

Sensor-Free Millimeter-Level Water Depth Estimation for Flood-Aware Route Optimization

Chau-Tan-Cuong Nguyen[✉], Nguyen-Anh-Khoa Doan[✉], Vu-Hoang Tran^{*✉}
Ho Chi Minh City University of Technology and Engineering, Vietnam

*Corresponding author. Email: hoangtv@hcmute.edu.vn

ARTICLE INFO		ABSTRACT
Received:	05/06/2026	Urban areas frequently face the compounded challenges of severe traffic congestion and localized flash flooding, rendering traditional navigation systems inadequate. Existing vision-based approaches either address traffic monitoring or flood detection in isolation, and none achieves the millimeter-level physical precision required for safe vehicle routing. This study proposes a novel vision-based sensing framework for dynamic route optimization that simultaneously estimates traffic density and physical water depth utilizing solely standard traffic camera feeds, requiring no additional sensors or infrastructure. The framework integrates a crowd-counting network for vehicle density estimation and a real-time segmentation network for real-time flood segmentation, achieving a mean Intersection over Union (mIoU) of 93.1%. This is coupled with a baseline-referenced contour depth method that leverages static anchor objects, extracted using the Segment Anything Model 3 (SAM 3), to reduce scale ambiguity. Furthermore, a two-step statistical filtering pipeline combining median-based interquartile range (IQR) outlier rejection and spatial coherence enforcement stabilizes depth measurements under adverse conditions. This approach achieves millimeter-level accuracy with a Mean Absolute Error (MAE) of 1.53 mm and 4.07 mm on the two test videos (Vid1 and Vid2), respectively, and demonstrates improved performance compared with state-of-the-art monocular depth estimators and Vision-Language Models. Finally, these dynamic measurements are incorporated into a multi-objective cost function, optimized via grid search, and solved using the A* pathfinding algorithm to dynamically recalculate safe trajectories. Simulation results show that the system redirects vehicles away from hazardous intersections while limiting the additional travel time and distance, providing a robust, sensor-free solution for Intelligent Transportation Systems (ITS) to enhance urban mobility and safety during extreme weather events.
Revised:	13/07/2026	
Accepted:	11/09/2026	
Online First:	25/09/2026	
Published:		
KEYWORDS		
Traffic Density Estimation;		
Water Depth Estimation;		
Segmentation;		
Route Optimization;		
Vision-Based Sensing.		

DOI: <https://doi.org/10.54644/jte.2026.2503>

Copyright © JTE. This is an open access article distributed under the terms and conditions of the [Creative Commons Attribution-NonCommercial 4.0 International License](#) which permits unrestricted use, distribution, and reproduction in any medium for non-commercial purpose, provided the original work is properly cited.

1. Introduction

Urban mobility faces severe disruptions from sudden flash flooding and traffic congestion, which conventional GPS navigation often fails to dynamically address. Traditionally, flood monitoring and traffic management are treated in isolation. A systematic review of existing literature reveals three primary research directions. First, remote sensing (e.g., Sentinel-1/2) provides broad flood mapping [1] but generally lacks the high temporal frequency and spatial resolution required for real-time urban flash floods. Conversely, physical Internet of Things (IoT) water-level sensors offer precision but incur prohibitive installation and scalability costs.

Second, to overcome hardware constraints, vision-based models employ semantic segmentation to identify inundated zones [2]. However, extracting absolute physical depth remains challenging under varying illumination conditions [3]. Third, reference-based optical methods estimate depth using submerged portions of traversing vehicles [4]. While effective, this approach relies on the continuous presence of vehicles and suffers from scale ambiguity. Concurrently, traffic density estimation has

evolved from traditional bounding-box detection to robust crowd-counting Convolutional Neural Networks (CNNs) [5]–[7], effectively handling severe visual occlusion.

Research Gap: Despite advancements in vision-based urban monitoring [8]–[11], a critical gap persists. Existing systems address traffic and flooding in isolation, and none achieves millimeter-level depth precision without relying on physical sensors or dynamic vehicles and without suffering from optical scale drift. A dynamic routing framework tightly integrating frame-level traffic density and precise, scale-calibrated physical water depth remains largely unexplored.

To bridge this gap, this paper proposes a framework leveraging existing street cameras. It employs Congested Scene Recognition Network (CSRNet) [12] for vehicle density estimation and Bilateral Network with Guided Aggregation (BiSeNet V2) [13] for real-time water segmentation. By integrating these models with the Segment Anything Model 3 (SAM 3) [14], the proposed architecture extracts static anchor objects from a dry baseline to calculate millimeter-accurate physical depth, mitigating illumination vulnerabilities. Finally, these continuous metrics are formulated into a multi-objective cost function to dynamically compute safe, flood-avoiding trajectories.

2. Methodology

This section presents the comprehensive methodology for the proposed dynamic route optimization system. As illustrated in Figure 1, the overall architecture uses existing urban traffic camera networks and is divided into three primary processing modules: traffic density estimation, water depth estimation, and dynamic route optimization.

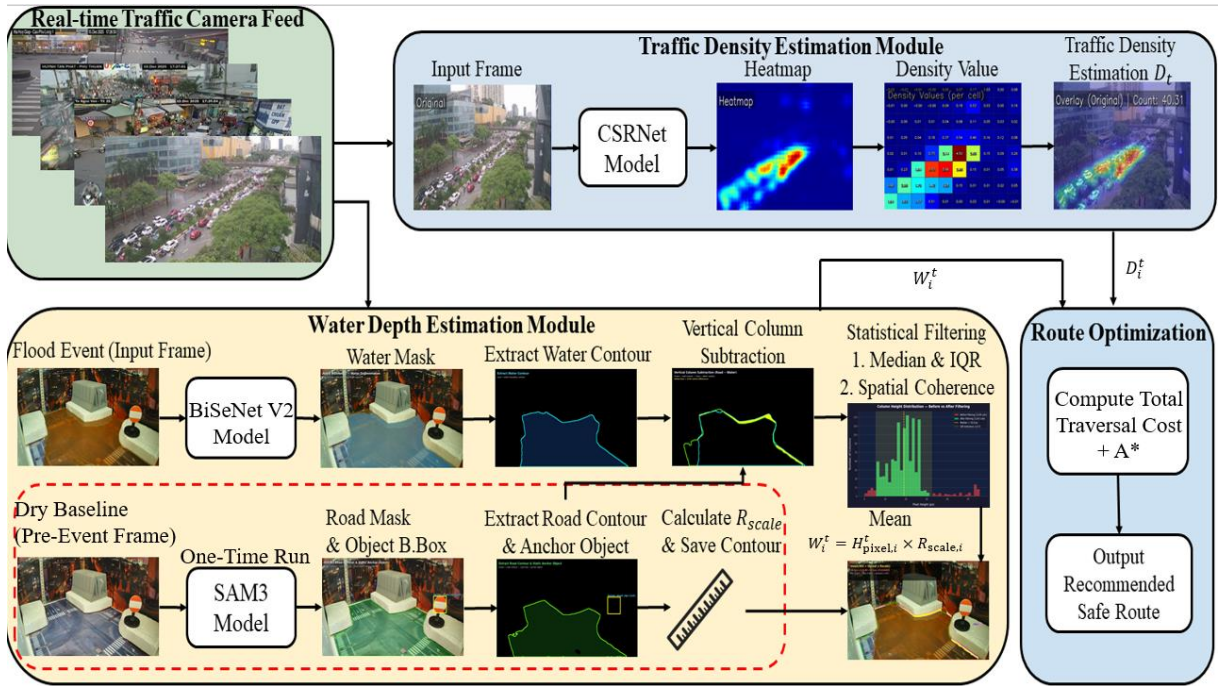


Figure 1. Overall architecture of the proposed framework.

2.1. Traffic Density Estimation

To effectively estimate the number of vehicles in highly congested intersections, where severe occlusion and visual overlap frequently occur, the proposed framework adapts a crowd-counting paradigm rather than relying on traditional bounding-box object detection [7]. Specifically, the CSRNet [12] is deployed to generate high-resolution density maps from real-time traffic surveillance feeds.

The architecture of CSRNet is designed to process highly dense scenes without relying on excessive pooling layers, which typically degrade spatial resolution. The model comprises two primary components: a front-end for 2D feature extraction and a back-end for density map generation. The front-

end utilizes the first ten layers of a pre-trained VGG-16 network [15], acting as a robust feature extractor that maintains an optimal balance between accuracy and computational efficiency.

To map these extracted features into a continuous density map while preserving spatial resolution, the back-end employs a series of dilated convolutional layers. Unlike standard convolution, dilated convolution expands the receptive field of the network by inserting spaces between kernel elements, enabling the model to capture broader contextual information while preserving the spatial resolution of the density map [12]. For each road segment i , the corresponding camera generates a density map for each frame t . By integrating the pixel values over this density map, the system estimates the total number of vehicles within the camera's field of view. This estimated vehicle count is defined as the traffic density factor D_i^t of road segment i at frame t . The value of D_i^t is updated for each incoming frame and subsequently used as a dynamic input to the route optimization module.

Furthermore, to ensure uninterrupted system operation, the CSRNet model is deployed to continuously estimate vehicle density across all weather conditions, including both normal dry periods and adverse rainy or flooded events. This continuous monitoring ensures the routing algorithm consistently receives real-time traffic data. Consequently, the system can evaluate congestion even when the road surface is partially or fully obscured by floodwater.

2.2. Water Depth Estimation

The proposed framework uses a granular estimation approach to calculate physical water depth in millimeters (mm) and centimeters (cm). As illustrated in the comprehensive system architecture (Figure 1), the process strategically decouples static scene analysis (One-Time Run) from real-time dynamic processing to optimize computational resources. Before any flooding event, the system captures a dry baseline image from the target camera. As depicted in the dashed red "One-Time Run" block of Figure 1, the SAM 3 [14] is applied exclusively to this pre-event frame to identify the road mask and isolate static reference objects, such as traffic poles and street signs. Because the physical dimensions of these objects are known a priori, a real-world-to-pixel scale ratio R_{scale} is initially computed. To robustly account for perspective distortion—where objects closer to the camera appear disproportionately larger than those further away—the system identifies multiple reference objects distributed across the field of view. Their individual scale ratios are then spatially interpolated to derive a highly accurate, perspective-calibrated R_{scale} that minimizes geometric errors caused by the camera's viewing angle. Once R_{scale} is established, SAM 3 is deactivated, which significantly reduces computational overhead during the subsequent inference phase.

During the active monitoring phase, the BiSeNet V2 [13] processes the real-time video feed. Chosen for its optimal balance of high-speed inference and spatial accuracy, BiSeNet V2 is specifically trained to isolate flooded regions. Upon detecting a flood event, the network outputs a dynamic water mask, precisely delineating the visible extent of the surface water over the roadway. To calculate the absolute water depth, the system extracts and compares the geometric boundaries of the pre-defined road baseline and the current water surface, executing a vertical column-by-column contour subtraction. This vertical pixel difference directly represents the rising water level from the camera's fixed perspective.

However, raw contour subtraction often yields a set of columns with varying pixel heights due to camera noise, dynamic surface waves, or partial occlusions. To mitigate these anomalies and ensure measurement precision, a two-step statistical filtering process is applied directly to these raw columns. Initially, the system utilizes the Median and Interquartile Range (IQR) method to identify the majority cluster, effectively filtering out extreme outliers. Subsequently, a spatial coherence filter is enforced, predicated on the physical premise that a localized water surface remains flat; thus, neighboring vertical columns must exhibit highly similar heights. This two-stage filtering process enables the selection of a robust and representative vertical pixel height $H_{\text{pixel},i}^t$ for road segment i at frame t . Finally, the physical water depth W_i^t is calculated by multiplying the filtered pixel height by the perspective-calibrated scale ratio $R_{\text{scale},i}$:

$$W_i^t = H_{\text{pixel},i}^t \times R_{\text{scale},i} \quad (1)$$

2.3. Dynamic Route Optimization

The route optimization module evaluates the traversal cost of each road segment i using three inputs available at the same time t : its physical length L_i , the traffic density D_i^t estimated from the corresponding camera using CSRNet, and the water depth W_i^t estimated by the depth estimation module (Figure 2). Thus, for a given frame t , the system obtains the traffic and water conditions of all monitored road segments and computes an individual traversal cost for each segment. For each road segment i , the algorithm first compares W_i^t with the predefined safe wading threshold of 250 mm. If the water depth exceeds this threshold, a maximum penalty is assigned to the segment, effectively preventing it from being selected for routing. Otherwise, the traversal cost is calculated as:

$$C_i^t = \alpha \times L_i + \beta \times D_i^t + \gamma \times W_i^t \quad (2)$$

where α , β , and γ are empirical weights for distance, traffic congestion, and water depth, respectively, normalized such that $\alpha + \beta + \gamma = 1$. The optimal coefficients were determined via a grid search (step size 0.1) across 10 simulated urban scenarios with varying rainfall intensities, optimizing for minimum hazard encounters while maintaining practical traversal times. As summarized in Table 1, distance- and traffic-priority configurations incurred 42 and 35 critical hazard encounters, respectively, while the Extreme Safety configuration eliminated hazards at the cost of impractical detour times (24.6 min). The proposed Safety-First configuration ($\alpha = 0.2$, $\beta = 0.3$, $\gamma = 0.5$) achieves the optimal trade-off, resulting in only 5 hazard encounters, all on segments with water below the safe wading threshold, with an average traversal time of 18.4 min. At each time t , the routing graph is updated with the segment costs C_i^t for all road segments, and the A* algorithm [16] is then executed on the updated graph to output the safest and most efficient route.

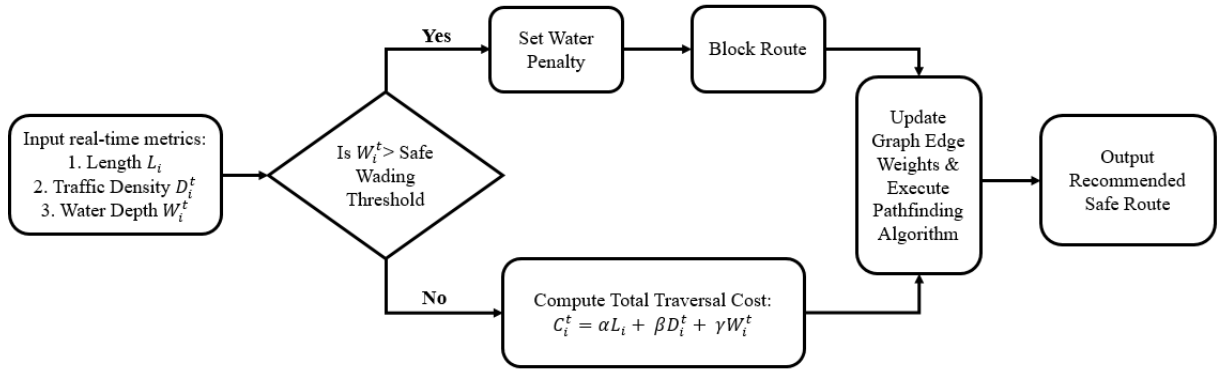


Figure 2. Flowchart of the dynamic route optimization algorithm incorporating traffic density and water depth constraints.

Table 1. Grid-search optimization results for cost function coefficients across 10 scenarios.

Configuration Strategy	α	β	γ	Avg. Traversal Time (min)	Avg. Distance (km)	Total Critical Hazard Encounters
Distance-Priority	0.7	0.2	0.1	14.5	4.20	42
Traffic-Priority	0.2	0.7	0.1	15.2	4.85	35
Balanced	0.3	0.3	0.4	16.8	4.50	11
Safety-First (Proposed)	0.2	0.3	0.5	18.4	4.92	5
Extreme Safety	0.1	0.1	0.8	24.6	6.10	0

3. Results and Discussion

3.1. Experimental Setup and Evaluation Metrics

The proposed system was developed, trained, and evaluated on a high-performance workstation equipped with an Intel Core i9-10900K processor (3.70 GHz) and an NVIDIA GeForce RTX 3090 GPU. To ensure that the models handle real-world urban complexities robustly, they were extensively fine-tuned on domain-specific data. For the traffic density estimation module, the CSRNet model [12] was trained for 200 epochs using a comprehensive, multi-source dataset comprising UA-DETRAC [20], VisDrone [21], CARPK [22], and TRANCOS [23], alongside a proprietary dataset specifically capturing Vietnamese urban traffic conditions. For the dynamic water segmentation task, BiSeNet V2 [13] was trained for 380 epochs on a custom-curated flood dataset to accurately distinguish diverse water textures and reflections. SAM 3 [14] was retained with its original pre-trained weights to preserve its zero-shot generalization capability for static road baseline extraction.

To rigorously validate the system quantitatively, specific evaluation metrics were established for each processing module. The performance of the traffic density estimation was evaluated using Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). For the water depth estimation, the module was initially verified qualitatively on real-world surveillance feeds and subsequently tested in a controlled physical environment where physical Ground Truth (GT) measurements were available. The depth measurement accuracy across two distinct test videos (Vid1 and Vid2) was assessed using MAE, Mean Absolute Percentage Error (MAPE), Measurement Bias, and Standard Deviation (Std). Furthermore, physical GT was established using a commercial millimeter ruler (± 0.5 mm tolerance). Because this is a proof-of-concept study, rigorous checkerboard camera calibration was bypassed; instead, scale calibration relied on the R_{scale} derived from static anchor objects. Measurement repeatability was assessed empirically by examining the variation in predicted depth across a continuous window of frames captured while the true water level remained stable, as detailed in Section 3.3.4. Finally, the dynamic route optimization algorithm was evaluated based on computational latency and the capability to safely bypass hazardous flooded segments.

3.2. Evaluation of Traffic Density Module

To effectively quantify real-time traffic conditions, the CSRNet module [12] was evaluated on its ability to generate accurate density maps from traffic surveillance feeds. The model was specifically trained to adapt to the perspective and scale of the target urban traffic cameras.

During the training phase, the model demonstrated stable convergence, with both training and validation losses steadily decreasing over the epochs without signs of overfitting. Quantitatively, the optimized CSRNet achieved an MAE of 2.0778 and an RMSE of 2.9960 on the testing sequence from the MIO-TCD dataset [24], which was not included in the training set. This indicates that the model's estimation closely aligns with the actual number of vehicles on the road segment.

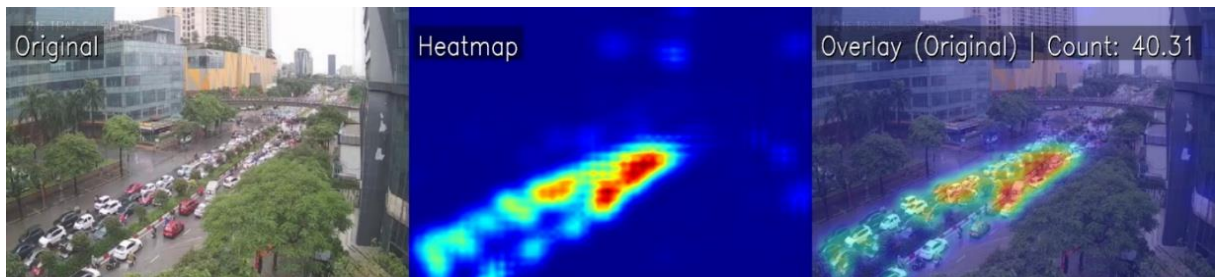


Figure 3. Qualitative visualization of the CSRNet outputs. Left: The original traffic camera input. Middle: The generated continuous density heatmap accurately highlighting congested regions. Right: Heatmap overlaid onto the original traffic camera image.

Qualitatively, the model's spatial understanding is illustrated in Figure 3. Instead of relying on individual bounding-box detections, which often fail under heavy occlusion, CSRNet generates a highly detailed continuous density map. The heatmap visualization explicitly shows that the model correctly

assigns higher density weights (red regions) to congested vehicle clusters while accurately suppressing background noise. This estimated vehicle count is subsequently used as the traffic density factor D_i^t and seamlessly fed into the multi-objective routing algorithm.

3.3. Evaluation of Water Depth Estimation

The water depth estimation module is the core of the proposed system. Its performance relies on both the accuracy of the semantic segmentation and the robustness of the geometric projection algorithm.

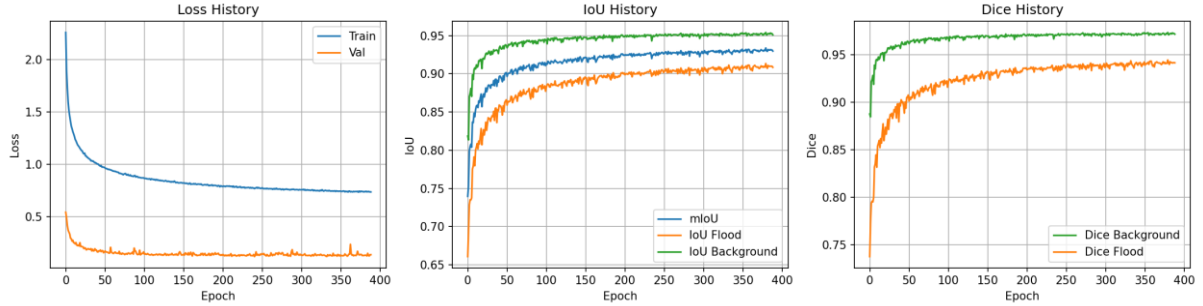


Figure 4. Training and validation metrics of the BiSeNet V2 model over 380 epochs.

3.3.1. Segmentation Performance of BiSeNet V2

As shown in Figure 4, BiSeNet V2 [13] was trained for 380 epochs, with the training loss decreasing steadily to ~ 0.73 and the validation loss stabilizing at ~ 0.13 , demonstrating robust convergence without overfitting. The proposed model achieved an overall mean Intersection over Union (mIoU) of 93.1% on the validation set. Specifically, the per-class IoU reached 95.1% for the background and 91.1% for the flooded regions. In terms of Dice coefficients, the model attained 97.1% for the background and 94.1% for flooded areas, demonstrating strong segmentation performance across both classes. These precise pixel-level boundaries are critical, as any segmentation deviation directly compromises the precision of the subsequent contour subtraction.

3.3.2. Quantitative Comparison of Water Depth Estimation

The proposed M1 method was benchmarked against three state-of-the-art approaches: M2 (Depth Anything 3, a dense monocular depth estimation (MDE) network) [17], M3 (SD-VLM [18], a spatial-reasoning vision-language model (VLM) using either an external Depth Anything 3 (DA3) depth map or its internal Depth Anything V2 (DA2) [25] module), and M4 (DepthLM, a VLM using visual point-prompting) [19]. Unlike M1, which relies on a pre-event dry baseline, these models estimate depth directly from single 2D images. Results across two test sequences are presented in Table 2. Each baseline method is described in detail below.

M2 — Depth Anything 3 (DA3) [17] is a state-of-the-art MDE network that produces dense, high-resolution relative depth maps from a single RGB image. DA3 leverages a large-scale Vision Transformer (ViT) backbone pre-trained on a diverse mixture of labeled and unlabeled data, enabling strong zero-shot generalization to unseen scenes. However, as a relative depth estimator, its raw output lacks a known metric scale. To enable fair quantitative comparison, each DA3 prediction is converted to a physical measurement via the Geometric Projection in Equation (3). Two measurement positions on the water contour are evaluated: *Center* (the midpoint of the visible water surface) and *Rim* (the outermost edge of the flooded contour), under both Aligned (scale-calibrated with R_{scale}) and Unaligned (raw depth output) configurations.

M3 — SD-VLM [18] is a spatial-reasoning Vision-Language Model that integrates depth priors into the visual reasoning pipeline to produce metric-aware spatial measurements from free-form language queries. It operates in two configurations: *VLM Map*, in which an external DA3 depth map is provided as an additional input channel to enrich the model's geometric context; and *VLM Internal*, in which the model relies solely on its own internal depth estimation module (DA2) without external depth input.

Both configurations are evaluated in Aligned mode, using R_{scale} from SAM 3 [14] to calibrate the predicted depth into metric units.

M4 — DepthLM [19] is a Vision-Language Model designed for metric depth estimation via visual point-prompting, in which the user specifies target regions of interest directly in the image using point coordinates. DepthLM grounds metric depth predictions by attending to both the visual content and the spatial prompt, aiming to recover absolute depth at specified locations without requiring external calibration. It is evaluated under two conditions: *Aligned* (scale-calibrated with R_{scale}) and *Unaligned* (using the model’s raw metric output directly).

Table 2. Quantitative comparison of water depth estimation errors across various baseline configurations.

Method	MAE (mm)		MAPE (%)		Bias (mm)		Std (mm)	
	Vid1	Vid2	Vid1	Vid2	Vid1	Vid2	Vid1	Vid2
M1 Contour Depth (Proposed)	1.53	4.07	19.55	66.41	0.25	0.28	2.06	9.08
M2 Aligned Center	8.36	7.99	61.35	91.57	5.06	4.22	11.05	15.16
M2 Aligned Rim	14.87	10.32	130.43	117.78	12.83	1.97	12.34	21.49
M2 Unaligned Center	19.76	9.33	123.20	98.80	−1.28	0.29	35.37	18.80
M2 Unaligned Rim	109.45	19.28	1286.38	166.29	109.45	18.86	42.49	35.72
M3 SD-VLM Map	50.85	42.59	664.18	764.88	44.29	41.17	36.21	40.07
M3 SD-VLM Internal	43.96	17.99	632.31	207.84	38.26	14.76	34.54	23.56
M4 DepthLM Aligned	17.04	40.31	195.26	851.33	−11.57	39.26	15.81	48.72
M4 DepthLM Unaligned	15.35	54.97	100.00	959.50	−15.35	52.80	7.79	64.32

For comparison, a Geometric Projection mechanism converts the estimated depth difference ΔZ_i^t (in meters) into physical water depth W_i^t using camera intrinsic parameters:

$$W_i^t = |v_i^t - c_{y,i}| \times \frac{\Delta Z_i^t}{f_{y,i}}, \quad (3)$$

where $f_{y,i}$ and $c_{y,i}$ denote the vertical focal length and the vertical principal-point coordinate of the camera associated with road segment i , respectively, and v_i^t denotes the vertical image coordinate of the water-mask boundary at frame t . In the absence of calibrated intrinsic parameters, $f_{y,i}$ is approximated as 0.8 times the image width, while $c_{y,i}$ is approximated as half of the image height.

The quantitative results indicate the fundamental limitations of applying state-of-the-art MDE and VLM to dynamic water surfaces. Among the baselines, DA3 [17] (M2 Aligned Center) outperformed language-based models, achieving an MAE of 7.99 mm on Vid2. This indicates that its low-level visual features are relatively more resilient to the textureless nature of water. However, DA3’s accuracy degrades significantly at the water boundary (M2 Aligned Rim: 14.87 mm on Vid1), demonstrating that pure MDE networks typically struggle with specular reflections that distort neural geometric depth cues. Furthermore, unaligned MDE configurations suffered severely from inherent scale ambiguity, with M2 Unaligned Rim producing entirely impractical physical measurements reaching 109.45 mm MAE on Vid1.

Conversely, VLM (M3 SD-VLM [18] and M4 DepthLM [19]) exhibited severe vulnerabilities. Despite possessing strong semantic reasoning capabilities to identify flooded scenes, they struggle to ground precise metric depth on specular surfaces. Specifically, through visual point-prompting on a mirror-like texture (M4 DepthLM Aligned) yields highly unstable outputs (40.31 mm MAE on Vid2) because the model cannot establish a reliable spatial prior. This limitation becomes more pronounced without external calibration, as seen when M4 DepthLM Unaligned produced an MAE of 54.97 mm on

Vid2, confirming that language-guided reasoning cannot mathematically resolve physical scale on reflective liquids.

In stark contrast, the proposed M1 method clearly outperformed, reducing the MAE to 1.53 mm on Vid1 and 4.07 mm on Vid2. This superiority stems from a fundamental paradigm shift: instead of forcing a neural network to infer 3D depth directly from a deceptive, reflective 2D water surface, M1 mathematically reduces the problem to robust 2D semantic segmentation combined with rigid geometric projection. By relying strictly on the known dimensions of static anchor objects from a pre-event dry baseline rather than interpreting the water's optical texture, M1 successfully neutralizes the optical illusions that confound direct depth estimators.

Table 3. Comparison of Experimental Conditions and Optical Image Characteristics Between Vid1 and Vid2.

Metric	Camera angle	Brightness	Contrast	Saturation	Sharpness
Vid1	~10°	98.81	54.06	141.49	95.34
Vid2	~15°	96.94	50.41	130.13	47.26

The MAPE increased from 19.55% (Vid1) to 66.41% (Vid2) due to degraded optical conditions, notably reduced sharpness (47.26 vs. 95.34) as shown in Table 3. Furthermore, temporal analysis in Figure 5 reveals that Vid2's error spikes are strictly localized to the turbulent water-pouring phase (Frames 55–130), where dynamic ripples cause minor segmentation misalignments. This confirms that Vid2's inflated error is driven by transient high-frequency surface noise rather than systematic scale drift. Crucially, the impact of Vid2 varied substantially across methods: M2 Aligned Center reached a MAPE of 91.57%, while M4 DepthLM Aligned and M4 DepthLM Unaligned exhibited substantially higher MAPE values of 851.33% and 959.50%, respectively. In contrast, several baselines showed improved MAE on Vid2, including M2 Unaligned Rim, which decreased from 109.45 to 19.28 mm, and M3 SD-VLM Internal, which decreased from 43.96 to 17.99 mm. This comparative evidence demonstrates that the proposed method exhibited the highest resilience among all evaluated methods under these test conditions. Future work will integrate temporal smoothing and multi-camera cross-verification to further stabilize measurements under adverse conditions.

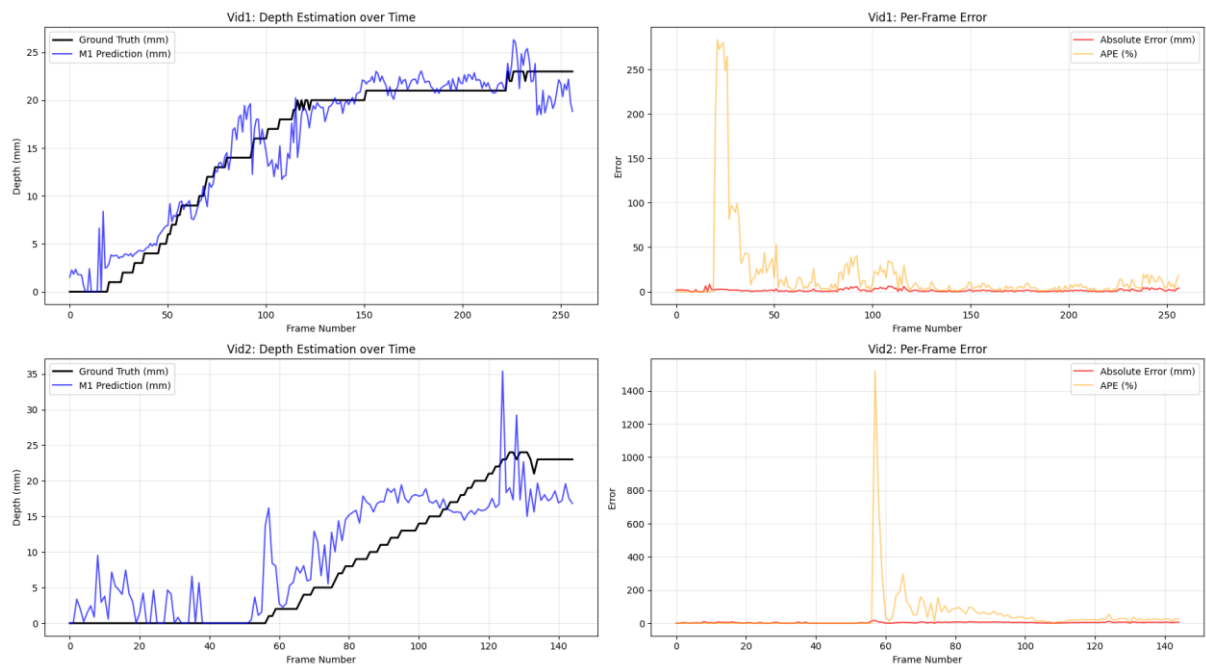


Figure 5. Temporal depth estimation and per-frame error analysis for Vid1 and Vid2.

3.3.3. Qualitative Analysis of the Contour-Based Mechanism

Qualitatively, the proposed contour-based mechanism operates by sequentially isolating the road area and flooded region, projecting them into a unified coordinate space, extracting a static reference object for scale calibration, and ultimately mapping the final vertical pixel difference. During this process, the Median-IQR and spatial coherence filters effectively suppress surface noise caused by reflections and ripples, contributing to a highly stable contour extraction. To further quantify the contribution of each filtering stage, an ablation study was conducted by independently enabling each filter and evaluating the resulting measurement performance.

As shown in Table 4, the two filtering stages do not contribute independently in equal measure. Applied alone, Median-IQR substantially reduces error on Vid1 (MAE: 6.40→2.94 mm) but provides no measurable benefit on Vid2 (16.19→16.43 mm), while Spatial Coherence alone closely tracks the unfiltered baseline on both videos (e.g., Bias of 6.37 mm vs. 6.39 mm on Vid1). This asymmetry reflects the sequential nature of the two corrections, consistent with the processing order described in Section 2.2. Median-IQR first removes statistically extreme outlier columns, establishing a locally reliable contour. Once these gross corruptions are removed, the Spatial Coherence filter can meaningfully enforce neighborhood consistency, since its reference window would otherwise be contaminated by the same outliers it is designed to smooth. The two filters therefore target distinct and complementary failure modes—statistically extreme deviations versus spatially localized inconsistency—and only their combination reduces MAE, MAPE, Bias, and Standard Deviation to their lowest values across both test videos, confirming that both stages are necessary components of the proposed stabilization pipeline.

Table 4. Ablation study of the proposed stabilization filters.

Filter Configuration	MAE (mm)		MAPE (%)		Bias (mm)		Std (mm)	
	Vid1	Vid2	Vid1	Vid2	Vid1	Vid2	Vid1	Vid2
No Filter	6.40	16.19	64.10	80.38	6.39	14.46	4.12	28.57
Median-IQR Only	2.94	16.43	32.50	76.18	1.87	14.33	3.62	29.16
Spatial Coherence Only	6.38	16.20	63.67	80.38	6.37	14.46	4.13	28.57
Proposed (Both Filters)	1.53	4.07	19.55	66.41	0.25	0.28	2.06	9.08

3.3.4. Linearity and Error Margin Analysis

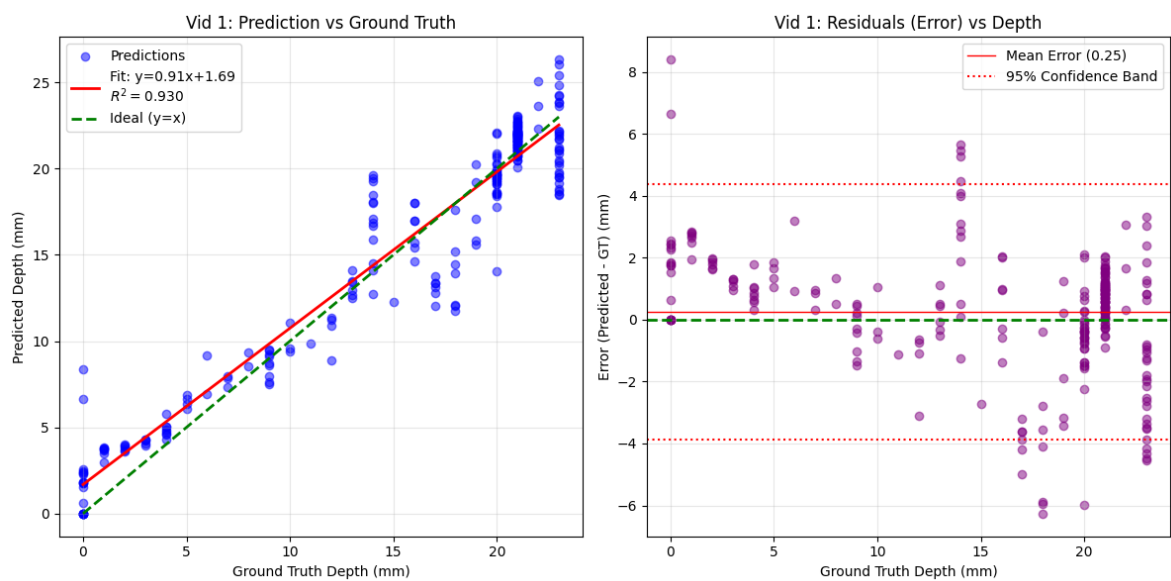


Figure 6. Linearity and error margin analysis (Vid1). Left: Linear regression $R^2 = 0.93$. Right: Residual plot showing stable bounds and improved precision at deeper levels.

Linear regression in Figure 6 demonstrates a strong correlation between predicted and ground-truth depths ($R^2 = 0.93$), where y represents the predicted depth (mm) and x represents the ground-truth depth (mm). The fitted equation $y = 0.91x + 1.69$ confirms scale fidelity, with the 1.69 mm intercept reflecting baseline surface irregularities.

Residual analysis yields a stable mean bias of +0.25 mm. Notably, absolute error does not diverge with depth. Shallow water (10–20 mm) exhibits a slightly elevated mean error (2.38 mm) due to surface turbulence. However, in deeper zones (> 20 mm), precision improves significantly (mean error: 1.21 mm, Std: 1.07 mm). This counterintuitive improvement is likely because deeper water forms a higher-contrast waterline, enabling sharper boundary segmentation.

As the proposed pipeline contains no stochastic inference components, re-processing an identical frame is trivially reproducible by construction. Repeatability is therefore assessed empirically instead, by examining the variation in predicted depth across a continuous window of frames captured from Vid1 while the true water level remained stable (frames 120–220 as shown in Figure 5). As shown in Table 5, this yields a standard deviation of 0.91 mm around the mean prediction, reflecting the system’s robustness to real sensor and environmental noise rather than mere computational determinism.

Table 5. Repeatability metrics during a stable water phase (Vid1, Frames 120–220).

GT	MAE (mm)	MAPE (%)	Bias (mm)	Std (mm)
19.0–21.0 mm	0.84	4.05	+0.41	0.91

3.4. Dynamic Route Optimization Case Study

To demonstrate practical applicability, a route optimization scenario was simulated for a vehicle navigating an urban network during heavy rainfall (Figure 7). A conventional navigation system identifies the baseline Shortest Path (18.6 km, highlighted in red). However, the proposed water depth estimation module measures $W_i^t = 530$ mm along this route—exceeding a safe wading threshold of 250 mm. The hard safety constraint is triggered, applying a maximum penalty and effectively blocking this segment from route selection. The multi-objective cost function in Equation (2) subsequently evaluates the remaining candidates to identify the Proposed Dynamic Route (20.0 km, highlighted in green).



Figure 7. Web-based UI demonstration of the dynamic route optimization, illustrating the system actively avoiding a severely flooded segment.

As shown in Table 6, the dynamic route is 1.4 km longer and has a slightly higher traffic density (77.4 vs. 71.4). Under normal conditions, a standard navigator would prefer the shorter route. However, the 530 mm flood depth triggers an absolute safety override, eliminating the flooded segment before the

cost function evaluates alternatives. This case study illustrates the Safety-First configuration: the system avoided the flooded segment at the cost of a 1.4 km detour.

Table 6. Route comparison under simulated heavy rainfall conditions.

Metrics	Shortest Path	Proposed Dynamic Route
Total Distance	18.6 km	20.0 km
Max Water Depth Encountered	530 mm	0 mm
Average Traffic Density Factor	71.4 vehicles/road segment	77.4 vehicles/road segment
Safety Status	Engine Stall Risk	Cleared for Wading

3.5. Discussion and Limitations

The proposed framework achieves millimeter-level accuracy in real time by grounding measurements on static anchors from a pre-event baseline. To thoroughly evaluate computational efficiency, a detailed latency breakdown of the individual processing modules is presented in Table 7.

A key architectural advantage of the active monitoring phase is the parallel execution of the deep learning models. By applying multithreading, CSRNet [12] and BiSeNet V2 [13] process the incoming frames concurrently. Consequently, the computational bottleneck is dictated by the maximum latency of the parallel branches rather than their sequential sum. Together with the highly efficient contour projection and A* pathfinding [16], the end-to-end latency is just 34.0 ms per frame (29.4 frames per second FPS). This real-time visual architecture effectively neutralizes the scale ambiguity inherent in MDE and VLM on dynamic water surfaces, serving as a robust prerequisite for downstream safety-first routing applications.

Table 7. Breakdown of computational latency across the active processing pipeline.

Algorithm / Task	Execution Mode	Average Latency (ms)
BiSeNet V2 (Water Segmentation)	Concurrent (Thread 1)	28.0
CSRNet (Traffic Density Estimation)	Concurrent (Thread 2)	30.0
Contour Projection & Filters	Sequential	3.0
Multi-objective Cost & A*	Sequential	1.0
End-to-End System	Parallel Optimized	34.0 (29.4 FPS)

However, certain operational limitations remain. First, strict reliance on a stable baseline makes the system sensitive to camera vibrations, necessitating recalibration whenever the camera shifts. Second, high-frequency surface noise (e.g., vehicle wakes, floating debris) can degrade segmentation precision. Third, pure RGB sensing is vulnerable to adverse conditions like dense fog, poor nighttime illumination, or intense specular reflections.

Finally, empirical evaluations face constraints. Direct benchmarking against several contemporary methods was precluded by unavailable source code and proprietary datasets, restricting comparisons to open-source architectures. Additionally, the two initial test sequences cannot fully capture real-world complexities. Future work will expand evaluations across comprehensive datasets encompassing diverse urban scenarios and extreme weather to confirm large-scale generalizability.

4. Conclusions

This paper presents a comprehensive vision-based sensing framework designed to extract precise environmental metrics for dynamic route optimization. While the system ultimately culminates in a navigation tool, the primary scientific contribution of this research lies in the measurement methodology itself: integrating real-time traffic density analysis with a novel, baseline-referenced contour subtraction method. This visual architecture achieves accurate water depth estimation, with a MAE of 1.53 mm on Vid1 and 4.07 mm on Vid2. These results demonstrate that the proposed approach maintains millimeter-

level estimation accuracy across different video conditions while addressing the geometric limitations of conventional deep learning depth models. Furthermore, the practical routing case study serves as a downstream application, demonstrating the system's real-time efficiency in dynamically recalculating safe, flood-avoiding trajectories. Ultimately, this research provides a robust and accurate architecture with potential for integration into modern Intelligent Transportation Systems (ITS) to enhance urban safety and navigation resilience.

Acknowledgments

This research was funded by the Ho Chi Minh City University of Technology and Engineering, Vietnam, under grant No. SV2026-444.

Conflict of Interest

The authors declare that there is no conflict of interest in this paper.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT for English language editing. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

- [1] P. Hofmann, M. Wieland, and S. Martinis, "Improving Reliability in Global Flood Mapping by Generating a Seasonal Reference Water Mask Using Sentinel-1/2 Time-Series Data," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLIII-B3-2022, pp. 1127–1132, 2022.
- [2] R. J. Pally and S. Samadi, "Application of Image Processing and Convolutional Neural Networks for Flood Image Classification and Semantic Segmentation," *Environmental Modelling & Software*, vol. 145, Art. no. 105285, 2021.
- [3] P. Jiang, G. Xu, G. Liu, and Y. Mao, "Visual Measurement of Water Level under Complex Illumination Conditions," *Sensors*, vol. 19, no. 19, Art. no. 4141, 2019.
- [4] J. H. Park et al., "Computer Vision-Based Estimation of Flood Depth in Flooded-Vehicle Images," *Journal of Computing in Civil Engineering*, vol. 35, no. 2, Art. no. 04020067, 2021, doi: 10.1061/(ASCE)CP.1943-5487.0000956.
- [5] T. N. M. Thu, V. L. Q. Phuong, L. B. B. Duy, and P. N. Khang, "Estimating Traffic Density Using Convolutional Neural Networks Based on Crowd Counting," in *Communications in Computer and Information Science*, vol. 2226, Springer, 2024, pp. 282–296, doi: 10.1007/978-981-96-0434-0_20.
- [6] D. P. Mien, T. T. Vu, and N. V. Si, "Estimating Traffic Density in Uncertain Environment," *UD - Journal of Science and Technology: Issue on Information and Communications Technology*, vol. 21, no. 6.2, pp. 43–48, 2023, doi: 10.51138/jst-ud.v21i6.2.8305.
- [7] L. Ciampi, C. Santiago, J. P. Costeira, C. Gennaro, and G. Amato, "Domain Adaptation for Traffic Density Estimation," in *Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2021) - Volume 5: VISAPP*, 2021, pp. 185–195, doi: 10.5220/0010303401850195.
- [8] J. Jang and Y. W. Kwon, "Real-Time Urban Flood Detection Using YOLOv5-Seg," in *Proceedings of the 16th International Conference on Ubiquitous and Future Networks (ICUFN)*, Lisbon, Portugal, 2025, pp. 174–176, doi: 10.1109/ICUFN65234.2025.11170106.
- [9] B. Alizadeh, D. Li, Z. Zhang, and A. H. Behzadan, "Feasibility Study of Urban Flood Mapping Using Traffic Signs for Route Optimization," *arXiv preprint arXiv:2109.11712*, 2021.
- [10] D. Kim and S. Lee, "Computer Vision-Based Estimation of Inundation Depth for Urban Flood Monitoring," in *Computing in Civil Engineering 2024*, ASCE, 2025, pp. 518–525, doi: 10.1061/9780784486139.056.
- [11] W. Jung et al., "Video-Based Deep Learning Approach for Water Level Monitoring in Reservoirs," *Water*, vol. 17, no. 17, Art. no. 2525, 2025.
- [12] Y. Li, X. Zhang, and D. Chen, "CSRNet: Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 1091–1100.
- [13] C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, and N. Sang, "BiSeNet V2: Bilateral Network with Guided Aggregation for Real-time Semantic Segmentation," *International Journal of Computer Vision*, vol. 129, pp. 3051–3068, 2021.
- [14] N. Carion et al., "SAM 3: Segment Anything with Concepts," in *International Conference on Learning Representations (ICLR)*, 2026.
- [15] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *International Conference on Learning Representations (ICLR)*, 2015.
- [16] P. E. Hart, N. J. Nilsson, and B. Raphael, "A Formal Basis for the Heuristic Determination of Minimum Cost Paths," *IEEE Transactions on Systems Science and Cybernetics*, vol. 4, no. 2, pp. 100–107, 1968.
- [17] H. Lin et al., "Depth Anything 3: Recovering the Visual Space from Any Views," in *International Conference on Learning Representations (ICLR)*, 2026.
- [18] P. Chen et al., "SD-VLM: Spatial Measuring and Understanding with Depth-Encoded Vision-Language Models," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.
- [19] Z. Cai et al., "DepthLM: Metric Depth from Vision Language Models," in *International Conference on Learning Representations (ICLR)*, 2026.

- [20] L. Wen et al., "UA-DETRAC: A New Benchmark and Protocol for Multi-Object Detection and Tracking," in *Computer Vision and Image Understanding*, vol. 193, p. 102907, 2020.
- [21] A. Jadhav et al., "Aerial Multi-Object Tracking by Detection Using Deep Association Networks," *arXiv preprint arXiv:1909.01547*, 2019.
- [22] M.-R. Hsieh, Y.-L. Lin, and W. H. Hsu, "Drone-Based Object Counting by Spatially Regularized Regional Proposal Network," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 1437–1445.
- [23] R. Guerrero-Gómez-Olmedo et al., "Extremely Overlapping Vehicle Counting," in *Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA), Lecture Notes in Computer Science*, vol. 9117, Springer, 2015, pp. 423–431.
- [24] Z. Luo et al., "MIO-TCO: A new benchmark dataset for vehicle classification and localization," *IEEE Transactions on Image Processing*, vol. 27, no. 10, pp. 5129–5141, 2018.
- [25] L. Yang et al., "Depth Anything V2," *arXiv preprint arXiv:2406.09414*, 2024.

Chau-Tan-Cuong Nguyen (Student ID: 23146007) is currently a fourth-year Mechatronics Engineering student at Ho Chi Minh City University of Technology and Engineering, Vietnam, and is conducting research at the UTE-AI Lab. His research interests focus on autonomous vehicles and traffic density estimation.

Email: 23146007@student.hcmute.edu.vn. ORCID:  <https://orcid.org/0009-0007-0665-3255>.

Nguyen-Anh-Khoa Doan (Student ID: 23134030) is currently a fourth-year Robot & AI Engineering student at Ho Chi Minh City University of Technology and Engineering, Vietnam, and is conducting research at the UTE-AI Lab. His research interests focus on autonomous vehicles and traffic density estimation.

Email: 23134030@student.hcmute.edu.vn. ORCID:  <https://orcid.org/0009-0009-3482-4941>.

Vu-Hoang Tran received the B.S. degree in Automation Engineering from Ho Chi Minh City University of Technology and Engineering, Vietnam in 2012, the Master degree in Manufacturing and Management from National Kaohsiung University of Applied Sciences, Taiwan, in 2015, and the Ph.D. degree in Electrical Engineering from National Chung Cheng University, Taiwan, in 2018. He is currently a lecturer in Faculty of Electrical and Electronics Engineering, Ho Chi Minh City University of Technology and Engineering, Vietnam. His research interests are in image processing, pattern recognition, computer vision, machine learning, and transfer learning. He has published research papers in a variety of prestigious journals and top conferences, such as, IEEE Transactions on Circuits and Systems for Video Technology (TCSVT), International Conference on Computer Vision (ICCV), IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Asian Conference on Computer Vision (ACCV), Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), IEEE International Conference on Pattern Recognition (ICPR), IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), IEEE International Conference on Systems, Man, and Cybernetics (SMC), IEEE International Conference on Image Processing (ICIP).

Email: hoangtv@hcmute.edu.vn. ORCID:  <https://orcid.org/0000-0002-1352-4853>.