Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.078756

ARTICLE

A Large-Scale Dataset for Real-Time Vehicle Detection in Vietnamese Urban Traffic Scenes

Quang Dong Nguyen Vo[1], Gia Nhu Nguyen[1] and Hoang Vu Tran[2,*]

[1]School of Computer Science and Artificial Intelligence, Duy Tan University, Danang, Vietnam
[2]The University of Danang-University of Technology and Education, Danang, Vietnam

*Corresponding Author: Hoang Vu Tran. Email: thvu@ute.udn.vn

Received: 07 January 2026; Accepted: 12 February 2026; Published: 08 May 2026

ABSTRACT: Reliable vehicle detection in urban traffic environments remains challenging, particularly for fixed-view CCTV systems deployed in Southeast Asian cities, where heterogeneous traffic composition, high traffic density, frequent occlusions, and complex visual conditions are prevalent. The absence of large-scale datasets tailored to such mixed-traffic environments poses a significant limitation to the performance and generalization capability of existing object detection models. To address this gap, this paper presents a large-scale traffic image dataset for real-time vehicle detection in Vietnamese urban environments. The proposed dataset comprises 23,364 images collected from fixed-view CCTV traffic cameras deployed across Da Nang City, a representative urban area exhibiting mixed-traffic patterns commonly observed in Southeast Asian cities. The data cover diverse temporal periods, weather conditions, and traffic density levels encountered in real-world traffic monitoring scenarios. To comprehensively characterize these conditions, over 1.1 million instances are annotated across multiple traffic-related categories, including pedestrians, bicycles, motorbikes, cars, buses, trucks, and traffic lights with explicit signal-state labels. Such fine-grained, multi-class annotations support not only object-level detection but also higher-level traffic scene analysis relevant to intelligent transportation system (ITS) applications, such as traffic flow analysis and signal control. To balance annotation accuracy and scalability, a semi-automatic labeling pipeline is employed. Initial object annotations are generated using a pretrained YOLOv11m model and subsequently refined through systematic manual verification using the CVAT platform. Comprehensive experiments are conducted under the same experimental protocol, using the same YOLOv11m architecture, comprising a pretrained baseline and a version fine-tuned on the proposed dataset with domain-specific data augmentation and optimized hyperparameter settings tailored to fixed-view CCTV conditions. Under the same evaluation setting, the pretrained YOLOv11m achieves a mean Average Precision (mAP) of 0.409; in contrast, fine-tuning on the proposed dataset improves the mAP to 0.788. These results underscore the necessity of localized, context-aware datasets such as the one presented in this work for robust real-time traffic perception in Vietnam and similar Southeast Asian urban contexts.

KEYWORDS: Deep learning; ITS; real-time vehicle detection; Vietnamese traffic dataset; traffic detection

1 Introduction

In recent decades, rapid urbanization has significantly increased the complexity of urban traffic systems. This growth has led to dynamic interactions among diverse vehicles and pedestrians within mixed-traffic environments commonly observed in many developing countries. Such complexity arises from the coexistence of cars, buses, trucks, motorcycles, bicycles, and pedestrians sharing limited road infrastructure.

These environments are typically associated with weak lane discipline and heterogeneous driving behaviors, which further complicate traffic dynamics.

Traffic complexity is further exacerbated by external factors such as time of day, weather conditions, and varying traffic densities, making real-time traffic monitoring and management a persistent challenge. In Vietnam, motorbikes dominate urban transportation and contribute to highly unstructured traffic scenes, highlighting the need for accurate and scalable traffic perception systems to enhance road safety, traffic efficiency, and smart mobility initiatives. However, conventional surveillance-based approaches often lack the adaptability and contextual awareness required to cope with such complex traffic dynamics [1–5]. These characteristics pose significant challenges for vision-based traffic perception systems, particularly those relying on fixed-view CCTV surveillance.

Recent advances in artificial intelligence and computer vision have enabled reliable real-time object detection in complex traffic environments. Deep learning-based detection frameworks, such as YOLO and Faster R-CNN, have demonstrated strong performance in recognizing and classifying traffic participants with high accuracy and computational efficiency. Nevertheless, their effectiveness strongly depends on the availability of high-quality, region-specific datasets that reflect local traffic characteristics [6–9]. Most publicly available traffic datasets are collected in Western cities or structured traffic environments. As a result, they fail to adequately capture dense motorbike flows, weak lane discipline, and persistent occlusions commonly observed under fixed-view CCTV settings. Consequently, these datasets provide limited representation of heterogeneous traffic composition, dense flows, and fixed-view CCTV surveillance conditions typical of Southeast Asian urban contexts.

To address this limitation, this paper introduces a large-scale dataset designed for real-time vehicle detection in Vietnamese urban traffic environments. The proposed dataset consists of 23,364 images captured from fixed-view CCTV traffic cameras installed at major intersections in Da Nang City, Vietnam. This urban setting is representative of mixed traffic flow with high motorbike density. The dataset spans diverse temporal periods, weather conditions, and traffic density levels commonly encountered in real-world deployments. It provides detailed annotations for multiple traffic participants, as well as traffic lights with explicit state labels. To balance annotation quality and scalability, a semi-automatic labeling pipeline is adopted. This pipeline combines initial detections generated by a pre-trained YOLOv11m model with systematic manual verification and refinement.

This work contributes a large-scale, fine-grained, and publicly available dataset tailored to Vietnamese urban traffic scenes. The dataset serves as a valuable resource for advancing research in intelligent traffic perception and smart city applications within mixed-traffic environments. Experimental evaluations show that models fine-tuned on the proposed dataset achieve improved detection performance in dense and visually complex traffic scenes. These findings highlight the limitations of general-purpose datasets and emphasize the importance of localized data for domain-specific adaptation. Although the dataset is collected in Vietnam, the traffic characteristics it captures, such as heterogeneous vehicle composition, high motorbike density, mixed traffic flow, and fixed-view CCTV surveillance, are widely observed across many Southeast Asian cities. Similar characteristics have been reported in studies conducted in Thailand, Indonesia, and Malaysia. Recent vision-based traffic research further supports this regional similarity; for example, Xu and Liu [10] demonstrate the effectiveness of fixed-camera, vision-based deep learning frameworks under heterogeneous vehicle structures and complex urban traffic conditions. While cross-country generalization is not explicitly evaluated in this work, the shared traffic characteristics suggest that the proposed dataset provides a solid foundation for intelligent transportation system (ITS) applications, including traffic surveillance, congestion analysis, and smart city development in comparable urban contexts. Fig. 1 illustrates

the overall technical roadmap of the proposed framework, covering the complete workflow from real-world CCTV data acquisition to dataset construction, model evaluation, and smart traffic applications. This framework highlights the practical and system-oriented nature of the proposed dataset for intelligent transportation research.

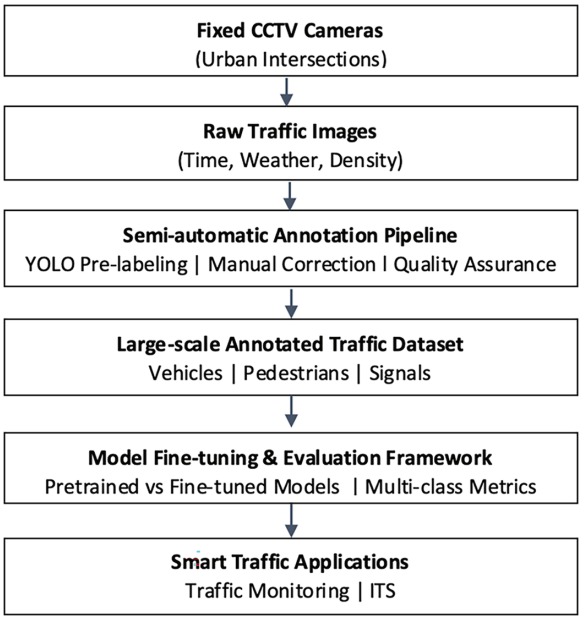


Figure 1: Overall technical roadmap of the proposed framework, from CCTV data acquisition and semi-automatic annotation to dataset construction, model evaluation, and smart traffic applications.

2 Related Work

Over the past decade, the object detection task has witnessed significant advancements, with numerous studies [11–13] leveraging deep learning-based approaches on widely adopted benchmark datasets, including UA-DETRAC and Microsoft COCO. These datasets have played a pivotal role in advancing object recognition and traffic scene understanding. However, their applicability to Southeast Asian urban environments, particularly Vietnam, remains limited due to substantial differences in traffic composition, infrastructure heterogeneity, and environmental conditions. This mismatch underscores the need for large-scale, context-specific datasets to support reliable real-time vehicle detection and ITS in the region.

The Common Objects in Context (COCO) dataset, introduced by Lin et al. [14], provides extensive annotations across diverse object categories, including vehicles and pedestrians, making it a foundational resource for training generic urban perception models. Nevertheless, COCO predominantly reflects Western traffic scenarios and does not adequately capture the characteristics of Vietnamese mixed traffic. Moreover, although traffic signals are included, the absence of explicit traffic light state annotations limits its utility for intersection-level traffic analysis and control in Vietnam.

UA-DETRAC, proposed by Wen et al. [15], offers high-quality annotations with over 1.2 million manually labeled bounding boxes and covers diverse environmental and illumination conditions, including rainy and nighttime scenes. It supports both dense and sparse traffic scenarios and provides multi-object tracking benchmarks, making it partially relevant for urban traffic analysis. However, UA-DETRAC exhibits critical limitations in Vietnamese contexts, most notably the absence of motorbike, pedestrian, and bicycle

annotations, as well as the lack of traffic light state labeling. In addition, its data are collected under more structured traffic regulations, reducing its representativeness of Vietnam's mixed and informal traffic flow.

VisDrone, developed by Zhu et al. [16], is a large-scale aerial dataset designed for traffic monitoring and surveillance, featuring detailed annotations under varied weather and lighting conditions. While effective for high-level traffic analysis, its top-down drone perspective differs substantially from fixed, street-level CCTV viewpoints commonly deployed in urban traffic monitoring. Furthermore, the lack of motorbike representation and traffic light annotations limits its applicability to Vietnamese urban traffic scenarios.

Trinh et al. [17] introduced the UIT-VinaDeveS22 dataset, which comprises 1364 CCTV images collected in Vietnam and captures diverse traffic conditions across different times of day and weather scenarios. Its local origin makes it more representative of Vietnamese traffic patterns than existing international benchmarks. However, its relatively small scale, limited number of video sources, potential data leakage due to video-level overlap, and the absence of traffic light annotations constrain its generalization capability and suitability for intersection-level analysis.

Recent object detection frameworks, including YOLOv7, YOLOv11, and Faster R-CNN, have demonstrated strong performance on standard benchmarks [18–20]. Nevertheless, their effectiveness in Vietnamese urban environments remains underexplored, largely due to the lack of large-scale, domain-representative datasets that reflect local traffic characteristics.

Beyond dataset construction, prior studies have explored methodological strategies to improve robustness in challenging traffic conditions, such as image-to-image translation for nighttime en-hancement and occlusion-aware modeling based on keypoints and spatio-temporal reasoning. Notably, Xu et al. [21] proposed a monocular framework integrating object detection and keypoint estimation to handle severe occlusion. Despite their effectiveness, such approaches remain highly dependent on representative, domain-specific datasets for training and validation, particularly in heterogeneous traffic environments.

In contrast to existing resources, this work introduces a large-scale CCTV-based dataset specifically tailored to Vietnamese urban traffic scenes. The proposed dataset provides fine-grained annotations for multiple traffic participants and explicit traffic light states, capturing mixed traffic flow, high motorbike density, and diverse environmental conditions. As such, it offers a solid foundation for robust real-time vehicle detection and the development of ITS applications in complex urban settings.

3 Dataset Construction and Annotation

3.1 Dataset Collection and Camera Configuration

The proposed dataset was collected from fixed-view CCTV traffic cameras deployed at major intersections in Da Nang City, Vietnam. Each intersection is monitored by four cameras corresponding to the incoming traffic directions. The cameras are installed at heights ranging from 6 to 10 m, with downward viewing angles ranging from 30° and 45°, enabling a comprehensive coverage of vehicle movements and interactions within the intersection area. In total, 23,364 images were extracted from continuous video streams, capturing diverse temporal, environmental, and traffic conditions, as illustrated in Fig. 2.

Video frames were extracted from the continuous CCTV streams using a fixed temporal sampling strategy to reduce redundancy while preserving traffic dynamics. Specifically, shorter sampling intervals were applied during peak traffic periods, whereas longer intervals were used under low-density conditions. All extracted frames were resized to a standardized resolution of 640 × 640 pixels to ensure compatibility with the input requirements of the YOLOv11m detection framework.

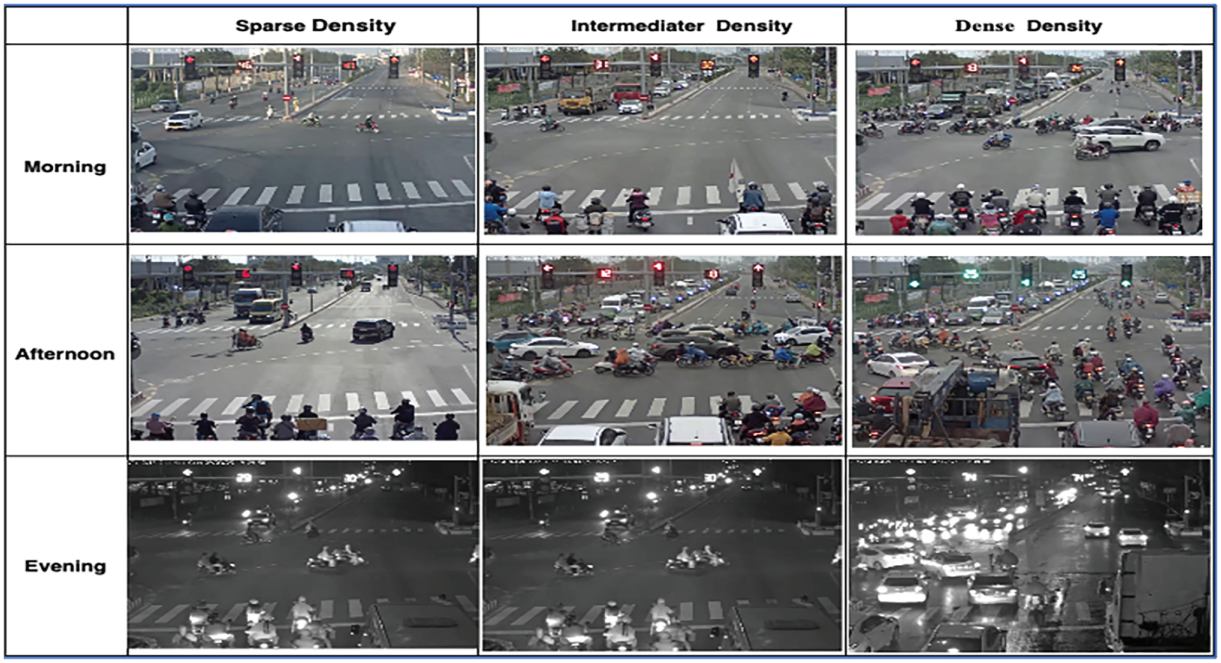


Figure 2: Representative sample frames from the Vietnamese urban traffic dataset.

3.2 Spatial and Temporal Coverage

To capture realistic variations in urban traffic conditions, the dataset was categorized along multiple spatial–temporal dimensions.

Time Variability: The dataset is divided into morning, afternoon, and evening periods to capture temporal variations in traffic flow and illumination conditions in urban environments.

Weather Scenarios: Traffic scenes are further categorized into fine, sunny, and rainy conditions, reflecting common weather variations that affect visibility, reflections, and motion blur.

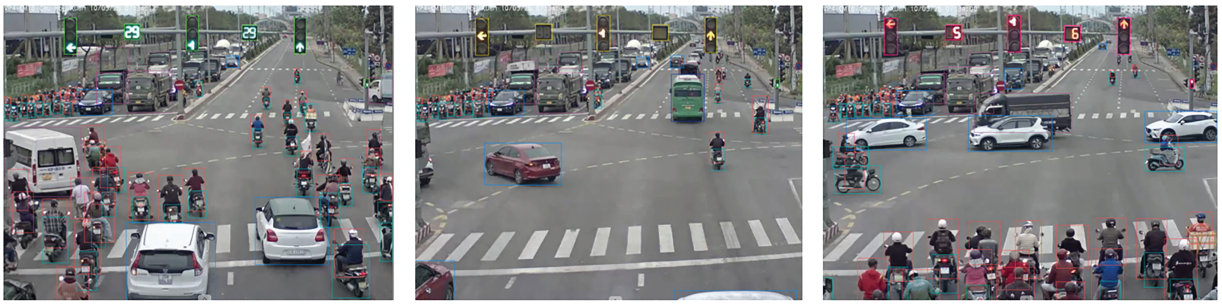
Traffic Density: It is classified into sparse and dense levels based on vehicle counts per scene, enabling evaluation of detection performance under low- and high-congestion conditions with varying degrees of occlusion.

3.3 Annotation Categories and Dataset Partitioning

All samples were annotated with bounding boxes and class labels covering seven traffic categories: pedestrian, bicycle, motorcycle, car, bus, truck, and traffic light. Traffic lights were additionally annotated by signal state to support traffic signal analysis within ITS. As shown in Fig. 3, traffic-light annotations include active states (green, yellow, and red) and the corresponding permitted movement directions, which are critical for intersection-level traffic analysis. For multi-head traffic signals, each clearly identifiable signal head was treated as an independent instance, while arrow-based signals were labeled according to the illuminated direction and inactive arrows were ignored.

In scenes containing multiple traffic signal groups, annotations were limited to signal heads relevant to the primary traffic flow, as determined by lane orientation and intersection geometry. To ensure annotation reliability, traffic-light instances were labeled only when both signal state and movement direction were unambiguous under fixed-camera viewpoints; instances affected by severe occlusion, motion blur, glare, or adverse weather conditions were excluded. Although explicit visibility-level annotations are not included in

the current release, this design choice prioritizes label accuracy and consistency, with potential extensions considered in future work.



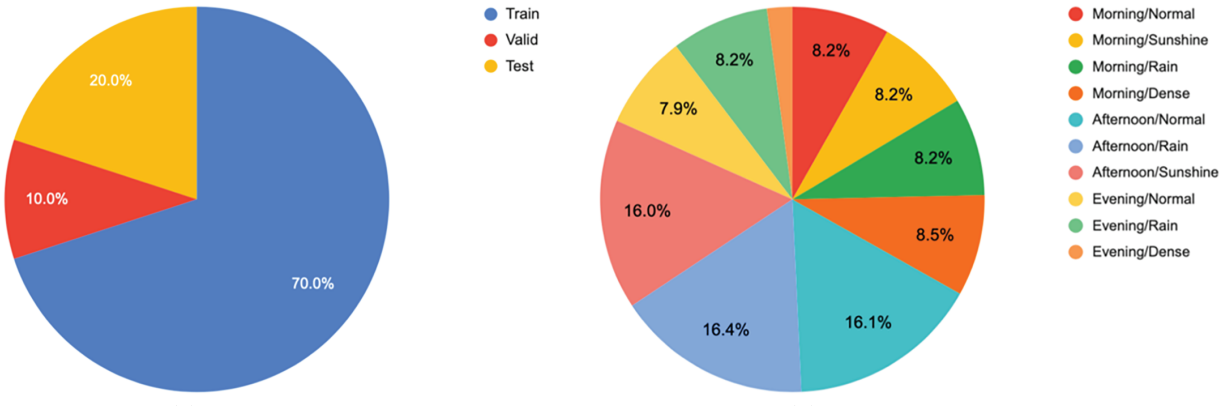
(a) (b) (c)

Figure 3: Representative examples of traffic-light annotations in the proposed dataset. (**a**) Illustrates a green traffic signal, including the remaining movement time and the corresponding permitted movement direction. (**b**) Shows a yellow traffic signal. (**c**) Presents a red traffic signal displaying a countdown timer indicating the remaining waiting time.

The dataset was partitioned into training (70%), validation (10%), and test (20%) subsets using a video-level split strategy to prevent temporal leakage. Dataset statistics are summarized in Table 1 and illustrated in Fig. 4.

Table 1: Distribution of dataset samples across training, validation, and test sets.

Subset	Sample Count
Training Set (70%)	16,353
Validation Set (10%)	2341
Test Set (20%)	4670
Total	**23,364**



(a) (b)

Figure 4: Pie charts illustrating dataset distribution. (**a**) Proportional breakdown of samples across training, validation, and test subsets. (**b**) Distribution of samples across temporal conditions, weather scenarios, and traffic-density levels.

3.4 Annotation Workflow

To construct a high-quality dataset for real-time vehicle detection in Vietnamese urban environments, we designed a semi-automatic annotation pipeline that integrates modern deep learning-based models with human-in-the-loop refinement. This hybrid strategy balances annotation efficiency with the accuracy required for downstream computer vision research. The overall annotation workflow is illustrated in Fig. 5.

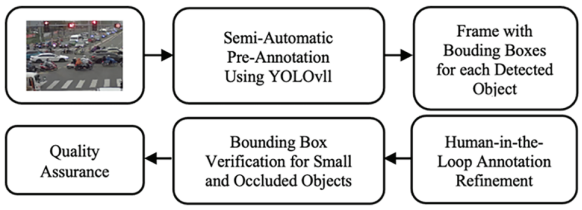


Figure 5: Overview of the proposed semi-automatic annotation pipeline.

3.4.1 Semi-Automatic Pre-Annotation Using YOLOv11m

The YOLOv11m model [22] was adopted for automatic pre-annotation. As a recent advancement in the Ultralytics YOLO family, YOLOv11m incorporates improvements in backbone architecture, feature aggregation, and training optimization (Fig. 6). These enhancements provide a favorable accuracy–speed trade-off, making the model suitable for dense urban traffic scenes with small and heavily occluded objects.

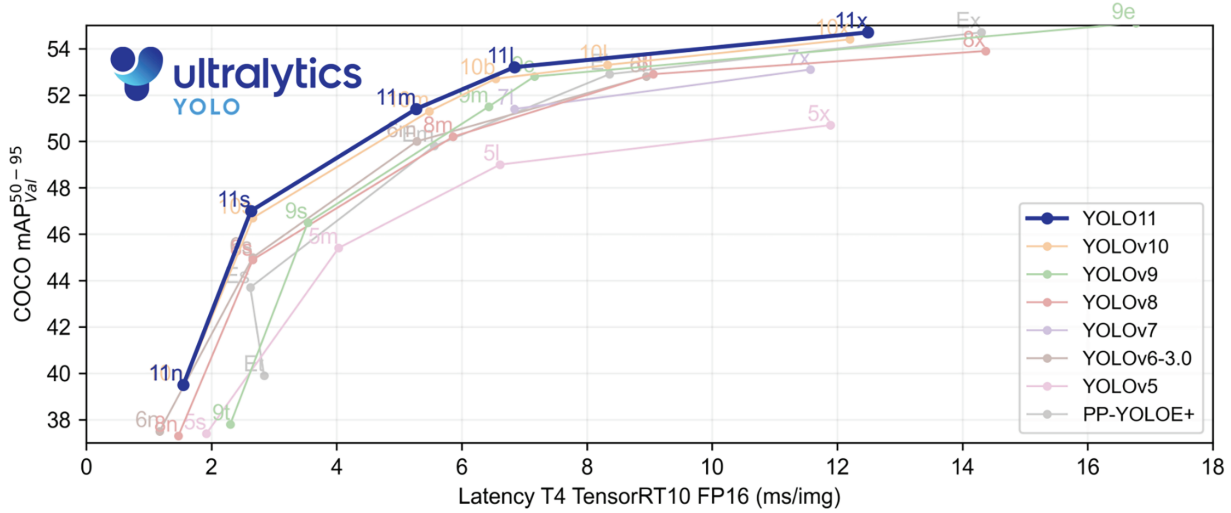


Figure 6: Comparison of representative YOLO model variants (adapted from [23]).

A total of 23,364 CCTV images were processed in a GPU-accelerated Google Colab environment and resized to 640 × 640 pixels. Initial annotations were generated using a YOLOv11m model pretrained on the COCO dataset and exported in YOLO format with associated model metadata to support reproducibility. This automated pre-labeling step reduced manual annotation effort and provided a consistent baseline for subsequent refinement.

3.4.2 Human-in-the-Loop Annotation Refinement

Automated annotations were further refined using a human-in-the-loop procedure to mitigate errors arising from domain discrepancies between Vietnamese urban traffic scenes and COCO-style datasets. Manual refinement was conducted using the CVAT platform, following standardized annotation guidelines that define class taxonomies, occlusion handling, and minimum bounding-box requirements (see Fig. 7). Annotators were trained to correct localization inaccuracies, supplement missed detections—particularly for small or occluded objects, resolve class ambiguities (e.g., bus vs. truck), and remove false positives caused by background clutter.

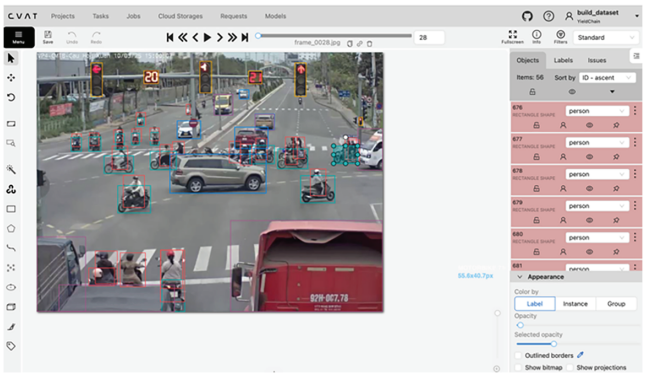


Figure 7: CVAT annotation interface used for manual refinement of object annotations.

3.4.3 Bounding-Box Verification for Small and Occluded Objects

To better capture heavily occluded, partially visible, and small-scale objects, an additional verification pass was performed. This review focused on sub-20-pixel instances, vehicles in highly congested intersection areas, and objects partially occluded by other traffic participants or scene elements. Such targeted refinement enhances the representation of challenging cases, thereby improving the dataset's suitability for training robust real-world detection models.

3.4.4 Quality Assurance and Consistency Checks

To ensure dataset reliability, a multi-stage quality control procedure was employed (Fig. 8 presents examples after refinement). A subset of 10% of the annotations was cross-checked by an independent annotator. Inter-annotator agreement was evaluated using IoU, with cases below 0.5 subject to additional review. Further consistency checks enforced uniform bounding-box tightness, object boundaries, and class definitions, while random sampling was conducted to identify potential systematic annotation biases.

In addition, all annotators underwent a structured training process prior to large-scale annotation. This training included detailed annotation guidelines defining class taxonomies, occlusion handling, and minimum bounding-box requirements, followed by pilot annotation sessions with expert feedback. To further reduce systematic bias, challenging samples—such as small objects, heavily occluded vehicles, and dense traffic scenes—were subject to targeted re-inspection. Annotation reliability was evaluated using both class-level agreement, measured by Cohen's Kappa, and box-level consistency based on IoU statistics. These quantitative checks complemented the qualitative review process and helped maintain consistent labeling across different annotators.

Figure 8: Sample frames after applying the proposed semi-automatic labeling approach.

4 Experimental Setup and Model Training

4.1 Benchmark Protocol and Dataset Usage

To evaluate the practical utility of the proposed dataset, benchmark experiments were conducted using the YOLOv11m object detection model. A video-level data partitioning strategy was adopted to prevent temporal overlap and ensure an unbiased evaluation.

The distribution of annotated instances is reported in Table 2, while Fig. 9 illustrates the dataset composition across different contextual factors, including time of day, weather conditions, and traffic density. This experimental setting facilitates a systematic assessment of detection performance under diverse urban traffic scenarios, supporting applications such as multi-class vehicle detection and traffic light state recognition.

4.2 Model Configuration and Training Strategy

The YOLOv11m model was selected as the benchmark detector due to its favorable accuracy–speed trade-off and suitability for real-time urban traffic surveillance. Two configurations were evaluated: (i) a pretrained model initialized with COCO weights, and (ii) a fine-tuned model trained on the proposed dataset. All images were uniformly resized to 640 × 640 pixels throughout pre-annotation, training, and evaluation to ensure pipeline consistency.

Table 2: Distribution of annotated instances across object classes.

Class	Train	Validation	Test	Total
Person	299,929	42,394	84,883	427,206
Bicycle	651	104	151	906
Motorbike	104,669	14,922	29,697	149,288
Car	223,380	31,658	63,784	318,822
Bus	4975	716	1441	7132
Truck	40,648	5961	11,684	58,293
Traffic Light	109,564	15,780	31,479	156,823
Total	**783,816**	**111,535**	**223,119**	**1,118,470**

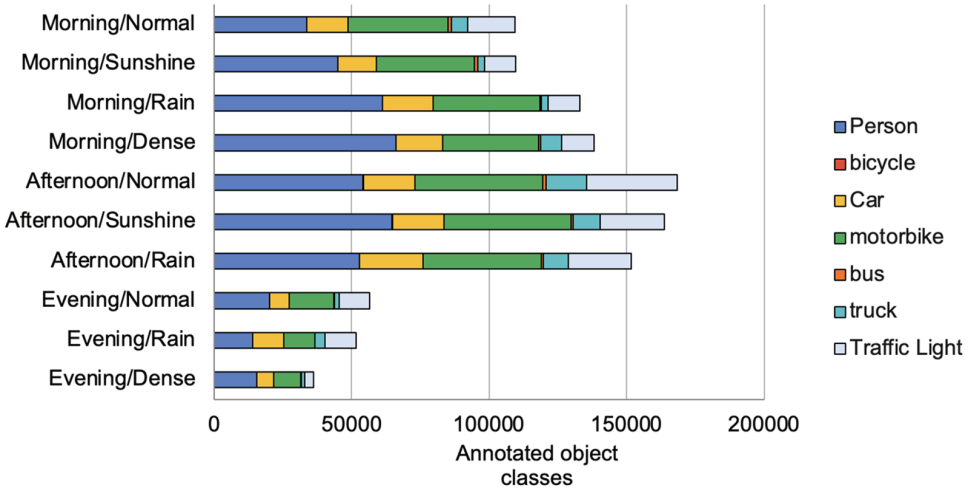


Figure 9: Dataset statistics categorized by time of day, weather conditions, and traffic-density levels.

To improve robustness and generalization under complex traffic conditions, a data augmentation strategy and corresponding hyperparameter configuration were incorporated into the YOLO training framework. The augmentation design was tailored specifically for fixed-view CCTV traffic cameras, ensuring that all transformations remained physically plausible while introducing sufficient variability for effective learning.

The augmentation pipeline integrates three categories of transformations: photometric, geometric, and composition-based augmentations. Photometric augmentations (Fig. 10) were applied to simulate real-world illumination variations commonly observed in urban traffic scenes, including shadows, nighttime lighting, and headlight glare. These include controlled adjustments of hue (0.02–0.04), saturation (0.5–0.8), and brightness (0.3–0.6).

Geometric augmentations were intentionally restricted to preserve realistic CCTV perspectives. Only minor translation (0.05–0.1), scaling (0–0.5), and limited rotation (up to 45°) were applied, reflecting the fixed-camera setup and avoiding unrealistic distortions. In addition, mosaic and copy-paste augmentations (Fig. 11) were employed to increase scene complexity and enhance robustness in dense and occlusion-heavy traffic scenarios, using four-image mosaic blending (0.5–1) and limited vehicle instance copy-paste (0–0.2).

(a) (b) (c) (d)

Figure 10: Photometric augmentations and geometric augmentations training method. (**a**) Original. (**b**) Horizontal Flip. (**c**) Brightness(HSV_v). (**d**) Saturation(HSV_s).

(a) (b) (c)

Figure 11: Mosaic augmentation samples used in the training process. (**a**) Sample 1. (**b**) Sample 2. (**c**) Sample 3.

All augmentation operations were integrated directly into the YOLO training pipeline using a predefined hyperparameter search space. The model was trained for 50 epochs with a batch size of 32, using the AdamW optimizer, an initial learning rate of 0.01, and standard YOLO weight decay parameter grouping. The complete augmentation configuration and hyperparameter settings are summarized in Table 3. By constraining augmentation intensities to match CCTV imaging characteristics, the training process achieves stable optimization while improving robustness across diverse environmental conditions and traffic congestion levels.

Table 3: Summarizes the key augmentation techniques and hyperparameter settings.

Argument	Type	Default	Proposed Values	Hyperparameter Tuning (Best Parameter)	Range
hsv_h	Float	0.015	0.02–0.04	0.021	0.0–1.0
hsv_s	Float	0.7	0.5–0.8	0.6	0.0–1.0
hsv_v	Float	0.4	0.3–0.6	0.38	0.0–1.0
Degrees	Float	0	0–45	45	0.0–180
Translate	Float	0.1	0.05–0.1	0.08	0.0–1.0
Scale	Float	0.5	0–0.5	0.1	≥ 0.0
Mosaic	Float	1	0.5–1	0.7	0.0–1.0
Copy_paste	Float	0	0–0.2	0.1	0.0–1.0

4.3 Benchmark Evaluation

4.3.1 Assessment Metric

To assess the effectiveness of the proposed dataset, benchmark experiments were conducted using the YOLOv11m object detection model. The mean Average Precision (mAP) serves as a standard measure for evaluating model effectiveness in terms of accuracy. mAP is computed as the mean of the Average Precision (AP) values across all object classes, where AP is evaluated at specific Intersection over Union (IoU) thresholds, including AP50 and AP75

- IoU quantifies the overlap between a predicted bounding box and the corresponding ground-truth annotation, as defined in Eq. (1).

$$IoU = \frac{Area\ of\ Overlap}{Area\ of\ Union} \tag{1}$$

- Precision (P) reflects the proportion of predicted bounding boxes that correspond to correct detections, as defined in Eq. (2).

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

where TP (True Positive): correctly detected objects and FP (False Positive): incorrect detections.

- Recall (R) measures the proportion of ground-truth objects that are successfully detected, as defined in Eq. (3).

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

where *FN* (False Negative): missed ground-truth objects.

- *AP* summarizes the precision–recall curve into a single scalar value, as defined in Eq. (4).

$$AP = \int_0^1 P(R)dR \tag{4}$$

where *PR* is precision–recall curve. *AP* is usually reported at a fixed IoU threshold, such as: $AP50$: $IoU \geq 0.50$ and $AP75$: $IoU \geq 0.75$

- mAP is computed as the mean of AP scores across all object classes, as defined in Eq. (5).

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{5}$$

where N is the total number of object classes, AP_i is the Average Precision for class i.

4.3.2 Experimental Environment Setup

The experiments were conducted on a Vast.ai cloud instance equipped with an NVIDIA GeForce RTX 3090 GPU, as detailed in Table 4.

Table 4: Hardware configuration of the experimental environment.

Vast.ai Cloud	Values
GPU Name	1x RTX 3090, Intel Xeon E5-2683 v3
CPU	56 CPU/64GB
GPU (vRAM)	24 GB
GPU Architecture	Ampere architecture
CUDA Compute Capability	8.6 (Max CUDA 13.0)
CUDA Version	12.x
Shared Memory	Up to 100 KB/block

This environment was used to train and evaluate both the pretrained and fine-tuned YOLOv11m models under a unified hardware and software configuration, ensuring fair and reproducible comparisons. The setup supported data augmentation and domain-specific hyperparameter tuning required for effective training. Inference speed was evaluated at an input resolution of 640 × 640 pixels. The fine-tuned YOLOv11m model achieved an average inference speed of approximately 35 frames per second (FPS) for single-stream CCTV input, including preprocessing and postprocessing overhead, satisfying the real-time requirements of urban traffic monitoring and ITS.

4.3.3 Experimental Results

The objective of the experimental evaluation in this work is to validate the quality, diversity, and practical relevance of the proposed dataset rather than to compare detection architectures or achieve state-of-the-art performance. Accordingly, a single strong and widely adopted real-time detector, YOLOv11m, is employed as a representative baseline to evaluate object detection performance across all annotated classes. To examine robustness under different conditions, the dataset was stratified by time of day, weather, and traffic density. Both pretrained and fine-tuned models were evaluated on a dedicated test set of 4670 images to ensure a fair comparison. As shown in Figs. 12 and 13, domain-specific training leads to consistent performance improvements across diverse urban traffic scenarios.

4.4 Results Analysis and Discussion

4.4.1 Overall Performance Analysis

Across all object classes, the fine-tuned model consistently outperforms the pretrained YOLOv11m baseline. The baseline model, trained on the general-purpose COCO dataset, shows clear limitations in Vietnamese urban traffic scenes, especially under high traffic density, small object scales, and frequent occlusions. Fine-tuning on the proposed domain-specific dataset leads to improved localization accuracy and more reliable discrimination among visually similar vehicle categories. Qualitative comparisons in Figs. 12 and 13 further illustrate the performance gap between the pretrained and fine-tuned detectors.

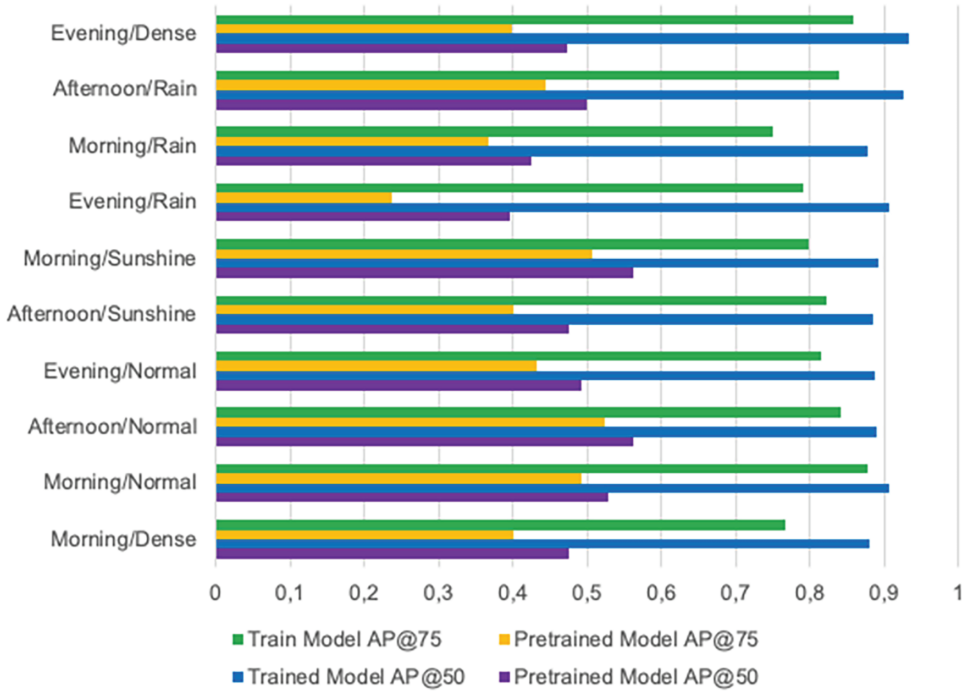


Figure 12: Comparison of Average Precision at IoU thresholds of 0.50 (AP50) and 0.75 (AP75) between the pretrained and fine-tuned models.

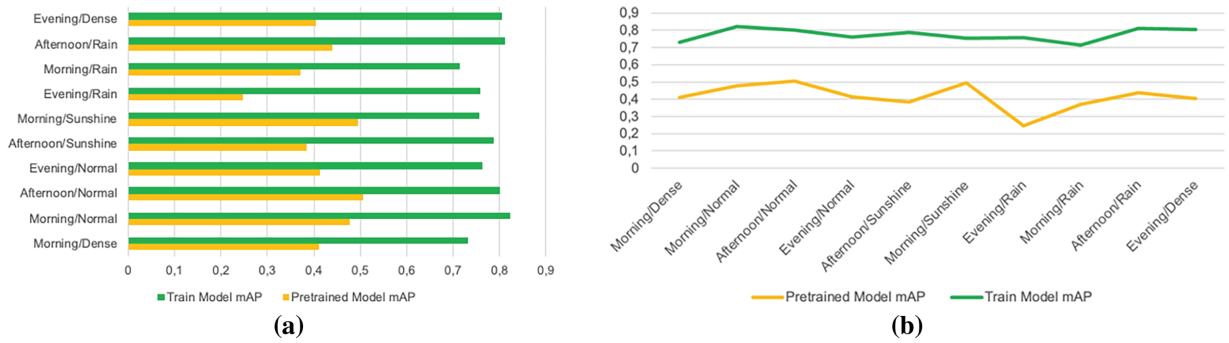


Figure 13: Comparison of mean Average Precision (mAP) between the pretrained and fine-tuned models. (**a**) mAP comparison illustrated using a clustered bar chart. (**b**) mAP trends of the pretrained and fine-tuned models shown using a line chart.

4.4.2 Per-Class Detection Performance Analysis

To complement the overall mAP evaluation, we analyze detection performance at the per-class level to reveal class-specific strengths and limitations, particularly for small and less frequent traffic participants that are critical in urban environments. Table 5 reports Average Precision (AP) at IoU thresholds of 0.50 and 0.75 for each class, comparing the pretrained and fine-tuned YOLOv11m models.

For the pretrained model, performance varies substantially across classes. Large and visually distinctive objects such as cars achieve relatively high AP (0.823 at AP50), whereas smaller or infrequent classes—including bicycles (0.132 AP50), bus (0.305 AP50), and traffic lights (0.321 AP50)—exhibit severe

performance degradation. This highlights the limited transferability of COCO-pretrained detectors to Vietnamese mixed-traffic scenes characterized by small object scales and frequent occlusion.

Table 5: Per-class AP comparison between pretrained and fine-tuned models.

Class	AP50		AP75	
	Pretrained	Fine-Tuned	Pretrained	Fine-Tuned
Person	0.559	0.922	0.377	0.745
Bicycle	0.132	0.681	0.09	0.597
Motorbike	0.461	0.929	0.245	0.806
Car	0.823	0.975	0.731	0.958
Bus	0.305	0.863	0.283	0.835
Truck	0.61	0.953	0.499	0.917
Traffic light	0.321	0.992	0.27	0.989
Total/Average	**0.459**	**0.902**	**0.356**	**0.835**

After fine-tuning on the proposed dataset, substantial improvements are observed across all categories. Notably, bicycle AP50 increases to 0.681, pedestrian AP50 to 0.922, and traffic light detection reaches 0.992 AP50. Despite these gains, stricter localization at AP75 remains more challenging for small or occluded objects, reflecting the combined effects of class imbalance, object scale, and dense traffic interactions.

4.4.3 Condition-Based Performance Analysis

To further understand how these class-wise behaviors manifest under realistic surveillance conditions, we analyze detection performance stratified by time of day, weather conditions, and traffic density.

Time of Day: Table 6 shows that the pretrained model degrades notably in evening scenes due to illumination changes, particularly for small and infrequent objects. After fine-tuning, performance becomes more stable across all periods, with substantial gains for bicycles and traffic lights, indicating improved robustness under low-light conditions.

Table 6: Class-wise mAP@ [50–95] performance by time of day.

Class	Pretrained			Fine-Tuned		
	Morning	Afternoon	Evening	Morning	Afternoon	Evening
Person	0.3205	0.439667	0.294333	0.7065	0.768	0.667667
Bicycle	0.115	0.121667	0.0813333	0.474	0.563667	0.435333
Motorbike	0.25375	0.271667	0.154333	0.74425	0.750333	0.66
Car	0.673	0.688667	0.550667	0.91625	0.935667	0.888667
Bus	0.35925	0.252333	0.117333	0.85025	0.758	0.764
Truck	0.25	0.523333	0.217333	0.86475	0.883667	0.854333
Traffic light	0.12525	0.357	0.250667	0.41925	0.984333	0.959

Weather conditions: Table 7 indicates that the pretrained model degrades significantly in rainy conditions. especially for small or low-contrast objects due to blur and reflections. Fine-tuning yields consistently high performance across all weather types, with notable gains in rain, demonstrating improved robustness to weather-related visual challenges.

Table 7: Class-wise mAP@ [50–95] performance by weather conditions.

Class	Pretrained			Fine-Tuned		
	Fine	Rain	Sunshine	Fine	Rain	Sunshine
Person	0.386	0.338333	0.382	0.768667	0.689667	0.7375
Bicycle	0.0753333	0.123	0.074	0.586667	0.333333	0.4695
Motorbike	0.250333	0.178667	0.239	0.739667	0.664	0.7475
Car	0.633	0.574333	0.6895	0.909667	0.89	0.933
Bus	0.204	0.278667	0.2695	0.74	0.835	0.7905
Truck	0.274333	0.189667	0.289	0.873333	0.868	0.8515
Traffic light	0.347667	0.209	0.012	0.984	0.889333	0.496

Traffic Density: Table 8 shows that the pretrained model degrades markedly in dense traffic, especially for small and occluded objects. After fine-tuning, performance improves substantially, with the largest gains in congested scenes, indicating enhanced robustness to occlusion and object overlap in dense urban traffic. Overall, the condition-based analysis demonstrates the robustness of the proposed dataset while revealing its remaining limitations in challenging urban scenarios.

Table 8: Class-wise mAP@ [50–95] performance by traffic density.

Class	Pretrained		Fine-Tuned	
	Sparse	Dense	Sparse	Dense
Person	0.386	0.2735	0.768667	0.6415
Bicycle	0.0753333	0.163	0.586667	0.597
Motorbike	0.250333	0.264	0.739667	0.751
Car	0.633	0.7045	0.909667	0.9365
Bus	0.204	0.2795	0.74	0.8305
Truck	0.452333	0.3475	0.873333	0.873
Traffic light	0.347667	0.327	0.984	0.4475

4.4.4 Summary of Experimental Findings

The experimental results indicate that the proposed large-scale, domain-specific dataset plays a central role in the observed performance improvements over the pretrained YOLOv11m model. Fine-tuning on this dataset yields consistent gains across object classes and evaluation scenarios, highlighting the limitations of general-purpose datasets when applied to mixed and highly congested urban traffic environments. Improved robustness under challenging conditions, including nighttime scenes, rainy weather, and dense traffic, is observed in the experimental results. This robustness is primarily attributed to the dataset's comprehensive coverage of real-world scenarios and the careful refinement of annotations for small and occluded objects. Standard data augmentation techniques and commonly adopted hyperparameter settings were applied to

ensure stable training. Isolated ablation studies were not conducted, as the primary focus of this work is dataset construction rather than training strategy optimization. Similarly, aspects related to deployment efficiency, hardware-specific optimization, and large-scale multi-stream scalability are beyond the scope of this experimental study, as they depend strongly on application-specific system configurations. Overall, the proposed dataset provides a reliable benchmark for real-time vehicle detection and establishes a solid foundation for future ITS research in Southeast Asian urban contexts.

5 Conclusion

In this paper, we addressed the limitations of existing object detection datasets in representing the complexity of Vietnamese urban traffic. This complexity is characterized by mixed vehicle types, dense traffic flows, and challenging environmental conditions. We introduced a new large-scale dataset comprising 23,364 images collected from fixed-view CCTV traffic cameras at major intersections in Da Nang City, Vietnam. The dataset is designed to support real-time vehicle detection in urban environments and covers diverse temporal conditions, weather scenarios, and traffic density levels. It is extensively annotated to include key traffic participants, particularly motorbikes, as well as fine-grained traffic light states essential for intelligent traffic monitoring systems.

To ensure annotation accuracy and scalability, we employed a semi-automatic annotation pipeline combining YOLOv11m-based pre-annotation with manual verification and refinement using CVAT. This hybrid workflow significantly reduced labeling effort while preserving high-quality annotations, especially for small, densely packed, and partially occluded objects. The final dataset contains over 1.1 million labeled instances with consistent and reliable annotations.

Experimental evaluations using the pretrained YOLOv11m baseline and a model fine-tuned on the proposed dataset show that general-purpose pretrained models perform poorly when directly applied to Vietnamese traffic scenes. In contrast, fine-tuning on the proposed dataset yields substantial im-provements across all object classes and evaluation conditions, including nighttime, rainy weather, and dense traffic. These results highlight the importance of localized, context-aware datasets for robust traffic perception.

Although standard training enhancements such as data augmentation and hyperparameter tuning were applied, the primary contribution of this work lies in the dataset rather than training strategy optimization. Overall, the proposed dataset and annotation framework provide a strong benchmark for AI-driven traffic perception in Vietnam. We acknowledge that all data were collected from Da Nang City, which may introduce a degree of geographical bias. However, Da Nang represents a typical Vietnamese urban environment characterized by mixed traffic flow, high motorbike density, and diverse weather conditions. As such, the dataset captures fundamental traffic characteristics common across many Vietnamese and Southeast Asian cities. Future work will focus on expanding the dataset to additional locations, exploring advanced detection architectures, and investigating techniques such as domain adaptation, self-supervised learning, and multi-camera fusion.

Acknowledgement: Not applicable.

Funding Statement: This work was supported in part by the Ministry of Science and Technology (MOST), The University of Danang—University of Technology and Education, and the School of Computer Science, Duy Tan University, Da Nang City, Vietnam.

Author Contributions: Quang Dong Nguyen Vo: Conceptualization, methodology, software, investigation, writing-original. Gia Nhu Nguyen: Conceptualization, investigation, writing-review and editing. Hoang Vu Tran:

Conceptualization, investigation, writing-review and editing, supervision. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The dataset is available for non-commercial research use upon reasonable request to the corresponding author and is provided under controlled access to ensure ethical and privacy compliance. Commercial use requires explicit permission from the authors.

Ethics Approval: The dataset was collected from publicly operated urban traffic CCTV systems and used solely for research purposes. Personally identifiable information, including faces and license plates, was anonymized when visible. The dataset contains only bounding-box annotations without identity tracking or personal metadata and complies with applicable Vietnamese data protection regulations.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Majstorović Ž, Tišljarić L, Ivanjko E, Carić T. Urban traffic signal control under mixed traffic flows: literature review. Appl Sci. 2023;13(7):4484. doi:10.3390/app13074484.
2. Huu DN, Ngoc VN. Analysis study of current transportation status in Vietnam's urban traffic and the transition to electric two-wheelers mobility. Sustainability. 2021;13(10):5577. doi:10.3390/su13105577.
3. Nguyen T. Crowd-AI sensing based traffic analysis for Ho Chi Minh City planning simulation. Dir Comput Inf Sci Eng. 2020;20(2025234):25234.
4. Buch N, Velastin SA, Orwell J. A review of computer vision techniques for the analysis of urban traffic. IEEE Trans Intell Transport Syst. 2011;12(3):920–39. doi:10.1109/tits.2011.2119372.
5. Tamizh Selvi A, Domilin Shyni I, Rexiline Sheeba I, Jayasudha FV, Sanju IMS. Real-time traffic monitoring and analysis using YOLO-based object detection. In: Proceedings of the 2025 International Conference on Next Generation Computing Systems (ICNGCS); 2025 Aug 21–22; Coimbatore, India. p. 1–6. doi:10.1109/ICNGCS64900.2025.11183064.
6. Aswini N, Hegde R. Real time traffic monitoring using YOLO V5. In: Proceedings of the 2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA); 2023 Aug 3–5; Coimbatore, India. p. 572–6. doi:10.1109/ICIRCA57980.2023.10220896.
7. Maity M, Banerjee S, Sinha Chaudhuri S. Faster R-CNN and YOLO based vehicle detection: a survey. In: Proceedings of the 2021 5th International Conference on Computing Methodologies and Communication (ICCMC); 2021 Apr 8–10; Erode, India. p. 1442–7. doi:10.1109/ICCMC51019.2021.9418274.
8. Abbasi M, Shahraki A, Taherkordi A. Deep learning for network traffic monitoring and analysis (NTMA): a survey. Comput Commun. 2021;170(3):19–41. doi:10.1016/j.comcom.2021.01.021.
9. Almukhalfi H, Noor A, Noor TH. Traffic management approaches using machine learning and deep learning techniques: a survey. Eng Appl Artif Intell. 2024;133(3):108147. doi:10.1016/j.engappai.2024.108147.
10. Xu B, Liu C. Keypoint detection-based and multi-deep learning model integrated method for identifying vehicle axle load spatial-temporal distribution. Adv Eng Inform. 2024;62:102688. doi:10.1016/j.aei.2024.102688.
11. He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV); 2017 Oct 22–29; Venice, Italy. p. 2980–8. doi:10.1109/ICCV.2017.322.
12. Ge Z, Liu S, Wang F, Li Z, Sun J. Yolox: exceeding yolo series in 2021. arXiv:2107.08430.2107. 2021. doi:10.48550/arXiv.2107.08430.
13. Lyu C, Zhang W, Huang H, Zhou Y, Wang Y, Liu Y, et al. RTMDet: an empirical study of designing real-time object detectors. arXiv: 2212.07784. 2022.
14. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D. Microsoft COCO: common objects in context. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T, editors. Computer VIsion—ECCV 2014. ECCV 2014. Lecture notes in computer science. Vol. 8693. Cham, Switzerland: Springer; 2014. p. 740–55. doi:10.1007/978-3-319-10602-1_48.
15. Wen L, Du D, Cai Z, Lei Z, Chang MC, Qi H, et al. UA-DETRAC: a new benchmark and protocol for multi-object detection and tracking. Comput Vis Image Underst. 2020;193(9):102907. doi:10.1016/j.cviu.2020.102907.

16. Zhu P, Wen L, Du D, Bian X, Fan H, Hu Q, et al. Detection and tracking meet drones challenge. arXiv:2001.06303. 2020.
17. Trinh T, Nguyen K. A Vietnamese benchmark for vehicle detection and real-time empirical evaluation. Tho Univ J Sci. 2022;14(3):45–52. doi:10.22144/ctu.jen.2022.042.
18. Sapkota R, Calero MF, Qureshi R, Badgujar C, Nepal U, Poulose A, et al. YOLO advances to its genesis: a decadal and comprehensive review of the You Only Look Once (YOLO) series. arXiv:2406.19407. 2024. doi:10.48550/arXiv.2406.19407.
19. Sun Y, Li Y, Li S, Duan Z, Ning H, Zhang Y. PBA-YOLOv7: an object detection method based on an improved YOLOv7 network. Appl Sci. 2023;13(18):10436. doi:10.3390/app131810436.
20. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell. 2017;39(6):1137–49. doi:10.1109/tpami.2016.2577031.
21. Xu B, Liu X, Feng G, Liu C. A monocular-based framework for accurate identification of spatial-temporal distribution of vehicle wheel loads under occlusion scenarios. Eng Appl Artif Intell. 2024;133(1):107972. doi:10.1016/j.engappai.2024.107972.
22. Jegham N, Koh CY, Abdelatti M, Hendawi A. Evaluating the evolution of YOLO (You Only Look Once) models: a comprehensive benchmark study of YOLO11 and its predecessors. arXiv:2411.00201. 2024.
23. Ultralytics. Ultralytics yolov11; 2025 [cited 2026 Feb 19]. Available from: https://www.ultralytics.com/blog/comparing-ultralytics-yolo11-vs-previous-yolo-models.