IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

Establishing Reality-Virtuality Interconnections in Urban Digital Twins for Superior Intelligent Road Inspection and Simulation

Yikang Zhang, Chuang-Wei Liu, Jiahang Li, Yingbing Chen, Jie Cheng, Rui Fan

***Abstract*—Road inspection is crucial for maintaining road serviceability and ensuring traffic safety, as road defects gradually develop and compromise functionality. Traditional inspection methods, which rely on manual evaluations, are labor-intensive, costly, and time-consuming. While data-driven approaches are gaining traction, the scarcity and spatial sparsity of real-world road defects present significant challenges in acquiring high-quality datasets. Existing simulators designed to generate detailed synthetic driving scenes, however, lack models for road defects. Moreover, advanced driving tasks that involve interactions with road surfaces, such as planning and control in defective areas, remain underexplored. To address these limitations, we propose a multi-modal sensor platform integrated with an urban digital twin (UDT) system for intelligent road inspection. First, hierarchical road models are constructed from real-world driving data collected using vehicle-mounted sensors, resulting in highly detailed representations of road defect structures and surface elevations. Next, digital road twins are generated to create simulation environments for comprehensive analysis and evaluation of algorithm performance. These scenarios are then imported into a simulator to facilitate both data acquisition and physical simulation. Experimental results demonstrate that driving tasks, including perception and decision-making, benefit significantly from the high-fidelity road defect scenes generated by our system.**

***Index Terms*—Simulation and animation, robot safety, sensor fusion, and constrained motion planning.**

I. Introduction

ROAD condition assessment is essential for ensuring optimal vehicle dynamics and driving performance [1]. Defects such as cracks and potholes not only induce vibrations but also accelerate the wear of vehicle components [2]. Timely detection and repair of these defects are therefore crucial for ensuring traffic safety [3]. Nevertheless, current road inspection methods face significant challenges [4]. Traditional manual visual inspection, carried out by certified inspectors, is both labor-intensive and hazardous [5]. It also causes substantial disruption to traffic flow, making it impractical for large-scale road condition assessment [6]. In recent years, data-driven approaches have gained prominence. By fully leveraging extensive driving datasets, deep neural networks (DNNs) can achieve outstanding performance [7]. However, despite the unlimited data that can be collected, their effectiveness remains limited by the availability of high-quality ground-truth annotations [8].

Recent technological advancements have enabled the development of intelligent road inspection systems supported by urban digital twins (UDT). A UDT system reconstructs the semantic and geospatial properties of urban entities, such as roads and buildings [9], providing indispensable digital replicas for real-time monitoring and simulation. However, existing simulators typically represent roads as 2D planar surfaces, which cannot capture surface unevenness. As road surfaces directly interact with vehicles and significantly influence driving safety, it is necessary to reconstruct their 3D structures using real-world measurements.

To address these challenges, we develop a UDT system that enables reality-virtuality interconnection (RVI), specifically tailored for intelligent road inspection. The system not only creates digital twins of road entities but also builds simulation environments containing road defects to support the evaluation of perception and decision-making algorithms. For perception tasks such as semantic scene parsing [10] and 3D geometry reconstruction [11], it not only replicates real-world road environments but also generates novel scenarios by randomly combining road models to enhance generalization. Experiments show that semantic segmentation and stereo matching networks pre-trained on our synthetic data transfer effectively to real-world datasets. For decision-making tasks, we explore a new paradigm for handling road defects. Instead of following conventional methods that treat defects as obstacles to be avoided entirely, our approach enables either bypassing or gliding over certain defects without full detours, leveraging detailed road surface geometry from the UDT. By incorporating tire-level collision detection rather than vehicle-body bounding box checks, the system supports more flexible and efficient obstacle avoidance strategies.

The contributions of this study span five key aspects: equipment, algorithm, simulator, dataset, and benchmark. As illustrated in Fig. 1, we develop a portable, multi-sensor ex-

Manuscript received May 4, 2025; Revised: August 19, 2025; Accepted: November 23, 2025. This paper was recommended for publication by Editor Hyungpil Moon upon evaluation of the Associate Editor and Reviewers' comments. This work was supported by the National Natural Science Foundation of China under Grant 62473288, the Fundamental Research Funds for the Central Universities, and the Xiaomi Young Talents Program. (*Corresponding author: Rui Fan*)

Yikang Zhang, Chuang-Wei Liu, Jiahang Li, and Rui Fan are with the College of Electronic and Information Engineering, Shanghai Institute of Intelligent Science and Technology, Shanghai Research Institute for Intelligent Autonomous Systems, Shanghai Key Laboratory of Intelligent Autonomous Systems, State Key Laboratory of Autonomous Intelligent Unmanned Systems, and Frontiers Science Center for Intelligent Autonomous Systems (Ministry of Education), Tongji University, Shanghai 201804, China. (email: yikangzhang, cwliu, lijiahang617, rfan}@tongji.edu.cn)

Yingbing Chen and Jie Cheng are with the Department of Electronic and Computer Engineering, the Hong Kong University of Science and Technology, Hong Kong SAR. (email: {ychengz, jchengai}@connect.ust.hk)

Digital Object Identifier (DOI): xxxxx.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

perimental setup, equipped with a Livox Mid-360 LiDAR, two FLIR BFS-U3-31S4C cameras, and a DETA100D4G GNSS RTK module. This setup can be mounted on any vehicle, enabling multi-modal, high-quality road data collection in real-world scenarios. To model the interconnection between reality and virtuality for urban entities, we develop two key modules: a hierarchical road model creator and a digital road twin generator. The former module extracts road defect structures and undamaged surface elevations from real-world sensor data. Real-world road entities across different scales, from coarse-grained road surfaces that shape the overall driving environment to fine-grained defects that directly affect vehicle stability and comfort, are segmented and integrated into unified road surface mesh models, replacing the simplified planar road assets for more realistic road scenes. The latter module creates high-fidelity scenes for synthetic data collection and physical simulation. Given the challenges inherent in large-scale road surfaces and sporadically occurring road defects in the real world, our system enables seamless integration of any defect model with any surface model smoothly, even if the models originate from different locations. For perception tasks, we create a comprehensive dataset containing semantic and instance-level annotations, along with ground truth data for depth, event, and surface normals, to benchmark state-of-the-art (SoTA) DNNs for intelligent road inspection. For decision-making tasks, we design an experimental framework that treats road defects as negative obstacles, enabling the evaluation of diverse trajectory planning and speed control algorithms. The framework provides physical simulation for both defect-avoidance and traversal strategies that optimize driving comfort, incorporating wheel-level collision constraints for more precise decisions.

In a nutshell, our main contributions are as follows:

- We design a UDT system for road inspection, including a portable multi-modal sensor equipment for real-world data collection, and simulation environments containing road defects for data synthesis and physical simulation.
- We propose a pipeline to autonomously construct hierarchical road models and generate digital road twins, which both replicate high-fidelity real-world environments and enable the creation of novel simulation scenarios.
- We provide a comprehensive benchmark to evaluate perception and decision-making algorithms leveraging our urban digital twins, demonstrating their effectiveness for downstream driving tasks.

II. Related Works

A. Intelligent Road Inspection

Traditional road inspection is typically conducted by structural engineers or certified inspectors, a process that is often subjective, inefficient, and sometimes hazardous [4]. To address these challenges, researchers have developed intelligent road inspection systems capable of automatically collecting data and identifying road defects [5]. Among the most direct effects of road defects is vibration. Measurement systems analyze vehicle dynamics as the vehicle traverses road defects [12]. While these systems provide high precision, they are limited to detecting defects within the tire tracks. Other approaches mount visual sensors on the rear of the vehicle with a downward-facing view, maintaining a known distance from the road surface, including single cameras [13], stereo camera pairs [14], and laser scanners [15]. However, such specialized sensor configurations are restricted to dedicated inspection vehicles, limiting scalability and widespread applicability. In contrast, crowd-sourcing approaches collect large-scale data from daily driving using forward-facing sensors, enabling dual functionality for both driving perception and road inspection. For instance, LiDAR-based methods collect data for road unevenness [16], while stereo camera-based approaches reconstruct 3D road surfaces [17]. Smartphone applications have also emerged as a practical data source.

Numerous algorithms have been developed for road defect detection. Classical 2D image processing methods have been widely explored, extracting damaged road areas from segmented foregrounds based on geometric and textural assumptions [2]. However, such methods are sensitive to environmental factors that violate the underlying assumptions. 3D point cloud methods have also been employed to detect road irregularities. Stereo vision-based approaches reconstruct dense 3D road point clouds and interpolate them into planar or quadratic surfaces. LiDAR-based methods are utilized for road roughness perception [18], while road patches traversed by a vehicle's tires are segmented to identify irregularities such as bumps and potholes [19]. With the rise of deep learning, data-driven approaches have become the dominant techniques [20]. However, the performance heavily depends on the dataset quality. Due to the rarity and sparse distribution of road defects, capturing sufficient annotated data for training remains a significant challenge.

B. Urban Digital Twins

Digital twin, originally developed for cyber-physical integration in factories, connects real and virtual products through advanced data communication technologies [21]. Leveraging advantages such as real-time monitoring, fast simulation, and troubleshooting, digital twin technology applied to urban entities integrates online sensor data and creates virtual models to address road anomalies [22]. These capabilities are exemplified by applications such as intelligent road inspection. For example, [23] proposed a pavement crack segmentation approach based on 3D edge detection within digital twins. Recent studies have also achieved large-scale reconstruction of explicit road surface meshes, recovering both geometry and texture [24], [25]. However, most research treats digital twins of road defects and surfaces as individual reconstruction tasks. In contrast, our work integrates defective road entities at the scene level, creating comprehensive simulation environments specifically tailored for autonomous driving applications.

III. Urban Digital Twin System

A. Hierarchical Road Model Creator

The hierarchical road model creator consists of two reconstruction streams, one for road defects and the other for non-defective surfaces, as illustrated in Fig. 2. The coarse-grained

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

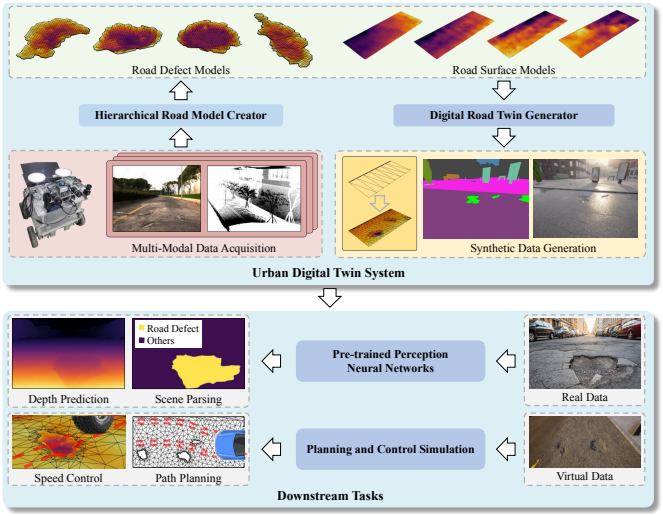

Fig. 1. The UDT system is composed of two components. The hierarchical road model creator autonomously reconstructs 3D defect models and non-defective road surfaces from the physical world, while the digital road twin generator produces virtual entities for a diverse, well-annotated environment.

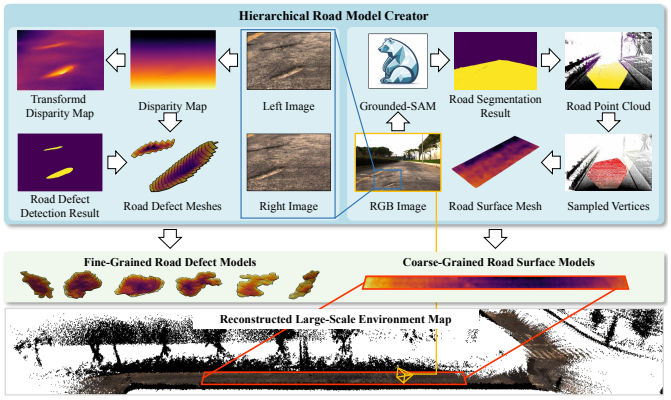

Fig. 2. Our hierarchical road model creator consists of two reconstruction streams. In the coarse-grained stream, semantic annotations are generated by Grounded-SAM, which filters LiDAR points for road surface mesh generation. The fine-grained stream reconstructs road defect models using transformed disparity maps derived from stereo images. The reconstructed results are stored in a model library for future use.

stream, focused on the reconstruction of non-defective road surfaces, follows a straightforward process. Semantic masks are generated from camera images using Grounded-SAM [26], with prompts such as "road" or "pavement". LiDAR points are segmented based on these masks and fused across multiple frames. To achieve coordinate alignment, a horizontal ground plane is fitted to points from multiple road segments, assuming local elevation variations while maintaining overall flatness. A 3D downsampling operation is then applied to refine the vertices before meshing, accounting for density variations in the point clouds caused by redundant or missing observation views. Given the simple geometry of non-defective road surfaces, elevation can be decoupled from the planar coordinates. The mesh model is reconstructed by directly triangulating the 2D projection of vertices onto the horizontal plane. Vertex elevations are subsequently restored from the original point cloud, ensuring that the reconstructed surface accurately reflects real-world topography.

The stream for fine-grained road defects, unlike the non-defective road surfaces, requires dense geometric details. Although many neural networks are proposed for road defect detection from monocular images, semantic cues at the boundaries between defective and non-defective surfaces are often ambiguous, leading to artifacts such as isolated pixels or incomplete segmentation. Therefore, defect regions are extracted based on geometric cues, performed by transformed disparity maps derived from stereo images, which highlight spatial deviations from the planar patches of the road surface. Road defect instances are identified as connected components within the parsed scene. Instances with regular shapes, as shown in Fig. 2, can be reconstructed directly through grid-based sampling, similar to the first stream. However, defects with irregular shapes, as illustrated in Fig. 3, may introduce structural discontinuities, even when pixel connectivity is preserved. To address this issue, the sampled point set is expanded iteratively until the entire defect region is covered. Expanded points, belonging to the defect but lying on the non-defective region, are shared between adjacent meshes. If reconstructed separately, these duplicate boundary points could cause elevation mismatches, resulting in gaps or overlaps. Therefore, the boundary of the defect mesh is extracted and assigned a pseudo-height, which becomes active only when assembled with neighboring road surface meshes. A boundary edge is defined as an edge connected to only one face in the mesh, indicating that it forms the outer boundary or "hole" of the surface. All faces in the road defect model are traversed to identify such edges. Since boundary points are guaranteed to lie on the non-defective road surface, they can be used to fit a transformation matrix that aligns the defect model with the ground plane, ensuring that the pseudo-height is set to zero.

The outputs from the two reconstruction streams are compiled into a road model library. This design, with varying levels of granularity, ensures that the road meshes remain memory-efficient while retaining sufficient details.

B. Digital Road Twin Generator

Although road scenes can be accurately replicated, road defects are relatively rare in the physical world. To generate more diverse scenes, the over-simplified road models provided by simulators must be updated with our road defect models, as illustrated in Fig. 4. First, the original planar road segment assets are exported from the simulator to maintain the scene's road topology. The generated road defect models are then sampled with random poses and scales for integration. When merging the mesh models, intersecting triangular faces must be removed. However, because road segments are not always convex, simply removing all faces and re-meshing could compromise their structural integrity. Therefore, defect models are projected vertically onto the road surface to accurately identify and remove only the intersecting triangles. The Möller-Trumbore algorithm [27], typically used to detect such ray-

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

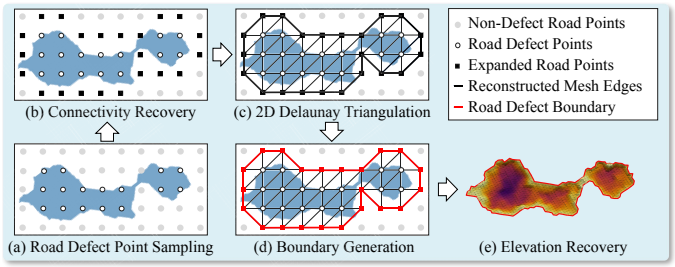

Fig. 3. Example of the road defect reconstruction process. Points are iteratively sampled until the entire road defect structure is covered, ensuring topological connectivity. After triangulation, the boundary vertices are stored in a list with pseudo-heights for further structural alignment in the digital road twin generator.

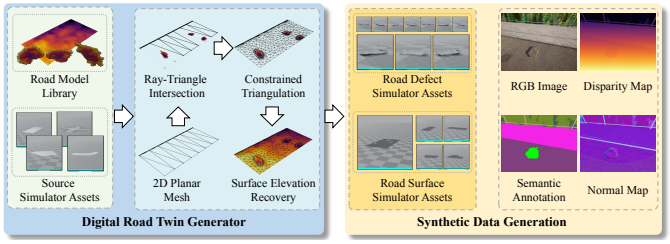

Fig. 4. The Digital Road Twin Generator integrates simulator assets with reconstructed road models. Road defect models are projected onto the 2D road mesh to eliminate intersecting faces. The 2D structure is reorganized using constrained Delaunay triangulation, preserving model boundaries. For compatibility with the simulator, the generator's output is disassembled into individual road surface and defect asset groups, providing semantic references for the rendering pipeline.

triangle intersections, is expressed as follows:

$$\boldsymbol{r}(t) = \boldsymbol{o} + t\boldsymbol{d} = (1-u-v)\boldsymbol{p_0} + u\boldsymbol{p_1} + v\boldsymbol{p_2}, \tag{1}$$

$$s.t. \begin{cases} 0 \le u \le 1, \\ 0 \le v \le 1, \\ 0 \le 1-u-v \le 1, \end{cases} \tag{2}$$

where ray $\boldsymbol{r}$ originates from $\boldsymbol{o}$ in direction $\boldsymbol{d}$ with length t, and the triangle is determined by points $\boldsymbol{p_0}$, $\boldsymbol{p_1}$, and $\boldsymbol{p_2}$. Rewriting (1) leads to the following expression:

$$\boldsymbol{o} - \boldsymbol{p_0} = (\boldsymbol{p_1} - \boldsymbol{p_0})u + (\boldsymbol{p_2} - \boldsymbol{p_0})v - t\boldsymbol{d}, \tag{3}$$

which can be solved using Cramer's rule, where

$$t = \frac{(\boldsymbol{o} - \boldsymbol{p_0}) \times (\boldsymbol{p_1} - \boldsymbol{p_0}) \cdot (\boldsymbol{p_2} - \boldsymbol{p_0})}{(\boldsymbol{o} - \boldsymbol{p_0}) \times (\boldsymbol{p_2} - \boldsymbol{p_0}) \cdot (\boldsymbol{p_1} - \boldsymbol{p_0})}, \tag{4}$$

with similar solutions for u and v. With all intersecting faces removed, the road defect models are placed into the corresponding holes of the planar road mesh. The planar road polygons are then reorganized, typically through triangulation, which iteratively divides edges, samples vertices, and generates faces. However, triangle mismatches may occur at the boundaries if adjacent meshes are reconstructed separately. For improved realism and diversity, real-world road surface elevations are restored from the library. The height of each vertex is efficiently queried through a k-d tree built from the surface model. The final integrated 3D mesh is then split into individual assets, as required by the simulator, to provide semantic references.

Leveraging the hierarchical road model creator and the digital road twin generator described above, our UDT system bridges reality-virtuality interconnections to generate diverse, well-annotated simulation scenes featuring road defects. The virtual environment not only provides synthetic data to enhance perception networks that depend on comprehensive road surface data, but also enables physical simulation for decision-making tasks that interact directly with road surfaces. In the following section, we demonstrate these advancements by presenting perception tasks trained on synthetic data and decision-making tasks performed within our virtual environment.

IV. Experimental Results

A. Experimantal Setups

The sensor setup (Fig. 1) is mounted on a remotely controlled vehicle for continuous data collection over campus roads. Images and LiDAR point clouds are captured at 10Hz, while coarse GNSS coordinates are recorded to enable defect localization and support large-scale maintenance in the future. The GNSS module also outputs signals for hardware-level synchronization. The sensor poses are estimated using FAST-LIO [28], which builds and updates a coarse-grained, large-scale 3D road point cloud. The points belonging to drivable areas are filtered by Grounded-SAM, while stereo matching is applied to generate dense 3D reconstructions of detected defects [29]. Based on sensor data collected from real-world defective roads, our hierarchical road model creator generates a library containing 53 distinct road defect models. These models are integrated into scenes in the simulator, replacing the original planar road surfaces. For example, roads in CARLA Town01 are replaced by 646 synthetic road segments and 1,000 defect instances using our digital road twin generator. In the simulation, an autopiloted vehicle navigates through road defect scenes populated with randomly placed surrounding vehicles and pedestrians. To improve data variability and mitigate the sim-to-real gap, we introduce controlled randomness into the data acquisition pipeline. Sensor poses are sampled with randomized viewing angles rather than fixed trajectories, reducing view-dependent bias. Road surfaces and defects are rendered with diverse materials, including asphalt, cobblestones, and cement, while dynamic weather conditions are continuously simulated to further enhance data variability. To demonstrate the effectiveness of our synthetic data, we selected models from MMSegmentation[1] with diverse backbones to ensure architectural diversity. For stereo matching, we employed representative state-of-the-art general-purpose models to provide a fair and rigorous baseline.

B. Perception Tasks

1) Semantic Segmentation: We first evaluate single-modal semantic segmentation networks using real-world RGB images. UDTIRI [9], a real-world road defect dataset containing

[1]MMSegmentation: https://github.com/open-mmlab/mmsegmentation

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

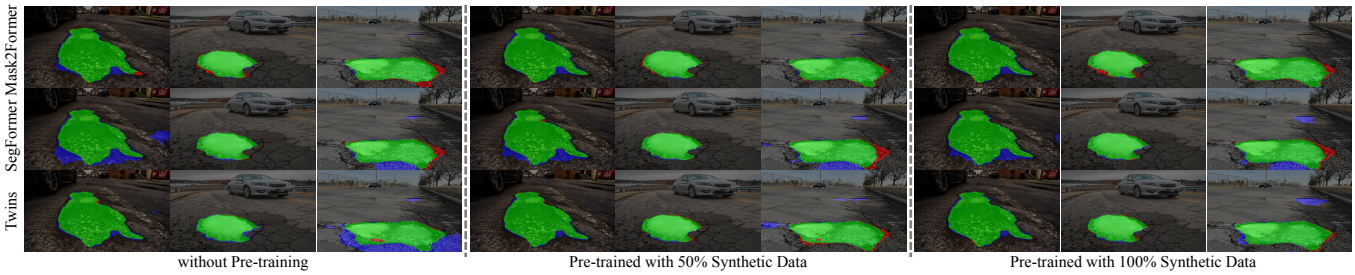

Fig. 5. Qualitative experimental results of semantic segmentation based on RGB images. The green areas in the image represent true-positive predictions, the blue areas represent false-positive predictions, and the red areas represent false-negative predictions.

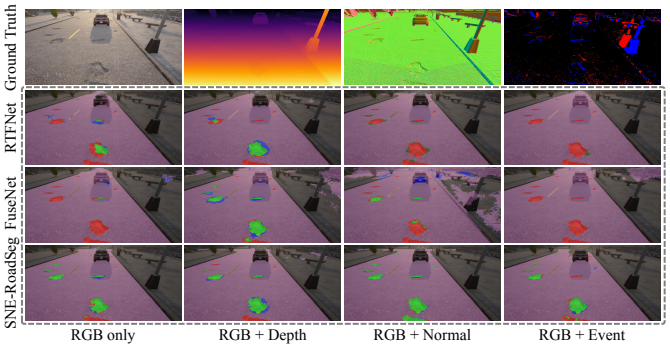

Fig. 6. Qualitative comparison between SoTA multi-modal networks on the synthetic data. The predicted drivable area is represented in purple. True-positive, false-positive, and false-negative classifications of road defects are shown in green, blue, and red, respectively.

TABLE I
MIOU(%) ↑ OF SINGLE-MODAL SEMANTIC SEGMENTATION RESULTS WITH DIFFERENT RATIOS OF SYNTHETIC DATA

Method	Synthetic/Real data ratio		
	0%	50%	100 %
Mask2Former [30]	77.08	79.79	**80.07**
PSPNet [31]	71.42	71.44	**74.29**
Twins [32]	78.40	**79.10**	79.06

1,000 semantically annotated images, is used for evaluation. All networks are pre-trained for 40,000 epochs by controlled ratios of synthetic data, followed by fine-tuning on 600 images from the real-world dataset. Network configurations follow the default settings of the MMSegmentation benchmark. To ensure fair comparisons, all hyperparameters are kept identical across experiments, with the only difference being the data for pre-training. Qualitative and quantitative results are presented in Fig. 5 and Table I, respectively. It is evident that pre-training with our synthetic data significantly enhances segmentation accuracy, providing more precise boundaries and reducing the occurrence of false-positive regions. The results show that synthetic data positively impacts image-based road defect detection, increasing mIoU by up to 3%. This improvement alleviates the challenge posed by the limited availability of annotated road defect data for real-world inspection models.

Single-modal prediction results are often vulnerable to varying weather and illumination. In contrast, multi-modal fusion is widely regarded as a solution to enhance perception

TABLE II
MIOU(%) ↑ OF MULTI-MODAL SEMANTIC SEGMENTATION RESULTS.

Method	RGB only	RGB+Depth	RGB+Normal	RGB+Event
RTFNet [33]	67.5	74.6	69.2	63.6
FuseNet [34]	66.6	78.7	53.9	58.5
SNE-RoadSeg [35]	67.4	78.8	94.5	73.5

robustness by utilizing auxiliary information sources. However, capturing these modalities remains challenging. Leveraging the G-buffer in the rendering pipeline, we are able to evaluate auxiliary modalities for road inspection, focusing on image segmentation for drivable areas, road defects, and backgrounds, as shown in Fig. 6. As shown in Table II, each additional modality is fused with RGB images in an RGB+X format. RGB+Depth provides the most generalized results, outperforming those trained using only RGB images by 4.5% to 12.1%. It also generates the most accurate true-positive predictions, demonstrating robust performance across all cases. This result is intuitive, as depth images directly capture the surface structure. Normal maps, which are highly sensitive to surface geometry, improve performance by 1.7% to 27.1% across most networks, except for FuseNet. Although useful in many cases, this sensitivity also introduces noise, leading to convergence issues. Simple fusion of event images across network channels is unstable. Although event images improve SNE-RoadSeg by 6.1%, they degrade FuseNet's performance by 8.1%. This instability is likely due to event data being influenced by the vehicle's driving speed, whereas RGB images assume static objects during exposure. To effectively utilize this modality, additional speed priors are required.

2) Stereo Matching: For stereo matching evaluation, we compare five SoTA networks to demonstrate the improvements enabled by our system. Similar to the segmentation experiments, all hyperparameters are kept fixed according to the released default settings, with the only difference being whether the models are pre-trained on synthetic UDT data. Unlike individual road surface images that are available from crowd-sourcing applications, stereo images of road surface defects are exceedingly rare. To address this limitation, we fine-tune the networks using the KITTI dataset [41], a widely used driving dataset that does not contain road defects. Experiments are conducted on both synthetic data with ground truth annotations and the real-world Stereo-Road dataset [42]

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

TABLE III
COMPARISONS OF SOTA STEREO MATCHING NETWORKS WITH AND WITHOUT OUR PROPOSED SYNTHETIC DATA.

Method	Evaluation on Synthetic UDT Data						Evaluation on Stereo-Road Dataset (Zero-shot)		
	PEP(%) ↓		EPE(pixel) ↓	SSIM ↑	MSE ↓	PSNR ↑	SSIM* ↑	MSE* ↓	PSNR* ↑
	$\delta = 0.5$	$\delta = 1$							
PSMNet [36]	53.50	22.00	2.57	0.82	152.2	28.3	0.84	129.0	28.3
PSMNet+synthetic	**7.61**	**3.59**	**0.67**	**0.90**	**93.5**	**30.9**	**0.92**	**64.7**	**31.1**
AANet [37]	37.60	23.7	4.79	0.85	118.2	29.7	0.77	198.7	25.9
AANet+synthetic	**9.36**	**5.30**	**0.61**	**0.91**	**93.4**	**30.8**	**0.93**	**67.7**	**30.7**
LacGwc [38]	15.00	8.17	1.35	0.89	113.4	30.4	0.92	75.8	30.3
LacGwc+synthetic	**6.11**	**2.92**	**0.55**	**0.90**	**96.0**	**30.7**	**0.93**	**63.3**	**31.1**
IGEV [39]	8.14	4.99	0.83	0.90	102.9	30.8	0.93	64.2	31.1
IGEV+synthetic	**3.17**	**1.59**	**0.23**	**0.91**	**92.5**	**30.9**	**0.93**	**60.7**	**31.3**
ViTAStereo [40]	7.61	5.08	1.31	0.90	124.4	29.6	0.92	79.6	29.9
ViTAStereo+synthetic	**3.14**	**1.61**	**0.22**	**0.91**	**91.6**	**30.9**	**0.93**	**60.0**	**31.3**

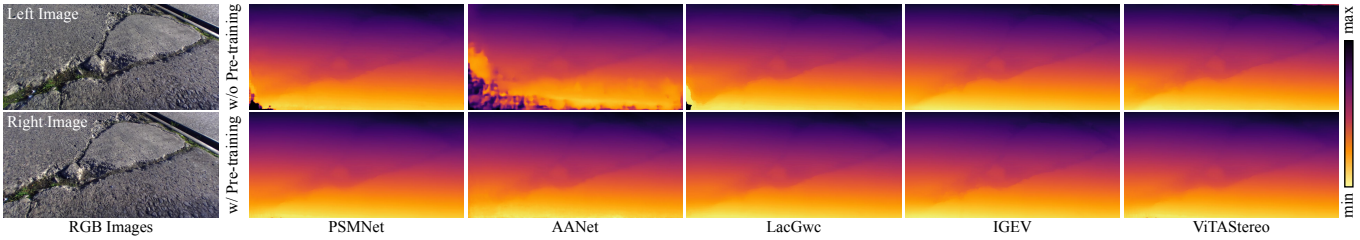

Fig. 7. Depth visualizations for stereo matching on real-world images, comparing models with and without pre-training on synthetic road defect data.

to evaluate zero-shot performance. Several metrics are selected for evaluation, including percentage of error pixels (PEP), which represents the percentage of incorrect disparities with respect to a tolerance of δ pixels, and end-point error (EPE), which measures the average disparity estimation error. Additionally, mean squared error (MSE), structural similarity index (SSIM), and peak signal-to-noise ratio (PSNR) are compared between the predicted and ground truth disparity maps. The loss function, learning rate, and optimizer settings are consistent with those reported in the corresponding publications. Table III presents the quantitative results tested on synthetic data. Nearly all networks pre-trained with our synthetic data demonstrate performance improvements. EPE is improved across all networks, with reductions ranging from 0.60 to 3.18. These results confirm that our UDT system significantly enhances model performance for stereo matching tasks.

C. Decision-Making Tasks

Two types of experiments are conducted: one for 2D path planning, where road defects are avoidable, and the other for speed control, where the vehicle must traverse defects.

1) Path Planning: In scenarios with ample road space and minimal opposing traffic, avoiding defects by following an alternative path is generally the preferred strategy. Unlike traditional obstacle constraints, road defects primarily impact the wheels rather than the vehicle body, allowing the vehicle to either bypass or glide over the defects. As shown in Fig. 8(a), multiple defects are distributed along the road. The drivable area and defective regions are modeled as grid maps with a resolution of 0.2 m, and the vehicle model assumes a width of 2.0 m and a rear-to-front wheelbase of 3.0 m. Four widely used planning algorithms, A*, RRT*, Lattice, and Hybrid A*, are compared. Evaluation metrics follow those from classical obstacle avoidance tasks. Path deviation quantifies the length difference between the planned path and the shortest possible path, typically the road's midline. Path smoothness is calculated by averaging the angle changes along the trajectory, while obstacle clearance measures the mean distance between the vehicle's wheels and the nearest road defects. Quantitative and qualitative results of the generated paths are presented in Table IV and Fig. 8(a). Notably, A* and RRT* do not account for vehicle kinodynamic constraints, resulting in relatively poor path smoothness. However, their discrete path points remain topologically reasonable. The paths generated by Hybrid A* and Lattice glide over defects P_0 and P_1, which are justified in the given context, respectively. Lattice generates the shortest and smoothest path, with a path deviation of 0.87 and a smoothness value of 0.047. However, because its grids and curves are pre-sampled, obstacles may not always be perfectly avoided. In contrast, Hybrid A* expands nodes iteratively at each step, achieving an obstacle clearance of 3.74, which ensures better safety distances. The trade-off, however, is that heuristic terms must be manually designed to suit specific scenarios.

2) Speed Control: In scenarios where lane changes pose a high risk, the vehicle must traverse road defects directly. In such cases, longitudinal speed becomes the only controllable variable to minimize vibrations. We customize these scenarios using our UDT system, where real-world road surface models are cropped into 40-meter sections and integrated with randomly distributed road defects. With the aid of simulators, road defects along the wheel trajectories can be precisely lo-

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

TABLE IV
COMPARISON OF PATH PLANNING ALGORITHMS TO AVOID ROAD DEFECTS

Planner	path dev.(%)	path smoothness(rad/m)	obstacle clearance(m)
A*	7.46	0.294	3.69
RRT*	21.2	0.402	3.72
Lattice	**0.87**	**0.047**	2.62
Hybrid A*	1.05	0.054	**3.74**

cated, enabling informed speed control strategies. To measure vehicle vibrations caused by defective road surfaces, collision detection is performed at the level of individual triangular faces between the road and tires during physical simulation. As shown in Fig. 8(b), six-axis IMU data are recorded as the vehicle drives at a constant speed. During the process of a wheel falling into and exiting a road defect, it can be observed that the most affected parameters are vertical acceleration a_z, rotational velocity in roll ω_x and pitch ω_y. Subsequently, the vibration degree is defined as $g = \sqrt{{\omega_x'}^2 + {\omega_y'}^2 + \alpha {a_z'}^2}$, where α is set to 0.1 in this scenario. Quantitative results for traversing three road defects at different speeds are presented in Fig. 8(c). The results indicate that, for large and severe defects such as P_2, slowing down is the most effective strategy for driving comfort. Interestingly, for small and shallow defects, such as P_1, increasing speed can also reduce vibrations. Although initially counterintuitive, this observation aligns with practical experience upon closer analysis.

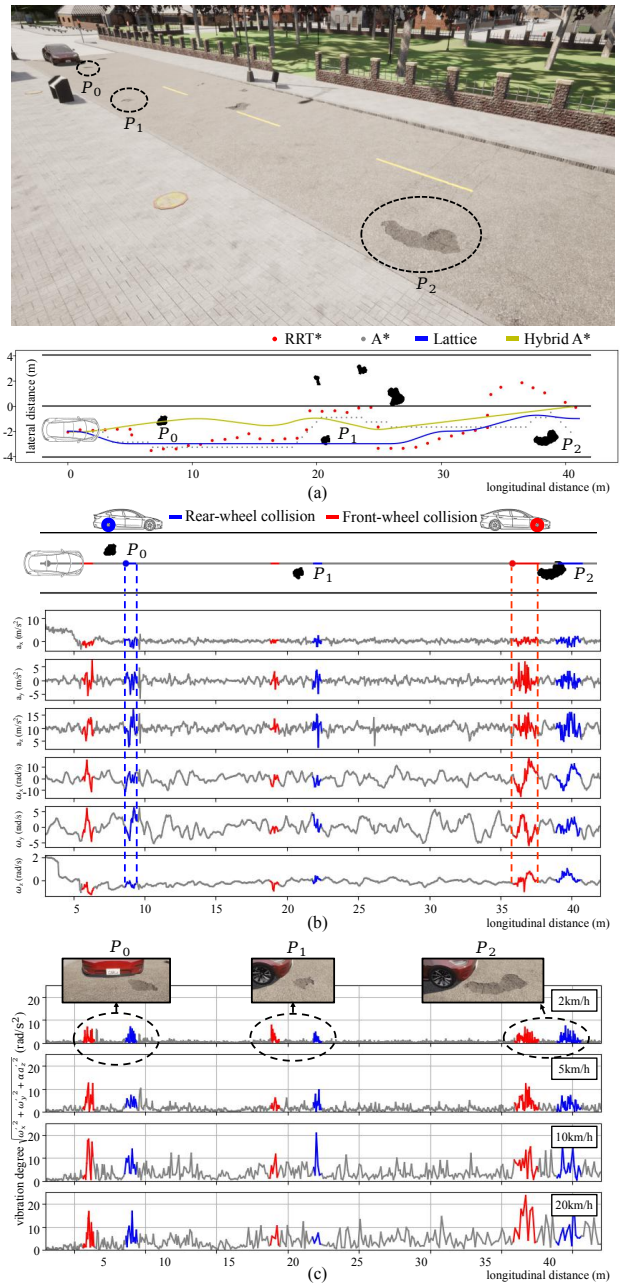

Fig. 8. Experiments on planning and control tasks using our UDT system. (a) Path planning for road defect avoidance, focusing on tire-ground collision constraints instead of the vehicle body. (b) Inertial measurements recorded while traversing road defects at a constant speed of 5 km/h. (c) Vibration degree evaluated over road defects with different traversing speeds.

V. DISCUSSION AND FUTURE WORK

Despite these advancements, several challenges remain. First, the defect taxonomy is not yet comprehensive. The current library primarily models geometric surface deformations such as potholes and depressions, whereas finer-scale surface distresses, including cracking, raveling, and bleeding, are not yet reconstructed. Achieving a more complete taxonomy will require larger-scale data acquisition and systematic defect surveying in future deployments. Second, loss of geometric and texture fidelity is inevitable during the reconstruction process, leading to domain discrepancy between reconstructed virtual environments and real-world observations. As a result, synthetic data is primarily limited to pre-training and must be supplemented with real-world images for fine-tuning. Future work could explore domain adaptation techniques and advanced rendering methods to reduce the sim-to-real gap. Another promising direction is to adopt novel view synthesis (NVS), which learns neural scene representations directly from multi-view imagery without explicit geometric reconstruction. Unlike traditional geometry pipelines, NVS has the potential to mitigate fidelity loss during reconstruction and deliver more photorealistic renderings for downstream tasks. Another limitation concerns deployment. Due to constraints of our single prototype and the campus-scale environment, we have not yet explored large-scale, in-the-wild deployment scenarios that are essential for real-world road inspection. In the future, the generated synthetic data can be exploited for model fine-tuning and knowledge distillation, thereby enabling the development of lightweight models capable of real-time operation in large-scale road monitoring applications.

VI. CONCLUSION

This letter presented a UDT system that exploits reality–virtual interconnections of road entities for intelligent road inspection. The system captures data from defective roads using a custom sensor setup, constructs detailed defect and surface models through a hierarchical road model creator, and integrates them into simulation scenes via a digital road

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

twin generator to recover simplified planar road assets. The proposed system provides unlimited, multi-modal, and well-annotated data for training road inspection networks, as well as a physical simulation environment for testing autonomous driving tasks under defective road conditions. Experiments showed that pre-training semantic segmentation networks on our synthetic data substantially improves performance, alleviating the shortage of annotated real-world defect data. The precise modeling of road defects enables both trajectory planning to bypass defects and speed control to traverse them smoothly, balancing safety and comfort. Finally, we discussed current limitations in addressing the sim-to-real gap and challenges in in-the-wild deployment, outlining future directions for improved realism and large-scale applications.

REFERENCES

[1] H. Guo *et al.*, "A review of estimation for vehicle tire-road interactions toward automated driving," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 49, no. 1, pp. 14–30, 2018.

[2] R. Fan *et al.*, "Rethinking road surface 3-D reconstruction and pothole detection: From perspective transformation to disparity map segmentation," *IEEE Transactions on Cybernetics*, vol. 52, no. 7, pp. 5799–5808, 2021.

[3] R. Fan and M. Liu, "Road damage detection based on unsupervised disparity map segmentation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 11, pp. 4906–4911, 2019.

[4] T. Kim and S.-K. Ryu, "Review and analysis of pothole detection methods," *Journal of Emerging Trends in Computing and Information Sciences*, vol. 5, no. 8, pp. 603–608, 2014.

[5] N. Ma *et al.*, "Computer vision for road imaging and pothole detection: a state-of-the-art review of systems and algorithms," *Transportation Safety and Environment*, vol. 4, no. 4, p. tdac026, 2022.

[6] A. Wang *et al.*, "The two-step method of pavement pothole and raveling detection and segmentation based on deep learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 6, pp. 5402–5417, 2024.

[7] J. Zhao *et al.*, "DRMNet: A multi-task detection model based on image processing for autonomous driving scenarios," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 12, pp. 15 341–15 355, 2023.

[8] T. Yin *et al.*, "Promoting automatic detection of road damage: A high-resolution dataset, a new approach, and a new evaluation criterion," *IEEE Transactions on Automation Science and Engineering*, vol. 22, pp. 2472–2484, 2024.

[9] S. Guo *et al.*, "UDTIRI: An online open-source intelligent road inspection benchmark suite," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 8, pp. 9920–9931, 2024.

[10] J. Li *et al.*, "RoadFormer: Duplex Transformer for RGB-normal semantic road scene parsing," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 7, pp. 5163–5172, 2024.

[11] C.-W. Liu *et al.*, "These maps are made by propagation: Adapting deep stereo networks to road scenarios with decisive disparity diffusion," *IEEE Transactions on Image Processing*, vol. 34, pp. 1516–1528, 2025.

[12] T. Dózsa *et al.*, "Road abnormality detection using piezoresistive force sensors and adaptive signal models," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–11, 2022.

[13] P. Prasanna *et al.*, "Automated crack detection on concrete bridges," *IEEE Transactions on Automation Science and Engineering*, vol. 13, no. 2, pp. 591–599, 2014.

[14] H. Karunasekera *et al.*, "Energy minimization approach for negative obstacle region detection," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 12, pp. 11 668–11 678, 2019.

[15] J. Han *et al.*, "Enhanced road boundary and obstacle detection using a downward-looking LIDAR sensor," *IEEE Transactions on Vehicular Technology*, vol. 61, no. 3, pp. 971–985, 2012.

[16] T. Zhao *et al.*, "A hierarchical scheme of road unevenness perception with LiDAR for autonomous driving comfort," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 1, pp. 2439–2448, 2023.

[17] F. Oniga and S. Nedevschi, "Processing dense stereo data using elevation maps: Road surface, traffic isle, and obstacle detection," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 3, pp. 1172–1182, 2009.

[18] T. Zhao *et al.*, "RoadBEV: Road surface reconstruction in bird's eye view," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 11, pp. 19 088–19 099, 2024.

[19] T. Zhao *et al.*, "A road surface reconstruction dataset for autonomous driving," *Scientific Data*, vol. 11, 2024, doi: 10.1038/s41597-024-03261-9.

[20] Z. Feng *et al.*, "MAFNet: Segmentation of road potholes with multi-modal attention fusion network for autonomous vehicles," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–12, 2022.

[21] Y. Lu *et al.*, "Communication-efficient federated learning for digital twin edge networks in industrial IoT," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 8, pp. 5709–5718, 2020.

[22] X. Wang *et al.*, "Applications and challenges of digital twin intelligent sensing technologies for asphalt pavements," *Automation in Construction*, vol. 164, p. 105480, 2024.

[23] T. Cao *et al.*, "Pavement crack detection based on 3D edge representation and data communication with digital twins," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 7, pp. 7697–7706, 2022.

[24] R. Mei *et al.*, "RoMe: Towards large scale road surface reconstruction via mesh representation," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 7, pp. 5173–5185, 2024.

[25] W. Wu *et al.*, "EMIE-MAP: Large-scale road surface reconstruction based on explicit mesh and implicit encoding," in *European Conference on Computer Vision (ECCV)*, 2024, pp. 370–386.

[26] T. Ren *et al.*, "Grounded SAM: Assembling open-world models for diverse visual tasks," *arXiv preprint arXiv:2401.14159*, 2024.

[27] T. Möller and B. Trumbore, "Fast, minimum storage ray/triangle intersection," in *ACM SIGGRAPH 2005 Courses*, 2005, pp. 7–es.

[28] W. Xu and F. Zhang, "FAST-LIO: A fast, robust LiDAR-inertial odometry package by tightly-coupled iterated Kalman filter," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 3317–3324, 2021.

[29] R. Fan *et al.*, "Pothole detection based on disparity transformation and road surface modeling," *IEEE Transactions on Image Processing*, vol. 29, pp. 897–908, 2019.

[30] B. Cheng *et al.*, "Masked-attention mask transformer for universal image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 1290–1299.

[31] H. Zhao *et al.*, "Pyramid scene parsing network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2881–2890.

[32] X. Chu *et al.*, "Twins: Revisiting spatial attention design in vision Transformers," in *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, 2021, pp. 9355–9366.

[33] Y. Sun *et al.*, "RTFNet: RGB-thermal fusion network for semantic segmentation of urban scenes," *IEEE Robotics and Automation Letters*, vol. 4, no. 3, pp. 2576–2583, 2019.

[34] C. Hazirbas *et al.*, "FuseNet: Incorporating depth into semantic segmentation via fusion-based CNN architecture," in *Asian conference on computer vision*. Springer, 2016, pp. 213–228.

[35] R. Fan *et al.*, "SNE-RoadSeg: Incorporating surface normal information into semantic segmentation for accurate freespace detection," in *European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 340–356.

[36] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 5410–5418.

[37] H. Xu and J. Zhang, "AANet: Adaptive aggregation network for efficient stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 1959–1968.

[38] X. Jing *et al.*, "End-to-end stereo matching network with two-stage partition filtering for full-resolution depth estimation and precise localization of kiwifruit for robotic harvesting," *Computers and Electronics in Agriculture*, vol. 225, p. 109333, 2024.

[39] G. Xu *et al.*, "Iterative geometry encoding volume for stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 21 919–21 928.

[40] C.-W. Liu *et al.*, "Playing to vision foundation model's strengths in stereo matching," *IEEE Transactions on Intelligent Vehicles*, pp. 1–12, 2024, doi: 10.1109/TIV.2024.3467287.

[41] A. Geiger *et al.*, "Vision meets robotics: The KITTI dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[42] R. Fan *et al.*, "Road surface 3D reconstruction based on dense subpixel disparity map estimation," *IEEE Transactions on Image Processing*, vol. 27, no. 6, pp. 3025–3035, 2018.