2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Towards Kbps-level Vehicle Teleoperation via Persistent-Transient Environment Modelling

Chunyang Zhao, Zeyu Zhou, Haoran Liu, Dogan Kircali, Guoyi Chi,
Hongming Shen, Yuanzhe Wang, Danwei Wang *Fellow, IEEE*

***Abstract*— Traditional teleoperation technologies based on video streaming are facing several challenges in practical applications, including limited bandwidth, constrained spatial awareness, and sensitivity to illumination. Existing studies have not adequately addressed these issues. This paper presents a novel non-video based teleoperation framework for autonomous vehicles operating in bandwidth-limited environments. To reduce the amount of data being transmitted, a persistent-transient environment model is proposed for telepresence. Initially, a digital twin of the environment is preconstructed, containing only persistent environmental information. Subsequently, transient information captured by onboard sensors, such as vehicle state and dynamic objects, necessitate real-time transmission. Based on this model, a 3D virtual scene is rendered in front of the teleoperator, offering any desired virtual viewpoint to enhance spatial awareness. This telepresence model only requires real-time transmission of minimal data, i.e., vehicle state and detected objects, and remains unaffected by illumination conditions, enabling teleoperation even in applications with Kbps-level bandwidth constraints. Experimental results showcase the substantial potential of the proposed framework in bandwidth-limited settings.**

I. Introduction

In recent years, tremendous efforts have been put into developing autonomous vehicles. Yet, full autonomy still remains a goal to be achieved, and additional time is expected for testing and approval of regulation [1]. Existing autonomous technologies cannot handle all situations in the real world and are always challenged by various corner cases [2], thus requiring human intervention. Nevertheless, deploying a human operator on site or keeping a safety driver onboard is not economically efficient. In this case, teleoperation could be a more cost-efficient solution and could expand the operational domain of autonomous vehicles [3]. Teleoperation has already been deployed in a variety of applications, from commercial activities, such as car sharing, mining [4], and remote timber terminal operations [5], to social services, including remote earthwork [6], disaster rescue [7], and nuclear disaster response [8]. Most of them [4], [9]–[17] rely on video streaming for situational awareness at the remote end. However, video-based teleoperation technologies face certain challenges in practical applications. Firstly, online video streaming occupies significant bandwidth. As pointed out by [18], at least 3Mbps uplink is required for safe teleoperation using video streaming. A 7.5Mbps uplink is required for high-resolution (720P) video streaming [19]. This hinders the application of video-based teleoperation technologies in bandwidth-limited environments, especially those with multiple end users. Secondly, video-based teleoperation constrains the spatial awareness of the teleoperator. The inherent nature of video feedback offers only a 2D representation of the environment, potentially causing motion sickness and confusion in depth perception [20], [21]. Besides, the fixed positioning of the camera further obscures the teleoperator's sense of direction [22]. Thirdly, video-based teleoperation is sensitive to illumination variations, especially in shadows and non-daylight conditions [23], as well as low-visibility scenarios such as foggy [24] and rainy [23] weather, thereby limiting the operational scope of teleoperation.

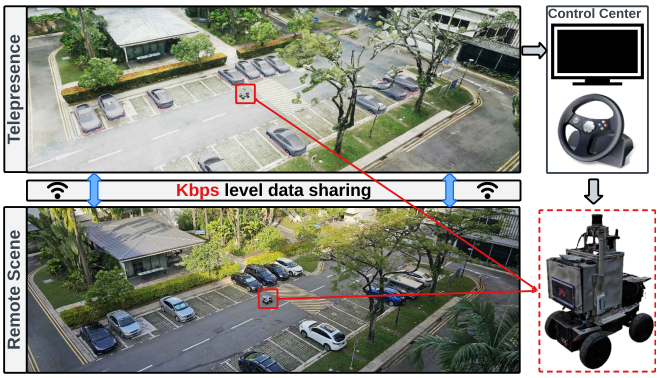


Fig. 1: Telepresence using a 3D digital twin with Kbps-level data transmission

To overcome the challenges outlined above, this paper investigates the minimal environmental information feedback required for teleoperation. In practice, the majority of environmental elements remain unchanged over a long period of time, which thus can be collected in specific digital forms, such as point cloud, neural radiance fields (NeRF) [25], 3D Gaussian Splatting [26] and presented with the help of

This research is supported by the National Research Foundation, Singapore, under the NRF Medium Sized Centre scheme (CARTIN), the Agency for Science, Technology and Research (A*STAR) under its National Robotics Programme (Project No. M22NBK0109), and National Research Foundation, Singapore and Maritime and Port Authority of Singapore under its Maritime Transformation Programme (Project No. SMI-2022-MTP-04). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore and Maritime and Port Authority of Singapore.

C. Zhao, Z. Zhou, H. Liu, D. Kircali, G. Chi, H. Shen, and D. Wang are with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798.

Y. Wang is with the School of Control Science and Engineering, Shandong University, Jinan, China, 250061, and was with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, 639798.

Corresponding author: Yuanzhe Wang, wang0951@e.ntu.edu.sg

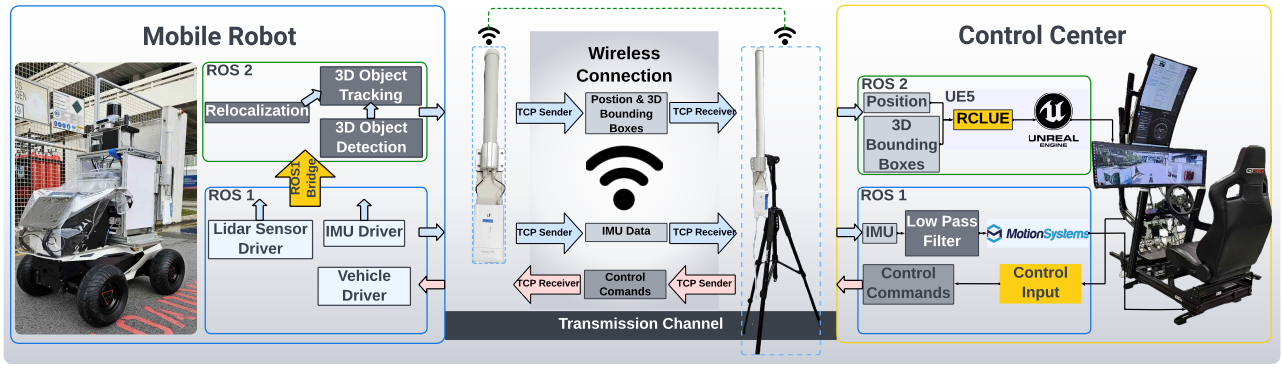


Fig. 2: The schematic diagram of the proposed teleoperation framework

specific graphic engines like Unreal Engine or Unity. As a result, only the information of transient environmental elements such as the vehicle state and dynamic object information needs to be transmitted with merely Kbps-level bandwidth consumption. Inspired by this finding, we propose a new non-video based teleoperation framework for vehicle teleoperation in low-bandwidth environments. The proposed framework adopts a persistent-transient model, which divides environmental elements into two categories: *persistent* and *transient*. The persistent elements are reconstructed in the form of digital twin and stored at the teleoperation platform, while only the transient element information, including vehicle state and perceived dynamic objects, are transmitted in real-time, which only consumes Kbps-level bandwidth. With both persistent and transient information, a 3D virtual scene (see Fig. 1) can be generated by a game engine (e.g. Unreal Engine 5), providing an arbitrary field-of-view (FOV) of the environment. Vehicle motion is recovered by a teleoperation platform at the same time. Experimental results show that the proposed framework provides an immersive telepresence experience for teleoperators with only Kbps-level bandwidth consumption. The contribution of this paper is the development of a new non-video based teleoperation technology with successful implementations on a real vehicle under various lighting conditions. The proposed teleoperation framework has the following salient features:

- *Kbps-level bandwidth consumption*. With the proposed persistent-transient environment model, only the information of transient elements needs to be transmitted to the teleoperation platform, which requires only Kbps-level bandwidth.
- *Realistic spatial awareness*. By providing an adjustable 3D virtual view of the scene, the teleoperator can immersively perceive the remote environment and get a comprehensive spatial understanding from arbitrary views.
- *Insensitive to illumination*. In the proposed framework, vehicle localizes itself with respect to the preconstructed digital twin of the environment based on observations of onboard LiDAR, which is not subject to illumination variations.

The remainder of this paper is organized as follows. Sec. II reviews existing studies on telepresence and efforts to mitigate bandwidth burden. Sec. III introduces the persistent-transient telepresence model. Sec. IV presents the proposed framework. Experimental results are given in Sec. V, while Sec. VI presents conclusions and future work recommendations.

II. Literature Review

This section reviews existing studies on two aspects: telepresence and data transmission.

A. Telepresence

Telepresence is the technology providing the feeling for the teleoperator of being present at a place other than the true location. Many existing related studies concentrate on interface design. For example, reference [17] evaluates the performance of a few display devices and video canvases, while reference [4] develops a panoramic display system to improve situational awareness with path recognition and sense of direction. Moreover, some research efforts [13], [20], [27]–[29] use head-mounted displays (HMDs) to provide immersive feedback. Reference [13] reports an improvement in the operator's task performance by using a mixed reality HMD. Reference [28] compares the performance of monoscopic views and stereoscopic views with a virtual reality (VR) HMD in teleoperation, showcasing the advantages of VR-assisted stereoscopic views in obstacle avoidance. Although significant progress has been made, these works rely heavily on 2D video streaming, where the FOV is constrained by camera specifications. Expanding the FOV would increase the bandwidth burden significantly. Moreover, 2D video cannot represent 3D environments accurately, thereby limiting the depth perception and overall situational awareness.

Some other existing works try to enhance the operator's situational awareness by using 3D visual feedback with multi-modal sensors. Reference [29] develops a modular immersive teleoperation system that incorporates heterogeneous sensor modalities, including 3D lidar, 2D lidar, fisheye camera, audio, and haptic, providing a 3D visualization of the remote scene. By fusing observations from an omnidirectional camera and a 3D lidar, reference [30] generates 3D virtual

scenes using colorized point clouds to improve the spatial awareness of the operator. Another study [31] reconstructs the remote environment using color point clouds, created from depth images and color images of multiple RGB-D cameras, and visualizes the virtual environment using Unity and a VR headset. A VR-based system is designed in [32] for immersive 3D telepresence by presenting the remote scene using a 3D truncated signed distance fields (TSDFs) map, reconstructed from RGB-D data in real-time. Although more realistic scenes are provided, these works introduce an additional bandwidth burden by incorporating information from multiple sensor modalities. Increased bandwidth burden limits their application domain in practice.

B. Data Transmission

Balancing high-fidelity telepresence and bandwidth consumption continues to be a challenge in teleoperation. To provide a higher resolution and a wider FOV for telepresence requires more data, resulting in increased bandwidth usage. In the literature, several approaches have been proposed to address the bandwidth issue, such as the streaming parameter reconfiguration and bitrate allocation in multi-camera streaming systems [9], [15], [16]. In [16], a region of interest (ROI) mask is used to adjust the streaming parameter adaptively for a multi-camera system, such that the bitrate can be reduced. Similarly, reference [9] divides the video into critical and non-critical parts, and allocates a lower bitrate to the latter. Nevertheless, these methods primarily focus on optimizing streaming parameters, which still require substantial bandwidths and degrades the performance of visual presence. With the emergence of high-resolution digital twin, it is now becoming a potential solution for telepresence. Our previous work [33] defines a new teleoperation system, which models and pre-stores the static environment using digital twins and only updates safety-critical information (including dynamic objects and vehicle pose), thereby reducing bandwidth consumption. A similar hypothesis is proposed in [34] that the bandwidth can be reduced by only transmitting processed data, such as segmented images or object lists instead of video stream. These elements are then overlaid with background images from a digital twin to create video stream for teleoperation. However, rather than achieving a practical implementation of a real teleoperation system, this research mainly focuses on verifying the hypothesis by comparing the data rate of ground truth simulator data (e.g. semantic segmented image) and video stream, which limits the study's applicability. To the best of our knowledge, there are no digital twin based teleoperation research works in the literature that have been demonstrated by real-vehicle experiments.

III. Persistent-Transient Telepresence Model

Current teleoperation technologies primarily rely on video streaming for telepresence. However, transmitting a high-resolution video with a large FOV consumes huge communication bandwidth. To address the bandwidth challenge, a new telepresence model is proposed, which categorizes the environmental information into two parts: persistent information $\mathcal{PI}$ and transient information $\mathcal{TI}$. $\mathcal{PI}$ includes static or relatively stable elements such as buildings and roads, while $\mathcal{TI}$ comprises dynamic or temporal elements such as pedestrians and vehicles. The overall environmental information $\mathcal{I}(t)$ is a combination of both parts:

$$\mathcal{I}(t) = \mathcal{PI}(t) \cup \mathcal{TI}(t) \tag{1}$$

In practice, persistent information dominates in the environmental information for telepresence, i.e.,

$$\mathcal{PI}(t) \gg \mathcal{TI}(t) \tag{2}$$

Therefore, communication burden can be substantially released through pre-storing $\mathcal{PI}$ offline and updating $\mathcal{TI}$ online.

A. Persistent Information

Persistent information is processed and pre-stored at a prior moment t_0, which characterizes the environment's persistent features and can be represented by specific forms:

$$\mathcal{D}(t_0) = \{\mathcal{F}_m(\mathcal{O}(\mathcal{I}(t_0)))\} \tag{3}$$

where $\mathcal{F}_m$ represents the pipeline to create the environment statically and digitally (e.g., [35], [36]). $\mathcal{O}$ denotes the sensor observation model used for mapping. After processing, persistent information can be stored in the form of a 3D point cloud map [37], NeRF [25], or 3D Gaussian Splatting [26], and can be visualized with a rendering engine in real-time during teleoperation.

B. Transient Information

Transient information, such as vehicle state and dynamic objects, need to be presented to the teleoperator in real-time. It is the only information that needs to be transmitted and rendered for visualization to enhance the teleoperator's situational awareness. Main information includes

- *Vehicle State* $\mathcal{X}_v(t)$ comprises the vehicle's current pose and motion states, represented as:

$$\mathcal{X}_v(t) = \{P_v(t), K_v(t)\} \tag{4}$$

  where $P_v(t) = \{T_v(t), Q_v(t)\}$ represents the vehicle's pose using translation and orientation in quaternion, and $K_v(t)$ denotes that the kinematics of the vehicle, such as linear acceleration and angular velocity, obtained from IMU.
- *Dynamic Objects* $\mathcal{S}_v(t)$ consists of the perceived dynamic objects in the workspace:

$$\mathcal{S}_v(t) = \{\mathcal{F}_{det}(\mathcal{O}(\mathcal{I}(t)))\} \tag{5}$$

  where $\mathcal{F}_{det}$ is the onboard perception function, and $\mathcal{O}$ denotes the observation model of onboard sensors.

C. Scene Reconstruction

Scene reconstruction comprises two key modules: scene rendering $\mathcal{R}$ and motion synchronization $\mathcal{M}$. Firstly, the scene rendering module integrates the prebuilt digital twin $\mathcal{D}(t_0)$, current vehicle state $\mathcal{X}_v(t)$, and perceived dynamic objects $\mathcal{S}_v(t)$ to provide a virtual scene for the teleoperator. Secondly, the motion synchronization module simulates the vehicle's motion based on the vehicle's attitude measurements. With these two modules, a virtual scene $\mathcal{S}$ is presented to the teleoperator immersively:

$$\mathcal{S} = \{\mathcal{R}(\mathcal{D}(t_0), \mathcal{X}_v(t), \mathcal{S}_v(t)), \mathcal{M}(\mathcal{X}_v(t))\} \tag{6}$$

D. Data Transmission

By using the proposed persistent-transient telepresence model, only the operator's control commands $\mathcal{C}_v(t)$ and transient information $\mathcal{TI}(t)$ are transmitted in the communication channel, which generally takes up Kbps-level bandwidth. In detail,

$$\mathcal{T}rans(t) = \{\mathcal{X}_v(t), \mathcal{S}_v(t), \mathcal{C}_v(t)\} \tag{7}$$

In this way, the communication burden has been released significantly.

IV. Framework

This section presents the framework of our proposed teleoperation system, which comprises three key components: human-machine interface, vehicle platform, and communication module.

A. Human-Machine Interface

To minimize the information required for transmission, the human-machine interface is developed based on our proposed persistent-transient telepresence model. Firstly, persistent environmental information is constructed as a virtual reality of the environment and stored in the teleoperation platform. Secondly, transient environmental information is obtained and transmitted in real-time. Persistent and transient information is processed to present a virtual reality scene to the teleoperator.

1) Persistent Information for Telepresence: Persistent information is used to build a virtual model of the environment in two steps. Firstly, the geometric structure of the environment is constructed in the form of the point cloud map. Secondly, the photogrammetry mesh is applied to improve visual appearance.

- *Geometric Structure*. Firstly, we use Point LIO [38] to construct a point cloud map using Lidar and IMU measurements. Considering that dynamic elements within the point cloud map may degrade re-localization accuracy, we project the point cloud of each scan onto a semantic image segmented by [39] and filter out points corresponding to pixel classes that are potentially movable, such as vehicles and motorcycles. The filtered point cloud map, which represents the geometric structure of the environment, is stored on the vehicle, and used for re-localization afterward.
- *Visual Appearance*. In real implementations, the point cloud map provides precise geometric information of the environment but lacks consistent visual representation. Moreover, using point cloud maps with a high resolution increases computational burden, resulting in large time delays during rendering. Consequently, a high-resolution 3D color mesh is employed as the visual appearance model to provide consistent visualization. To generate a 3D mesh model that closely matches the real scene, we use photogrammetry technique to reduce visual discrepancies. In detail, we collect high-resolution images in the workspace. To match the visual appearance model with the geometric structure, both images and the prebuilt point cloud map are processed by iTwin Capture. Aerotriangulation is conducted with manually selected 3D and 2D point pairs as initial correspondences between images and point clouds. Then, a high-resolution mesh model can be generated by iTwin Capture with the mesh model's coordinate aligned with the point cloud map. The mesh model is subsequently refined using Blender by manual cleanup and editing. Then, we import the mesh model into Unreal Engine 5. With the pose estimation information from the re-localization module, we render the photogrammetry mesh accordingly, ensuring a seamless integration of the vehicle within its virtual representation.

2) Transient Information for Telepresence: To ensure safe teleoperation, transient information, including vehicle state and dynamic objects nearby, needs to be presented to the teleoperator.

- *Vehicle State*. The vehicle ego pose with respect to the prebuilt geometry map, i.e. the point cloud map, is estimated by a Lidar-IMU based pose tracking approach [40], which is implemented through an Unscented Kalman Filter (UKF) [41]. In the prediction phase, vehicle pose is predicted based on linear acceleration and angular velocity measurements from IMU. This serves as an initial estimate for the NDT scan matching algorithm, which aligns the current LiDAR scan with the prebuilt map. Subsequently, vehicle pose estimation is corrected based on the matching results. Additionally, the vehicle's motion state, including linear acceleration and angular velocity, is measured by the onboard IMU.
- *Dynamic Objects*. Dynamic objects, such as vehicles and pedestrians, threaten vehicle safety during teleoperation, which thus needs to be perceived accurately and presented to the teleoperator in real-time. In the developed system, we have implemented a 3D object detection and tracking module engaged with Autoware Universe [42]. A lidar-based 3D object detection model, CenterPoint [43], is trained on the nuScenes dataset [44], which contains a variety of categories that are prevalent in driving scenarios such as cars, trucks, pedestrians, and so forth. This model has been trained on nuScenes for 20 epochs and gained the mAP of 0.5725 in the test set. The module is integrated with a

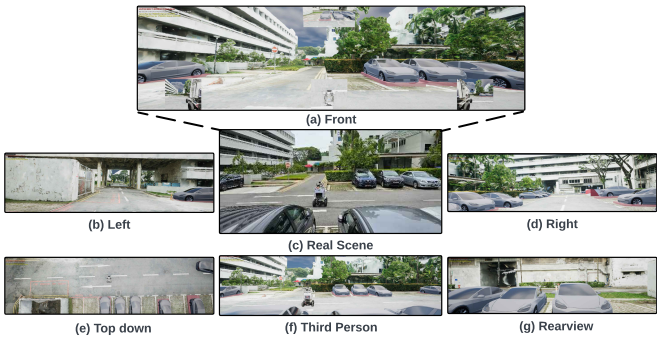


Fig. 3: The developed telepresence system with multiple adjustable virtual viewpoints

360° horizontal Field of View (FOV) 32-line spinning LiDAR, RoboSense Helios. To ensure real-time inference, we export this model from PyTorch to the ONNX model and subsequently convert it to FP16-engine files of TensorRT. The final inference time with TensorRT is reduced to approximately 40 ms when deployed on the mobile robot's edge computing unit with a GeForce RTX 2060 Mobile GPU. After detection, a multi-object tracking module is adopted to correlate 3D objects across both temporal and spatial dimensions, enabling the measurement of speed and motion trajectories. This module employs the min-cost flow algorithm, muSSP [45], to associate the detected objects across frames, and uses the Extended Kalman Filter (EKF) to track objects, employing distinct motion models for different object categories. The entire detection and tracking pipeline is executed within 100ms, with a frame rate exceeding 10 frames per second, thereby achieving real-time detection and tracking.

3) Scene Reconstruction from Hybrid Representation: With both persistent and transient information, vehicle-side scenes can be recreated at the teleoperation platform. Scene reconstruction comprises two key parts: scene rendering and motion synchronization.

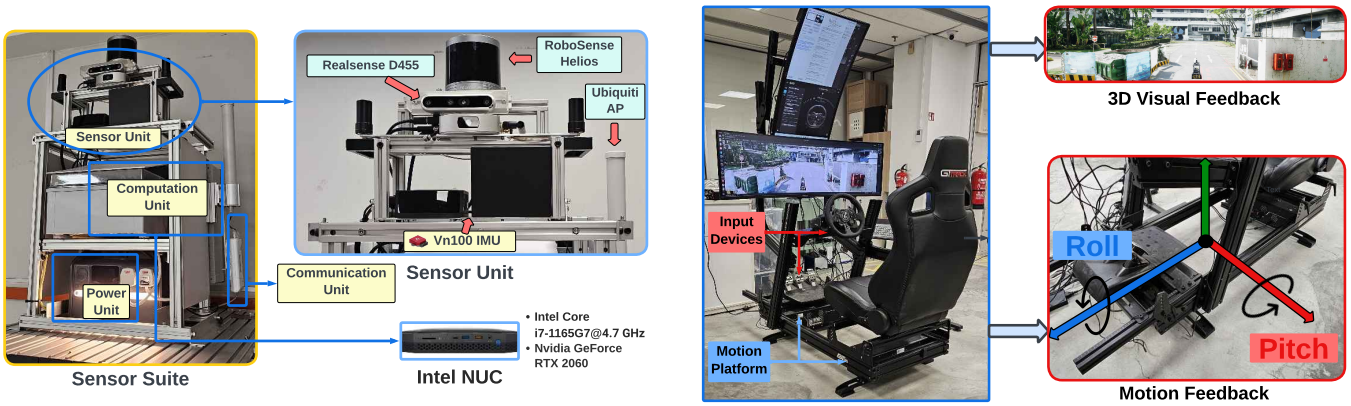
- *Scene Rendering*. The rendering system is developed based on Unreal Engine 5 to provide a 3D virtual reality of the remote environment, providing the functionality of arbitrary viewpoint selection. The operator can switch between different viewpoints, as shown in Fig. 3. Functionalities including zoom in, zoom out, and shift are supported as well. Moreover, the rendering system also incorporates pose, speed, and scale information of detected objects into the synthesis process. A pre-set vehicle model will be rendered for objects detected as vehicles with a safety boundary below and a speed display on top. The scale of the safety boundary represents the real scale of the vehicle, and the color of the boundary indicates whether the vehicle is moving or not, as shown in Fig. 5c. For pedestrians, body animations such as running, walking, and standing are added to the pedestrian model based on the speed of the pedestrians. If an object cannot be classified, a red-white box of its real scale will be put into place to signify unknown obstacles. In this way, the operator gains a better spatial awareness, understanding of the surroundings, and insight into the intentions of objects through an adjustable immersive 3D view of the surrounding scenes and objects.
- *Motion Synchronization*. The remote scene is finally presented to the teleoperator on a specialized platform, as illustrated in Fig. 4b. This setup offers an immersive experience by providing both 3D visual and motion feedback. The operator has the flexibility to configure the 3D scene with an optimal viewpoint. Moreover, they can feel vehicle's acceleration and vibrations through the platform's pitch and roll movements. These motions are synthesized by transforming the IMU measurements from the remote vehicle. To improve the quality of feedback, a low-pass frequency filter is applied in the motion platform to attenuate high-frequency noise inherent in the IMU data.

B. Vehicle Platform

The developed system is general for a variety of vehicle platforms. In this paper, we have used the developed teleoperation system to control an Agilex Hunter 2.0 UGV, an Ackermann steering robot as shown in Fig. 2. A perception and localization sensor suite has been developed for the teleoperation system, as shown in Fig. 4a. This sensor suite is equipped with an Intel NUC with a mobile GPU for data processing, an Ubiquiti Wireless AP for communication, and a 512Wh power station for power supply. In addition, the sensor suite comprises a few sensors with multiple modalities, including a RoboSense Helios Lidar, a RealSense D455 camera, and a VN100 IMU. Extrinsic calibration has been conducted between the Helios Lidar and the D455 camera using the method proposed in [46], while calibration between the Helios Lidar and the VN100 IMU has been conducted using the method provided in [47].

C. Communication Module

The communication between the vehicle and the teleoperation platform, as shown in Fig. 2, involves transmitting control commands from the platform to the vehicle and updating the platform with real-time information on the vehicle's state and detected dynamic objects. Our system adopts a hybrid ROS structure due to certain modules running in ROS 1, such as sensor drivers, while others operate in ROS 2, e.g., the Rendering module. On the vehicle side, sensor drivers use ROS 1 to publish data. Specifically, IMU data is sent to the platform using a TCP sender. A ROS 1-ROS 2 bridge (*ros1_bridge*) is utilized to facilitate the conversion of sensor data from ROS 1 message to ROS 2 message, which are then fed into our localization and perception modules (ROS 2 nodes). Subsequently, both positional data and 3D object detection results, encapsulated in ROS 2 messages, are then transmitted through a TCP sender. On the platform side, the TCP receiver dedicated to IMU data publishes this data via ROS 1 upon reception. The motion platform then



(a) Perception and localization sensor suite

(b) Human interface for telepresence and control

Fig. 4: The overview of the developed teleoperation system

processes the IMU ROS messages and simulates vehicle motion. Simultaneously, another TCP receiver receives vehicle position and detection data and publishes it via ROS2, which is further processed by Unreal Engine 5 through *rclUE* (a middleware that acts as a bridge between Unreal Engine 5 and ROS 2). Unreal Engine 5 utilizes this information to render the vehicle's surroundings and generate 3D mesh models of detected objects. Besides, control signals from the operator are relayed back to the vehicle via TCP as ROS Twist messages to facilitate vehicle control.

V. Experimental Results

A. Experimental Setup

To validate the proposed teleoperation framework, a series of experiments have been conducted at Nanyang Technological University, Singapore. A Hunter robot equipped with our designed sensor suite was used in the experiments, as detailed in Sec. IV-B. A specialized teleoperation platform is equipped with a steering wheel, a braking system, a widescreen monitor, two electric motors for motion synchronization, a desktop computer with an Intel Core i9 13900K processor, an NVIDIA GeForce RTX 4090 GPU, and 64GB of RAM. The communication system comprises 3 Ubiquiti Wireless Access Points (APs), with a mobile AP mounted on the vehicle, a base station AP at the teleoperation platform, and an intermediary AP as a relay to bridge the communication link between the vehicle and the teleoperation platform.

Experiments were conducted at two different locations, respectively. One is an open carpark surrounded by campus buildings, and the other is a descending urban roadway with two distinct crosswalks. The traditional video streaming based teleoperation technology was used as the baseline method for bandwidth comparison. The streaming function is developed based on GStreamer, with a RealSense D455 Camera mounted on the vehicle, providing 1280×800 resolution images with FOV of $90° \times 65°$. The streaming pipeline encodes videos using x264 codec.

B. Discussions

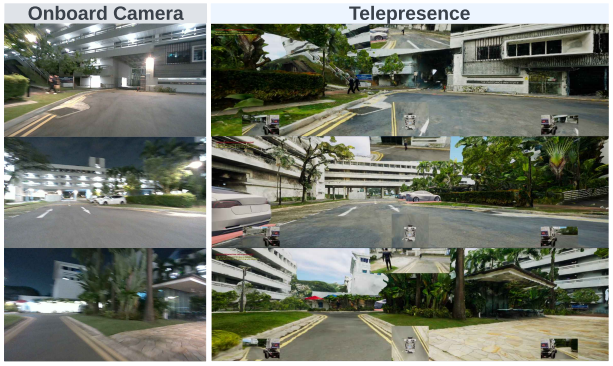
Experimental results of our proposed telepresence framework are shown in Fig. 5. The left images in each subfigure show the snapshots of the onboard camera, while the right images are the corresponding virtual scene presented to the teleoperator. It can be observed that our proposed framework can provide a larger FOV with adjustable viewpoints (see Fig. 3), which enhances the spatial awareness of the teleoperator. Figs. 5a and 5b showcase the teleoperation results of the vehicle during nighttime at a carpark and on a public road, respectively. Obvious visibility degradation of the camera can be observed, which was caused by motion blur and overexposure. In contrast, our proposed telepresence framework is insensitive to lighting conditions. A realistic virtual scene with daytime lighting conditions can be generated even at night, thereby enhancing the teleoperator's situation awareness.

C. Quantitative Analysis

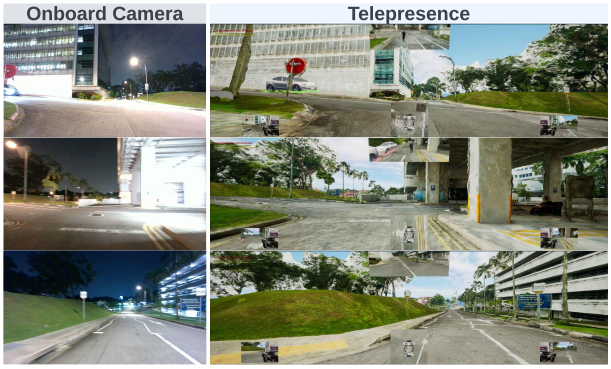
The bandwidth consumption of our proposed teleoperation framework has been quantitatively evaluated in the public road scenario under day and night lighting conditions, respectively. Our proposed framework and the video streaming framework have been tested respectively with the vehicle moving along the same route and the same network setup. The bandwidth data was collected from the Ubiquiti Wireless AP's control panel. As detailed in Tab. I, nighttime tests reveal that our proposed method has achieved an average bandwidth consumption of 598.389Kbps, in contrast to 2.251Mbps for the conventional single front camera based video streaming based method. Daytime evaluations show similar results, with the 453.269Kbps bandwidth consumption of our method versus the 2.224Mbps bandwidth consumption of the video streaming based method. The results show a significant bandwidth reduction in our system when compared to the video streaming based method.

VI. Conclusions

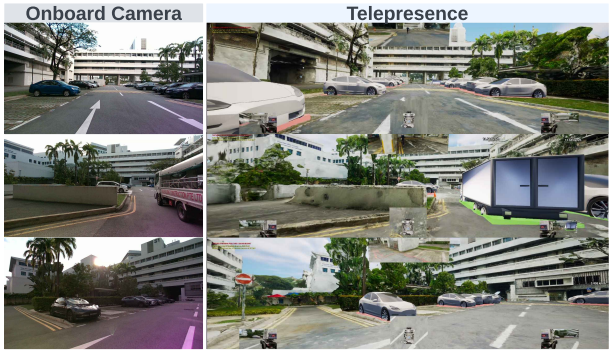
This paper has introduced a new non-video based teleoperation framework for bandwidth-limited applications. A



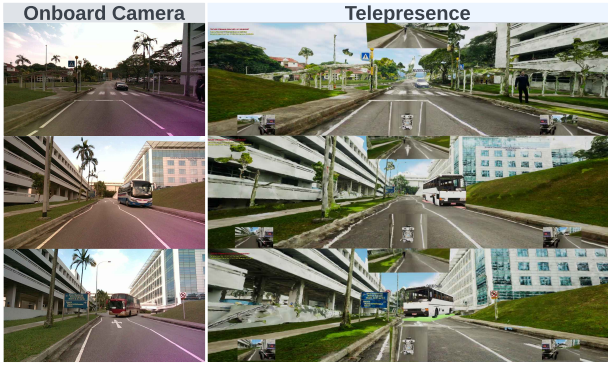
(a) Teleoperation with Hunter UGV at a carpark, nighttime



(b) Teleoperation with Hunter UGV along a road, nighttime



(c) Teleoperation with Hunter UGV at a carpark, daytime



(d) Teleoperation with Hunter UGV along a road, daytime

Fig. 5: Experimental results in various conditions

TABLE I: The comparative results of bandwidth consumption

Telepresence System	Nighttime		Daytime	
	Average Bandwidth	Std. Deviation	Average Bandwidth	Std. Deviation
Ours	**598.389 Kbps**	**208.895**	**453.269 Kbps**	**162.323**
Single Camera Streaming	2251.016 Kbps	186.183	2224.745 Kbps	141.900

persistent-transient environment model has been proposed for scene reconstruction at the remote end. Within this framework, a digital twin of the environment is constructed and stored at the teleoperation platform, using persistent information gathered in advance, while only transient information, such as vehicle state and dynamic objects, necessitates real-time transmission. The proposed framework enables the provision of a realistic 3D virtual representation of the remote environment to the teleoperator, consuming only Kbps-level bandwidth. Furthermore, the framework demonstrates resilience to illumination variations. Real vehicle experiments have been conducted to validate the proposed framework. Future research will explore its potential applications, including environmental service, port operation, and smart mining.

REFERENCES

[1] T. Litman, "Autonomous vehicle implementation predictions: implications for transport planning," 2020.

[2] J. Zhou and J. Beyerer, "Corner cases in data-driven automated driving: definitions, properties and solutions," in *2023 IEEE Intelligent Vehicles Symposium (IV)*, 2023, pp. 1–8.

[3] T. Zhang, "Toward automated vehicle teleoperation: vision, opportunities, and challenges," *IEEE Internet of Things Journal*, vol. 7, no. 12, pp. 11 347–11 354, 2020.

[4] C. A. James, T. P. Bednarz, K. Haustein, L. Alem, C. Caris, and A. Castleden, "Tele-operation of a mobile mining robot using a panoramic display: an exploration of operators sense of presence," in *2011 IEEE International Conference on Automation Science and Engineering*, 2011, pp. 279–284.

[5] A. Mahmood, S. F. Abedin, M. O'Nils, M. Bergman, and M. Gidlund, "Remote-timber: an outlook for teleoperated forestry with first 5g measurements," *IEEE Industrial Electronics Magazine*, 2023.

[6] K. You, C. Zhou, L. Ding, W. Chen, R. Zhang, J. Xu, Z. Wu, and C. Huang, "Earthwork digital twin for teleoperation of an automated bulldozer in edge dumping," *Journal of Field Robotics*, vol. 40, no. 8, pp. 1945–1963, 2023.

[7] K. D. Katyal, C. Y. Brown, S. A. Hechtman, M. P. Para, T. G. McGee, K. C. Wolfe, R. J. Murphy, M. D. Kutzer, E. W. Tunstel, M. P. McLoughlin, and M. S. Johannes, "Approaches to robotic teleoperation in a disaster scenario: from supervised autonomy to direct control," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2014, pp. 1874–1881.

[8] M. Chiou, G.-T. Epsimos, G. Nikolaou, P. Pappas, G. Petousakis, S. Mühl, and R. Stolkin, "Robot-assisted nuclear disaster response: report and insights from a field exercise," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 4545–4552.

[9] S. Neumeier, V. Bajpai, M. Neumeier, C. Facchi, and J. Ott, "Data rate reduction for video streams in teleoperated driving," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 10, pp. 19 145–19 160, 2022.

[10] X. Shen, Z. J. Chong, S. Pendleton, G. M. James Fu, B. Qin, E. Frazzoli, and M. H. Ang, "Teleoperation of on-road vehicles via

immersive telepresence using off-the-shelf components," in *Intelligent Autonomous Systems 13: Proceedings of the 13th International Conference IAS-13*. Springer, 2016, pp. 1419–1433.

[11] L. Kang, W. Zhao, B. Qi, and S. Banerjee, "Augmenting self-driving with remote control: challenges and directions," in *Proceedings of the 19th International Workshop on Mobile Computing Systems & Applications*, 2018, pp. 19–24.

[12] J. Y. Chew, M. Kawamoto, T. Okuma, E. Yoshida, and N. Kato, "Adaptive attention-based human machine interface system for teleoperation of industrial vehicle," *Scientific Reports*, vol. 11, no. 1, p. 17284, 2021.

[13] A. Hosseini and M. Lienkamp, "Enhancing telepresence during the teleoperation of road vehicles using hmd-based mixed reality," in *2016 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2016, pp. 1366–1373.

[14] Y. Yu and S. Lee, "Remote driving control with real-time video streaming over wireless networks: design and evaluation," *IEEE Access*, vol. 10, pp. 64 920–64 932, 2022.

[15] A. Schimpe, S. Hoffmann, and F. Diermeyer, "Adaptive video configuration and bitrate allocation for teleoperated vehicles," in *2021 IEEE Intelligent Vehicles Symposium Workshops (IV Workshops)*, 2021, pp. 148–153.

[16] M. Hofbauer, C. B. Kuhn, M. Khlifi, G. Petrovic, and E. Steinbach, "Traffic-aware multi-view video stream adaptation for teleoperated driving," in *2022 IEEE 95th Vehicular Technology Conference:(VTC2022-Spring)*. IEEE, 2022, pp. 1–7.

[17] J.-M. Georg, E. Putz, and F. Diermeyer, "Longtime effects of videoquality, videocanvases and displays on situation awareness during teleoperation of automated vehicles," in *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2020, pp. 248–255.

[18] S. Neumeier, E. A. Walelgne, V. Bajpai, J. Ott, and C. Facchi, "Measuring the feasibility of teleoperated driving in mobile networks," in *2019 Network Traffic Measurement and Analysis Conference (TMA)*. IEEE, 2019, pp. 113–120.

[19] A. Gaber, W. Nassar, A. M. Mohamed, and M. K. Mansour, "Feasibility study of teleoperated vehicles using multi-operator lte connection," in *2020 International Conference on Innovative Trends in Communication and Computer Engineering (ITCE)*, 2020, pp. 191–195.

[20] Y. Luo, J. Wang, R. Shi, H.-N. Liang, and S. Luo, "In-device feedback in immersive head-mounted displays for distance perception during teleoperation of unmanned ground vehicles," *IEEE Transactions on Haptics*, vol. 15, no. 1, pp. 79–84, 2021.

[21] D. R. Tyczka, R. Wright, B. Janiszewski, M. J. Chatten, T. A. Bowen, and B. Skibba, "Study of high-definition and stereoscopic head-aimed vision for improved teleoperation of an unmanned ground vehicle," in *Unmanned Systems Technology XIV*, vol. 8387. SPIE, 2012, pp. 226–243.

[22] E. J. Fabris, V. A. Sangalli, L. P. Soares, and M. S. Pinho, "Immersive telepresence on the operation of unmanned vehicles," *International Journal of Advanced Robotic Systems*, vol. 18, no. 1, p. 1729881420978544, 2021.

[23] S. Neumeier, S. Stapf, and C. Facchi, "The visual quality of teleoperated driving scenarios how good is good enough?" in *2020 International Symposium on Networks, Computers and Communications (ISNCC)*, 2020, pp. 1–8.

[24] Z. Hong, Q. Zhang, X. Su, and H. Zhang, "Effect of virtual annotation on performance of construction equipment teleoperation under adverse visual conditions," *Automation in Construction*, vol. 118, p. 103296, 2020.

[25] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[26] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," *ACM Transactions on Graphics*, vol. 42, no. 4, 2023.

[27] K. Doki, K. Suzuki, A. Torii, S. Mototani, Y. Funabora, and S. Doki, "Application of augmented reality based on sensing data to teleoperation system for operator support," in *2021 22nd IEEE International Conference on Industrial Technology (ICIT)*, vol. 1, 2021, pp. 19–24.

[28] Y. Luo, J. Wang, H.-N. Liang, S. Luo, and E. G. Lim, "Monoscopic vs. stereoscopic views and display types in the teleoperation of unmanned ground vehicles for object avoidance," in *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 2021, pp. 418–425.

[29] J.-M. Georg and F. Diermeyer, "An adaptable and immersive real time interface for resolving system limitations of automated vehicles with teleoperation," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, 2019, pp. 2659–2664.

[30] M. Oehler and O. von Stryk, "A flexible framework for virtual omnidirectional vision to improve operator situation awareness," in *2021 European Conference on Mobile Robots (ECMR)*. IEEE, 2021, pp. 1–6.

[31] D. Wallace, Y. H. He, J. C. Vaz, L. Georgescu, and P. Y. Oh, "Multimodal teleoperation of heterogeneous robots within a construction environment," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 2698–2705.

[32] P. Stotko, S. Krumpen, M. Schwarz, C. Lenz, S. Behnke, R. Klein, and M. Weinmann, "A vr system for immersive teleoperation and live exploration with a mobile robot," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 3630–3637.

[33] D. Wang, Y. Yue, D. Kircali, and S. R. Mharolkar, "High fidelity teleoperation," Patent 2021-041-01-SGPRV, 2021, inventors: Wang Danwei, Yue Yufeng, Kircali Dogan, and Mharolkar Sanat Rajesh.

[34] P. Kremer, N. Nourani-Vatani, and S. Park, "A digital twin for teleoperation of vehicles in urban environments," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 12 521–12 527.

[35] G. Kim and A. Kim, "Remove, then revert: static point cloud map construction using multiresolution range images," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 10 758–10 765.

[36] H. Lim, S. Hwang, and H. Myung, "Erasor: egocentric ratio of pseudo occupancy-based dynamic object removal for static 3d point cloud map building," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 2272–2279, 2021.

[37] C. Zheng, Q. Zhu, W. Xu, X. Liu, Q. Guo, and F. Zhang, "Fast-livo: fast and tightly-coupled sparse-direct lidar-inertial-visual odometry," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 4003–4009.

[38] D. He, W. Xu, N. Chen, F. Kong, C. Yuan, and F. Zhang, "Point-lio: robust high-bandwidth light detection and ranging inertial odometry," *Advanced Intelligent Systems*, p. 2200459, 2023.

[39] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1290–1299.

[40] K. Koide, J. Miura, and E. Menegatti, "A portable three-dimensional lidar-based system for long-term and wide-area people behavior measurement," *International Journal of Advanced Robotic Systems*, vol. 16, no. 2, p. 1729881419841532, 2019.

[41] E. Wan and R. Van Der Merwe, "The unscented kalman filter for nonlinear estimation," in *Proceedings of the IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium (Cat. No.00EX373)*, 2000, pp. 153–158.

[42] Autoware Foundation, "Autoware.universe," 2020. [Online]. Available: \url{https://autowarefoundation.github.io/autoware-documentation/}

[43] 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. 11 784–11 793.

[44] 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.

[45] C. Wang, Y. Wang, Y. Wang, C.-T. Wu, and G. Yu, "Mussp: efficient min-cost flow algorithm for multi-object tracking," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[46] G. Yan, Z. Liu, C. Wang, C. Shi, P. Wei, X. Cai, T. Ma, Z. Liu, Z. Zhong, Y. Liu, *et al.*, "Opencalib: A multi-sensor calibration toolbox for autonomous driving," *Software Impacts*, vol. 14, p. 100393, 2022.

[47] F. Zhu, Y. Ren, and F. Zhang, "Robust real-time lidar-inertial initialization," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 3948–3955.