RESEARCH-ARTICLE

Eye on the Street: Computer Vision for Spatial-Temporal Mapping of Street Safety Elements

QIAN ZHANG, University of Toronto, Toronto, ON, Canada

CAMELLIA ZAKARIA, University of Toronto, Toronto, ON, Canada

MARIANNE HATZOPOULOU, University of Toronto, Toronto, ON, Canada

JUNSHI XU, The University of Hong Kong, Hong Kong, Hong Kong

TATE HUBKARAO

STEVE TCHANA, University of Michigan, Ann Arbor, Ann Arbor, MI, United States

View all

Published: 12 October 2025

UbiComp '25:The 2025 ACM International Joint Conference on Pervasive and Ubiquitous Computing / ISWC ACM International Symposium on Wearable Computers
October 12 - 16, 2025
Espoo, Finland

Conference Sponsors:
SIGMOBILE
SIGCHI

https://doi.org/10.1145/3714394.3756280
ISBN: 9798400714771

Eye on the Street: Computer Vision for Spatial-Temporal Mapping of Street Safety Elements

Qian Zhang
Dalla Lana School of Public Health, University of Toronto
Toronto, Canada
erica.zhang2025@outlook.com

Camellia Zakaria
Dalla Lana School of Public Health, University of Toronto
Toronto, Canada
camellia.zakaria@utoronto.ca

Marianne Hatzopoulou
Civil and Mineral Engineering, University of Toronto
Toronto, Canada
marianne.hatzopoulou@utoronto.ca

Junshi Xu
Department of Geography, University of Hong Kong
Hong Kong, China
junshixu@hku.hk

Tate HubkaRao
Humber River Health Research Institute
Toronto, Canada
tate.hubkarao@utoronto.ca

Steve Tchana
University of Michigan
Ann Arbor, United States
sttcha@umich.edu

Linda Rothman
Toronto Metropolitan University
Toronto, Canada
linda.rothman@torontomu.ca

Aryan Sadeghi
Dalla Lana School of Public Health, University of Toronto
Toronto, Canada
aryan.sadeghi@mail.utoronto.ca

Brice Batomen
Dalla Lana School of Public Health, University of Toronto
Toronto, Canada
brice.kuimi@utoronto.ca

Abstract

Understanding when and where traffic calming measures are implemented is essential to assess their impact and plan future safety interventions for vulnerable road users. Yet, such records are often incomplete or unavailable. To support accurate, automated, and large-scale efforts to address these critical data gaps, we propose a computer vision–based framework to detect such measures from historical street view imagery that captures real-world urban complexity. We share key preliminary results demonstrating the effectiveness of our framework in overcoming visual challenges within these images, providing a solid foundation as we continue to improve and progress toward full implementation.

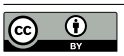
CCS Concepts

• **Computing methodologies** → **Supervised learning by classification**; *Object recognition*; Neural networks.

Keywords

road safety; traffic calming measures; public health; deep learning

ACM Reference Format:
Qian Zhang, Camellia Zakaria, Marianne Hatzopoulou, Junshi Xu, Tate HubkaRao, Steve Tchana, Linda Rothman, Aryan Sadeghi, and Brice Batomen. 2025. Eye on the Street: Computer Vision for Spatial-Temporal Mapping of Street Safety Elements. In *Companion of the 2025 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp Companion '25), October 12–16, 2025, Espoo, Finland*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3714394.3756280



UbiComp Companion '25, Espoo, Finland

ACM ISBN 979-8-4007-1477-1/2025/10
https://doi.org/10.1145/3714394.3756280

1 Introduction

Vulnerable road users (VRUs), such as pedestrians and cyclists, are at greater risk of injury or death in traffic collisions [27, 28]. In 2021 alone, VRUs accounted for over half of the estimated 1.19 million road traffic deaths worldwide, with pedestrians making up 22% of the total; in some countries, this proportion is as high as two-thirds [27]. Compared to motorized vehicle users, per distance traveled, cyclists face a higher rate of injury or death compared to motorized vehicle users, especially in areas where they share roads with motor vehicles [15, 28].

Given these persistent risks—and the notable health and environmental benefits of walking and cycling—ensuring a safe road network remains a critical priority for urban centers [8]. Cities worldwide have gradually adopted Vision Zero [5, 26], a road safety initiative that seeks to eliminate traffic fatalities and severe injuries among all road users [30]. Unlike traditional approaches, it recognizes that human errors are unavoidable and prioritizes road system designs that prevent fatalities and minimize crash severity [11]. A central principle involves lowering vehicle speeds, as higher speeds significantly increase the likelihood of death or serious harm [21]. To achieve this, cities are implementing traffic calming measures (TCMs), physical road modifications that lower vehicle speeds and/or volumes, thus improving the safety of all road users [4].

However, recent studies [3, 22] have found that many Canadian cities lack accurate, complete, and timely records of where and when TCMs have been implemented. Without this, we cannot assess to what extent existing TCMs are improving VRU safety, nor can we identify locations that require additional safety interventions. Filling this gap is essential to enable more integrated, data-driven, and equitable urban health strategies.

Given that manually filling missing TCM records for each city is resource-intensive, we propose an automated and scalable computer vision-based approach. By leveraging a vision foundation model

and Google Street View (GSV) images (2011-2023) from Toronto and Montreal, we developed a multi-label classification system to detect eight TCMs that are more commonly implemented in North America [17]. These include curb extensions, cycle tracks, median islands, mid-block narrowings, raised crosswalks, speed bumps, speed cushions, and speed humps. Cross-city training supports model generalization across Canadian urban areas. Applying it to historical images will enable the creation of a geodatabase tracking when and where infrastructure was implemented, allowing for impact evaluation. While the project is ongoing, we present preliminary results and discuss challenges in scaling. Our code and dataset are available at https://github.com/BbriceK/EyeOnTheStreet

2 Motivation and Background

2.1 Data Deficits in Traffic Calming Studies

Over 25 Canadian cities have adopted Vision Zero [5], yet most lack detailed and accurate records on TCM locations, types, and installation dates. For example, in Montreal, half the boroughs lacked TCM implementation years [3]. Where data do exist, they are often limited to a snapshot of current conditions without historical context [16]. Moreover, in cases where retrospective data are available, accuracy remains a concern. In Calgary, only 45% of cycling network implementation dates matched street view imagery [22].

The absence of complete, retrospective data hinders both the refinement of local VRU safety strategies and large-scale research needed for generalizable, evidence-based insights.

2.2 Computer Vision for Urban Road Features

Advancements in computer vision, including YOLO and transformer-based detectors, have enabled scalable detection of urban elements like vehicles [1] from imagery. However, TCM detection remains underexplored, with most work limited to speed humps and bumps. While methods have evolved from—Gaussian filters and handcrafted segmentation [7], to convolutional neural networks (CNNs)[2, 23, 25], and advanced object detection frameworks like YOLOv4 and YOLOv8 [6, 10]—these models are ill-suited for our task. They were typically trained on close-range images where target objects are prominently visible, resulting in strong performance on such images. Our goal is to identify when and where TCMs were installed citywide. Since TCM locations are unknown, acquiring close-up images for inference would require an impractical city-wide sweep. Therefore, inference must be done on images not specifically focused on the target TCMs, which they may appear small or partially occluded, posing significant challenges for these models.

Among various imagery sources, GSV is particularly well-suited for our task due to its broad temporal and spatial coverage. Prior work started to apply YOLO and its variants to detect traffic-related objects in GSV imagery [14, 19], while transformer-based detectors remain underexplored in this context. In our experiments, we applied YOLOv11 [24] and a transformer-based detector DINO [29], but as shown in Section 4.3.2 both struggle in our context, highlighting the need for a more tailored approach.

Key Takeaway Due to gaps in TCM records across Canadian cities and the limitations of current computer vision methods for detecting TCMs, there is a clear need for automated, scalable, and generalizable systems to address these data deficits and support large-scale longitudinal safety analyses for VRUs.

3 System

Figure 1 presents an overview of our proposed system, *Eye on the Street*. Our approach frames the task as a multi-label image classification problem. The process begins by extracting three types of image features using DINOv2 [18], a self-supervised vision transformer. Then, these features are concatenated and passed to a neural network classifier designed to predict the presence of the eight TCMs of interest. We describe our system components and design considerations in detail as follows. Note, our study was approved by the Research Ethics Board (00046079), University of Toronto.

3.1 Key System Challenges

Our approach must address several non-trivial challenges at different stages of the system pipeline.

3.1.1 Variation and Types of TCMs We anticipate that our modeling task will be challenged by the high visual variability across the eight TCM categories. Even within the same category (e.g., cycle tracks, as shown in Figure 2), appearances can differ significantly depending on the type of barrier used, such as bollards and concrete, making consistent detection difficult. Furthermore, the appearance of a given category may differ across cities due to local policies. Beyond appearance, the geographic distribution of TCM types also varies according to local priorities. For example, measures like curb extensions and cycle tracks may be prevalent in some areas but entirely absent in others. It is also possible for a location to contain only one type of measure. These factors can lead to class imbalance, which poses challenges for training models.

3.1.2 Temporal Environment Factors at Longitudinal Scale Leveraging historical data sources is central to our approach in determining when and where TCMs were implemented. However, the same location can differ markedly over time due to seasonal changes, weather conditions, lighting variations, or even updates to road infrastructure. As illustrated in Figure 3, these changes can obscure or alter key visual cues. Therefore, a robust model must be capable of adapting to these variations arising from diverse environmental contexts.

3.1.3 Inference Under Occlusion and Visual Clutter As discussed in Section 2.2, the lack of complete TCM location data makes collecting close-range images impractical—whether through visiting every city corner or manipulating camera position and zooming in API-accessible street-level imagery. As a result of relying on broad street-level imagery, TCMs may appear small, partially occluded, or embedded in cluttered scenes (see Figure 4). Therefore, the model must be able to detect TCMs from limited or ambiguous visual cues.

3.2 Input Data: Google Street View Images

Manually collecting street-level images across entire cities and across time is impractical, especially for historical evidence. Platforms such as GSV offer a compelling alternative, providing extensive geographic reach and archived images of streets in a consistent visual format. This availability enables retrospective detection of

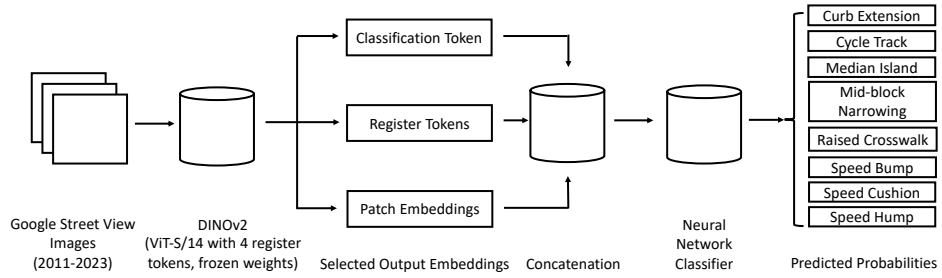


Figure 1: Overview of *Eye on the Street*, where inputs are a series of 640×640 GSV images, employing a frozen DINOv2 ViT-S/14 with 4 register tokens to extract visual features and a custom neural network classifier to generate probabilities of each TCM.

Figure 2: Two cycle tracks delineated by bollards (left) and concrete dividers (right).

Figure 3: Capturing two curb extensions of interest at the same place, over different time points and weather.

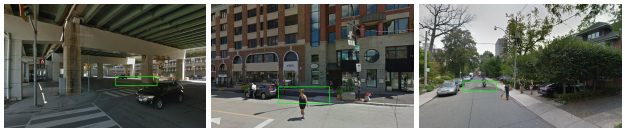


Figure 4: Small and partially occluded TCMs: a median island (left), curb extension (middle), and speed hump (right)

road infrastructure changes and supports future applications of our approach in other regions.

Table 1 summarizes the dataset we constructed, amounting to 126,718 (640×640) images from 17,395 sampling points across two selected cities between 2011 and 2023. Annotating the entire set requires substantial effort, and as a workaround, we curated a subset (n = 5,401), prioritizing locations where TCMs were more likely to appear, based on each city's Vision Zero map.

3.2.1 Data Collection Procedure We retrieved all images using the GSV Static API, with default camera parameters to ensure visual consistency. For each city, we followed a four-step process to generate sampling points for image retrieval. First, we selected six wards or boroughs based on traffic collision rates, choosing the three with the highest and the three with the lowest rates. Second, we used all road intersections within the selected areas as anchor points, since the majority of these measures and collisions occur at intersections [3]. Third, we excluded path types such as trails, expressways, and ferry routes due to the low likelihood of TCMs being present. Finally, for adjacent intersections more than 100 meters apart, additional sampling points were added at 100-meter intervals. This approach ensured broad and even spatial coverage within each city, helping capture sufficient images likely to contain TCMs and thus minimizing the uneven data distribution anticipated in the challenges described above (Section 3.1).

Table 1: Distribution of Images Collected Across Two Cities

Ward	Total (n)	Annotated (%)
Toronto: Toronto Center	11,405	7.58
Toronto: Spadina-Fork York	16,266	5.15
Toronto: Don Valley West	16,492	0.77
Toronto: Toronto St. Paul's	14,189	4.74
Toronto: University-Rosedale	22,652	1.58
Montreal: Ville-Marie	21,770	4.34
Montreal: Le Plateau-Mont-Royal	13,000	6.97
Montreal: Montreal-Nord	10,944	6.1

3.2.2 Annotation Procedure To construct the annotation subset, we selected the nearest sampling point for each marked TCM location from the Vision Zero safety maps and retrieved all associated images. Wards without TCM records on Vision Zero safety maps were excluded, resulting in five wards for Toronto and three for Montreal. Table 2 presents the category distribution in the final annotated dataset. We used Label Studio [9] for annotation. Two trained annotators independently labeled all 865 images from the Toronto Center ward, guided by a detailed protocol. They achieved a 94.5% agreement rate, with discrepancies resolved by three other reviewers. The remaining images from the annotation subset were split for independent labeling.

3.3 Multi-Label Classification Module

In addressing the challenges (Section 3.1) with limited training data, we applied transfer learning techniques. That is, rather than training from scratch, the model was initialized with weights learned on a large, general-purpose dataset. This enables the model to leverage previously acquired visual representations such as edges, textures,

Table 2: Percentage Distribution of Annotated TCMs

	Total (n)	Percentage (%)
Background	4614	85.8
Cycle Track	264	4.9
Curb Extension	251	4.7
Median Island	178	3.3
Speed Hump	56	1
Mid-block Narrowing	7	0.1
Speed Bump	5	0.1
Raised Crosswalk	3	0.1
Speed Cushion	0	0

and object shapes. We developed a multi-label classification approach to identify which TCMs are present in an image. As per Figure 1, *Eye on the Street* operates in three stages: it first extracts visual features using a pretrained backbone model, then aggregates these extracted features through a custom module, followed by a classifier to predict the presence of each TCM.

3.3.1 Feature Extraction In our work, we leverage the DINOv2 ViT-S/14 with 4 register tokens [18] as the backbone for feature extraction, utilizing its pretrained weight to obtain embeddings suitable for multi-label image classification. Our experiments showed that the best performance was achieved by freezing the entire pretrained backbone and fine-tuning only the classification head on our dataset. We also explored augmenting underrepresented categories during training to mitigate class imbalance.

The field of visual recognition has rapidly progressed from CNN-based models like AlexNet [12], which required large labeled datasets, to more scalable approaches such as self-supervised learning and vision transformers. These innovations have led to foundation models like DINOv2, which are trained on diverse, large-scale datasets and produce general-purpose visual representations that can transfer well across tasks, making them well-suited to our system challenges. Unlike common classification tasks involving clearly distinguishable objects (e.g., animals), TCM detection involves identifying small, context-dependent features that may be occluded, degraded, or embedded in complex urban scenes. Strong pretrained representations are therefore essential for capturing these subtle patterns with limited training data.

Although DINOv2 was originally developed for downstream tasks such as single-label classification and depth estimation, we adapted it for multi-label classification by leveraging all of the model's available output embeddings: (1) the **classification token**, which summarizes the global image context; (2) the **register tokens**, capturing complementary contextual and global information across the image; and (3) **patch embeddings**, which represent raw, localized features from distinct image regions. While (1) is the standard embedding used for classification, combining it with (2) and (3) adds complementary contextual, global, and local insights. Our experiments show that using all three together yields better performance.

3.3.2 Feature Concatenation These extracted embeddings capture both coarse global context and fine-grained local details of the image. To fully utilize this multi-scale information, we integrate these embeddings into a unified representation, enabling the model to learn complex patterns inherent to the task.

For the register tokens (4 tokens per image, each with 384 features), we applied a 1D convolutional layer (kernel size 4, stride 4) to reduce dimensionality while preserving relationships across the sequence of tokens. To capture higher-order spatial interactions and aggregate local information into a compact global representation, we processed the patch embeddings (16×16 patches, each with 384 features) using a lightweight CNN comprising two convolutional layers. Each layer was followed by batch normalization, which stabilizes and accelerates training, and then by a ReLU activation to introduce non-linearity and enhance the model's ability to learn complex patterns. Subsequently, an adaptive average pooling was applied to yield the global patch representation. These processed features were then concatenated with the classification token (384 dimensions), resulting in a unified feature vector of dimension 896.

3.3.3 Neural Network Classifier To model the intricate nonlinear patterns within the multiscale, spatially complex features, we employed a three-layer fully connected neural network. By progressively reducing dimensionality from 896 to 8, the network condenses rich feature representations into compact category-specific signals, enabling more accurate and robust predictions of TCMs. To enhance generalization and prevent overfitting, batch normalization, ReLU activation, and dropout (rate 0.7) were applied after the first two layers. The model was trained for 150 epochs using the AdamW optimizer (learning rate: 0.0001, batch size: 32).

3.3.4 Loss Function We used an asymmetric loss [20] to address the data imbalance in our dataset (see Section 3.1.1), which assigns higher penalties to errors in rare positive classes and focuses more on underrepresented TCMs. The asymmetric loss is defined as:

$$\mathcal{L}_{ASL} = -\sum_{i=1}^{C} \left[y_i (1-p_i)^{\gamma_+} \log(p_i) + (1-y_i) p_i^{\gamma_-} \log(1-p_i) \right]$$

where C is the number of classes, $y_i \in 0, 1$ is the ground-truth label for class i, p_i is the predicted probability, and γ_+ and γ_- are focusing parameters for positive and negative samples, respectively. We experimentally set $\gamma_+ = 0.3$ and $\gamma_- = 0.5$ to emphasize the learning on rare positive samples, with clipping thresholds at 0.01.

4 Evaluation

We report the preliminary evaluation results of our approach using the dataset described in Section 3.2.

4.1 Data Split

To avoid overwhelming the model with a large number of background images (i.e., images without any TCMs), we capped the proportion of background images to 10%. Future experiments will gradually increase this proportion to better reflect real-world class distributions. Our data splitting strategy has two main goals: (1) to ensure spatial independence by assigning all images from a single sampling point to the same split, thus avoiding data leakage between training and testing; and (2) to maintain class balance across training, validation, and test sets using stratified sampling based on the presence of each TCM, including background images. This resulted in 539 training, 151 validation, and 178 test images.

4.2 Performance Metric

Standard metrics like average precision and mean average precision are widely used in computer vision tasks, and they summarize model performance across confidence thresholds (CTs). However, our overarching goal is to build a geodatabase with yes/no labels indicating the presence of TCMs. Therefore, we fixed CTs to binarize predictions, choosing two levels: 0.5, which favors recall by capturing more true positives, and 0.9, which favors precision by reducing false positives. This contrast helps explore trade-offs relevant to different real-world deployments. The implications of our findings from these CT choices are discussed in Section 5.

In essence, we report *precision*, *recall*, and *F1 score*. These metrics effectively capture the model's ability to correctly identify objects of interest while accounting for both missing detections and false alarms. We excluded accuracy, as the large number of true negatives per category would inflate it, making it a less informative metric.

4.3 Results

4.3.1 Multi-label Classification of TCMs Tables 3 and 4 present the multi-label classification results at CTs of 0.5 and 0.9. Only four TCM categories are reported, as others lacked test instances after stratified sampling. Compared with these, additional augmentation—applied to the median island (one per image) and speed hump (three per image) during training—reduced the overall model performance in terms of Macro F1 score at CT 0.5 (52.54% vs. 55.86%) but slightly improved it at CT 0.9 (44.94% vs. 44.32%; Appendix A).

Overall, the model captures a reasonable number of true positives across diverse instances within each TCM category at both CTs. At 0.5, F1 scores range from approximately 40% to 80%, with speed humps achieving the highest precision and recall, while cycle tracks exhibit the lowest performance. At 0.9, precision improves for all categories—reaching 100% for speed humps—though recall declines substantially, with cycle tracks again performing the poorest. Closer examination of the errors shows that mispredictions mainly stem from variability in TCM appearances across locations and partial visibility of interventions. For example, a series of missed curb extensions was linked to a red "No Entry" sign in front—an appearance that deviates from typical curb extensions. In another case, only a small portion of a median island was visible across images from the same location; while humans can infer its full presence from context, these partial and atypical views challenged the model. Similar issues were noted across other categories.

Table 3: Multi-label Classification Results at CT of 0.5

	Precision (%)	Recall (%)	F1 Score (%)
Curb Extension	53.45	50.82	52.1
Cycle Track	45.95	36.17	40.47
Median Island	65.22	41.67	50.85
Speed Hump	75	85.71	80

4.3.2 Comparison with Leading Models To understand our approach and assess its strengths, we benchmarked it against two types of models: (1) multi-label classification and (2) object detection. We included (2) because they predict both object presence and spatial information—the former aligning with our task goal—and because object detection is a well-established field with mature models. Query2Label [13] was selected as the representative for (1) due to its top ranking on several benchmark datasets. Within (2), we chose YOLOv11, a widely adopted and high-performing model, and DINO, a leading transformer-based detector.

Table 4: Multi-label Classification Results at CT of 0.9

	Precision (%)	Recall (%)	F1 Score (%)
Curb Extension	66.67	29.51	40.91
Cycle Track	47.37	19.15	27.27
Median Island	76.47	36.11	49.06
Speed Hump	100	42.86	60

We evaluated model performances using fixed thresholds. For object detection, this means applying two thresholds: intersection over union (IoU), which measures overlap between predicted and true boxes, and confidence score, which indicates prediction certainty. In our evaluation, the output from Query2Label showed atypical patterns. For example, it consistently predicted a probability of 1.0 for the presence of cycle tracks across all test images, while assigning near-zero probabilities to all other categories. Among all YOLOv11 configurations, at a moderate IoU of 0.5 combined with a CT of 0.5, the best results were as follows: cycle tracks (precision: 0.05, recall: 0.02, F1 score: 0.03), median islands (precision: 0.22, recall: 0.05, F1 score: 0.08), and other categories had zero on all metrics. At a higher CT (0.9), precision, recall, and F1 scores were zero across all categories. Similarly, DINO produced zero on all metrics across all categories and thresholds. This underscores the challenges of capturing subtle, small, and context-dependent visual features central to our task.

5 Discussion and Future Work

We proposed a promising novel approach to determine the spatial and temporal implementation of TCMs. Here we discuss the implications of our findings.

5.1 Discussion

Our work highlights key challenges in adapting computer vision techniques to detect TCMs in complex urban settings. Preliminary results demonstrate the ability of our model to identify key features for four target categories of TCM despite these challenges. The difficulty of the task and the relative strength of our approach are further underscored by comparison with leading methods. The relatively lower performance observed for cycle tracks is likely due to greater visual variability within this category, compounded by common issues such as small size and occlusion that affect all categories. In contrast, categories with distinctive features, such as speed humps, tend to achieve more reliable detection. Moreover, augmentations like vertical flipping and large-angle rotation increase intra-class variability and introduce unrealistic distortions. While not perfect, the performance is sufficiently strong and interpretable to serve as a practical baseline model for future improvements.

The fixed CTs approach offers a flexible framework for tailoring geodatabase construction to specific analytical or operational needs. At lower thresholds, the model generates a more comprehensive

but noisier TCM geodatabase, with more false positives that can be manually filtered. At higher thresholds, the geodatabase becomes more precise but may miss valid cases.

Our work offers a valuable resource for road safety research by introducing an automated method to identify TCMs across time and locations. Using only historical street view imagery and intersection coordinates, this framework can accelerate efforts to fill critical data gaps and complements traditional, labor-intensive data collection. It represents a key step toward scalable, data-driven infrastructure monitoring, enabling continuous and cost-effective urban analysis.

5.2 Future work

Several additional strategies, such as converting images to grayscale and increasing input resolution, were tested but yielded no substantial improvements. This suggests that the model may have reached a performance plateau within the existing framework. As indicated by the error analysis, the task requires sophisticated reasoning to infer TCM presence from complex visual cues. To address this, the next step involves exploring multimodal large language models that integrate image understanding with language-based reasoning. In parallel, we propose continued development of performance metrics tailored to fill missing data via multi-label classification, aiming to further improve database quality and reduce the need for manual filtering.

Acknowledgments

This work was supported by the Data Sciences Institute at the University of Toronto, DSI-CGY3R1P40, and the Natural Sciences and Engineering Research Council of Canada, RGPIN-2025-06914.

References

[1] Mortda A. A. Adam and Jules R. Tapamo. 2025. Survey on Image-Based Vehicle Detection Methods. *World Electric Vehicle Journal* 16, 6, Article 303 (2025). https://doi.org/10.3390/wevj16060303

[2] J. Arunpriyan, V. V. S. Variyar, K. P. Soman, and S. Adarsh. 2020. Real-Time Speed Bump Detection Using Image Segmentation for Autonomous Vehicles. In *Intelligent Computing, Information and Control Systems*. Vol. 1039. 413–423. https://doi.org/10.1007/978-3-030-30465-2_35

[3] Brice Batomen, Marie-Soleil Cloutier, Mabel Carabali, Brent Hagel, Andrew Howard, Linda Rothman, Samuel Perreault, Patrick Brown, Erica Di Ruggiero, and Susan Bondy. 2023. Traffic-Calming Measures and Road Traffic Collisions and Injuries: A Spatiotemporal Analysis. *American Journal of Epidemiology* 193, 5 (2023), 707–717. https://doi.org/10.1093/aje/kwad136

[4] Olivier Bellefleur and François Gagnon. 2012. *Urban Traffic Calming and Health: A Literature Review*. Institut national de santé publique du Québec. Retrieved June 21, 2025 from https://www.ncchpp.ca/docs/ReviewLiteratureTrafficCalming_En.pdf Available in electronic format on NCCHPP and INSPQ websites.

[5] Parachute Canada. 2024. *Vision Zero Map*. https://parachute.ca/en/program/vision-zero/vision-zero-map/ Accessed: 2025-06-21.

[6] Payal Chandak, Shivamkumar Chaurasia, and Megharani Patil. 2025. Identification of Potholes and Speed Bumps Using SSD and YOLO. In *Intelligent Computing and Networking*. 1–20. https://doi.org/10.1007/978-981-97-8631-2_1

[7] W. Devapriya, Nelson K. B. Christopher, and T. Srihari. 2016. Real time speed bump detection using Gaussian filtering and connected component approach. In *2016 World Conference on Futuristic Trends in Research and Innovation for Social Welfare (Startup Conclave)*. 1–5. https://doi.org/10.1109/STARTUP.2016.7583981

[8] Gerson Ferrari, Clemens Drenowatz, Irina Kovalskys, Georgina Gómez, Attilio Rigotti, Lilia Y. Cortés, Martha Y. García, Rossina G. Pareja, Marianella Herrera-Cuenca, Ana P. Del'Arco, Miguel Peralta, Adilson Marques, Ana C. B. Leme, Kabir P. Sadarangani, Juan Guzmán-Habinger, Javiera L. Chaves, and Mauro Fisberg. 2022. Walking and cycling, as active transportation, and obesity factors in adolescents from eight countries. *BMC Pediatrics* 22, 1 (2022), 510. https://doi.org/10.1186/s12887-022-03577-8

[9] Heartex. 2024. *Label Studio: Data Labeling Tool*. https://labelstud.io/ Accessed: 2025-06-21.

[10] Abrar A. Hussein, Mariam M. Omar, Jawaher S. Altamimi, and Waleed M. Ead. 2024. An Intelligent Approach for Speed Bump Detection. In *2024 IEEE 3rd International Conference on Computing and Machine Intelligence (ICMI)*. 1–6. https://doi.org/10.1109/ICMI60790.2024.10586000

[11] Ellen Kim, Peter Muennig, and Zohn Rosen. 2017. Vision Zero: A Toolkit for Road Safety in the Modern Era. *Injury Epidemiology* 4 (2017), 1. https://doi.org/10.1186/s40621-016-0098-z

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2017. ImageNet classification with deep convolutional neural networks. *Commun. ACM* 60, 6 (2017), 84–90. https://doi.org/10.1145/3065386

[13] Shilong Liu, Lei Zhang, Xiao Yang, Hang Su, and Jun Zhu. 2021. Query2Label: A Simple Transformer Way to Multi-Label Classification. https://doi.org/10.48550/arXiv.2107.10834 arXiv:2107.10834

[14] Arenla Longchar, Misal D. Anna, and Rajesh Dhumal. 2023. A Comparative Analysis of Deep-Learning-Based YOLO Models (V8n and V8s) for Object Detection Using GSV Images. In *2023 International Conference on Integration of Computational Intelligent System (ICICIS) (ICICIS '23)*. IEEE, Pune, India, 1–8. https://doi.org/10.1109/ICICIS56802.2023.10430204

[15] Wesley E. Marshall and Nicholas N. Ferenchak. 2019. Why cities with high bicycling rates are safer for all road users. *Journal of Transport & Health* 13 (2019), 100539. https://doi.org/10.1016/j.jth.2019.03.004

[16] City of Montreal. 2023. *Vision Zero: Getting Around Safely by Foot, Bike, and Car*. https://montreal.ca/en/articles/vision-zero-getting-around-safely-foot-bike-and-car-14584 Accessed: 2025-06-22.

[17] Institute of Transportation Engineers. 2023. *Traffic Calming Measures*. https://www.ite.org/technical-resources/traffic-calming/traffic-calming-measures Accessed: 2025-06-21.

[18] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2023. DINOv2: Learning Robust Visual Features without Supervision. https://doi.org/10.48550/arXiv.2304.07193 arXiv:2304.07193

[19] Vung Pham, Du Nguyen, and Christopher Donan. 2022. Road Damage Detection and Classification with YOLOv7. In *2022 IEEE International Conference on Big Data (Big Data) (Big Data '22)*. IEEE, Osaka, Japan, 6416–6423. https://doi.org/10.1109/BigData55660.2022.10020856

[20] Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. 2021. Asymmetric Loss For Multi-Label Classification. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 82–91. https://doi.org/10.1109/ICCV48922.2021.00015

[21] Erik Rosén, Helena Stigson, and Ulrich Sander. 2011. Literature review of pedestrian fatality risk as a function of car impact speed. *Accident Analysis & Prevention* 43, 1 (2011), 25–33. https://doi.org/10.1016/j.aap.2010.04.003

[22] Linda Rothman, Konrad Samsel, Andrew Howard, Moreno Zanotto, Meghan Winters, Brent Hagel, Richard Wen, and Brice Batomen. 2024. 330 Pedaling Forward: The Evolution of Bicycling Infrastructure in 3 Canadian Cities from 2010 to 2022. *Injury Prevention* 30 (2024), A72. https://doi.org/10.1136/injuryprev-2024-SAFETY.169

[23] Anju Thomas, Harikrishnan P. M., Nisha J. S., Varun P. Gopi, and Palanisamy Ponnusamy. 2021. Pothole and Speed Bump Classification Using a Five-Layer Simple Convolutional Neural Network. In *Proceedings of International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications*. 491–499. https://doi.org/10.1007/978-981-15-7234-0_45

[24] Ultralytics. 2024. Ultralytics GitHub Repository. https://github.com/ultralytics/ultralytics Accessed: 2025-06-21.

[25] V. S. K. P. Varma, S. Adarsh, K. I. Ramachandran, and Binoy B. Nair. 2018. Real Time Detection of Speed Hump/Bump and Distance Estimation with Deep Learning using GPU and ZED Stereo Camera. *Procedia Computer Science* 143 (2018), 988–997. https://doi.org/10.1016/j.procs.2018.10.335

[26] Vision Zero Network. n.d.. *Vision Zero Network*. Retrieved June 21, 2025 from https://visionzeronetwork.org/

[27] World Health Organization. 2023. *Global Status Report on Road Safety 2023: Summary*. Retrieved June 21, 2025 from https://www.who.int/publications/i/item/9789240086456

[28] George Yannis, Dimitrios Nikolaou, Alexandra Laiou, Yvonne A. Stürmer, Ilona Buttler, and Dagmara Jankowska-Karpa. 2020. Vulnerable road users: Cross-cultural perspectives on performance and attitudes. *IATSS Research* 44, 3 (2020), 220–229. https://doi.org/10.1016/j.iatssr.2020.08.006

[29] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel Ni, and Heung-Yeung Shum. 2023. DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=3mRwyG5one

[30] Matts Åke Belin, Per Tillgren, and Evert Vedung. 2011. Vision Zero – a road safety policy innovation. *International Journal of Injury Control and Safety Promotion* 19, 2 (2011), 171–179. https://doi.org/10.1080/17457300.2011.635213

A Stratified Sampling with Data Augmentation

Table 5: Multi-label Classification Results at CT of 0.5

	Precision (%)	Recall (%)	F1 Score (%)
Curb Extension	57.41	50.82	53.91
Cycle Track	52.5	44.68	48.28
Median Island	65.22	41.67	50.85
Speed Hump	57.14	57.14	57.14

Table 6: Multi-label Classification Results at CT of 0.9

	Precision (%)	Recall (%)	F1 Score (%)
Curb Extension	54.29	31.15	39.58
Cycle Track	50	27.66	35.62
Median Island	81.25	36.11	50
Speed Hump	75	42.86	54.55