STRIDE: Street View-based Environmental Feature Detection and Pedestrian Collision Prediction

Cristina González[1,2]* Nicolás Ayobi[1,2]* Felipe Escallón[1,2]
Laura Baldovino-Chiquillo[3] Maria Wilches-Mogollón[2] Donny Pasos[4] Nicole Ramírez[2]
Jose Pinzón[3,6] Olga Sarmiento[3] D. Alex Quistberg[3,5] Pablo Arbeláez[1,2]

[1]Center for Research and Formation in Artificial Intelligence, Universidad de los Andes, Colombia
[2]School of Engineering, Universidad de los Andes, Colombia [3]School of Medicine, Universidad de los Andes, Colombia
[4]School of Economics, Universidad de los Andes, Colombia [5]Dornsife School of Public Health, Drexel University, USA
[6]School of Architecture and Design, Pontificia Universidad Javeriana, Colombia

Abstract

This paper introduces a novel benchmark to study the impact and relationship of built environment elements on pedestrian collision prediction, intending to enhance environmental awareness in autonomous driving systems to prevent pedestrian injuries actively. We introduce a built environment detection task in large-scale panoramic images and a detection-based pedestrian collision frequency prediction task. We propose a baseline method that incorporates a collision prediction module into a state-of-the-art detection model to tackle both tasks simultaneously. Our experiments demonstrate a significant correlation between object detection of built environment elements and pedestrian collision frequency prediction. Our results are a stepping stone towards understanding the interdependencies between built environment conditions and pedestrian safety.

1. Introduction

Autonomous driving systems rely on their ability to gather and interpret information from their surroundings [62], enabling them to anticipate future events and make situation-aware decisions without compromising road safety for all parties involved. A challenging task for autonomous vehicles (AVs) is the detection of pedestrians and other vulnerable road users to avoid pedestrian-motor vehicle collisions. While much of the prior research has focused on pedestrian-detection tasks, few studies have examined the role of road infrastructure and built environment features in improving pedestrian detection and thus reducing the chance of pedestrian collisions. In particular, it is well known that pedestrian safety is impacted by road design and the built environment [74, 65, 21, 34, 28, 69, 42], including objects that comprise defined pedestrian crossing areas (e.g., crosswalks, stop lines, speed bumps), traffic control (e.g., traffic signs, pedestrian signs, stop signs) and traffic speed (e.g., road width, traffic lanes, trees, street lights).

Figure 1: **STRIDE.** Given specific city coordinates (left) and the corresponding panoramic street view image (middle), we propose to predict the number of pedestrian collisions in those coordinates (right) by detecting built environment features (middle).

However, the influence of these features on pedestrian injuries can vary depending on the dynamics of road users within specific geographical locations. Consequently, the presence or absence of such features may have distinct effects in low and middle-income countries compared to high-income ones [65]. Therefore, it is crucial to prioritize advancing autonomous driving systems that can generalize to the unique characteristics of their environments worldwide.

Despite road features' crucial role in pedestrian injuries, limited research has explored the correlation between these objects and pedestrian collision frequency using visual information extracted from the street-level scene. Existing

*Equal contribution

frameworks often rely on precomputed data about the street environment or primarily focus on identifying and anticipating collisions rather than proactively preventing them. This research gap underscores the need for a comprehensive approach that leverages visual cues from the street to analyze the relationship between road features and pedestrian collision occurrences.

This paper introduces a novel dataset comprising more than 18k Google Street View [3] panoramic images, annotated with bounding boxes for 27 categories of common road environment objects that may affect pedestrian safety. Furthermore, we calculate the true incidence of pedestrian injuries for specific geographical points corresponding to the images within the dataset by leveraging public historical records from 2015 to 2021 from Bogota City in Colombia.

Most publicly available data for autonomous driving primarily originates from European, Asian, or North American countries [20, 60, 94, 35, 55], thus leaving Latin American and African countries significantly underrepresented. Therefore, focusing on the Latin American region bridges a crucial gap and provides specific insights into these countries and their unique challenges. The geographic diversity introduced by our dataset facilitates the development of more inclusive and robust models for autonomous driving and other related applications.

Thus, our approach addresses both the object detection task and predicts the frequency of pedestrian collisions. We introduce a baseline method that builds upon the state-of-the-art model DINO [98]. Our model can estimate the number of pedestrian collisions associated with a given location by leveraging actual visual features from the images and the corresponding geographical coordinates.

Our main contributions can be summarized as follows:

1. We propose the task of automated pedestrian collision prediction by considering the road-built environment in a specific location.

2. We establish an experimental framework for studying this problem in a city within the Latin American context, including frequencies for pedestrian collisions and detection labels for Google Street View panoramic images.

3. We empirically demonstrate that by training a multi-task model to detect objects in the urban built environment with a potential influence on pedestrian injuries, we can improve the predicting capabilities of the model.

To ensure the reproducibility of our results and to promote further research on predicting pedestrian collisions, we make all the resources of this paper publicly available on our project web page [1].

[1] https://github.com/BCV-Uniandes/STRIDE

2. Related Work

2.1. Built Environment Object Recognition in Autonomous Driving

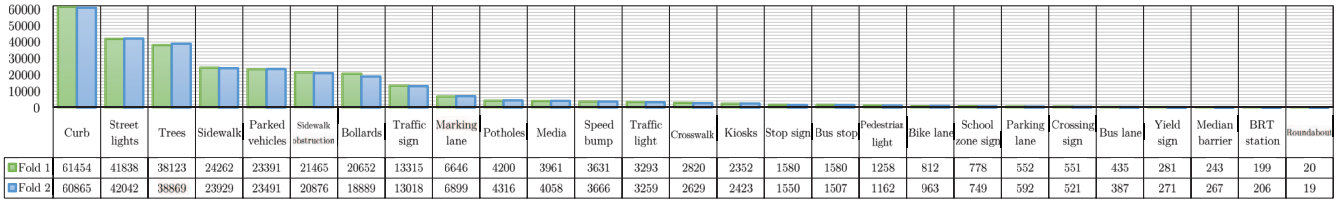
Object recognition for autonomous driving has been extensively studied. For instance, generic object detection benchmarks like MS COCO [50], and PascalVOC [33] include street images with annotated vehicles, pedestrians, and some general road elements. More specifically, pedestrian detection is a pioneering and well-studied task [22, 32, 26], explored with large and complex datasets [100, 9, 61, 18, 46], and even with 3D detection and tracking approaches [31, 35]. Broader benchmarks also cover detection of vehicles [27, 102, 12], traffic signs [27, 47, 30], and axis-aligned vehicle detection [8, 12]. However, these datasets focus mainly on dynamic agents and neglect the importance of static road infrastructure elements for comprehensive scene understanding.

On the contrary, standard benchmarks for urban static object identification include lane segmentation [81, 44], lane markings detection [44, 47], and multiple traffic signs detection [30]. Fine-grained urban scene parsing datasets like [10, 70, 25] provide pixel-level semantic annotations for street images, covering some static road infrastructure and dynamic agents. CityScapes [20] and KITTY [2] offer panoptic segmentation annotations, thus identifying instances within some semantic classes. Mapillary Vistas [60] extended this framework by including many more semantic categories and using highly variable user-uploaded data. BDD100K [94] extends these tasks through time by including segment tracking in videos, while ApolloScape [78], KITTI [35, 7, 6] and [80], focus on 3D point cloud segmentation. However, most of these benchmarks have limited classes for street infrastructure, often treating them as stuff categories, unlike our benchmark, which explicitly identifies and differentiates these objects to study their effect on pedestrian collision frequency.

2.2. Panoramic Street View Benchmarks

Initial frameworks for wide field of view autonomous driving tackle object detection of a few categories using panoramic images [47, 76, 59]. Further approaches use data gathered with fisheye and surround-view cameras annotated for semantic scene parsing [24, 92, 29, 43, 49], thus introducing significant distortion and deformation caused by this type of cameras. Other frameworks use synthetic data [71, 83] for the same task, which generates a considerable gap for real-world applications.

Fine-grained frameworks offer pixel-wise annotations on a few panoramas from annular lenses [84, 85] or Google Street Views [87, 86, 40, 99, 63] as testing sets for domain adaptation. Recently, Mapillary Metropolis [52] introduced the first large-scale panoramic panoptic segmenta-



	Curb	Street lights	Trees	Sidewalk	Parked vehicles	Sidewalk obstruction	Bollards	Traffic sign	Marking lane	Potholes	Media	Speed bump	Traffic light	Crosswalk	Kiosks	Stop sign	Bus stop	Pedestrian light	Bike lane	School zone sign	Parking lane	Crossing sign	Bus lane	Yield sign	Median barrier	BRT station	Roundabout
Fold 1	61454	41838	38123	24262	23391	21465	20652	13315	6646	4200	3961	3631	3293	2820	2352	1580	1580	1258	812	778	552	551	435	281	243	199	20
Fold 2	60865	42042	38869	23929	23491	20876	18889	13018	6899	4316	4058	3666	3259	2629	2423	1550	1507	1162	963	749	592	521	387	271	267	206	19

Figure 2: Number of labeled bounding boxes (y-axis) per class in each fold and their corresponding category name (x-axis).

tion benchmark with 360° panoramas of 4,000×8,000 resolution aligned with aerial images and 3D point clouds. Similarly, the Waymo Panoramic Video Panoptic Segmentation Dataset [53] used an extensive set of 220° panoramas for panoptic segmentation and segment tracking. Despite these advancements, many street infrastructure classes are still disregarded in these benchmarks, and only Mapillary Metropolis utilizes large panoramic images. On the contrary, our dataset considers multiple streets features categories on 4,000×13,312 images, surpassing any previous panoramic benchmark in image size.

2.3. Street Collision Prediction Benchmarks

Existing methods for collision frequency prediction rely primarily on tabular variables related to road properties, street conditions, traffic volume, and environmental factors [15, 14, 11, 79, 103]. More complex approaches incorporate historical collision records [17, 68], spatial and temporal relations among city regions and time windows [95, 77, 4, 96], and satellite images [56, 96]. Conversely, the US-Accidents benchmark [55, 54] provides information on the presence of general street components. However, these frameworks rely heavily on alternative or precalculated data rather than analyzing urban scene images captured from street-level perspectives.

In contrast, computer vision-based approaches predominantly employ detection and tracking methods to identify vehicle crashes [82, 88, 90, 64, 73, 75, 39] or to track vehicles and other agents to predict accidents and anomalies [57, 58, 89, 1, 72, 91, 37, 36, 101, 48]. Similarly, some frameworks estimate vehicle trajectories to anticipate collisions [19, 16, 38]. Other alternative benchmarks utilize reinforcement learning [5], forecasting in time series [41], and causality recognition [93] to anticipate accidents. Contrarily, some datasets specifically target pedestrian safety by predicting pedestrians' intentions to cross [66, 67], and some studies on this task have demonstrated that street infrastructure state considerably impacts pedestrian crossing prediction capabilities [42]. However, no previous framework has directly studied the relations between built environment elements and collision prediction using computer vision models.

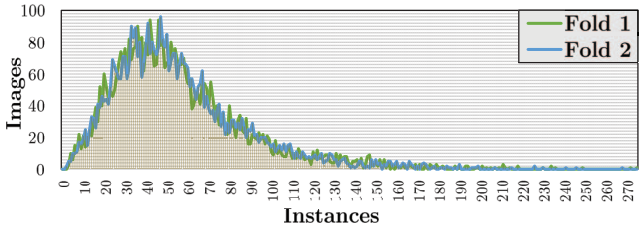


Figure 3: **Distribution of the number of bounding box annotations per image.** The figure shows the number of images (y-axis) with a certain number of annotated instances (x-axis). Our images contain a varying range of instances with a similar distribution among folds.

3. STRIDE Dataset

We introduce the Street View-based Environmental Feature and Pedestrian Collision Dataset (STRIDE), a novel challenging benchmark that studies the interrelations between built-environment elements and pedestrian collision frequency for scene awareness in autonomous driving. Figure 1 presents an overview of our benchmark. STRIDE combines multiple public data sources for two main tasks: (1) road-built-environment static object detection; and (2) image-guided pedestrian collision frequency estimation. In this section, we describe the details of our benchmark.

3.1. Image Gathering

First, we uniformly sampled random locations along the streets of Bogota City in Colombia, including specific points where statistical analysis indicated a higher incidence of pedestrian injuries. Secondly, to leverage complete 360° information, we utilize the 3D Google Street View service [97] to download panoramic images corresponding to the selected locations. The resulting dataset encompasses 18,036 panoramic images from different parts of the city. Our images have a high resolution of 4,000×13,312, making them the panoramic dataset with the most extensive images. We split our data into a training and validation set and a test set. Our training and validation set comprises 9,900 images on which we performed a 2-fold cross-validation to train and validate our experiments. The remaining 8,136 images are used for testing.

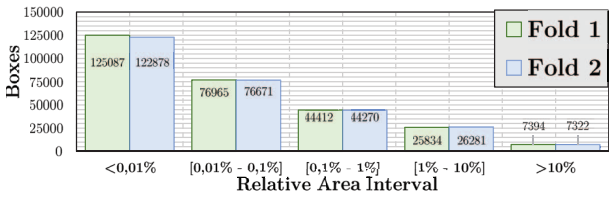


Figure 4: **Number of annotated bounding boxes (y-axis) per relative box area interval (x-axis).** Most instances in our dataset have small relative sizes due to the large scale of our images.

3.2. Detection Annotations

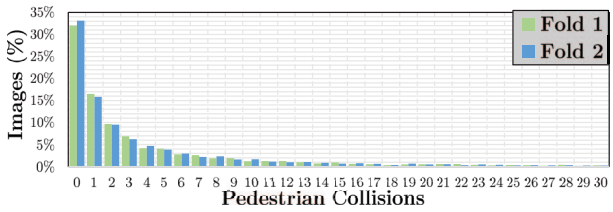
We selected a team of skilled annotators and trained them to draw bounding boxes around relevant static objects in the street infrastructure present in our panoramic images. Experts in built environment safety defined the classes used for annotation to encompass a comprehensive range of static objects that play crucial roles in pedestrian safety. We annotated all 9,900 images in our training set with 27 static street furniture classes indicated in Figure 2. For cross-validation splitting, we ensured that the distributions of street object classes, number of boxes per image, and area of boxes were maintained consistently across both folds.

3.3. Pedestrian Collision Annotations

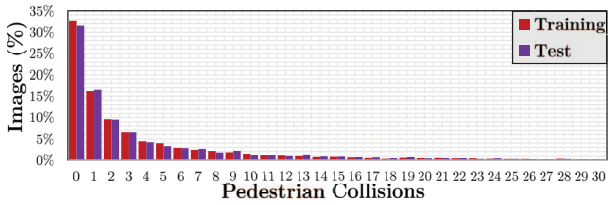
To correlate our dataset with real-world pedestrian collisions data, we leverage publicly available pedestrian collision records of Bogota City from 2015 until 2021. This data consists of records documenting the total number of vehicle collisions with pedestrians (pedestrian collisions) reported for each crossing point on the city's streets. By utilizing the corresponding geographic coordinates of each image, we associate each of our panoramic images with a single crossing point. We used ArcGIS 10.8 to perform geospatial analysis and geostatistics to identify each image's closest registered crossing point. Most of our images matched a registered point within a 30-meter distance; the rest were matched to points in a 100-meter radius. We filtered a few images outside of the urban area or with more than a 100-meter distance to a crossing point. We provide the distribution of distances in the Supplementary Material. In the end, 17,388 downloaded images were matched with a street segment, from which 9,252 belonged to detection annotated images. As for detection, we trained and cross-validated with these 9,252 images and tested on the remaining 8,136.

3.4. Dataset Statistics

Figure 3 presents the distribution of the number of annotated bounding boxes per image. We annotated a total of 557,115 objects. Our images contain an average of 56.5 annotations per image, a minimum of 2 boxes, and a maximum of 275 boxes per image. This diverse range of annotations per image allows for a robust representation of various street object distribution in different urban scenes. Furthermore, Figure 2 shows each semantic class's frequency distribution. We observe a considerable imbalance in the frequency distribution of our classes, as some parts of the street infrastructure, like curbs, street lights, and trees, are naturally more frequent than roundabouts, Bus Rapid Transit (BRT) Stations, and Median barriers. Hence, this long tail distribution is highly representative of real-world-urban scenes.



(a)



(b)

Figure 5: **Pedestrian Collision frequency distribution among training folds (a) and testing set (b).** Figures portray the percentage of images (y-axis) for each amount of pedestrian collisions (x-axis). Our dataset maintains a constant long-tail distribution among the training folds and the testing set. **Note:** The figures were cut to a maximum of 30 collisions for better visualization.

Additionally, our annotated bounding boxes exhibit highly varying sizes. Figure 4 portrays the distribution of the bounding box areas relative to the image size. The absolute areas of our annotated boxes range from 100 to $57.15e^6$ pixels with an average size of $4.37e^5$. Based on the MS COCO standards [50], the absolute areas of our annotations correspond mostly to large objects (see the Supplementary Material for more detail on box areas distribution according to MS COCO standards). However, due to the considerably large size of our images, most boxes have a low relative area with an average relative area of 0.82%. This characteristic makes our detection benchmark extremely challenging.

Finally, Figure 5(a) exhibits the histogram with the distribution of the number of pedestrian collisions per image in the training set folds. Figure 5(b) exhibits the distribution in the test set. The number of pedestrian collisions corresponding to each image presents a considerable imbal-

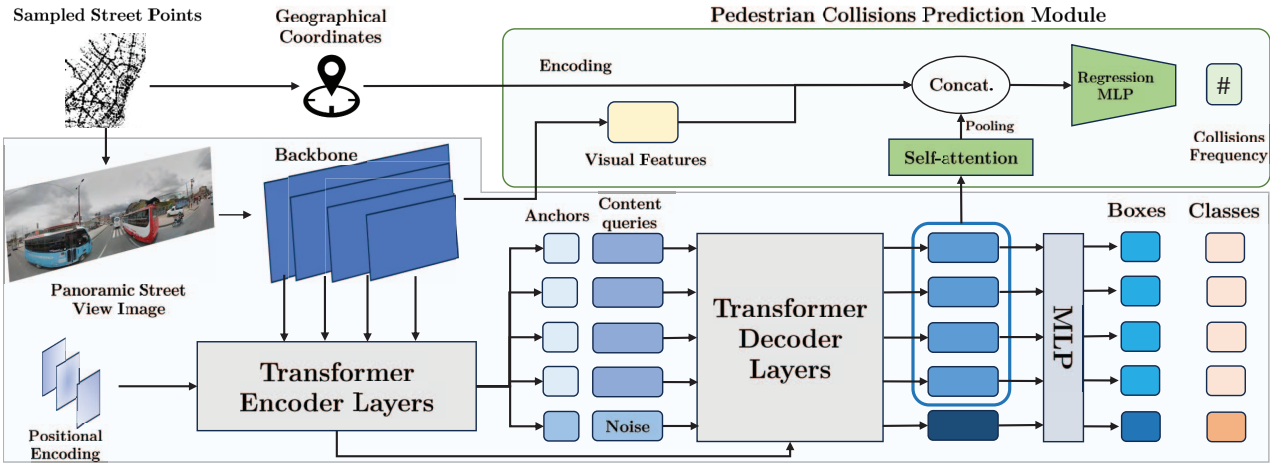


Figure 6: **STRIDE Baseline Method.** Our model uses DINO [98] (bottom) for object detection on an input panoramic image from a sampled street point. Our pedestrian collision prediction module (top-right) employs self-attention on the output embeddings of DINO's decoder to capture spatial and semantic relationships among object proposals. Additionally, we extract a visual feature from DINO's backbone, and we encode the geographical coordinates of the sampled street point. Finally, we pool the output features of the self-attention layer and concatenate them with the visual features and the encoded coordinates to perform a linear regression and estimate pedestrian collision frequency.

ance and an evident long tail in both sets as, naturally, most crossing points have a low pedestrian collision incidence. Our images have a minimum of 0 pedestrian collisions, a maximum of 193 pedestrian collisions, an average of 6.65 pedestrian collisions, and a standard deviation of 14.3. This imbalance in the pedestrian collisions distribution presents another challenge in our data.

3.5. Evaluation Metrics

We use the standard Average Precision (AP) metric from MS COCO [50] to evaluate the object detection task. For the pedestrian collisions prediction task, we adopt the root mean squared error (RMSE) as the primary evaluation metric as it has an increased sensitivity for large errors. Similarly, we avoid using the mean absolute error (MAE) metric since our unbalanced data easily biases it. As an alternative, we propose the Weighted Mean Absolute Error (WMAE) as a more stringent metric that severely penalizes underestimations, thereby addressing the specific challenges posed by our benchmark. We define WMAE as follows:

$$WMAE(y, \hat{y}) = \frac{\sum_i^N (y_i + 1)|y_i - \hat{y}_i|}{\sum_i^N (y_i + 1)} \quad (1)$$

Where y_i is the ground truth value of the ith image, $\hat{y}_i$ is the predicted value of the ith image, and N is the total number of images.

4. STRIDE Baseline

We propose a multi-task model capable of simultaneously detecting essential street infrastructure objects in panoramic street images and predicting the frequency of pedestrian collisions. Figure 6 depicts our model. The general intuition behind our method lies in leveraging the visual information captured in panoramic street images to identify built environment elements and their correlation with pedestrian collision occurrences. Thus, we employ a two-stage process; we first utilize the DINO [98] model to detect various street infrastructure objects in the images. Subsequently, we introduce a pedestrian collision prediction module that exploits self-attention to capture spatial and semantic relationships among the detected objects. By combining the extracted features from both stages, we aim to enhance the accuracy of pedestrian collision prediction.

4.1. Object Detector

We build upon DINO [98], a state-of-the-art efficient object detection model. As a DETR-like architecture [13], DINO utilizes a set-prediction approach to predict a fixed-size set z of $\mathcal{N}$ class probability-bounding box pairs using $\mathcal{N}$ input learnable queries. DINO combines multiple enhancements to the original DETR model [51, 45, 104]. First, it includes a mixed query selection approach that leverages information from the backbone and the encoder to initialize anchor boxes as positional embeddings. Additionally, a contrastive denoising training introduces negative and positive noise to ground truth boxes, facilitating bipartite matching and faster convergence. Finally, DINO also integrates box refinement and deformable attention from Deformable-DETR [104] to boost efficiency and performance. For further details on the DINO architecture, we re-

fer readers to the original DINO paper [98]. We adapt and optimize this architecture to detect and classify annotated built environment objects in our panoramic street images.

4.2. Pedestrian Collisions Prediction

We extend DINO with an additional pedestrian collisions prediction module (PCPM) as shown in figure 6. This module captures the relationships between the detected objects and the likelihood of pedestrian collisions in a given street image. On the one hand, the PCPM employs a self-attention layer on the output features of DINO's decoder to capture spatial and semantic information. This layer allows the module to focus on relevant regions and learn their interdependencies within identified objects. We exclude the features corresponding to noise queries from the module's processing as we only consider actual object proposals. On the other hand, we extract a visual embedding from the output feature map of DINO's backbone to capture general visual cues from the image. Furthermore, The output features from the self-attention layer are pooled and concatenated with the visual embedding. Additionally, we encode and concatenate the geographical coordinates of the image to include geospatial information. Finally, a regression multilayer perceptron (MLP) processes the concatenated embeddings and calculates the number of pedestrian collisions associated with the input image. This approach enables us to capture both the global context of the street image and localized details of detected objects to improve the predictive capabilities of our model.

4.3. Implementation Details

Object Detection: We train DINO with a ResNet50 backbone and DINO's *4-scale* implementation as we prioritize image resolution over model parameters. To fully exploit the resolution of our images, we train the model using a batch size of 1 and the largest image size allowed by our GPU resources, which is 1,800×5,990 pixels. We trained DINO for 50 epochs in 4 NVIDIA Quadro RTX GPUs with an SGD optimizer, and a learning rate of $1e^{-4}$ decayed by 0.1 after 12 epochs. We use random horizontal flips followed by either a random short side scale augmentation with a 1,600 to 1,800 range, or a short side rescaling between 1,920 and 3,000 pixels with a random crop of 1,800×5,990 pixels. Finally, we rescaled images to 1,800×5,990 for inference.

We discovered that using DINO's pretrained weights on the MS COCO dataset [50] is not beneficial for our specific task due to its substantial differences in class composition with our dataset (each dataset excludes most classes of the other). Hence, we pretrain our model for object detection in Mapillary Vistas [60] using only the categories with certain similarities to ours. We chose this dataset because it has the best coverage of static street infrastructure categories among previous datasets and provides highly diverse images that include some Latin American scenes.

Model	DINO [98]	Deformable-DETR [104]
AP	**32.59** $\pm$1.34	30.28 $\pm$0.07
AP_{50}	**50.53** $\pm$1.76	49.25 $\pm$0.22
AP_{75}	**34.46** $\pm$1.43	31.26 $\pm$0.15
AP_S	**18.11** $\pm$0.63	16.65 $\pm$0.29
AP_M	**45.10** $\pm$0.89	42.23 $\pm$0.18
AP_L	**56.40** $\pm$4.35	54.18 $\pm$0.30
Params.	47M	40M

Table 1: **Results in STRIDE's object detection task** of DINO [98] compared with Deformable DETR [104]. The best performances are shown in bold.

Pedestrian Collisions Prediction Module: We train our module for 20 epochs with a batch size of 5 on 4 Quadro RTX GPUs. We rescale images to 1,800×5,990 and use random horizontal flips for training. We use an $L2$ loss function with an SGD optimizer and a learning rate of $1e^{-4}$ decayed by 0.1 after 15 epochs.

5. Experiments

Experimental Setup: We train and validate DINO with all the 9,900 annotated images in our dataset and report the standard deviation error across folds. Additionally, we modify MS COCO's standards [50] of object sizes to our images. We define *small objects* as those with $< 0.01\%$ relative area, *medium* as 0.01% to 10% relative area, and *large* as $> 10\%$ relative area. We use the same cross-validation scheme for pedestrian collisions prediction and we test on the 8,136 images of the test set.

5.1. Detection

Table 1 shows the quantitative results of our best detection model. Our best-performing model achieves a considerably lower AP value compared to DINO's performance on generic benchmarks [98, 50]. This performance drop proves the challenging differences between objects' appearances in large panoramic images and standard images, underscoring the need for more specialized detection models tailored to our object detection task in panoramic urban environments. Moreover, our model has the lowest performance on small and medium objects, despite their high frequency in our dataset. This behavior is due to the image downscaling process, which strongly reduces the absolute area of objects, reaching areas of just 50 pixels. We provide qualitative results examples in the Supplementary Material. We note that DINO correctly identifies and locates most objects in our images but frequently produces false positive box predictions.

Predicted measure	Value	RMSE	WMAE
Mode	0	16.93 ±0.34	38.11 ±1.45
Median	2	16.24 ±0.36	36.37 ±1.50
Mean	6.70	16.32 ±1.03	36.11 ±3.80

Table 2: **Control experiment results for the pedestrian collisions prediction task** by constantly predicting the training set's mode, median, and mean to calculate the statistical lower bounds of our benchmark in our training folds.

Regarding per-class detection performance, we provide detailed results for each class in the Supplementary Material. Our model performs poorly in low frequent categories such as roundabouts and median barriers. However, we also observe that some highly frequent categories, such as curbs or sidewalks, do not yield high detection performances. We attribute this discrepancy to the high intra-class visual appearance variability of these objects, which are often occluded or appear in small sizes at distant parts of the images.

Finally, we compare our model with Deformable DETR [104] by adapting and optimizing it for our task using the same backbone and pretraining scheme. Table 1 also summarizes our overall results; detailed results are presented in the Supplementary Material. As expected, DINO outperforms Deformable-DETR in all metrics within fewer training iterations due to its contrastive denoising training. Nevertheless, regardless of the performance difference, we note that both detectors obtain similar value ranges and relative performances for all AP subtypes among both models. These consistencies validate the reliability of our results for object detection in our benchmark.

5.2. AutoML Regression

Initially, we conduct control experiments to establish reference points and statistical lower-bound metrics for our pedestrian collision prediction task. Specifically, we calculate our metrics for predicting statistical measurements of central tendencies, such as the training set's mean, mode, and median, as constant predictions for the validation process. Table 2 shows these results. Additionally, we perform baseline experiments using a model search with AutoML from Python's H2O library to predict pedestrian collisions. We show these results in Table 3. First, we run AutoML solely on the geographical coordinates of our images. Our results demonstrate a considerable correlation between the coordinates and pedestrian collision occurrences, which suggests that the model might learn to discern areas with elevated pedestrian collision frequency within the city. This observation also serves as an additional reference point to understand the relationship of geographical location information in pedestrian injury prediction.

Moreover, we study the impact of incorporating DINO's predictions of the number of instances per category in an image (object counts) for pedestrian collision prediction. We observe a notable increase in performance, indicating that a regression model can effectively leverage object presence information to predict pedestrian collisions. This finding underscores the importance of object recognition inputs, as they provide valuable cues to identify risk factors and assess the likelihood of pedestrian collisions. Similarly, we use ground truth counts and achieve a slight performance increase, therefore proving that the model benefits from improved object recognition. We obtain lower error values by including the geographical information as input, once again proving the impact of geographical location on pedestrian collisions prediction. Regardless of the promising performances of these baseline models, the performance of AutoML is highly limited by the lack of visual or spatial information processing.

Coordinates	Object Counts	RMSE	WMAE
✓	–	15.03 ±0.20	30.04 ±1.32
–	DINO	13.66 ±0.36	27.73 ±1.10
✓	DINO	13.58 ±0.36	27.65 ±1.25
–	Ground Truth	13.53 ±0.26	27.33 ±1.00
✓	Ground Truth	**13.44** ±0.26	**26.96** ±1.02

Table 3: **Results of AutoML experiments on the training folds** using geographical coordinates and the number of instances per class (object counts). We indicate the use (✓) or absence (–) of coordinates and counts and whether the counts are obtained from DINO predictions or ground truth annotations. The best results are shown in bold.

5.3. Pedestrian Collisions Prediction Module

To evaluate the impact of the detection task on pedestrian collision prediction, we first calculate pedestrian collisions directly from the sole backbone and the input coordinates using an MLP. We use the backbone pretrained in ImageNet [23] and train until convergence. The results of this experiment are shown in Table 4 as *Backbone Regression*. Using the backbone yields a better performance than AutoML's best model, thus proving the importance of visual information for pedestrian collision prediction as it provides general context features to identify risky environments.

Additionally, we explore various configurations for the pedestrian collisions predictions module to optimize the performance of our model. Initially, we use a linear layer on the outputs of DINO's decoder; this experiment is portrayed in Table 4 as *Linear Layer*. This configuration outperforms AutoML's best method and the *Backbone Regression* experiment. This result underscores the potential of object detection embeddings for pedestrian collision prediction. Subsequently, we introduce the self-attention layer (shown in Table 4 as *Self-Att. Layer*) instead of the linear

Module Design	RMSE	WMAE	Params. (M)
Backbone Regression	13.26 ±0.11	27.01 ±0.80	0.54
Linear Layer	13.11 ±0.49	26.07 ±1.68	0.11
Self-Att. Layer	12.79 ±0.38	25.51 ±0.78	0.44
Self-Att. + Visual Embd.	**12.67** ±0.33	**24.73** ±1.05	1.08

Table 4: **Results of the pedestrian collision prediction module in our training folds.** We compare different design choices of our pedestrian collision prediction module. We present the number of parameters corresponding just to the regression module. The best results are shown in bold.

layer, and we note a significant boost in performance for our model, as it captures relationships and dependencies among the spatial and semantic information encoded in the output embeddings of DINO's decoder. Hence, self-attention enables the model to understand better associations among predicted objects which is essential for pedestrian collision prediction.

Furthermore, we incorporate the visual embeddings extracted from DINO's backbone into the pedestrian collision prediction module (named *Self-Att. + Visual Embd.*). This modification led to a further increase in performance, especially in $WMAE$. The visual embeddings encapsulate general visual information before object location, providing valuable context cues for pedestrian collision prediction. These results demonstrate the importance of leveraging detection and visual information for accurate pedestrian collision estimation in street scenes. Finally, we explore multiple encoding techniques to incorporate geographical coordinates into the model. These techniques included linear projection or positional encoding. However, we observed no significant performance improvement compared to a simple normalization approach. Our results also demonstrate that the increase in model performance requires a significant computational cost as the number of parameters for each module design (shown in Table 4) increases with the inclusion of the self-attention and visual features.

5.4. Model Testing

We directly evaluate our best model for pedestrian collision prediction and compare it with our best AutoML method that takes the object counts predicted by DINO as input. Table 5 presents our results. Both models exhibit similar performance on the test and the cross-validation sets. The relative behavior of both models remains consistent, with the prediction module of DINO consistently outperforming the AutoML model. This observation validates the superiority of the prediction module and reinforces the efficacy of calculating pedestrian collisions directly from the visual, spatial, and semantic embeddings provided by DINO. Our findings confirm the validity of our proposal and underscore the value of leveraging visual information from DINO for accurate pedestrian collision prediction. We provide the distribution of the Exact Error among predictions on the test set in the Supplementary Material. Most of our model predictions achieve low error values with a tendency towards slight overestimations. However, the model fails the most in predicting large pedestrian collision values due to the reduced frequency of these samples.

Model	RMSE	WMAE
AutoML	13.78 ±0.02	28.55 ±0.12
STRIDE Baseline	**12.88** ±0.01	**23.41** ±0.40

Table 5: **Results of STRIDE's baseline on the test set.** We compare the performance of our best AutoML model with the best pedestrian collision prediction module configuration. The best results are shown in bold.

6. Conclusions

This paper introduces STRIDE, a novel benchmark to improve environmental awareness in autonomous driving by studying the relationship of built environment elements in pedestrian injury prediction. Our framework introduces a multitask approach to simultaneously detect relevant built-environment features and estimate pedestrian collision frequency. We present a new dataset that geographically associates public records of pedestrian collisions with large-scale panoramic street view images, manually annotated for built environment detection of 27 categories. By presenting multiple challenges representative of real-world situations, our benchmark provides a robust testing ground for image-guided pedestrian collision prediction models. To pave the way for future research, we propose a strong baseline that combines a state-of-the-art object detector with an additional collision prediction module. Our experimental validation demonstrates that our model can leverage our detection annotations by capturing interrelations among built environment features to estimate pedestrian collision frequency. Our benchmark promotes the development of autonomous agents capable of predicting pedestrian collision events solely from visual inputs and GPS coordinates, which holds the potential to enhance situation awareness and real-time active collision prevention. Hence, our work is a stepping stone towards improving security in autonomous driving systems, mitigating potential risks, and ensuring safer transportation.

Acknowledgements: Nicolás Ayobi acknowledges the support of the 2022 Uniandes-DeepMind scholarship. This research was supported by the Fogarty International Center of the National Institutes of Health (NIH) under award numbers K01TW011782 and 3K01TW011782-01S1. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

[1] Armstrong Aboah, Maged Shoman, Vishal Mandal, Sayedomidreza Davami, Yaw Adu-Gyamfi, and Anuj Sharma. A vision-based system for traffic anomaly detection using deep learning and decision trees. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4202–4207, 2021. 3

[2] Hassan Alhaija, Siva Mustikovela, Lars Mescheder, Andreas Geiger, and Carsten Rother. Augmented reality meets computer vision: Efficient data generation for urban driving scenes. *International Journal of Computer Vision (IJCV)*, 2018. 2

[3] Dragomir Anguelov, Carole Dulong, Daniel Filip, Christian Frueh, Stéphane Lafon, Richard Lyon, Abhijit Ogale, Luc Vincent, and Josh Weaver. Google street view: Capturing the world at street level. *Computer*, 43(6):32–38, 2010. 2

[4] Jie Bao, Pan Liu, and Satish V. Ukkusuri. A spatiotemporal deep learning approach for citywide short-term crash risk prediction with multi-source data. *Accident Analysis &amp Prevention*, 122:239–254, Jan. 2019. 3

[5] Wentao Bao, Qi Yu, and Yu Kong. Drive: Deep reinforced accident anticipation with visual explanation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7599–7608, 2021. 3

[6] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, J. Gall, and C. Stachniss. Towards 3D LiDAR-based semantic scene understanding of 3D point cloud sequences: The SemanticKITTI Dataset. *The International Journal on Robotics Research*, 40(8-9):959–967, 2021. 2

[7] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall. SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences. In *Proc. of the IEEE/CVF International Conf. on Computer Vision (ICCV)*, 2019. 2

[8] Karsten Behrendt. Boxy vehicle detection in large images. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 840–846, 2019. 2

[9] Markus Braun, Sebastian Krebs, Fabian B. Flohr, and Dariu M. Gavrila. Eurocity persons: A novel benchmark for person detection in traffic scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2019. 2

[10] Gabriel J. Brostow, Julien Fauqueur, and Roberto Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 30(2):88–97, 2009. Video-based Object and Event Analysis. 2

[11] Ciro Caliendo, Maurizio Guida, and Alessandra Parisi. A crash-prediction model for multilane roads. *Accident Analysis Prevention*, 39(4):657–670, 2007. 3

[12] Claudio Caraffi, Tomáš Vojíř, Jiří Trefný, Jan Šochman, and Jiří Matas. A system for real-time detection and tracking of vehicles from a single car-mounted camera. In *2012 15th International IEEE Conference on Intelligent Transportation Systems*, pages 975–982, 2012. 2

[13] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I*, page 213–229, Berlin, Heidelberg, 2020. Springer-Verlag. 5

[14] Li-Yen Chang. Analysis of freeway accident frequencies: Negative binomial regression versus artificial neural network. *Safety Science*, 43(8):541–557, 2005. 3

[15] Li-Yen Chang and Wen-Chieh Chen. Data mining of tree-based models to analyze freeway accident frequency. *Journal of Safety Research*, 36(4):365–375, 2005. 3

[16] Deesha Chavan, Dev Saad, and Debarati B. Chakraborty. COLLIDE-PRED: prediction of on-road collision from surveillance videos. *CoRR*, abs/2101.08463, 2021. 3

[17] Chao Chen, Xiaoliang Fan, Chuanpan Zheng, Lujing Xiao, Ming Cheng, and Cheng Wang. Sdcae: Stack denoising convolutional autoencoder model for accident risk prediction via traffic big data. In *2018 Sixth International Conference on Advanced Cloud and Big Data (CBD)*, pages 328–333, 2018. 3

[18] Yukyung Choi, Namil Kim, Soonmin Hwang, Kibaek Park, Jae Shin Yoon, Kyounghwan An, and In So Kweon. Kaist multi-spectral day/night data set for autonomous and assisted driving. *IEEE Transactions on Intelligent Transportation Systems*, 19(3):934–948, 2018. 2

[19] Richard Coll-Josifov, Albert Masip-Álvarez, and David Lavèrnia-Ferrer. Deep learning classification applied to traffic accidents prediction. In *XLIII Jornadas de Automática: libro de actas: 7, 8 y 9 de septiembre de 2022, Logroño (La Rioja)*, pages 964–971. Servizo de Publicacións da UDC, Sept. 2022. 3

[20] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[21] Curtis M Craig, Nichole L Morris, Ron Van Houten, and David Mayou. Pedestrian safety and driver yielding near public transit stops. *Transportation research record*, 2673(1):514–523, 2019. 1

[22] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893 vol. 1, 2005. 2

[23] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 7

[24] Liuyuan Deng, Ming Yang, Hao Li, Tianyi Li, Bing Hu, and Chunxiang Wang. Restricted deformable convolution-based road scene semantic segmentation using surround view cameras. *IEEE Transactions on Intelligent Transportation Systems*, 21:4350–4362, 2018. 2

[25] Li Ding, Jack Terwilliger, Rini Sherony, Bryan Reimer, and Lex Fridman. Mit driveseg (manual) dataset, 2020. 2

[26] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state

of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):743–761, 2012. 2

[27] Alex Dominguez-Sanchez, Sergio Orts-Escolano, Jose Garcia-Rodriguez, and Miguel Cazorla. A new dataset and performance evaluation of a region-based cnn for urban object detection. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2018. 2

[28] Joseph Donroe, Monica Tincopa, Robert H Gilman, Doug Brugge, and David AJ Moore. Pedestrian road traffic injuries in urban peruvian children and adolescents: case control analyses of personal and environmental risk factors. *PLoS One*, 3(9):e3166, 2008. 1

[29] Ciarán Eising, Jonathan Horgan, and Senthil Yogamani. Near-field perception for low-speed vehicle automation using surround-view fisheye cameras. *Trans. Intell. Transport. Sys.*, 23(9):13976–13993, sep 2022. 2

[30] Christian Ertler, Jerneja Mislej, Tobias Ollmann, Lorenzo Porzi, Gerhard Neuhold, and Yubin Kuang. The mapillary traffic sign dataset for detection and classification on a global scale. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 68–84, Cham, 2020. Springer International Publishing. 2

[31] Andreas Ess, Bastian Leibe, Konrad Schindler, and Luc Van Gool. A mobile vision system for robust multi-person tracking. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008. 2

[32] Andreas Ess, Bastian Leibe, and Luc Van Gool. Depth and appearance for mobile scene analysis. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8, 2007. 2

[33] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, Jan. 2015. 2

[34] Nick Foster, Christopher M Monsere, and Katherine Carlos. Evaluating driver and pedestrian behaviors at enhanced, multilane, midblock pedestrian crossings: Case study in portland, oregon. *Transportation Research Record*, 2464(1):59–66, 2014. 1

[35] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012. 2

[36] Hadi Ghahremannezhad, Hang Shi, and Chengjun Liu. A real time accident detection framework for traffic video analysis. 07 2020. 3

[37] Hadi Ghahremannezhad, Hang Shi, and Chengjun Liu. Real-time accident detection in traffic surveillance using deep learning. In *2022 IEEE International Conference on Imaging Systems and Techniques (IST)*, page 1–6. IEEE Press, 2022. 3

[38] Xiaohui Huang, Pan He, Anand Rangarajan, and Sanjay Ranka. Intelligent intersection: Two-stream convolutional networks for real-time near-accident detection in traffic video. *ACM Trans. Spatial Algorithms Syst.*, 6(2), jan 2020. 3

[39] Earnest Paul Ijjina, Dhananjai Chand, Savyasachi Gupta, and K. Goutham. Computer vision-based accident detection in traffic surveillance. In *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–6, 2019. 3

[40] Alexander Jaus, Kailun Yang, and Rainer Stiefelhagen. Panoramic panoptic segmentation: Insights into surrounding parsing for mobile agents via unsupervised contrastive learning. *Trans. Intell. Transport. Sys.*, 24(4):4438–4453, jan 2023. 2

[41] Eitan Kosman and Dotan Di Castro. Vision-guided forecasting – visual context for multi-horizon time series forecasting, 2021. 3

[42] Iuliia Kotseruba, Amir Rasouli, and John K. Tsotsos. Do they want to cross? understanding pedestrian intention for behavior prediction. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, pages 1688–1693, 2020. 1, 3

[43] Varun Ravi Kumar, Senthil Kumar Yogamani, Hazem Rashed, Ganesh Sitsu, Christian Witt, Isabelle Leang, Stefan Milz, and Patrick Mäder. Omnidet: Surround view cameras based multi-task visual perception network for autonomous driving. *IEEE Robotics and Automation Letters*, 6:2830–2837, 2021. 2

[44] Seokju Lee, Junsik Kim, Jae Shin Yoon, Seunghak Shin, Oleksandr Bailo, Namil Kim, Tae-Hee Lee, Hyun Seok Hong, Seung-Hoon Han, and In So Kweon. Vpgnet: Vanishing point guided network for lane and road marking detection and recognition. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1965–1973, 2017. 2

[45] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13619–13627, 2022. 5

[46] Xiaofei Li, Fabian Flohr, Yue Yang, Hui Xiong, Markus Braun, Shuyue Pan, Keqiang Li, and Dariu M. Gavrila. A new benchmark for vision-based cyclist detection. In *2016 IEEE Intelligent Vehicles Symposium (IV)*, pages 1028–1033, 2016. 2

[47] Yong Li, Guofeng Tong, Huashuai Gao, Yuebin Wang, Liqiang Zhang, and Huairong Chen. Pano-RSOD: A dataset and benchmark for panoramic road scene object detection. *Electronics*, 8(3):329, Mar. 2019. 2

[48] Yingying Li, Jie Wu, Xue Bai, Xipeng Yang, Xiao Tan, Guanbin Li, Shilei Wen, Hongwu Zhang, and Errui Ding. Multi-granularity tracking with modularlized components for unsupervised vehicles anomaly detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2501–2510, 2020. 3

[49] Yiyi Liao, Jun Xie, and Andreas Geiger. KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *Pattern Analysis and Machine Intelligence (PAMI)*, 2022. 2

[50] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and

C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. 2, 4, 5, 6

[51] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. DAB-DETR: Dynamic anchor boxes are better queries for DETR. In *International Conference on Learning Representations*, 2022. 5

[52] Mapillary. Mapillary metropolis dataset, 2021. 2

[53] Jieru Mei, Alex Zihao Zhu, Xinchen Yan, Hang Yan, Siyuan Qiao, Liang-Chieh Chen, and Henrik Kretzschmar. Waymo open dataset: Panoramic video panoptic segmentation. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIX*, page 53–72, Berlin, Heidelberg, 2022. Springer-Verlag. 3

[54] Sobhan Moosavi, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. A countrywide traffic accident dataset. *CoRR*, abs/1906.05409, 2019. 3

[55] Sobhan Moosavi, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. Accident risk prediction based on heterogeneous sparse data: New dataset and insights. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '19, page 33–42, New York, NY, USA, 2019. Association for Computing Machinery. 2, 3

[56] Alameen Najjar, Shun'ichi Kaneko, and Yoshikazu Miyanaga. Combining satellite imagery and open data to map road safety. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4524–4530. AAAI Press, 2017. 3

[57] Milind Naphade, Shuo Wang, David Anastasiu, Zheng Tang, Ming-Ching Chang, Xiaodong Yang, Liang Zheng, Anuj Sharma, Rama Chellappa, and Pranamesh Chakraborty. The 4th ai city challenge. pages 2665–2674, 06 2020. 3

[58] Milind Naphade, Shuo Wang, David C. Anastasiu, Zheng Tang, Ming-Ching Chang, Xiaodong Yang, Yue Yao, Liang Zheng, Pranamesh Chakraborty, Anuj Sharma, Qi Feng, Vitaly Ablavsky, and Stan Sclaroff. The 5th ai city challenge. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4258–4268, 2021. 3

[59] A. Nassar, S. Lefevre, and J. Wegner. Simultaneous multi-view instance detection with learned geometric soft-constraints. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6558–6567, Los Alamitos, CA, USA, nov 2019. IEEE Computer Society. 2

[60] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulò, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5000–5009, 2017. 2, 6

[61] Lukáš Neumann, Michelle Karg, Shanshan Zhang, Christian Scharfenberger, Eric Piegert, Sarah Mistr, Olga Prokofyeva, Robert Thiel, Andrea Vedaldi, Andrew Zisserman, and Bernt Schiele. Nightowls: A pedestrians at night dataset. In C. V. Jawahar, Hongdong Li, Greg Mori, and Konrad Schindler, editors, *Computer Vision – ACCV 2018*, pages 691–705, Cham, 2019. Springer International Publishing. 2

[62] Joshua Niemeijer, Paulin Pekezou Fouopi, Sascha Knake-Langhorst, and Erhardt Barth. A review of neural network based semantic segmentation for scene understanding in context of the self driving car. In *Student Conference on Medical Engineering Science*. Infinite Science Publishing, 2017. 1

[63] Semih Orhan and Yalin Bastanlar. Semantic segmentation of outdoor panoramic images. *Signal, Image and Video Processing*, 16(3):643–650, Apr 2022. 2

[64] Karishma Pawar and Vahida Attar. Deep learning based detection and localization of road accidents from traffic surveillance videos. *ICT Express*, 8(3):379–387, 2022. 3

[65] D Alex Quistberg, Thomas D Koepsell, J Jaime Miranda, Linda Ng Boyle, Brian D Johnston, and Beth E Ebel. The walking environment in lima, peru and pedestrian–motor vehicle collisions: an exploratory analysis. *Traffic injury prevention*, 16(3):314–321, 2015. 1

[66] Amir Rasouli, Iuliia Kotseruba, Toni Kunic, and John Tsotsos. Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6261–6270, 2019. 3

[67] Amir Rasouli, Iuliia Kotseruba, and John K. Tsotsos. Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 206–213, 2017. 3

[68] Honglei Ren, You Song, Jingwen Wang, Yucheng Hu, and Jinzhi Lei. A deep learning approach to the citywide traffic accident risk prediction. *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 3346–3351, 2017. 3

[69] Richard A Retting, Susan A Ferguson, and Anne T McCartt. A review of evidence-based traffic engineering measures designed to reduce pedestrian–motor vehicle crashes. *American journal of public health*, 93(9):1456–1463, 2003. 1

[70] Timo Scharwächter, Markus Enzweiler, Uwe Franke, and Stefan Roth. Efficient multi-cue scene segmentation. In Joachim Weickert, Matthias Hein, and Bernt Schiele, editors, *Pattern Recognition*, pages 435–445, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. 2

[71] A. R. Sekkat, Y. Dupuis, P. Vasseur, and P. Honeine. The omniscape dataset. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1603–1608, 2020. 2

[72] Ankit Parag Shah, Jean-Bapstite Lamare, Tuan Nguyen-Anh, and Alexander Hauptmann. Cadp: A novel dataset for cctv traffic camera based accident analysis. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–9, 2018. 3

[73] Dinesh Singh and Chalavadi Krishna Mohan. Deep spatio-temporal representation for detection of road accidents us-

ing stacked autoencoder. *IEEE Transactions on Intelligent Transportation Systems*, 20(3):879–887, 2019. 3

[74] Virginia P Sisiopiku and Darcin Akin. Pedestrian behaviors at and perceptions towards various pedestrian facilities: an examination based on observation and survey data. *Transportation research part f: traffic psychology and behaviour*, 6(4):249–274, 2003. 1

[75] Aparajith Srinivasan, Anirudh Srikanth, Haresh Indrajit, and Venkateswaran Narasimhan. A novel approach for road accident detection using detr algorithm. In *2020 International Conference on Intelligent Data Science Technologies and Applications (IDSTA)*, pages 75–80, 2020. 3

[76] Guofeng Tong, Huairong Chen, Yong Li, Xiance Du, and Qingchun Zhang. Object detection for panoramic images based on ms-rpn structure in traffic road scenes. *IET Computer Vision*, 13(5):500–506, 2019. 2

[77] Beibei Wang, Youfang Lin, Shengnan Guo, and Huaiyu Wan. GSNet: Learning spatial-temporal correlations from geographical and semantic aspects for traffic accident risk forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(5):4402–4409, May 2021. 3

[78] Peng Wang, Xinyu Huang, Xinjing Cheng, Dingfu Zhou, Qichuan Geng, and Ruigang Yang. The apolloscape open dataset for autonomous driving and its application. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 2

[79] Lu Wenqi, Luo Dongyu, and Yan Menghua. A model of traffic accident prediction based on convolutional neural network. In *2017 2nd IEEE International Conference on Intelligent Transportation Engineering (ICITE)*, pages 198–202, 2017. 3

[80] Jun Xie, Martin Kiefel, Ming-Ting Sun, and Andreas Geiger. Semantic instance annotation of street scenes by 3d to 2d label transfer. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3688–3697, 2016. 2

[81] Ping Luo Xiaogang Wang Xingang Pan, Jianping Shi and Xiaoou Tang. Spatial as deep: Spatial cnn for traffic scene understanding. In *AAAI Conference on Artificial Intelligence (AAAI)*, February 2018. 2

[82] Yajun Xu, Chuwen Huang, Yibing Nan, and Shiguo Lian. Tad: A large-scale benchmark for traffic accidents detection from video surveillance, 2022. 3

[83] Yuanyou Xu, Kaiwei Wang, Kailun Yang, Dongming Sun, and Jia Fu. Semantic segmentation of panoramic images using a synthetic dataset. In Judith Dijk, editor, *Artificial Intelligence and Machine Learning in Defense Applications*, volume 11169, page 111690B. International Society for Optics and Photonics, SPIE, 2019. 2

[84] Kailun Yang, Xinxin Hu, Luis M. Bergasa, Eduardo Romera, and Kaiwei Wang. Pass: Panoramic annular semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 21(10):4171–4185, 2020. 2

[85] Kailun Yang, Xinxin Hu, Hao Chen, Kaite Xiang, Kaiwei Wang, and Rainer Stiefelhagen. Ds-pass: Detail-sensitive panoramic annular semantic segmentation through swaftnet for surrounding sensing. *2020 IEEE Intelligent Vehicles Symposium (IV)*, pages 457–464, 2019. 2

[86] Kailun Yang, Xinxin Hu, and Rainer Stiefelhagen. Is context-aware cnn ready for the surroundings? panoramic semantic segmentation in the wild. *IEEE Transactions on Image Processing*, 30:1866–1881, 2021. 2

[87] Kailun Yang, Jiaming Zhang, Simon Reiß, Xinxin Hu, and Rainer Stiefelhagen. Capturing omni-range context for omnidirectional segmentation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[88] Yu Yao, Xizi Wang, Mingze Xu, Zelin Pu, Ella Atkins, and David Crandall. When, where, and what? a new dataset for anomaly detection in driving videos, 2020. 3

[89] Yu Yao, Xizi Wang, Mingze Xu, Zelin Pu, Yuchen Wang, Ella Atkins, and David J. Crandall. Dota: Unsupervised detection of traffic anomaly in driving videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):444–459, 2023. 3

[90] Yu Yao, Mingze Xu, Yuchen Wang, David J. Crandall, and Ella M. Atkins. Unsupervised traffic accident detection in first-person videos. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, page 273–280. IEEE Press, 2019. 3

[91] Yu Yao, Mingze Xu, Yuchen Wang, David J. Crandall, and Ella M. Atkins. Unsupervised traffic accident detection in first-person videos. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 273–280, 2019. 3

[92] Senthil Yogamani, Ciaran Hughes, Jonathan Horgan, Ganesh Sistu, Sumanth Chennupati, Michal Uricar, Stefan Milz, Martin Simon, Karl Amende, Christian Witt, Hazem Rashed, Sanjaya Nayak, Saquib Mansoor, Padraig Varley, Xavier Perrotton, Derek Odea, and Patrick Pérez. Woodscape: A multi-task, multi-camera fisheye dataset for autonomous driving. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9307–9317, 2019. 2

[93] Tackgeun You and Bohyung Han. Traffic accident benchmark for causality recognition. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 540–556, Cham, 2020. Springer International Publishing. 3

[94] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2633–2642, 2020. 2

[95] Le Yu, Bowen Du, Xiao Hu, Leilei Sun, Liangzhe Han, and Weifeng Lv. Deep spatio-temporal graph convolutional network for traffic accident prediction. *Neurocomputing*, 423:135–147, 2021. 3

[96] Zhuoning Yuan, Xun Zhou, and Tianbao Yang. Hetero-convlstm: A deep learning approach to traffic accident prediction on heterogeneous spatio-temporal data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, KDD '18, page 984–992, New York, NY, USA, 2018. Association for Computing Machinery. 3

[97] Amir Roshan Zamir and Mubarak Shah. Image geo-localization based on MultipleNearest neighbor feature matching UsingGeneralized graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(8):1546–1558, Aug. 2014. 3

[98] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 2, 5, 6

[99] Jiaming Zhang, Chaoxiang Ma, Kailun Yang, Alina Roitberg, Kunyu Peng, and Rainer Stiefelhagen. Transfer beyond the field of view: Dense panoramic semantic segmentation via unsupervised domain adaptation. *IEEE Transactions on Intelligent Transportation Systems*, 23(7):9478–9491, 2022. 2

[100] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Citypersons: A diverse dataset for pedestrian detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4457–4465, 2017. 2

[101] Jianfei Zhao, Zitong Yi, Siyang Pan, Yanyun Zhao, Zhicheng Zhao, Fei Su, and Bojin Zhuang. Unsupervised traffic anomaly detection using trajectories. In *CVPR Workshops*, 2019. 3

[102] Zhengping Che, Bo Jiang, Yiping Meng, Guangyu Li, Tracy Li, Ke Dong, Xinsheng Zhang, Xuefeng Shi, Ying Lyu, Guobin Wu, Yan Liu, Jian Tang, and Jieping Ye. D²-city: A large-scale dashcam video dataset of diverse traffic scenarios, 2021. 2

[103] Lei Zhu, Tianrui Li, and Shengdong Du. Ta-stan: A deep spatial-temporal attention learning framework for regional traffic accident risk prediction. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2019. 3

[104] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021. 5, 6, 7