Automated Large-Scale Labeling of Brazilian Urban Traffic Images with Model Ensembles

Carlos Loureiro, William Harada, Elloá B. Guedes, Carlos Maurício S. Figueiredo

[1]Grupo de Pesquisa em Sistemas Inteligentes
Universidade do Estado do Amazonas
Av. Darcy Vargas, 1200 – Manaus – Amazonas
{cenlf.eng22, wmch.eng22, ebgcosta, cfigueiredo}@uea.edu.br

Abstract. *Custom dataset creation remains a time-consuming bottleneck for training detection models. To advance urban mobility analysis in Brazil, we propose an autonomous, large-scale labeling pipeline using a vision model ensemble (Grounding DINO, OWLv2, YOLOv11) to annotate traffic surveillance images. Training the RT-DETR architecture on the resulting dataset yielded an mAP@0.5:0.95 of* 0.5265 *(*F_1*:* $73.59\,\%$*) on synthetic labels, and* 0.4651 *(*F_1*:* $74.02\,\%$*) on human-supervised ground truth. This strong real-world generalization offers a highly scalable alternative to manual annotation for intelligent urban traffic monitoring.*

1. Introduction

Automated vehicular traffic monitoring is an essential component of intelligent transportation systems applied to the context of Smart Cities (SCs) [Myagmar-Ochir and Kim 2023]. In Brazil, particularly, the circulating vehicle fleet has surpassed 129 Mi units [Brasil 2025], and escalating mobility challenges demand scalable and adaptive technological solutions. In this scenario, efforts have been directed toward the development of intelligent, Computer Vision-based systems capable of assisting real-time urban traffic management [Loureiro et al. 2025, Barbado et al. 2022, Souza et al. 2025], thereby enabling viable urban mobility solutions for SCs within the national context.

For these intelligent systems to operate satisfactorily in real-world environments, it is crucial that they are appropriately trained using data originating from the specific problem context. Regarding state-of-the-art models based on Deep Learning methods and techniques, there is a demand for massive amounts of labeled data, given that the Supervised Learning paradigm is predominant for urban traffic management tasks, such as vehicle type classification and vehicle detection in urban scenes, among others. Beyond mere quantity, the quality of these labels must also be highlighted; they must accurately reflect the real problem context to serve as a reliable learning foundation for the models.

In the domain of object detection, human annotations are frequently susceptible to noise, such as missing or ill-fitted bounding boxes, and incorrect occlusions [Chan et al. 2024]. It is estimated that the data corruption rate in real-world datasets ranges from $8.0\,\%$ to $38.5\,\%$, imposing a significant challenge, as traditional Machine Learning models assume clean distributions and accurate labels [Freire et al. 2024]. This discrepancy can induce biases that compromise detection performance in practical scenarios, since low-quality annotations degrade model performance in specific tasks [Velesaca et al. 2026]. Although manual labeling is often viewed as an alternative to achieve higher accuracy, it becomes unfeasible with the growing demand for data due to its high costs and the intense requirements for human, financial, and temporal resources [Wu 2024].

To mitigate the challenges of manual annotation and advance vehicle detection strategies for SC applications tailored to Brazilian urban traffic, this work introduces a large-scale labeling strategy for realistic images within this domain. The proposed approach utilizes diverse pre-trained models to perform inferences and unifies their predictions through a non-maximum suppression technique. As a result of this strategy, a dataset comprising $10\,000$ images from three Brazilian cities was developed, annotated across four categories, aiming to serve as a benchmark for vehicle detection in Brazilian urban traffic. To validate the proposed methodology, a state-of-the-art model Transformer-based detector was trained and achieved competitive performance, thereby consolidating a contribution that fosters numerous subsequent solutions for the SC ecosystem in Brazil.

To present the proposed approach, this paper is organized as follows. Related works are presented in Section 2. The proposed solution and its components are detailed in Section 3. The results are reported in Section 4. Finally, concluding remarks and future perspectives are discussed in Section 5.

2. Related Work

To comprehensively understand existing solutions and identify trends and gaps, an analysis of search results from Google Scholar between 2020 and 2026 was conducted. The inclusion criteria encompassed Computer Vision-based works for object detection in Brazilian urban traffic.

The work by Santos *et al.* [2023] addresses license plate detection within Brazilian universities. To this end, they utilized two datasets available in the literature to train single-stage YOLO detectors for license plate detection. The resulting detections were then fed into a pre-trained PP-OCR model for character recognition. The proposed solution was validated in two realistic scenarios, achieving an accuracy of $76.6\,\%$, with variations in performance depending on the monitoring camera's angle or the license plate type (legacy standard or Mercosul). Other works in the literature, based on the training and evaluation of YOLO models for detecting Brazilian urban traffic components, employ similar strategies but focus on traffic lights [Júnior and Guedes 2023], road damages [Carvalho et al. 2022], pedestrians, and similar elements.

In the context of vehicle monitoring, the work by Santos *et al.* [2020] explores vehicle detection and counting on Brazilian highways. The adopted methodology consists of a two-stage pipeline, where a pre-trained single-stage YOLO detector performs vehicle detection, and subsequently, the Deep SORT multiple object tracking algorithm re-identifies each vehicle across frames. This approach maintains the resilience of the vehicle count under adverse conditions such as temporary occlusions. The method was validated on public datasets comprising elevated-view videos, as well as on a proprietary database from the National Department of Transport Infrastructure (DNIT), which consists of side-view recordings more susceptible to inter-vehicle occlusions. In these scenarios, a precision of $99.15\,\%$ was observed on the public datasets and rates exceeding $90\,\%$ on the proprietary dataset, with the performance drop attributed primarily to the side-view perspective of the cameras.

In the same context of vehicle monitoring, yet focused on urban traffic, the work by Loureiro *et al.* [2025] proposes a real-time monitoring solution supported by the camera infrastructure provided by Brazilian municipalities. As an initial step, the authors constructed a custom dataset comprising 1000 images and 6932 annotations distributed across four vehicle classes, collected from different regions of the country. Upon this dataset, they evaluated the

performance of YOLOv11 variants, where the Medium version stood out after hyperparameter fine-tuning by achieving an mAP@0.5:0.95 of 0.7372 at 94 FPS. The model was subsequently integrated into a continuous capture infrastructure that extracts frames from public streams provided by local municipal governments and automatically forwards them to the detector, thereby enabling the intelligent monitoring of urban roads.

From a detailed analysis of the related work, a prevalence in the use of various versions of YOLO models for object detection in Brazilian urban traffic is observed. To the best of our knowledge, no datasets intended to serve as a benchmark for the vehicle detection problem within the national context were found, nor has the evaluation of more recent state-of-the-art models, such as those based on Transformers, been considered. The solution proposed in the following section presents strategies to mitigate these identified gaps.

3. Proposed Solution

The methodology adopted in this work encompassed stages focused on the development of an automatic image labeling method. The first stage consisted of dataset acquisition, performed according to the methodology of Loureiro *et al.* [2025], which leverages the traffic camera infrastructure made available by municipal governments on the internet, resulting in a dataset composed of traffic images from three Brazilian municipalities (Section 3.1). Next, the models comprising the detector ensemble were defined (Section 3.2), as well as the tuning set used to calibrate their hyperparameters. Once the ensemble composition was determined, a common taxonomy was established to standardize the predictions of the models within the classes of interest (Section 3.3). Subsequently, the detectors were employed to generate independent predictions for each image in the dataset, which were later unified into a single set of annotations through a non-maximum suppression strategy to remove duplicates. Finally, the resulting annotations constituted a new dataset upon which the RT-DETR model was trained to assess its performance (Section 4). This evaluation was conducted across two distinct test partitions comprising a tuning subset of human-annotated images and an independent subset of automatically labeled images, allowing for the assessment of performance differences between human-validated data and ensemble-generated labels.

3.1. Experimental Data

The acquisition of the experimental data followed the procedure employed by Loureiro et al. [2025], capturing images from real-time video streams of traffic cameras made available online by municipal governments. The collected dataset comprises $10\,000$ images from three cities: São Paulo (SP), Rio de Janeiro (RJ), and Manaus (AM), selected because they rank among the seven largest cities in the country and possess the aforementioned streaming infrastructure.

Figure 1 presents the distribution of the samples according to the capture time for each municipality, demonstrating that the images were recorded over a 24-hour period and that all municipalities contribute samples across all time intervals, thereby encompassing diverse lighting conditions. The images were also acquired from different perspectives within each municipality, originating from 5 cameras in São Paulo, 3 in Rio de Janeiro, and 10 in Manaus, installed at distinct traffic monitoring stations. This temporal and spatial diversity yields a highly varied dataset that covers multiple realistic scenarios, comprising images with diverse framings, traffic levels, and weather conditions, examples of which are presented in Figure 2.

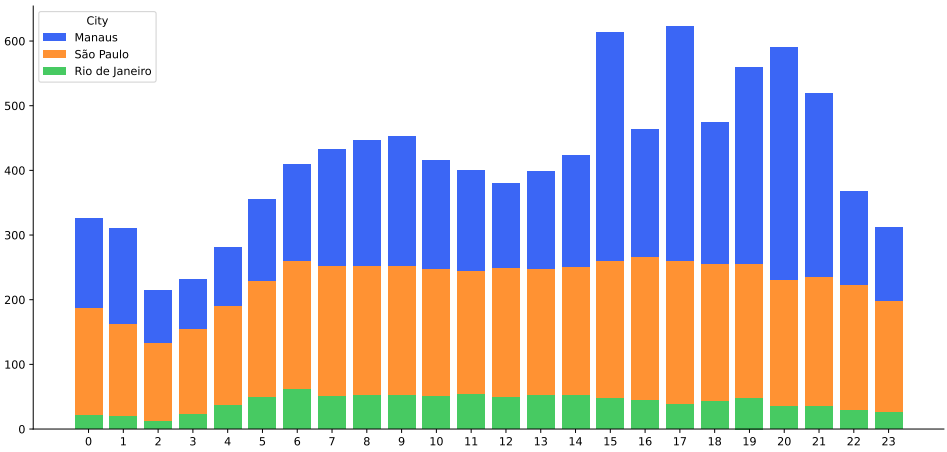


Figure 1: Histogram of the samples by city and time.



(a) Manaus **(b) São Paulo** **(c) Rio de Janeiro**

Figure 2: Sample image captures in each city.

3.2. Models Adopted

In this work, the following *Deep Learning* models were employed:

1. **YOLO (You Only Look Once) Family**. Operating as a single-stage detector, YOLO approaches object localization and classification as a direct spatial regression problem [Redmon et al. 2016]. The architecture processes image features at multiple scales to refine overlapping predictions and employs intrinsic regularization techniques, such as Mosaic Augmentation, to improve performance on small-scale objects [Vijayakumar and Vairavasundaram 2024]. In this work we used YOLOv11 Medium with pre-trained weights for four classes of Brazilian traffic objects as provided by Loureiro *et al.*[2025];
2. **Grounding DINO**. This model aligns visual data with textual descriptions to enable the detection of objects present in its vocabulary, even if corresponding visual examples were absent during training. Trained on massive amounts of data, it exhibits strong generalization capabilities and provides flexibility for including new objects in the annotation process. Its base version was employed, which utilizes Swin-B as the image backbone and BERT-base as the textual backbone, as it yielded better-fitted bounding boxes compared to its tiny counterpart. [Liu et al. 2024];
3. **OWLv2**. Based on the multimodal CLIP architecture, the model employs a Vision Transformer (ViT) and a masked self-attention Transformer to encode visual and textual features in a decoupled manner, performing detection through late alignment in the embedding space. By utilizing the self-training paradigm, the architecture massively scales its training data, which grants it high generalization capabilities for detection.

For this work, the version equipped with a 14-patch ViT was adopted, whose dense image partitioning maximizes the extraction of visual details. Additionally, this specific version underwent fine-tuning using human annotations, a fundamental step to ensure the refinement of the generated bounding boxes [Minderer et al. 2023];

4. **Real-Time Detection Transformer (RT-DETR)**. This architecture reformulates object detection as a direct set prediction task. Initial feature extraction is performed by a convolutional backbone, followed by attention mechanisms that weights the spatial and contextual relationships of the image before the decoder generates the final inferences [Carion et al. 2020]. To overcome the efficiency limitations of traditional Transformers, RT-DETR introduces a highly efficient hybrid encoder that optimizes multi-scale processing by integrating convolutional and attention blocks [Zhao et al. 2024]. In this work, this model is not employed for label generation; rather, it was selected as the target architecture for supervised training and performance evaluation, serving as a proxy to estimate the quality of the developed dataset. The RT-DETR Large version was utilized.

3.3. Class Taxonomy

The proposed method processes each image using an ensemble of detectors comprising two Vision-Language Models (VLMs), which utilize open-vocabulary textual classes, and a domain-specific YOLO model, which outputs fixed identifiers. To reconcile these incompatible prediction formats and enable the fusion of the ensemble's outputs, a unifying taxonomy was defined based on the four original YOLO training classes: Car, Motorcycle, Bus, and Vehicle. As presented in Table 1, this taxonomy maps 18 distinct textual descriptions to these four domain classes. These descriptions were subsequently used to formulate the prompts in the specific formats required by the VLMs to perform the detections.

Table 1: Taxonomy employed to ensure compatibility between the VLMs and the YOLO model.

YOLO Classes	Open-World Vocabulary
Car	*car, pickup, suv*
Motorcycle	*motorcycle, motorbike, scooter, moped*
Bus	*bus, school bus, microbus, shuttle bus, minibus, transit bus, passenger van*
Vehicle	*semi truck, box truck, lorry, cargo van, delivery van, ambulance, fire engine, police cruiser*

3.4. Model Parameterization and Fusion Strategy

Since the models performed inferences on out-of-distribution scenarios and on low-resolution data, the detection confidence scores were generally low, resulting in multiple redundant detections for the same objects. To maximize the performance of the detectors, a tuning set comprising 300 images sampled from the general dataset was established. This subset was utilized to calibrate hyperparameters such as the confidence score, the IoU threshold, and the text threshold (specific to Grounding DINO). Furthermore, this calibration phase was necessary to evaluate different aggregation strategies for the generated labels. To this end, Non-Maximum Suppression (NMS) and Weighted Box Fusion (WBF) [Solovyev et al. 2021] were tested, each with its respective configurations.

Each image was labeled by a human annotator and subsequently reviewed by a second annotator according to the classes of interest, aiming to minimize potential labeling errors within this set. Multiple parameter configurations were evaluated with the objective of maximizing the mAP and F_1-Score metrics, in an effort to reduce the number of false positives in the final annotation. Furthermore, the outputs of the best-performing model were evaluated through visual inspection to verify the quality of the generated labels.

4. Results and Discussion

The application of the proposed method resulted in the construction of a benchmark dataset for urban traffic object detection. The automatic generation of labels initially required calibrating the ensemble of detectors, as inferences were performed on scenarios absent from the models' training sets. To this end, on a tuning set sampled from the initial dataset, the hyperparameters of each expert model were adjusted to jointly maximize the mAP and F_1-Score: Grounding DINO (`grounding-dino-base`) performed best with a confidence threshold of 0.36 and an NMS of 0.45; OWLv2 (`owlv2-l-p14-finetuned`) with a confidence of 0.33 and an NMS of 0.45; and the custom YOLO with a confidence of 0.33 and an NMS of 0.35.

To consolidate the predictions from the three models into a single set of annotations, Non-Maximum Suppression (NMS) was adopted with a global IoU threshold of 0.4 and a global confidence cutoff of 0.33, a configuration that maximized the precision and recall of the ensemble and yielded the metrics reported in Table 2. While YOLO and OWLv2 had good results among different classes and supported each other, the only positive result for Grounding DINO was the Car class.

Table 2: Maximized metrics obtained on the tuning set.

Model	$\mathbf{F_1}$-Score	mAP@0.5	mAP@0.5:0.95
Grounding DINO	0.5532	0.1570	0.1232
YOLOv11	0.6380	0.4025	0.2675
OWLv2	0.7737	0.5656	0.3379
Ensemble	**0.7975**	**0.6177**	**0.3989**

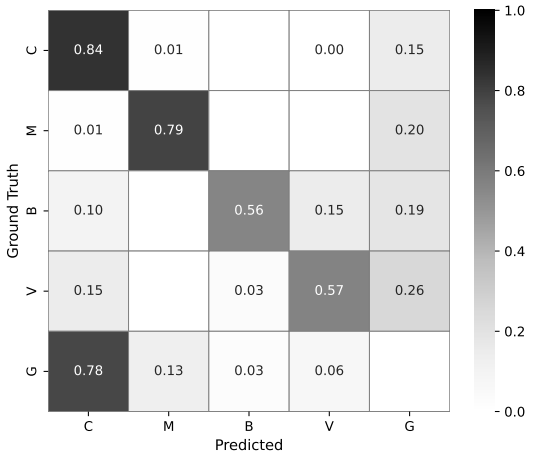


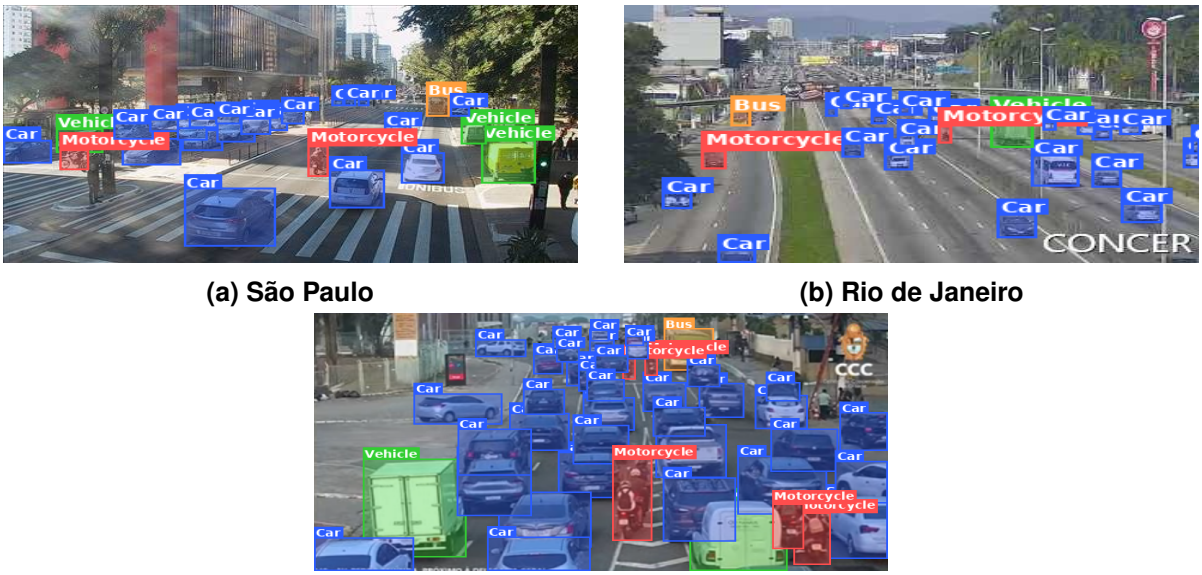
Figure 3: Ensemble confusion matrix on the tuning set

With the fusion configuration established, the automatic labeling of the initial dataset of 10 000 images was performed, resulting in 157 109 annotated instances and consolidating the proposed dataset, as illustrated in Figure 4. The class distribution of the generated dataset is markedly unbalanced, with the Car class accounting for 81.1 % of the instances, followed by Motorcycle (10.6 %), Vehicle (4.9 %), and Bus (3.5 %). The dataset was partitioned following a holdout strategy, allocating 60 % of the images for training, 10 % for validation, and 30 % for testing, as detailed in Table 3.

Table 3: Distribution of annotated objects by dataset partition

Class	Train	Validation	Test
Car	76 586	13 023	37 785
Motorcycle	10 111	1 641	4 852
Bus	3 243	567	1 626
Vehicle	4 459	750	2 466

The test partition comprised two subsets: (i) 300 images corresponding to the tuning set previously employed during the configuration of the zero-shot detector ensemble, and (ii) 2700 images randomly selected from the remaining data. Model performance was therefore analyzed both on the tuning subset and on the independent test subset, allowing the assessment of possible performance differences between images annotated by humans and the labels generated by the ensemble.



(a) São Paulo **(b) Rio de Janeiro**

(c) Manaus

Figure 4: Labels generated by the proposed ensemble.

The dataset composition is also reflected in the distribution of objects by time of day (Table 4), which demonstrates that the morning and afternoon periods concentrate the highest volume of objects, whereas the early morning hours register the lowest total. This pattern is consistent with the commercial peak hours encompassed within these periods, during which a natural intensification of vehicle flow occurs.

Table 4: Distribution of objects by period.

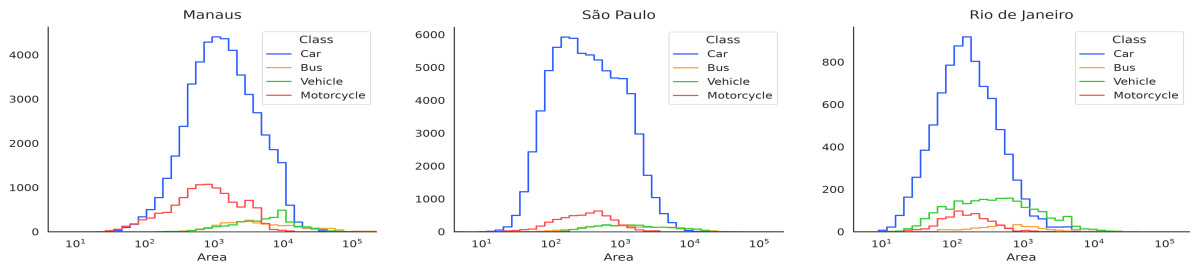
Class	Early Morning	Morning	Afternoon	Night
Car	6 560	38 627	49 689	32 518
Motorcycle	940	5 179	6 098	4 387
Bus	444	1 726	1 989	1 277
Vehicle	753	3 087	2 693	1 142

Regarding the traffic characteristics of each city, it can be observed from Table 5 that the three municipalities present distinct traffic profiles: São Paulo concentrates the highest number of cars, whereas Manaus exhibits a more diverse traffic flow with a strong presence of motorcycles, and Rio de Janeiro registers the lowest overall volume of annotations, due to both the smaller number of monitored cameras and the low resolution of its images, a factor that hinders detection.

Table 5: Number of objects by city and class.

City	Car	Motorcycle	Vehicle	Bus
Manaus	47 539	10 960	3035	2590
Rio de Janeiro	8223	631	2190	301
São Paulo	71 632	5013	2450	2545

From a spatial perspective, observing the bounding box area histogram in Figure 5, a predominance of small to medium-sized objects is noted, indicative of distant or partially occluded vehicles commonly found in heavy urban traffic. This characteristic is also evident when comparing the dimensions of the annotated objects across classes, as Buses are the largest objects in the dataset, with average dimensions of 72.6×59.1 pixels, followed by Vehicles, with 57.5×54.1 pixels. Cars, in turn, exhibit intermediate dimensions (34.1×25.3), whereas Motorcycles present the smallest average width of 22 pixels, which, combined with occlusions and dense traffic conditions, makes them particularly challenging targets for detection. This complexity is further reinforced by the density of the scenes shown in Figure 6, where multiple targets appear simultaneously in most frames.



(a) Manaus **(b) São Paulo** **(c) Rio de Janeiro**

Figure 5: Bounding box area histograms.

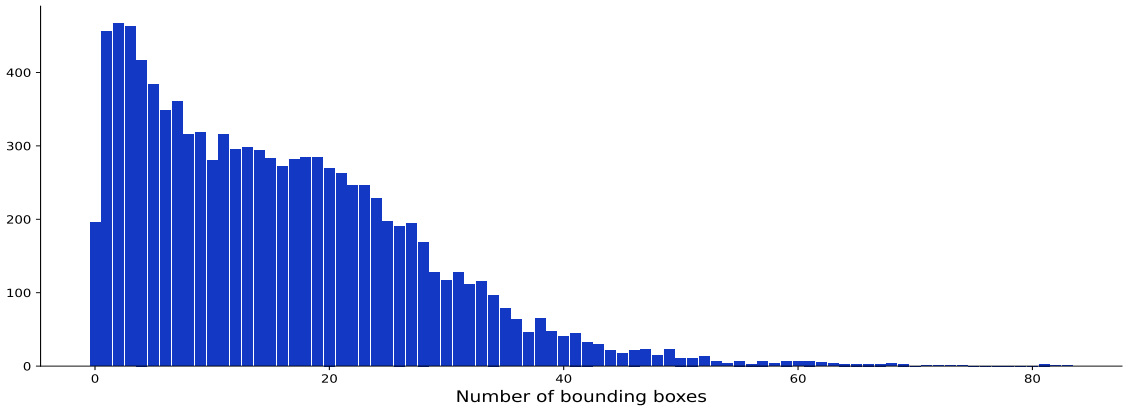


Figure 6: Distribution of the number of bounding boxes per image.

To establish a performance baseline on the proposed dataset and validate the capability of intelligent models to learn its labels, the RT-DETR model was trained in its Large variant. The optimization process was configured for a maximum of 300 epochs, utilizing a batch size of 8 and an initial learning rate of 0.01. To mitigate overfitting, an early stopping criterion with a patience of 40 epochs was adopted. The optimizer was set to the automatic mode within the Ultralytics framework.

The training was completed in approximately 4 hours and 34 minutes, having been conducted on a workstation equipped with an NVIDIA GB10 Grace Blackwell processor featuring 20 ARM cores, a clock speed of up to $2.52\,\mathrm{GHz}$, $128\,\mathrm{GB}$ of unified main memory, $4\,\mathrm{TB}$ of secondary memory, and an integrated NVIDIA Blackwell GB20B graphics card. Table 6 details the specific performance of the model for each object class.

Table 6: Detailed class-wise mAP@0.5:0.95 performance of the RT-DETR architecture.

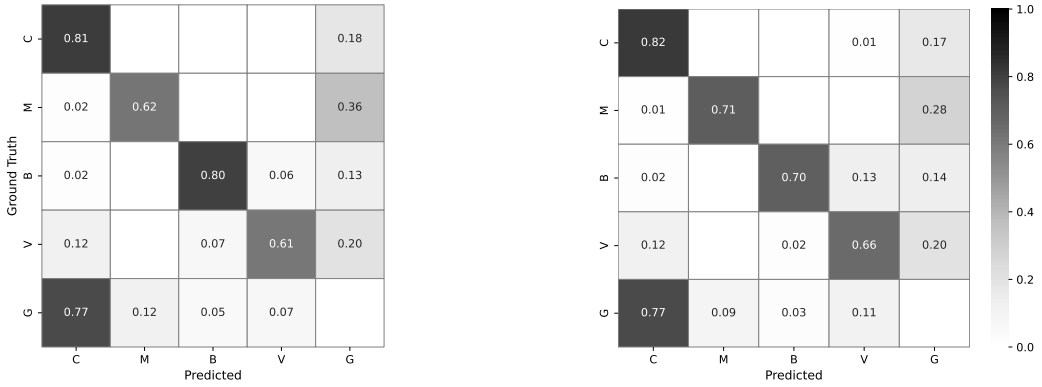
Class	Pseudo-Labels Set	Tuning Set
Car	0.6379	0.5426
Motorcycle	0.4481	0.3792
Bus	0.5663	0.5145
Vehicle	0.4538	0.4241

The evaluation confirms RT-DETR successfully distilled the ensemble's knowledge. On the 2700-image synthetic test set, it achieved an F_1-Score of 0.7359, an mAP@0.5 of 0.7480, and an mAP@0.5:0.95 of 0.5265. Validated against the tuning set for real-world generalization, performance remained robust (F_1-Score: 0.7402, mAP@0.5: 0.7573, mAP@0.5:0.95: 0.4651). Notably, the student model's mAP@0.5 on human ground truth outperformed the teacher ensemble's baseline (0.6177), proving it learned to generalize spatial features rather than merely memorizing pseudo-labels.

The normalized confusion matrices (Figure 7) detail this knowledge transfer. The model reliably detected larger objects on the synthetic set, accurately classifying Cars (0.81) and Buses (0.80). This accuracy translated consistently to the human-supervised labels (Cars: 0.82; Buses: 0.70). As expected in dense traffic, heavily occluded classes saw lower accuracy. Motorcycles were frequently misclassified as background ($36\,\%$ synthetic, improving to $28\,\%$ human), while the rare Vehicle class was occasionally confused with Cars ($12\,\%$) or

background ($20\,\%$) across both sets.

A notable trend emerges regarding background instances: across both test sets, $77\,\%$ of false-positive errors occur when the model predicts a Car where the ground truth indicates background. This strongly suggests the presence of unmapped vehicles in the original images that both the ensemble and human annotators missed. Consequently, the model demonstrated enough generalization to detect these omitted targets, paradoxically suffering precision penalties for making correct real-world identifications.



(a) Distillation Fidelity (2700 images) **(b) Real-World Generalization (300 images)**

Figure 7: RT-DETR row-normalized confusion matrices evaluated on the independent test sets.

Finally, although automated labeling inherently introduces noisy bounding boxes, the evaluation proves this does not undermine the methodology. Architectures like RT-DETR act as statistical filters, resisting label noise when accurate annotations predominate. The strong main-diagonal accuracy and the model's ability to identify unmapped objects confirm a highly favorable signal-to-noise ratio. Ultimately, the proposed pipeline generates annotations robust enough to train high-performance detectors for complex urban traffic environments.

5. Final Remarks

To overcome the manual annotation bottleneck in urban traffic monitoring, this study proposed an automatic labeling pipeline based on an ensemble of predictive models. This methodology produced a large-scale dataset comprising $10\,000$ images and $157\,109$ annotated objects across three major Brazilian cities. The dataset's viability was successfully validated by training an RT-DETR Large model. Demonstrating remarkable consistency, the model achieved an mAP@0.5 of 0.7480 on the 2700-image synthetic test set and 0.7573 on the 300-image human-supervised set (with F_1-Scores of $73.59\,\%$ and $74.02\,\%$, respectively). This confirms the architecture effectively acted as a statistical filter against pseudo-label noise, extracting consistent spatial features even in dense scenarios.

One of the practical contributions of this work is the provision of the developed dataset, accessible at `https://data.mendeley.com/datasets/z7sft43w88`, which stands out for capturing the geographical diversity and the particularities of the Brazilian vehicle fleet. Alongside the images, the predictions generated by the pipeline are made available, enabling other researchers to investigate the behavior of the ensemble. Furthermore, the final weights of the trained RT-DETR model are available in the same repository, allowing them to be utilized directly for inference or via transfer learning in new locations.

The striking performance similarity across both test sets strongly indicates successful real-world generalization without overfitting. Future work will focus on expanding the 300-image manually curated subset to establish a more comprehensive evaluation benchmark. Additionally, we intend to investigate new strategies for automatically identifying false positives in the ensemble outputs. The continuous improvement of these filtering techniques will further elevate the reliability of unsupervised labeling and promote the development of increasingly accurate detection models.

AI Transparency and Ethical Considerations

Gemini 3.1 Pro was utilized to support the writing and editing of this manuscript (organization, stylistic revision, clarity, and translation). The research conception, methodology, experiments, results analysis, and conclusions are the sole responsibility of the authors, who have comprehensively reviewed the content and assume full responsibility for its accuracy and originality.

Acknowledgements

The authors acknowledge the financial support provided by the Amazonas State Research Support Foundation (FAPEAM) and by the Amazonas State University through the PAIC 2025/2026 program. EBG acknowledges the financial support from CAPES through Project COFECUB-PJ3126170P. The authors also express their gratitude for the material support received from the Intelligent Systems Laboratory (LSI) at the Amazonas State University (UEA).

References

Barbado, L., Santos, L., Matrakas, M., and Moreira, J. (2022). Aplicação da rede convolucional yolo para análise de fluxo de veículos. In *Anais do XIX Congresso Latino-Americano de Software Livre e Tecnologias Abertas*, pages 43–49, Porto Alegre, RS, Brasil. SBC.

Brasil (2025). Ministério dos transportes – secretaria nacional de trânsito – frota de veículos. Disponível em `https://www.gov.br/transportes/pt-br/assuntos/transito/conteudo-Senatran/frota-de-veiculos-2025`. Acessado em: 15 maio 2026.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer.

Carvalho, R., Guedes, E., and Figueiredo, C. (2022). Detecção inteligente de falhas em pavimentações asfálticas com redes neurais convolucionais regionais. In *Anais do III Workshop Brasileiro de Cidades Inteligentes*, pages 119–130, Porto Alegre, RS, Brasil. SBC.

Chan, P. H., Li, B., Baris, G., Sadiq, Q., and Donzella, V. (2024). The inconvenient truth of ground truth errors in automotive datasets and dnn-based detection. *Data-Centric Engineering*, 5:e34.

Freire, A., de S. Silva, L. H., de Andrade, J. V., Azevedo, G. O., and Fernandes, B. J. (2024). Beyond clean data: Exploring the effects of label noise on object detection performance. *Knowledge-Based Systems*, 304:112544.

Júnior, P. S. D. S. F. and Guedes, E. B. (2023). Detecção de semáforos em ambientes urbanos para cidades inteligentes. In *Procedings do XVI Simpósio Brasileiro de Automação*

Inteligente / X Simpósio Brasileiro de Sistemas Elétricos, SBAI2023 / SBSE2023. SBA Sociedade Brasileira de Automática.

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., Zhu, J., and Zhang, L. (2024). Grounding dino: Marrying dino withnbsp;grounded pre-training fornbsp;open-set object detection. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XLVII*, page 38–55, Berlin, Heidelberg. Springer-Verlag.

Loureiro, C., Guedes, E., and Figueiredo, C. (2025). Monitoramento inteligente de tempo real para tráfego urbano em cidades brasileiras. In *Anais do XVII Simpósio Brasileiro de Computação Ubíqua e Pervasiva*, pages 41–50. SBC.

Minderer, M., Gritsenko, A., and Houlsby, N. (2023). Scaling open-vocabulary object detection. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Myagmar-Ochir, Y. and Kim, W. (2023). A survey of video surveillance systems in smart city. *Electronics*, 12(17):3567.

Redmon, J., Divvala, S., Girshick, R. B., and Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788.

Santos, A. M., Bastos-Filho, C. J. A., Maciel, A. M. A., and Lima, E. (2020). Counting vehicle with high-precision in Brazilian roads using YOLOv3 and Deep SORT. In *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. IEEE.

Santos, D., Claro, D. B., and Gondim, J. (2023). Monitoring vehicle plate detection in brazilian universities. In *Proceedings of the XIX Brazilian Symposium on Information Systems*, SBSI '23, page 268–275. ACM.

Solovyev, R., Wang, W., and Gabruseva, T. (2021). Weighted boxes fusion: Ensembling boxes from different object detection models. *Image and Vision Computing*, 107:104117.

Souza, A. N., Vassallo, R., and Abling, A. (2025). Reidentificação espaço-temporal de veículos para sistemas inteligentes de tráfego. In *Anais Estendidos do XLIII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, pages 370–375. SBC.

Velesaca, H. O., Mero, A., Singh, A., and Sappa, A. (2026). The hidden cost of poor annotations: How label quality affects camouflaged object detection performance. In Bebis, G., Ye, J., Wang, Y., Konaković Luković, M., Kalantari, N. K., Cho, I., Yang, Y., Dimara, E., and Brehmer, M., editors, *Advances in Visual Computing*, pages 285–296, Cham. Springer Nature Switzerland.

Vijayakumar, A. and Vairavasundaram, S. (2024). YOLO-based object detection models: A review and its applications. *Multimed. Tools Appl.*

Wu, T. (2024). Reducing the cost of test data labelling for deep-learning systems: An empirical study. Master's thesis, School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, Ontario, Canada. Disponível em `https://doi.org/10.20381/ruor-30788`. Acesso em: 15 de maio de 2026.

Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., and Chen, J. (2024). DETRs beat YOLOs on real-time object detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16965–16974. IEEE.