and Transportation Security Management. These flexible programs offer live online classes so that working transportation professionals can pursue an advanced degree regardless of their location.

Information and Technology Transfer

MTI utilizes a diverse array of dissemination methods and media to ensure research results reach those responsible for managing change. These methods include publication, seminars, workshops, websites, social media, webinars, and other technology transfer mechanisms. Additionally, MTI promotes the availability of completed research to professional organizations and works to integrate the research findings into the graduate education program. MTI's extensive collection of transportation-related publications is integrated into San José State University's world-class Martin Luther King, Jr. Library.

Report 26-28

Multimodal Solutions for the Detection of Safety-Related Railroad Objects under Adverse Weather: Architectures, Proof-of-Concepts, and Performance

Hayder Radha, PhD

September 2026

A publication of the
Mineta Transportation Institute
Created by Congress in 1991

College of Business
San José State University
San José, CA 95192-0219

DOI: 10.31979/mti.2026.2412

Mineta Transportation Institute
College of Business
San José State University
San José, CA 95192-0219

Email: mineta-institute@sjsu.edu

transweb.sjsu.edu/research/2412

This research was supported by the Federal Railroad Administration (FRA) through Consolidated Rail Infrastructure and Safety Improvements (CRISI) Grant No. 69A36523420190CRSCA.

Acknowledgments

The MSU team would like to express their sincere gratitude to Dr. Hilary Nixon, the Deputy Executive Director of the Mineta Transportation Institute at San José State University for her outstanding leadership and guidance throughout this project.

Contents

Acknowledgements iii
Executive Summary 1
1. Introduction 3
1.1 Background 3
1.2 Objectives 3
1.3 Overall Approach 4
1.3.1 Availability of Visual and 3D Lidar Point-Cloud Data 4
1.3.2 A Three-Prong Approach 4
1.4 Scope 6
1.5 Organization of the Report 6
2. Real-time Object Detectors' Performance Results using Publicly Available Datasets 7
2.1 Performance Results 8
2.2 SPARK Deep-Learning Architecture 10
3. Weather-Resilient Object Detectors' Performance Results using the MSU-Four Seasons Dataset 12
4. Next-Generation Lidar Hardware and Software Integration and Performance Testing 15
5. Cooperative Perception using Connected Vehicle Technologies 18
5.1 Faster HEAL: A New Framework for Cooperative Perception 19
6. Ongoing Research Effort and Future Work 25
6.1 Generative Models for Railroad Data 25
6.2 Comprehensive Dataset Collection and Annotation using Next-Generation Lidar-based System 26
6.3 Multimodal Domain-Adaptation 26
6.4 Cooperative Perception under Challenging Weather 26
7. Conclusion 28
8. References 29
Appendix A. A Generative Framework for Rail-Domain Perception Data 44

Illustrations

Figure 1. The three main subareas pursued by the MSU team for the development of real-time object detection methods for operation under adverse weather conditions 5

Figure 2. The speed (FPS) and the maximum detection accuracy gain, in terms of AP or AP with Heading (APH) achieved due to the proposed SPARK framework among different detectors, classes, and datasets 8

Figure 3. The tradeoff between detection accuracy and speed for the 3D object detectors evaluated in this study 10

Figure 4. Architecture and training pipeline of SPARK 11

Figure 5. Train-crossing images from the MSU-Four Seasons dataset 12

Figure 6. A multimodal lidar-camera scene from the MSU Four Seasons dataset captured under snowy winter conditions 13

Figure 7. (Top left) high-resolution 3D point-cloud frame of a railroad-crossing scene captured by a Next Generation (NG) lidar used by the MSU team. (Top right) The fusion of the lidar point-cloud frame with the corresponding RGB camera frame. (Bottom) A zoomed-in version of the same scene illustrating the high-resolution features captured by the lidar to the extent that one can read the text “STOP” on the traffic sign at the railroad crossing 15

Figure 8. Point-cloud 3D lidar frames captured by different lidar sensors at different temporal frame ratesof the same location around a railroad crossing.....16

Figure 9. High-resolution point-cloud frame captured by the Luminar lidar and the results for applying a SPARK-based PointPillars object detector onto the scene 17

Figure 10. Overview of Faster-HEAL 20

Figure 11. An example of synthetic rail domain data generated by the proposed JACL generative model.....26

Tables

Table 1. Object detection AP performance of four low-complexity real-time object detectors (*students*) with (and without) the proposed SPARK framework in conjunction with three high-complexity object detectors (*teachers*). 9

Table 2. The impact of the WISDOM framework on the Average Precision (AP) performance of two popular object detectors (Voxel-RCNN and PV-RCNN++) when tested under (a) Snowy and (b) Rainy conditions 14

Table 3. The performance impact of cooperative perception in terms of AP for different object detectors and sensor configuration models using the homogeneous phase of HEAL and the proposed Faster-HEAL frameworks 21

Table 4. Comparison of Faster-HEAL with other state-of-the-art cooperative perception methods including the original HEAL framework 24

Executive Summary

A team of researchers in the College of Engineering at Michigan State University (MSU) in East Lansing, Michigan developed proof-of-concepts of multi sensor-modalities (multimodal) solutions for real-time detection of safety-related railroad events and objects. The team conducted the research between September 2023 to March 2025.

The MSU team developed key components of multimodal solutions that rely on advanced deep learning methods and utilize object detection approaches to demonstrate the feasibility of real-time detection of objects and events that could represent major safety issues for moving trains. They developed and integrated different software and hardware platforms that are based on state-of-the-art sensor technologies and deep learning solutions. Detected objects could be vehicles, people, or both trespassing on a railroad track, or they could be hazardous objects such as falling rocks obstructing a moving train. The team developed multiple deep-learning based software algorithms that demonstrate the feasibility of a robust real-time railroad object detection safety system. The project also integrated automotive-grade lidars and cameras, which are used by Autonomous Vehicles (AVs) and Advanced Driver Assistance Systems (ADAS), with deep learning-based object detection solutions to demonstrate the feasibility of achieving high levels of object detection accuracy and robustness under challenging weather conditions. Using automotive-grade sensor technologies can lead to significant savings in cost for the railroad industry while enabling the deployment of accurate real-time detection solutions that have been improving over the past decade, and which continue to improve as new sensors, deep learning architectures, and other AI solutions emerge.

This report covers the design options of the corresponding multimodal solutions that were considered by the MSU team. The report also presents the performance results of different aspects of these solutions and their core components in terms of two main performance criteria: (1) speed (i.e., execution time) and (2) accuracy in terms of a system's precision to detect objects. Achieving both objectives, real-time execution and high-levels of accuracy, represent a major challenge for current deep learning-based object detection solutions. The best available object detectors tend to be quite complex and cannot run in real-time even when using powerful computers or servers. Meanwhile, popular object detectors that can run in real-time suffer from lower accuracy relative to the performance of more accurate high-complexity detectors. Consequently, one of the major challenges that the MSU team had to solve was improving the accuracy of real-time object detectors while maintaining their low complexity for viable real-time operations.

The MSU team faced another major challenge related to lack of railroad datasets, which are vital for performing supervised learning of deep learning-based object detectors. To address this major issue, the MSU team developed a three-prong strategy to execute the project and complete the team's major milestones in a timely manner. This three-prong strategy consisted of three paths pursued by the MSU team in parallel: (1) Exploiting publicly available datasets from the autonomous vehicle industry and research communities to train and test a broad range of low- and high-complexity object detectors. (2) Deploying MSU's existing AV platform, which is equipped with earlier generations of lidars, to collect sensor data under challenging weather conditions such as rain and snow. This provided vital data for training and testing leading object

detectors under these challenging conditions. (3) Integrating Next Generation (NG) automotive-grade lidar technology into a new multimodal system, which plays a significant role as a proof-of-concept in assessing the levels of improvements that can be achieved by deploying these NG high-resolution lidar sensors.

In addition to developing the key components of multimodal lidar-camera object detection methods and related deep learning solutions, the MSU team pursued another major effort: *cooperative perception*. Cooperative perception utilizes vehicle-to-vehicle and vehicle-to-infrastructure (V2X) solutions to enable trains, vehicles, and certain infrastructure locations to share sensor data to enhance the object detection performance of the overall system. The team's effort in this area focused on the development of a low-complexity cooperative perception framework.

The project made significant improvements to leading object detection solutions and demonstrated their effectiveness when integrated with the best available and NG automotive grade lidar sensor technologies. In particular, the project's outcomes demonstrate the feasibility of: (a) operating advanced deep learning-based 3D object detectors using lidar sensors at around 60 point-cloud frames-per-second (FPS), which is significantly faster than real-time, while maintaining accurate detections, (b) achieving high levels of detection accuracy in terms of Average Precision (AP) that can range between 80% for distant hard-to-detect objects to more than 90% AP for near easier-do-detect objects when operating in real-time under favorable conditions, (c) achieving both real-time speed at around 30 FPS and medium levels of accuracy that ranges between 40% to around 60% AP when operating under challenging weather conditions such as rain or snow, (d) achieving potentially-significant gains in accuracy when deploying NG lidars that can capture high resolution point-cloud frames, which can lead to both substantial improvements and changes in object detection solutions, and (e) improving accuracy under cooperative perception when multiple *agents* (trains, vehicles, infrastructure nodes, etc.) share and fuse their multimodal sensor data.

While the team made significant progress in advancing object detection on railroad tracks, the project also demonstrates that further work is needed to develop a viable and comprehensive railroad safety system that can be installed in the field and on moving trains. In particular, there are further needs for: (a) deploying advanced NG high-resolution lidar sensors and cameras on moving trains, vehicles, and on critical points within the railroad infrastructure to collect multimodal data under diverse weather conditions that include rain, snow, and fog. (b) Labeling such diverse-weather data for training and testing supervised deep learning methods. (c) Developing advanced domain-adaptation techniques to improve the performance of current object detection solutions under challenging conditions. (d) Exploring the area of cooperative perception further, especially under challenging conditions and a broader range of heterogeneous scenarios.

1 Introduction

A team of researchers in the College of Engineering at Michigan State University (MSU) in East Lansing, Michigan developed core components, software modules, deep-learning models, and integrated hardware sensors of multimodal solutions for real-time detection of safety-related railroad events and objects. The team started this project September 2023 and the project concluded March 2026. The developed solutions represent a combination of software and hardware platforms that are based on state-of-the-art sensor technologies and deep learning methods for the detection of objects and related events.

1.1 Background

The developed components and proof-of-concepts of the multimodal solutions rely on advanced deep learning methods and utilizes object detection approaches to demonstrate the feasibility of real-time detection of objects and events that could represent major safety issues for moving trains. Therefore, the primary motivation for developing the proposed solutions and related object detection technologies is addressing major railroad safety issues that include hazardous events such as falling rocks on railroad tracks, railroad crossings' safety-violation events, and trespassing incidents, which are the leading cause of fatalities in the US for the railroad industry. Another important goal is to explore new hardware, software, and deep learning technologies that lead to viable and practical solutions for implementation in the field such as moving trains and critical points within the railroad infrastructure.

1.2 Objectives

To achieve these goals, the MSU team identified three key research objectives that guided the project since its inception:

a) **Object Detection in Real-time**
A key objective of the MSU team effort was to develop the ability of the pursued solutions to detect safety-related objects in *real-time*. Achieving this objective is critical for demonstrating the viability of the proposed solutions in terms of enabling realistic deployment on moving trains while in operation.

b) **Object Detection under Diverse Weather Conditions**
The proposed solutions should be able to operate and detect objects under any weather conditions, such as rain and snow, and not only under favorable weather, such as clear and sunny conditions. This is another critical objective for enabling viable and realistic deployment of practical solutions in the field.

c) **Object Detection using Next-Generation Automative-Grade Sensors** Another key objective was to employ the best-available automotive-grade sensors. In that context, there are two operative subobjectives under this overall objective: "Next-Generation" and "automotive grade". Sensor technologies are continuously improving, and it became

imperative to develop a system that exploits the best available sensor technologies [1]. In particular, the latest trend in automotive grade laser-based sensors are high-resolution lidars that can capture significantly denser point clouds than earlier lidar sensors. We refer to these high-resolution laser sensors as Next-Generation (NG) lidars. The second subobjective under this overall objective is using automotive-grade sensors. This subobjective is quite critical for pursuing solutions that are cost effective, and which can be deployed by train operators broadly.

1.3 Overall Approach

The MSU team pursued a three-prong approach that represents three parallel yet interconnected efforts when conducting the project. Before describing this three-prong approach, it is critical to highlight the major challenges that shaped the team's effort and influenced its overall approach. In particular, the team faced two major issues that needed to be addressed from the beginning of the project.

1.3.1 Availability of Visual and 3D Lidar Point-Cloud Data

The primary challenge encountered by the team from the very early stages of the project is the lack of railroad-related visual and 3D lidar data that can be used for developing viable object detection solutions. It is important to note that these solutions require massive amount of data for training supervised-learning based techniques including state-of-the-art deep learning algorithms. In addition, datasets capturing hazardous natural events such as falling rocks on railroad tracks are virtually nonexistent.

1.3.2 A Three-Prong Approach

To achieve the project's primary objectives, development of real-time object detectors that can operate under challenging weather conditions using state-of-the-art and next generation lidar sensor technologies with advanced deep learning solutions, while addressing the main challenge regarding the lack of publicly available railroad-related lidar data the MSU team pursued a three prong approach that is illustrated in Figure 1.

[1] During the early stages of the project, it became clear that employing the best available sensor technology has its own challenges including issues related to procurement and availability. The MSU team identified this issue early on by following a three-prong plan as explained later in this report.

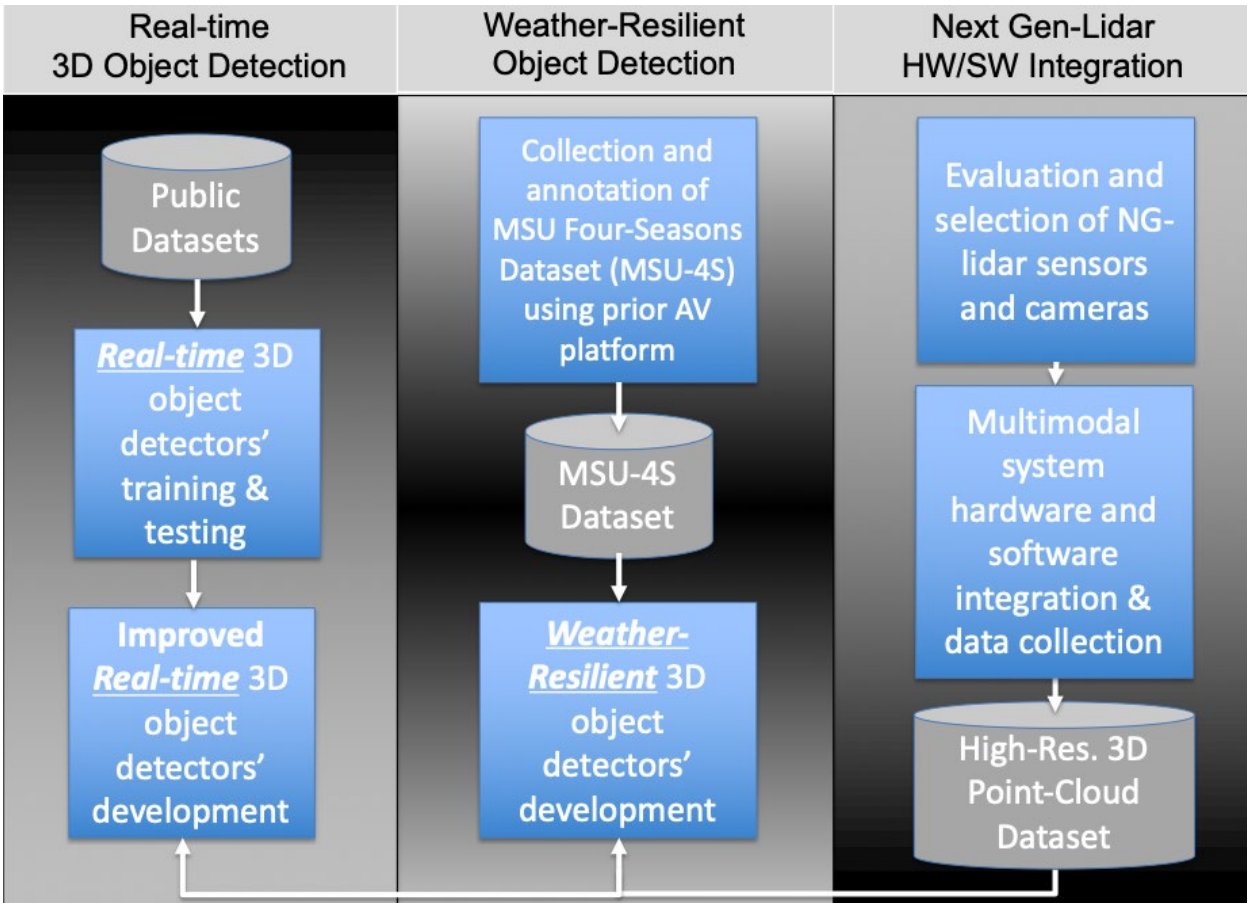


Figure 1. The three main subareas pursued by the MSU team for the development of real-time object detection methods for operation under adverse weather conditions.

This three-prong framework enabled the team to pursue three development efforts in parallel, and hence, the team avoided significant delays for the completion of the project.

Under this strategy, the team accomplished the three primary objectives of the project in parallel and completed the key milestones of the project in a timely manner. These three parallel paths are:

a) Development of *real-time* 3D object detectors with improved accuracy using publicly available data from the autonomous-vehicle industry and research community.

b) Development of robust 3D object detectors that can operate under challenging weather conditions using a dataset collected by the MSU team over four seasons (summer, spring, winter and fall). The MSU Four Seasons (MSU-4S) Dataset included challenging conditions such as rain and snow.

c) Integration of state-of-the-art Next-Generation (NG) lidar sensor technologies and all necessary software and hardware modules into a multimodal system that was used to collect a new high-resolution lidar dataset.

The three parallel efforts are described in detail in the remainder of this report.

1.4 Scope

The MSU team explored, analyzed, and developed advanced deep learning methods and object detection approaches to demonstrate the feasibility of real-time detection of objects and events that could represent major safety issues for moving trains. The MSU team focused on mobility and railroad-related objects that could be involved in safety-related obstructions, such as vehicles or people trespassing a railroad crossing. The researchers also focused on integrating automotive-grade lidars and cameras, which are readily deployed by Autonomous Vehicles (AVs) and Advanced Driver Assistance Systems (ADAS), with deep learning-based object detection solutions to achieve high levels of accuracy and robustness under challenging weather conditions.

1.5 Organization of the Report

Section 2 describes the development of a real-time object detection framework with improved Average Precision (AP) using publicly available data. Section 3 covers the research team's effort in developing weather-resilient object detection using a novel domain adaptation framework. Section 4 describes a proof-of-concept prototype that integrates a NG, automotive grade, high-resolution lidar and an RGB camera. Section 5 describes cooperative perception and the proposed low-complexity framework that the MSU team developed for sharing and fusing multimodal sensor data. Section 6 covers ongoing research and proposed future directions that can be built on the outcome of this project. Finally, Section 7, which provides a summary of the major outcomes of the project.

2 Real-time Object Detectors' Performance Results using Publicly Available Datasets

The development of real-time object detectors with improved accuracy using publicly available data was a critical objective of this project from its early stages. Given the lack of railroad lidar data, the MSU team exploited the vast amount of data publicly available to the research community by leading AV companies such as Waymo. Despite the availability of such data, there were two challenges that needed to be addressed. First, the team needed to evaluate the execution time (or speed) of state-of-the-art deep learning-based object detectors. This evaluation led to the observation that the best available object detectors that exhibit the highest level of accuracy tend to be very complex and cannot run in real-time.

On the other hand, many popular object detectors such as PointPillars, Voxel R-CNN, and SECOND can run in real-time, however, their accuracy is noticeably lower than the more complex detectors. Consequently, it became imperative to improve the performance of low-complexity real-time object detectors to meet the two key objectives, real-time execution and higher accuracy. Therefore, the second challenge that needed to be addressed is how to improve the object detection accuracy of low-complexity detectors while maintaining its real-time speed. To address this challenge, the MSU team developed a novel new framework that enabled low-complexity object detectors to *learn* from more advanced and more complex detectors during off-line training. We refer to this novel framework as SPARK, which is a *knowledge distillation* (KD) strategy where the complex more-accurate detectors act as *teachers* whereas the low-complexity less-accurate detectors behave like *students* that need to learn from their corresponding teachers. The SPARK framework was built on prior state-of-the-art research in very broad areas including object detection, data collection, knowledge distillation, and related technologies [1]-[73].

Overall, the research team employed four state-of-the-art high-complexity object detectors as teachers and used them to teach four low-complexity object detectors (i.e., students) that can run in real-time or faster than real-time. The team employed four object detectors as teachers: VirConv (Virtual Sparse Convolution), PV-RCNN++ (Point-Voxel RCNN), TED (Transformation-Equivariant Detection), and CasA (Cascade Attention network). For the students, the team used four low-complexity detectors: SECOND, Voxel-RCNN, SECOND-IoU, and PointPillars. All these low-complexity detectors can run in real-time or faster than real-time. Here, the criterion for being real-time is achieving faster than 10 Frames-Per-Second (FPS) speed on a common Nvidia-GPU hardware platform. In practice, the speed must be higher than 10 frames per second to leave room for other tasks to run on the same GPU platform.

Figure 2 shows a summary of the results. As shown in the figure, the fastest detector is PointPillars, which can process and detect objects at almost 60 Frames Per Second (FPS). Using our distillation framework, i.e. SPARK, the team was able to improve PointPillars accuracy by more than 6%, which is quite significant, especially when considering that the baseline for PointPillars' performance (without SPARK) was less than 47%. Here, accuracy is measured in Average Precision (AP) or AP with Heading (APH) information for the Waymo dataset.

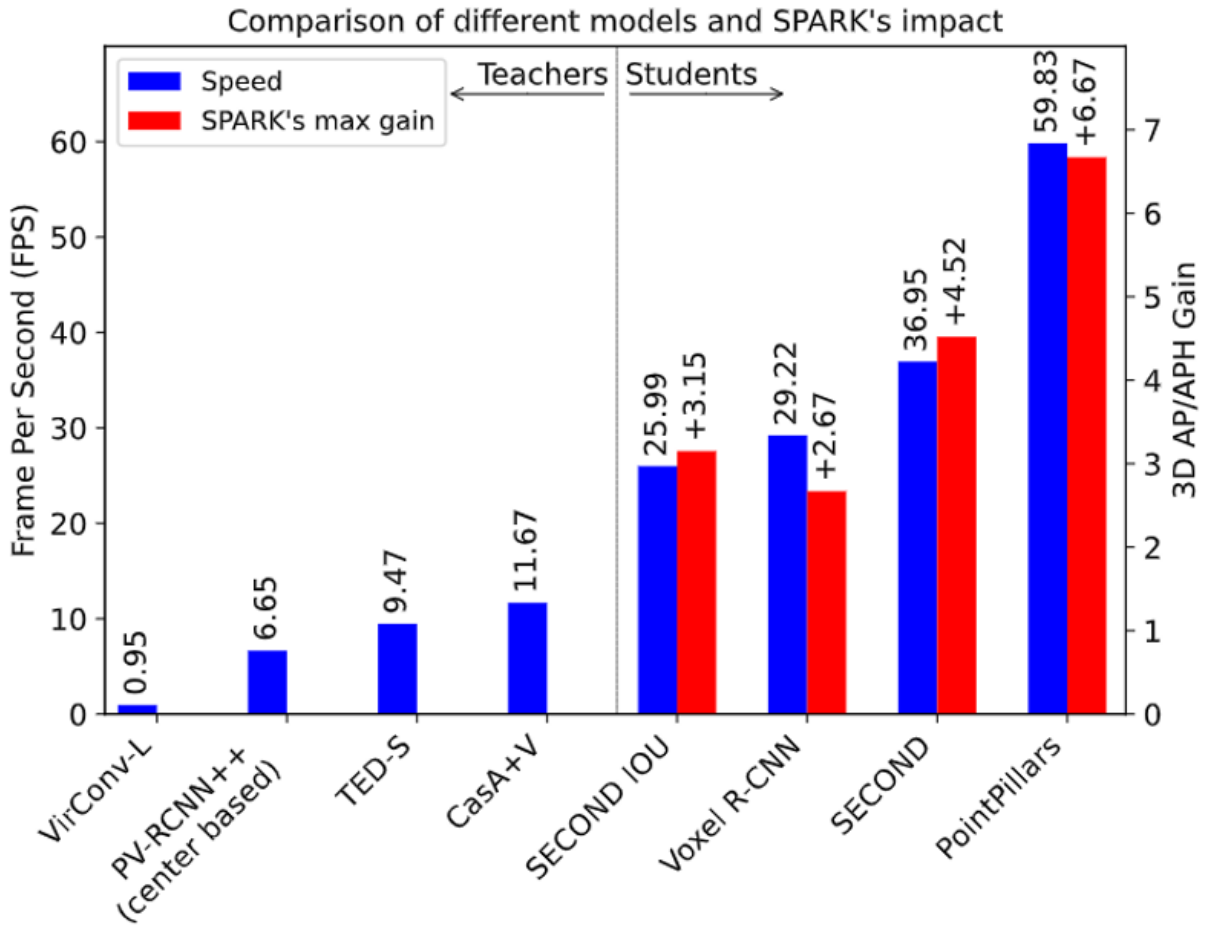


Figure 2. The speed (FPS) and the maximum detection accuracy gain, in terms of AP or AP with Heading (APH) achieved due to the proposed SPARK framework among different detectors, classes, and datasets.

In Figure 2, the blue bars represent the speed for a variety of state-of-the-art *teachers* and *students* object detectors. The red bars represent the maximum precision gain achieved by the student after applying a SPARK-based knowledge distillation from one of the teachers. SPARK can provide significant gains in AP/APH while maintaining the real-time speed of student detectors. Speeds are based on inference-time using an Nvidia 3090 GPU.

2.1 Performance Results

The MSU team used the KITTI dataset, which is the most popular dataset in the area of object detection, to train and evaluate the performance of the real-time object detectors considered in this study. The Average Precision (AP) results are shown in Table 1, which is divided into four major columns representing the four low-complexity real-time object detectors: SECOND-IoU, SECOND, PointPillars, and Voxel R-CNN. The table also shows the impact of the proposed real-time SPARK-based framework on improving the performance of these detectors. These results are based on employing three teachers: T1 (VirConv), T2 (TED), and T3 (CasA). There are three levels of difficulties for the objects that are being detected: Easy, Moderate, and Hard.

In general, these levels range from nearby objects that are easy to detect to faraway objects that are hard to detect. The *Baseline* performance represent the AP results without applying the SPARK framework. Overall, PointPillars, which is the least complex and fastest detector, benefits the most from the SPARK framework. SECOND and SECOND-IoU also improve their AP performance quite noticeably while both have faster than real-time speed at around 26 and 37 FPS (Figure 2), respectively.

Table 1. Object detection AP performance of four low-complexity real-time object detectors (*students*) with (and without) the proposed SPARK framework in conjunction with three high-complexity object detectors (*teachers*).

Students	**SECOND IOU**			**SECOND**			**PointPillars**			**Voxel R-CNN**		
	Easy	**Moderate**	**Hard**	**Easy**	**Moderate**	**Hard**	**Easy**	**Moderate**	**Hard**	**Easy**	**Moderate**	**Hard**
Baseline	91.54	82.37	79.63	90.55	81.61	78.61	87.75	78.41	75.19	92.38	85.29	82.86
SPARK: S+T1	93.38	85.28	82.87	92.14	84.68	81.83	90.96	81.60	78.38	93.53	86.18	83.64
Δ (vs Baseline)	**+1.84**	**+2.91**	**+3.24**	**+1.59**	**+3.07**	**+3.22**	**+3.21**	**+3.19**	**+3.19**	**+1.15**	**+0.89**	**+0.78**
SPARK: S+T2	93.25	85.19	82.78	91.62	84.94	82.09	90.82	81.38	78.84	93.49	86.00	85.53
Δ (vs Baseline)	**+1.71**	**+2.82**	**+3.15**	**+1.07**	**+3.33**	**+3.48**	**+3.07**	**+2.97**	**+3.65**	**+1.11**	**+0.71**	**+2.67**
SPARK: S+T3	92.35	84.76	82.64	91.23	83.79	81.14	89.72	80.57	77.93	92.77	85.94	83.45
Δ (vs Baseline)	**+0.81**	**+2.39**	**+3.01**	**+0.68**	**+2.18**	**+2.53**	**+1.97**	**+2.16**	**+2.74**	**+0.39**	**+0.65**	**+0.59**

Figure 3 provides a summary of the tradeoff between speed and AP accuracy for the different 3D object detectors evaluated. The red crosses represent the performance of low-complexity real-time object detectors prior to applying the SPARK framework. For these real-time detectors (students) the blue circles represent their improved performance after applying SPARK. While PointPillars provides the least complex and highest speed option, it also has the lowest accuracy. Meanwhile, among the other three real-time detectors, Voxel-RCNN provides slightly higher accuracy than the two detectors (SECOND and PointPillars). The rail operator can determine the final selection for which object detector to deploy in the field. For example, operators who need to deploy low-complexity detectors that can run much faster than real-time but at the expense of some acceptable degradation in accuracy, they may choose SPARK-PointPillars because of its significantly higher speed than other detectors. Alternatively, operators who target the highest possible accuracy for real-time object detection, then SPARK-Voxel R-CNN would probably be their best choice.

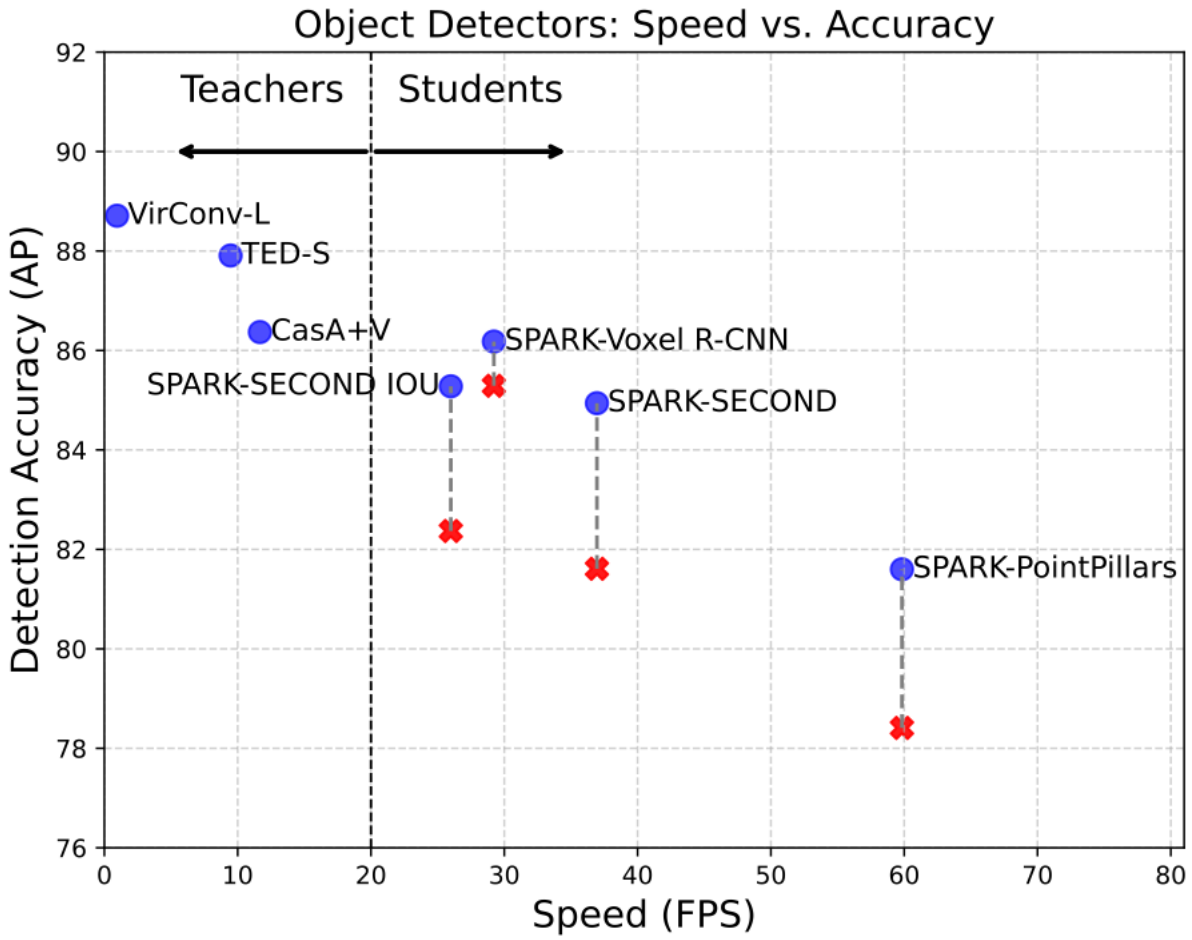


Figure 3. The tradeoff between detection accuracy and speed for the 3D object detectors evaluated in this study.

SPARK Deep-Learning Architecture

Figure 4 shows the high-level architecture for the proposed SPARK framework. SPARK consists of three major parts. First, the research team trained an advanced high-complexity highly accurate *teacher* using a set of training data. This results in the *teacher model*. The team then used this pre-trained teacher model to generate pseudo-labels for training a low-complexity *student model* that can run in real-time. Finally, the team trained low-complexity student model (object detector) using the pseudo labels generated by the teacher model. This three-stage process improves accuracy of the student model. The novelty of knowledge distillation from the teacher to the student is rooted in a new loss function, Symmetrically Weighted Inter-Focal-loss from Teacher-to-student (SWIFT) loss function, introduced by the team.

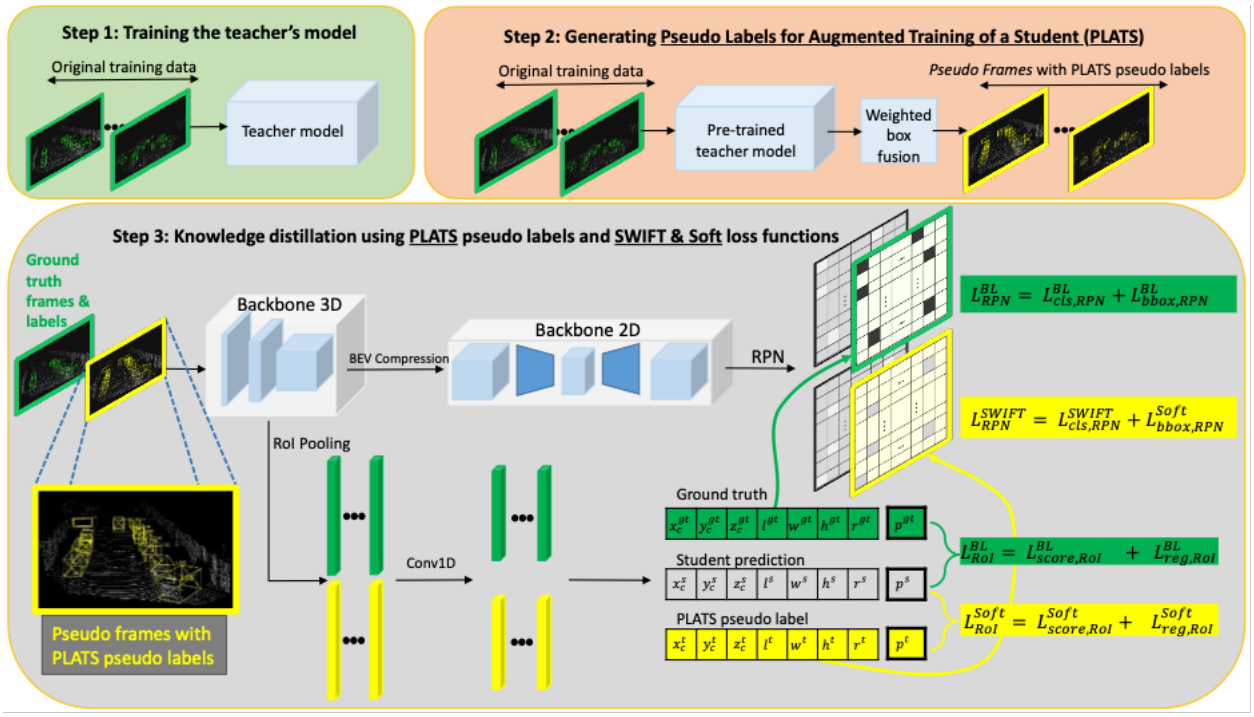


Figure 4. Architecture and training pipeline of SPARK.

As shown in Figure 4, SPARK consists of three steps. Under Step 1, a teacher model is trained. In Step 2, the teacher model generates pseudo labels (PLATS - Pseudo Labels for Augmented Training of a Student) that will be used for training the student. Under Step 3, the pseudo frames (yellow) with the (PLATS), and the original ground-truth frames/labels (green) are input into the student detector. The Region Proposal Network (RPN) applies the Symmetrically Weighted Inter-Focal-loss from Teacher-to-student (SWIFT) and Soft loss functions for Knowledge Distillation (KD). If the student supports a Region-of-Interest (RoI) network, KD is achieved through Soft loss functions. The grid cells of the RPN output with potential objects are shown in gray representing the teacher's confidence score, while the black cells represent the ground-truth foreground.

3 Weather-Resilient Object Detectors' Performance Results using the MSU-Four Seasons Dataset

A primary research objective of this project is the development of object detection solutions that can operate under challenging weather conditions such as rain and snow. The accuracy of deep learning solutions in general, and object detectors trained under favorable conditions in particular, can degrade significantly when tested under other conditions. This phenomenon is known as the *domain shift* problem since the training data (source domain) covers a different domain than the test data (target domain).

To achieve this research objective, the MSU team needed a comprehensive dataset that covers different challenging weather conditions. The lack of such data motivated the team to pursue a concerted effort to collect the MSU Four Seasons (MSU-4S) dataset [74], which was captured using an already available vehicle platform that is equipped with earlier generations of lidar sensors. The collection of the Four Seasons dataset provided much-needed data that enabled the team to move forward with meeting its objectives in a timely manner avoiding significant delays while pursuing the other critical path of acquiring and integrating the NG lidar sensor within the multimodal system that is described later in this report.

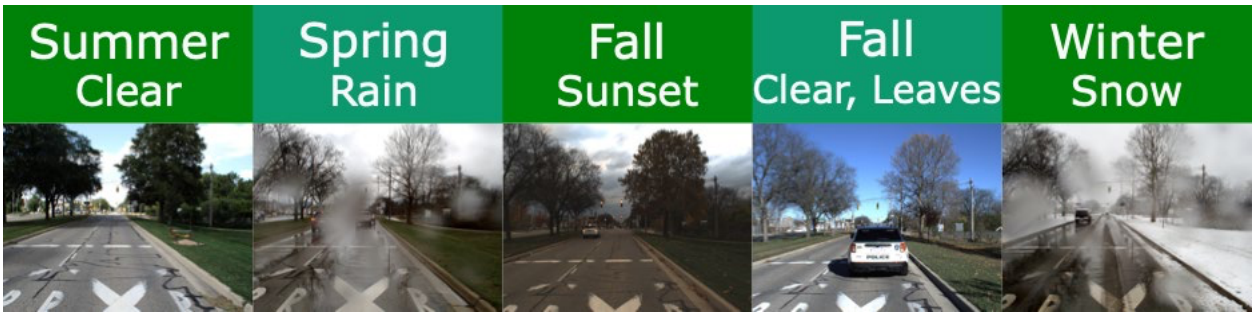


Figure 5. Train-crossing images from the MSU-Four Seasons dataset.

Figure 5 show sample images of the same train-crossing scene captured under different seasons and weather conditions. In Figure 6, the lower image shows a lidar point cloud frame from the MSU-4S dataset with ground-truth bounding boxes (red for vehicles and green for pedestrians). The upper image shows the lidar samples superimposed over the corresponding RGB picture captured by the MSU-4S multimodal system.

With availability of the MSU Four Seasons data, the team developed a novel weather-resilient object detection framework that addresses the domain shift problem between the performance of object detector models trained and tested under favorable weather (e.g., sunny or summer data) and the performance of the same model but tested under challenging weather such as rain or snow. In particular, the MSU team developed a novel framework referred to as WISDOM (Weather-driven and IntenSity Distribution-based Object Modulation). Under this framework, favorable weather (source-domain) lidar data are transformed into challenging weather (target-domain) lidar data using a modulation function. The modulation function is derived from actual source and target domain probability distributions. This strategy provided significant gains

relative to the baseline performance as shown in Table 2. The table shows the performance when the *baseline* object detector model (without WISDOM) and the WISDOM models for snow and rain are tested under snowy and rainy conditions, respectively. WISDOM represents a departure from prior work in weather-related detection, which was traditionally built on numerous other works in domain adaptation, 3D object detection, and simulation-based and physics-based LiDAR data transformation [75]-[174].

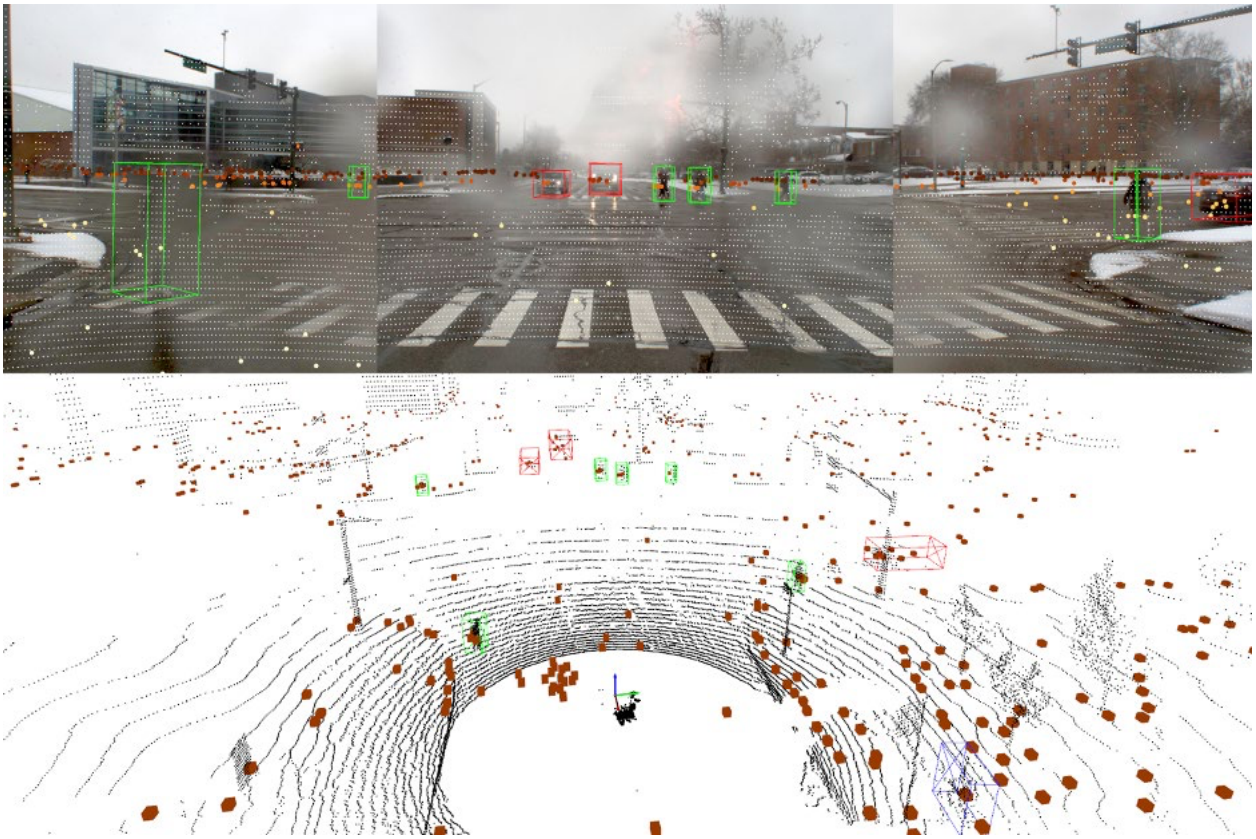


Figure 6. A multimodal lidar-camera scene from the MSU Four Seasons dataset captured under snowy winter conditions.

As shown in Table 2, the impact of the WISDOM framework could be quite significant on the Average Precision (AP) performance of the two popular object detectors, Voxel R-CNN and PV-RCC++, when tested under snowy and rainy conditions. In these tables, the *baseline model* implies the object detector was trained using the favorable weather (sunny) conditions. It is important to note that WISDOM-snow and WISDOM-rain represent two different models that were trained using two different modulating functions. For example, the WISDOM-snow model was based on a modulation function that was derived using both favorable weather (with ground truth labels) and snowy data (with pseudo labels). Moreover, WISDOM can achieve significant gains in AP performance. As an example, and relative to the baseline performance (without WISDOM), the WISDOM-based rain model can achieve more than 18% in AP gain for the pedestrian class. This level of performance is quite significant and encouraging, and it can open

the door for further improvements by combining the WISDOM framework with other solutions that address the domain shift problem.

Table 2. The impact of the WISDOM framework on the Average Precision (AP) performance of two popular object detectors (Voxel-RCNN and PV-RCNN++) when tested under (a) Snowy and (b) Rainy conditions.

a) Detectors' models tested on <u>Snowy</u> Conditions	Car	Pedestrian	Bike
Voxel-RCNN (baseline)	21.02	29.49	22.49
+ WISDOM-snow (Δ)	+2.98	+16.38	+5.89
+ WISDOM-rain (Δ)	+3.63	+18.67	+5.33
PV-RCNN++ (baseline)	23.46	42.68	14.73
+ WISDOM-snow (Δ)	+3.15	+6.36	+6.29
+ WISDOM-rain (Δ)	+0.29	+7.91	+7.30

b) Detectors' models tested on <u>Rainy</u> Conditions	Car	Pedestrian	Bike
Voxel-RCNN (baseline)	16.51	33.52	2.67
+ WISDOM-snow (Δ)	+1.53	+9.48	+4.79
+ WISDOM-rain (Δ)	+2.91	+8.54	+4.55
PV-RCNN++ (baseline)	20.62	40.23	2.41
+ WISDOM-snow (Δ)	+0.52	+3.25	+2.59
+ WISDOM-rain (Δ)	+2.52	+5.44	+4.43

4 Next-Generation Lidar Hardware and Software Integration and Performance Testing

One of the primary objectives of the project has been the evaluation of the performance of Next-Generation (NG) automotive-grade lidar sensors in the context of 3D object detection. These NG lidars can capture much higher-resolution 3D point-cloud frames when compared with earlier lidar technologies. Specifically, NG lidars are built using 1550nm laser technology, and hence these lidars can produce densely sampled data. With the higher point density, very fine features at unprecedented levels of details can become detectable or recognizable, whereas the same information were virtually impossible to discern by previous lidar systems. With NG lidars, and at maximum data density, even the text on signs is visible, opening new possibilities for new detection capabilities.

Figure 7. (Top left) high-resolution 3D point-cloud frame of a railroad-crossing scene captured by a Next Generation (NG) lidar used by the MSU team. (Top right) The fusion of the lidar point-cloud frame with the corresponding RGB camera frame. (Bottom) A zoomed-in version of the same scene illustrating the high-resolution features captured by the lidar to the extent that one can read the text "STOP" on the traffic sign at the railroad crossing.

Figure 8. Point-cloud 3D lidar frames captured by different lidar sensors at different temporal frame rates of the same location around a railroad crossing.

In Figure 8, the lidar frames are fused with the corresponding RGB camera images. The top image of the figure shows a lidar frame captured by an 850nm wavelength lidar sensor at 10 frames/second (10 Hz). The middle and bottom images in Figure 8 show two lidar frames captured by a 1550nm lidar sensor at two different frame rates, 10 Hz and 1 Hz, respectively. The difference in sample density and resolution is quite clear between the two lidar sensors.

After evaluating and assessing different features and capabilities of state-of-the-art lidar sensors, the MSU team selected the Luminar Iris 1550nm lidar. The team integrated the Luminar sensor with a state-of-the-art camera into a multimodal system that provided impressive performance in terms of sample density and high-resolution data. Figure 7 shows examples of the data captured by this NG lidar based system. These examples clearly show the high-density and resolution of this NG lidar to the extent that some text of traffic signs can be read rather easily by viewing the lidar point-cloud frame, which is an unprecedented capability that was not feasible by using prior automotive-grade lidar sensor technologies.

The MSU team employed the new multimodal system to capture a new set of data at different locations around the MSU campus including at railroad crossings and railroad tracks. The team then used this new data for testing the SPARK-based 3D object detector they developed. It is important to note that the team trained the SPARK-based PointPillars detector, which can operate in real time, using other publicly available datasets and not the new multimodal data captured by the new Luminar lidar. Therefore, the training data was very different from the testing data in this case. Despite a significant domain-shift between the training data and testing data, the SPARK-based PointPillars object detector was able to detect many objects in the scenes that were used for testing. This successful outcome is attributed to both the robustness of the 3D object detector and the high-resolution of the new data captured by the Luminar lidar.

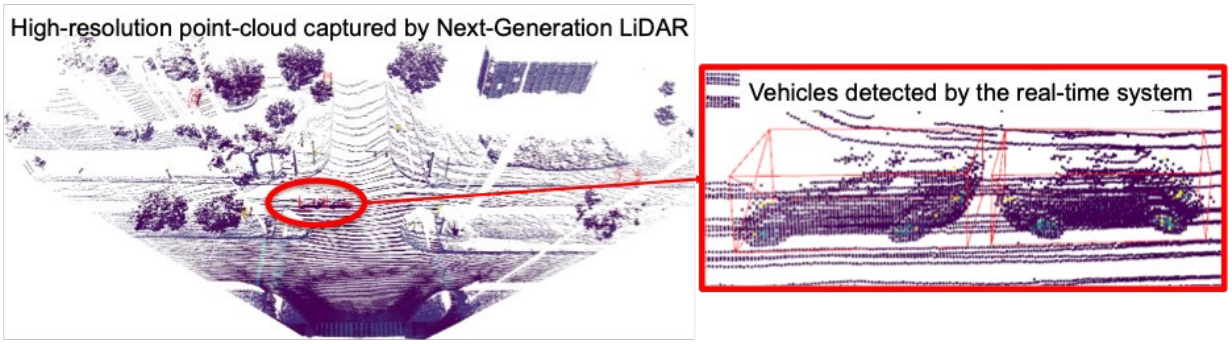
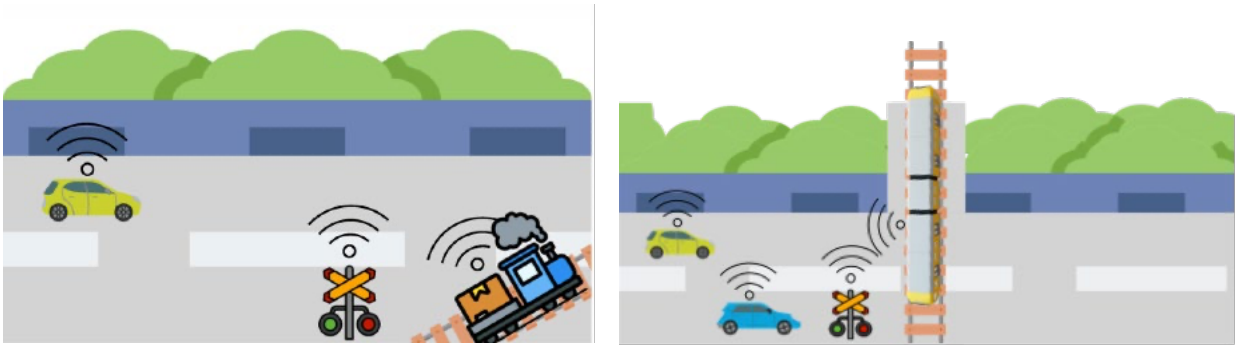


Figure 9. High-resolution point-cloud frame captured by the Luminar lidar and the results for applying a SPARK-based Point Pillars object detector onto the scene.

5 Cooperative Perception using Connected Vehicle Technologies

The accuracy of object detectors is inherently constrained by the coverage provided by the sensor data captured at a given location and the visible surroundings of that sensor's vantage point. In other words, an object detector can only detect objects that are within the field-of-view of its sensors. Meanwhile, there are many safety-critical scenarios where objects or events are not within the field-of-view of the sensor, or they might be occluded by other objects that are within the field-of-view of the scene. In such situations, *connected vehicle* technologies can play a crucial role in enhancing the object detectors' capabilities in terms of improving their perception accuracy. In particular, Vehicle-to-Vehicle (V2V) and Vehicle-to-Infrastructure (V2I) capabilities can improve the operational capabilities of vehicles' perception and understanding of their surroundings, effectively expanding the detection range by incorporating data from other vehicles or from certain locations within the infrastructure. Collectively, known as V2X solutions, vehicle connectivity technologies enable an *ego vehicle* which is the vehicle under consideration, to receive sensor data and information from other nearby vehicles or from a set of sensors installed within the infrastructure to complement its own sensor information.

For this project specifically, V2X can provide earlier warnings to other trains and to emergency personnel about potential hazards or obstacles obstructing a moving railroad vehicle. Besides this capability, MSU researched applications that go beyond basic message-based standards such as those defined in SAE J2735. They pursued novel approaches for sharing sensor and processed perception data either among mobile vehicles or between vehicles and stationary platforms within the infrastructure. The sharing of sensor information and detection data among different vehicles and infrastructure locations is emerging as an important area of research known as ***cooperative perception*** **or** ***collaborative perception*** [2]. In that context, the different cooperative perception entities, i.e., connected vehicles and infrastructure platforms that are cooperating, are known as ***agents***.



[2] Cooperative perception and collaborative perception are two expressions that are used interchangeably in the literature. We also use these two expressions interchangeably in this document.

Previous research pursued in this area focused mainly on sharing raw data or using basic compression techniques. On the other hand, deep learning frameworks offer the capability of sharing *deep features,* which tend to be compact representation of the scene and objects within that scene. This can be thought of as *feature-based compression and fusion* among cooperating and connected vehicles and infrastructure nodes. This form of feature-based cooperative perception could potentially be used directly by compatible object detectors used by different agents without the need to convert the deep features back into the original data, which can potentially reduce perception latency. It is worth noting that reducing latency is an important goal in any system that is safety critical. Furthermore, novel compression techniques suited for cooperative vehicles' and infrastructure platforms' perception may alleviate much of the increased bandwidth requirements; without compression, raw sensor data would require massive amounts of bandwidth, which in turn complicates the selection of appropriate wireless technology that can be used to implement cooperative perception.

5.1 Faster HEAL: A New Framework for Cooperative Perception

As explained above, cooperative perception is a promising paradigm for improving situational awareness by overcoming the limitations of single-agent perception. However, most existing approaches assume homogeneous agents, that is agents that have the exact set of sensors, using the same object detectors, and virtually the same sensor configurations. This homogeneous-agents assumption restricts the applicability of cooperative perception in real-world scenarios where vehicles and infrastructure nodes usually use diverse sets of sensors and perception models. Such heterogeneity introduces a feature domain gap that degrades detection performance. Bridging the feature domain gap among heterogeneous agents is therefore essential for building robust and safety-critical cooperative perception systems suitable for real-world autonomous driving.

Prior research on heterogeneous collaborative perception can broadly be categorized into two approaches. The first approach seeks to close the domain gap by retraining parts of the ego-vehicle or neighboring agents' models to project features into a *unified feature space*. In particular, the HEterogeneous ALliance (HEAL) approach consists of two phases: an initial phase during which the ego vehicle and other homogeneous agents are trained using a traditional cooperative perception strategy, and a second phase where new agents are introduced and are required to retrain their encoders among other parts of the model to project the input data into the unified feature space. While effective, this strategy poses practical challenges: retraining is required for each newly introduced agent type, which is infeasible for real-time deployment and may compromise both vehicle privacy and single-agent perception performance. Since most autonomous driving stacks are proprietary, sharing internal parameters or retraining on private data is often impractical. The second approach introduces feature interpreters to translate or align heterogeneous features across agents, thereby preserving single-agent performance and protecting model confidentiality. However, these interpreter-based methods typically rely on larger, more computationally expensive models and still require retraining or storage expansion whenever new heterogeneous agents join the system. These limitations highlight the need for approaches that can efficiently generalize to unseen agents while maintaining accuracy, privacy, and computational efficiency.

To address the above limitations of state-of-the-art cooperative perception strategies, the MSU team propose *Faster-HEAL*, a heterogeneous cooperative perception framework that efficiently bridges the semantic domain gap between heterogeneous agents while preserving privacy and maintaining significantly higher computational efficiency relative to the original HEAL method. Our approach, which deploys the initial stage of HEAL, introduces a novel *lightweight interpreter for feature transformation (**LIFT**)* in the second phase of the architecture. In particular, LIFT aligns intermediate features from newly joining neighbor agents with the unified feature space of the ego agent.

Therefore, Faster-HEAL is trained in two stages. During the first phase, the ego agent is trained on a homogeneous collaborative perception base to construct a unified feature space and fusion model. This phase, which is common between HEAL and Faster-HEAL, provides significant gain in AP performance when compared with the AP performance when different agents operate without any collaboration.

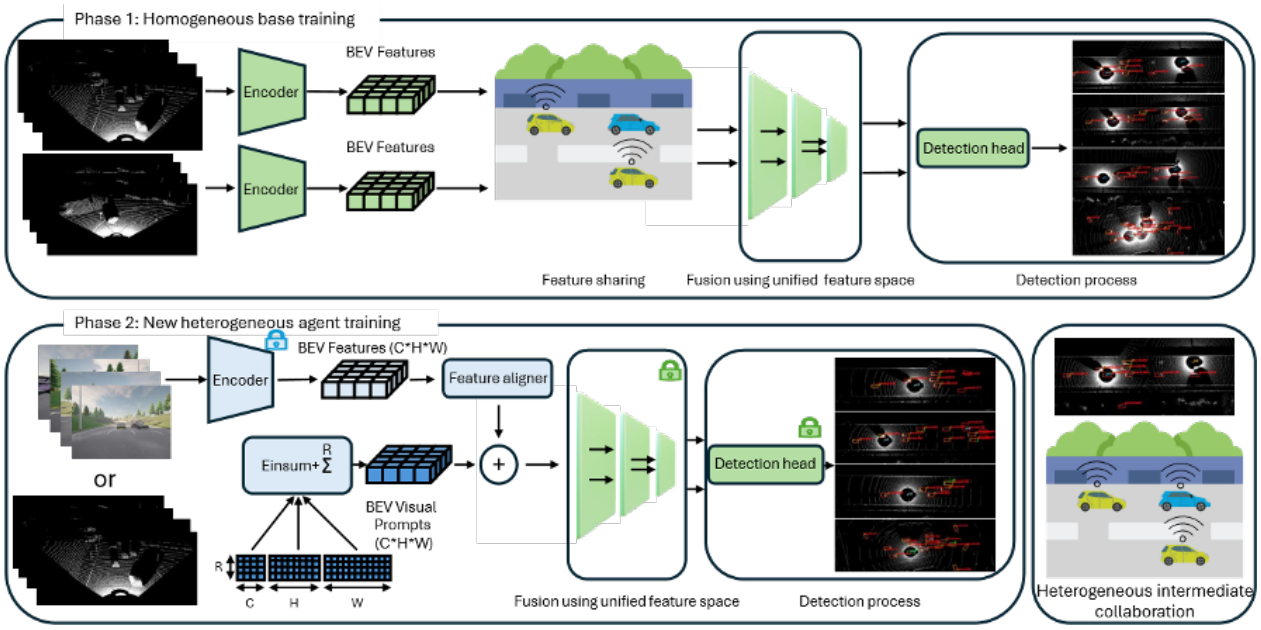


Figure 10. Overview of Faster-HEAL.

As shown in Figure 10, under Faster-HEAL's Phase 1, a *Homogeneous Base Training* constructs a unified feature space using pyramid fusion and trains the detection head on collaborative data from homogeneous agents. Under Phase 2, for each new heterogeneous agent type, the team froze the pretrained encoder, pyramid fusion, and detection head, and trained a lightweight feature aligner and visual prompts to map its intermediate features into the unified space. This design ensures low training cost, preserves the privacy of participating agents by requiring only their intermediate features, and enables scalable heterogeneous collaboration. In this framework, LiDAR is used for homogeneous agents, while camera or LiDAR could be used for new heterogeneous agents.

To evaluate the performance of the proposed Faster-HEAL framework, the team conducted a comprehensive set of experiments using the OPV2V-H [3] dataset and showed that Faster-HEAL achieves state-of-the-art detection performance with significantly lower computational overhead, offering a practical solution for scalable heterogeneous collaborative perception. Table 3 shows the performance impact of cooperative perception in terms of APAP for different object detectors and sensor configuration models using the homogeneous phase of Faster-HEAL frameworks. For example, in Table 3, Agent M1 uses a 64-laser beam lidar and the popular 3D object detector PointPillars whereas Agent M2 employs an RGB camera with 384x768 resolution images and a 2D object detector EfficientNet. These results are shown for two Intersection-over-Union (IoU) criteria, IoU-50 and IoU-70 that correspond to performance results AP50 and AP70, respectively. The "Single Agent" column shows the AP performance without cooperative perception. The "Homogeneous Collaborative Perception" column shows the performance when two agents using the same sensor and object detector models collaborate with each other.

Table 3. The performance impact of cooperative perception in terms of AP for different object detectors and sensor configuration models using the homogeneous phase of HEAL and the proposed Faster-HEAL frameworks.

Agent Model ID	**Agent Sensor(s) & Object Detector/ Backbone Models**	**AP50 Single Agent**	**AP50 *Homogeneous* Collaborative Perception**	**AP70 Single Agent**	**AP70 *Homogeneous* Collaborative Perception**
M1	64-channel LiDAR	0.83	0.95	0.72	0.91
	PointPillars				
M2	Camera with 384 x 768 images	0.34	0.61	0.18	0.43
	EfficientNet				
M3	32-channel LiDAR	0.76	0.91	0.66	0.86
	SECOND				
M4	Camera with 336 x 672 images	0.28	0.51	0.11	0.32
	ResNet-50				

[3] OPV2V is an open dataset for Vehicle-to-Vehicle cooperative perception. The original OPV2V dataset contains more than simulated 11K frames and more than 232K annotated 3D vehicle bounding boxes. The dataset was collected from 8 towns in CARLA and a digital town of Culver City, Los Angeles. OPV2V-H complements the original dataset with supplementary data and more diverse set of sensor configurations.

During the second phase of Faster-HEAL, the lightweight interpreter LIFT is trained to map new heterogeneous agents' features into the unified feature space without retraining any other model components. This design preserves the privacy and single-agent accuracy of newly added agents, avoids costly retraining and storage overhead, and enables scalable deployment. By leveraging a unified feature space together with a parameter-efficient lightweight interpreter, Faster-HEAL mitigates the performance degradation typically observed in heterogeneous collaborative settings while maintaining computational feasibility.

The MSU Research Team leverage visual prompts as the core mechanism for feature interpretation. Analogous to language prompts, visual prompts guide the model by emphasizing task-relevant information in the feature representation. Specifically, the team reused: (a) each new agent's encoder and backbone that have been trained under a single-agent scenario, along with (b) the pretrained pyramid fusion module and detection head of the ego agent, and train only the vision prompt parameters to align heterogeneous features. By freezing all pretrained components, our approach preserves single-agent detection accuracy and protects the privacy of newly introduced agents. Although naively training visual prompts can be computationally expensive due to their high dimensionality, the team design a low-rank prompt representation that reduces the number of trainable parameters by an order of magnitude, significantly improving training efficiency.

Consequently, the key capabilities and contributions of the proposed cooperative perception Faster-HEAL framework are:

1. Faster-HEAL introduces a lightweight feature interpreter (LIFT) for newly emerging heterogeneous agent types, enabling single-stage feature alignment with the ego agent's unified feature space. Only a visual prompt is fine-tuned for each new agent type with a different modality. Hence, the ego agent only requires intermediate features from collaborators, allowing each vehicle to keep its sensor configuration and model parameters private.

2. By introducing low-rank visual prompts, our approach achieves fast and accurate feature adaptation while minimizing the number of trainable parameters. Moreover, the method reduces training computation cost by an order of magnitude and requires the ego agent to store only the visual prompts corresponding to different agent types.

3. The team conducted extensive experiments that show that Faster-HEAL significantly reducing computational overhead by at least 90% while suffering no degradation in accuracy. In fact, they showed that Faster-HEAL not only does not degrade the AP of the original HEAL approach, but it improves the AP by at least 3%. The significant reduction in computational cost is achieved while representing a practical step toward scalable and privacy-preserving collaborative perception in real-world autonomous driving.

Table 4 shows the performance results of Faster-HEAL in comparison with other state-of-the-art cooperative perception methods including the original HEAL framework. These results are based on employing agent M1 as the ego agent, which is the same agent shown in Table 3, and which deploys a 64-laser beam lidar and PointPillars as an object detector. Other agents, M2, M3, and

M4, were introduced sequentially to the system. Hence, the column (+M2) in the table represents the performance when the ego vehicle (M1) adds the perception information provided by the new agent M2. Then, the column +M3 represents the performance when the ego agent further adds the information from the new agent M3 in addition to the already received information from M2. Hence, +M3 represents the fusion of three deep features' vectors (M1+M2+M3). Similarly, +M4 represents the fusion of four features' vectors (M1+M2+M3+M4). The "No Fusion" row shows the AP results for the ego agent (M1) when no cooperative perception is used. By adding a new agent (+M2) to the ego agent (M1) clear AP improvements can be achieved for all cooperative perception approaches. Further gains can be observed by adding a third agent (+M3) to the system. Adding a fourth agent (+M4) can also lead to some improvements. However, the gain in performance tend to saturate as further new agents cannot provide much of new information that the system already has. By comparing the proposed framework with other state-of-the-art approaches shown in the table, Faster-HEAL outperforms all other methods including HEAL under all scenarios while maintaining very significant reduction in complexity. For example, Faster-HEAL can reduce training complexity by 99% in terms of the number of parameters that need to be retrained when adding the new agent M2 to the system in comparison with the original HEAL architecture.

Table 4. Comparison of Faster-HEAL with other state-of-the-art cooperative perception methods including the original HEAL framework.

Approach **Based on** **M1** **Model as the Ego Agent** **(see** Table 3**)**	**Average Precision** **Minimum** **Intersection-over-Union** **IoU = 50** **AP50** ***The higher the number the better* (↑)**			**Average Precision** **Minimum** **Intersection-over-Union** **IoU= 70** **AP70** ***The higher the number the better* (↑)**			**Model Complexity** **Number of Training Parameters** **(in millions)** ***The smaller the number the better* (↓)**		
New Agent →	**+M2**	**+M3**	**+M4**	**+M2**	**+M3**	**+M4**	**+M2**	**+M3**	**+M4**
No Fusion	0.748	0.748	0.748	0.606	0.606	0.606	/	/	/
Late Fusion	0.775	0.833	0.834	0.599	0.685	0.685	/	/	/
F-Cooper	0.778	0.742	0.761	0.628	0.517	0.494	30.70	39.59	49.22
DiscoNet	0.798	0.833	0.830	0.653	0.682	0.695	30.80	39.67	49.29
AttFusion	0.796	0.821	0.813	0.635	0.685	0.659	30.78	39.60	49.22
V2X-ViT	0.822	0.888	0.882	0.655	0.765	0.753	36.17	44.99	54.62
CoBEVT	0.822	0.885	0.885	0.671	0.742	0.742	33.24	42.05	51.68
HM-ViT	0.813	0.871	0.876	0.646	0.743	0.755	47.71	65.08	83.34
HEAL	0.826	0.892	0.894	0.726	0.812	0.813	15.01	1.08	1.91
Faster-HEAL (ours)	**0.893**	**0.930**	**0.927**	**0.797**	**0.863**	**0.857**	**0.113**	**0.113**	**0.113**
Improvement w.r.t. HEAL	**↑ 0.067**	**↑ 0.038**	**↑ 0.033**	**↑ 0.071**	**↑ 0.051**	**↑ 0.044**	**↓ 99.2%**	**↓ 89.5%**	**↓ 94.1%**

6 Ongoing Research Effort and Future Work

While the team made significant progress as highlighted above, they also concluded that several areas need continued efforts and further work to develop a viable railroad safety system that can be installed in the field and on moving trains. The team identified several areas for further research that could be pursued as a follow-up to the current project.

6.1 Generative Models for Railroad Data

The lack of railroad related datasets was a major challenge for this project. This motivated the team to explore emerging foundational models to generate railroad data that could be used for training deep learning neural networks that can be deployed for railroad safety systems. The use of such data is motivated by the fact that scene perception algorithms for image classification, object detection, and semantic segmentation require massive datasets for training, validation, and testing. Hence, the lack or scarcity of *real* railroad data makes it necessary to employ novel approaches for generating new data using emerging *generative models*. Image segmentation is of particular interest, as a properly segmented scene provides far more granular scene understanding than bounding box or whole image labels. However, the railroad domain has limited public data available, leading to much more limited research and development of computer vision applications in this area. Synthetic datasets have been used in other applications to mitigate the dearth of data, and with modern open-source image generation techniques now widely available, the question remains not if, but how, these new techniques can be used to fill the gap. To that end, the team developed a methodology for leveraging a pre-trained diffusion-based image generation network to create training data targeted at railroad domains for image segmentation networks, using limited amounts of real-world data. In particular, the MSU team integrated two generative-models' tools: ControlNet and Low-Rank Adaptation (LoRA). The team refers to the integrated model as the Joint Augmented Controlnet-LoRA (JACL) framework. The first neural network tool, ControlNet, works as an extension to a large generative model, more specifically Stable Diffusion. In that context, ControlNet provides a mechanism for guiding (or controlling) the large generative model to create content that is influenced by visual prompts such as images rather than simple text input. This provides a platform for dictating certain structure of the output content, which can influence the shape and layout of the generated visual data.

LoRA is a highly efficient fine-tuning tool that can be employed with generative models (e.g., Stable Diffusion or Large Language Models – LLMs). LoRA's efficiency is achieved by freezing the massive number of weights and parameters of the base generative model while adding a very small "low-rank" neural-network layers for further training and fine-tuning of the model. This eliminates the need for retraining a massive amount of model parameters and weights while retraining the model to perform a specific task of interest. This strategy reduces computational complexity, storage size, and training time rather significantly. And at the same time, LoRA enables rapid retraining and customization of the model to focus on a certain application domain, which in this case railroad data. Using a Joint Augmented Controlnet-LoRA (JACL) framework, we both bias the network towards our target-application domain (i.e, railroad) while enabling precise control of output generation using pre-defined image segmentation masks as input, which then become the ground truth segmentation mask for the output image. So far, the team demonstrated that this architecture could produce realistic railroad imagery, with additional

broad control over weather conditions and time of day, using simple text prompt modification. The next major objective is to demonstrate that the resultant data can be used to train image segmentation networks. Further details regarding this framework are provided in Appendix A.

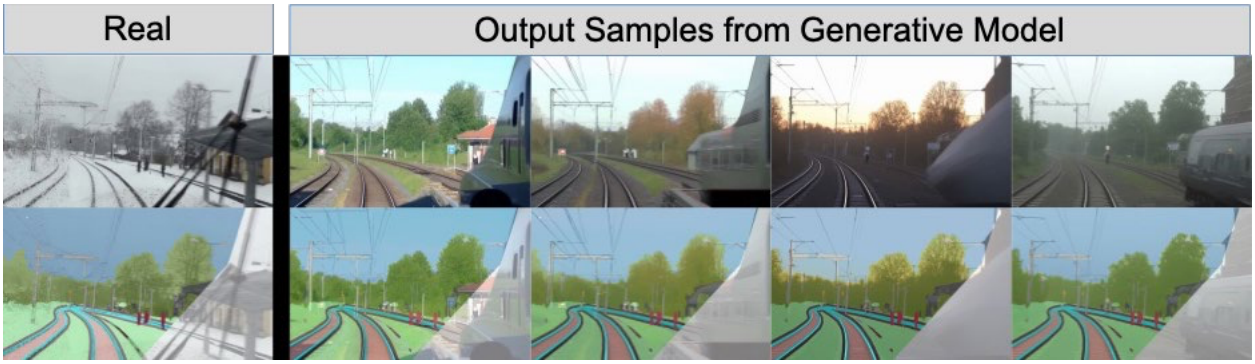


Figure 11. An example of synthetic rail domain data generated by the proposed JACL generative model.

6.2 Comprehensive Dataset Collection and Annotation using Next-Generation Lidar-based System

Under this effort, the MSU team aimed to achieve two major objectives:

(a) Deploying advanced NG high-resolution lidar sensors and cameras on moving trains, vehicles, and on critical points within the railroad infrastructure to collect multimodal data under diverse weather conditions that include rain, snow, and fog.

(b) Labeling the collected data under (a) above for training and testing supervised deep learning methods. This step is equally important to collecting the data itself since ground-truth labels and annotations play a critical role for training, validation, and testing most deep-learning algorithms, especially those that are used by object detectors.

6.3 Multimodal Domain-Adaptation

Under future work, MSU team will aim to develop advanced domain-adaptation techniques to improve the performance of current object detection solutions under challenging conditions. The team made significant performance gains in terms of AP under the novel WISDOM framework developed, however, there is still room for improvements relevant to the *oracle* performance. Here, oracle performance is the system precision when the object detector is trained and tested on data collected under the challenging condition (target domain), such as snow or rain, and when ground truth labels are available for such data.

6.4 Cooperative Perception under Challenging Weather

Here, the team's objective is to explore the area of cooperative perception further, especially under challenging conditions and a broader range of scenarios. The proposed Faster-HEAL improved the performance relative to the single-agent scenario. However, the results were based

on training and testing under normal (favorable) conditions, and no test was conducted under challenging conditions.

7 Conclusion

Under this project, the MSU team developed key components of a proof-of-concept multimodal sensor system capable of real-time detection of safety-related railroad events and objects. The developed multimodal system's components rely on advanced deep learning methods and utilize object detection approaches to demonstrate the feasibility of real-time detection of objects and events that could represent major safety issues for moving trains. One critical aspect of the project has been the integration of automotive-grade lidars and cameras, which are readily deployed by Autonomous Vehicles (AVs) and Advanced Driver Assistance Systems (ADAS), with deep learning-based object detection solutions to achieve high levels of accuracy and robustness under challenging weather conditions. Furthermore, they developed a cooperative perception system, Faster-HEAL. Cooperative perception utilizes vehicle-to-vehicle and vehicle-to-infrastructure (V2X) communications to enable trains, vehicles, and certain infrastructure locations to share sensor data to enhance the object detection performance of the overall system.

The project made significant improvements to leading object detection solutions and demonstrated their effectiveness when integrated with the best available and NG automotive grade lidar sensor technologies. In particular, the project's outcomes demonstrate the feasibility of: (a) operating advanced deep learning-based 3D object detectors using lidar sensors at around 60 point-cloud FPS, which is significantly faster than real-time, while maintaining accurate detections, (b) achieving high levels of detection accuracy in terms of AP that can range between 80% for distant hard-to-detect objects to more than 90% AP for near easier-do-detect objects when operating in real-time under favorable conditions, (c) achieving both real-time speed at around 30 FPS and medium levels of accuracy that ranges between 40% to around 60% AP when operating under challenging weather conditions such as rain or snow, (d) potentially achieving significant gains in accuracy when deploying NG lidars that can capture high resolution point-cloud frames, which can lead to both substantial improvements and potential changes in object detection solutions, and (e) improving accuracy under cooperative perception when multiple systems share and fuse their multimodal sensor data.

While the team made significant progress as highlighted above, the project also demonstrates that further work is needed to develop a viable railroad safety system that can be installed in the field and on moving trains. In particular, there are further needs for: (a) Deploying advanced NG high-resolution lidar sensors and cameras on moving trains, vehicles, and/or on critical points within the railroad infrastructure to collect multimodal data under diverse weather conditions that include rain, snow, and fog. (b) Labeling such diverse-weather data for training and testing supervised deep learning methods. (c) Developing advanced domain-adaptation techniques to improve the performance of current object detection solutions under challenging conditions. (d) Exploring the area of cooperative perception further, especially under challenging conditions and a broader range of scenarios.

8 References

Real-time Object Detection and SPARK

X. Wu *et al.*, 'Sparse Fuse Dense: Towards High Quality 3D Detection with Depth Completion', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5418–5427.

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.

S. Shi et al., 'PV-RCNN++: Point-voxel feature set abstraction with local vector representation for 3D object detection', arXiv preprint arXiv:2102. 00463, 2021.

Y. Chen, Y. Li, X. Zhang, J. Sun, and J. Jia, 'Focal Sparse Convolutional Networks for 3D Object Detection', in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5428–5437.

S. Pang, D. Morris, and H. Radha, 'CLOCs: Camera-LiDAR object candidates fusion for 3D object detection', in 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2020, pp. 10386–10393.

J. Deng, S. Shi, P. Li, W. Zhou, Y. Zhang, and H. Li, 'Voxel r-cnn: Towards high performance voxel-based 3d object detection', in Proceedings of the AAAI Conference on Artificial Intelligence, 2021, vol. 35, pp. 1201–1209.

S. Shi et al., 'PV-RCNN++: Point-voxel feature set abstraction with local vector representation for 3D object detection', International Journal of Computer Vision, vol. 131, no. 2, pp. 531–551, 2023.

H. Wu, C. Wen, W. Li, X. Li, R. Yang, and C. Wang, 'Transformation-equivariant 3D object detection for autonomous driving', in Proceedings of the AAAI Conference on Artificial Intelligence, 2023, vol. 37, pp. 2795–2802.

H. Wu, J. Deng, C. Wen, X. Li, C. Wang, and J. Li, 'CasA: A cascade attention network for 3-D object detection from LiDAR point clouds', IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1–11, 2022.

S. Shi, X. Wang, and H. Li, 'Pointrcnn: 3d object proposal generation and detection from point cloud', in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 770–779.

B. Cheng, L. Sheng, S. Shi, M. Yang, and D. Xu, 'Back-tracing representative points for voting-based 3d object detection in point clouds', in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 8963–8972.

H. Wang et al., 'Rbgnet: Ray-based grouping for 3d object detection', in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 1110–1119.

C. R. Qi, L. Yi, H. Su, and L. J. Guibas, 'Pointnet++: Deep hierarchical feature learning on point sets in a metric space', Advances in neural information processing systems, vol. 30, 2017.

C. R. Qi, H. Su, K. Mo, and L. J. Guibas, ‘Pointnet: Deep learning on point sets for 3d classification and segmentation’, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 652–660.

Lang, A. Manor, and S. Avidan, ‘Samplenet: Differentiable point cloud sampling’, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 7578–7588.

Y. Yan, Y. Mao, and B. Li, ‘Second: Sparsely embedded convolutional detection’, Sensors, vol. 18, no. 10, p. 3337, 2018.

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.

W. Zheng, W. Tang, L. Jiang, and C.-W. Fu, ‘SE-SSD: Self-ensembling single-stage object detector from point cloud’, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 14494–14503.

Z. Yang, Y. Sun, S. Liu, and J. Jia, ‘3dssd: Point-based 3d single stage object detector’, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 11040–11048.

C. Chen, Z. Chen, J. Zhang, and D. Tao, ‘Sasa: Semantics-augmented set abstraction for point-based 3d object detection’, in Proceedings of the AAAI Conference on Artificial Intelligence, 2022, vol. 36, pp. 221–229.

Z. Yang, Y. Sun, S. Liu, X. Shen, and J. Jia, ‘Std: Sparse-to-dense 3d object detector for point cloud’, in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 1951–1960.

X. Li et al., ‘LoGoNet: Towards Accurate 3D Object Detection with Local-to-Global Cross-Modal Fusion’, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 17524–17534.

H. Wu, C. Wen, S. Shi, X. Li, and C. Wang, ‘Virtual Sparse Convolution for Multimodal 3D Object Detection’, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 21653–21662.

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, pp. 3354–3361.

Q. Xu, Y. Zhong, and U. Neumann, ‘Behind the curtain: Learning occluded shapes for 3d object detection’, in Proceedings of the AAAI Conference on Artificial Intelligence, 2022, vol. 36, pp. 2893–2901.

H. Yang et al., ‘Graph r-cnn: Towards accurate 3d object detection with semantic-decorated local graph’, in European Conference on Computer Vision, 2022, pp. 662–679.

X. Wu *et al.*, ‘Sparse Fuse Dense: Towards High Quality 3D Detection with Depth Completion’, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5418–5427.

Q. Xia et al., '3-D HANet: A Flexible 3-D Heatmap Auxiliary Network for Object Detection', IEEE Transactions on Geoscience and Remote Sensing, vol. 61, pp. 1–13, 2023.

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.

S. Shi, Z. Wang, J. Shi, X. Wang, and H. Li, 'From points to parts: 3d object detection from point cloud with part-aware and part-aggregation network', IEEE transactions on pattern analysis and machine intelligence, vol. 43, no. 8, pp. 2647–2664, 2020.

OpenPCDet, 'OpenPCDet: An Open-source Toolbox for 3D Object Detection from Point Clouds', 2020. [Online]. Available: https://github.com/open-mmlab/OpenPCDet.

P. Sun et al., 'Scalability in perception for autonomous driving: Waymo open dataset', in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 2446–2454.

J. Gou, B. Yu, S. J. Maybank, and D. Tao, 'Knowledge distillation: A survey', International Journal of Computer Vision, vol. 129, pp. 1789–1819, 2021.

X. Dai et al., 'General instance distillation for object detection', in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 7842–7851.

L. Zhang and K. Ma, 'Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors', in International Conference on Learning Representations, 2020.

Z. Kang, P. Zhang, X. Zhang, J. Sun, and N. Zheng, 'Instance-conditional knowledge distillation for object detection', Advances in Neural Information Processing Systems, vol. 34, pp. 16468–16480, 2021.

Chawla, H. Yin, P. Molchanov, and J. Alvarez, 'Data-free knowledge distillation for object detection', in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 3289–3298.

P. Passban, Y. Wu, M. Rezagholizadeh, and Q. Liu, 'Alp-kd: Attention-based layer projection for knowledge distillation', in Proceedings of the AAAI Conference on artificial intelligence, 2021, vol. 35, pp. 13657–13665.

D. Chen et al., 'Cross-layer distillation with semantic calibration', in Proceedings of the AAAI Conference on Artificial Intelligence, 2021, vol. 35, pp. 7028–7036.

N. Passalis, M. Tzelepi, and A. Tefas, 'Heterogeneous knowledge distillation using information flow modeling', in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 2339–2348.

W. Park, D. Kim, Y. Lu, and M. Cho, 'Relational knowledge distillation', in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 3967–3976.

M. Klingner et al., 'X3KD: Knowledge Distillation Across Modalities, Tasks and Stages for Multi-Camera 3D Object Detection', in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 13343–13353.

Y. Liu, W. Zhang, and J. Wang, ‘Adaptive multi-teacher multi-level knowledge distillation’, Neurocomputing, vol. 415, pp. 106–113, 2020.

J. Guo et al., ‘Distilling object detectors via decoupled features’, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2154–2164.

T. Feng and M. Wang, ‘Response-based distillation for incremental object detection’, arXiv preprint arXiv:2110. 13471, 2021.

J. Yang, S. Shi, R. Ding, Z. Wang, and X. Qi, ‘Towards efficient 3d object detection with knowledge distillation’, Advances in Neural Information Processing Systems, vol. 35, pp. 21300–21313, 2022.

L. Huang et al., ‘On the Importance of Pretrained Knowledge Distillation for 3D Object Detection’. 2023.

G. Chen, W. Choi, X. Yu, T. Han, and M. Chandraker, ‘Learning efficient object detection models with knowledge distillation’, Advances in neural information processing systems, vol. 30, 2017.

Z. Wang, D. Li, C. Luo, C. Xie, and X. Yang, ‘Distillbev: Boosting multi-camera 3d object detection with cross-modal knowledge distillation’, in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 8637–8646.

X.-F. Han, J. S. Jin, M.-J. Wang, W. Jiang, L. Gao, and L. Xiao, ‘A review of algorithms for filtering the 3D point cloud’, Signal Processing: Image Communication, vol. 57, pp. 103–112, 2017.

Z. Liu, T. Huang, B. Li, X. Chen, X. Wang, and X. Bai, ‘EPNet++: Cascade bi-directional fusion for multi-modal 3D object detection’, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022.

T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, ‘Focal loss for dense object detection’, in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988.

N. J. Beaudry and R. Renner, ‘An intuitive proof of the data processing inequality’, arXiv preprint arXiv:1107. 0740, 2011.

X. Bai et al., ‘Transfusion: Robust lidar-camera fusion for 3d object detection with transformers’, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 1090–1099.

Z. Liu et al., ‘Bevfusion: Multi-task multi-sensor fusion with unified bird’s-eye view representation’, in 2023 IEEE international conference on robotics and automation (ICRA), 2023, pp. 2774–2781.

H. Wang et al., ‘Unitr: A unified and efficient multi-modal transformer for bird’s-eye-view representation’, in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 6792–6802.

H. Sheng et al., ‘Improving 3d object detection with channel-wise transformerbai2022transfusion’, in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 2743–2752.

H. A. Hoang, D. C. Bui, and M. Yoo, ‘TSSTDet: Transformation-based 3-D Object Detection via a Spatial Shape Transformer’, IEEE Sensors Journal, 2024.

H. Zhang, L. Liu, Y. Huang, X. Lei, L. Tong, and B. Wen, ‘InstKD: Towards Lightweight 3D Object Detection With Instance-Aware Knowledge Distillation’, IEEE Transactions on Intelligent Vehicles, pp. 1–13, 2024.

L. Zhang, R. Dong, H.-S. Tai, and K. Ma, ‘Pointdistiller: structured knowledge distillation towards efficient and compact 3d detection’, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 21791–21801.

H. Cho, J. Choi, G. Baek, and W. Hwang, ‘itkd: Interchange transfer-based knowledge distillation for 3d object detection’, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 13540–13549.

P. Palmer, M. Krüger, S. Schütte, R. Altendorfer, G. Adam, and T. Bertram, ‘LEROjD: Lidar Extended Radar-Only Object Detection’, in European Conference on Computer Vision, 2024, pp. 379–396.

Y. Li, S. Xu, M. Lin, J. Yin, B. Zhang, and X. Cao, ‘Representation disparity-aware distillation for 3d object detection’, in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 6715–6724.

S. Zhou, W. Liu, C. Hu, S. Zhou, and C. Ma, ‘Unidistill: A universal cross-modality knowledge distillation framework for 3D object detection in bird’s-eye view’, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 5116–5125.

J. Li, Z. Liu, J. Hou, and D. Liang, ‘Dds3d: Dense pseudo-labels with dynamic threshold for semi-supervised 3d object detection’, arXiv preprint arXiv:2303. 05079, 2023.

H. Wang, Y. Cong, O. Litany, Y. Gao, and L. J. Guibas, ‘3dioumatch: Leveraging iou prediction for semi-supervised 3d object detection’, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 14615–14624.

J. Zhang and J. Liu, ‘Voxel-to-pillar: Knowledge distillation of 3d object detection in point cloud’, in Proceedings of the 4th European Symposium on Software Engineering, 2023, pp. 99–104.

L. Li et al., ‘Detkds: Knowledge distillation search for object detectors’, in Forty-first International Conference on Machine Learning, 2024.

Z. Zheng et al., ‘Localization distillation for dense object detection’, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 9407–9416.

Z. Zhang et al., ‘Atf-3d: Semi-supervised 3d object detection with adaptive thresholds filtering based on confidence and distance’, IEEE Robotics and Automation Letters, vol. 7, no. 4, pp. 10573–10580, 2022.

X. Hao et al., ‘Mbfusion: A new multi-modal bev feature fusion method for hd map construction’, in 2024 IEEE International Conference on Robotics and Automation (ICRA), 2024, pp. 15922–15928.

X. Bai et al., ‘Sgdet3d: Semantics and geometry fusion for 3d object detection using 4d radar and camera’, IEEE Robotics and Automation Letters, 2024.

J. Song, L. Zhao, and K. A. Skinner, ‘Lirafusion: Deep adaptive lidar-radar fusion for 3d object detection’, in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 18250–18257.

Weather-Resilient Object Detection, MSU Four Seasons Dataset, and WISDOM

D. Kent et al., ‘MSU-4S - The Michigan State University Four Seasons Dataset’, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 22658–22667.

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.

S. Shi et al., ‘PV-RCNN++: Point-voxel feature set abstraction with local vector representation for 3D object detection’, arXiv preprint arXiv:2102. 00463, 2021.

Y. Chen, Y. Li, X. Zhang, J. Sun, and J. Jia, ‘Focal Sparse Convolutional Networks for 3D Object Detection’, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5428–5437.

S. Pang, D. Morris, and H. Radha, ‘CLOCs: Camera-LiDAR object candidates fusion for 3D object detection’, in 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2020, pp. 10386–10393.

J. Deng, S. Shi, P. Li, W. Zhou, Y. Zhang, and H. Li, ‘Voxel r-cnn: Towards high performance voxel-based 3d object detection’, in Proceedings of the AAAI Conference on Artificial Intelligence, 2021, vol. 35, pp. 1201–1209.

S. Shi et al., ‘PV-RCNN++: Point-voxel feature set abstraction with local vector representation for 3D object detection’, International Journal of Computer Vision, vol. 131, no. 2, pp. 531–551, 2023.

H. Wu, C. Wen, W. Li, X. Li, R. Yang, and C. Wang, ‘Transformation-equivariant 3D object detection for autonomous driving’, in Proceedings of the AAAI Conference on Artificial Intelligence, 2023, vol. 37, pp. 2795–2802.

H. Wu, J. Deng, C. Wen, X. Li, C. Wang, and J. Li, ‘CasA: A cascade attention network for 3-D object detection from LiDAR point clouds’, IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1–11, 2022.

S. Shi, X. Wang, and H. Li, ‘Pointrcnn: 3d object proposal generation and detection from point cloud’, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 770–779.

B. Cheng, L. Sheng, S. Shi, M. Yang, and D. Xu, ‘Back-tracing representative points for voting-based 3d object detection in point clouds’, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 8963–8972.

H. Wang et al., ‘Rbgnet: Ray-based grouping for 3d object detection’, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 1110–1119.

C. R. Qi, L. Yi, H. Su, and L. J. Guibas, ‘Pointnet++: Deep hierarchical feature learning on point sets in a metric space’, Advances in neural information processing systems, vol. 30, 2017.

C. R. Qi, H. Su, K. Mo, and L. J. Guibas, ‘Pointnet: Deep learning on point sets for 3d classification and segmentation’, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 652–660.

I. Lang, A. Manor, and S. Avidan, ‘Samplenet: Differentiable point cloud sampling’, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 7578–7588.

Y. Yan, Y. Mao, and B. Li, ‘Second: Sparsely embedded convolutional detection’, Sensors, vol. 18, no. 10, p. 3337, 2018.

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.

W. Zheng, W. Tang, L. Jiang, and C.-W. Fu, ‘SE-SSD: Self-ensembling single-stage object detector from point cloud’, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 14494–14503.

Z. Yang, Y. Sun, S. Liu, and J. Jia, ‘3dssd: Point-based 3d single stage object detector’, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 11040–11048.

C. Chen, Z. Chen, J. Zhang, and D. Tao, ‘Sasa: Semantics-augmented set abstraction for point-based 3d object detection’, in Proceedings of the AAAI Conference on Artificial Intelligence, 2022, vol. 36, pp. 221–229.

Z. Yang, Y. Sun, S. Liu, X. Shen, and J. Jia, ‘Std: Sparse-to-dense 3d object detector for point cloud’, in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 1951–1960.

X. Li et al., ‘LoGoNet: Towards Accurate 3D Object Detection with Local-to-Global Cross-Modal Fusion’, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 17524–17534.

H. Wu, C. Wen, S. Shi, X. Li, and C. Wang, ‘Virtual Sparse Convolution for Multimodal 3D Object Detection’, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 21653–21662.

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, pp. 3354–3361.

Q. Xu, Y. Zhong, and U. Neumann, ‘Behind the curtain: Learning occluded shapes for 3d object detection’, in Proceedings of the AAAI Conference on Artificial Intelligence, 2022, vol. 36, pp. 2893–2901.

H. Yang et al., ‘Graph r-cnn: Towards accurate 3d object detection with semantic-decorated local graph’, in European Conference on Computer Vision, 2022, pp. 662–679.

Q. Xia et al., ‘3-D HANet: A Flexible 3-D Heatmap Auxiliary Network for Object Detection’, IEEE Transactions on Geoscience and Remote Sensing, vol. 61, pp. 1–13, 2023.

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.

S. Shi, Z. Wang, J. Shi, X. Wang, and H. Li, ‘From points to parts: 3d object detection from point cloud with part-aware and part-aggregation network’, IEEE transactions on pattern analysis and machine intelligence, vol. 43, no. 8, pp. 2647–2664, 2020.

OpenPCDet, ‘OpenPCDet: An Open-source Toolbox for 3D Object Detection from Point Clouds’, 2020. [Online]. Available: https://github.com/open-mmlab/OpenPCDet.

P. Sun et al., ‘Scalability in perception for autonomous driving: Waymo open dataset’, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 2446–2454.

J. Gou, B. Yu, S. J. Maybank, and D. Tao, ‘Knowledge distillation: A survey’, International Journal of Computer Vision, vol. 129, pp. 1789–1819, 2021.

X. Dai et al., ‘General instance distillation for object detection’, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 7842–7851.

L. Zhang and K. Ma, ‘Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors’, in International Conference on Learning Representations, 2020.

Z. Kang, P. Zhang, X. Zhang, J. Sun, and N. Zheng, ‘Instance-conditional knowledge distillation for object detection’, Advances in Neural Information Processing Systems, vol. 34, pp. 16468–16480, 2021.

A. Chawla, H. Yin, P. Molchanov, and J. Alvarez, ‘Data-free knowledge distillation for object detection’, in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 3289–3298.

P. Passban, Y. Wu, M. Rezagholizadeh, and Q. Liu, ‘Alp-kd: Attention-based layer projection for knowledge distillation’, in Proceedings of the AAAI Conference on artificial intelligence, 2021, vol. 35, pp. 13657–13665.

D. Chen et al., ‘Cross-layer distillation with semantic calibration’, in Proceedings of the AAAI Conference on Artificial Intelligence, 2021, vol. 35, pp. 7028–7036.

N. Passalis, M. Tzelepi, and A. Tefas, ‘Heterogeneous knowledge distillation using information flow modeling’, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 2339–2348.

W. Park, D. Kim, Y. Lu, and M. Cho, ‘Relational knowledge distillation’, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 3967–3976.

M. Klingner et al., ‘X3KD: Knowledge Distillation Across Modalities, Tasks and Stages for Multi-Camera 3D Object Detection’, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 13343–13353.

Y. Liu, W. Zhang, and J. Wang, ‘Adaptive multi-teacher multi-level knowledge distillation’, Neurocomputing, vol. 415, pp. 106–113, 2020.

J. Guo et al., ‘Distilling object detectors via decoupled features’, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2154–2164.

T. Feng and M. Wang, ‘Response-based distillation for incremental object detection’, arXiv preprint arXiv:2110. 13471, 2021.

J. Yang, S. Shi, R. Ding, Z. Wang, and X. Qi, ‘Towards efficient 3d object detection with knowledge distillation’, Advances in Neural Information Processing Systems, vol. 35, pp. 21300–21313, 2022.

L. Huang et al., ‘On the Importance of Pretrained Knowledge Distillation for 3D Object Detection’. 2023.

G. Chen, W. Choi, X. Yu, T. Han, and M. Chandraker, ‘Learning efficient object detection models with knowledge distillation’, Advances in neural information processing systems, vol. 30, 2017.

Z. Wang, D. Li, C. Luo, C. Xie, and X. Yang, ‘Distillbev: Boosting multi-camera 3d object detection with cross-modal knowledge distillation’, in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 8637–8646.

X.-F. Han, J. S. Jin, M.-J. Wang, W. Jiang, L. Gao, and L. Xiao, ‘A review of algorithms for filtering the 3D point cloud’, Signal Processing: Image Communication, vol. 57, pp. 103–112, 2017.

Z. Liu, T. Huang, B. Li, X. Chen, X. Wang, and X. Bai, ‘EPNet++: Cascade bi-directional fusion for multi-modal 3D object detection’, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022.

T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, ‘Focal loss for dense object detection’, in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988.

N. J. Beaudry and R. Renner, ‘An intuitive proof of the data processing inequality’, arXiv preprint arXiv:1107. 0740, 2011.

X. Bai et al., ‘Transfusion: Robust lidar-camera fusion for 3d object detection with transformers’, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 1090–1099.

Z. Liu et al., ‘Bevfusion: Multi-task multi-sensor fusion with unified bird’s-eye view representation’, in 2023 IEEE international conference on robotics and automation (ICRA), 2023, pp. 2774–2781.

H. Wang et al., ‘Unitr: A unified and efficient multi-modal transformer for bird’s-eye-view representation’, in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 6792–6802.

H. Sheng et al., ‘Improving 3d object detection with channel-wise transformerbai2022transfusion’, in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 2743–2752.

H. A. Hoang, D. C. Bui, and M. Yoo, ‘TSSTDet: Transformation-based 3-D Object Detection via a Spatial Shape Transformer’, IEEE Sensors Journal, 2024.

D. Kent et al., ‘MSU-4S-The Michigan State University Four Seasons Dataset’, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 22658–22667.

Y. Ganin et al., ‘Domain-adversarial training of neural networks’, Journal of machine learning research, vol. 17, no. 59, pp. 1–35, 2016.

M. Hnewa and H. Radha, ‘Integrated multiscale domain adaptive YOLO’, IEEE Transactions on Image Processing, vol. 32, pp. 1857–1867, 2023.

C. Zhang et al., ‘Robust-FusionNet: Deep multimodal sensor fusion for 3-D object detection under severe weather conditions’, IEEE Transactions on Instrumentation and Measurement, vol. 71, pp. 1–13, 2022.

A. T. Do and M. Yoo, ‘LossDistillNet: 3D object detection in point cloud under harsh weather conditions’, IEEE Access, vol. 10, pp. 84882–84893, 2022.

J. Ho, A. Jain, and P. Abbeel, ‘Denoising diffusion probabilistic models’, Advances in neural information processing systems, vol. 33, pp. 6840–6851, 2020.

X. Huang, H. Wu, X. Li, X. Fan, C. Wen, and C. Wang, ‘Sunshine to rainstorm: Cross-weather knowledge distillation for robust 3d object detection’, in Proceedings of the AAAI Conference on Artificial Intelligence, 2024, vol. 38, pp. 2409–2416.

J. Gao, J. Zhang, X. Liu, T. Darrell, E. Shelhamer, and D. Wang, ‘Back to the source: Diffusion-driven adaptation to test-time corruption’, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 11786–11796.

C. A. Diaz-Ruiz et al., ‘Ithaca365: Dataset and Driving Perception Under Repeated and Challenging Weather Conditions’, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 21383–21392.

Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. Van Gool, ‘Domain Adaptive Faster R-CNN for Object Detection in the Wild’, in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 3339–3348.

D. Hegde, V. Kilic, V. Sindagi, A. B. Cooper, M. Foster, and V. M. Patel, ‘Source-free unsupervised domain adaptation for 3d object detection in adverse weather’, in 2023 IEEE International Conference on Robotics and Automation (ICRA), 2023, pp. 6973–6980.

V. Kilic, D. Hegde, V. Sindagi, A. B. Cooper, M. A. Foster, and V. M. Patel, 'Lidar light scattering augmentation (lisa): Physics-based simulation of adverse weather conditions for 3d object detection', arXiv preprint arXiv:2107. 07004, 2021.

M. Hahner, C. Sakaridis, D. Dai, and L. Van Gool, 'Fog simulation on real LiDAR point clouds for 3D object detection in adverse weather', in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 15283–15292.

G. Wu, T. Cao, B. Liu, X. Chen, and Y. Ren, 'Towards universal LiDAR-based 3D object detection by multi-domain knowledge transfer', in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 8669–8678.

Q. Xu, Y. Zhou, W. Wang, C. R. Qi, and D. Anguelov, 'Spg: Unsupervised domain adaptation for 3d object detection via semantic point generation', in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 15446–15456.

I. Goodfellow et al., 'Generative adversarial networks', Communications of the ACM, vol. 63, no. 11, pp. 139–144, 2020.

A. Q. Nichol and P. Dhariwal, 'Improved denoising diffusion probabilistic models', in International conference on machine learning, 2021, pp. 8162–8171.

J. Song, C. Meng, and S. Ermon, 'Denoising diffusion implicit models', arXiv preprint arXiv:2010. 02502, 2020.

H. Ran, V. Guizilini, and Y. Wang, 'Towards realistic scene generation with lidar diffusion models', in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 14738–14748.

Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. Van Gool, 'Domain adaptive faster r-cnn for object detection in the wild', in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 3339–3348.

M. Hnewa and H. Radha, 'Integrated Generative-Model Domain-Adaptation for Object Detection under Challenging Conditions', in 2022 IEEE 95th Vehicular Technology Conference:(VTC2022-Spring), 2022, pp. 1–5.

M. Hahner et al., 'Lidar snowfall simulation for robust 3d object detection', in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 16364–16374.

A. Piroli, V. Dallabetta, M. Walessa, D. Meissner, J. Kopp, and K. Dietmayer, 'Robust 3d object detection in cold weather conditions', in 2022 IEEE Intelligent Vehicles Symposium (IV), 2022, pp. 287–294.

X. Huang et al., 'L4dr: Lidar-4dradar fusion for weather-robust 3d object detection', in Proceedings of the AAAI Conference on Artificial Intelligence, 2025, vol. 39, pp. 3806–3814.

K. Qian, S. Zhu, X. Zhang, and L. E. Li, 'Robust multimodal vehicle detection in foggy weather using complementary lidar and radar signals', in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 444–453.

T. Vattem, G. Sebastian, and L. Lukic, ‘Rethinking lidar object detection in adverse weather conditions’, in 2022 International Conference on Robotics and Automation (ICRA), 2022, pp. 5093–5099.

N. A. M. Mai, P. Duthon, P. H. Salmane, L. Khoudour, A. Crouzil, and S. A. Velastin, ‘Camera and LiDAR analysis for 3D object detection in foggy weather conditions’, in 2022 12th International Conference on Pattern Recognition Systems (ICPRS), 2022, pp. 1–7.

M. L. Menéndez, J. A. Pardo, L. Pardo, and M. del C. Pardo, ‘The jensen-shannon divergence’, Journal of the Franklin Institute, vol. 334, no. 2, pp. 307–318, 1997.

X. Lu and H. Radha, ‘DALI: Domain Adaptive LiDAR Object Detection via Distribution-Level and Instance-Level Pseudolabel Denoising’, IEEE Transactions on Robotics, vol. 40, pp. 3866–3878, 2024.

V. Kilic, D. Hegde, A. B. Cooper, V. M. Patel, and M. Foster, ‘Lidar light scattering augmentation (lisa): Physics-based simulation of adverse weather conditions for 3d object detection’, in ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2025, pp. 1–5.

H. Caesar et al., ‘nuscenes: A multimodal dataset for autonomous driving’, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 11621–11631.

Q. Hu, D. Liu, and W. Hu, ‘Density-insensitive unsupervised domain adaption on 3d object detection’, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 17556–17566.

J. Yang, S. Shi, Z. Wang, H. Li, and X. Qi, ‘St3d: Self-training for unsupervised domain adaptation on 3d object detection’, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 10368–10378.

J. Yang, S. Shi, Z. Wang, H. Li, and X. Qi, ‘St3d++: Denoised self-training for unsupervised domain adaptation on 3d object detection’, IEEE transactions on pattern analysis and machine intelligence, vol. 45, no. 5, pp. 6354–6371, 2022.

Y. Wang et al., ‘Train in germany, test in the usa: Making 3d object detectors generalize’, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11713–11723.

M. Bijelic et al., ‘Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather’, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11682–11692.

Y.-C. Shih, W.-H. Liao, W.-C. Lin, S.-K. Wong, and C.-C. Wang, ‘Reconstruction and synthesis of lidar point clouds of spray’, IEEE Robotics and Automation Letters, vol. 7, no. 2, pp. 3765–3772, 2022.

Z. Li, Y. Yao, Z. Quan, L. Qi, Z.-H. Feng, and W. Yang, ‘Adaptation via proxy: Building instance-aware proxy for unsupervised domain adaptive 3d object detection’, IEEE Transactions on Intelligent Vehicles, vol. 9, no. 2, pp. 3478–3492, 2023.

M. K. Wozniak, M. Hansson, M. Thiel, and P. Jensfelt, 'Uada3d: Unsupervised adversarial domain adaptation for 3d object detection with sparse lidar and large domain gaps', IEEE Robotics and Automation Letters, 2024.

J. Lin, 'Divergence measures based on the Shannon entropy', IEEE Trans. Inf. Theory, vol. 37, pp. 145–151, 1991.

X. Zhai, H. Huang, N. N. Sze, Z. Song, and K. K. Hon, 'Diagnostic analysis of the effects of weather condition on pedestrian crash severity', Accident Analysis & Prevention, vol. 122, pp. 318–324, 2019.

Z. Yihan et al., 'Learning transferable features for point cloud detection via 3d contrastive co-training', Advances in Neural Information Processing Systems, vol. 34, pp. 21493–21504, 2021.

X. Lu and H. Radha, 'DALI: Domain Adaptive LiDAR Object Detection via Distribution-level and Instance-level Pseudo Label Denoising', IEEE Transactions on Robotics, 2024.

C. Saltori, S. Lathuiliére, N. Sebe, E. Ricci, and F. Galasso, 'Sf-uda 3d: Source-free unsupervised domain adaptation for lidar-based 3d object detection', in *2020 International Conference on 3D Vision (3DV)*, 2020, pp. 771–780.

Cooperative Perception and Faster-HEAL

T. Huang, J. Liu, X. Zhou, D. C. Nguyen, M. R. Azghadi, Y. Xia, Q. L. Han, and S. Sun, V2X cooperative perception for autonomous driving: Recent advances and challenges. arXiv preprint arXiv:2310.03525, 2023.

Y. Hu, S. Fang, Z. Lei, Y. Zhong, S. Chen, Where2comm: Communication-efficient collaborative perception via spatial confidence maps. Advances in neural information processing systems, 2022.

Z. Lei, S. Ren, Y. Hu, W. Zhang, and S. Chen, Latency-aware collaborative perception. In Computer Vision–ECCV 2022: 17th European Conference, 2022.

Y. Lu, Q. Li, B. Liu, M. Dianat, C. Feng, S. Chen, and Y. Wang, "Robust collaborative 3D object detection in presence of pose errors," in Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), London, United Kingdom, 2023, pp. 4812–4818.

Y. C. Liu, J. Tian, N. Glaser, and Z. Kira, 2020. When2com: Multi-agent perception via communication graph grouping. In Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition (pp. 4106-4115).

Y. Hu, J. Peng, S. Liu, J. Ge, S. Liu, and S. Chen. Communication efficient collaborative perception via information filling with codebook. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 15481–15490, 2024.

S. Su, Y. Li, S. He, S. Han, C. Feng, C. Ding, and F. Miao, "Uncertainty quantification of collaborative detection for self-driving," in Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), London, UK, 2023, pp. 5588–5594.

X. Li, J. Yin, W. Li, C. Xu, R. Yang, and J. Shen, “Di-v2x: Learning domain-invariant representation for vehicle-infrastructure collaborative 3d object detection,” in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, n. 4, 2024, pp. 3208–3215.

H. Xiang, R. Xu, and J. Ma, “HM-ViT: Hetero-modal vehicle-to-vehicle cooperative perception with vision transformer,” in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2023, pp. 284–295.

P. Zhi, L. Jiang, X. Yang, X. Wang, H. W. Li, Q. Zhou, et al.. Cross-domain generalization for lidar-based 3d object detection in infrastructure and vehicle environments. Sensors. January 2025. Y. Lu *et al.*, “An extensible framework for open heterogeneous collaborative perception,” *arXiv:2401.13964*, Jan. 2024.

R. Xu, J. Li, X. Dong, H. Yu, and J. Ma. Bridging the domain gap for multi-agent perception. In 2023 IEEE International Conference on Robotics and Automation (ICRA), pages 6035–6042, 2023.

L. Xin, G. Zhou, Z. Yu, D. Wang, T. Luo, X. Fu, J. Li, PnPDA+: A Meta Feature-Guided Domain Adapter for Collaborative Perception. World Electric Vehicle Journal. January 2025.

Y. Xia, Q. Yuan, G. Luo, X. Fu, Y. Li, X. Zhu, et al., One is Plenty: A Polymorphic Feature Interpreter for Immutable Heterogeneous Collaborative Perception. In Proceedings of the Computer Vision and Pattern Recognition Conference 2025 (pp. 1592-1601).

M. Singha, H. Pal, A. Jha, and B. Banerjee. Ad-clip: Adapting domains in prompt space using clip. In 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), pages 4357–4366, 2023.

M. Cai, H. Liu, S. K. Mustikovela, G. P. Meyer, Y. Chai, D, Park, and Y. J. Lee. Vipllava: Making large multimodal models understand arbitrary visual prompts. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.

D. Cheng, Z. Xu, X. Jiang, N. Wang, D. Li, and X. Gao. Disentangled prompt representation for domain generalization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 23595–23604, 2024.

X. Zhang, F. Wei, Y. Wang, W. Zhao, F. Li, X. Chu. UPRE: Zero-Shot Domain Adaptation for Object Detection via Unified Prompt and Representation Enhancement. arXiv preprint arXiv:2507.00721, 2025.

Y. A. Farrukh, S. Wali, I. Khan, N. D. Bastian. Prmpt2Adpt: Prompt-Based Zero-Shot Domain Adaptation for Resource-Constrained Environments. arXiv preprint arXiv:2506.16994. 2025 Jun 20.

H. R. Medeiros, A. Belal, S. Muralidharan, E. Granger, M. Pedersoli. Visual Modality Prompt for Adapting Vision-Language Object Detectors. arXiv preprint arXiv:2412.00622. 2024.

S. Xie, R. Girshick, P. Doll´ar, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2017.

T. Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár. Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision 2017 (pp. 2980-2988).

R. Girshick. Fast r-cnn. In Proceedings of the IEEE international conference on computer vision 2015 (pp. 1440-1448).

Z. Liu, H. Mao, C. YuanWu, C. Feichtenhofer, T. Darrell, and S. Xie. A convnet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11976–11986, 2022.

R. Xu, H. Xiang, X. Xia, X. Han, J. Li, and J. Ma. Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to- vehicle communication. In 2022 International Conference on Robotics and Automation (ICRA), pp. 2583–2589. IEEE, 2022.

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, pp. 12697–12705, 2019.

Y. Yan, Y. Mao, and B. Li. Second: Sparsely embedded convolutional detection. Sensors, 18(10):3337, 2018.

M. Tan and Q. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning, pp. 6105–6114. PMLR, 2019.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778, 2016.

J. Philion and S. Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16, pp. 194–210. Springer, 2020.

Q. Chen, X. Ma, S. Tang, J. Guo, Q. Yang, S. Fu. F-cooper: Feature based cooperative perception for autonomous vehicle edge computing system using 3D point clouds. InProceedings of the 4th ACM/IEEE Symposium on Edge Computing 2019 Nov 7 (pp. 88-100).

Y. Li, S. Ren, P. Wu, S. Chen, C. Feng, W. Zhang. Learning distilled collaboration graph for multi-agent perception. Advances in Neural Information Processing Systems. 2021 Dec 6;34:29541-52.

R. Xu, H. Xiang, Z. Tu, X. Xia, M. H. Yang, J. Ma. V2x-vit: Vehicle-to-everything cooperative perception with vision transformer. In European conference on computer vision 2022 Oct 23 (pp. 107-124). Cham: Springer Nature Switzerland.

R. Xu, Z. Tu, H. Xiang, W. Shao, B. Zhou, and J. Ma. Cobevt: Cooperative bird's eye view semantic segmentation with sparse transformers. arXiv preprint arXiv:2207.02202, 2022.

J. D. Carroll, J. J. Chang. Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. Psychometrika. 1970 Sep;35(3):283-319.

Appendix A. A Generative Framework for Rail-Domain Perception Data

JACL: A Generative Foundational Model for Rail-Domain Vehicles' Perception Data Using Visual Semantic Segmentation

Daniel Kent
Electrical and Computer Engineering Department
Michigan State University College of Engineering
East Lansing, Michigan, USA
kentdan3@msu.edu

Hayder Radha
Electrical and Computer Engineering Department
Michigan State University College of Engineering
East Lansing, Michigan, USA
radha@msu.edu

***Abstract*—Semantic image segmentation provides extremely granular image understanding for intelligent and autonomous vehicle platforms. A major factor in the development of these algorithms comes courtesy of the sheer scale of labeled data publicly available. However, more niche domains, such as rail vehicles, face a shortfall of publicly available data. In this paper, we address shortages of semantically labeled data in such thinly-sampled domains by adapting modern diffusion based generative foundational models. Our technique, which we call Joint Augmented ControlNet-LoRA (JACL), combines two widely used diffusion fine-tuning methods in a novel manner to jointly bias the network towards the target domain. JACL differs from existing techniques which aim to label synthetic imagery during or after generation; instead, JACL flips the problem on its head, using the desired ground truth labels as the input, resulting in imagery that closely conforms with the input semantic labels. Furthermore, JACL can easily generate synthetic imagery in various weather and lighting conditions by modifying input text prompts, with no additional training. We qualitatively and quantitatively demonstrate how the output imagery from JACL conforms to the desired scene structure while exhibiting statistically significant differences that can augment downstream algorithm development and testing.**

***Index Terms*—synthetic data, diffusion, semantic segmentation, rail**

I. Introduction

With the introduction of neural networks, especially convolutional neural networks, machine learning-based computer vision algorithms are now capable of accomplishing many tasks that have previously been challenging to accomplish at scale and speed with high accuracy, including image classification, object detection, and semantic segmentation. Of these tasks, semantic segmentation arguably requires the most granular labels of these three tasks. Semantic segmented imagery, with its pixel-precision labels, has many uses; of particular interest in ground vehicle research and development, semantically labeled camera imagery can more easily be fused with data from other sensors such as lidar and radar, providing precise perception capabilities that are required for robust and reliable systems.

This work was supported by a Department of Transportation/Federal Railroad Administration grant.

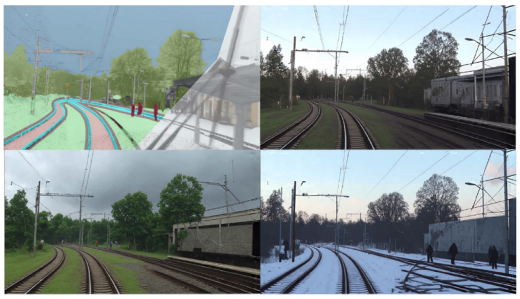


Fig. 1. Examples of synthetic rail domain data generated by JACL, with text prompts used to generate specific conditions, including "Spring Morning" (top right), "Summer Thunderstorms" (bottom left), and "Winter Sunset" (bottom right). The input ground truth segmentation mask overlayed on the corresponding ground truth image appears in the top left. All generated images precisely follow the structural components of the input visual semantics, such as the curvature of the train tracks, while exhibiting strong and diverse generative capabilities in terms of ambient conditions. The white area in the ground truth is a "don't care" labeled region.

Large amounts of data are required to train and test semantic segmentation networks, and for certain domains such datasets are widely available. Millions of labeled samples are available among the most popular publicly accessible semantic segmentation datasets, including COCO [1] (200k), OpenImages [2], [3] (944k), and the SA-1B dataset [4] (11M). Certain application domains such as autonomous driving also have a large amount of data available; between KITTI-360 [5] (78k), nuScenes [6] (93k), and CityScapes [7] (25k) there are over 100,000 samples available for training semantic segmentation networks intended for vehicular applications. However, more niche domains such as railroad environments do not enjoy the same scale of publicly available data: we only identified one such dataset, RailSem19 [8], which contains a mere 8500 total samples, of which only 3500 were used in their dataset analysis. When training and testing image detection methods, the limited scope of the underlying dataset raises questions

as to the robustness of methods that are trained and/or tested on these data. Additional data would alleviate much of these concerns, though establishing a minimum amount of data required to fully validate image detection algorithms remains an open question.

Techniques to bolster training data for dataset-limited domains have evolved and advanced significantly over time. Early techniques focused on data augmentation using intrinsic image regularization and randomization, including randomized image transformations, noise injection, and distortion. Generative Adversarial Networks (GANs), were later developed as a means to learn the underlying statistics from an underlying dataset to produce statistically similar synthetic samples [9], which was soon applied to image datasets to create some of the first modern synthetic image generation networks [10]. Shortcomings in GAN architecture for image generation eventually led to the development of autoregression and diffusion-based networks [11], including Stable Diffusion, a popular architecture employing latent-space based techniques to reduce required system resources [12].

While modern diffusion-based image generation networks can output realistic synthetic imagery with little to no fine tuning, there are several problems that must be addressed before these data can be used to train and test semantic image segmentation algorithms for our intended domain. First, pretrained stable diffusion networks are trained on a variety of image data that includes non-photographic imagery, posing a potential issue with generated data not conforming to our desired target domain. Second, even with potentially photorealistic synthetic imagery, labeling these images requires additional effort.

In this paper, we introduce a novel framework that aims at generating realistic and structurally precise imagery by exploiting semantic segmentation information. The proposed framework, JACL, is particularly useful for domains that have major shortages in available data. In that context, we focus in this paper on the rail-domain to demonstrate the effectiveness of JACL. The main contributions of this works can be summarized as follows:

- We develop a novel architecture that combines the Low-Rank Adaptation (LoRA) [13] and ControlNet [14], [15] architectures to fine-tune our model for our specific domain. Once trained, our method can create novel synthetic imagery using a combination of text labels and a ground truth image segmentation label to produce realistic synthetic imagery that closely conforms to the input data structure and semantics.
- In addition, we demonstrate that our method can be further biased towards specific target weather, time, and lighting domains using straightforward input prompt changes without further training.
- Finally, we show that our method can synthesize sufficiently novel data by demonstrating that real data sharing the same underlying ground truth as samples generated with that ground truth have high PSNR and low MSE (demonstrating structural similarity) while also showing significant differences in Variational Autoencoder (VAE) activation compared to simpler whitening techniques.

II. Related Work

A. Labeling Imagery vs. Imagery From Labels

Prior work on creating semantically labeled synthetic imagery can be taxonomized into two broad categories: generative image labeling, and label-based sample generation. The first category, generative image labeling, uses the final synthetic sample, or information generated during sample generation, to label the synthetic sample. Generative image labeling can naively be accomplished using the same techniques as real imagery: either by employing human labelers, or running off-the-shelf semantic segmentation networks. The former is extremely labor- and time-intensive, and the latter, despite advancements in semantic image segmentation, may not be sufficiently accurate for critical applications. While recent techniques have been developed to use information from the underlying image generation process [16], an accuracy gap still exists, which may limit its use to automatically label generated imagery.

Instead of labeling synthetic data post-generation, either using existing techniques or by adapting generation information to create more accurate semantic segmentation masks, there exist techniques to use semantic segmentation labels themselves as an input for synthetic data generation. In theory, if the synthetic data generation algorithm is reasonably accurate, the input labels then automatically become the ground truth labels for the synthetic imagery, which eliminates the costly labeling step. However, image generation networks typically use text-based conditioning, which are converted by text embedding networks into conditioning weights, to bias the network towards generation of specific imagery. In order to successfully generate realistic imagery that conforms to a desired ground truth label, more precise guidance of the underlying algorithm is required.

B. Image Synthesis with Diffusion-Based Models

While there are many technologies capable of synthetic image generation as outlined in Section I, Stable Diffusion and its later variants are popular due to its relatively open architecture. As per its namesake, Stable Diffusion is based on diffusion, a machine-learning architecture developed by Sohl-Dickstein . This technique works by training a model to denoise data by first adding noise to training data, processing the image. After training, when a diffusion network is provided random noise as input, the network effectively generates a synthetic sample [17].

As with all diffusion-based methods, Stable Diffusion works by training an architecture to learn a probabilistic distribution by taking real data, adding noise, then attempting to reverse the noising. In the process, such architectures can encode spatial representations very well, allowing them to synthesize realistic imagery from multiple domains, including photographs, artwork, sketches, and even stylized 2D or 3D computer graphics. Stable Diffusion's main distinction is its use of latent space to

reduce the network's overall complexity; ordinarily, a diffusion network operating in RGB pixel space would require an extraordinary amount of resources to train and run inference; by leveraging the lower dimensionality latent space, Stable Diffusion and its later variants can output high resolution imagery using far lower computational resources compared to methods that operate directly in the pixel space [12].

C. Generative Image Control

As stated earlier, precisely controlling the output imagery to conform to specific feature types and target domains is critical to our desired application. While fine-tuning a baseline model's weights is possible, advances in lighter-weight methods such as Low-Rank Adaptation (LoRA) [13] have become increasingly popular. By instead training decomposition matrices that can then be injected into a baseline model, language models such as the input text embedding network within Stable Diffusion can effectively fine tune the network with far fewer parameters, which brings both runtime and storage advantages. LoRA fine-tuning has become a popular technique on Stable Diffusion based networks.

However, even LoRA cannot, by itself, enable pixel-level semantic control over output imagery. This shortcoming is instead addressed with ControlNet, a method that combines text-based conditioning with an image-based conditioning layer, which is processed together using a U-net architecture mirroring a fine-tuned first half of the Stable Diffusion U-net. In developing ControlNet, Zhang specifically demonstrated semantic segmentation as one control modality that can be used with ControlNet [14]. ControlNet++ further demonstrates the feasibility of semantic segmentation-based input control, especially when trained with additional loss functions intended to improve image coherence [15].

The potential of combining a pretrained Stable Diffusion network with a LoRA trained on domain-specific data and a ControlNet for pixel-level semantic specifity has been recognized for several applications. In plant science, Hartley utilized a LoRA trained on plant data, combined with a ControlNet trained on synthetic plant data generated using traditional 3D rendering techniques to achieve leaf-specific instance segmentation; when combined, they were able to create a generative plant identification dataset that was superior to existing CycleGAN-based techniques [18]. Similarly, Deng used similar techniques, albeit with a bounding-box based ControlNet, to augment an existing weed identification dataset, while demonstrating that downstream training on the augmented dataset exhibited superior performance up to a 3:2 synthetic to real data ratio [19].

III. Methodology

Taking inspiration from prior work on synthetic data generation using Stable Diffusion, we created Joint Augmented ControlNet-LoRA (JACL), which combines existing work towards image synthesis domain biasing and control into a unique architecture that maximizes the limited amount of publicly available rail domain data. In effect, we are augmenting

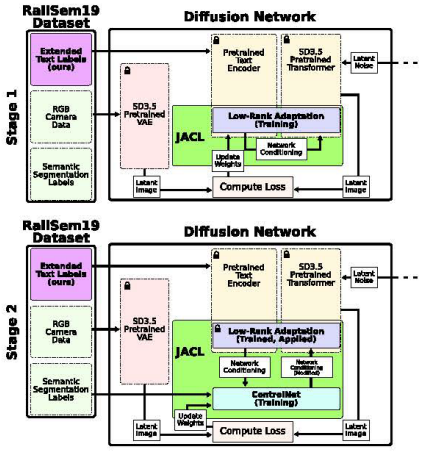


Fig. 2. JACL training process. In stage 1, the LoRA network is trained using the text labels and RGB camera data. In stage 2, the ControlNet network is trained with the LoRA applied to the underlying layers.

a limited dataset by leveraging the highly compressed data that is embedded into the parameters of an image synthesis network, by using our limited dataset to bias the pretrained network towards our target domain.

A. JACL Architecture

The JACL architecture is based on a pretrained Stable Diffusion 3.5 architecture, which contains three text encoders, a transformer that operates as the diffusion backbone, and a pretrained variational autoencoder. This architecture is fairly lightweight, and can generate 1-megapixel imagery using a single GPU and 20GB of VRAM; with optimizations, this architecture could likely operate in inference mode on a consumer-grade GPU. JACL itself consists of two subcomponents: a LoRA that helps bias the dataset towards the target domain (in our case, railroad data when using RailSem19 data), while the ControlNet component biases individual pixels towards specific semantic classifications.

B. JACL Training and Inference

Ordinarily, training an architecture that contains a LoRA and a ControlNet can be performed separately, and then combined later during inference. However, given the limited amount of data available, we were interested in maximizing the bias of our resultant architecture towards rail domain data. To ensure that the LoRA and ControlNet components were closely coupled during the training process, we investigated means by which their training could be more closely coupled. Instead of performing true joint training, we found an alternate approach that theoretically can accomplish the same overall composition without modifying the underlying training infrastructure. First, we trained the LoRA network on its own, using our added labels as the input text for training. After separate evaluation

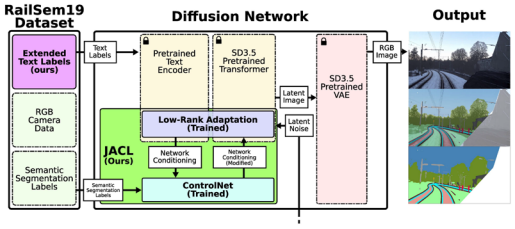


Fig. 3. JACL Architecture. Existing data and pretrained algorithms appear with a dashed outline. The JACL-specific networks and augmented dataset components appear with a solid outline and bold text.

of the LoRA, we then trained the ControlNet component, with the LoRA network active and applied to the baseline network weights. This technique effectively allows the ControlNet weights to inherit the baseline fine-tuning from LoRA, and allows us to split the training process of JACL into two stages, which minimizes the amount of parameters that are simultaneously trained, requires fewer system resources, and allows evaluation of each training stage independently. Once JACL is trained, generating synthetic data only requires two inputs: the desired semantic segmentation mask, and a text description of the scene.

IV. Experiments

Our experiments on JACL are split into two key phases: the training phase, and the data evaluation phase. The training phase, as described in Section III, adapts a pretrained Stable Diffusion model to output the desired imagery with strong adherence to input segmentation masks, using a two-stage training process. We also perform limited subjective evaluation to determine whether the trained JACL architecture is performing well. The second phase analyzes the resultant imagery to determine whether the output synthetic imagery is suitable for training downstream tasks such as semantic segmentation algorithms. To determine the suitability of the data, we evaluate the objective structural image similarity between real-synthetic data pairs, as well as the difference in VAE activation between the ground truth and

A. Dataset Processing and Preparation

In this work, we used the RailSem19 dataset for our experiments. This dataset contains 8500 samples of semantically labeled data taken from cameras mounted to the front of inter-urban and intra-urban rail vehicles; we use the 3000 sample training subset and 500 testing subset that was used in the authors' analysis, further limiting the number of training samples available. To maximize the pixel-level difference in the labels, we converted the semantic label masks to RGB colors based on RailSem19's suggested label color palette. Then, we created captions for all the images in the dataset; while neither LoRA nor ControlNet explicitly require text labels, we determined that relying on classifier-free guidance would not produce as reliable. We chose to generate text labels formatted as comma separated lists of key properties, generated deterministically using the contents of the segmentation masks, as human annotation would be extremely labor intensive, and machine learning-based labeling may introduce unwanted errors. We also generated perturbed separate ground truth segmentation masks when generating synthetic samples to reduce the likelihood that the network produce exact copies of the input data. These perturbations include random scaling, offset, and x-axis mirroring.

B. Training and Inference Processes

Our training process, as previously described, occurs in two stages. The first stage, the LoRA stage, trains a LoRA network on 3000 samples of RailSem19 camera data, with our augmented text labels used for a majority of the samples. We trained blocks 12 through 24 of LoRA to reduce the effect of the LoRA on fine detail generation, as we want the pretrained network to have more influence on smaller details. We limited the training to 3 epochs and a learning rate of 1e-4 with a restarting cosine scheduler; in our experiments, we found that training beyond 3 epochs did not significantly change the bias of the output imagery towards RailSem19-like images. After training the LoRA, we then trained the ControlNet, using a learning rate of 1e-5 with 128-step gradient accumulation with 500 total training steps overall, or approximately 21.33 epochs. While Zhang claim ControlNet is resistant to overfitting [14], we found overfitting artifacts visible before 100 training epochs. To avoid overfitting, we limited ControlNet training to 20 epochs.

For sample generation, we used the previously generated perturbed segmentation masks and their associated text labels as inputs to the JACL network, using a 35-step denoising sampler with uniform Stochastic Galerkin Method scheduling. We also generated a set using non-perturbed segmentation masks using the same sampling parameters to perform a study on the resultant image similarity and statistical divergence.

C. Results

We performed a brief study on the effects of selectively activating the LoRA and ControlNet networks individually; one sample of which appears in Figure 4. While the ControlNet clearly affects JACL's ability to conform to the input masks, the LoRA does have more subtle effects; with only the LoRA active, the synthetic imagery contains geometry that does not make physical sense, such as poles appearing in the middle of rail tracks. In addition, human subjects tend to appear further away, which conforms to the imagery biases present in the underlying RailSem19 dataset.

While the JACL architecture can generate images that appear to be structurally similar to the input ground truth images, we must also ensure there are statistically significant differences between real and generated samples. To measure the difference in real and synthetic samples, we used JACL to generate a set of 10 samples based on a specific ground truth image, sample `rs03012`, which was taken in winter conditions. We compare these images to the baseline ground

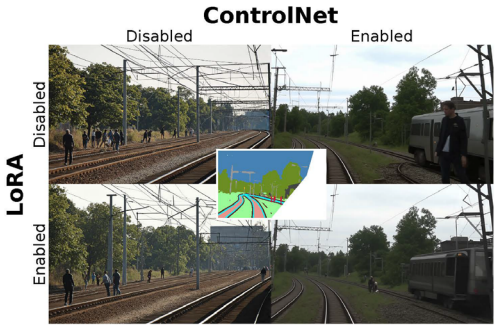


Fig. 4. Comparison of network outputs when LoRA and ControlNet components are selectively enabled. Subtle details introduced with LoRA include poles not appearing in the middle of rails. All samples are generated using the same inputs, with the input segmentation mask appearing inset in the middle.

truth RGB image using three image similarity metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity (SSIM), and the Jensen-Shannon Divergence (JSD) of the image luminance histogram, and the median difference in element-wise difference in VAE latent space, Δ_{VAE}, taking the average of 10 synthetic samples. We compare the metrics of these baseline samples (i.e., using the ground truth segmentation and text labels) with both ground truth samples that have been distorted with Gaussian noise, pixel perturbation, and gamma correction. We also generated additional samples with additional text keywords corresponding to winter, fall, and spring conditions.

As shown in Table I, we found that the synthetic samples from JACL exhibited surprisingly high PSNR. As the underlying ground truth was taken in snowy winter conditions, we are unsurprised that PSNR for this subset was the highest among the generated samples. Structural similarity was higher for all generative samples compared with the Gaussian noised ground truth, but still noticeably lower than the gamma and distorted images. Average luminance histogram JSD was higher in all generative samples compared with the distorted and noised image, but was lower than the gamma-adjusted ground truth sample. We observed that the median difference in latent activation was higher for generative samples compared to the distorted sample, and comparable to both the noise- and gamma-adjusted ground truth images. Overall, these results demonstrate that the generative samples produced by JACL exhibit an objectively high divergence from comparable ground truth images, suggesting these samples should work well in augmenting datasets for downstream perception tasks.

Finally, we compare the semantic segmentation performance of a pretrained SegFormer [20] network that is fine-tuned on real data versus purely synthetic data. For both tests, we start with a SegFormer model that was pretrained on the CityScapes dataset [7], both of which are trained over 10 epochs, one using 3000 samples of real data and the other one using the same number of generative samples constructed by JACL, with both networks tested against 500 samples of real RailSem19 data. As shown in Table II, many object classes had minimal performance differences between the synthetic and real data. For the construction class, the synthetic dataset actually performed *better* than the real data, possibly owing to the high representation of buildings in the pretrained diffusion network. While many classes exhibited more significant performance differences, rail tracks and rail vehicles exhibited a severe performance drop on the synthetic dataset, despite the rail imagery exhibiting high subjective realism. The performance gaps in these classes contribute to the overall mean accuracy difference between the networks trained on real versus synthetic data.

V. Conclusion

In this paper, we established a methodology for generating realistic synthetic rail domain data from a limited set of ground truth data by leveraging diffusion-based networks with domain- and context conditioning control though Joint Augmented ControlNet-LoRA (JACL). We demonstrate this architecture can produce convincing imagery that closely conforms to an input semantic segmentation map, which can be used to use existing segmentation labels to generate new imagery, or to generate new imagery from scratch using existing or machine augmented segmentation masks. We also demonstrate that the generated imagery is structurally similar to the ground truth, which indicates close conformance with the input segmentation mask, while also exhibiting statistical divergence from the ground truth image, which should help downstream algorithms improve performance.

A. Limitations and Caveats

While our analysis demonstrates that synthetic samples could be used to successfully train a segmentation network, there are some limitations that should be disclosed. First, for our experiments, we have the full benefit of a sizable dataset with known ground truth; evaluating this technique to domains with true data scarcity will present a significant challenge. We advocate for those interested in using our proposed technique to thoroughly test on real data before wide deployment. As more data are collected, they can be used to iteratively evaluate the accuracy of the system, which can then be used to retrain the ControlNet and LoRA, generate new data, and then finally to train a new set of parameters for the target semantic segmentation network.

Another limitation with our approach is the limited variation in underlying ground truth segmentation masks used to condition the JACL framework. However, we hypothesize that the output of a partially trained image segmentation network, even if there are significant performance gaps, could then be used to create new masks to generate additional imagery. This approach would dovetail with our previously advocated testing approach. Another approach would be to develop or adapt an existing simulation framework to output appropriate masks for JACL; simulators such as the CARLA simulator can already

TABLE I
COMPARISON OF WHITENED GROUND TRUTH IMAGES WITH GENERATIVE SAMPLES. HIGHER JSD AND Δ_{VAE} ARE DESIREABLE.

	GT + Distortion	GT + Gamma	GT + Noise	Gen, Base (10-mean)	Gen, Winter (10-mean)	Gen, Fall (10-mean)	Gen, Spring (10-mean)
PSNR	45.837	38.052	33.867	33.434	36.353	33.815	32.907
SSIM	0.984	0.963	0.716	0.778	0.885	0.798	0.741
JSD, Lum	0.002	0.362	0.226	0.383	0.243	0.234	0.333
Δ_{VAE}	0.118	0.321	0.389	0.383	0.308	0.368	0.375

TABLE II
COMPARISON OF SEMANTIC SEGMENTATION ACCURACY (%) BETWEEN REAL AND SYNTHETIC RAILSEM19 DATASETS

	Mean	car	construction	human	road	sidewalk	sky	on-rails	rail-track
Real	**65.36%**	**83.76%**	82.54%	**75.11%**	58.85%	**59.94%**	**97.75%**	**69.74%**	**80.68%**
Synthetic	43.41%	81.82%	**85.76%**	58.26%	**69.94%**	50.26%	95.64%	47.96%	41.79%

output segmentation masks which, while intended for training algorithms with corresponding simulated RGB images, could be used by JACL for novel segmentation masks.

B. Future Work

Our experiments in Section IV demonstrated the potential for our system to not only produce rail-centric synthetic data, but to also have the capability to generate more specific data based on minimal modification to the text conditioning in statistically significant manners. This capability suggests that our system, as well as others that are based on our proposed framework, can be used to synthesize data across a wider domain. We would like to focus future work on characterizing the extent of the domain shifts that can be accomplished with our framework, and to what extent the training of the pretrained network affects this capability. We also would like to investigate the discrepancy between the subjective quality of the rails within the synthetic images, versus the significant performance gap in semantic segmentation compared with real data.

REFERENCES

[1] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*. Springer, 2014, pp. 740–755.

[2] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov *et al.*, "The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale," *International journal of computer vision*, vol. 128, no. 7, pp. 1956–1981, 2020.

[3] R. Benenson, S. Popov, and V. Ferrari, "Large-scale interactive object segmentation with human annotators," in *CVPR*, 2019.

[4] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4015–4026.

[5] Y. Liao, J. Xie, and A. Geiger, "Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3292–3310, 2022.

[6] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.

[7] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[8] O. Zendel, M. Murschitz, M. Zeilinger, D. Steininger, S. Abbasi, and C. Beleznai, "Railsem19: A dataset for semantic rail scene understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[9] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[10] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681–4690.

[11] C. Zhang, C. Zhang, M. Zhang, and I. S. Kweon, "Text-to-image diffusion models in generative ai: A survey," *arXiv preprint arXiv:2303.07909*, 2023.

[12] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[13] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[14] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3836–3847.

[15] H. K. J. W. Z. W. X. X. C. C. Ming Li, Taojiannan Yang, "Controlnet++: Improving conditional controls with efficient consistency feedback," in *European Conference on Computer Vision*, 2024.

[16] J. Tian, L. Aggarwal, A. Colaco, Z. Kira, and M. Gonzalez-Franco, "Diffuse attend and segment: Unsupervised zero-shot segmentation using stable diffusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 3554–3563.

[17] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International conference on machine learning*. pmlr, 2015, pp. 2256–2265.

[18] Z. K. Hartley, R. J. Lind, M. P. Pound, and A. P. French, "Domain targeted synthetic plant style transfer using stable diffusion lora and controlnet," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 5375–5383.

[19] B. Deng and Y. Lu, "Weed image augmentation by controlnet-added stable diffusion," in *Synthetic Data for Artificial Intelligence and Machine Learning: Tools, Techniques, and Applications II*, vol. 13035. SPIE, 2024, pp. 163–175.

[20] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in neural information processing systems*, vol. 34, pp. 12 077–12 090, 2021.

MTI FOUNDER

Hon. Norman Y. Mineta

MTI BOARD OF TRUSTEES

Directors

MINETA TRANSPORTATION INSTITUTE