Explainable, automated urban interventions to improve pedestrian and vehicle safety

C. Bustos[a,*], D. Rhoads[a,*], A. Solé-Ribalta[a], D. Masip[a], A. Arenas[c], A. Lapedriza[a,b], J. Borge-Holthoefer[a,*]

[a]*Internet Interdisciplinary Institute (IN3), Universitat Oberta de Catalunya, Barcelona 08860, Catalonia, Spain*
[b]*Media Lab, Massachusetts Institute of Technology, 02139 Cambridge, MA*
[c]*Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili, 43007 Tarragona, Spain*

Abstract

At the moment, urban mobility research and governmental initiatives are mostly focused on motor-related issues, e.g. the problems of congestion and pollution. And yet, we can not disregard the most vulnerable elements in the urban landscape: pedestrians, exposed to higher risks than other road users. Indeed, safe, accessible, and sustainable transport systems in cities are a core target of the UN's 2030 Agenda. Thus, there is an opportunity to apply advanced computational tools to the problem of traffic safety, in regards especially to pedestrians, who have been often overlooked in the past. This paper combines public data sources, large-scale street imagery and computer vision techniques to approach pedestrian and vehicle safety with an automated, relatively simple, and universally-applicable data-processing scheme. The steps involved in this pipeline include the adaptation and training of a Residual Convolutional Neural Network to determine a hazard index for each given urban scene, as well as an interpretability analysis based on image segmentation and class activation mapping on those same images. Combined, the outcome of this computational approach is a fine-grained map of hazard levels across a city, and an heuristic to identify interventions that might simultaneously improve pedestrian and vehicle safety. The proposed frame-

*Corresponding author
Email addresses: mbustosro@uoc.edu (C. Bustos), drhoads@uoc.edu (D. Rhoads), jborgeh@uoc.edu (J. Borge-Holthoefer)

work should be taken as a complement to the work of urban planners and public authorities.

Keywords: Deep Learning, Google Street View, Mapillary, Pedestrian, Traffic safety

1. Introduction

In the last century, the accelerated growth of urban areas has given rise to challenges at a variety of levels. Among these, mobility stands out. The ability to efficiently move people and goods is critical to a city's social and economic success [1–3]. It is unsurprising, then, the enormous amount of economic and engineering effort that urban planners have devoted to enhance the efficiency of road networks, bus lines, and metro systems [4]. Unlike transportation modes that operate in exclusive spaces, such as metro lines, the uncontrolled rise in urban automotive mobility has gone hand in hand with the degradation of other modes of transportation. Of all these alternative modes, walking has suffered the most, due in large part to the fact that the amount of the streetscape allotted to vehicles invades and interferes with the pedestrian space. Nevertheless, cities exhibit a growing tendency to stop and reverse this process by fostering more active, citizen-friendly transportation modes –foot, bike and personal mobility vehicles, which compete for this public space [5].

One logical consequence of this paradigm shift, is the increased level of interaction between pedestrians and motor vehicles, largely due to the overlapping use of common (or adjacent) spaces such as roads, sidewalks, and zebra-crossings. Such increase gives rise to an important, negative side-effect: a growth in pedestrian injuries and fatalities. Data from the National Highway Traffic Safety Administration (NHTSA) of the United States indicate that the number of pedestrian fatalities per year is rising in the U.S. [6]. After a steady decline from the mid-1990's to a low in 2009, there has been a clear and consistent reversal until 2017 (the last year of available data), when pedestrian fatalities surpassed a previous 23-year high in 1995.

Traditionally, pedestrian safety research has focused on the impact of structural factors (e.g. road lanes [7], traffic network structure [8, 9], existence of direct line-of-sight between objects [10, 11], etc.). In addition, socio-behavioral factors may be concomitant, e.g. the change of individual behavior related to the use of new, distraction-causing technologies [12], in-

side and outside of vehicles, which is not likely to diminish in the future. Also, demographic variables (socio-economic status, race, gender) may play a role as well [13]. Nonetheless, crashes that involve motor vehicles and pedestrians are understudied, and, at the micro level, much less so outside intersections [14].

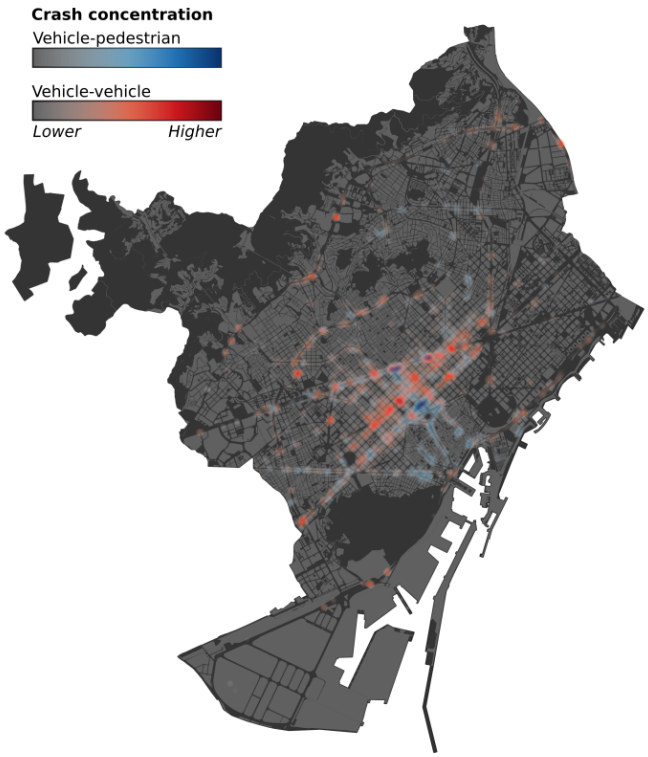

Figure 1: **Accident distribution in Barcelona.** Relative concentration of accidents by type (vehicle-to-pedestrian, vehicle-to-vehicle).

An enlightening example, built upon real accident data, is shown in Figure 1. Quite clear even to the naked eye, accidents involving vehicles may happen throughout a city. However, when a distinction is introduced (vehicle-to-vehicle *vs.* vehicle-to-pedestrian), the spatial patterns where these accidents occur are mostly non-overlapping, suggesting that the configuration of the public space –the scene where the accident happens– matters, see as well Figure S1 in the Supplementary Information (SI). All in all, the strategies

for the safe coexistence of pedestrians and vehicles demand a separate and careful examination.

The combination of increasingly available street-level imagery sources and city open data portals, together with advances in the field of computer vision and larger training datasets [15, 16], has opened up promising new opportunities for facing challenges in urban science. Examples include the quantification of physical change and pattern identification in cities [17–19], road safety assessment [20], the prediction of human-perceived features of street scenes [21, 22], the automated estimation of demographic variables across the United States [23] and Great Britain [24], or the beautification of urban images through the generation of prototypes [25]. Turning to transportation research, however, computer vision has focused mostly on traffic control and surveillance [26], and automatic detection and collision prevention [27, 28] for autonomous vehicles. Outside scene analysis, the Deep Learning paradigm has been exploited mostly on motor traffic [29–33] , so far leaving aside its potential to tackle pedestrian safety.

Here, we address the complexities of vehicle-to-pedestrian interaction combining the structural (scene elements) and perceptual (scene composition) aspects of the problem. Overall, the contributions of the present work can be summarized as follows:

1. Creating a dataset of urban street-level images labelled according to accidentality, based on open data municipal accident records.
2. Developing a deep learning architecture, adapted from Deep Residual Networks (ResNet), for hazard index estimation in urban images, that works for both pedestrian and vehicle accidents, and is capable of producing city-wide hazard level landscapes at an unprecedented resolution of one value every 15-20 meters.
3. Proposing a set of interpretability analyses to extract human meaning from the outputs of the classification, through customized implementations of Pyramid Scene Parsing networks (PSPNet), Gradient-weighted class activation mapping (GradCam++), radar plots, and a new measure of scene disorder.
4. Designing a greedy heuristic to propose realistic urban interventions, based on scene segmentation, class activation mapping and k-nn algorithm, which constitutes an informed guide for planners to pedestrian safety improvements.

Taken together, these points constitute a novel and comprehensive deep

learning pipeline for estimating vehicle and pedestrian hazard in urban scenes, and recommending feasible physical improvements to make those same scenes safer. The building blocks of the pipeline are tailored variants of different state-of-the-art deep learning/machine learning models and techniques (Deep Residual Networks (ResNet), Pyramid Scene Parsing network (PSPNet), Gradient-weighted class activation mapping (GradCam++)).

The remainder of the paper is organized as follows: in Section 2, data (collection, processing techniques and labelling) and methods (pipeline components) are described in detail; then, in Section 3, the results on the hazard index and landscape, its connection to scene composition, and intervention heuristic are presented and discussed. Finally, Section 4 summarizes the work and discusses possible gaps and lines of development.

2. Materials and Methods

In this Section we provide the details about the datasets and Deep Learning methods that are used throughout the work. For an introduction to the Deep Learning paradigm, with a focus on transportation systems, we refer to Wang *et. al.* [32].

2.1. Dataset collection and curation

To feed the proposed framework, we use two types of real urban data: historical accident statistics and street-level urban imagery.

In the case of Madrid and Barcelona, historical accident records for the years 2010-2018 are available from the open data portals of the respective municipal governments [34, 35]. For San Francisco, data was available from 2015-2017 and it was filtered from the University of California, Berkeley's Transport Injury Mapping System (TIMS) of California traffic accidents [36]. In total, the Barcelona dataset was made up of 86,414 accidents, 10,240 being pedestrian and 76,174 being vehicle accidents. The Madrid dataset had 76,026 accidents (12,533 pedestrian, 63,492 vehicle). In San Francisco, the dataset was made up of 15,492 accidents (3331 pedestrian, 12,161 vehicle). All data points are geolocated with their corresponding GPS coordinates. Besides location, due the detonating causes may be different, we distinguish between accidents where a vehicle and a pedestrian were involved (simply 'pedestrian', or P, onwards), from vehicle-to-vehicle accidents (simply 'vehicle', or V, onwards). The spatial distribution of empirical accident data for both vehicles and pedestrians can be seen in the SI Figure S1.

Street-level imagery was extracted from two data sources. The Google StreetView (GSV) [37] API was used for Barcelona and Madrid. In these dataset, images are, on average, 15 meters away from each other. As we wanted to capture the view of the driver, we limited our queries to images facing directly down the direction of traffic of the street. The result of this process was a comprehensive and homogeneous set of images for both cities.

For the city of San Francisco, images were provided by Mapillary [38], a crowd-sourced alternative to GSV. With Mapillary, all user-uploaded images are available under the CC-BY-SA license. As images are uploaded by private individuals working with different equipment, different setup, different light conditions, different vehicles, and without central coordination, several distinct challenges were presented by this dataset. Firstly, for each point provided, usually a single image was available. Occasionally, this image did not fit our criteria of facing down the direction of traffic, and had to be discarded. Secondly, data was only available from a smaller part of the city, corresponding to the area covered by the Mapillary contributors. The part of San Francisco available in the dataset, consisting mostly of high-traffic streets, is shown in Figure S2 of the SI.

Combining data from different sources (GSV and Mapillary) allows us to test the robustness of our methods when dealing with similar, but not equally distributed, data . All the collected images, both for GSV and Mapillary, contain GPS locations in their metadata, which allows us to assign each street image a binary accident category ("safe" vs. "dangerous"). We categorize a point as "dangerous" if one or more accidents have occurred with a 50 meter radius of its location. Otherwise, the point is categorized as "safe". More details on the creation of the image dataset can be found in Section S1 of the SI, along with a more extended discussion of the trade-offs of using a radius to assign accidents to images in Section S4.

The large collection of images tagged according to accident category was divided in 6 different datasets, resulting from the combination of the three targeted cities and two accident types (V and P). The characteristics of each dataset (number of images per dataset and category) are detailed in Table 1.

Notice that the San Francisco datasets are much smaller than Barcelona and Madrid datasets. For the 6 datasets, data was randomly split into train and test sets, containing 90% and 10% of the images respectively.

2.2. Hazard index estimation with Deep Learning

		Vehicle (V)		**Pedestrian** (P)	
City	Total	Accident	No accident	Accident	No accident
Barcelona	177645	61.8%	38.2%	48.1%	51.9%
Madrid	704950	48.3%	51.7%	29.1%	70.9%
San Francisco	162530	35.7%	64.3%	17.4%	82.6%

Table 1: Image dataset properties. Comparing the relative proportion of points with and without accidents across the various cities. In all 3 cities, there is a higher proportion of points with vehicle-to-vehicle accidents than vehicle-to-pedestrian accidents. Relatively less accident points in San Francisco reflects the smaller amount of accident data for that city.

A variety of Deep Learning architectures have shown to be remarkably effective for many computer vision tasks [39, 40]. In this work we use a Residual Neural Network (ResNet) [41], a particular architecture of Convolutional Neural Network (CNN), to estimate the *hazard index* (H) in new, unseen images. The main characteristic of ResNets is the implementation of "shortcut connections" that skip blocks of convolutional layers, allowing the network to learn residual mappings between layers that mitigate the vanishing gradients problem. For this critical step, all of the elements used were created from scratch – training and test datasets, weight learning stage, etc. – as is detailed in the following.

We define our *hazard index* (H) as the probability that a target image is classified as 'dangerous' by the ResNet. For this objective, we train the ResNet to first classify images between the two defined accident categories: 'dangerous' and 'safe'. For each street-level image, the classifier delivers a value H in the range of $[0, 1]$. When $H \approx 1$, the point where the image was taken is considered as dangerous. On the contrary, when $H \approx 0$, the corresponding point is considered as safe. The hazard index is defined as the output of the Softmax activation function (between 0 and 1) of the last layer of the classifier architecture:

$$H = \frac{e^{z_i}}{\sum_{j=0}^{K} e^{z_j}} \tag{1}$$

where z is the output logits of the last ResNet layer, i is the index of 'dangerous' class and K is the number of classes. H can be interpreted as the probability that the point related to a given image is hazardous.

To successfully train our ResNet architecture for the required classifica-

tion task, we start with a pre-trained network that considers the Imagenet dataset [42], and then, via 'Transfer learning' techniques, we fine-tune the network using our data. At this stage, we remove the connections from the last layer of the pre-trained ResNet model, replace it with a new layer with two outputs (categories *dangerous* and *safe*), and randomly initialize the layer's weights. We re-trained (fine-tuned) this last layer, leaving the rest of the CNN static. To compensate for class imbalance during training stage, class weights were adjusted in the objective cross entropy loss function according to inverse class frequency:

$$w_i = \frac{1}{ln(c + r_i)} \tag{2}$$

with w_i as the weight assigned to each class, c is a parameter to control the range of the valid values, and r_i is the ratio of the number of samples from each class respect the total of samples, and then

$$Loss = \frac{1}{N} \sum_{i=1}^{N} w_i \cdot (y_i \cdot log(\hat{y_i}) + (1 - y_i) \cdot log(1 - \hat{y_i})) \tag{3}$$

where N is the number of samples, and y_i and $\hat{y_i}$ are the true label and the prediction for i class, respectively. In accordance with the defined accident types (V and P), we train our ResNet to estimate two subtypes of hazard index: H_V and H_P, corresponding to the hazard indices for vehicle-to-vehicle and vehicle-to-pedestrian accidents, respectively. Therefore, we end up training 6 models in total, two per city.

2.3. Hazard index interpretability

One of the main shortcomings of Deep Learning techniques is (the lack of) interpretability. Certainly, deep neural networks can provide a high level of discriminative power, but at the cost of introducing many model variables, which eventually hinders the interpretability of their black-box representations [43]. This difficulty is especially pertinent in our case: improving pedestrian safety sometimes demands changes in the urban landscape, the question being *which* changes are pertinent. Here, we address this by using two different interpretability techniques. The first, scene disorder, is used to assess image complexity and the second, Class Activation Mapping (CAM), to assess which areas are more informative for the estimation of the hazard

index. In particular, CAM methods have been recently shown to be successful for interpretability tasks in several fields [44–47], including medicine [48].

2.3.1. Urban scene segmentation and scene disorder

First, in order to identify what objects are in the scene, and where they are positioned, we use urban scene segmentation. The goal of the semantic image segmentation task is to assign a category label to each pixel of an image. Segmentation provides a comprehensive breakdown of the physical elements visible in the scene. It predicts the label, location and mask for each object. For this task, we used a high-performance method called Pyramid Scene Parsing Network (PSPNet) [49] architecture, pre-trained with the Cityscapes dataset [50]. PSPNet is a state-of-the-art deep learning model that exploits the capability of both global and local context information aggregation through several pyramid pooling layers. It has shown outstanding performance on several semantic segmentation benchmarks. Cityscapes is a real-world, vehicle-egocentric dataset for semantic urban scene understanding which contains 25K pixel-annotated images taken in different weather conditions. Images in Cityscapes are annotated with 30 urban object categories, but we used a subset of those (19) in our image repository segmentation – those that are common and relevant in driver-perspective scenes (e.g. "car", "road", "sidewalk", "person", "traffic light", etc.; see right-most labels in Figure 4).

On top of the image segmentation outcome, we propose a measure of scene disorder inspired by the gray-tone spatial-dependence matrix [51], also known as Gray-level co-occurrence matrix (GLCM), which captures the amount of transitions between adjacent pixels labelled with different categories. It is known that complex images (related to scene disorder) may cause a division of attention [52–55] and, as a consequence, reduce attention towards objects that are relevant to urban hazard.

Originally, GLCM characterizes the texture of an image by calculating how often pairs of pixels with specific values are adjacent in a specified spatial configuration. In our measure of scene disorder, the frequency of pair of pixels of different values is calculated over the segmented image, where the value of a pixel corresponds to an urban object category, instead of a gray intensity like the usual GLCM. We perform the calculation as follows:

$$SD = \sum_{i=0}^{m} \sum_{j=0}^{n} \delta \left[I(i,j) \neq I(i+\Delta i, j+\Delta j) \right] \tag{4}$$

where $\delta[x]$ is the Kronecker delta, valued 1 if the condition x is met, and 0 otherwise; and Δi and Δj represent an offset of 1, to compute the amount of pixel value transitions in two directions (right and below). With this definition, the measure SD is incremented by 1 for every pair of neighboring pixels that have differing values. Examples of scene disorder measures can be seen in Figure 2.

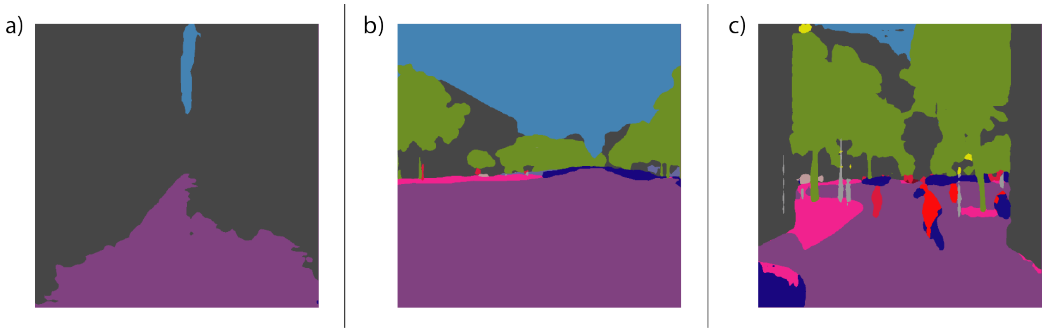

Figure 2: **Illustrating the concept of scene disorder.** Segmented images with low $SD = 0.15$ scene disorder (a); mild $SD = 0.39$ scene disorder (b); and high scene disorder $SD = 0.81$ (c).

2.3.2. Interpretability through Activation Mapping

Moving on to the second step of our interpretability process, Class Activation Mapping (CAM) [56] and related techniques (e.g. gradient-weighted class activation mapping (GradCAM++) [57, 58]) are used to interpret, visually, the patterns of images that are informative of a specific image category [43, 59], meaning, in our case, the regions that have influenced the most about the decision taken by the classifier for a certain class, in our case, classifying an image as 'dangerous'.

GradCAM++ was used to identify the regions of the image that are dangerous. Given an input image and a our trained CNN model, GradCAM++ generates a localization map by the use of the gradient information of the specific target class 'dangerous' to compute the target class weights of each

feature map of the last convolutional layer of the CNN before the final classification. The final localization map is synthesized from the aggregated sum of these target class weights. Generating a GradCAM++ map for the 'dangerous' class helps to visually identify the specific patterns and objects learned by the CNN in order to differentiate between 'safe' and 'dangerous' scenes. Since the images have been fully segmented, we can retrieve the objects that overlap with the dangerous regions. Analyzing frequencies, we can recover what object categories are more relevant to determine H_V or H_P. Figure 4 shows one example per city in the first column and visualizations of the described techniques in the other columns. In particular, second and third column display H_P and H_V, respectively, with the corresponding Class Activation Map. Areas in red color are those that are more relevant to the hazard index, that is, areas that strongly contribute to increase the hazard indexes. Last column shows the automatic segmentation of the images.

2.4. A greedy heuristic to improve H

The combination of the Class Activation Mapping and image segmentation described in the previous section gives us insight into which regions and objects of a scene contribute most to its estimated hazard level. While this information is already relevant, it provides users with no concrete recommendations for structural changes to the scene that might make it safer. Accordingly, as a final step in the pipeline, we propose a strategy to exploit the large pool of images available in order to identify, for each scene, realistic and potentially low-cost physical alterations that would diminish H_P and H_V the most.

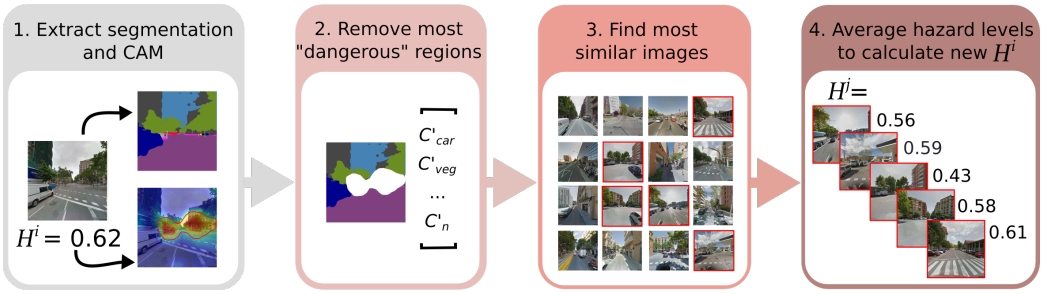

Figure 3: **Image hazard reduction flowchart.** Processing pipeline to improve the most hazardous parts of a street-level image i, comparing the new image with similar partner images j, and arriving at a new H_P and H_V for the original image.

To this end, we take advantage of the methodologies developed in the previous steps. On the one hand, the segmentation task allows us to identify which objects among C categories are present in a given scene (and to what extent). On the other, CAM provides information regarding which regions of the scene contribute most to the estimated hazard score. With this information at hand, for every image i we build a vector of characteristics $v_i \in \mathbb{R}^C$, containing information of the relative area of category C in i. For the target scene (the one for which we intend to reduce the hazard levels), we construct an additional surrogate vector of characteristics, $\tilde{v}_i$, in which we discard those regions that contribute most to H_P, i.e. we only consider regions of i where the class activation is mild-to-low (< 0.7), see first and second blocks in Figure 3. Next, we deploy an exhaustive search to find the five mirror images j for $\tilde{v}_i$, with their respective vectors of characteristics v_j, such that their hazard index is lower:

$$\begin{aligned} &\text{argmin}_j ||\tilde{v}_i - v_j||_2 \\ &H_P^j < H_P^i \\ &H_V^j < H_V^i \end{aligned} \tag{5}$$

In other words, we seek the most similar locations in the city that have smaller H_P and H_V than i, see Fig. 3 for a schematic representation of the process. The search for mirror images is limited to structurally similar scenes (compared to the original one), in order to promote simple and feasible interventions. We emphasize that this strategy is designed to be used in tandem with human users, who will be able to judge which recommendations are realistic. The choice of five images allows for some diversity in the range of interventions recommended.

Finally, we remark that our approach is very similar to the regressive k-nearest neighbor (k-nn) algorithm [60], as opposed to a more sophisticated, Deep Learning-based mechanism for image "safe-fication" (following the concept of "beautification" in ref. [25]). These techniques lie beyond the scope of the present work.

3. Experiments and Results

3.1. Hazard index Estimation

We begin the results section by assessing how well our trained ResNet performs the required classification task for the six datasets we have defined,

considering the cities of Barcelona, Madrid, and San Francisco. Images belonging to the 'dangerous' class are defined as positive, while those belonging to the 'safe' class are defined as negative. In the training stage, the parameter c of the loss function was experimentally assigned as 1. For our results, we focus on the following measures: recall, precision and accuracy; and the indicators: FP (False positives), TP (True Positives), TN (True Negatives) and FN (False negatives). Recall refers to the fraction of samples detected as dangerous over the total number of dangerous samples in the dataset (TP over TP+FN). Precision is the fraction of the true dangerous points detected, over the number of points detected as dangerous by the ResNet (TP over TP+FP). Accuracy measures how good the system is at detecting dangerous points (TP+TN over all the samples).

As we can see in Table 2, the obtained accuracy is outstanding for all datasets, considering that the CNN training stage relies only on visual information, along with a binary tag indicating the occurrence (or not) of an accident within a 50m radius (sensitivity with respect to radii is discussed in Section S4.1 and Figure S7 of the SI). As illustrated examples of hazard index estimation, see the scores in the central columns of Figure 4.

	Recall	Prec.	Acc.	FP	TP	TN	FN
Barcelona P	0.86	0.72	0.75	17.8%	45.4%	29.8%	7%
Barcelona V	0.77	0.84	0.82	7.1%	37.9%	44.1%	10.9%
Madrid P	0.76	0.75	0.75	12.4%	37.5%	38%	12.1%
Madrid V	0.73	0.74	0.75	12%	35.2%	40.1%	12.7%
San Francisco P	0.63	0.81	0.76	6.6%	29%	47.7%	16.7%
San Francisco V	0.61	0.82	0.74	6.3%	30.1%	44.7%	18.9%

Table 2: Results of the Deep Learning approach for accident prediction, considering a 50 meters radius. Rows labelled as P and V correspond to pedestrian-to-vehicle and vehicle-to-vehicle accident dataset, respectively. Results for other radii can be seen on Table S1 of the SI.

Additionally, we compared the performance of different ResNet and other state-of-the-art architectures against the Barcelona dataset. Metrics like F1-score, area under the Precision and Recall (PR) curve, and the area under the Receiver Operating Characteristic (ROC) curve were used for comparison as well. The F1-measure provides a balance between precision and recall in a

single score:

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{6}$$

Whereas the PR curve represents the balance between the measures precision and recall through different thresholds between 0 and 1. The ROC curve plots the false positive rate versus the true positive rate through different thresholds, like the PR curve. The results presented in Table 3 show that the ResNet-v2-50 offers the highest performance for this particular image classification task.

Discerning between safe and dangerous locations in a binary fashion might be limiting in several practical scenarios, such as the prioritization of urban interventions to improve pedestrian safety. To assess to what extent we can produce finer results, we have also implemented the method in [61] to learn an ordinal regressor. In this case, the Barcelona pedestrian dataset was divided in four rating classes: *no-danger*, *mild-danger*, *danger* and *high-danger*. Images tagged as 'no-danger', correspond those images where no accidents were observed. Images in the class 'mild-danger' had one accident nearby, images in class 'danger' have between 2 and 5 accidents nearby. Finally, images belonging to class 'high-danger' have more than 5 accidents in their vicinity. The dataset proportions were approximately 85k, 34k, 40k and 17k images samples, respectively. The method in [61] relies on several binary classifiers. We used our same ResNet architecture for each of those binary classifiers. After training, we obtained a balanced accuracy of 0.47 (with a the dummy classifier accuracy of 0.25) which is comparable to the performance reported in [20] for a similar task. That is, the ResNet architecture can also provide competitive results for a finer assessment of pedestrian safety.

3.2. Urban hazard landscape

The first remarkable outcome of the described methodology (in particular, Section 2.2) is a fine-grained map of hazard indices throughout the cities under study. The Deep Learning approach, together with the short distance intervals between consecutive images, allows us to quantify the safety of all city locations at a microscopic level, i.e. every 15 meters approximately (see Figures S3 and S4 in the SI), independently of whether accidents have occurred at a given site or not.

To give a complete picture of hazard for pedestrians and vehicles, and to highlight their differences, Figure 5 shows the spatial distribution of points that were identified as very hazardous for pedestrians ($H_P \geq 0.66$), but with

Model	Acc.	Prec.	Rec.l	F1-Score	PR	ROC
VGG16 [62]	0.61	0.58	0.96	0.72	0.78	0.59
VGG19 [62]	0.68	0.73	0.62	0.67	0.77	0.68
Inception-V3 [63]	0.70	0.70	0.75	0.72	0.79	0.70
Inception-V4 [64]	0.57	0.80	0.24	0.37	0.72	0.59
Mobilenet [65]	0.62	0.77	0.39	0.52	0.74	0.63
ResNet-v1-50 [66]	0.61	0.80	0.35	0.49	0.75	0.63
ResNet-v1-101 [66]	0.59	0.56	0.99	0.71	0.78	0.57
ResNet-v1-152 [66]	0.67	0.71	0.62	0.66	0.76	0.67
ResNet-v2-50 [41]	**0.75**	0.72	0.87	**0.78**	**0.82**	**0.74**
ResNet-v2-101 [41]	0.72	0.75	0.70	0.72	0.80	0.72
ResNet-v2-152 [41]	0.72	0.74	0.72	0.73	0.80	0.72

Table 3: Results of the Deep Learning approach for accident prediction, considering different classification architectures.

low-to-moderate hazard for vehicles ($H_V < 0.66$), and vice-versa. As can be seen, in both Madrid and Barcelona, areas of high hazard for pedestrians alone are highly concentrated in the denser, older city centers. High levels of vehicle hazard tend to be distributed around arterial roads, as well as some distinct neighborhoods (e.g. Sant Martí-Poble Nou, middle right corner in Barcelona). San Francisco presents an interesting case in which the two spatial distributions are nearly homogeneous. This can likely be explained by the bias towards residential, medium-density areas in our image coverage for the city (see Materials and Methods for further discussion). Notably, we lacked image coverage in high-density downtown San Francisco, as well as peripheral low-density districts. With the inclusion of such zones, it is possible that clearer spatial patterns would emerge, although they might be distinct from those of denser European cities like Barcelona and Madrid [67]. Nevertheless, it should be noted that competitive levels of precision and accuracy were still achieved in San Francisco, indicating that our method is robust to relatively homogeneous training data. Furthermore, it shows that the classifier need not only be applied to comprehensive collections of images from an entire city, but can function well on sufficiently rich, spatially homogeneous samples of images. Separate visualizations for pedestrian and vehicle hazards are available in the SI, Figure S3.

Worth highlighting, there has been no previous attempt to associate a given street image with traffic hazard levels –unlike other urban attributes

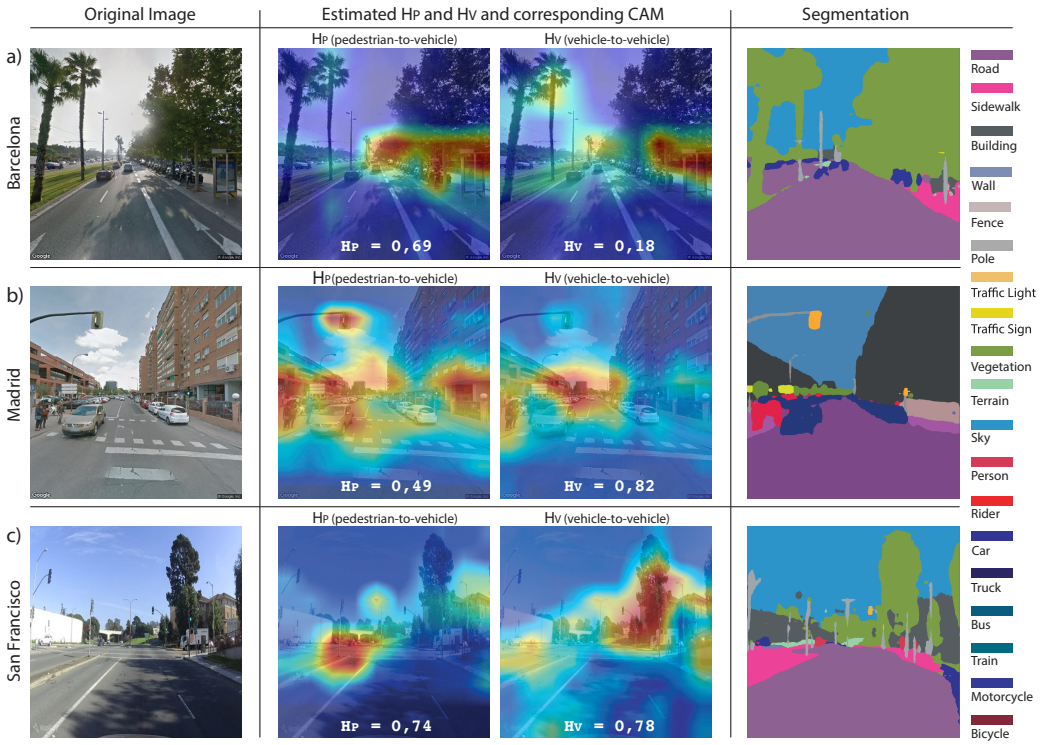

Figure 4: **Deep Learning approach: classification, segmentation and interpretability.** The figures display image examples from Barcelona, San Francisco and Madrid, one location per row. First column shows the original street view image. Second and third columns correspond to the obtained CAM for pedestrian and vehicle datasets, respectively. The last column corresponds to the outcome of the segmentation task. The example in Barcelona location (top row) is classified as dangerous for pedestrians (note the score in each picture), but safe for vehicles. The second example, corresponding to a Madrid location, is classified as dangerous for vehicles, but safe for pedestrians. Finally, the third example, corresponds to a San Francisco location. Notice that, in this last case, the location is dangerous for both pedestrian and vehicle, but the CAM highlights different regions: areas increasing the hazard for pedestrians may not coincide with those increasing hazard for vehicles. Images courtesy of Google, Inc. and Mapillary.

(e.g. beauty [17, 68], or security [21]). Here, we do so under the assumption that street-level imagery is a good proxy for both the structural and perceptual complexity of the city landscape. Typically, traffic-related risk is either aggregated to the macro-level (neighborhoods, census tracts, even counties)[7, 69, 70], or painstakingly micro-tailored to very specific settings

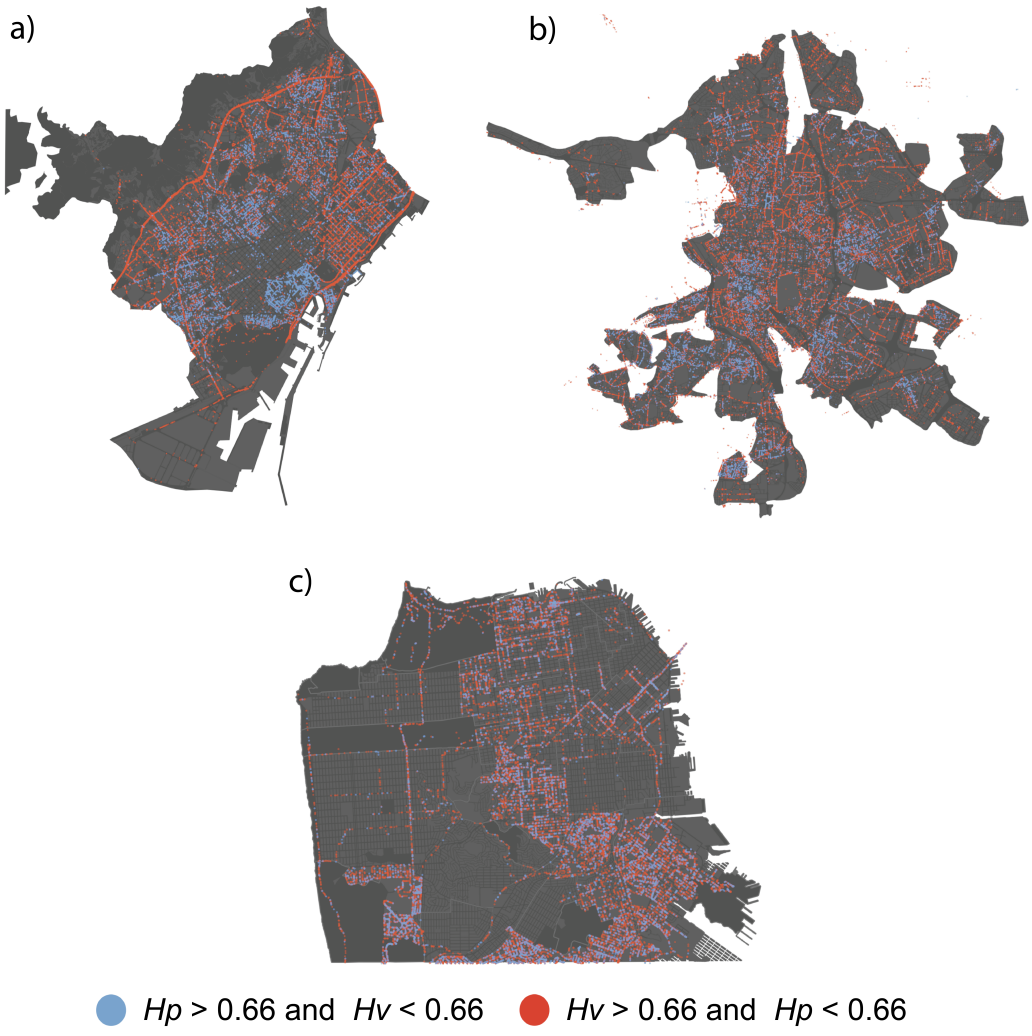

Figure 5: **Spatial distribution of hazard index**. Distribution of high-hazard points for pedestrians and vehicles across all three cities of study. Points displayed are those for which hazard is high for pedestrians (vehicles) but not for vehicles (pedestrians).

(e.g. considering only zebra-crossings [71]). However, initiatives like Vision Zero, involving governments and organizations worldwide, demand new streams of data and methodologies that help address the street safety challenge at the finest level *and* at scale. This is achieved here combining images and accident data.

3.3. Mapping safety to scene composition

The second (segmentation) and third (Class Activation Mapping, CAM) processing steps complete the data analysis pipeline, linking hazard indices, H_P and H_V, to specific objects found in street-level images. In practice, such link is established combining the information in the central and right columns of Figure 4. Mapping each pixel label (e.g. "road", "sidewalk", etc.) to its corresponding activation level (heatmap in central columns of Figure 4) provides a quantification of the contribution of that pixel to the overall hazard score of the image. Thus, at the city level, we can obtain a global perspective of the categories that most contribute to the hazard index.

Figure 6 (panels a and b) illustrates this for the central area of Barcelona. These radar plots show the level of object fixation of the CAM model for pedestrians (a) and cars (b). In both cases, the blue line represents safe scenes ($H < 0.33$), while dangerous ones ($H > 0.66$) are shown in red. Specifically, we plot the ratio between the amount of CAM fixation on a given category (in safe and dangerous scenes), with respect to the CAM fixation on that category across all the images of the dataset. Thus, values below 1 in the radar plots are underrepresented, while those above 1 are overrepresented. We would like to highlight that we have restricted the analysis to the city center, to avoid an exaggeration of the presence of natural elements (vegetation and sky) in low accident risk images. Remarkably, the presence of people in a scene is correlated to a dangerous classification for both vehicle-to-pedestrian and vehicle-to-vehicle predictions. Low buildings and/or wide streets (tantamount to a clear vision of the sky) correlate to safer scenes for pedestrians, whereas the presence of buildings implies a safer environment for vehicles. Also, the absence of vegetation, such as trees, could be contributing to a safe classification for vehicles.

Radar plots for Madrid (see SI, Fig. S5) show high resemblance to the Barcelona ones, while those for San Francisco (Fig. S6) show completely different patterns: for pedestrians, the presence of sidewalks –and not people– is identified as the strongest driver for high H_P. Again, the distinct layouts

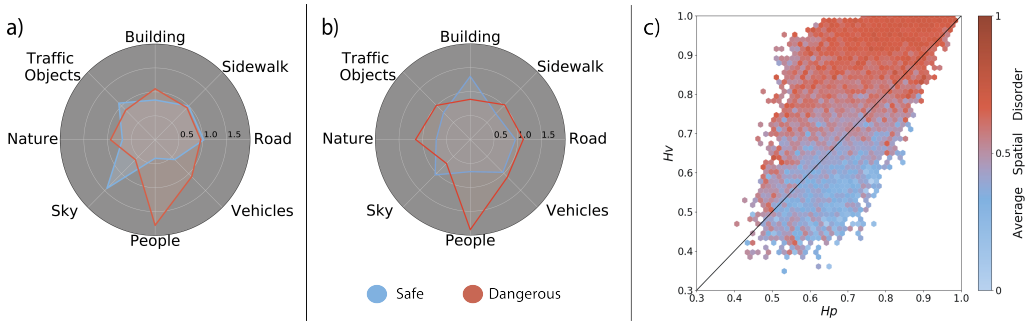

Figure 6: **Hazard level interpretability. Top:** Radar plots showing the level of object fixation of the CAM model for pedestrian (a) and cars (b). For both, the blue area corresponds to images classified as safe ($H < 0.33$), while scenes classified as dangerous ($H > 0.66$) are mapped on the plot as red. To build these radars, each individual image is mapped to the radar categories (a relevant subset of those detected by the segmentation task), and the average of such mappings is shown. **(c)** The plot shows the triple relationship between H_P, H_V and the color-coded level of disorder (adapted from [51]) –which increases towards warmer colors as the levels of hazard increase. The plot corresponds to Barcelona.

and walking habits of European and North American cities may be directly related to these emergent patterns.

Moving further, we can relate hazard levels to the scene complexity. While the radar plots show interesting information, they are blind to specific scene compositions in urban scenes, i.e. whether categories appear in a clustered or fragmented way. To grasp this information, we quantify scene disorder (SD) as defined in Equation 4, see Methods above. Figure 6c shows an hexbin scatter plot of hazard indices (H_V against H_P), with a color-coded third dimension that corresponds to scene disorder, normalized in the range $[0, 1]$. A first observation is that H_P and H_V are positively correlated. More interestingly, it is clear that more complex scenes (warmer colors) correspond to more dangerous ones. In Figure S5c of the SI, an even clearer trend is shown for Madrid. On the other hand, the level of disorder in San Francisco scenes is high when $H_P \approx H_V \approx 1$, but not clearly related to either H_P or H_V for the rest of values, see Figure S6c. All in all, the connection between image complexity and hazard (especially for vehicles) suggests that more research is needed in this direction. While certain distractions are very explicit (e.g. attending the mobile phone), the perils of scene disorder are subtle and implicit (in the sense that they are not obvious on visual

inspection).

3.4. An informed guide to pedestrian safety improvements

A precipitate analysis of Figure 6 may render unfeasible interventions: substitution of built space with larger green areas, building height reduction, or street widening would suffice to improve pedestrian safety, but they do not represent a realistic approach. Instead, we resort on the greedy strategy developed in Section 2.4 to propose interventions conducive to scene alterations that diminish H_P and H_V most.

Figure 7a shows the results of the application of this optimization to the set of images in Barcelona (Figure S8 in SI for Madrid and San Francisco). In some occasions the hazard index cannot be reduced (points near the $(1, 1)$ coordinate). And yet, many locations present a potential to decrease the hazard levels, even observing, for some scenarios, extreme improvements (points near the $(0, 0)$ coordinate). The grey intensity in Fig. 7a reflects the density of observations in that area. To provide a baseline for comparison, panel b shows alternative results considering a dummy k-nn regressor, that does not take our hazard index into account. Ratios larger than 1 indicate an increase in H_V or H_P, and ratios lower than 1 indicate a decrease. The average in both dimensions is close to zero, evidencing that, with a dummy regressor, we have no guarantee of reducing either pedestrian or vehicle hazard. Figure 7c shows a selection of two targets and their most similar mirror image, illustrating some common interventions proposed by the heuristic (more examples, for the three cities under study, can be found in Figure S9 of the SI). Visually, all of them seem to point at simplifications of the original image – mostly removing objects on sidewalks.

Finally, Figure 7d provides a visual overview of the most frequent interventions predicted by our optimization scheme, in the case of Barcelona. The color of the link connecting two categories expresses the source of that link. The most notable changes point –perhaps unsurprisingly– to the need to reconfigure urban scenes towards greener and wider spaces: indeed, both categories 'road' and 'building' contribute largely to 'nature', while the latter does the same towards 'sky'. Madrid presents an almost identical trend, while San Francisco shows a less clear pattern (although the relevance of 'nature' and 'sky' is still clear). Both diagrams are available in the SI, Figure S10. Overall, the estimations and insights from the panels in Fig. 7 can provide initial indications to urban planners about achieving potential reductions of a local hazard score, both in terms of which items could be removed

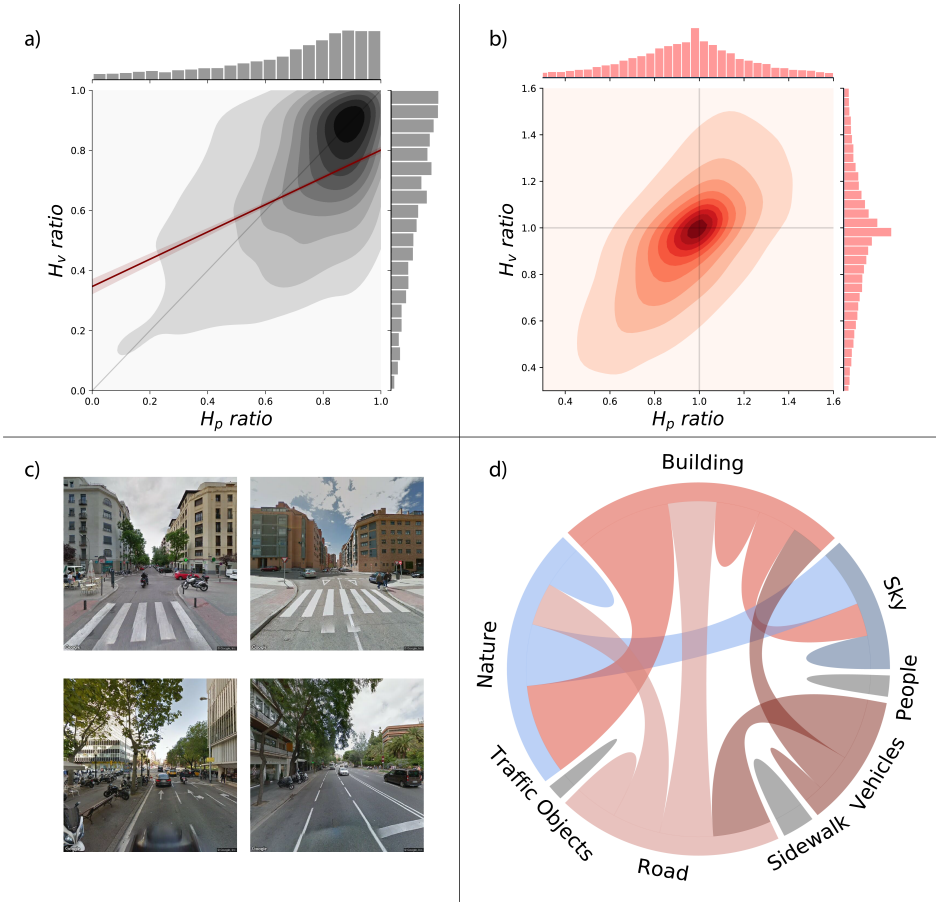

Figure 7: **Hazard reduction: results.** **(a)** Expected improvement for pedestrian and vehicle hazards, with respect to their original values. The horizontal axis corresponds to the ratio between the improved and the original pedestrian hazard index, $\tilde{H}_P/H_P$; while the vertical axis represents the equivalent ratio for vehicles, $\tilde{H}_V/H_V$. Grey intensity represents the density of observations in a given area of the plot. **(b)** Expected improvement of a dummy k-nn algorithm that only considers similarity between images. This can be regarded as a baseline for results in panel (a) **(c)** Examples of original and mirror images in Barcelona and Madrid. **(d)** Chord diagram representing an aggregate overview of proposed interventions in Barcelona. The most notable outcome from the diagram is the propensity to reduce the space allotted to roads and buildings, exchanging it emptier, greener scenes.

or relocated.

4. Discussion

As cities become increasingly populated, the interactions among pedestrians and motorized vehicles become permanent. This translates into a growing number of pedestrian-vehicle accidents. Complementary to the efforts by urban planners, public authorities and sensor technology designers, we present here an automated scheme that exploits a wide range of Computer Vision methods (classification, segmentation and interpretability techniques) to reduce traffic-related fatalities. The proposed processing pipeline, conveniently fed with rich sources of open data, renders an holistic characterization of a city's hazard landscape, capturing the physical (scene structure) and perceptual (scene complexity) characteristics from a car driver's point of view. Beyond its informative value, the hazard landscape provides actionable insights to planners.

The main strength of our proposal lies in its simplicity, and its potentially universal applicability out of a comprehensive street image collection and a rich accident dataset. Even crowd-sourced imagery, which is unavoidably diverse and often sparse, provides a solid starting point to quantify safety at a below-segment level. A global, automated, data-driven endeavour towards improving pedestrian safety is not out of reach, considering the advances in cities' public data portals, and the wide coverage of proprietary services like Google Street View or open initiatives like Mapillary.

Our approach opens a promising line of development. The hazard landscape is defined at an unprecedented, sub-segment resolution level –roughly a hazard score every 15 meters– through an automated and scalable classification process. This is well beyond macroscale approaches (e.g. crash hotspots), and extends the emphasis on intersections [14]. Such fine-grained map adds a valuable geoinformation layer to those already in use –traffic and pollution levels [72], land and underground transportation systems, crime, etc.– enabling better route design: safe paths, along with clean, beautiful, or shortest ones.

Additionally, segmentation and interpretability methods unveil the relationship between potential danger and specific objects in urban scenes. What's more, the disposition of those objects is related to hazard indices, adding a perceptual-attentional link to other possible concomitant variables that affect vehicle and pedestrian safety. Along this line, our work can be used in conjunction with other similar pipelines, such as [20], which automates road safety assessment in terms of infrastructure and estimates road

attributes, or may contribute to more focused analysis, relating what a person pays attention to while driving [73]. Additionally, further information such as temporal accident data, or factors known to influence accident rate (e.g. weather, lighting condition, distraction, asphalt conditions, road signaling) could be included by using, for instance, a multi-branch convolutional neural network, to obtain a richer prediction model.

On the other hand, the step from descriptive (hazard landscape) to actionable insights paves the way to automatized, computer-aided prioritization of urban interventions. The proposed heuristic towards safety improvements can serve as a novel tool for planners and policy makers, and might trigger the development of more sophisticated approaches such as the use of Generative Adversarial Networks to produce virtual, plausible alternatives to target scenes (seeking for instance "safe-fication", instead of "beautification" [25]). These techniques could be complemented with intervention cost quantification, considering as well cost-safety gain trade-offs.

Acknowledgements

All authors acknowledge financial support from the Dirección General de Tráfico (Spain), Project No. SPIP2017-02263, as well as TIN2015-66951-C2-2-R and RTI2018-095232-B- C22 grants from the Spanish Ministry of Science, Innovation and Universities (FEDER funds). CB and DR acknowledge as well the support of a doctoral grant from the Universitat Oberta de Catalunya (UOC). CB, DM and AL acknowledge the NVIDIA Hardware grant program. Street network data copyrighted OpenStreetMap contributors and available from https://www.openstreetmap.org.

References

[1] M. De Domenico, A. Solé-Ribalta, S. Gómez, A. Arenas, Navigability of interconnected networks under random failures, Proceedings of the National Academy of Sciences 111 (2014) 8351–8356.

[2] S. Jiang, Y. Yang, S. Gupta, D. Veneziano, S. Athavale, M. C. González, The timegeo modeling framework for urban mobility without travel surveys, Proceedings of the National Academy of Sciences 113 (2016) E5370–E5378.

[3] S. Abbar, T. Zanouda, J. Borge-Holthoefer, Structural robustness and service reachability in urban settings, Data Mining and Knowledge Discovery 32 (2018) 830–847.

[4] R. Gakenheimer, Urban mobility in the developing world, Transportation Research Part A: Policy and Practice 33 (1999) 671 – 689.

[5] R. Cervero, M. Duncan, Walking, bicycling, and urban landscapes: Evidence from the san francisco bay area, American Journal of Public Health 93 (2003) 1478–1483. PMID: 12948966.

[6] National Highway Traffic Safety Administration, Fatality analysis reporting system (fars) encyclopedia, `https://www-fars.nhtsa.dot.gov/Main/index.aspx`, 2018. Accessed: 2019-06-27.

[7] S. Ukkusuri, L. F. Miranda-Moreno, G. Ramadurai, J. Isa-Tavarez, The role of built environment on pedestrian crash frequency, Safety Science 50 (2012) 1141–1151.

[8] S. M. Rifaat, R. Tay, A. De Barros, Effect of street pattern on the severity of crashes involving vulnerable road users, Accident Analysis & Prevention 43 (2011) 276–283.

[9] M. Moeinaddini, Z. Asadi-Shekari, M. Z. Shah, The relationship between urban street networks and the number of transport fatalities at the city level, Safety Science 62 (2014) 114–120.

[10] G. Mecredy, I. Janssen, W. Pickett, Neighbourhood street connectivity and injury in youth: a national study of built environments in canada, Injury Prevention 18 (2012) 81–87.

[11] T. Fu, W. Hu, L. Miranda-Moreno, N. Saunier, Investigating secondary pedestrian-vehicle interactions at non-signalized intersections using vision-based trajectory data, Transportation Research Part C: Emerging Technologies 105 (2019) 222–240.

[12] J. Nasar, P. Hecht, R. Wener, Mobile telephones, distracted attention, and pedestrian safety, Accident analysis & prevention 40 (2008) 69–75.

[13] K. K. Mukoko, S. S. Pulugurtha, Examining the influence of network, land use, and demographic characteristics to estimate the number of bicycle-vehicle crashes on urban roads, IATSS Research (2019).

[14] Y. Hu, Y. Zhang, K. S. Shelton, Where are the dangerous intersections for pedestrians and cyclists: A colocation-based approach, Transportation Research Part C: Emerging Technologies 95 (2018) 431–441.

[15] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, A. Oliva, Learning deep features for scene recognition using places database, in: Advances in neural information processing systems, pp. 487–495.

[16] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, A. Torralba, Places: A 10 million image database for scene recognition, IEEE transactions on pattern analysis and machine intelligence 40 (2017) 1452–1464.

[17] N. Naik, S. D. Kominers, R. Raskar, E. L. Glaeser, C. A. Hidalgo, Computer vision uncovers predictors of physical urban change, Proceedings of the National Academy of Sciences 114 (2017) 7571–7576.

[18] A. Albert, J. Kaur, M. C. Gonzalez, Using convolutional networks and satellite imagery to identify patterns in urban environments at a large scale, in: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, ACM, pp. 1357–1366.

[19] I. Seiferling, N. Naik, C. Ratti, R. Proulx, Green streets- quantifying and mapping urban trees with street-level imagery and computer vision, Landscape and Urban Planning 165 (2017) 93–101.

[20] W. Song, S. Workman, A. Hadzic, X. Zhang, E. Green, M. Chen, R. Souleyrette, N. Jacobs, Farsa: Fully automated roadway safety assessment, in: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, pp. 521–529.

[21] N. Naik, J. Philipoom, R. Raskar, C. Hidalgo, Streetscore-predicting the perceived safety of one million streetscapes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 779–785.

[22] L. Liu, E. A. Silva, C. Wu, H. Wang, A machine learning-based method for the large-scale evaluation of the qualities of the urban environment, Computers, Environment and Urban Systems 65 (2017) 113–125.

[23] T. Gebru, J. Krause, Y. Wang, D. Chen, J. Deng, E. L. Aiden, L. Fei-Fei, Using deep learning and google street view to estimate the demographic

makeup of neighborhoods across the united states, Proceedings of the National Academy of Sciences 114 (2017) 13108–13113.

[24] E. Suel, J. W. Polak, J. E. Bennett, M. Ezzati, Measuring social, environmental and health inequalities using deep learning and street imagery, Scientific Reports 9 (2019) 6229.

[25] T. Kauer, S. Joglekar, M. Redi, L. M. Aiello, D. Quercia, Mapping and visualizing deep-learning urban beautification, IEEE Computer Graphics and Applications 38 (2018) 70–83.

[26] Z. M. Fadlullah, F. Tang, B. Mao, N. Kato, O. Akashi, T. Inoue, K. Mizutani, State-of-the-art deep learning: Evolving machine intelligence toward tomorrow's intelligent network traffic control systems, IEEE Communications Surveys & Tutorials 19 (2017) 2432–2455.

[27] L. Zhang, L. Lin, X. Liang, K. He, Is faster r-cnn doing well for pedestrian detection?, in: European Conference on Computer Vision, Springer, pp. 443–457.

[28] S. Zhang, R. Benenson, M. Omran, J. Hosang, B. Schiele, How far are we from solving pedestrian detection?, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1259–1267.

[29] N. G. Polson, V. O. Sokolov, Deep learning for short-term traffic flow prediction, Transportation Research Part C: Emerging Technologies 79 (2017) 1–17.

[30] Y. Wu, H. Tan, L. Qin, B. Ran, Z. Jiang, A hybrid deep learning based traffic flow prediction method and its understanding, Transportation Research Part C: Emerging Technologies 90 (2018) 166–180.

[31] Z. Zhang, Q. He, J. Gao, M. Ni, A deep learning approach for detecting traffic accidents from social media data, Transportation Research Part C: Emerging Technologies 86 (2018) 580–596.

[32] Y. Wang, D. Zhang, Y. Liu, B. Dai, L. H. Lee, Enhancing transportation systems via deep learning: A survey, Transportation Research Part C: Emerging Technologies 99 (2019) 144–163.

[33] Z. Zhang, M. Li, X. Lin, Y. Wang, F. He, Multistep speed prediction on traffic networks: A deep learning approach considering spatio-temporal dependencies, Transportation Research Part C: Emerging Technologies 105 (2019) 297–322.

[34] Ayuntamiento de Madrid, Portal de datos abiertos del ayuntamiento de madrid, https://datos.madrid.es/portal/site/egob/, 2019. Accessed: 2019-04-20.

[35] Ajuntament de Barcelona, Open data bcn, https://opendata-ajuntament.barcelona.cat/en/, 2019. Accessed: 2019-04-20.

[36] Safe Transportation Research and Education Center, University of California, Berkeley, Transportation injury mapping system (tims), 2019. Accessed: 2019-06-27.

[37] D. Anguelov, C. Dulong, D. Filip, C. Frueh, S. Lafon, R. Lyon, A. Ogale, L. Vincent, J. Weaver, Google street view: Capturing the world at street level, Computer 43 (2010) 32–38.

[38] Mapillary contributors, Mapillary - Street-level imagery, powered by collaboration and computer vision , https://www.mapillary.com/app, 2019.

[39] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436.

[40] J. Schmidhuber, Deep learning in neural networks: An overview, Neural networks 61 (2015) 85–117.

[41] K. He, X. Zhang, S. Ren, J. Sun, Identity mappings in deep residual networks, in: European conference on computer vision, Springer, pp. 630–645.

[42] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, pp. 1097–1105.

[43] A. Adadi, M. Berrada, Peeking inside the black-box: A survey on explainable artificial intelligence (xai), IEEE Access 6 (2018) 52138–52160.

[44] H. Fukui, T. Hirakawa, T. Yamashita, H. Fujiyoshi, Attention branch network: Learning of attention mechanism for visual explanation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 10705–10714.

[45] J. Wagner, J. M. Kohler, T. Gindele, L. Hetzel, J. T. Wiedemer, S. Behnke, Interpretable and fine-grained visual explanations for convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9097–9107.

[46] S. Desai, H. G. Ramaswamy, Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization, in: 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, pp. 972–980.

[47] B. N. Patro, M. Lunayach, S. Patel, V. P. Namboodiri, U-cam: Visual explanation using uncertainty based class activation maps, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 7444–7453.

[48] Z. Wang, J. Yang, Diabetic retinopathy detection via deep convolutional networks for discriminative localization and visual explanation, arXiv preprint arXiv:1703.10757 (2017).

[49] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2881–2890.

[50] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, The cityscapes dataset for semantic urban scene understanding, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3213–3223.

[51] R. M. Haralick, K. Shanmugam, I. H. Dinstein, Textural features for image classification, IEEE Transactions on Systems, Man, and Cybernetics (1973) 610–621.

[52] N. Moray, Attention in dichotic listening: Affective cues and the influence of instructions, Quarterly journal of experimental psychology 11 (1959) 56–60.

[53] D. Kahneman, Attention and effort, volume 1063, Citeseer, 1973.

[54] G. A. Alvarez, P. Cavanagh, The capacity of visual short-term memory is set both by visual information load and by number of objects, Psychological science 15 (2004) 106–111.

[55] J. E. Richards, The development of attention to simple and complex visual stimuli in infants: Behavioral and psychophysiological measures, Developmental Review 30 (2010) 203–219.

[56] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2921–2929.

[57] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626.

[58] A. Chattopadhay, A. Sarkar, P. Howlader, V. N. Balasubramanian, Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks, in: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, pp. 839–847.

[59] C. Ventura, D. Masip, A. Lapedriza, Interpreting cnn models for apparent personality trait regression, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 55–63.

[60] P. Harrington, Machine learning in action, Manning Publications Co., 2012.

[61] E. Frank, M. Hall, A simple approach to ordinal classification, in: European Conference on Machine Learning, Springer, pp. 145–156.

[62] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).

[63] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2818–2826.

[64] C. Szegedy, S. Ioffe, V. Vanhoucke, A. A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: Thirty-first AAAI conference on artificial intelligence.

[65] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: Efficient convolutional neural networks for mobile vision applications, arXiv preprint arXiv:1704.04861 (2017).

[66] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

[67] R. Louf, M. Barthelemy, A typology of street patterns, Journal of The Royal Society Interface 11 (2014) 20140924.

[68] D. Quercia, R. Schifanella, L. M. Aiello, The shortest path to happiness: Recommending beautiful, quiet, and happy routes in the city, in: Proceedings of the 25th ACM conference on Hypertext and social media, ACM, pp. 116–125.

[69] H. Huang, M. A. Abdel-Aty, A. L. Darwiche, County-level crash risk analysis in florida: Bayesian spatial modeling, Transportation Research Record 2148 (2010) 27–37.

[70] P. Chen, J. Zhou, Effects of the built environment on automobile-involved pedestrian crash frequency and risk, Journal of Transport & Health 3 (2016) 448–456.

[71] P. Olszewski, I. Buttler, W. Czajewski, P. Dabkowski, C. Kraśkiewicz, P. Szagała, A. Zielińska, Pedestrian safety assessment with video analysis, Transportation Research Procedia 14 (2016) 2044–2053.

[72] Y. Xu, S. Jiang, R. Li, J. Zhang, J. Zhao, S. Abbar, M. C. González, Unraveling environmental justice in ambient pm2. 5 exposure in beijing: A big data approach, Computers, Environment and Urban Systems 75 (2019) 12–21.

[73] A. Palazzi, D. Abati, F. Solera, R. Cucchiara, et al., Predicting the driver's focus of attention: the dr (eye) ve project, IEEE Transactions on Pattern Analysis and Machine Intelligence 41 (2018) 1720–1733.

[74] OpenStreetMap contributors, Planet dump retrieved from https://planet.osm.org , `https://www.openstreetmap.org`, 2017.

[75] G. Boeing, Osmnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks, Computers, Environment and Urban Systems 65 (2017) 126–139.