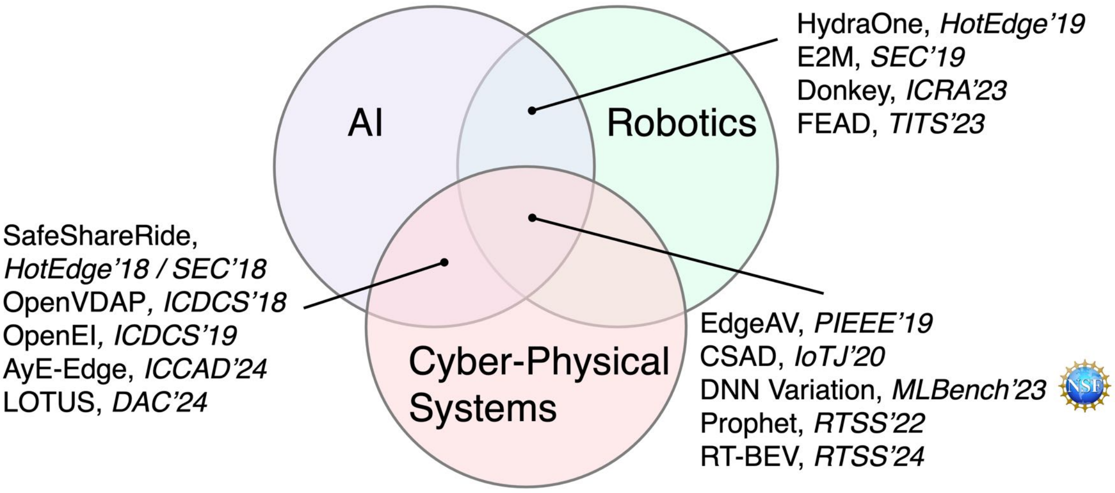
Research Statement

Research Objective

My research vision is to develop **real-time cyber-physical systems** that enhance **autonomy**, **efficiency**, and **safety** in transportation and robotics. By integrating cyber-physical systems, real-time processing, and edge computing, I aim to create robust and scalable solutions to address the challenges of autonomous driving and intelligent transportation. My goal is to bridge the gap between theory and practice, ensuring that advanced transportation systems are both innovative and reliable in real-world applications.



Research Highlights

My research interests lie in computing systems for autonomous driving, cyber-physical systems, robotics, and edge computing. By delivering robust AI system solutions with significant societal impact, I aim to enhance the **safety**, **predictability**, and **efficiency** of **real-time cyber-physical systems**, such as autonomous vehicles, trucks, and mobile robots. My work, published in top-tier venues like RTSS, RTAS, ICCAD, ICRA, SEC, HotEdge, TITS, IWC, and IoTJ, addresses these objectives in the autonomous CPS sphere.

Notably, I am among the first in the computing systems community to explore the issue of DNN inference time variations in autonomous driving, leading to a **$600,000** **NSF Grant** awarded to me as a **Co-Principal Investigator** for research on multi-tenant DNN inference time variations. I have developed testbeds like HydraOne and Donkey, which serve as programmable and energy-efficient platforms for autonomous mobile robots. Additionally, my research on fuel-efficient autonomous trucking has resulted in a 7% reduction in fuel consumption in practical deployments compared to experienced drivers. As of December 9, 2024, my work has received **1,960** citations, with an H-Index of **17** and an i10-Index of **20**.

Prior Research

Here I will introduce my prior research with three projects: predictable perception systems for autonomous vehicles (AVs), energy-efficient autonomous systems for robots/trucks, and vehicular edge computing for public safety.

1. Predictable Perception for AVs

Autonomous vehicles (AVs) depend on sensors and deep neural networks (DNNs) to perceive their environment and make real-time maneuver decisions. However, ensuring predictability in the perception pipeline presents significant challenges [1]. To address this issue, we conducted a comprehensive profiling of DNN inference time variations and developed novel solutions to achieve predictable perception for AVs. First, *Prophet* mitigates substantial time variations in most DNN models within an AV system through novel early-exit and coordination mechanisms. Second, *PP-DNN* achieves a real-time and predictable pipeline by employing region-of-interest (ROI) processing. Third, *RT-BEV* addresses multi-camera synchronization and detection issues through a co-design of communication and detection, coupled with ROI-based processing. Finally, *AyE-Edge* presents a general solution for power-efficient real-time object detection via the co-design of keyframe selection, model pruning, and CPU-GPU configuration.

DNNs Inference Time Variations Profiling [2, 3]: Understanding the time variations of the DNN inference becomes a fundamental challenge in real-time scheduling for autonomous driving. Therefore, we develop a reconfigurable testbed for DNN inference profiling [3]. On top of that, we **analyze the time variation in DNN inference in fine granularity from six perspectives: data, I/O, model, runtime, hardware, and end-to-end perception system**. Six insights are derived in understanding the time variations for DNN inference [2]. This work contributes to two research papers and one **NSF Grant** for studying *DNN inference time variations for multi-tenant DNNs*.

Prophet [4]: Through a comprehensive empirical study, we found that the inference time variations for a single DNN model are **mainly caused by the DNN's multi-stage/multi-branch structure**, which has a dynamic number of proposals or raw points. In addition, we found that the **uncoordinated contention and cooperation** are the roots of the time variations for multi-tenant DNNs inference. Based on these insights, we proposed the `Prophet` system that addresses the time variations in the AV perception system in two steps [4]. The first step is to predict the time variations based on the intermediate results like proposals and raw points. The second step is coordinating the multi-tenant DNNs to ensure the execution progress is close to each other. The evaluation on the KITTI dataset show effective performance in reducing inference time variations. This paper is published at **IEEE RTSS 2022**.

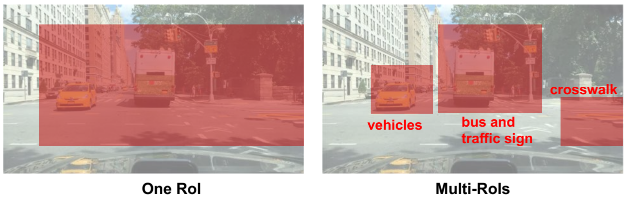


Figure 1: An example of environment-aware dynamic ROIs.

PP-DNN [5]: Existing studies primarily focus on optimizing the DNN inference time to achieve faster perception, for example, by compressing the DNN model with pruning and quantization. In contrast, we present a Predictable Perception system with DNNs (`PP-DNN`) that reduce the amount of image data to be processed while maintaining the same level of accuracy for multi-tenant DNNs by **dynamically selecting 'critical' frames and regions of interest (ROIs)** [5]. `PP-DNN` is based on our key insight that **critical frames and ROIs for AVs vary with the AV's surrounding environment**, as shown in Figure 1. However, it is challenging to identify and use critical frames and ROIs in multi-tenant DNNs for predictable inference. Given image-frame streams, `PP-DNN` leverages an ROI generator to identify critical frames and ROIs based on the similarities of consecutive frames and traffic scenarios. `PP-DNN` then leverages a FLOPs predictor to predict multiply-accumulate operations (MACs) from the dynamic critical frames and ROIs. The ROI scheduler coordinates the processing of critical frames and ROIs with multiple DNN models. Finally, we design a detection predictor for the perception of non-critical frames. We have implemented `PP-DNN` in an ROS-based AV pipeline and evaluated it with the BDD100K and the nuScenes dataset. `PP-DNN` is observed to significantly enhance perception predictability. This paper is currently under review by **RTAS 2025**.

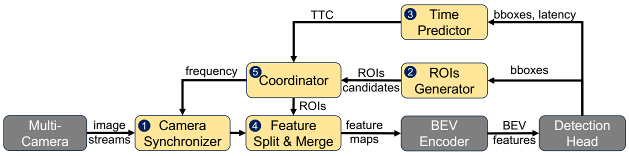


Figure 2: An overview of `RT-BEV` system.

RT-BEV [6]: For vision-centric Bird's Eye View (BEV) perception, prior work either compresses the dense detection model to reduce computation which can hurt accuracy and assume images are well synchronized, or focuses on worst-case communication delay. To meet this challenge, we propose `RT-BEV`, the first framework designed to **co-optimize message communication and object detection** to improve real-time e2e BEV perception without sacrificing accuracy [6]. Figure 2 shows an overview of `RT-BEV`. The main insight of `RT-BEV` lies in **generating traffic environment- and context-aware Regions of Interest (ROIs) for AV safety, combined with ROI-aware message communication**. `RT-BEV` features an ROI-aware Camera Synchronizer that adaptively determines message groups and allowable delays based on ROIs' coverage. We also develop a ROIs Generator to model context-aware ROIs and a Feature Split & Merge component to handle variable-sized ROIs effectively. Furthermore, a Time Predictor forecasts timelines for processing ROIs, and a Coordinator jointly optimizes latency and accuracy for the entire e2e pipeline. We have implemented `RT-BEV` in a ROS-based BEV perception pipeline and evaluated it with the nuScenes dataset. `RT-BEV` is shown to significantly enhances real-time BEV perception. This paper is published at **IEEE RTSS 2024**.

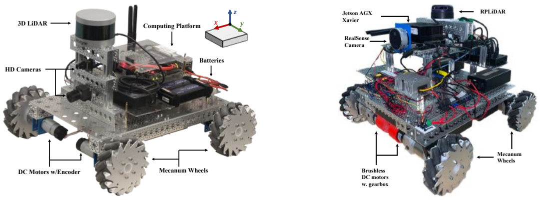
AyE-Edge [7]: Power-Efficient Real-Time Object Detection. Object detection on the edge is in growing demand thanks to its ever-broad application prospects. However, the development of this field is rigorously restricted by the deployment dilemma of simultaneously achieving high accuracy, excellent power efficiency, and meeting strict real-time requirements. To tackle this dilemma, we propose AyE-Edge, the first-of-this-kind development tool that explores automated algorithm-device deployment space search to realize accurate yet power-efficient real-time object detection on the edge [7]. Through a **collaborative exploration of keyframe selection, CPU-GPU configuration, and DNN pruning strategy**, AyE-Edge excels in extensive real-world experiments conducted on a mobile device. The results consistently demonstrate AyE-Edge's effectiveness, realizing outstanding real-time performance, detection accuracy, and notably, a remarkable 96.7% reduction in power consumption, compared to state-of-the-art (SOTA) competitors. This paper is published at **ICCAD 2024**.

2. Energy-Efficient Autonomous System

Energy-efficiency has been a fundemental issue for mobile robots and heavy-duty trucks. Autonomous driving techniques bring potential for saving energy through proper coordination of software stack (perception, planning, and control) and hardware platforms (sensor, computing device, motors, engine management system). Figure 3 shows platforms we built for energy efficinet autonomous system. More specifically, on top of the HydraOne platform [8], E2M coordinates the perception module with computing devices and sensors [9]; Donkey platform extends this to the whole navigation stack with sensor, computing device, and motors [10]; pNav achieves real-time power-efficient navigation for AMRs [11]; FEAD coordinates planning and control with the truck's engine management systems [12].

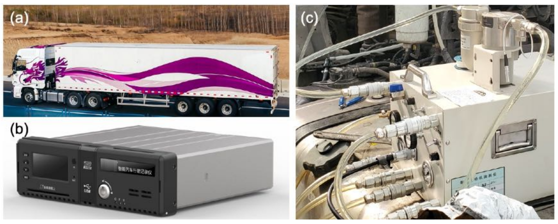
E2M [9]: By analyzing the breakdown of power dissipation for the execution of computer-vision applications on AMRs and **discover three main root causes of energy inefficiency**: *uncoordinated access to sensor data*, *performance-oriented model inference execution*, and *uncoordinated execution of concurrent jobs*. To address three inefficiencies, we propose E2M, an energy-efficient middleware software stack for autonomous mobile robots [9]. First, E2M regulates the access of different processes to sensor data, e.g., camera frames, so that the amount of data actually captured by concurrently executing jobs can be minimized. Second, based on a predefined per-process performance metric (e.g., safety, accuracy) and desired target, E2M manipulates the process execution period to find the best energy-performance trade off. Third, E2M coordinates the execution of the concurrent processes to maximize the total contiguous sleep time of the computing hardware for maximized energy savings. We have implemented a prototype of E2M on HydraOne [8], a real-world AMR. Our experimental results show that, compared to several baselines, E2M leads to 24% energy savings for the computing platform, which translates into an extra 11.5% of battery time and 14 extra minutes of robot runtime, with a performance degradation lower than 7.9% for safety and 1.84% for accuracy. This paper is published at **ACM/IEEE SEC 2019**.

Energy-Efficient Path Planning for AMRs [10]. Through empirical studies on real AMRs, we have identified a lack of coordination between computation and control as a major source of energy inefficiency. In this work, we propose **a comprehensive energy prediction model that provides real-time energy consumption** for each component of the AMR [10]. Additionally, we propose three path models to address the obstacle avoidance problem for AMRs. To evaluate the performance of our energy prediction and path models, we have developed a customized AMR called Donkey, which has the capability for fine-grained (millisecond-level) end-to-end power profiling. Our energy prediction model demonstrated an accuracy of over 90% in our evaluations. Finally, we applied our energy prediction model to obstacle avoidance and guided energy-efficient path selection, resulting in up to a 44.8% reduction in energy consumption compared to the baseline. This paper is published at **ICRA 2023**.



HydraOne Platform: E2M

Donkey Platform: energy-efficient path planning, pNav



Inceptio Autonomous Trucking: FEAD

Figure 3: Platforms for Energy Efficient Autonomous System.

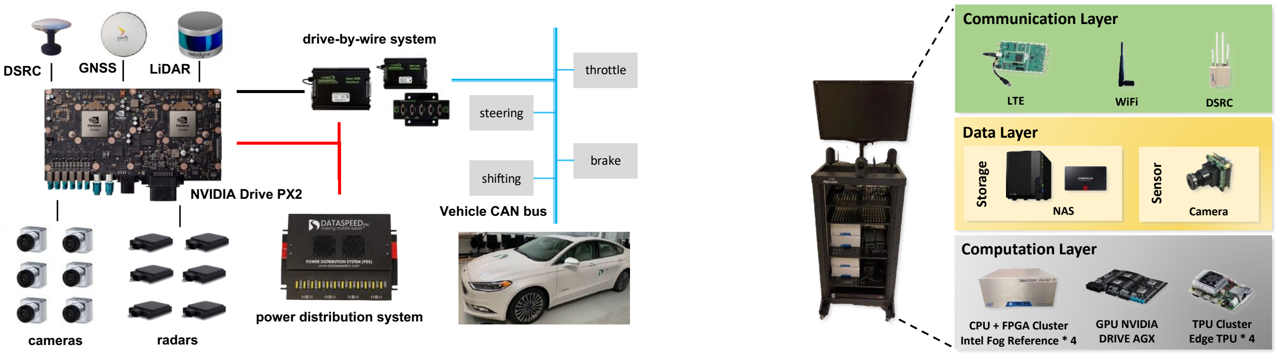
pNav: Power-Efficient Navigation of AMRs [11]: `pNav` is a novel power-management system designed to enhance the power/energy-efficiency of Autonomous Mobile Robots (AMRs) by **jointly optimizing their physical/mechanical and cyber subsystems** [11]. From power profiling on robots, we identified three types of inefficiency in current AMRs designs that combine cyber (C) and physical (P) subsystems: (1) unawareness of power-breakdown variations, (2) uncoordinated C and P subsystems, and (3) neglect of navigation localities. To achieve proper coordination of C and P subsystems for power-efficiency, `pNav` employs a multi-faceted approach. First, `pNav` integrates millisecond-level power consumption prediction for both C and P subsystems. Second, `pNav` includes novel real-time modeling and monitoring of spatial and temporal navigation localities for AMRs. Third, `pNav` supports dynamic coordination of AMR software (navigation, detection) and hardware (motors, DVFS driver) configurations. `pNav` is prototyped using the Robot Operating System (ROS) Navigation Stack, 2D LiDAR, and camera. Our in-depth evaluation with Donkey and Gazebo environments demonstrates a >96% accuracy in predicting power consumption and a 38.1% reduction in power consumption without compromising navigation accuracy and safety. This paper is currently under review by **EuroSys 2025**.

FEAD [12]: Fuel Efficient Autonomous Driving. Fuel cost contributes significantly to the high operation cost of heavy-duty trucks. Developing fuel rate prediction models is the cornerstone of fuel consumption optimization approaches for heavy-duty trucks. However, limited by accurate features directly related to the truck's fuel consumption, state-of-the-art models show poor performance and are rarely deployed in practice. In this work, we **use the truck's engine management system (EMS) and Instant Fuel Meter (IFM)** to collect a three-month dataset during the period of December 2019 to June 2020 [12]. Seven prediction models, including linear regression, polynomial regression, MLP, CNN, LSTM, CNN-LSTM, and AutoML, are investigated and evaluated to predict real-time fuel rate. The evaluation results show that the EMS and IFM dataset help to improve the coefficient of determination of traditional linear/polynomial models from 0.87 to 0.96, while learning-based approach AutoML improves the coefficient of determination to attain 0.99. Besides, the actual deployment of fuel rate prediction and path planning show up to 7% fuel saving on real Inceptio trucks [13]. This paper is published at **IEEE TITS 2023**.

3. Vehicular Edge Computing for Public Safety

Connected and Autonomous Vehicles (CAVs) is becoming more and more popular for automobile academic and industry community. I have worked on prototyping and profiling real autonomous driving vehicles as well as road-side unit, including the Hydra AV Testbed, and the Equinox platform. On top of Equinox and Hydra, we profile the communciation performance for V2X communciations. Futhermore, I proposed SafeShareRide, which is an edge computing-enabled system for protecting both driver and passenger's safety.

Hydra AV Testbed [14]. At Wayne State University, we **built a level-4 autonomous driving vehicles called *Hydra***. Figure 4a shows an overview for it [14]. An NVIDIA Drive PX2 is used as the vehicle computation unit (VCU). Multiple sensors, including six cameras, six radars, one LiDAR, one GNSS antenna, and one DSRC antenna, are installed for sensing and connected with VCU. The CAN bus is used to transmit messages between different ECUs controlling steering, throttle, shifting, brake, etc. Between the NVIDIA Drive PX2 and the vehicle's CAN bus, a drive-by-wire system is deployed as an actuator of the vehicle control commands from the computing system. Additionally, a power distribution system is used to provide extra power for the computing system.



(a) Hydra Autonomous Driving Testbed.

(b) Equinox Platform.

Figure 4: Testbed for level 4 autonomous driving and road-side unit.

Vehicular Communication Comparison [15]. Communication mechanisms play an important role in CAVs applications and services. However, lack of detailed comparison of different communication mechanisms is the main obstacle for the deployment of CAVs applications and services. In this work, we **built an end-to-end prototype by integration of Equinox** [16] (Figure 4b) and HydraOne [8] which supports WiFi, LTE, and DSRC based communications and evaluate the performance in latency, power dissipation, and system utilization [15]. This paper is published at **USENIX HotEdge 2020**.

SafeShareRide [17]. Ridesharing services, such as Uber and Didi, are enjoying great popularity; however, a big challenge remains in guaranteeing the safety of passenger and driver. We propose an edge-based attack detection in ridesharing services, namely `SafeShareRide`, which can **detect dangerous events happening in the vehicle in near real-time** [18, 17]. `SafeShareRide` is implemented on both drivers' and passengers' smartphones.

`SafeShareRide` consists of three stages: speech recognition, driving behavior detection, and video capture and analysis. Abnormal events detected during the stages of speech recognition or driving behavior detection will trigger the video capture and analysis in the third stage. The video data processing is also redesigned: video compression is conducted at the edge to save upload bandwidth while video analysis is conducted in the cloud. We implement the `SafeShareRide` system by leveraging open source algorithms. Our experiments include a performance comparison between `SafeShareRide` and other edge-based and cloud-based approaches, CPU usage and memory usage of each detection stage, and a performance comparison between stationary and moving scenarios. Finally, we summarize several insights into smartphone based edge computing systems. This work leads to paper publications at **USENIX HotEdge 2018** and **ACM/IEEE SEC 2018**.

FUTURE RESEARCH DIRECTIONS

In the future, my research will concentrate on addressing critical challenges related to safety and efficiency in real-time cyber-physical systems and intelligent transportation systems. By advancing methodologies in these areas, I aim to enhance the robustness, reliability, and performance of such systems in complex and dynamic environments.

End-to-End Testing of CAVs. I plan to develop comprehensive end-to-end testing methodologies for connected and autonomous vehicles (CAVs) that cover every aspect of their operation—from perception to control. This holistic approach is essential for ensuring robustness and safety across a wide range of real-world scenarios. A key component of this work will be the creation of universal metrics applicable to various security challenges, allowing for accurate assessment of different strategies' strengths and weaknesses. These metrics will evaluate multiple facets of the AV pipeline, enabling identification of where attacks and defenses are most effective and quantifying their impact on functional safety. By providing deeper insights into potential threats and vulnerabilities, this research will contribute to the development of more secure and reliable autonomous vehicle systems.

Predictable Offroad Autonomous Driving. I also aim to tackle the unique challenges presented by offroad autonomous driving, where unpredictable conditions—such as fast-moving objects, varying terrains, and sudden changes in elevation—impose stringent and dynamic timing and accuracy requirements on machine learning-based perception pipelines. To overcome these obstacles, I envision a collaborative approach that employs a complex sensor suite, including drones and advanced sensing technologies, to enhance environmental perception. By integrating diverse sensors and developing robust systems capable of real-time processing under dynamic conditions, my goal is to design autonomous vehicles that can reliably and safely navigate offroad environments. This research will advance autonomous driving technology and enable practical applications in areas like agriculture, disaster response, and exploration.

Efficient Foundation Models in Cyber-Physical Systems. Foundational models in vision, language, and audio are increasingly being deployed in cyber-physical systems (CPS) for applications such as autonomous driving, smart healthcare, and precision agriculture. These models enable CPS to perceive and interpret complex environments, leading to enhanced functionality and smarter decision-making processes. However, integrating these computationally intensive models into resource-constrained devices presents significant challenges, particularly regarding energy efficiency and real-time performance. I believe that a co-design approach involving both software and hardware optimizations is essential to overcome these challenges. By jointly optimizing algorithms, system architectures, and hardware implementations, we can achieve energy-efficient model inference and training. This approach will facilitate the practical deployment of sophisticated models in CPS applications while maintaining high performance and low energy consumption.

REFERENCES

[1] Liangkai Liu, Sidi Lu, Ren Zhong, Baofu Wu, Yongtao Yao, Qingyang Zhang, and Weisong Shi. Computing systems for autonomous driving: State of the art and challenges. IEEE Internet of Things Journal (IoTJ), 8(8):6469–6486, 2020.

[2] Liangkai Liu, Yanzhi Wang, and Weisong Shi. Understanding time variations of DNN inference in autonomous driving. arXiv preprint arXiv:2209.05487, 2022.

[3] Liangkai Liu, Yanzhi Wang, and Weisong Shi. CPT: A configurable predictability testbed for DNN inference in AVs. Tsinghua Science and Technology, 30(1):87–99, 2024.

[4] Liangkai Liu, Zheng Dong, Yanzhi Wang, and Weisong Shi. Prophet: Realizing a predictable real-time perception pipeline for autonomous vehicles. In IEEE Real-Time Systems Symposium (RTSS), pages 305–317. IEEE, 2022.

[5] Liangkai Liu, Kang G. Shin, Chengmo Yang, and Weisong Shi. Predictable DNN inference for autonomous vehicles. In under review by RTAS 2025.

[6] Liangkai Liu, Jinkyu Lee, and Kang G. Shin. RT-BEV: Enhancing real-time BEV perception for autonomous vehicles. In 45th IEEE Real-Time Systems Symposium (RTSS), 2024.

[7] Liangkai Liu*, Chao Wu*, Yifan Gong*, Mengquan Li, Yushu Wu, Xuan Shen, Zhimin Li, Geng Yuan, Weisong Shi, and Yanzhi Wang. AyE-Edge: Automated deployment space search empowering accuracy yet efficient real-time object detection on the edge. In International Conference on Computer-Aided Design (ICCAD), 2024.

[8] Yifan Wang, Liangkai Liu, Xingzhou Zhang, and Weisong Shi. HydraOne: An indoor experimental research and education platform for CAVs. In 2nd USENIX Workshop on Hot Topics in Edge Computing (HotEdge 19), 2019.

[9] Liangkai Liu, Jiamin Chen, Marco Brocanelli, and Weisong Shi. E2M: An energy-efficient middleware for computer vision applications on autonomous mobile robots. In Proceedings of the 4th ACM/IEEE Symposium on Edge Computing (SEC), pages 59–73, 2019.

[10] Liangkai Liu, Ren Zhong, Aaron Willcock, Nathan Fisher, and Weisong Shi. An open approach to energy-efficient autonomous mobile robots. In IEEE International Conference on Robotics and Automation (ICRA), 2023.

[11] Liangkai Liu, Weisong Shi, and Kang G. Shin. Power-efficient navigation of autonomous mobile robots. In under review by EuroSys 2025.

[12] Liangkai Liu, Wei Li, Dawei Wang, Yi Wu, Ruigang Yang, and Weisong Shi. Fuel rate prediction for heavy-duty trucks. IEEE Transactions on Intelligent Transportation Systems (TITS), 24(8):8222–8235, 2023.

[13] Inceptio Completes Delivery of 300 Autonomous Heavy-Duty Trucks to YTO Express. `https://en.inceptio.ai/news/99012.html`.

[14] Hydra In Action. `https://www.youtube.com/watch?v=QJ54VSeqJuY&t=6s`.

[15] Liangkai Liu, Baofu Wu, and Weisong Shi. A comparison of communication mechanisms in vehicular edge computing. In 3rd USENIX Workshop on Hot Topics in Edge Computing (HotEdge 20), 2020.

[16] Liangkai Liu, Yongtao Yao, Ruijun Wang, Baofu Wu, and Weisong Shi. Equinox: A road-side edge computing experimental platform for CAVs. In International Conference on Connected and Autonomous Driving (MetroCAD), pages 41–42. IEEE, 2020.

[17] Liangkai Liu, Xingzhou Zhang, Mu Qiao, and Weisong Shi. Safeshareride: Edge-based attack detection in ridesharing services. In IEEE/ACM Symposium on Edge Computing (SEC), pages 17–29. IEEE, 2018.

[18] Liangkai Liu, Xingzhou Zhang, Mu Qiao, and Weisong Shi. SafeShareRide: Edge-based attack detection in ridesharing services. In USENIX Workshop on Hot Topics in Edge Computing (HotEdge 18), Boston, MA, 2018. USENIX Association.