# Toward Intelligent Transportation With Pedestrians and Vehicles In-the-Loop: A Surveillance Video-Assisted Federated Digital Twin Framework

Xiaolong Li, Jianhao Wei, Haidong Wang, Li Dong, Ruoyang Chen, Changyan Yi, Jun Cai, Dusit Niyato, and Xuemin Shen

## Abstract

In intelligent transportation systems (ITSs), integrating pedestrians and vehicles into traffic management models is essential for developing realistic and safe solutions. However, current systems often fail to simulate complex, real-world scenarios due to the absence of a comprehensive digital twin framework across diverse traffic environments and effective modeling of pedestrian-vehicle interactions. In this article, we propose a surveillance video-assisted federated digital twin (SV-FDT) framework to enhance ITSs by incorporating pedestrians and vehicles into the control loop. SV-FDT improves computational efficiency and communication performance by transmitting only semantic data and agent parameters, rather than raw video streams. The proposed framework adopts three-layer architecture and constructs detailed pedestrian-vehicle interaction models using multi-source traffic surveillance videos. The three-layer architecture includes: (i) an end layer that collects surveillance videos from multiple sources; (ii) an edge layer that performs self-supervised semantic segmentation to extract interactions, converts them into executable traffic codes, and generates local digital twin systems (LDTSs) for regional traffic modeling; and (iii) a cloud layer that integrates LDTSs into a real-time global digital twin model. Key design considerations, challenges, and practical implementation guidelines are discussed for SV-FDT, and a testbed evaluation is used to show that SV-FDT improves traffic flow, reduces mirroring delay, and enhances recognition accuracy and system efficiency compared to traditional terminal-server frameworks. Finally, we outline open challenges and potential directions for future research in digital twin-enabled ITS.

## Introduction

Intelligent transportation systems (ITSs) are envisioned to utilize information and communication technologies to enhance road safety and traffic management. Since pedestrians and vehicles are the natural and primary participants in transportation systems, incorporating both entities into real-time decision-making and control processes becomes necessary for ITSs, giving rise to the concept of pedestrian and vehicle in-the-loop. Such an in-the-loop scheme requires to continuously collect and process data from various sources, dynamically adjusting traffic operations based on real-time pedestrian-vehicle interactions, with both vehicles and pedestrians acting as active participants and decision-makers in the traffic management process. Compared to traditional ITSs, which primarily focus on vehicle modeling and often neglect pedestrians and their impacts [1], [2], [3], ITSs with pedestrians and vehicles in the loop can model and analyze the mutual influence between vehicles and pedestrians to provide safer, more efficient, and human-satisfying traffic management. However, enabling ITS applications with pedestrians and vehicles in-the-loop is non-trivial. It requires a cohesive system to characterize real-world traffic environments and conduct behavioral simulations of pedestrians and vehicles.

Recently, digital twin (DT) has been increasingly recognized as a powerful tool for building ITSs, offering real-time virtual representations of traffic operations to support data-driven decision-making and optimize traffic management [4]. The federated digital twin (FDT), which can be seen as an advanced form of DT [1], [2], allows the integration of geographically dispersed DTs into a unified framework. FDT offers significant potential for ITSs to achieve pedestrians and vehicles in-the-loop, even when traffic participants, environments, and infrastructures are spread across different regions, each with numerous features and data sources. In addition, FDT allows data processing to be distributed, and thus can improve scalability and reduce latency. Furthermore, FDT enhances ITS' ability to predict future events using real-time data updates from multiple sources. All these capabilities are vital for proactive decision-making, such as pedestrian crossing safety and traffic signal optimization over a large area with multiple correlated intersections.

*Xiaolong Li, Jianhao Wei (corresponding author), Haidong Wang (corresponding author), and Li Dong are with the Xiangjiang Laboratory and the Hunan Provincial General University Key Laboratory of IoT Intelligent Sensing and Distributed Collaborative Optimization, Hunan University of Technology and Business, Changsha 410205, China; Ruoyang Chen and Changyan Yi are with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China; Jun Cai is with the Department of Electrical and Computer Engineering, Concordia University, Montreal, QC H3G 1M8, Canada; Dusit Niyato is with the College of Computing and Data Science, Nanyang Technological University, Singapore 639798; Xuemin Shen is with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada.*

Digital Object Identifier: 10.1109/MNET.2025.3574228
Date of Current Version: 18 November 2025
Date of Publication: 27 May 2025

In recent years, several DT frameworks have been developed to advance ITS applications [1], [2], [3], [5], [6]. For example, Wang et al. [5] proposed a cloud-edge-device collaborative DT framework for mobility services, but it did not provide implementation details. A DT co-simulation framework for pedestrians and vehicles was introduced in [6], but it lacked twin agent modeling for these traffic participants. Khan et al. [1] and Yu et al. [2] proposed two general FDT architectures for ITSs; however, neither included a technical system implementation. Tang et al. [3] presented an FDT-enabled collision warning framework for autonomous driving, focusing on a semi-asynchronous federated learning scheme to reduce training delays. Despite these contributions, most prior works focus on high-level architectural designs or isolated applications, and share several critical limitations: (1) Lack of pedestrian-vehicle interaction modeling. Existing DT solutions primarily emphasize vehicle modeling while overlooking pedestrian-vehicle interactions and their impacts on traffic dynamics; (2) High communication costs and privacy risks. Centralized ITS architectures require processing large volumes of surveillance video data, leading to excessive bandwidth usage, latency, and privacy concerns; (3) Absence of practical FDT implementations. No existing framework integrates pedestrian-vehicle interaction modeling with real-time FDT representations in complex traffic environments. These gaps highlight the urgent need for a holistic framework that addresses these intertwined challenges.

To address these limitations, in this article, we propose a surveillance video-assisted FDT implementation framework (SV-FDT) for ITSs with pedestrians and vehicles in-the-loop. The proposed SV-FDT is built on a cloud-edge-end collaborative architecture and introduces the agent concept to construct pedestrian-vehicle interaction models. The framework integrates semantic segmentation technology [7] for processing surveillance videos, which allows pix-level object recognition and offers granular insights into video data. SV-FDT can achieve timely video acquisition and data integration from multiple sources, precise modeling of pedestrian-vehicle twin agents at different positions, and seamless aggregation of local DT models across different regions. To our best knowledge, SV-FDT is the first DT implementation framework to enable ITSs with pedestrian and vehicle in-the-loop. Our key contributions are summarized as follows:

- We propose SV-FDT, a novel cloud-edge-end collaborative framework for implementing ITSs with pedestrians and vehicles in-the-loop. Its architecture includes i) an end layer that collects surveillance videos from widespread traffic surveillance cameras, ii) an edge layer that performs visual understanding, agent-based interaction modeling, and LDTS creation in local regions, and iii) a cloud layer that integrates LDTSs across geographically dispersed regions to construct a global DT model in real-time. To unlock its full potential, SV-FDT seamlessly integrates algorithms for semantic segmentation, semantic-to-code transformation, and agent-based modeling.
- We analyze key design requirements and challenges for semantic segmentation-based FDT construction in SV-FDT, offering practical guidance for deploying FDTs via traffic video surveillance. We also highlight potential but promising solutions to these challenges, inspiring effective SV-FDT implementation in dynamic transportation.
- We validate SV-FDT through a case study in the CARLA simulation environment to optimize traffic light control in traffic management. Our results show that SV-FDT outperforms traditional terminal-server frameworks in terms of mirroring delay, recognition accuracy, and subjective evaluations.
- We summarize the open challenges in this field and propose future research directions.

## Framework of SV-FDT

### System Architecture and Key Techniques

The proposed SV-FDT system architecture consists of three layers: end, edge, and cloud, as illustrated in Fig. 1.

**End Layer:** This layer consists of terminal devices such as surveillance cameras, radars, and in-vehicle sensors. Its primary function is to collect large-scale, real-time traffic data, particularly surveillance videos. For instance, high-definition (HD) surveillance cameras capture detailed visual data on traffic scenarios, such as vehicle movement, pedestrian crossings, and road conditions. Radars detect and track vehicles and pedestrians, providing information on their speed, location, and movement. In-vehicle sensors, e.g., GPS and ultrasonic sensors, contribute precise real-time localization and movement data. In addition to real-time data collection, these devices with integrated processing modules perform data fusion. Combining information (e.g., the locations and speeds of vehicles and pedestrians) from multiple sources can reduce uncertainties from solely relying on a single data source.

**Edge Layer:** At the edge layer, edge nodes, equipped with GPUs or NPUs, are responsible for data preprocessing tasks, including i) redundancy elimination to video data uploaded by terminal devices for abundant storage and bandwidth; ii) data aggregation to obtain unified data from heterogeneous sources, e.g., traffic cameras, radars, and in-vehicle sensors, by combining radar measurements and in-vehicle sensor data with the visual analysis results; and iii) synchronization of temporal data to minimize latency discrepancies [8]. More importantly, the edge nodes should conduct semantic segmentation to classify each pixel within each frame of surveillance video, enabling effective identification of pedestrians, vehicles, road markings, traffic signs, barriers, and weather conditions. Powered by advanced vision encoders, it extracts detailed visual features for precise pixel-level classification, continuously generating structured semantic data that accurately represents current traffic scenes. After that, the edge layer adopts a semantic-to-code transformation module that converts semantic data into machine-readable traffic codes, enabling their direct use in traffic simulation systems and facilitating the DT creation and update of physical environments by edge nodes. Meanwhile,

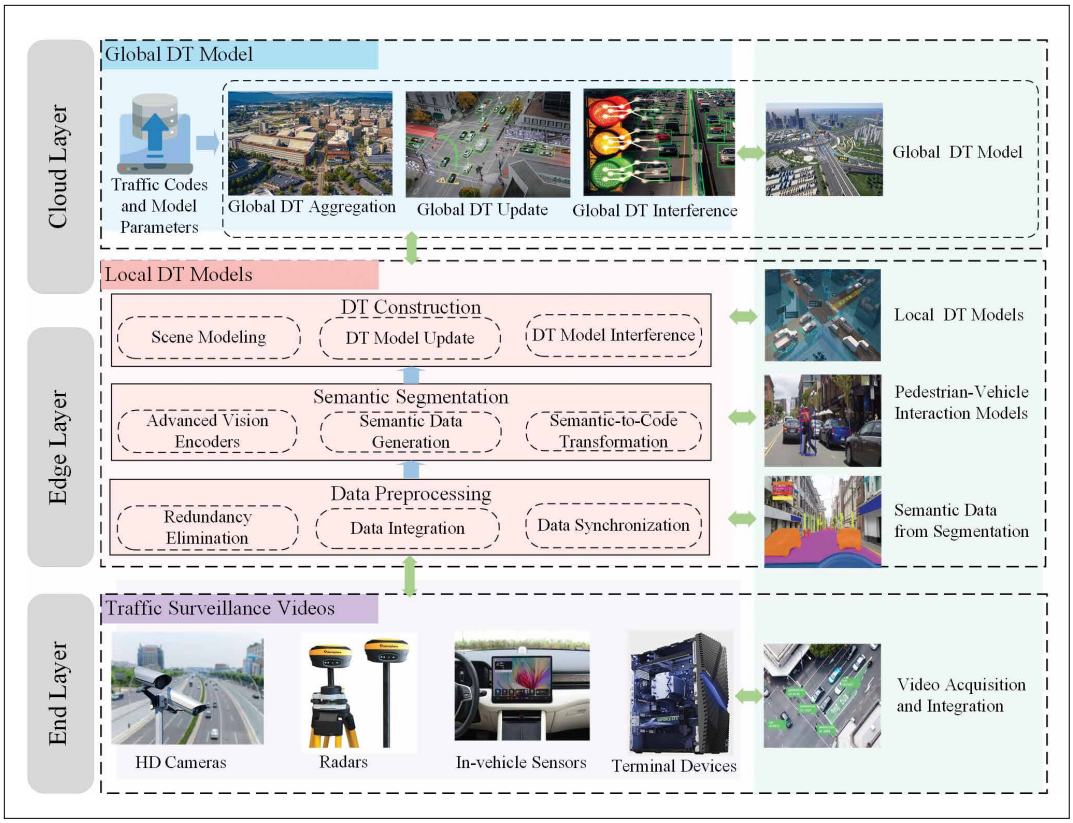


**FIGURE 1.** The overall architecture of the proposed SV-FDT system. SV-FDT consists of three layers: the end layer contains multiple terminal devices; the edge layer handles tasks for data preprocessing, semantic segmentation, and local DT construction; and the cloud layer constructs the global model and conducts model inferences for ITS applications.

the edge nodes transmit traffic codes and model parameters of twin agents to the cloud layer for aggregation to maintain spatiotemporal consistency between local and global DT models. Decision-making and inference results are returned to end-layer terminal devices, enabling adaptive signal control, path optimization, and personalized guidance to improve pedestrian safety and vehicle efficiency.

**Cloud Layer:** The cloud layer aggregates local DT models from distributed edge devices, creating an up-to-date global DT representation using traffic codes and agent-based model parameters across the transportation network. The global DT model continuously adapts to updates from local models, capturing new patterns in pedestrian-vehicle interactions. Only traffic codes and model parameters are transmitted to protect data privacy, safeguarding sensitive information about pedestrians and vehicles. After updating the global model, the cloud layer distributes global DT model parameters to the edge layer for synchronized updates of local models, enhancing decision-making accuracy.

## Example Scenarios

Here, we present three typical application scenarios of the proposed SV-FDT framework for ITSs with pedestrians and vehicles in-the-loop.

**Extreme Pedestrian-Vehicle Flow Testbed In-the-Loop:** SV-FDT can effectively implement ITSs with pedestrians and vehicles in-the-loop by providing a dynamic testbed to manage their interactions. The framework can support various automated driving levels, from Level 0 (fully manual) to Level 5 (fully autonomous). In each scenario, SV-FDT simulates real-world traffic challenges, such as sudden pattern changes and adverse weather conditions, and adapts by incorporating real-time data from diverse sources. This integration allows the testbed to reflect the varying capabilities of autonomous systems across different levels. By leveraging predictive analytics and collaborative decision-making, FDTs proactively identify and manage potential pedestrian-vehicle conflicts, unlike traditional testbeds that can only react after these situations occur.

**Emergency and Disaster Management:** Integrating pedestrians and vehicles into SV-FDT can strengthen emergency and disaster management within ITSs. Traditional DT systems without in-the-loop characteristics rely on historical traffic data or static models for delayed emergency responses. Unlike these systems, SV-FDT can continuously monitor and simulate real-world traffic scenarios, and adapt vehicles to unexpected traffic events and situations, including traffic incidents, road blockages, vehicle breakdowns, and sudden pedestrian crowds. Additionally, the system uses predictive models to anticipate emergencies, enabling the accurate assessment of traffic management strategies such as rerouting vehicles and safety notification of pedestrians. The in-the-loop characteristic means that the behaviors of pedestrians and vehicles are integrated into the system's

decision-making processes, allowing for proactive rather than reactive management.

**Customized Navigation and Travel Guidance:** Without in-the-loop FDTs, conventional navigation guidance systems rely on static or historical traffic data. However, they do not account for real-time conditions such as pedestrian activities, sudden traffic disruptions, or evolving road conditions. SV-FDT can be instrumental in enabling personalized navigation systems that cater to the specific needs of both drivers and pedestrians. By integrating data from various sources, such as surveillance cameras and in-vehicle sensors, FDTs create a dynamic, user-centered environment that adapts to individual preferences. For example, commercial drivers can receive guidance to avoid narrow streets or select routes suitable for heavy vehicles. Likewise, family drivers are provided with optimized routes that prioritize safety and efficiency. This adaptability allows FDTs to provide personalized, context-aware travel guidance that enhances the user experience and facilitates seamless interactions between pedestrians and vehicles in ITSs.

## Design Requirements and Challenges

To establish a cloud-edge-end collaborative framework for real-time FDT construction in ITS, we identify the following key design requirements and challenges:

**1) Ultra-Reliable and Agile Global DT Model Aggregation With Pedestrian-Vehicle Interactions:** Real-time pedestrian-vehicle interactions require ultra-reliable data transmission to ensure safety and enable quick adjustments to traffic signals and vehicle routing. Therefore, achieving "ultra-reliability" and "agility" in aggregating global DT models from local data is paramount.

***Timely Video Acquisition and Integration With Precise Co-Camera Control:*** Real-time video data collection and integration from potentially heterogeneous terminal cameras present challenges in synchronization and coordination [9]. Therefore, a distributed network protocol is essential for timely video acquisition and integration, ensuring that video data is aligned temporally and processed without delays for real-time and accurate DT modeling [10]. Coordination techniques, such as distributed scheduling and synchronization, facilitate seamless communication and data sharing among devices, ensuring efficient real-time video data aggregation from multiple cameras. To ensure smooth HD streaming, the end devices, typically the surveillance camera, must achieve a minimum video transmission rate of 10 Mbps and a frame rate of 30 frames per second, supporting H. 264 or H. 265 video coding. Besides, cameras can be retrofitted with GPU-powered hardware modules to improve efficiency by using lightweight semantic segmentation algorithms to extract granular motion features of traffic elements (e.g., speeds, locations, and directions). Driven by these technologies, co-camera control enables multiple cameras to collaboratively capture videos from various angles and locations, ensuring comprehensive data integrity.

***Multi-Source Data Processing and Distributed Data Synchronization With Privacy Protection:*** Leveraging multimodal data collected from diverse sources in ITS, e.g., HD cameras and radars, provides valuable insights into ITS' full feature spaces, including the appearance, locations, directions, and speed of both pedestrians and vehicles. Consolidating such information requires advanced multimodal data processing and distributed synchronization techniques to create unified feature profiles. This data processing and synchronization process must occur efficiently, ensuring quick updates to the global DT model. However, collecting and transmitting multimodal data across numerous sources may face significant privacy challenges. From a cross-layer security perspective, privacy-aware access control and signal protection in both the physical and data link layer, user indistinguishability enhancement, cross-device authentication, and customized packet encryption in upper layers are essential. Advanced technologies, such as physical layer security, local differential privacy, zero-trust architecture, and blockchain, can effectively mitigate these risks.

***Elastic Resource Orchestration for Fine-Grained DT Visualization Under Traffic Dynamics:*** Edge servers play a critical role in constructing DTs with low-latency requirements yet face significant challenges. High latency can hinder the real-time monitoring of pedestrians and vehicles, leading to uncoordinated interactions and safety risks. Therefore, an integrated approach to task scheduling, communication spectrum management, computing resources, and storage allocation becomes essential [11]. Improving data transmission reliability requires employing ultra-reliable, low-latency communication (URLLC) technologies such as time-sensitive networks and mobile edge computing. These technologies support DT services, including immersive virtual reality and collaborative autonomous driving. For instance, immersive virtual reality demands high-rate video delivery from end to edge, requiring 99.999999% reliability and a latency of no more than 1 ms [12].

**2) Customized and Sustainable Multi-Agent Maintenance:** To enable customized pedestrian-vehicle interactions in ITSs, traffic participants are modeled as independent twin agents, containing virtual models replicating the physical objects and artificial intelligence for predictive decision-making. "Customization" addresses diverse and complex user needs, while "sustainability" focuses on the long-term evolution of FDT systems. Computational optimization is crucial for achieving customized and sustainable multi-agent maintenance, ensuring that FDT systems adapt to user needs and support sustainable development over time.

***Achieving Lightweight and Redundant-Free Replication of Large-Scale Video Footage:*** In DTs, videos uploaded by multiple terminals often contain significant redundancy, consuming excessive storage space and network bandwidth while increasing the risk of exposing sensitive pedestrian-vehicle information. Redundancy can hinder semantic segmentation, reducing the responsiveness and safety of pedestrian-vehicle interactions. Therefore, de-redundancy algorithms are essential for constructing DTs and reducing storage consumption. To obtain lightweight and redundant-free replication, techniques such as data deduplication algorithms, model pruning, knowledge distillation, video compression, and

semantic communication can be employed, promoting effective management of both pedestrians and vehicles.

***Balancing Standardized and Personalized Model Extraction for Pedestrian-Vehicle Agents:*** Achieving a balance between standardization and personalization is crucial in pedestrian-vehicle interactions. Standardized agent models should support some industry standards such as UL-4600[1] and can capture general features, such as typical pedestrian behavior patterns and vehicle operation modes. In contrast, personalized agent models focus on individual differences, including walking patterns, driving styles, and vehicle configurations. Although commonalities exist, significant individual variations must be considered to model pedestrian-vehicle interactions precisely. Standardized models depend on extensive data collection, whereas personalized models require individual feature extraction via deep neural networks. An imbalance between the standardized and personalized DT model construction in ITS may result in low-quality global and local models within FDT construction. This significantly degrades the accuracy of both i) the global decision-making, e.g., traffic control and emergency management; and ii) the local decision-making, e.g., customized navigation and travel planning.

***Forward-Thinking and Spatially Versatile Cross-Model Migration Under Ubiquitous Mobility:*** Predicting user behaviors in DT models for highly mobile users, such as pedestrians and drivers, is challenging due to its inherent unpredictability. For instance, drivers might deviate from suggested routes due to personal preferences or unforeseen circumstances, complicating the transfer of models. To tackle this issue, generative AI (GAI) technology can simulate potential user behaviors and decision-making paths, aiding in scenario forecasting. Transfer learning techniques can also be employed to personalize and refine the model, enhancing its adaptability to individual user needs. This approach facilitates effective management of interactions between pedestrians and vehicles, ultimately delivering more intelligent and tailored services.

## System Implementation

The SV-FDT framework presented in the section "Framework of SV-FDT" shows that cloud-edge-end collaboration is critical for real-time data integration and distributed processing in dynamic traffic environments, and continuous model refinement and data management can ensure local responsiveness and global coherence. In this section, we describe in detail the implementation of the proposed SV-FDT framework.

### DT Construction by Cloud-Edge-End Collaboration

Fig. 2 illustrates an implementation of the SV-FDT framework, which incorporates seven key steps: real-time video acquisition and integration, time-sensitive transmission, self-supervised semantic segmentation, pedestrian-vehicle interaction modeling, and local and global DT model construction. It also explicitly visualizes the data flow across functional modules at the end, edge, and cloud layers. SV-FDT leverages the CARLA simulation environment to create traffic scenes and develop local and global DT models. Prior systems primarily consist of real-time video acquisition, time-sensitive transmission, and DT model construction. SV-FDT addresses key gaps by incorporating self-supervised semantic segmentation, semantic-to-code transformation, pedestrian-vehicle interaction modeling, and real-time digital twin construction. CARLA, an open-source simulator built on Unreal Engine, generates high-fidelity traffic environments and accurately simulates traffic participant behaviors. SV-FDT utilizes CARLA's map editor to design immersive traffic scenarios featuring vehicles, pedestrians, and detailed surroundings.

**The end layer** is responsible for real-time video data acquisition via HD cameras and sensors, ensuring timely traffic perception. At the end layer, SV-FDT employs HD cameras and other terminal devices to collect and integrate real-time surveillance videos, establishing a foundation for constructing local and global DT models.

**The edge layer** handles critical processes, such as self-supervised semantic segmentation, semantic-to-code transformation, pedestrian-vehicle interaction modeling, and local DT model construction, enabling local-area traffic DT. At the edge layer, before semantic segmentation, edge nodes perform essential data preprocessing, including redundancy elimination, data integration, and synchronization. After that, semantic segmentation algorithms are applied to identify traffic elements in the video, such as pedestrians, vehicles, road markings, traffic signs, barriers, and weather conditions. Both traffic elements' static attributes (e.g., appearance, location, and direction) and dynamic attributes (e.g., speed, behaviors, and movement trajectories) can be captured. Based on these attributes, semantic segmentation generates continuous, structured semantic data to support a granular understanding of visual inputs. The semantic-to-code transformation module then translates semantic data into executable traffic codes, which are used to generate digital twinning of real-world scenes. Such traffic codes define attributes of vehicles and pedestrians, including locations, speeds, and movement directions. Edge nodes continuously update local DT models and simulate pedestrian and vehicle behaviors by successively executing generated traffic codes.

**The cloud layer** is a central hub for global DT model aggregation via semantic data and cross-region decision-making. At the cloud layer, servers constantly receive semantic data and model parameters of pedestrian-vehicle agents from edge nodes to achieve a unified global DT representation while ensuring spatiotemporal coherence without transmitting sensitive visual data. We could exploit the global DT model to develop and implement ITS applications running on cloud servers, such as traffic signal control and traffic flow prediction.

## Implementation of Local and Global Digital Twins

**Real-Time Video Acquisition and Integration:** Video compression standards, such as H.264 and H.265, are applied to reduce bandwidth consumption while maintaining high video quality. Adaptive bitrate streaming adjusts video quality

[1] UL-4600 is available at https://www.intertek.com/automotive/ul-4600

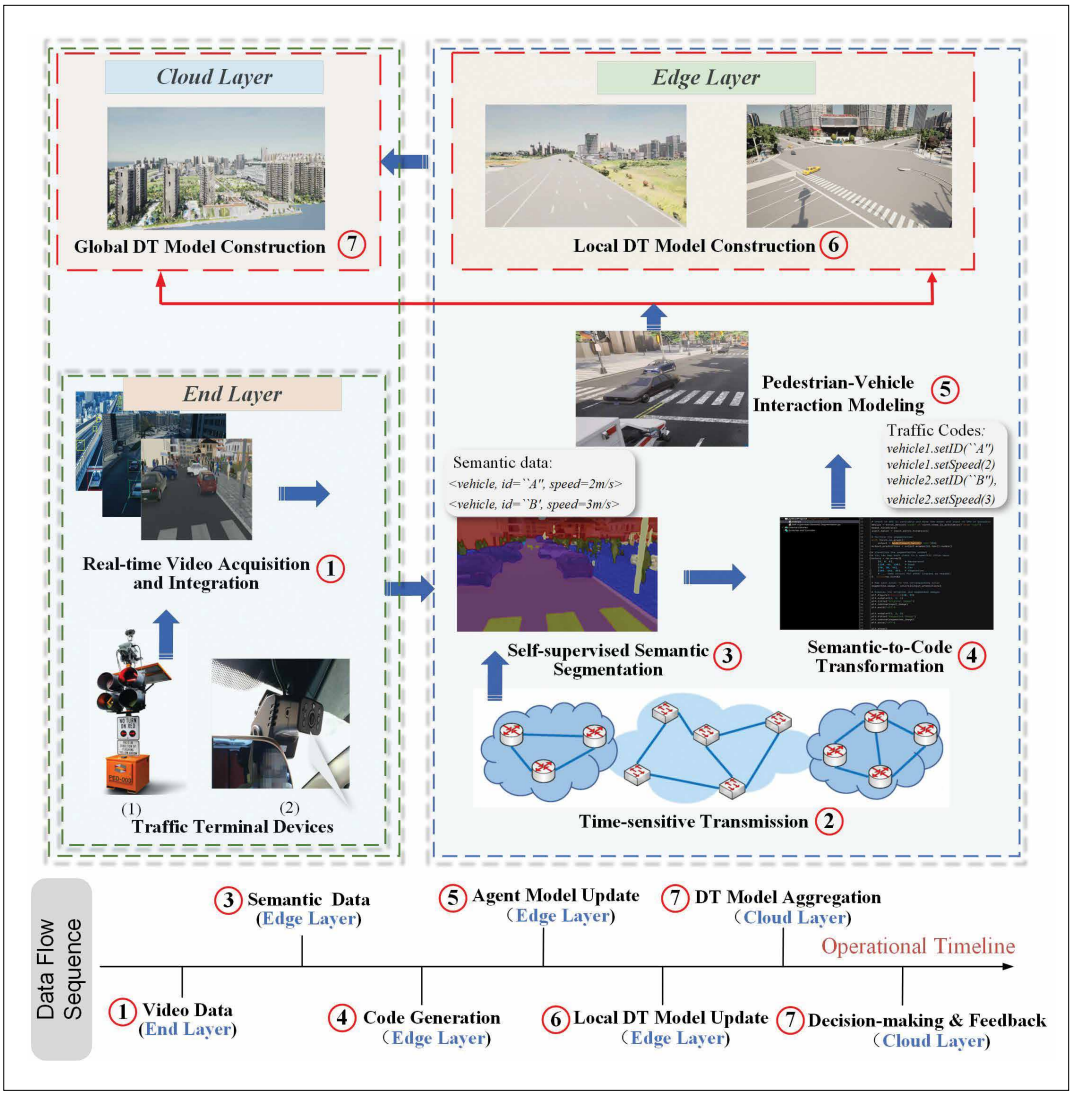


**FIGURE 2.** Systematic framework of the proposed SV-FDT. The semantic segmentation module extracts vehicle data, e.g., *<vehicle, id=A, speed=2m/s>*. The semantic-to-code transformation module converts this data into executable traffic codes: *vehicle1.setID(A), vehicle1.setSpeed(2))*. These codes are subsequently processed and executed by both local and global DT models.

dynamically according to bandwidth availability, ensuring seamless streaming. With the time synchronization protocol, all video frames and sensor data are timestamped for accurate synchronization during analysis. Multi-sensor fusion algorithms, such as Kalman Filters, are used to integrate data from various sources to ensure synchronized object representation across modalities.

**Time-Sensitive Transmission:** URLLC technologies are employed to enable immersive interactions in DTs. Edge computing further minimizes latency by positioning computational resources closer to data sources (e.g., HD cameras and radars), enabling parallel processing of multimodal data and reducing transmission delays. Based on latency requirements and importance, we prioritize data packets such as real-time perception, semantic understanding, and signal timing control, ensuring critical packets are transmitted first. We dynamically adjust transmission rates and route paths according to real-time network conditions and communication needs.

**Self-Supervised Semantic Segmentation:** We prefer self-supervised semantic segmentation over unsupervised methods due to its ability to deliver real-time, pixel-level insights into traffic elements without manual labeling, which is time-consuming and costly. Advanced models such as DINO[2] excel at capturing long-range dependencies and contextual relationships, making them ideal for complex traffic scenarios. Specifically, we enhance semantic perception through a combination of contrastive learning for discriminative feature extraction, temporal correlation modeling for consistent object tracking, and context-aware refinement to suppress background interference. To mitigate the effects of occlusion and limited viewpoints, we implement cross-camera panoramic feature fusion and Transformer-based attention mechanisms that recover missing spatial information. Additionally, granularity-controllable segmentation allows precise customization of output detail to align with user requirements [7]. Techniques such as contrastive learning enhance the model's ability to identify key attributes at the

[2] https://github.com/IDEA-Research/MaskDINO

pixel level. For example, optical flow measures pedestrian and vehicle speeds, while pixel-level annotations precisely position traffic elements. By extracting static and dynamic attributes through semantic segmentation, we generate continuous, structured semantic data to support a detailed visual understanding of traffic scenes. This semantic data is then converted into traffic codes via the semantic-to-code transformation module.

**Semantic-to-Code Transformation:** The semantic-to-code transformation module is a core component of the framework, converting structured semantic data into executable traffic codes for CARLA simulation. This process uses rule-based algorithms and advanced deep learning techniques, e.g., generative AI and large language models (LLMs), to generate traffic codes based on traffic elements and their attributes, as illustrated in Fig. 2. The transformation logic ensures accurate scene replication within CARLA environments by adhering to CARLA's APIs and command structures. We continuously monitor traffic code performance in CARLA to maintain accuracy, refining the transformation logic as needed to correct discrepancies. This iterative process ensures that the traffic codes are both precise and reliable.

**Pedestrian-Vehicle Interaction Modeling:** In the CAR-LA environment, the vehicular motion model is based on a kinematic bicycle model, enabling realistic simulation of acceleration, deceleration, lane changes, and turning behaviors. To explicitly model pedestrian dynamics, a collision-free velocity model is employed [13], allowing pedestrians to exhibit natural walking patterns such as acceleration, deceleration, hesitation, and spontaneous crossings in response to nearby vehicles and infrastructure layouts. We develop pedestrian-vehicle twin agents with both standardized and personalized features to simulate realistic interactions. Using model-driven agent modeling, we treat traffic participants as independent, autonomous agents, combining standardized traits for predictable behaviors in typical scenarios with personalized features derived from individual-specific data. Each twin agent operates autonomously in the CARLA environment, independently perceiving its surroundings, making decisions, and executing actions. This hybrid approach balances standardized and personalized behaviors, enabling agents to dynamically adjust and produce realistic, responsive interactions in complex traffic environments. To enhance the system's ability to model highly unpredictable behaviors of pedestrians and vehicles, such as sudden pedestrian road crossings and abrupt stops, the agent models integrate two core technologies: semantic-level interaction representation and a probabilistic behavior inference module. The semantic-level interaction representation captures high-level intentions of both pedestrians and vehicles, including actions like waiting, crossing, and hesitating. To address inherent randomness and spontaneity, the probabilistic behavior inference module employs Gaussian process regression to generate behavioral predictions. A context-aware attention mechanism further refines these predictions in real time by adapting to interactions with surrounding agents. These components enable SV-FDT to effectively capture and respond to unpredictable behaviors, enhancing the system's realism and adaptability in complex urban environments.

**Local and Global DT Model Construction:** Local and global DT models are constructed and updated using semantic data and model parameters of twin agents. Edge nodes build 3D local panoramic DT models of roads, buildings, and infrastructure, either online or offline, and continuously update them by executing traffic codes. These traffic codes dynamically adjust attributes, such as weather, vehicle positions, and pedestrian presence, ensuring that local DT models remain accurate and up-to-date. The global DT model aggregates semantic data and agent parameters from distributed edge nodes, providing a unified, large-scale view of the transportation network. Data exchanges between edge nodes and cloud servers are designed to maintain privacy and spatiotemporal coherence by omitting sensitive visual information. As local DT models update in real time, the global DT model synchronizes these updates to preserve a consistent and coherent representation of traffic conditions across regions. Local DT models mirror specific traffic regions in ITS applications, such as navigation guidance, by incorporating pedestrian-vehicle interactions and evolutionary reasoning based on real-time updates. The global DT integrates these updates to provide a comprehensive network view for guiding vehicles and pedestrians along the most efficient or safest routes.

To meet the real-time demands of ITS applications, we introduce a latency-aware synchronization mechanism to coordinate local and global DT models. The total update delay, $T_{\text{update}}$, includes transmission, processing, and aggregation delays. To ensure timely responsiveness, we impose the constraint:

$$T_{\text{update}} \leq T_{\text{threshold}} \tag{1}$$

where $T_{\text{threshold}}$ is empirically set to 0.5 seconds based on safety-critical latency requirements in ITS environments. To satisfy this constraint, SV-FDT adopts URLLC protocols and packet prioritization techniques to reduce transmission delays. Time-sensitive packet scheduling dynamically prioritizes critical semantic data and traffic control messages to ensure they are transmitted first. This mechanism enables SV-FDT to maintain reliable synchronization and real-time DT updates across system layers.

## Case Study and Experimental Results

Given the overall architecture of the proposed SV-FDT framework, this section presents a pioneering case study on optimizing traffic light time setting, addressing design requirements and challenges as presented in the section "Framework of SV-FDT."

### Experimental Setup and System Design

As illustrated in Fig. 3, we developed a testbed platform based on SV-FDT to optimize traffic signal timing in ITS. The end layer includes HD cameras, radars, and traffic lights. The cameras and radars collect visual data, which they preprocess and integrate locally. The edge layer comprises a desktop computer serving as the traffic signal control center and five laptops

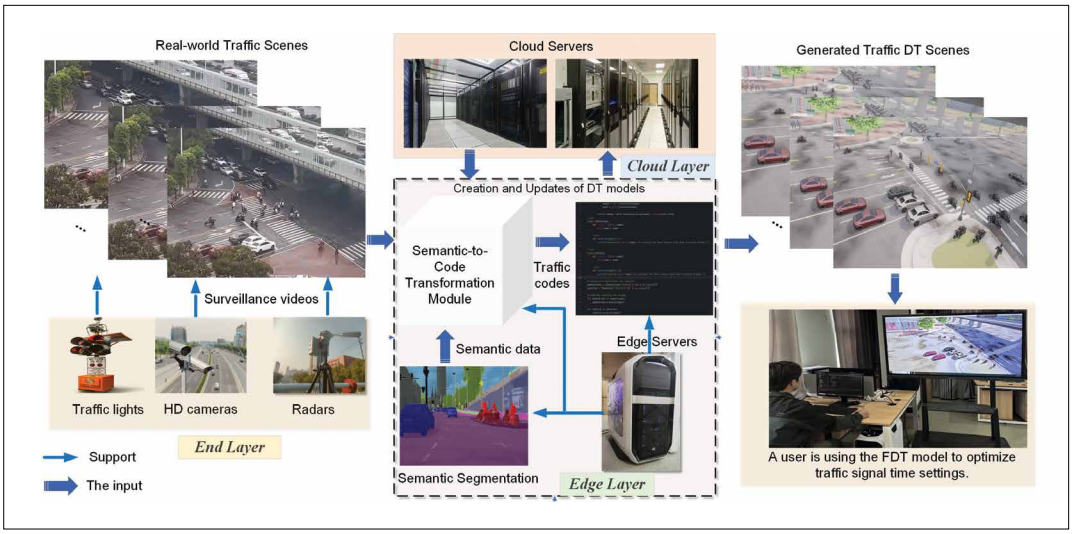


**FIGURE 3.** The SV-FDT framework-based FDT testbed platform for optimizing traffic signal timing.

equipped with NVIDIA RTX 4060 GPUs and i9-14900HX processors, functioning as edge nodes. To enable flexible deployment of edge nodes under resource-constrained conditions, we employ lightweight semantic segmentation algorithms and large language models with significantly reduced parameter sizes based on available edge resources. One edge node runs a lightweight semantic segmentation algorithm to monitor pedestrian and vehicle positions, generating real-time semantic data on their speeds and trajectories. A second node handles semantic-to-code transformation, converting this data into executable traffic control commands for the CARLA simulation environment. A third node is responsible for 3D scene modeling, generating a panoramic representation of the road and surrounding structures. The fourth and fifth edge nodes maintain local DT models for two subregions, updating them in real time. The traffic signal control center uses these local DT models to optimize signal settings dynamically.

The traffic signal control center employs a fuzzy clustering algorithm combined with a density-based spatial clustering algorithm (FCA-DBSCAN) [14] to estimate pedestrian and vehicle densities. FCA-DBSCAN leverages traffic structure, real-time pedestrian and vehicle densities, road width, and pedestrian walking speed to enable adaptive signal control. The cloud layer hosts two servers: one for generating the global DT model and the other for executing global traffic signal optimization. Fig. 4 illustrates the deployment of the SV-FDT system at the intersection of Yuelu Ave. and Wanglong Rd. in Yuelu District, Changsha, China. Real-time federated digital twin scenes are generated from traffic surveillance videos captured by three cameras. Due to substantial traffic fluctuations throughout the day, this location serves as an ideal urban testing site. To evaluate SV-FDT's dynamic signal control capabilities, experiments were conducted at 07:00, 14:00, and 18:00, representing the morning peak, off-peak, and evening peak periods, respectively. SV-FDT dynamically adjusts signal timings based on road width, pedestrian walking speed, and traffic volume to enhance safety and optimize traffic flow.

The experiments utilized 4-megapixel cameras capturing traffic with a bandwidth of 10 Mbps, supporting real-time data collection and processing at 30 frames per second. Drone aerial footage provided offline video data to the scene modeling node, enabling the pre-construction of a static 3D panoramic DT model. End nodes connect to edge nodes via Gigabit Ethernet for efficient video transfer, while edge-to-cloud communication occurs over 5G networks at speeds up to 200 Mbps. Lightweight UniMatch V2+ DINOv2 encoders[3] perform semantic segmentation to track real-time pedestrian and vehicle flow, generating structured semantic data at 30 times per second sent to the semantic-to-code transformation node. This transformation node utilizes the WizardCoder language model as its foundational pretraining model. After fine-tuning with semantic data and traffic codes, WizardCoder generates executable traffic codes from structured semantic inputs. The local DT node loads the 3D panoramic model from the scene modeling node and uses CARLA to execute these codes, simulating real-world pedestrian-vehicle interactions at the frequency of 30. Meanwhile, the global DT modeling node integrates semantic data from all local DT models, creating a comprehensive global DT model within CARLA. The aggregation and updates of the global DT are done less frequently than those of the local DTs. Instead, updates occur periodically every minute or when significant changes in traffic patterns, such as traffic events, are detected. Ultimately, SV-FDT dynamically adjusts signal timings to improve road safety. The case study showcases the feasibility and validity of the SV-FDT framework.

## Performance Evaluation

Performance evaluation of the proposed platform is based on seven key metrics: 1) **Mirroring Delay**: The time required to replicate real-world traffic conditions within the digital twin system. 2) **Recognition Accuracy**: The proportion of correctly identified vehicles and pedestrians in

[3] https://github.com/LiheYoung/UniMatch-V2

surveillance video frames. 3) **Traffic Flow Under Different Signal Modes**: The volume of vehicles passing through the intersection under various traffic signal control strategies. 4) **Mirroring Delay Intervals Under Varying Confidence Levels**: The distribution of mirroring delays at different confidence thresholds. 5) **Delay Under Different Resource Constraints**: The latency performance under limited computational or bandwidth resources. 6) **Signal Timing Settings Under Varying Recognition Accuracy Ratios:** Traffic signal duration settings corresponding to different recognition accuracy ratios. 7) **Subjective Evaluation:** Based on user feedback regarding video fluency, cyber-physical consistency, and movement synchronization. Based on experimental data at the intersection of Yuelu Ave. and Wanglong Rd., Fig. 5 compares SV-FDT with the traditional terminal-server framework across the above performance metrics.

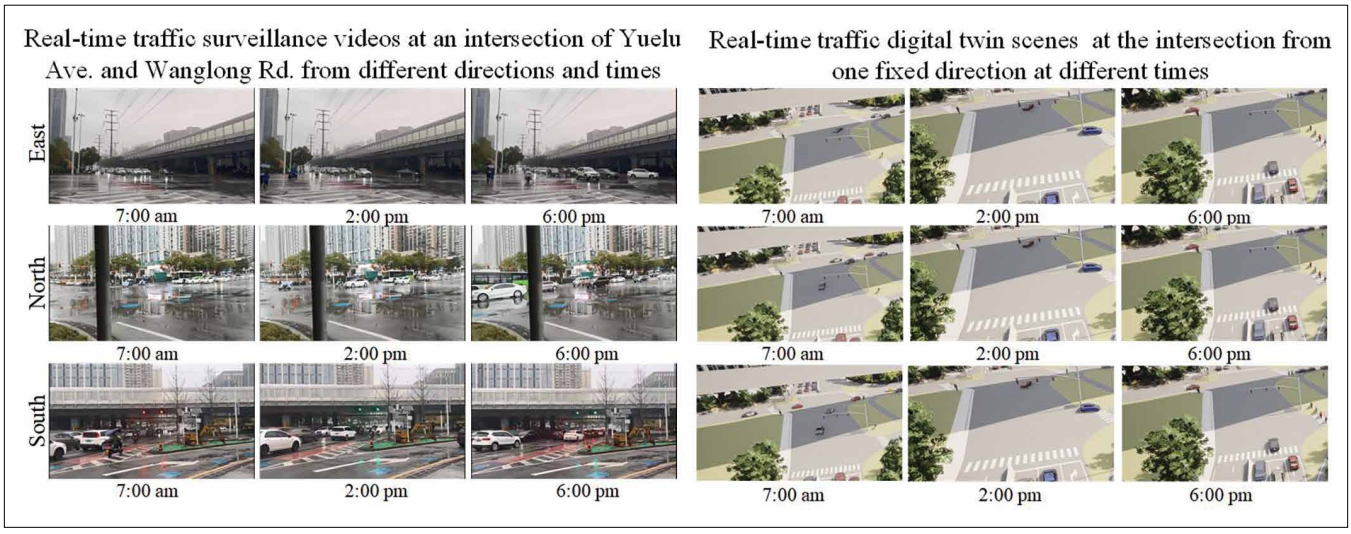


**FIGURE 4.** Real-time traffic digital twin scenes at the intersection of Yuelu Ave. and Wanglong Rd. at three different times of day (Time: 7:00 AM, Morning Peak Hour; Time: 2:00 PM, Midday Off-Peak; Time: 6:00 PM, Evening Peak Hour), generated by the proposed SV-FDT model based on real-time traffic surveillance videos from three cameras.

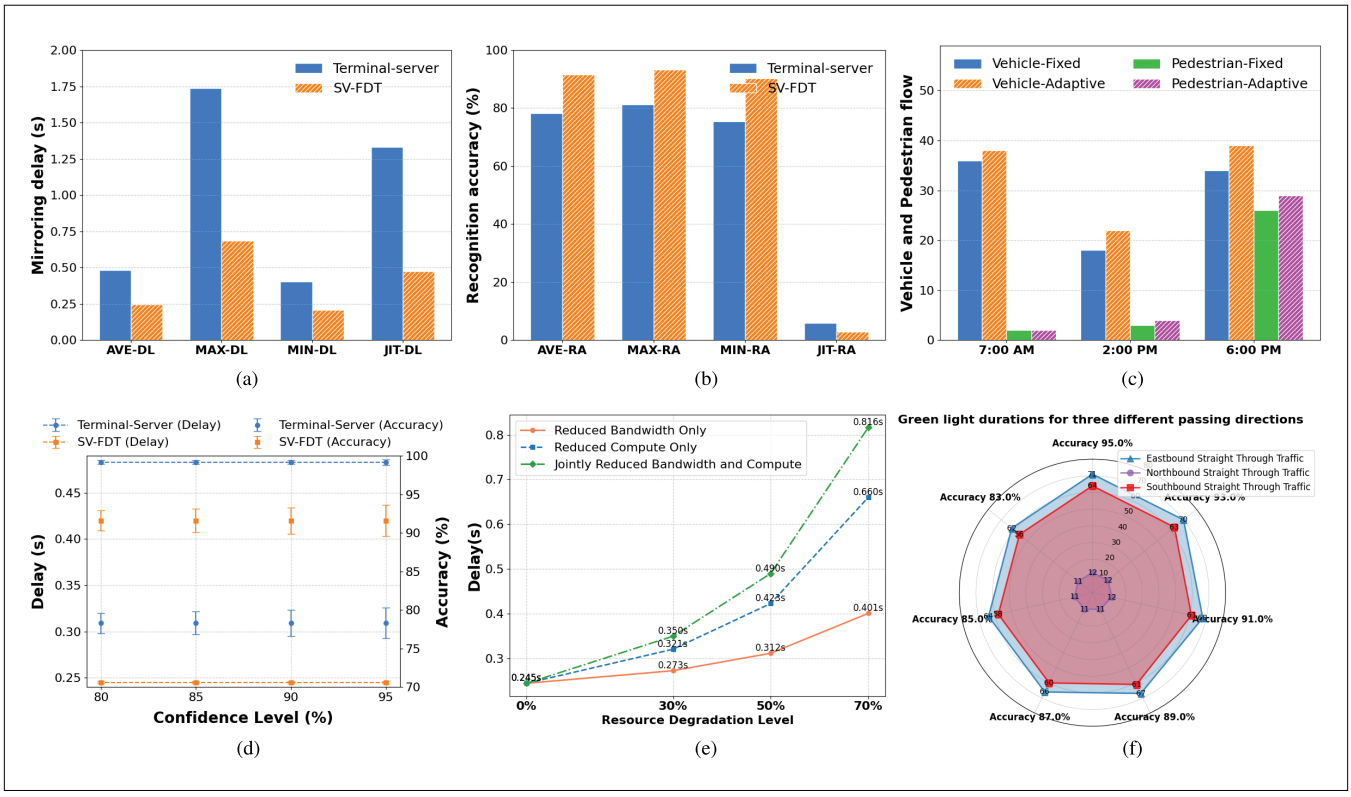


**FIGURE 5.** Comparisons on objective measurements. a) Mirroring delay comparison. b) Recoginition accuracy comparison. c) Traffic flow in different signal modes d) Delay intervals under varying confidence levels. e) Delay under different resource constraints. f) Signal timing vs. recognition accuracy.

As shown in Fig. 5(a), the SV-FDT framework outperforms the traditional terminal-server framework in DT modeling, achieving significantly lower average delay (AVE-DL), maximum delay (MAX-DL), minimum delay (MIN-DL), and jitter (JIT-DL). This improvement is attributed to two key reasons in the proposed SV-FDT: (i) data processing occurs at edge nodes closer to the data source; and (ii) only traffic codes and model parameters of twin agents, rather than raw surveillance videos, are transmitted to the cloud. Fig. 5(b) shows the recognition accuracy comparison, where SV-FDT also performs better, demonstrating higher average (AVE-RA), maximum (MAX-RA), and minimum recognition accuracy (MIN-RA), along with lower jitter (JIT-RA). This is because the semantic segmentation algorithm in SV-FDT is executed at edge nodes, which can continuously learn from local traffic environments, ensuring overall recognition accuracy. Fig. 5(c) illustrates traffic flow variations under fixed and adaptive signal modes at different times of the day. Under the adaptive signal mode, the average vehicle and pedestrian densities during peak hours are 1.8× and 10× higher than those during off-peak periods, and 1× and 10× higher than those during morning peak hours, respectively. Compared to the fixed signal mode, the adaptive mode effectively enhances traffic flow by dynamically adjusting the signal timings, significantly increasing vehicle and pedestrian throughput during peak hours. Fig. 5(d) presents the mirroring delay intervals at different confidence levels. At confidence levels above 80%, SV-FDT achieves a mirroring delay of approximately 0.24 seconds while maintaining recognition accuracy within the range of [0.90, 0.93]. Fig. 5(e) illustrates mirroring delay variations under three different resource constraint modes. The results show that the total mirroring delay remains below 0.5 seconds even when computational or bandwidth resources are degraded to 50%, individually or jointly. Fig. 5(f) shows the relationship between signal timing settings and recognition accuracy ratios. For simplicity, green light durations assigned to three different passing directions are selected as illustration samples. When recognition accuracy ratios are higher, signal timing configurations remain almost unchanged. In summary, the SV-FDT model outperforms the terminal-server framework in both delay and recognition accuracy, while demonstrating high robustness and stability. Compared with fixed traffic signal settings, SV-FDT achieves significant performance improvements by adopting adaptive traffic signal timing. Finally, subjective evaluations were conducted by 10 volunteers on the test platform. Table 1 shows that the terminal-server framework encountered issues related to DT modeling delay, inconsistencies, and asynchronous movement, whereas the SV-FDT framework operated seamlessly. These findings confirm that our framework provides an outstanding and immersive quality of experience (QoE) in both objective and subjective measurements.

| Framework | Video fluency | DT consistence | Sync. |
|---|---|---|---|
| Terminal-server | Choppy | Discrepant | Asynchronous |
| SV-FDT | Fluent | Consistent | Synchronous |

**TABLE 1.** Comparisons on subjective measurements.

## Conclusion and Future Research Directions

This article presents a novel cloud-edge-end collaborative framework, SV-FDT, designed to revolutionize ITSs by integrating pedestrians and vehicles in the loop. The proposed framework includes a comprehensive system architecture and highlights key design requirements and challenges for semantic segmentation-based FDT construction. We demonstrate several potential applications, such as an extreme pedestrian-vehicle flow testbed and emergency and disaster management. A case study conducted in CARLA simulation environments validates the effectiveness of SV-FDT in optimizing traffic management. Our results show that SV-FDT outperforms traditional terminal-server frameworks in terms of mirroring delay, recognition accuracy, and subjective evaluations. Furthermore, we identify and outline several open challenges.

- *High-Precision DT Extraction via Multimodal Fusion:* Extracting and fusing features from multimodal traffic data (e.g., images, audio, text, videos) enhances the accuracy of FDT models in heterogeneous ITS environments. While traditional methods rely on single-modal data, multimodal models require real-time data acquisition and distributed storage, posing privacy and security challenges. Future work will focus on building FDT models using knowledge graphs and deep learning to improve security and accuracy.
- *Seamless DT Synchronization Across Diverse Vehicular Communication Systems:* Vehicular communication devices vary in transmission rates, coverage, reliability, and interference. Real-time DT synchronization across these vehicular communication systems is challenging. Future efforts will leverage 6G technology for efficient wireless transmission and explore visible light communication (VLC), radio frequency identification (RFID), millimeter wave (mmWave), acoustic communication, and low Earth orbit (LEO) satellites to optimize DT synchronization.
- *Extending FDT to Future ITS With Hyper-Spatial Mobility:* As AI advances, future ITS will evolve into hyper-spatial systems supporting efficient and secure transportation across rail, subway, drone, and autonomous vehicle networks. Future research will investigate deep learning and intelligent control technologies to develop hyper-spatial FDT models for autonomous decision-making, route planning, state monitoring, performance evaluation, and fault prediction across multiple transport modes. In addition, we will explore integrating resource scheduling strategies, such as TD3 [15], into these systems to optimize computational resource allocation, improve data processing and communication efficiency, and ensure real-time performance with reduced latency in large-scale, multi-source environments.
- *Optimal Traffic Signal Control in Self-Evolving FDT Systems:* Errors in DT models can significantly impair traffic signal accuracy and decision-making, making optimal

FDT coordination highly complex. Future research will focus on developing mechanisms to detect and correct inaccuracies in traffic DT systems, applying filtering techniques, and leveraging federated learning and collaborative optimization to enhance data sharing and coordinated decision-making. For instance, in multi-intersection signal control, local SV-FDT agents will use reinforcement learning to adjust signal timings based on real-time pedestrian-vehicle interactions. Instead of transmitting raw data, only model updates are shared via federated learning, enabling the cloud to construct a global control policy. Distributed, multi-agent learning strategies will be also developed to adaptively respond to real-time traffic, improving both traffic flow and road safety. Additionally, we will explore adaptive learning mechanisms that allow SV-FDT to dynamically respond to diverse traffic scenarios. In future work, SV-FDT will be equipped with reinforcement learning capabilities and semantic feedback loops to enable real-time policy updates based on local observations of traffic patterns, weather conditions, pedestrian flow, and vehicle behaviors. These localized models will be aggregated via federated learning to refine global optimization strategies, allowing SV-FDT to adapt to varying conditions across different regions while preserving data privacy.

## Acknowledgment

This work was supported in part by the Major Program Project of Xiangjiang Laboratory under Grant 23XJ01001 and Grant 22XJ01001, in part by the Innovation Fund of QiYuan Laboratory under Grant 2022-JCJQ-LA-001-088, in part by the IoT Intelligent Sensing Support Project for Science and Technology Innovation Teams in Hunan Province, in part by the Key Research and Development Program of Hunan Province under Grant 2024JK2007, in part by the Hunan Provincial Natural Science Foundation of China under Grant 2023JJ40237, in part by the Guangxi Science and Technology Major Program under Grant 2024AA15007, and in part by the Postgraduate Research and Practice Innovation Program of Jiangsu Province under Grant KYCX25_0631.

## References

[1] L. U. Khan et al., "Federated learning for digital twin-based vehicular networks: Architecture and challenges," *IEEE Wireless Commun.*, vol. 31, no. 2, pp. 156–162, Apr. 2024.
[2] T. Yu et al., "Internet of federated digital twins: Connecting twins beyond borders for society 5.0," *IEEE Internet Things Mag.*, vol. 7, no. 5, pp. 64–71, Sep. 2024.
[3] L. Tang et al., "Digital twin-enabled efficient federated learning for collision warning in intelligent driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 3, pp. 2573–2585, Mar. 2024.
[4] B. Lu et al., "Cooperative perception aided digital twin model update and migration in mixed vehicular networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 26, no. 2, pp. 2293–2308, Feb. 2025.
[5] Z. Wang et al., "Mobility digital twin: Concept, architecture, case study, and future challenges," *IEEE Internet Things J.*, vol. 9, no. 18, pp. 17452–17467, Sep. 2022.
[6] Z. Wang et al., "Towards next generation of pedestrian and connected vehicle in-the-loop research: A digital twin co-simulation framework," *IEEE Trans. Intell. Vehicles*, vol. 8, no. 4, pp. 2674–2683, Apr. 2023.
[7] Y. Zhao et al., "GraCo: Granularity-controllable interactive segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2024, pp. 3501–3510.
[8] J. Chen et al., "A revolution of personalized healthcare: Enabling human digital twin with mobile AIGC," *IEEE Netw.*, vol. 38, no. 6, pp. 234–242, Nov. 2024, doi: 10.1109/MNET.2024.3366560
[9] Y. Yang et al., "Dynamic human digital twin deployment at the edge for task execution: A two-timescale accuracy-aware online optimization," *IEEE Trans. Mobile Comput.*, vol. 23, no. 12, pp. 12262–12279, Dec. 2024.
[10] J. Chen et al., "Networking architecture and key supporting technologies for human digital twin in personalized healthcare: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 26, no. 1, pp. 706–746, 1st Quart., 2024.
[11] L. Dong et al., "Deep progressive reinforcement learning-based flexible resource scheduling framework for IRS and UAV-assisted MEC system," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 2, pp. 2314–2326, Feb. 2025, doi: 10.1109/TNNLS.2023.3341067
[12] K. S. Kim et al., "Ultrareliable and low-latency communication techniques for tactile Internet services," *Proc. IEEE*, vol. 107, no. 2, pp. 376–393, Feb. 2019.
[13] Q. Xu, M. Chraibi, and A. Seyfried, "Anticipation in a velocity-based model for pedestrian dynamics," *Transp. Res. C, Emerg. Technol.*, vol. 133, Dec. 2021, Art. no. 103464.
[14] A. S. Akopov and L. A. Beklaryan, "Traffic improvement in Manhattan road networks with the use of parallel hybrid biobjective genetic algorithm," *IEEE Access*, vol. 12, pp. 19532–19552, 2024.
[15] A. Mohajer, J. Hajipour, and V. C. M. Leung, "Dynamic offloading in mobile edge computing with traffic-aware network slicing and adaptive TD3 strategy," *IEEE Commun. Lett.*, vol. 29, no. 1, pp. 95–99, Jan. 2025.

## Biographies

Xiaolong Li (Member, IEEE) (lxl@hutb.edu.cn) is currently a Professor with the Hunan University of Technology and Business, Changsha. His research interests include deep learning, intelligent transportation systems, and the Internet of Things.

Jianhao Wei (jianhao@hnu.edu.cn) is currently an Associate Professor with the Hunan University of Technology and Business, Changsha. His research interests include deep learning, intelligent transportation systems, and the Internet of Things.

Haidong Wang (whd@hutb.edu.cn) is currently a Lecturer with the Hunan University of Technology and Business, Changsha. His research interests include deep learning, intelligent transportation systems, and digital twins.

Li Dong (Dlj2017@hunnu.edu.cn) is currently an Associate Professor with the Hunan University of Technology and Business, Changsha. Her research interests include deep learning, intelligent transportation systems, and digital twins.

Ruoyang Chen (ruoyangchen@nuaa.edu.cn) is currently pursuing the Ph.D. degree with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, China. His research interests include edge computing and digital twin.

Changyan Yi (Senior Member, IEEE) (changyan.yi@nuaa.edu.cn) is currently a Professor with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, China. His research interests include edge computing, the Industrial IoT, digital twin, 5G, and beyond.

Jun Cai (Senior Member, IEEE) (Jun.cai@concordia.ca) is currently a Professor and the Perform Centre Research Chair with the Department of Electrical and Computer Engineering, Concordia University, Canada. His research interests include edge/fog computing and eHealth.

Dusit Niyato (Fellow, IEEE) (dniyato@ntu.edu.sg) is currently the President's Chair Professor with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. His research interests include edge intelligence, machine learning, and incentive mechanism design.

Xuemin (Sherman) Shen (Fellow, IEEE) (sshen@uwaterloo.ca) is currently a University Professor with the Department of Electrical and Computer Engineering, University of Waterloo, Canada. His research focuses on network resource management, wireless network security, social networks, and vehicular ad hoc networks. He is a Canadian Academy of Engineering Fellow, a Royal Society of Canada Fellow, and a Chinese Academy of Engineering Foreign Fellow.