

Deep Learning and Computer Vision-Based Modeling for a Real-Time Traffic Congestion Monitoring Framework

Neil Dustin Benedict A. Agner[✉], Angelo B. Dela Cruz[✉],

Aster Benedict A. Mangabat[✉], Raphael B. Alampay[✉], Patricia Angela R. Abu[✉]

Ateneo Laboratory for Intelligent, Visual Environments, Ateneo de Manila University, Philippines

Abstract—As intelligent traffic systems evolve to manage complex urban mobility, conventional congestion estimation techniques, such as the time-windowed Volume-to-Capacity (V/C) ratio, fail to capture the real-time traffic situation. Because these methods rely on vehicles crossing a specific point, they often fail to register stopped cars, creating a ‘zero flow’ paradox where complete gridlock is incorrectly categorized as an empty, free-flowing road. To address these limitations, this study proposes an end-to-end computer vision framework utilizing the Real-Time DEtection TRansformer (RT-DETR) model to quantify vehicle volume and immediately classify a roadway’s Level of Service (LOS) through spatial density. In processing video feeds from the Ateneo de Manila University Campus Safety and Mobility Office, the framework calculates an instantaneous, frame-level Spatial Density Index across six distinct vehicle classes. Results show that the proposed SDI framework outperforms the traditional time-windowed method in estimating traffic congestion and reduces percentage error margins to between -0.22% and 2.85%. The proposed framework successfully resolves the zero flow paradox by analyzing frame-wide congestion rather than fixed line crossings prescribed in the time-windowed method. It also offers a highly responsive and robust alternative that detects immediate spikes in vehicle density, which makes it potentially suitable for integration into intelligent traffic systems.

Keywords—Traffic congestion estimation; level of service; volume-to-capacity ratio; RT-DETR; real-time object detection

I. INTRODUCTION

Traffic congestion remains a critical challenge in many urban areas around the world. It affects the efficiency of transportation, quality of life, and economic productivity [1]. Due to growing urbanization, there is a need for intelligent transport systems that are capable of monitoring and analyzing traffic conditions in real-time [2]. Computer vision (CV), a field of artificial intelligence that is focused on enabling machines to extract meaningful information from visual data, has emerged as an effective solution to address these challenges [3].

Traffic congestion estimation plays a key role in managing and optimizing transport systems. It often involves analyzing vehicle volume and capacity to infer congestion levels [4]. Traditional methods for traffic congestion estimation utilize various technologies for collecting traffic information, such as inductive loop detectors, Light Detection and Ranging (LiDAR) sensors, and Global Positioning System (GPS)-based tracking [5], [6], [7]. However, these methods have drawbacks. GPS-based systems have limited coverage as they rely on vehi-

cles being equipped with tracking devices [6]. Loop detectors require road installation and maintenance, which can be costly and disruptive [8]. LiDAR sensors’ performance can be heavily affected by different weather conditions as they interfere with the laser signals and reduce accuracy [9].

With the increase in deployment of surveillance cameras across urban centers, their integration into intelligent transport system provides a highly beneficial method to address traffic congestion in real-world scenarios. Deep learning models in computer vision have shown significant promise in addressing challenges associated with traffic congestion [10]. Among its many applications, object detection, a technique for identifying and locating objects within images, has shown potential in enhancing current intelligent transport systems and providing a non-intrusive mechanism to quantify traffic flow.

Despite these capabilities, current CV approaches often fall short of providing a complete end-to-end analysis for traffic management. Existing frameworks typically output raw vehicle counts which require further manual processing to derive standard congestion measures [11]. Currently, congestion detection is treated as a visual classification problem through the discretization of the degrees of congestion [12]. Furthermore, few studies have shown a framework that would transform vehicle detection counts, obtained via CV-based solutions, into a measure of traffic congestion.

This study addresses this limitation by proposing an end-to-end framework that uses object detection to quantify vehicle volume and classify the Level of Service (LOS) of a roadway. While baseline evaluations utilize traditional time-windowed V/C ratios, this study introduces a Spatial Density Index (SDI) to address their inherent temporal lag. The proposed SDI provides an immediate classification of LOS through frame-level evaluation of spatial occupancy of a road by vehicles.

II. REVIEW OF RELATED LITERATURE AND STUDIES

A. Traffic Congestion Estimation

Traffic congestion is defined as a condition where the demand for road space exceeds the available capacity; this leads to a decrease in vehicle speed, longer travel times, and the overall inefficiency of the road network [8]. It can be caused by various factors such as high traffic volumes, accidents, roadworks, or inefficient traffic management. When congestion occurs, the flow of vehicles is hindered, causing

delays and contributing to environmental pollution through increased emissions [13].

To accurately estimate congestion, various parameters are considered. These parameters, such as traffic flow, speed, density, and travel time, are essential for assessing congestion, since they offer measurable data points that reflect real-time and overall road network performance [14].

For the purpose of this study, the parameters (see Table I) defined by the Philippine Department of Public Works and Highways (DPWH) Highway Planning Manual [4] for the estimation of traffic congestion were used to classify the level of congestion detected in traffic scenes. Level of Service is a qualitative measure describing the operational conditions within a traffic stream and their perception by motorists and passengers. It is categorized into different levels according to the value of the Volume-to-Capacity (V/C) ratio of the road, visualized in Fig. 1.

As for the quantitative parameters, Volume is defined as the number of vehicles passing at a point or section of roadway per lane for a given time period (number of vehicles/time), usually measured by hour. Volume is a weighted measure describing how much space those vehicles occupy, and that depends on the Flat Terrain Passenger Car Equivalent Factor (PCEF) (see Table II) of the different vehicle classes that compose the traffic. On the other hand, capacity is the maximum number of vehicles that have a reasonable expectation of passing over a given section of lane or roadway in one direction or in both directions during one hour under prevailing road and traffic conditions (see Table III). It is typically expressed as passenger car unit per hour (PCU/hr).



Fig. 1. LOS Examples [15].

TABLE I. LEVEL OF SERVICE FOR PHILIPPINE HIGHWAYS [4]

Level of Service	V/C Ratio	Description
A	< 0.20	Free Flow Traffic
B	0.21–0.50	Free Flow Traffic
C	0.51–0.70	Moderate Traffic
D	0.71–0.85	Moderate/Heavy Traffic
E	0.86–1.00	Heavy Traffic
F	> 1.00	Forced Flow, Stop and Go

The V/C ratio defined by the DPWH provides a standardized metric for congestion estimation. However, its reliance on fixed time-windows introduces a temporal lag. Since the V/C ratio is calculated using aggregate per-hour or per-minute data, the resulting LOS reflects the average past traffic states rather than real-time road conditions.

B. Vehicle Detection in Traffic Scenes

Vehicle detection is the process of identifying and recognizing vehicles within a specific environment, such as roadways or parking areas [16]. Accurate vehicle detection is important for estimating congestion because it provides real-time data on the number, position, and movement of vehicles on roadways. This information helps in calculating important traffic parameters for determining the degree of congestion [17]. Accurate vehicle detection allows traffic management systems to improve urban mobility by detecting bottlenecks and anticipating traffic jams, and facilitating the implementation of more effective intelligent transport systems such as adaptive traffic signal control or route recommendations to reduce congestion [18].

Traditional vehicle detection methods rely on sensor-based technologies that are installed within road infrastructure. These methods include inductive loop detectors (ILDs), infrared sensors, microwave radar, and LiDAR-based approaches, all of which have been widely used in intelligent transport systems for vehicle identification and tracking. Despite their effectiveness for basic tracking, these sensor-based methods face several challenges and limitations. ILDs and infrared sensors are effective for basic tracking, but they are hindered by high installation costs, road surface disruption, and sensitivity to environmental factors like fog or rain. Similarly, while microwave radar and LiDAR offer resilience in adverse weather and high-precision 3D mapping, they face challenges with stationary vehicle detection and significant computational and financial overhead. Additionally, many of these methods require physical infrastructure changes, which makes them difficult to deploy in rapidly evolving traffic environments.

To address these limitations, CV-based vehicle detection has emerged as a more flexible and scalable alternative. Advancements in image processing and machine learning have led to the adoption of CV techniques for vehicle detection. Vehicle detection in CV involves identifying and locating vehicles within images or video frames, typically captured in traffic scenes [19]. These methods primarily utilize camera feeds from surveillance systems. This addresses the challenge sensor-based methods face, as it eliminates the need for extensive physical infrastructure. Modern object detection algorithms, such as convolutional neural networks, have significantly improved the accuracy and efficiency of vehicle detection, making CV an increasingly viable solution for traffic monitoring, congestion estimation, and even autonomous vehicle systems [20].

However, despite the significant improvement from tradi-

TABLE II. PCEF VALUES [4]

Vehicle Type	Description	Flat Terrain PCEF
1	Motor tricycle	2.5
2	Passenger car, Motorcycle	1.0
3–5	Passenger and goods utility and small bus	1.5
6	Large bus	2.0
7	Rigid truck, 2 axles	2.0
8	Rigid truck, 3 axles	2.5
9	Truck semi-trailer, 3 and 4 axles	2.5
10	Truck semi-trailer, 5+ axles	2.5
11	Truck trailers, 4 axles	2.5
12	Truck trailers, 5+ axles	2.5

TABLE III. BASIC HOURLY CAPACITY, PCU [4]

Carriage Width (m)	Urban Hourly PCU
Single less than 4	600
4.0-5.0	1,200
5.1-6.0	1,600
6.1-6.7	1,700
6.8-7.3	1,800
2 × 6.7 or 2 × 7.3	6,700

tional methods, vision-based vehicle detection systems still face significant challenges that hinder its effectiveness in real-world traffic scenarios. One of the most challenging problems is occlusion, which occurs when portions of a vehicle are occluded by other objects like other cars, pedestrians, road construction, or natural environments like trees and shadows. The accuracy and reliability of vehicle detection systems can be heavily affected by occlusion, which can result in partial or complete vehicle misdetection, especially in dense traffic settings [21]. Additionally, viewpoint variations also pose a problem, as vehicles appear differently depending on the distance and angles of the cameras. While deep learning models have improved object detection significantly, they still struggle to generalize well under these conditions without extensive training on diverse datasets [22].

C. End-to-End DETection TRansformers

The current CV landscape is dominated by several state-of-the-art architectures, most notably the YOLO (You Only Look Once) family, SAM (Segment Anything Model), and DETR (DETECTION TRAnsformer) [23]. DETR-based models, with their global attention mechanisms, are inherently better equipped to handle the occlusions of vehicles in dense traffic scenes.

The Real-Time DETECTION TRAnsformer (RT-DETR) is a new approach to object detection that aims to address the limitations of current real-time detectors like the YOLO series and earlier DETRs. The architecture of RT-DETR consists of a backbone, a hybrid encoder, and a Transformer decoder with auxiliary prediction heads [24]. RT-DETR demonstrates state-of-the-art real-time object detection performance, outperforming previously advanced YOLO detectors in both speed (in terms of frames per second) and accuracy (in terms of mean average precision) on the Common Objects in Context (COCO) dataset [24]. Real-time, in this case, is defined as an object detector with an inference speed of 20 frames per second or 50 ms per frame [25]. RT-DETR also significantly outperforms other DETR variants in terms of speed, achieving comparable or better accuracy with higher frames per second [24].

III. METHODOLOGY

A. Image Dataset Creation

The experimental workflow began with the acquisition and preparation of the raw data from the Ateneo de Manila University Campus Safety and Mobility Office (CSMO) video feeds which were turned into a dataset.

For this study, the CSMO provided static video feeds from Gates 2, 2.9, 3, and 3.5 of the university, running at 15 frames

per second (see Fig. 2). The feeds were taken on 2025-02-20, 06:00 AM to 2025-02-21, 06:00 AM; 2025-04-03, 06:00 AM to 2025-04-04, 06:00 AM; and 2025-10-07, 06:00 AM to 2025-10-08, 06:00 AM.

All of the frames were sampled using FFmpeg. To avoid excessively long training times while still ensuring sufficient representation of the video footage, two frame samples were taken per minute on the footage. For the training process, frames were only extracted from the feeds on 2025-02-20 to 2025-02-21 and 2025-04-03 to 2025-04-04. The footage on Gates 2.9, 3, and 3.5, frame samples were taken from 06:00 AM of the first date up until 10:00 PM, and from 05:00 AM of the succeeding date up until 06:00 AM. For the footage on Gate 2, the frame samples were taken from 06:00 AM on the first date up until 06:00 PM and from 05:00 AM on the succeeding date up until 06:00 AM. Conceptually, 11,760 frames should be sampled. However, due to the fragmented nature of the footage, a total of 11,783 frames were sampled by FFmpeg, with 23 extra samples. Since these 23 extra samples were only a very small proportion of the projected total of 11,760 frames, it was decided that the frames would still be kept for the annotation process.

For the testing process, frames were sampled from the footage on 2025-10-07 to 2025-10-08. For the footage on Gates 2.9, 3, and 3.5, samples were taken from 06:00 AM to 01:00 PM and 04:00 PM to 10:00 PM on 2025-10-07. For the footage on Gate 2, samples were taken from 06:00 AM to 01:00 PM and 02:00 PM to 06:00 PM on 2025-10-07, and 05:00 AM to 06:00 AM on 2025-10-08. Following this process, an additional 3,060 frames were sampled from the given footage with no extra frames. Combined together, there were a total of 14,843 sampled images in the dataset.

The sampled images were in 1920 x 1080 pixels, but was resized to 640 x 640 pixels when fed into the model for training and validation, similar to the image sizes in COCO dataset, which the base RT-DETR was trained on [24]. Moreover, having a consistent size for all images allows for more effective feature extraction as it helps effectively capture relevant details and patterns, which is especially important for detecting occlusions where details may be subtle.

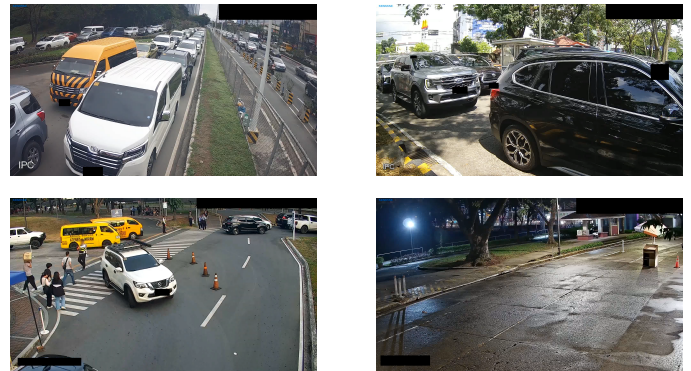


Fig. 2. Sample images from CSMO-provided feeds.

The sampled images were then labeled by an object annotation tool. Among the commonly used tools like CVAT, Roboflow Annotate, LabelStudio, and Make Sense, CVAT was

selected as it offers a free, open-source, and feature-packed solution for annotations [26].

In annotating the images, the six vehicle classes defined by the DPWH Highway Planning Manual [4], namely car, motorcycle, tricycle, bus, van, and truck were used. To guide the annotators in annotating the images, the sample images for each of the categories shown in the Philippines' Land Transportation Office [27] Filipino Driver's Manual were used as the visual basis.

The generated image annotations were saved in a JavaScript Object Notation (JSON) file format using the COCO annotation format. Sample annotation images can be seen in Fig. 3a and Fig. 3b.



Fig. 3. Sample annotated frames.

B. Model Training and Testing

The RT-DETR model with the ResNet50 backbone was then trained on the training data subset with 11,760 frames and then tested on the testing data subset with 3,060 frames.

To test the counting and traffic congestion estimation capabilities of the model, video footage was taken and trimmed from the raw CSMO footage on 2025-10-07 to 2025-10-08. 12-hour segments from 06:00 AM to 06:00 PM were chosen from the video footage as these represent the operational hours of the school roads. After trimming these segments from the raw footage, the number of vehicles that passed through the gate were counted according to their category. These counts served as the ground truth versus the model's counting capability.

C. Vehicle Counting and Ground Truth Annotation

1) *Ground truth validation*: Manual vehicle counts were established by the researchers to serve as the ground truth. To ensure measurement reliability and mitigate annotation bias, the manual counts underwent a rigorous verification stage. Annotations were all performed by the researchers. Initial counts were self-validated two times by a primary annotator and subsequently cross-validated two times by a secondary annotator. In instances of discrepancy, the cross-validation count was established as the final ground truth.

2) *Automated counting pipeline*: The valid vehicle types to be counted were selected based on the availability of PCEF values in the DPWH Highway Planning Manual [4] and its specified classes, namely, cars, motorcycles, tricycles, buses, vans, and trucks.

To accommodate the comparative analysis, two distinct vehicle counting pipelines were implemented:

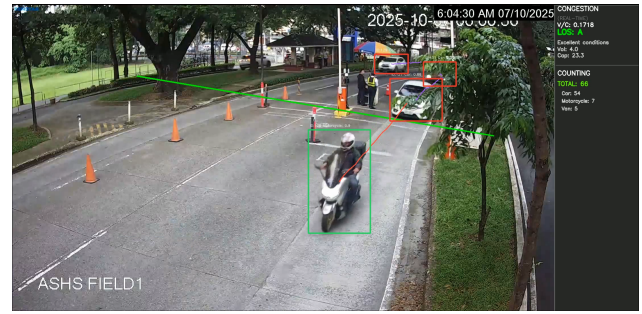


Fig. 4. Vehicle counting sample.

- **Time-Windowed Pipeline**: For the traditional metric, the ByteTrack Multi-Object Tracking Algorithm [6] was used. A vehicles was only counted when its tracked trajectory crossed a user-defined line on the roadway, as illustrated in Fig. 4.
- **Frame-Level Spatial Pipeline**: For the proposed spatial metric, the line-crossing logic was discarded. Instead, the model executed a frame-level count by aggregating all visible vehicles in a single frame simultaneously.

D. Traffic Estimation Metrics

In this study, two model frameworks were used to translate the automated vehicle detections and counts into congestion measures. This time-windowed framework was based on the standard Volume-to-Capacity ratio established by the DPWH. The LOS was determined based on equivalent traffic levels of the computed V/C ratio (see Table I).

1) *Time-windowed volume-to-capacity ratio*: The Volume-to-Capacity ratio is a quantitative measure of the demand of vehicles placed in a roadway compared to its ability to handle the traffic [4].

Under this time-windowed framework, raw vehicle counts are transformed into Passenger Car Units using the equivalent Passenger Car Equivalent Factors to account for the varying impact of different vehicle types on traffic flow [4]. Volume pertains to the number of vehicles that are using the road per hour. Volume is calculated by aggregating the weighted counts of six distinct vehicle classes: cars, vans, motorcycles, tricycles, buses, and trucks. Hence, the window-based volume is defined in Eq. 1:

$$\text{Volume}_{tw} = \sum_{i=1}^6 (n_i \times PCEF_i) \quad (1)$$

where, n_i represents the raw count of vehicles in class i , and $PCEF_i$ represents the corresponding factor for that specific class. Note that in this implementation, volume is calculated over the past n -minutes.

On the other hand, capacity is the maximum number of vehicles per hour that the road can handle before congestion occurs. A roadway's capacity (Capacity_{tw}) is based on its carriage width. The DPWH [4] defines the hourly capacity for a given road's carriage width in Table III. Hence, the V/C ratio is defined in Eq. 2:

$$\text{V/C Ratio} = \frac{\text{Volume}_{tw}}{\text{Capacity}_{tw}} \quad (2)$$

The capacity was obtained via a user-defined road width which was mapped to its corresponding PCU. This capacity was then scaled down for the shorter time intervals used in the video inference.

2) *Frame-level spatial density index*: While the DPWH-based framework provides a standardized metric for assessing traffic, its reliance on fixed time-windows introduces a temporal lag. Because the V/C ratio is calculated using aggregate data from the preceding n -minutes or hours, the resulting LOS reflects past traffic states rather than current conditions.

To address the limitations of the time-windowed V/C ratio, a Spatial Density Index (SDI) was formulated based on the instantaneous frame-level spatial occupancy. In this approach, frame-based volume is defined in Eq. 3 as the weighted sum of vehicle counts detected within a single frame:

$$\text{Volume}_{fl} = \sum_{i=1}^6 (n_{i,frame} \times PCEF_i) \quad (3)$$

Meanwhile, the modified capacity is defined in Eq. 4:

$$\text{Capacity}_{fl} = k_j \times n_l \times L \quad (4)$$

where, k_j is the jam density, n_l is the number of lanes in focus, L is the length of the roadway in focus in kilometers. Jam density is defined as the maximum possible vehicle density on the road when traffic is completely stopped [28]; its units are expressed in veh/km/lane. Theoretical simulations by Knoop and Daamen [28] identified that the theoretical jam density was 145 veh/km/lane.

The resulting SDI is then expressed in Eq. 5:

$$\text{SDI} = \frac{\text{Volume}_{fl}}{\text{Capacity}_{fl}} \quad (5)$$

Both of these ratios are dimensionless since both the PCE-weighted numerator and capacity denominator share the same unit, namely the number of vehicles.

3) *Percent error*: Accuracy is an important evaluation metric that can be used to evaluate the performance of the proposed method. Percent error represents the difference between an estimated value and the true value as a percentage of the true value [29]. It can be signed when analyzing bias or unsigned when expressing accuracy. It can be computed using the formula,

$$\text{Signed \% Error} = \frac{\text{Estimated Value} - \text{True Value}}{\text{True Value}} \times 100\% \quad (6)$$

For this study, this metric will be used to provide insights into the model's ability to underestimate or overestimate in its counting of vehicles and computation of the level of traffic congestion estimation based on the Volume-to-Capacity ratio or the Spatial Density Index.

IV. RESULTS AND DISCUSSION

A. Base Model Performance and Inference Latency

Before evaluating the congestion estimation pipelines, the underlying RT-DETR model's object detection reliability was briefly assessed. As seen in Table IV, the model demonstrated highly robust baseline detection capabilities for the selected vehicle classes.

TABLE IV. MODEL PERFORMANCE OF RT-DETR

Precision	Recall	mAP@0.5	mAP@0.5:0.95
0.999	0.989	0.934	0.795

To validate the real-time applicability of both the traditional and proposed end-to-end framework, the computational complexity of the model was measured. Using an NVIDIA RTX 3090 GPU for the complete pipeline, it can be noted that the system achieved an inference speed of 29.59 frames per second (FPS) (see Table V). This exceeds the standard 20 FPS threshold required for real-time traffic monitoring systems [25].

TABLE V. COMPUTATIONAL COMPLEXITY OF RT-DETR

Parameters	GFLOPs	Inference (FPS)	Memory (GB)
42,710,790	67.91	29.59	2.121

B. Time-Windowed Traffic Congestion Estimation

The model's ability to count vehicles over 12 hours (06:00 AM to 06:00 PM) was evaluated against the manually counted ground truth, as shown in Table VI. Across most gates, the model showed a strong performance in counting the number of vehicles, maintaining an error rate of around $\pm 10\%$. However, Gate 2 showed a significant outlier with a -15.84% error in vehicle counting.

TABLE VI. COMPARISON OF THE GROUND TRUTH, AND RT-DETR VEHICLE COUNT

Gate	Ground Truth	RT-DETR	% Error
Gate 2	3953	3643	-7.84%
Gate 2.9	5171	5303	2.55%
Gate 3	7442	7396	-0.62%
Gate 3.2	2424	2556	5.45%
Gate 3.5	4084	4261	4.33%

TABLE VII. COMPARISON OF THE GROUND TRUTH, AND RT-DETR V/C RATIO (TIME-WINDOWED)

Gate	Ground Truth	RT-DETR	% Error
Gate 2	0.196	0.157	-19.97%
Gate 2.9	0.448	0.453	1.12%
Gate 3	0.304	0.294	-3.23%
Gate 3.2	0.145	0.135	-6.75%
Gate 3.5	0.978	0.973	-0.51%

The V/C ratios derived from the counts in Table VII typically had error margins within -6.75% to 1.12%. Gate 2

was also a notable exception with a -19.97% error in the V/C ratio.

C. Strengths and Limitations of the Time-Windowed Method

The traditional time-windowed method [4] shows significant strengths primarily oriented toward long-term planning as the V/C ratio, as defined in this method, can be used to predict the exact years that a roadway may reach a saturation point and require engineering interventions [30]. The primary strength of this approach is its focus on average throughput over a long period, such as an hour. As such, it is able to ignore brief fluctuations such as drivers stopping. Hence, the primary utility of the method is in the analysis of historical congestion data to identify persistent trends.

However, the method has significant limitations for real-time applications. One critical flaw is its reliance on point-based volume counts. When the road is empty, only a few vehicles or no vehicles will pass over the counting line. Similarly, when the road is completely gridlocked, only a few vehicles or no vehicles will also pass over the counting line. In both cases, the volume and, consequently, the V/C ratio become 0, which is equivalent to LOS A. Therefore, an empty and a gridlocked roadway are both considered as free-flowing traffic, resulting in the 'zero flow' paradox (see Fig. 7). The paradox lies in the inability of the method to distinguish between the two scenarios without supporting context from, for example, a live camera. Furthermore, the traditional V/C ratio requires calculation over n -minute or hour intervals to obtain an average flow rate. This delay means V/C ratios fail to reflect rapid changes in congestion, making them unsuitable for real-time control. As seen in Fig. 5, while the average vehicle count from 3:00 PM to 3:10 PM is at a high 16.59, the average LOS was calculated at A, indicating free flow. This means that at 3:00 PM to 3:10 PM, it was not able to properly represent the LOS of the past 10 minutes as no cars were moving past the counting line.

A notable outlier observed in both the vehicle counts and the V/C ratio computations was Gate 2, where the error was significantly higher. Upon checking the video feed, the camera at Gate 2 was observed to be positioned in a ground-level, side-view manner, with vehicles in the foreground frequently occluding those in the background. This left the model with insufficient visual features to identify and count trailing vehicles. This highlights the critical impact of camera placement on model detection accuracy, as accuracy relies on an object having the appropriate resolution at a given distance [31]. Going beyond the optimal distance makes objects look smaller and more difficult to separate from noise [31]. Moreover, optimal placement ensures that the maximum possible area is covered while minimizing blind spots or redundant overlaps between cameras [32]. Hence, it can be observed here how improper placement led to significant undercounting, which skewed the Level of Service computations for a given gate.

D. Frame-Level Traffic Congestion Estimation

The proposed frame-level framework demonstrates a notable performance improvement over the traditional time-windowed approach. As shown in Table VIII, the proposed

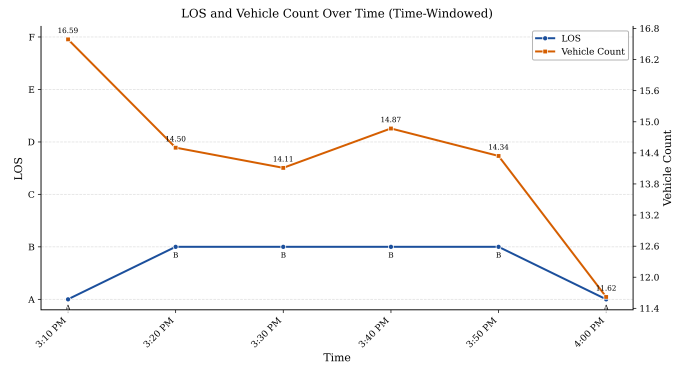


Fig. 5. LOS and vehicle count over time (time-windowed).

TABLE VIII. COMPARISON OF THE GROUND TRUTH AND RT-DETR SDI (FRAME-LEVEL)

Gate	Ground Truth	RT-DETR	% Error
Gate 2	0.400	0.355	-11.08%
Gate 2.9	0.794	0.815	2.69%
Gate 3	0.179	0.178	-0.77%
Gate 3.5	0.334	0.333	-0.22%

method reduced the percentage error margin to between -0.22% and 2.85%, a substantial improvement over the time-windowed measurements with range of -6.75% to 1.12%. Notably, the framework also demonstrated resilience to suboptimal camera configurations in its LOS computations compared to the time-windowed method. For instance, at Gate 2, despite a highly problematic viewing angle, the model reduced the error relative to the ground truth from a baseline of -19.97% in the time-windowed method to -11.08%. While the negative percentage value indicates that there is still an underestimation in congestion, the proposed framework mitigates the severe miscalculations inherent in the time-windowed method.

E. Strengths and Limitations of the Frame-Level Estimation Method

A primary strength of the proposed method is its ability to solve the zero flow paradox. This proposed method, by analyzing the entire frame, instantly distinguishes between gridlocked and free-flowing traffic states by detecting the presence of stationary vehicles (see Fig. 7). As seen in Fig. 6, the average vehicle count at the interval from 3:00 to 3:10 PM is 16.59, with the LOS being at level D. When contrasted with Fig. 5, this shows that the frame-level method can distinguish a congested road, whereas the time-windowed method cannot. Furthermore, the method provides a significant advantage in temporal resolution versus the time-windowed method. The proposed method is able to detect spikes in vehicle density immediately, offering data that is far more responsive for applications like adaptive traffic signal control.

However, the method has limitations dependent on camera perspective and scene complexity. Because it detects all vehicles within the entire video frame, it can be inaccurate for frames containing multiple traffic routes, such as in Gates 3.2 and 3.5, where distinguishing the specific lane or road segment is challenging. Without Region of Interest (ROI) masking, the

model may aggregate the traffic density of the entire frame, making it difficult to isolate and assign specific congestion metrics to each individual gate. Another significant constraint is occlusion. A common issue with single, oblique-angle CCTV views is that larger vehicles, like trucks, can obscure smaller vehicles behind them, such as cars or motorcycles. This leads to an underestimation of true density. A top-down perspective from a drone or satellite would mitigate this [33], but it is often not available for most traffic monitoring infrastructures.

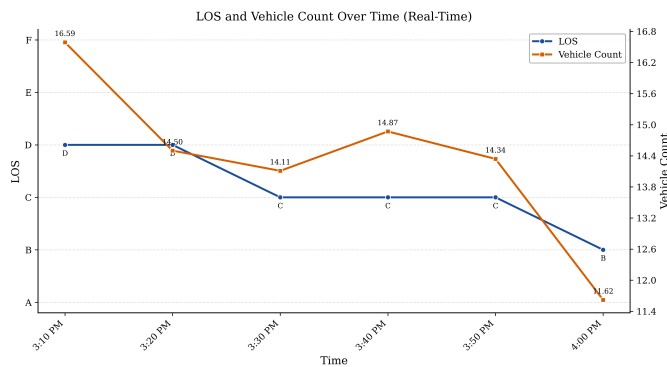


Fig. 6. LOS and vehicle count over time (frame-level).



Fig. 7. Time-windowed method (top) vs. frame-level method (bottom) detection comparison in congested states.

V. CONCLUSION

In conclusion, this study demonstrates that frame-wide spatial density analysis offers a robust alternative to the traditional time-windowed and line-based traffic congestion estimation by directly resolving the zero flow paradox inherent in severely congested conditions. By processing instantaneous vehicle density through the proposed frame-level method, the framework outperformed the traditional methods and effectively reduced percentage errors by approximately half while

maintaining a more accurate representation of the true congestion state. Furthermore, unlike traditional flow-based V/C ratios that require substantial time-windows, the ability of the SDI to detect immediate spikes in congestion makes it highly applicable for integration into adaptive traffic signal control systems as it uses real-time congestion data; this enables control systems to autonomously optimize signal timing and alleviate sudden congestion [34].

VI. RECOMMENDATIONS FOR FUTURE WORK

The current framework utilizes jam density thresholds derived from the literature based on traffic conditions in the Netherlands. However, Philippine traffic is highly heterogeneous, characterized by unique vehicle dimensions (e.g., jeepneys, motorcycles, etc.), unstructured lane discipline, and tighter vehicle packing during severe congestion. Consequently, applying foreign baselines may skew maximum density estimations. Future research should consider establishing localized jam density metrics. Defining these local parameters will critically calibrate the model's congestion thresholds and improve the accuracy of Level of Service (LOS) classifications in similar developing urban centers.

To address the limitations of processing multi-directional traffic within a single camera feed, future research should investigate integrating ROI masking. Applying explicit spatial boundaries will allow the model to isolate specific road segments and lane directions. This targeted extraction is crucial for preventing data aggregation in complex scenes and enabling the framework to provide accurate, road-specific congestion estimations rather than generalized frame-wide density.

To address the limitation of vehicle occlusion inherent in ground-level camera angles, future implementations may look into prioritizing top-down camera perspectives. As noted, ground angles frequently result in larger vehicles (e.g. vans, trucks) obscuring smaller trailing vehicles, which inherently leads to an underestimation of true traffic congestion.

While this study successfully serves as a proof-of-concept for the SDI, certain methodological limitations must be acknowledged. First, due to the limited sampling windows of the university vehicular gates, the current evaluation lacks the statistical volume necessary for cross-validation. Consequently, the reported error margins require further validation across diverse environmental conditions, including varying weather and lighting. Second, the SDI has to be adapted into discrete DPWH Level of Service categories since the LOS categories outlined by the DPWH are based on the V/C ratio. Future studies must consider both of these before considering the generalized deployment of this system.

ACKNOWLEDGMENT

The authors would like to thank the Campus Safety and Mobility Office (CSMO) for providing the dataset used in this study.

Special thanks are extended to Social and Environmental Engagement for Development and Sustainability (SEEDS) for the provision of the computer used for model training and to Research, Creative Work, and Innovation (RCWI) for their administrative assistance and organizational support throughout this research.

VII. DECLARATION ON GENERATIVE AI

During the preparation of this work, the authors used Google Gemini 3.1 Pro to improve the clarity of the paper and aid in proofreading. After the use of this tool, the authors carefully reviewed and edited the content as needed to align with the SAI paper format. The authors take full responsibility and accountability for the content in this work.

REFERENCES

- [1] M. Akhtar and S. Moridpour, "A review of traffic congestion prediction using artificial intelligence," *Journal of Advanced Transportation*, vol. 2021, no. 1, Jan. 2021. [Online]. Available: <https://doi.org/10.1155/2021/8878011>
- [2] R. Ravish and S. R. Swamy, "Intelligent traffic management: A review of challenges, solutions, and future perspectives," *Transport and Telecommunication Journal*, vol. 22, no. 2, pp. 163–182, Apr. 2021. [Online]. Available: <https://doi.org/10.2478/ttj-2021-0013>
- [3] H. Ghahremannezhad, H. Shi, and C. Liu, "Object detection in traffic videos: A survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 7, pp. 6780–6799, Jul. 2023. [Online]. Available: <https://doi.org/10.1109/TITS.2023.3258683>
- [4] D. of Public Works and Highways, "Highway planning manual," 2007.
- [5] R. R. Sai, A. K. Tangirala, and L. Vanajakshi, "Road traffic analysis using 2d lidar," in *2024 16th International Conference on COMMunication Systems & NETworkS (COMSNETS)*, ser. COMSNETS 2024. New York, NY, USA: IEEE, 2024, pp. 228–233. [Online]. Available: <https://doi.org/10.1109/COMSNETS59351.2024.10426892>
- [6] Z. Zhang, S. Zhang, L. Mo, M. Guo, F. Liu, and X. Qi, "Traffic volume estimate based on low penetration connected vehicle data at signalized intersections: A bayesian deduction approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 10 596–10 609, Aug. 2022. [Online]. Available: <https://doi.org/10.1109/TITS.2021.3094933>
- [7] N. K. Singh, A. K. Tangirala, and L. D. Vanajakshi, "A multivariate analysis framework for vehicle detection from loop data under heterogeneous and less lane disciplined traffic," *IEEE Access*, vol. 9, pp. 143 580–143 591, Oct. 2021. [Online]. Available: <https://doi.org/10.1109/ACCESS.2021.3120470>
- [8] F. H. A. Research and Technology, "Chapter 1, traffic detector handbook: Third edition—volume i - fhwa-hrt-06-108," 2006, retrieved from <https://www.fhwa.dot.gov/publications/research/operations/its/06108/01.cfm>. [Online]. Available: <https://www.fhwa.dot.gov/publications/research/operations/its/06108/01.cfm>
- [9] J. Kim, B.-j. Park, and J. Kim, "Empirical analysis of autonomous vehicle's lidar detection performance degradation for actual road driving in rain and fog," *Sensors*, vol. 23, no. 6, Mar. 2023. [Online]. Available: <https://doi.org/10.3390/s23062972>
- [10] Q. Xue, "Advancements in object detection: From machine learning to deep learning paradigms," in *Proceedings of the 2nd International Conference on Software Engineering and Machine Learning (CONF-SEML 2024)*, ser. CONF-SEML 2024. Oxford, United Kingdom: EWA Publishing, 2024, pp. 154–159. [Online]. Available: <https://doi.org/10.54254/2755-2721/75/20240530>
- [11] N. R. Dangal, D. Wangmo, S. Meena, and S. Kumar, "Study of traffic flow for estimation of level of service," *International Journal for Research Trends and Innovation*, vol. 8, no. 4, pp. 1348–1353, Apr. 2023. [Online]. Available: <https://www.ijrti.org/viewpaperforall?paper=IJRTI2304220>
- [12] X. Li, T. Hao, X. Jin, B. Huang, and J. Liang, "Fine traffic congestion detection with hierarchical description," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 12, pp. 24 439–24 453, Dec. 2022. [Online]. Available: <https://doi.org/10.1109/TITS.2022.3206583>
- [13] P. Krüttli, D. Nef, P. Kangethe, D. Etongo, and M. Stauffacher, "Sustainable land use – seychelles' transport system," ETH Zurich, Zurich, Switzerland, Tech. Rep., 2022.
- [14] M. Z. Mehdi, H. M. Kammoun, N. G. Benayed, D. Sellami, and A. D. Masmoudi, "Entropy-based traffic flow labeling for cnn-based traffic congestion prediction from meta-parameters," *IEEE Access*, vol. 10, pp. 16 123–16 133, Feb. 2022. [Online]. Available: <https://doi.org/10.1109/ACCESS.2022.3149059>
- [15] T. R. Board, "Us highway capacity manual," 2010. [Online]. Available: <https://www.jpautoceste.ba/wp-content/uploads/2022/05/Highway-Capacity-Manual-2010-PDFDrive-.pdf>
- [16] Z. Yang and L. S. Pun-Cheng, "Vehicle detection in intelligent transportation systems and its applications under varying environments: A review," *Image and Vision Computing*, vol. 69, pp. 143–154, Jan. 2018. [Online]. Available: <https://doi.org/10.1016/j.imavis.2017.09.008>
- [17] R. Bauza and J. Gozalvez, "Traffic congestion detection in large-scale scenarios using vehicle-to-vehicle communications," *Journal of Network and Computer Applications*, vol. 36, no. 5, pp. 1295–1307, Sep. 2013. [Online]. Available: <https://doi.org/10.1016/j.jnca.2012.02.007>
- [18] L. E. Y. Mimbela and L. A. Klein, "Summary of vehicle detection and surveillance technologies used in intelligent transportation systems," 2007. [Online]. Available: <https://www.fhwa.dot.gov/policyinformation/pubs/vdstits2007/vdstits2007.pdf>
- [19] H. Song, H. Liang, H. Li, Z. Dai, and X. Yun, "Vision-based vehicle detection and counting system using deep learning in highway scenes," *European Transport Research Review*, vol. 11, Dec. 2019. [Online]. Available: <https://doi.org/10.1186/s12544-019-0390-4>
- [20] L. Jiao, F. Zhang, F. Liu, S. Yang, L. Li, and Z. Feng, "A survey of deep learning-based object detection," *IEEE Access*, vol. 7, pp. 128 837–128 868, Sep. 2019. [Online]. Available: <https://doi.org/10.1109/ACCESS.2019.2939201>
- [21] S. Gilroy, E. Jones, and M. Glavin, "Overcoming occlusion in the automotive environment—a review," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 1, pp. 23–35, Jan. 2021. [Online]. Available: <https://doi.org/10.1109/TITS.2019.2956813>
- [22] Z. Deng, H. Sun, S. Zhou, J. Zhao, and H. Zou, "Toward fast and accurate vehicle detection in aerial images using coupled region-based convolutional neural networks," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 8, pp. 3652–3664, May 2017. [Online]. Available: <https://doi.org/10.1109/JSTARS.2017.2694890>
- [23] X. M. Li, X. H. Pu, W. C. Ling, and X. D. Song, "Yolo-sam an end-to-end framework for efficient real time object detection and segmentation," *Scientific Reports*, vol. 15, Nov. 2025. [Online]. Available: <https://doi.org/10.1038/s41598-025-24576-6>
- [24] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen, "Detrs beat yolos on real-time object detection," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, ser. CVPR 2024. New York, NY, USA: IEEE, 2024, pp. 16 965–16 974. [Online]. Available: <https://doi.org/10.1109/CVPR52733.2024.01605>
- [25] V. K. Sharma and R. N. Mir, "A comprehensive and systematic look up into deep learning based object detection techniques: A review," *Computer Science Review*, vol. 38, Nov. 2020. [Online]. Available: <https://doi.org/10.1016/j.cosrev.2020.100301>
- [26] CVAT.ai, "Computer vision annotation tool (cvat)," 2018, retrieved from <https://github.com/cvat-ai/cvat>. [Online]. Available: <https://github.com/cvat-ai/cvat>
- [27] P. L. T. Office, "Filipino driver's manual," 2021. [Online]. Available: <https://lto.gov.ph/wp-content/uploads/2023/10/FDM-Vol.-1-2nd-Edition.pdf>
- [28] V. L. Knoop and D. Winnie, "Automatic fitting procedure for the fundamental diagram," *Transportmetrica B: Transport Dynamics*, vol. 5, no. 2, pp. 129–144, Nov. 2016. [Online]. Available: <https://doi.org/10.1080/21680566.2016.1256239>
- [29] LibreTextsChemistry, "3.13: Percent error." [Online]. Available: [https://chem.libretexts.org/Bookshelves/Introductory_Chemistry/Introductory_Chemistry_\(CK-12\)/03%3A_Measurements/3.13%3A_Percent_Error](https://chem.libretexts.org/Bookshelves/Introductory_Chemistry/Introductory_Chemistry_(CK-12)/03%3A_Measurements/3.13%3A_Percent_Error)
- [30] J. Arliansyah, M. R. Prasetyo, and A. Y. Kurnia, "Planning of city transportation infrastructure based on macro simulation model," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 7, no. 4, pp. 1262–1267, Aug. 2017. [Online]. Available: <https://doi.org/10.18517/ijaseit.7.4.2444>

- [31] AllGoVision, "Optimal camera placement for object detection accuracy," AllGoVision, Bangalore, India, Tech. Rep., n.
- [32] D. Khanal, "Traffic monitoring camera locations optimization study for interstate in idot district 8," Master's thesis, Southern Illinois University Edwardsville, 2023.
- [33] X. Liu and Z. Zhang, "A vision-based target detection, tracking, and positioning algorithm for unmanned aerial vehicle," *Wireless Communications and Mobile Computing*, vol. 2021, 2021. [Online]. Available: <https://doi.org/10.1155/2021/5565589>
- [34] M. Gupta, H. Miglani, P. Deo, and A. Barhatte, "Real-time traffic control and monitoring," *e-Prime - Advances in Electrical Engineering, Electronics and Energy*, vol. 5, p. 100211, 2023. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2772671123001067>