Vol. 11, No. 2, Oktober 2026, Page. 506-520
https://journal.uin-alauddin.ac.id/index.php/instek
p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

YOLO11m-Based Vehicle Counting for Standardized CO_2 and CO Emission Estimation from Semarang CCTV

Moch Adi Setiawan[*1], Galuh Wilujeng Saraswati[2], Erba Lutfina[3], Danny Oka Ratmana[4], Resha Meiranadi Caturkusuma[5]

[1,2, 3, 5] Information Systems Department, Faculty of Computer Science, Dian Nuswantoro University, Indonesia
[4]Informatics Engineering Department, Faculty of Computer Science, Dian Nuswantoro University, Indonesia
Email: [1]612202300092@mhs.dinus.ac.id, [2]galuhwilujengs@dsn.dinus.ac.id,
[3]erba.lutfina@dsn.dinus.ac.id, [4]rdannyoka@dsn.dinus.ac.id, [5]resha.meiranadi@adm.dinus.ac.id

Abstract

This study evaluated YOLO11m-based vehicle counting in Pantau Semar CCTV recordings and analyzed how class-specific counting differences affect standardized CO_2 and CO estimates. The dataset comprised 12,000 images: 9,120 for training, 2,280 for validation, and 600 for testing. Model development used main training and corrective fine-tuning with Ultralytics. A Python pipeline combined ByteTrack, class stabilization, and line-crossing counting. Automated counts were compared with manual reference counts across six daytime and nighttime recordings from three locations. The detector achieved precision of 0.904, recall of 0.895, mAP_{50} of 0.947, and $mAP_{50–95}$ of 0.833. The system counted 7,444 vehicles compared with 7,667 manually, a difference of −2.91%; daytime and nighttime differences were +0.88% and −6.55%, respectively. Using identical emission factors and a one-kilometer reference distance, system-derived estimates were 1.70–3.88% lower for CO_2 and 2.79–4.94% lower for CO across six scenarios. The study shows how class-specific counting errors and emission factors produce emission-estimate differences that cannot be inferred from total-count differences alone. Findings are limited to the evaluated recordings and assumptions; the results are standardized comparative estimates rather than measurements of actual emissions.

Keywords: computer vision, emission estimation, vehicle counting, vehicle detection, YOLO11m

This work is an open access article and licensed under a Creative Commons Attribution-NonCommercial ShareAlike 4.0 International (CC BY-NC-SA 4.0)

1. INTRODUCTION

Road traffic contributes to urban emissions through fuel combustion, with the amount released depending on vehicle activity, fleet composition, and operating conditions. Carbon dioxide (CO_2) is associated with fuel consumption, while carbon monoxide (CO) is produced through incomplete combustion. Traffic-related air pollution has also been linked to adverse cardiovascular and respiratory outcomes [1]. Emission inventories commonly combine activity data with pollutant-specific emission factors [2], including factors applied in an Indonesian road-corridor study [3]. The usefulness of these estimates depends partly on how well the activity data represent the vehicles observed at a particular location and time. Class-specific vehicle counts are therefore an important input when different factors are assigned to different vehicle categories.

Manual observation can provide reference counts but becomes demanding when several locations and recording periods must be reviewed. Computer vision offers an alternative for extracting vehicle classes and traffic volumes from images and video [4], [5]. Applications have used cameras on transit buses [6] and existing traffic surveillance cameras [7], although these observation platforms differ from one another in viewpoint and scene coverage. Studies have investigated counting from prerecorded video [8], joint counting and speed estimation [9], fixed-camera detection and counting [10], and multiobject multicamera monitoring [11]. These applications show that traffic observation involves several stages beyond frame-level detection. Vehicles must be associated across successive frames, assigned a class, and counted according to a consistent spatial rule. Errors introduced during these stages can affect the resulting counts even when individual-frame detection performance is satisfactory.

Previous emission studies illustrate different relationships between vehicle observations and environmental estimates. Valdés-Montenegro et al. [12] used the International Vehicle Emissions model to estimate on-road emissions with attention to vehicle technology and local conditions. This approach incorporates fleet information that cannot be established from broad visual categories alone. Pinzon-Acosta et al. [13] integrated YOLO-based vehicle observations with an EMEP Tier 1 method to examine the spatial distribution of road-traffic emissions. Together, these studies demonstrate the importance of both activity data and the assumptions used to translate those data into emissions. They also distinguish inventory-oriented estimation from direct measurement of exhaust emissions or ambient pollutant concentrations.

Closer links between automated counting and emission estimates have been investigated by Yavuz et al. [14] and Suparwito et al. [15]. Yavuz et al. combined YOLOv7 with SORT and calculated Tier 3 emission estimates using both ground-truth data and algorithm-derived observations. Their work establishes that comparing vision-derived and ground-truth-derived emission estimates is already an existing research approach. Suparwito et al. used YOLOv9 to identify motorcycles, cars, buses, and trucks and related the observed counts to CO estimation. Their class-level detection results and comparisons of actual and predicted counts also illustrate that performance varies across vehicle categories. These studies provide a basis for examining the environmental consequences of counting discrepancies, but their different activity definitions, pollutants, and estimation assumptions prevent direct numerical comparison with the present study. Here, the analysis focuses on the contribution of each class-count difference to CO_2 and CO estimates when both count sources use identical factors and a common reference distance.

The reliability of those counts remains sensitive to the visual conditions represented in the data. CCTV detection can be affected by oblique views and occlusion [16], headlight glare and obstruction by large vehicles [17], changes in luminance [18], and dense, heterogeneous traffic [19]. Comparisons of vehicle detectors have also reported differences among model configurations and vehicle classes [20]. Such differences matter because the relative importance of a counting discrepancy is not determined by its numerical size alone. A small undercount in a class assigned a large emission factor may contribute substantially to the estimated emission difference, while a larger discrepancy in another class may have a smaller effect. Moreover, false counts and omissions can partly cancel in the aggregate vehicle total without canceling in the corresponding pollutant estimates.

Against this background, the question addressed in this study is how class-specific differences between automated and manual counts propagate into standardized CO_2 and CO estimates in the selected daytime and nighttime Pantau Semar CCTV recordings. The analytical gap concerns the information lost when performance is summarized only by detector metrics, a net vehicle-count difference, or a total emission estimate. Examining the signed contribution of each class makes that relationship explicit. Using six alternative emission-factor combinations further shows how the interpretation changes with assumptions about passenger-car technology and truck type. The contribution is thus an evaluation of the count-to-emission relationship under stated local observation conditions and factor assumptions; using YOLO11m itself is not claimed as a new detection architecture.

Wicaksana et al. [21] evaluated YOLO11 classification variants on the HAM10000 dataset and reported improved skin-lesion classification performance after data augmentation. Their study provides methodological context on augmentation, although its image-classification task differs from the vehicle detection and counting tasks evaluated here. YOLO11m was implemented through the Ultralytics framework [22] as an existing medium-capacity detector. It was selected as a practical configuration for training with 1,024-pixel inputs and a batch size of two on the available NVIDIA GeForce RTX 4060 8 GB GPU. These hardware and computational considerations explain the model choice without establishing its superiority over other architectures. Main training and corrective fine-tuning adapted the pretrained detector to the study data, while its performance was evaluated separately from video-based counting.

The study had three objectives: to evaluate the trained detector on a separate image test subset; to compare automated and manually prepared ground-truth (GT) counts by vehicle class, location, and observation period; and to quantify how those counting differences affect standardized CO_2 and CO estimates. The dataset contained 12,000 Pantau Semar images representing buses, passenger cars, motorcycles, and trucks. Video evaluation used six approximately 15-minute recordings from three Semarang locations, with one daytime and one nighttime recording per location. ByteTrack [23], track-

level class stabilization, and a predefined line-crossing rule converted detections into vehicle counts. The two count sources were then processed with identical emission factors and a one-kilometer reference distance per counted vehicle in each of the six scenarios. This common basis isolates the effect of differences between the count sources within a scenario. The resulting values are comparative estimates for the observed recordings, rather than direct measurements of physical emissions, ambient pollutant concentrations, or total emissions along the monitored road corridors.

2. RESEARCH METHODS

This study used a quantitative experimental design comprising image-dataset preparation, two-stage YOLO11m training, bounding-box evaluation, vehicle tracking and line-crossing counting, aggregate counting validation, and standardized emission estimation. The image-based evaluation measured detector performance on a locked test set, whereas the video-based evaluation compared automated and manually prepared vehicle counts from six CCTV recordings. Within each sensitivity scenario, the same emission factors and reference distance were applied to both count sources, so differences between their estimates arose from differences in vehicle counts. Figure 1 summarizes the complete procedure.

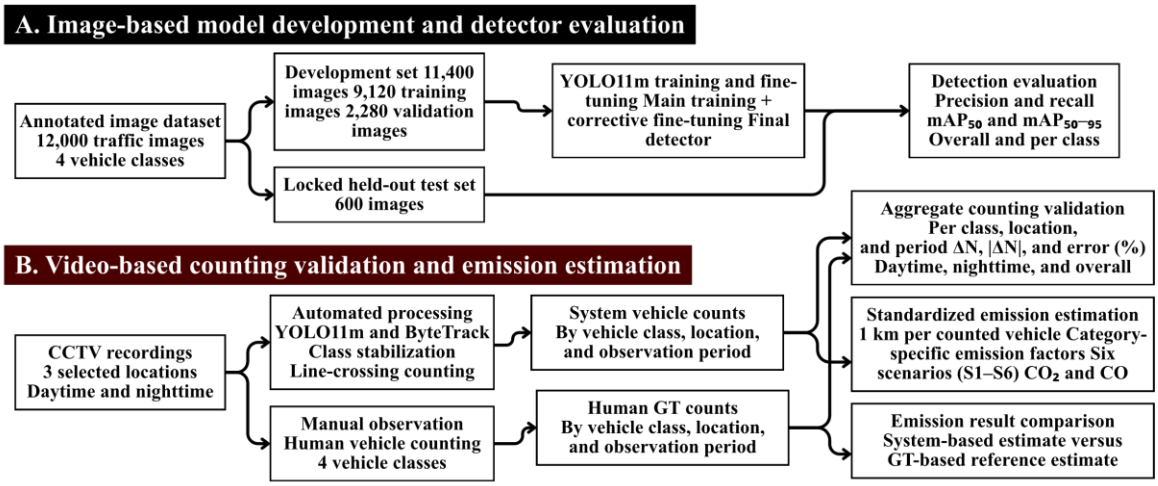


Figure 1. Overall framework for vehicle detection, aggregate counting validation, and standardized emission estimation.

2.1. Vehicle Detection, Tracking, and Counting

Before image acquisition, publicly viewable Pantau Semar CCTV streams were screened for operational stability and visual conditions. Twelve cameras were selected to cover differences in viewing angle, field of view, image clarity, illumination, and traffic scenes; inactive or unstable streams were excluded. Following an initial acquisition trial on 21 August 2026, images were collected during morning, daytime, afternoon, and nighttime periods. Custom Python scripts invoked FFmpeg to save JPEG snapshots, generally at approximately one image per camera every 60 seconds. Supplementary acquisitions used different intervals, including targeted nighttime and early-morning sampling to increase the representation of trucks. Images with severe blur, irrelevant content, or inadequate visibility were removed during quality screening. The final dataset contained 12,000 images annotated in four vehicle classes: buses, passenger cars, motorcycles, and trucks.

Dataset preparation began with a gold set of 1,800 images whose initial annotations were manually audited and corrected in a private Roboflow workspace. A separate 50-image pilot set, outside the gold set and not previously used to train the preliminary model, was used to inspect its predictions and annotation consistency. After the pilot stage, these 50 images entered the final development pool: 40 were assigned to training and 10 to validation. None entered the final test subset. The gold set and pilot set therefore describe stages of dataset preparation, rather than additional partitions to be added to the final dataset total.

After the early gold-set and pilot stages, 600 images were selected for the final detector test before main training of the final model. Selection was deterministic with random seed 42 and stratification by camera and acquisition period, with 50 images from each camera. Preliminary-model predictions supplied initial labels, which were manually checked and corrected before the final evaluation. These 600 images were excluded from the final training and validation lists.

A separate Python script divided the remaining 11,400 images into training and validation partitions. Within each camera, the script ranked the 950 non-test image keys using a SHA-256 hash constructed from seed 42 and the image key, assigned the first 190 images to validation, and allocated the remaining 760 to training. This produced 9,120 training images and 2,280 validation images. Roboflow supported annotation review and export; the Python script constructed the final training-validation split. Table 1 summarizes the final image partitions.

Table 1. Final image partitions used for detector development and evaluation

Dataset subset	Unique images	Images per camera	Purpose
Training	9,120	760	Model training
Validation	2,280	190	Performance monitoring and checkpoint selection
Test	600	50	Detector evaluation; excluded from training and validation
Total	12,000	1,000	Complete annotated dataset

Manifest checks found no image keys shared between the three partitions. However, images from the same cameras occur in all partitions, and the split did not hold out entire cameras or acquisition-time blocks. Non-overlapping image keys therefore do not establish temporal or visual independence between the partitions. The detector results describe performance on separate images from the represented camera sources, with limited evidence for generalization to unseen cameras.

During development-data quality review, the human-reviewed test annotations were also consulted when selecting thresholds for a heuristic that flagged potentially duplicated bounding boxes in development annotations. The heuristic generated candidates for human inspection rather than automatically deleting annotations. Thus, although the test images were excluded from training and validation, not all annotation-review decisions were independent of the test annotations. This use of test information is a limitation of the detector evaluation and is distinguished here from direct inclusion of test images in model training.

The CCTV recordings were obtained from publicly accessible Pantau Semar streams that could be viewed without logging in or creating an account. All cameras used in this study captured traffic from distant viewpoints rather than providing close-up views of individuals. The viewing distance, limited image resolution, and blur restricted the detail available for facial identification and license-plate reading. The analysis focused exclusively on vehicle classes and traffic counts and did not involve facial identification or license-plate recognition. No additional face or license-plate masking was applied. The complete image dataset, annotation collection, raw recordings, individual vehicle trajectories, trained model weights, and original project repository remain private and are not publicly released or distributed.

A restricted supporting package contains five image–annotation pairs, code for selected counting operations, aggregate count comparisons, standardized emission arithmetic, and selected model information. Access may be granted upon request, subject to individual approval by the authors. The materials support inspection of the supplied operations and verification of basic calculations, but do not enable independent reproduction of the complete training, detection, tracking, and video-counting pipeline. The complete dataset, trained model weights, and original project repository remain private.

Model training was performed using Python 3.11.0, PyTorch 2.11.0+cu128 (CUDA 12.8 build), and Ultralytics 8.4.129 [22] on a computer equipped with an Intel Core i5-13600KF processor, an NVIDIA GeForce RTX 4060 GPU with 8 GB of memory, and 32 GB of RAM. The study used the existing YOLO11m architecture in two successive stages: main training followed by corrective fine-

tuning. Main training was initialized with the original pretrained YOLO11m weights, whereas corrective fine-tuning was initialized with the best checkpoint obtained during main training.

Main training used the 9,120-image training partition, with the 2,280-image validation partition used for performance monitoring and checkpoint selection. It was configured for a maximum of 100 epochs, early-stopping patience of 20 epochs, an input size of 1,024 pixels, and a batch size of two. The random seed was 42, deterministic operation and automatic mixed precision were enabled, and optimizer selection was delegated to Ultralytics through the optimizer=auto setting. The configured warm-up period was three epochs. The checkpoint selected through validation became the pre-corrective model.

Preliminary application to diagnostic CCTV footage prompted further review of the development annotations, including suspected duplicate boxes and other labeling problems. After the identified annotations were corrected, corrective fine-tuning continued from the pre-corrective best checkpoint. This stage was configured for a maximum of 20 additional epochs, early-stopping patience of seven epochs, and random seed 0. It used stochastic gradient descent with an initial learning rate of 0.001, momentum of 0.937, weight decay of 0.0005, and a one-epoch warm-up period. The input size remained 1,024 pixels and the batch size remained two, with deterministic operation and automatic mixed precision enabled.

The corrective training list retained all 9,120 unique training images and added targeted repetitions of difficult cases and images containing trucks or buses. It contained 10,496 entries: 8,156 unique images appeared once, 714 appeared twice, 169 appeared three times, and 81 appeared five times. These repetitions increased sampling frequency without adding new images to the dataset. The validation partition retained 2,280 distinct images, and the 600-image test subset remained outside both training lists.

Both training stages used HSV augmentation gains of 0.015, 0.7, and 0.4 for hue, saturation, and value, respectively; translation of 0.1; scale variation of 0.5; horizontal-flip probability of 0.5; and mosaic probability of 1.0. Rotation, shear, perspective transformation, vertical flipping, and mixup were disabled. Mosaic augmentation was configured to stop during the final ten scheduled epochs of main training and the final three scheduled epochs of corrective fine-tuning. The best checkpoint from corrective fine-tuning was used as the final detector for image evaluation and subsequent video processing.

The final detector was evaluated on the separate 600-image test subset using precision, recall, mAP_{50}, and $mAP_{50–95}$, reported overall and by vehicle class. Precision and recall quantify agreement between predicted bounding boxes and annotated objects. The mAP_{50} metric uses an intersection-over-union threshold of 0.50, whereas $mAP_{50–95}$ averages performance over thresholds from 0.50 to 0.95 in increments of 0.05. These image-level metrics evaluate detection and localization; aggregate vehicle-count validation is described separately in Section 2.2.

The video evaluation used six recordings: one daytime and one nighttime clip from each of Fatmawati, JPO Ahmad Yani, and Tugu Muda 2. Each clip lasted approximately 15 minutes, giving approximately 90 minutes of footage in total. These recordings were distinct from the earlier footage used for model diagnosis and parameter development and were evaluated without retuning the detector or counting parameters on them. Recording resolutions were 1,920 × 1,080 pixels at Fatmawati, 2,688 × 1,520 at JPO Ahmad Yani, and 2,560 × 1,440 at Tugu Muda 2. Source frame rates ranged from 24.353 to 25.012 frames per second. The observations were limited to the selected camera locations and recording periods.

Video processing used Python 3.11.0 and OpenCV (opencv-python 4.11.0.86), with the final corrective YOLO11m checkpoint supplying detections to ByteTrack [23]. All recorded frames were processed, corresponding to a frame stride of one. Detector input size was 1,024 pixels, with a confidence threshold of 0.10 and non-maximum-suppression IoU threshold of 0.60. ByteTrack used high- and low-confidence association thresholds of 0.10 and 0.05, a new-track threshold of 0.25, a track buffer of 90 frames, and a matching threshold of 0.90, with score fusion enabled. Figure 2 illustrates detection, track association, stabilized classes, and the counting-line configuration.

Vol. 11, No. 2, Oktober 2026, Page. 506-520
https://journal.uin-alauddin.ac.id/index.php/instek
p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

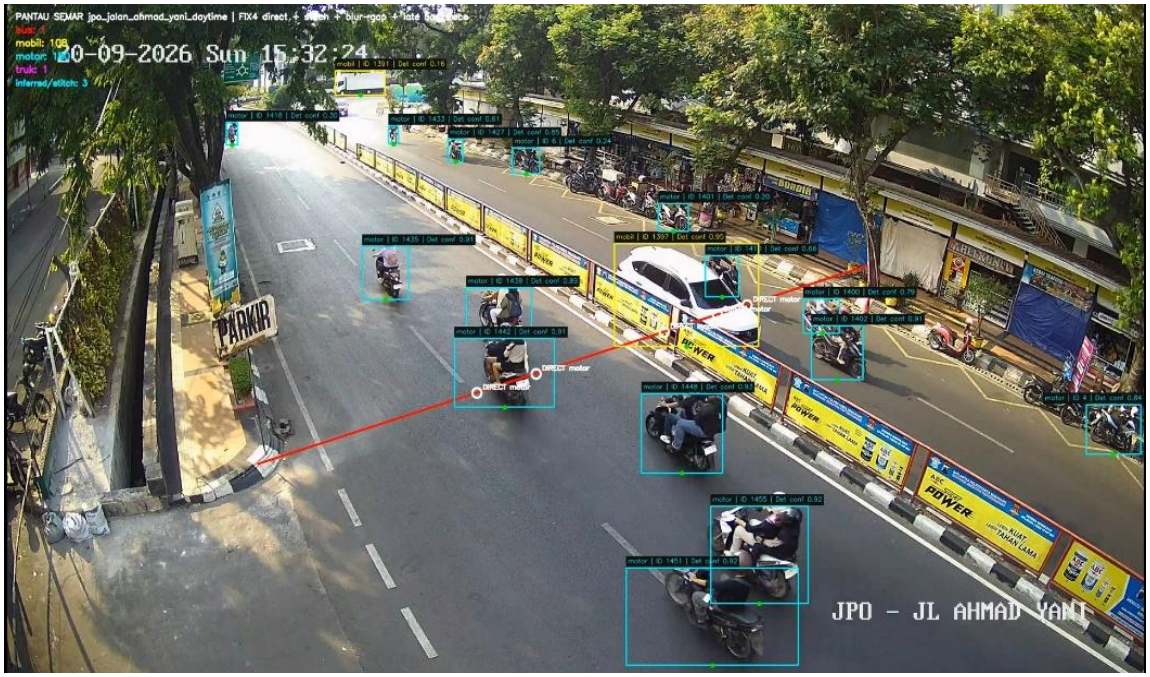


Figure 2. YOLO11m vehicle detection, ByteTrack association, and line-crossing counting in a Pantau Semar CCTV recording.

A stable vehicle class was assigned to each track using quality-weighted temporal voting. Votes incorporated detection confidence and bounding-box area relative to observations within the same track, giving greater weight to larger, higher-confidence observations. Smaller or more distant observations still contributed with lower weights. The stabilized class was used for counting, while the original frame-level class predictions were retained. This stage assigned one class per track without deleting or merging tracks.

Vehicle position was represented by the bottom-center point of the bounding box as a proxy for contact with the road. A finite counting line was defined for each recording, and direct crossing events were identified from trajectories passing between opposite sides outside a six-pixel deadband. Each directly crossing track was counted once. Additional rules addressed interrupted trajectories: compatible track fragments could be joined across gaps of up to 0.75 seconds; short-gap linear extrapolation was limited to 0.45 seconds; and a motorcycle-specific late-backtrace rule allowed backward trajectory projection for up to 1.35 seconds when the first reliable detection appeared beyond the line. These inferred events were distinguished from directly observed crossings in the event records. Motorcycle tracks were not stitched to passenger-car, bus, or truck tracks.

Duplicate-event handling used previously identified persistent duplicate-track pairs together with temporal and spatial evidence. Direct crossings were prioritized over matching inferred events within 0.35 seconds and 120 pixels. A further suppression rule required the same crossing direction, a time difference of at most 0.18 seconds, crossing locations within 60 pixels, and strong bounding-box overlap or containment. Remaining suspected duplicate events within a one-second window were flagged for review rather than automatically removed. The retained events were aggregated by vehicle class, location, and observation period for comparison with the manual reference counts.

2.2. Aggregate Counting Validation

Human ground-truth (GT) counts were prepared by one observer, the first author, from the same six recordings used for automated counting. A Python playback interface displayed the predefined counting line and provided separate keys for motorcycles, passenger cars, buses, and trucks. Each side of the road was reviewed separately, and vehicles were entered manually by class during playback. Pausing, seeking, replay, and slower playback were used to inspect vehicle passages and recheck uncertain counts. The manual procedure did not require the observer to identify the exact frame at which an individual vehicle's bottom-center point crossed the line.

The manual records were reviewed by the same observer. No independent second annotation was available, so inter-observer agreement was not measured and the manual reference remained subject to observer error. A recorded timestamp indicated the video position at which the observer entered a count, rather than a precisely annotated crossing frame. The evaluation reported here therefore compared aggregate counts rather than using event-level temporal matching as its validation criterion.

Manual and automated records were aggregated using the same four classes, locations, and observation periods. Comparisons were reported by class, by daytime and nighttime period, and across all six recordings. For comparison group i, the signed count difference was calculated as

$$\Delta N_i = N_{sys,i} - N_{GT,i} \quad (1)$$

Here, $N_{sys,i}$ and $N_{GT,i}$ denote the automated and manual reference counts, respectively. Positive differences indicate overcounting and negative differences indicate undercounting relative to the manual reference. The absolute count difference was calculated as

$$|\Delta N_i| = \left|N_{sys,i} - N_{GT,i}\right| \quad (2)$$

The signed count error percentage relative to the manual reference was calculated as

$$CE_i = \frac{N_{sys,i} - N_{GT,i}}{N_{GT,i}} \times 100\% \quad (3)$$

Percentages were calculated only when the corresponding manual count was nonzero. Overall error was calculated from the summed system and GT counts, rather than by averaging class-level or recording-level percentages. Likewise, the overall absolute difference was calculated from the two total counts rather than by summing the absolute differences across classes.

2.3. Standardized Emission Estimation

System-generated and human-GT vehicle counts were converted separately into standardized CO_2 and CO estimates. The calculation followed the activity-times-emission-factor framework described in the EMEP/EEA guidebook [2], while the numerical vehicle-specific factors were adopted from the published Kalimalang corridor study [3]. Fuel type, engine characteristics, and emission-standard compliance were not measured from the CCTV recordings. The assigned factors were therefore treated as explicit assumptions rather than verified characteristics of the Semarang fleet.

A reference distance of one kilometer was assigned to each counted vehicle because individual travel distances along the road corridors were not established from the recordings. Applying the same distance to both count sources provided a common basis for comparing their estimates using factors expressed in grams per vehicle-kilometer. This was a standardization assumption, not the length of the visible road segment or a measured journey. For count source q, pollutant p, scenario s, and observation group i, the estimate was calculated as

$$E_{q,p,s,i} = d_{ref} \sum_c N_{q,c,i} EF_{p,c,s} \quad (4)$$

Here, q denotes the automated system or human GT, $E_{q,p,s,i}$ is the standardized emission estimate, $N_{q,c,i}$ is the number of vehicles in class c, and $EF_{p,c,s}$ is the corresponding emission factor. The reference distance was $d_{\mathrm{ref}} = 1$ km. The class index c covers buses, passenger cars, motorcycles, and trucks. Observation group i represents a location-period combination or an aggregate of recordings. Equation (4) produces estimates in grams, which were converted to kilograms by dividing by 1,000.

Six scenarios were formed by combining three passenger-car assumptions with two truck assumptions. S1 combined a medium diesel passenger car with a light cargo truck; S2, a medium diesel passenger car with a heavy truck; S3, a small gasoline passenger car with a light cargo truck; S4, a small gasoline passenger car with a heavy truck; S5, a medium gasoline passenger car with a light cargo truck; and S6, a medium gasoline passenger car with a heavy truck. The bus and motorcycle factors remained fixed. Within each scenario, the selected passenger-car factor was applied to all counted passenger cars

and the selected truck factor to all counted trucks. Table 2 lists the factors and their scenario applicability.

Table 2. Vehicle-specific CO_2 and CO emission factors used in sensitivity scenarios S1–S6 (all values in g/vehicle-kilometer)

Applicable scenarios	Vehicle class	Factor assumption	CO_2	CO	Source
S1–S6	Bus	Average bus factor	815.513	1.622	[3]
S1–S6	Motorcycle	Four-stroke motorcycle <250 cm³, Euro III	53.467	1.148	[3]
S1, S2	Passenger car	Medium diesel passenger car, Euro IV	168.739	0.092	[3]
S3, S4	Passenger car	Small gasoline passenger car, Euro IV	182.346	1.050	[3]
S5, S6	Passenger car	Medium gasoline passenger car, Euro IV	214.942	0.954	[3]
S1, S3, S5	Truck	Light diesel cargo vehicle N1-III, Euro IV	255.543	0.389	[3]
S2, S4, S6	Truck	Average heavy truck	639.210	1.419	[3]

Each scenario was evaluated separately using identical factors for the system and human-GT counts. The scenarios represented alternative factor assumptions, not observed fuel types, vehicle technologies, or the actual fleet composition in Semarang; their results were therefore not summed. The range across scenarios describes sensitivity to the selected passenger-car and truck factors, rather than a statistical confidence interval. The fixed bus and motorcycle factors were not varied in this analysis.

For example, the aggregate human-GT counts under S1 comprised 51 buses, 1,963 passenger cars, 5,424 motorcycles, and 229 trucks. Using the one-kilometer reference distance and the corresponding CO_2 factors gives

$$\begin{aligned} E_{\mathrm{GT,CO_2,S1,all}} &= (51 \times 815.513) + (1{,}963 \times 168.739) \\ &\quad + (5{,}424 \times 53.467) + (229 \times 255.543) \\ &= 721{,}350.175 \text{ g } \mathrm{CO_2} \end{aligned} \tag{5}$$

Dividing by 1,000 yields 721.350 kg CO_2 after rounding to three decimal places, as reported in Table 6. The signed difference between estimates derived from the two count sources was calculated as

$$\Delta E_{p,s,i} = E_{sys,p,s,i} - E_{GT,p,s,i} \tag{6}$$

Here, $E_{sys,p,s,i}$ and $E_{GT,p,s,i}$ denote the system-based and human-GT-based estimates. Positive values indicate a higher system-based estimate, while negative values indicate a lower estimate. The relative difference was calculated as

$$RE_{p,s,i} = \frac{E_{sys,p,s,i} - E_{GT,p,s,i}}{E_{GT,p,s,i}} \times 100\% \tag{7}$$

Relative differences were calculated only for groups with a nonzero human-GT-based estimate. All calculations used unrounded values, with conversion to kilograms and rounding performed for reporting. Since both count sources used identical factors and distance within each scenario, differences between their estimates arose from the class-specific count differences. The common distance scales the absolute estimates but cancels from their relative difference.

The human-GT-based emission estimate served as a count-derived reference rather than a measurement of emitted pollutant mass. Both results are standardized comparative estimates based on the adopted factors and one-kilometer reference distance. They do not represent directly measured exhaust emissions, ambient pollutant concentrations, or total emissions along the observed road corridors.

3. RESULTS AND DISCUSSIONS

This section presents the final detector evaluation, aggregate vehicle-count comparisons, daytime and nighttime results, and standardized emission estimates. Detection metrics and count differences are interpreted separately before examining how the class-specific count differences affect the emission estimates.

3.1. Vehicle Detection Performance

The final corrective YOLO11m detector was evaluated on the 600-image test subset containing 13,542 annotated vehicle instances. These images were excluded from the training and validation lists. Table 3 reports the overall and class-specific detection results.

Table 3. Detection performance of the final corrective YOLO11m model

Vehicle class	Precision	Recall	mAP_{50}	$mAP_{50–95}$
Bus	0.891	0.810	0.889	0.807
Passenger car	0.940	0.930	0.979	0.883
Motorcycle	0.921	0.928	0.973	0.800
Truck	0.862	0.912	0.946	0.844
Overall	0.904	0.895	0.947	0.833

The overall precision and recall were 0.904 and 0.895, respectively. The mAP_{50} of 0.947 indicates strong detection performance at an IoU threshold of 0.50, while the lower $mAP_{50–95}$ of 0.833 reflects the greater difficulty of meeting stricter localization thresholds. Passenger cars achieved the highest values across all four metrics. Buses had the lowest recall, at 0.810, whereas trucks had the lowest precision, at 0.862. Motorcycles achieved a recall of 0.928 and an mAP_{50} of 0.973, but their $mAP_{50–95}$ of 0.800 was the lowest among the four classes. In Figure 3, the labels mobil, motor, and truk denote passenger car, motorcycle, and truck, respectively.

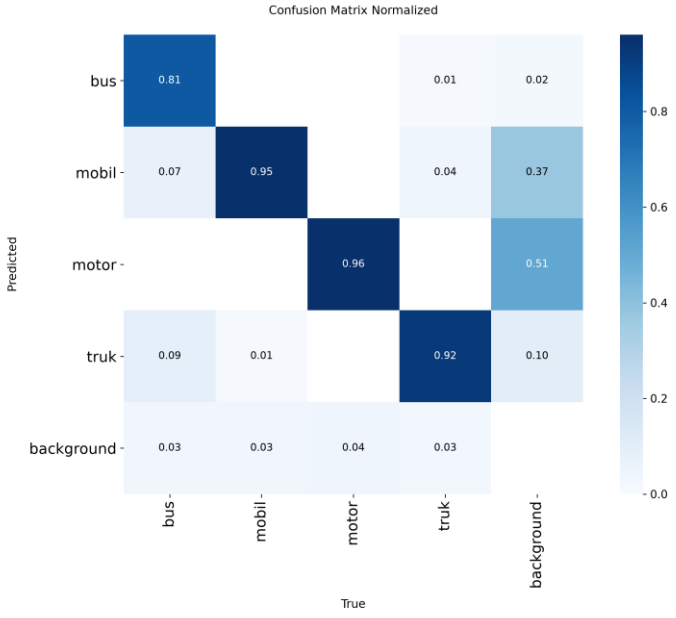


Figure 3. Normalized confusion matrix for the locked held-out test set.

Figure 3 shows that buses were confused with passenger cars and trucks. This pattern complements the class-level metrics by identifying categories that were difficult to distinguish. It does not identify whether a vehicle subsequently produced a missed, duplicated, or incorrectly classified crossing event. Those outcomes depend on track association, class stabilization, and the counting rules in addition to the detector.

Findings from research on heterogeneous traffic [19] and comparisons of vehicle detectors [20] likewise support examining class-level behavior. Direct numerical comparisons with those studies are not used here because datasets and evaluation procedures differ. The present image test also shares camera sources with the development partitions, and its annotation-review dependency is described in Section 2.1. The results therefore characterize detection on the represented image sources, while counting performance is evaluated separately below.

3.2. Aggregate Vehicle-Counting Performance

Table 4 compares automated and human-GT counts across the six recordings using the measures defined in Section 2.2. The overall absolute difference is calculated from the two total counts, rather than by summing the class-specific absolute differences.

Table 4. Aggregate vehicle-counting performance across all CCTV recordings

Vehicle class	Human-GT count	System count	Signed difference	Absolute difference	Count error (%)
Bus	51	45	−6	6	−11.76
Passenger car	1,963	2,027	+64	64	+3.26
Motorcycle	5,424	5,181	−243	243	−4.48
Truck	229	191	−38	38	−16.59
Overall	7,667	7,444	−223	223	−2.91

The system counted 7,444 vehicles compared with 7,667 in the manual reference, producing a net difference of −223 vehicles (−2.91%). Motorcycles contributed the largest numerical shortfall, with 243 fewer vehicles, while trucks had the largest relative difference, at −16.59%. Buses were undercounted by six vehicles (−11.76%), and passenger cars were overcounted by 64 (+3.26%). The smaller bus reference count of 51 makes its percentage particularly sensitive to a small number of counting discrepancies.

The passenger-car overcount partly offset the undercounts in the other classes. Consequently, the −2.91% result describes net agreement in the total and should not be interpreted as 97.09% vehicle-level accuracy. Agreement in the total does not establish that individual vehicle events were counted correctly or assigned the correct class.

The counting results also cannot be predicted directly from Table 3. For example, trucks had a detector recall of 0.912 but an aggregate count difference of −16.59%, whereas motorcycles had a smaller relative undercount despite producing the largest numerical shortfall. The image and video evaluations use different observations and assess different stages of the pipeline. The available aggregate reference counts do not isolate how much of each discrepancy arose from detection, track continuity, class assignment, duplicate-event handling, or manual counting.

3.3. Daytime and Nighttime Counting Performance

Table 5 compares the three daytime recordings with the three nighttime recordings, retaining the same four vehicle classes.

Table 5. Daytime and nighttime vehicle-counting performance

Period	Vehicle class	Human GT	System	ΔN	|ΔN|	Error (%)
Daytime	Bus	31	28	−3	3	−9.68
	Passenger car	1,066	1,087	+21	21	+1.97
	Motorcycle	2,562	2,583	+21	21	+0.82
	Truck	102	96	−6	6	−5.88
	Overall	3,761	3,794	+33	33	+0.88

Period	Vehicle class	Human GT	System	ΔN	|ΔN|	Error (%)
	Bus	20	17	−3	3	−15.00
	Passenger car	897	940	+43	43	+4.79
Nighttime	Motorcycle	2,862	2,598	−264	264	−9.22
	Truck	127	95	−32	32	−25.20
	Overall	3,906	3,650	−256	256	−6.55

Daytime counts differed by +33 vehicles (+0.88%), whereas nighttime counts differed by −256 (−6.55%). Motorcycles changed from a slight daytime overcount of 21 vehicles to a nighttime undercount of 264. Trucks were undercounted in both periods, with the larger relative difference occurring at night (−25.20%). Passenger cars were overcounted in both periods. These class-level differences show why the period totals should be interpreted together with their composition.

The discrepancy also varied across locations. Daytime differences were +32 vehicles at Fatmawati, −1 at JPO Ahmad Yani, and +2 at Tugu Muda 2; nighttime differences were +14, −177, and −93, respectively. The nighttime JPO Ahmad Yani recording contained shortfalls of 174 motorcycles and 12 trucks, while Tugu Muda 2 contained shortfalls of 101 motorcycles and 11 trucks. These two recordings accounted for the main nighttime undercount, partly offset by the positive difference at Fatmawati.

Headlight glare, reduced visibility, and changing luminance are recognized challenges for traffic vision [17], [18]. However, the observed period differences cannot be attributed to illumination alone. Traffic composition, density, occlusion, and scene characteristics also varied. Each location contributed only one clip per period, so the results do not establish a general nighttime effect or its statistical significance. The percentages are descriptive results for these six recordings and should not be extrapolated to all Pantau Semar cameras or observation periods.

3.4. Standardized CO_2 and CO Emission Estimates

System-generated and human-GT counts were converted into standardized CO_2 and CO estimates using the same emission factors and a reference distance of one kilometer per counted vehicle. Table 6 presents the aggregate results under the six scenarios defined in Table 2.

Table 6. Aggregate standardized CO_2 and CO emission estimates across all CCTV recordings

Scenario	Human-GT CO_2 (kg)	System CO_2 (kg)	Human-GT CO (kg)	System CO (kg)
S1	721.350	704.553	6.579	6.282
S2	809.210	777.834	6.815	6.478
S3	748.061	732.135	8.460	8.223
S4	835.920	805.415	8.696	8.420
S5	812.047	798.207	8.271	8.029
S6	899.906	871.487	8.507	8.226

Human-GT-based CO_2 estimates ranged from 721.350 to 899.906 kg, compared with 704.553 to 871.487 kg for the system-based estimates. S1 produced the lowest CO_2 values and S6 the highest. The corresponding CO ranges were 6.579–8.696 kg and 6.282–8.420 kg, with the lowest values in S1 and the highest in S4. These patterns reflect the passenger-car and truck factors assigned to each scenario, while the bus and motorcycle factors remained fixed. The ranges describe sensitivity to the selected factor assumptions; they are not statistical confidence intervals or bounds on actual emissions.

The system-based estimates were lower for both pollutants in all six scenarios. Because each vehicle class had a different emission factor, the magnitude of the difference depended on the distribution of counting errors rather than only on the aggregate count difference. The scenarios were evaluated separately as alternative assumptions and were not summed. None represents a measured fuel mix or vehicle-technology distribution in Semarang, and the one-kilometer reference distance does not establish actual vehicle travel distances or total emissions along the observed corridors.

3.5. Effect of Counting Error on Emission Estimates

Table 7 reports the signed and relative differences between the system-based and human-GT-based standardized emission estimates. Negative values indicate lower estimates derived from the automated counts. The human-GT-based estimates provide a comparison calculated from manual counts; they are not independently measured exhaust emissions.

Table 7. Effect of aggregate vehicle-counting differences on standardized emission estimates

Scenario	CO_2 difference (kg)	CO_2 difference (%)	CO difference (kg)	CO difference (%)
S1	−16.797	−2.33	−0.298	−4.52
S2	−31.376	−3.88	−0.337	−4.94
S3	−15.926	−2.13	−0.236	−2.79
S4	−30.505	−3.65	−0.275	−3.17
S5	−13.840	−1.70	−0.242	−2.93
S6	−28.419	−3.16	−0.282	−3.31

Note: Differences were calculated using unrounded emission estimates. Emission differences are reported to three decimal places, and relative differences to two decimal places.

To identify the contributions of individual vehicle classes, the signed count differences in Table 4 were multiplied by the corresponding S1 factors in Table 2 and the one-kilometer reference distance. The resulting differences were converted from grams to kilograms. Table 8 presents this decomposition for S1. S1 was used as an illustrative case to explain how class-specific differences combine into the aggregate emission difference, while Tables 6 and 7 retain the results for all six scenarios. It was not selected as the most accurate scenario or as a representation of the observed fleet. The same decomposition can be applied to the other scenarios using their respective passenger-car and truck factors.

Table 8. Class-specific contributions to system–human-GT differences in standardized emission estimates under S1

Vehicle class	Count difference (system − GT)	CO_2 contribution (kg)	CO contribution (kg)
Bus	−6	−4.893	−0.010
Passenger car	+64	+10.799	+0.006
Motorcycle	−243	−12.992	−0.279
Truck	−38	−9.711	−0.015
Overall	−223	−16.797	−0.298

Note: Overall values were calculated before rounding. S1 is a sensitivity assumption, not a description of the observed fleet composition.

As shown in Table 8, the system counted 64 more passenger cars than the human GT, but counted 6 fewer buses, 243 fewer motorcycles, and 38 fewer trucks. Under S1, passenger-car overcounting increased the system-based CO_2 estimate by 10.799 kg. This positive contribution partly offset the lower estimates associated with buses (−4.893 kg), motorcycles (−12.992 kg), and trucks (−9.711 kg). Together, these class-specific contributions produced a system-based CO_2 estimate 16.797 kg below the human-GT-based estimate. Thus, the overall difference cannot be inferred from the net vehicle-count difference alone.

The contributions differed for CO because its emission factors have a different pattern across vehicle classes. Under S1, the motorcycle count difference contributed −0.279 kg CO, accounting for most of the total −0.298 kg difference. The passenger-car overcount contributed only +0.006 kg CO

under the medium diesel passenger-car assumption used in S1. These results illustrate why the same set of vehicle-count differences can have different effects on CO_2 and CO estimates.

The relationship was also visible within individual recordings. During daytime at JPO Ahmad Yani, the system counted only one fewer vehicle than the human GT, yet its S1-based CO_2 estimate was 1.210 kg higher. At Fatmawati during nighttime, the system counted 14 more vehicles, yet its S1-based CO_2 estimate was 1.627 kg lower. In both cases, differences among vehicle classes changed the direction of the emission-estimate difference relative to the total count difference. Close agreement in total counts can therefore conceal class-specific differences relevant to emission estimation.

Across all six recordings, the aggregate vehicle-count difference was −2.91%. In comparison, Table 7 shows that the CO_2 estimates were 1.70–3.88% lower and the CO estimates were 2.79–4.94% lower than their human-GT-based counterparts. The smallest CO_2 shortfall occurred in S5 and the largest in S2; the smallest CO shortfall occurred in S3 and the largest in S2. Using the larger truck factors in S2, S4, and S6 increased the negative contribution from the truck undercount. The higher passenger-car CO factors in the gasoline passenger-car scenarios increased the positive contribution from passenger-car overcounting and partly offset negative contributions from other classes.

Yavuz et al. [14] also reported lower estimates derived from deep-learning-based vehicle activity than from ground-truth data. Direct numerical comparison is inappropriate because the studies used different models, datasets, pollutants, and estimation procedures. Suparwito et al. [15] demonstrated the conversion of class-specific computer-vision counts into CO estimates. In the present study, applying the same emission factors and one-kilometer reference distance to automated and human-GT counts makes the effect of their class-specific counting differences explicit.

These results indicate that evaluating a counting system for emission estimation requires attention to individual vehicle classes and their assigned pollutant factors. Under S1, the motorcycle shortfall dominated the CO difference, whereas buses and trucks also made substantial negative contributions to CO_2 despite their smaller numerical count differences. Accordingly, reducing the net total-count difference alone would not necessarily reduce the emission difference. The relevant class contributions would need to be assessed for the pollutant and factor assumptions used.

The findings remain descriptive and conditional on the evaluated data and assumptions. Six approximately 15-minute recordings from three locations provide limited evidence about variation across dates, traffic conditions, and other camera locations. They do not establish citywide performance or statistical certainty for daytime–nighttime differences. Human-GT counts were prepared by one observer, without an independent second-observer assessment, and the reported validation did not include event-level temporal matching. Aggregate differences therefore cannot identify how many individual vehicles were correctly counted or separate missed events, duplicate events, class confusion, tracking errors, and manual-counting errors.

The emission calculations introduce additional limitations. Vehicle fuel types, technologies, operating speeds, and actual travel distances were not measured, and the six scenarios vary only the selected passenger-car and truck factors while keeping the bus and motorcycle factors fixed. They therefore do not capture the full uncertainty of actual traffic emissions. Applying common factors and a common reference distance isolates the arithmetic effect of the observed count differences within each scenario, but agreement between the two calculated estimates does not establish accuracy against physical emissions. The results should consequently be interpreted as standardized comparative estimates, not measurements of exhaust emissions, ambient pollutant concentrations, or total corridor emissions.

4. CONCLUSION

This study evaluated an existing YOLO11m detector combined with ByteTrack, track-level class stabilization, and line-crossing counting to examine how vehicle-counting differences affect standardized CO_2 and CO estimates from Pantau Semar CCTV recordings. On the 600-image test subset, the final detector achieved a precision of 0.904, recall of 0.895, mAP_{50} of 0.947, and $mAP_{50–95}$ of 0.833. These bounding-box metrics were evaluated separately from vehicle counting. Across six recordings from three locations, the system counted 7,444 vehicles compared with 7,667 in the human GT, a difference of −223 vehicles (−2.91%). The daytime and nighttime differences were +0.88% and −6.55%, respectively. These descriptive differences cannot be attributed solely to illumination.

Applying identical emission factors and a one-kilometer reference distance to both count sources produced system-based estimates 1.70–3.88% lower for CO_2 and 2.79–4.94% lower for CO across six scenarios evaluated separately. The analysis showed that the direction and magnitude of the emission differences depended on the affected vehicle classes and their assigned factors. Passenger-car overcounting partly offset the effects of undercounting other classes, while the motorcycle shortfall dominated the CO difference under S1. Consequently, agreement in total vehicle counts alone did not establish comparable agreement in standardized emission estimates.

The findings are limited to the evaluated recordings and factor assumptions. The six scenarios represent alternative emission-factor assignments rather than the observed fleet composition, and one kilometer is a common reference distance rather than a measured travel distance. The estimates are standardized comparisons, not measurements of exhaust emissions, ambient pollutant concentrations, or total corridor emissions. Generalization is constrained by the small video sample, shared camera sources across image partitions, and the use of test annotations during development-data quality review. Counting validation also relied on one observer and aggregate comparisons without event-level temporal matching. Future work should expand camera locations and observation periods, use camera- or time-separated image evaluation with an independently maintained test set, strengthen manual validation, and incorporate verified local fleet characteristics and travel activity.

5. REFERENCES

[1] H. Boogaard *et al.*, "Long-term exposure to traffic-related air pollution and selected health outcomes: A systematic review and meta-analysis," *Environ. Int.*, vol. 164, p. 107262, Apr. 2022, doi: 10.1016/j.envint.2022.107262.

[2] European Environment Agency, "EMEP/EEA air pollutant emission inventory guidebook 2023 : technical guidance to prepare national emission inventories," Publications Office of the European Union, Luxembourg, 2023. doi: 10.2800/795737.

[3] Ivan Imanuel, Prihartono, and Krisnanda Teguh Saputra, "Traffic-Based Emission Assessment and Modal Shift Analysis Using a Harmonized Emission Factor Approach: A Case Study of the Kalimalang Corridor," *PRESUNIVE Civil Engineering Journal*, vol. 4, no. 1, pp. 26–39, Apr. 2026, doi: 10.33021/pcej.v4i1.6647.

[4] S. B. Neamah and A. A. Karim, "Vehicle Classification and Counting for Traffic Analysis based on Single-stage YOLOv8 Model," *Iraqi Journal of Computers, Communications, Control and Systems Engineering*, vol. 24, no. 2, pp. 41–53, Jun. 2024, doi: 10.33103/uot.ijccce.24.2.4.

[5] P. Premaratne, I. Jawad Kadhim, R. Blacklidge, and M. Lee, "Comprehensive review on vehicle Detection, classification and counting on highways," *Neurocomputing*, vol. 556, p. 126627, Aug. 2023, doi: 10.1016/j.neucom.2023.126627.

[6] K. A. Redmill, E. Yurtsever, R. G. Mishalani, B. Coifman, and M. R. McCord, "Automated Traffic Surveillance Using Existing Cameras on Transit Buses," *Sensors*, vol. 23, no. 11, p. 5086, May 2023, doi: 10.3390/s23115086.

[7] Z. Hu, W. H. K. Lam, S. C. Wong, A. H. F. Chow, and W. Ma, "Turning traffic surveillance cameras into intelligent sensors for traffic density estimation," *Complex and Intelligent Systems*, vol. 9, no. 6, pp. 7171–7195, Jun. 2023, doi: 10.1007/s40747-023-01117-0.

[8] M. Majumder and C. Wilmot, "Automated Vehicle Counting from Pre-Recorded Video Using You Only Look Once (YOLO) Object Detection Model," *J. Imaging*, vol. 9, no. 7, p. 131, Jun. 2023, doi: 10.3390/jimaging9070131.

[9] S. A. Kamal, A. Gaber, and M. A. Shehata, "Automated Vehicle Counting and Speed Estimation Using YOLOv8 and Computer Vision," *Journal of Al-azhar University Engineering Sector*, vol. 19, no. 73, pp. 1382–1395, Oct. 2024, doi: 10.21608/auej.2024.280422.1642.

[10] A. Gomaa, T. Minematsu, M. M. Abdelwahab, M. Abo-Zahhad, and R. Taniguchi, "Faster CNN-based vehicle detection and counting strategy for fixed camera scenes," *Multimed. Tools Appl.*, vol. 81, no. 18, pp. 25443–25471, Mar. 2022, doi: 10.1007/s11042-022-12370-9.

[11] P. N. Huu, B. N. Anh, and Q. T. Minh, "Proposing Smart System for Detecting and Monitoring Vehicle Using Multiobject Multicamera Tracking," *International Journal of Digital Multimedia Broadcasting*, vol. 2024, p. 6667738, Apr. 2024, doi: 10.1155/2024/6667738.

[12] A. Valdes-Montenegro, F. Gonzalez-Olivardia, S. Thepanondh, and C. Pinzon-Acosta, "Estimation of on-road mobile emissions based on the vehicle technology in a high-traffic avenue in Panama City, Panama," in *E3S Web of Conferences*, EDP Sciences, May 2024, p. 01003. doi: 10.1051/e3sconf/202453001003.

[13] C. Pinzon-Acosta, F. Gonzalez-Olivardia, M. Hidalgo-Rodríguez, A. Valdés-Montenegro, J. Asprilla-Gonzalez, and E. Cruz, "Spatial estimation of on-road mobile emissions during morning hours using machine learning: A case study in Panama city," *Ecol. Inform.*, vol. 95, p. 103731, Mar. 2026, doi: 10.1016/j.ecoinf.2026.103731.

[14] E. Yavuz, A. Öztürk, N. G. N. Balkanlı, Ş. N. Engin, and S. L. Kuzu, "Development of Artificial Intelligent-Based Methodology to Prepare Input for Estimating Vehicle Emissions," *Applied Sciences (Switzerland)*, vol. 14, no. 23, p. 11175, Nov. 2024, doi: 10.3390/app142311175.

[15] H. Suparwito, B. H. G. Prakoso, R. A. Kumalasanti, and A. M. Polina, "Real-Time Vehicle Detection and Air Pollution Estimation Using YOLOv9," *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, vol. 14, no. 1, pp. 23–30, Jan. 2025, doi: 10.32736/sisfokom.v14i1.2329.

[16] Y. Guo, Y. Yamamoto, H. Yaginuma, and Y. Taniguchi, "Vehicle detection in CCTV with global-guided self-attention and convolution," *Complex and Intelligent Systems*, vol. 11, p. 458, Sep. 2025, doi: 10.1007/s40747-025-02073-7.

[17] Y. Iwao, Y. Yamamoto, H. Yaginuma, and Y. Taniguchi, "Traffic volume measurement using nonlinear count-lines," *Mach. Vis. Appl.*, vol. 36, p. 52, Mar. 2025, doi: 10.1007/s00138-025-01668-x.

[18] Y. Wei, Z. Zeng, T. He, S. Yu, Y. Du, and C. Zhao, "An Adaptive Vehicle Detection Model for Traffic Surveillance of Highway Tunnels Considering Luminance Intensity," *Sensors*, vol. 24, no. 18, p. 5912, Sep. 2024, doi: 10.3390/s24185912.

[19] D. M. Tan and L. M. Kieu, "TRAMON: An automated traffic monitoring system for high density, mixed and lane-free traffic," *IATSS Research*, vol. 47, no. 4, pp. 468–481, Oct. 2023, doi: 10.1016/j.iatssr.2023.10.001.

[20] A. Sundaresan Geetha, M. A. R. Alif, M. Hussain, and P. Allen, "Comparative Analysis of YOLOv8 and YOLOv10 in Vehicle Detection: Performance Metrics and Model Efficacy," *Vehicles*, vol. 6, no. 3, pp. 1364–1382, Aug. 2024, doi: 10.3390/vehicles6030065.

[21] I. C. Wicaksana, R. Anggi Pramunendar, G. W. Saraswati, and G. Alfa Trisnapradika, "Skin Lesion Classification Using YOLOv11 on the HAM10000 Dataset," *Inform : Jurnal Ilmiah Bidang Teknologi Informasi dan Komunikasi*, vol. 10, no. 1, pp. 45–52, Jan. 2025, doi: 10.25139/inform.v10i1.9310.

[22] G. Jocher and J. Qiu, "Ultralytics YOLO11," *Ultralytics*. Accessed: Sep. 27, 2026. [Online]. Available: https://docs.ultralytics.com/models/yolo11/

[23] Y. Zhang *et al.*, "ByteTrack: Multi-object Tracking by Associating Every Detection Box," in *Computer Vision – ECCV 2022*, Springer, 2022, pp. 1–21. doi: 10.1007/978-3-031-20047-2_1.