The version of record of this article, published in *Journal of Urban Technology*, is available online at the publisher's website: https://doi.org/10.1080/10630732.2026.2665592.

Cite as: Liu, P., Yap, W., Hou, Y., Lei, B., Stouffs, R., & Biljecki, F. (2026). A graph-based human-centered GeoAI for understanding street-level environment and traffic accident frequency with street view imagery. *Journal of Urban Technology*. https://doi.org/10.1080/10630732.2026.2665592

A Graph-Based Human-Centered GeoAI for Understanding Street-Level Environment and Traffic Accident Frequency with Street View Imagery

Pengyuan Liu[a,b], Winston Yap[c,f], Yujun Hou[d,f], Binyu Lei[e,f], Rudi Stouffs[f], Filip Biljecki[f,g]

[a]*Urban Analytics Subject Group, Urban Studies & Social Policy Division, University of Glasgow, United Kingdom*
[b]*Future Cities Lab Global, Singapore-ETH Centre, Singapore*
[c]*Cornell Engineering, Cornell University, United States*
[d]*Department of Geography and Planning , University of Liverpool, United Kingdom*
[e]*School of Geography, Earth and Environmental Sciences, University of Birmingham Dubai, United Arab Emirates*
[f]*Department of Architecture, National University of Singapore, Singapore*
[g]*Department of Real Estate, National University of Singapore, Singapore*

Abstract

Traffic accidents are a major global concern, highlighting the need for advanced traffic analysis and predictive techniques. The emergence of crowdsourced street view imagery (SVI) platforms has transformed public participation in collecting urban data, initiating the development of human-centered traffic analytics. This paper investigates the relationship between visual urban understanding derived from Mapillary SVI and the frequency of urban traffic accidents. We analyzed SVI's visual complexities using the Mask2Former image segmentation model and provided spatial reasoning of the urban objects (e.g., cars, buildings, trees) by estimating visual distances between those objects and the drivers using Dist-YOLOv5. These distances were encapsulated as edge weights, with urban objects and the drivers as nodes, creating human-centered graphs at each SVI location. We propose a framework that integrates a graph-based deep learning approach, GAT-LSTM, to capture the spatial-temporal dynamics of these urban objects for modeling traffic accident frequency. Our results indicate that this model outperforms a traditional machine learning method by over 70% in mean absolute percentage error (MAPE) and demonstrates superior performance compared to other deep learning-based methods. Additionally, we introduce a twostep Explainable AI (XAI) method to identify key factors associated with roads with higher traffic accident rates, thereby improving the interpretability and practicality of our research for understanding the urban safety environment.

Keywords: Graph Neural Network, GeoAI, Road Safety, Urban Sensing, Transportation

1. Introduction

Cities are intricate systems comprising networks and connections (Batty, 2009), with roads being a vital component that facilitates the daily commuting needs of urban residents (Lyons and Chatterjee, 2008). As urbanization continues to rise globally, the increasing complexity of urban environments highlights the need for advanced analytics to better understand and manage these dynamic systems. By examining how the built environment, human everyday activities, and transport networks interact, cities can foster smarter, more sustainable transportation systems that are equipped to meet the demands of modern urban living (Zhang et al., 2023b; Lana et al., 2018) and support traffic safety analysis (Siddiqui et al., 2022; Haraguchi et al., 2022; Zhang et al., 2018; Lana et al., 2018; Du et al., 2014).

In recent years, Street View Imagery (SVI) has become an essential tool in various urban analytical tasks (Wang et al., 2024; Biljecki, 2023; Liu and Biljecki, 2022; Zhang and Huang, 2018), including comprehensive urban traffic analysis (Abdel-Aty et al., 2024; Hu et al., 2023b; Yang et al., 2023; Zhang et al., 2018; Prasolenko et al., 2015). SVI, with time series (i.e., timestamps) and locational information (i.e., geo-coordinates) attached, serves as a valuable proxy for representing the urban environment itself in a sequential manner (Zhang et al., 2023a). The spatial-temporality of SVIs captures key elements, such as road infrastructure, signage, pedestrian facilities, and the spatial arrangement of vehicles, which are associated with road safety outcomes. As such, researchers increasingly utilize SVI to conduct systematic and quantitative assessments of urban settings, as these images provide a consistent and scalable method to evaluate the physical conditions that are associated with traffic safety (Ye et al., 2025; Hu et al., 2023b; Cai et al., 2022; Edquist et al., 2012; Underwood, 2007).

Owing to the rapid advances in computer vision and artificial intelligence (AI), image segmentation and object detection have become key techniques for identifying urban features in SVI to evaluate the urban environment (Ma et al., 2024; Biljecki and Ito, 2021; Ibrahim et al., 2020). Object detection is essential for identifying and classifying objects in images, such as vehicles, pedestrians, and cyclists, which is essential for traffic monitoring and management (Zhou et al., 2025; Yue, 2024; Ghahremannezhad et al., 2023; Rezaei and EbrahimpourKomleh, 2021). Meanwhile, image segmentation partitions an image into separate segments to isolate and categorize different areas, like roads, pavements, and various types of vehicles. Such a process is often a critical step in extracting urban information that feeds into downstream tasks, such as accident analysis (Cui et al., 2025; Zhang et al., 2025; Liu et al., 2023b; Li and Luo, 2023; Zheng et al., 2023; Zhou et al., 2020) and urban environment understanding (Liu et al., 2025c; Zhang et al., 2023a; Biljecki and Ito, 2021). However, both techniques are limited to two-dimensional data. That is, while SVI offers detailed insights into the presence of urban objects, image segmentation and object detection lack the capability to utilize the three-dimensional information intrinsic to SVI, such as the distances of objects to drivers (Cai et al., 2022). Consequently, they do not fully grasp the spatial dynamics among moving entities in real-world scenarios.

It is also important to note that existing research using SVI to study urban traffic adopted, predominantly, Google Street View (GSV) or other commercial SVI providers such as Tencent and Baidu to explore urban traffic and road safety concerns (Cui et al., 2025; Zhang et al., 2025; Cai et al., 2022; Biljecki and Ito, 2021; Tanprasert et al., 2020; Zhang et al., 2017). While these platforms often offer high-resolution panoramic images that cover extensive geographic and temporal ranges at a high quality, the less frequently discussed licensing challenges present significant risks (Helbich et al., 2024), akin to a "sword of Damocles", hanging over future research (Biljecki and Ito, 2021; Gallo and Kettani, 2020). Meanwhile, the emergence of crowdsourced SVI platforms such as Mapillary and KartaView, which allow citizens to voluntarily contribute their visual experiences of urban surroundings, offers promising alternatives that could supplement commercial sources in urban environmental studies (Hou and Biljecki, 2022). However, the potential of these crowdsourced SVIs for urban traffic research remains underexplored, with limited use cases, such as measuring greenery (Biljecki et al., 2023).

This study aims to explore the association between urban street-level environments captured by Mapillary SVI and traffic accident frequency by proposing a graph-based framework that integrates a spatio-temporal regression model with a two-step explainable AI (XAI) method. Note that although traffic accidents are influenced by many short-term and dynamic factors, such as weather, traffic density, or driver behavior, this study focuses on long-term structural risks embedded in the built environment, including visual complexity and the street physical environment, which can be correlated with elevated cognitive loads and reduced situational awareness for drivers (Tapiro et al., 2020; Cai et al., 2022; Miao et al., 2025). Such conditions, while not directly causing accidents, create persistently more hazardous environments. Therefore, this study is not a cognitive study focusing on drivers' experiences on roads per se, but rather a study using SVI and a human-centered GeoAI approach to uncover the association between the urban environment and traffic accidents, with a focus on urban patterns and their implications for planning a safer urban environment.

By using temporally-ordered street view imagery, instead of modeling instantaneous or event-level risk, our model captures these persistent spatial features and their association with accident frequency over time. Our graph-based model integrates a graph attention network (GAT) (Veliˇckovi´c et al., 2018) and a long short-term memory (LSTM) network (Graves and Graves, 2012) to capture the spatio-temporal interactions among the urban objects captured by time-series SVI and the cameras' viewpoints, based on the graphs created, thus, predicting the number of accidents per urban road. Meanwhile, we further introduce a two-step XAI component (an attention mechanism for the LSTM network and a graph-level perturbation-based method for GAT) as part of the framework to explain the regression results, exploring at which timestamps the SVI was taken and which urban physical infrastructure contributed to different traffic accident frequencies on the road. In a broader context, our study also demonstrates that widely available, free crowdsourced SVI can be effective for the traffic accident modeling task examined here, competing with commercial data, as our experiments show that they are suitable for their intended purpose. Furthermore, we consider our approach to be complementary to traditional traffic safety models in transportation engineering, which typically incorporate traffic volume, roadway design, and behavioral data (Hills, 1980; Abdel-Aty et al., 2024). By contrast, our focus is on environmental perception derived from street-level imagery, which is a growing but underutilized dimension in urban risk modeling. To summarize, in this paper, we:

- study the effects of urban environmental factors captured by crowdsourced SVI on traffic accident analysis through a proposed graph-based GeoAI approach;
- analyze which characteristics of the street-level environment are relevant to a safer urban traffic environment through an XAI-based framework; • advance the study of using crowdsourced SVI to support urban traffic analytics and beyond.

2. Background

Although in this paper, our research objective is to explore the association between urban street-level environment and traffic accident frequency using crowdsourced SVI, the initial idea was inspired by three distinct yet interrelated papers on the topic of urban mobility (Yap et al., 2023a), human-centred GeoAI (Liu et al., 2023c), and XAI in urban analytics (Liu et al., 2024). In the subsequent sections, we provide a detailed explanation of how each of these studies has informed and shaped our research and the proposed framework.

2.1. Visual Complexity and Urban Mobility

Visual stimuli share a complicated relationship with urban mobility, significantly influencing the frequency and duration of trips and users' visual and physiological comfort (Koehler et al., 2025; Wisutwattanasak et al., 2023). Visual complexity, a pivotal element of urban design, is the amount of sensory information that can be perceived per unit time during urban mobility (Florio et al., 2023). Such visual complexity stems from various factors, including buildings, street furniture, traffic signs, human activity, and urban greenery. A recent framework utilized a deep learning model alongside a network science workflow to explore the determinants impacting active mobility choices and route planning, drawing on data from OpenStreetMap and Mapillary (Yap et al., 2023a). In this context, the researchers employed a pre-trained image segmentation model to delineate 65 semantic categories from Mapillary SVI. They used visual complexity as a metric to assess urban road walkability, calculated using Shannon's information entropy theory (Shannon, 2001). This method evaluates the informational content of images, identifying high visual complexity in scenes rich in diverse, well-distributed visual elements such as trees, traffic signs, and buildings. The study quantified trip experiences using visual complexity, demonstrating heterogeneous distributions of semantic indicators across urban settings, and corroborated existing research linking dense visual clusters in urban environments to increased traffic accident risk (Zhang et al., 2022; Tapiro et al., 2020). Crucially, the findings advocate for reducing visual stimuli in high-complexity areas to enhance traffic safety and promote active mobility in urban neighborhoods.

In our study, we embraced this concept of visual complexity as a pivotal starting point and facilitated the integration of human-centered perspectives into the development of our model, enabling a more nuanced understanding of how visual elements can potentially correlate to the risks of traffic accidents.

2.2. Human-centered GeoAI

Although visual complexity is a crucial factor in modeling human visual assessments of urban settings, the immersive experience of urban mobility frequently necessitates sophisticated methods and models to enhance the in situ modeling of urban activities (Liu et al., 2025a; Abdel-Aty et al., 2024; Ma et al., 2024; Otkovi´c et al., 2020). Liu et al. (2023c) introduced a human-centered GeoAI framework that integrates two pivotal steps: human-centered graph construction and a graph-based LSTM model (i.e., GraphSAGE-LSTM in their study) to predict pedestrians' outdoor comfort. To encapsulate the dynamic nature of interactions between urban spatial entities (such as trees, cars, traffic poles, and buildings) and pedestrians, urban objects were conceptualized as nodes within the graphs. The connections to pedestrians were determined by their visibility in SVI at each image acquisition time step. That is, when a pedestrian is on the sidewalk, the traffic conditions, including cars and other pedestrians within their field of view, vary with each timestamp; thus, the graphs are updated accordingly. In their study, the variation in graphs is attributed not only to changing traffic conditions but also to the presence or absence of spatial objects along the route. Such a temporal evolution of the visibility of urban objects, together with other environmental variables (e.g., sun intensity, noise), is fed into the GraphSAGE-LSTM network to predict human outdoor comfort.

The human-centered graph conceptualization of urban objects formed another key aspiration of our framework proposed in this paper. While such a method (Liu et al., 2023c) effectively captures the dynamic nature of urban objects along pedestrian pathways through graph conceptualization, it falls short in incorporating the distances between these objects and pedestrians, which plays a crucial role in contributing to safer road conditions (Abdel-Aty et al., 2024; Cai et al., 2022; Debattista et al., 2017; Faure et al., 2016; Rudin-Brown et al., 2014; Hills, 1980). Hence, integrating distance metrics into human-centered graph construction has the potential to enhance the predictive accuracy and practical utility of our model in broader urban planning and safety contexts.

2.3. XAI in Urban Traffic Analysis

XAI comprises a suite of methods designed to enable humans to comprehend the decisions or predictions made by AI, which has recently become a prominent topic in urban analytics (Lei et al., 2025; Liu et al., 2024; Law et al., 2023; Liu et al., 2023d; Li, 2023). While much of the existing literature has employed techniques such as SHapley Additive exPlanations (SHAP) (Xu et al., 2023; Li, 2022) or saliency mapping (Hu et al., 2023a; Cheng et al., 2017) to study SVI in assessing urban environments, Liu et al. (2024) introduced a graph-based XAI method, GNNExplainer. It operates as a perturbation-based method that integrates spatial information into the model-explanation process, identifying which urban objects are associated with specific traffic volumes. Perturbation-based methods in XAI refer to techniques that assess the importance or influence of features in a model's decisions by systematically altering (perturbing) features and observing the resulting change in the model's output. In their study, they demonstrated that the concept of perturbation-based explanation opens opportunities for urban planning practices by augmenting the proportion of urban objects in SVI, thereby providing insights into planning strategies.

In line with the prevailing trend in XAI for urban analytics (Lei et al., 2025; Liu et al., 2024; Law et al., 2023), our adoption of a perturbation-based XAI method is a crucial component in investigating which urban street-level factors are associated with varying numbers of traffic accidents. By elucidating the complex interactions and influences within the urban model, XAI yields insights that could inform technology-driven urban traffic safety strategies.

3. Method

In this study, we introduce a multi-faceted approach to analyze and interpret urban environments through visual sensing and advanced machine learning techniques, as shown in Figure 1. In this framework, we first capture the visual complexity of urban settings using Mapillary SVI through the image segmentation method Mask2Former (Cheng et al., 2022) and the object detection-based method Dist-YOLOv5 (Vajgl et al., 2022), which can estimate distances between urban objects and SVI cameras. Such a visual assessment of the urban environment facilitates the development of a human-centered graph, wherein nodes represent key urban elements, and edges depict the interactions and proximities between these elements based on Dist-YOLOv5. We then apply a GAT-LSTM model to predict the number of urban traffic accidents by learning temporal dependencies and node-level relationships within our constructed graph. Lastly, to ensure the interpretability of our model predictions, we implement a two-step XAI approach. This approach involves an initial time-step-wise importance evaluation followed by graph-level perturbation-based explanation, enabling a deeper understanding of model reasoning in the context of urban environment analysis.

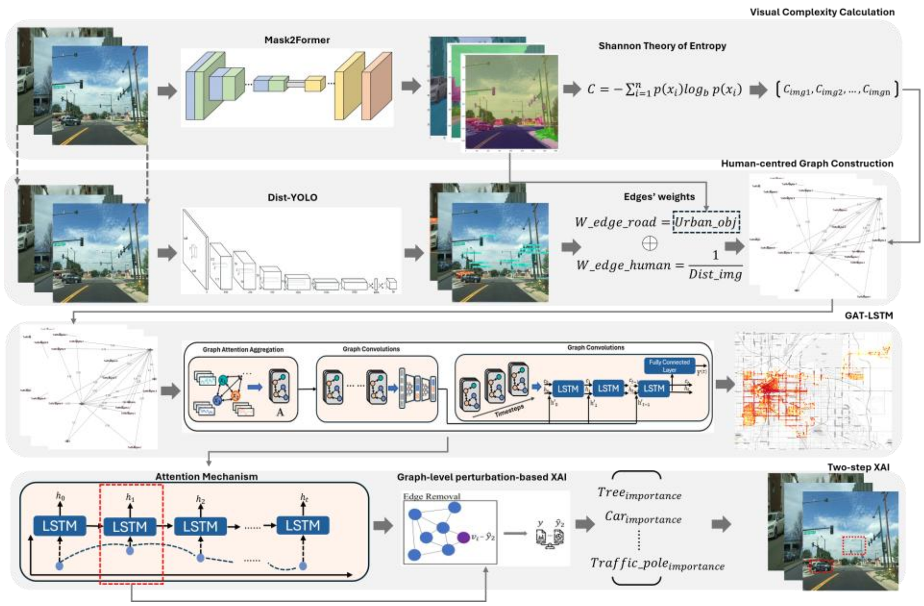

Figure 1: The diagram for the proposed framework. A detailed introduction of each component is provided in the respective subsection of the Method section. This diagram contains SVI retrieved from Mapillary under the Creative Commons (CC) license and a base map from OpenStreetMap licensed under the Open Data Commons Open Database License (ODbL) by the OpenStreetMap Foundation (OSMF).

3.1. Visual Sensing of Urban Environment

The visual sensing component of our study is designed to capture and analyze the complex visual features of urban environments from sequentially organized SVI. By leveraging the temporal dimension of SVI, as underscored by Zhang et al. (2023a), we gain not only a snapshot of the urban environment but also a more nuanced, time-sequenced understanding of how these environments evolve and interact with urban activities. To accurately simulate this visual complexity, we employ two advanced technologies: Mask2Former (Cheng et al., 2022) for image segmentation and Dist-YOLOv5 (Vajgl et al., 2022) for object detection.

Mask2Former, a leading deep learning model for image segmentation, combines transformer architectures (Vaswani et al., 2017) with mask prediction capabilities, enabling it to handle a range of segmentation tasks effectively. Utilizing a pre-trained Mask2Former model available on HuggingFace [1], we extracted diverse semantic classes from Mapillary SVI (the description of the data used in this study can be seen in the Study Area section). A Mask2Former model, trained on the Mapillary Vistas dataset (Neuhold et al., 2017), which includes 65

[1] https://huggingface.co/facebook/mask2former-swin-large-mapillary-vistas-semantic

semantic classes for its SVI, demonstrates its proficiency in capturing a wide range of urban elements, such as trees, roads, and buildings, as depicted in Figure 2.

Image ID (renamed)	Time-stamp	Geolocation	Compass Heading	Sequence ID (renamed)	Focal Length	Field of View	Resolution
Image 001	2020-06-25	39.642246° N 105.025620° W	90° (East)	seq123	4.5 mm	90°	2048 x 1536 px
Image 002	2021-08-07	39.833913° N 104.673576° W	67° (East)	seq234	4.5 mm	90°	1920 x 1080 px

Figure 2: Examples of Mask2Former output, illustrating segmentation of various urban elements. Images also contain an example of Mapillary metadata. To protect the privacy of the original Mapillary users who uploaded the SVI, we have anonymised all relevant metadata except for the timestamp and geolocation, which remain unchanged from the original data. SVI retrieved from Mapillary under CC licenses.

To further model the visual complexity as experienced by humans, we apply the Shannon Theory of Entropy, akin to the methodology used by Yap et al. (2023a,b), represented by the following formula (also illustrated in Figure 1):

$$C = -\sum_{i=1}^{n} P(x_i) log_b P(x_i) \tag{1}$$

where C represents the visual complexity, $P(x_i)$ is the probability of encountering the i-th segmentation class among the possible 65 classes from SVI, and $\log_b$ denotes the logarithm base b, which is set to 2 for this calculation. This entropy measure quantifies the information content in images, where scenes with a rich array of evenly distributed visual elements (e.g., diverse street objects) correspond to high visual complexity. Additionally, the use of Shannon Entropy facilitates data cleaning, which is crucial for the crowdsourced analysis of SVI. As per our methodology, images with a visual complexity score below 1.0 are excluded, indicating the dominance of singular visual elements, such as walls or open roads, which are less informative for our study objectives. The visual complexity is fed as an essential feature of the human-centered graph; details are introduced in the following section.

Dist-YOLOv5 is an advanced variant of the popular YOLO (You Only Look Once) object detection framework (Redmon et al., 2016) that incorporates distance estimation into the traditional YOLO architecture, enhancing its utility in applications where spatial awareness is critical. We used a pre-trained DistYOLOv5 (trained on the COCO dataset (Lin et al., 2014)) to identify urban objects in SVI, with distances estimated between the drivers' viewpoints and the objects. An example output of the Dist-YOLOv5 can be seen in the left part of Figure 3. Note that only object classes that co-exist in the Mapillary Vistas dataset (Neuhold et al., 2017) were used for the study to ensure consistent urban elements are used for human-centered graph construction later. Such an output is a critical step in developing human-centered graphs, which incorporate the spatial and contextual relationships between the driver and various urban elements. Nevertheless, it is well known that the backwards projection of photographs is recoverable only up to scale. Thus, the Dist-YOLOv5 model focuses on spatial relativity rather than absolute distances between objects. Consequently, the distances estimated by Dist-YOLOv5 are not entirely accurate. However, they still provide approximate spatial reasoning for the observed objects and can serve as useful inputs to the method proposed in this paper (see the sensitivity analysis in Section 5.2).

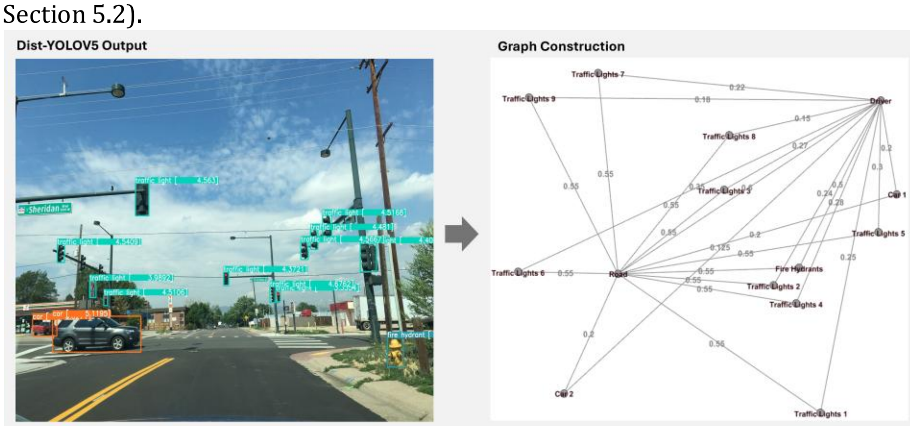

Figure 3: An example of human-centred graph construction.

3.2. Human-centered Graph Construction

As mentioned in Section 2, Liu et al. (2023c) developed a human-centered method for constructing graphs that captures the dynamic interactions between urban objects and pedestrians during their walking activities on urban sidewalks.

Building on their proposed method, we identified the urban objects detected by Dist-YOLOv5 as graph nodes and used the estimated distances as edge weights. Importantly, considering that urban objects closer to the viewpoint have a stronger influence on driving experiences (Cai et al., 2022), we incorporated a distance conversion step (Liu and De Sabbata, 2021) as follows (also depicted in Figure 1):

$$W_{edge_{human}} = \frac{1}{Dist_{img}} \tag{2}$$

where $W_{edge\ human}$ is the weight of the graph's edge; $Dist_{img}$ is the distance estimated between an urban object and the SVI camera's viewpoint. An illustration of the constructed graph is shown in Figure 3. Note that, as shown in the figure, we created an artificial node *Road* in the constructed graph. The creation of such a node involves, first, studying the impact of road conditions on traffic accidents, serving as a conceptual node that represents the road as an essential urban object within the broader context of understanding the urban environment. Second, from a technical perspective, this increases graph complexity by connecting all urban objects to this node, ensuring that a graph neural network (see the following subsection) has sufficient variability and structural features in the graph information to train the network effectively. Edge weights for those connections to the roads are assigned based on the image segmentation results, that is, the proportion of urban objects extracted from Mask2Former, denoted by $Urban_{obj}$ in Figure 1. These connections are intended to represent the spatial and functional relationships between the road and other urban elements, as observed in the SVI. It is important to note that while roads may occasionally be occluded by other participants (e.g., vehicles, pedestrians, or other obstacles), these occlusions reflect the complexity and dynamic nature of traffic. Such instances of road occupation provide valuable insights into how various elements coexist and interact.

Following the construction of this human-centered graph, we proceeded to assign features to its nodes. The node labeled *Driver* in the graph was assigned features based on the visual complexity calculated from the output of Mask2Former, as detailed in the previous section. For all other nodes, representing urban objects identified by Dist-YOLOv5, the proportional size of their bounding boxes relative to the SVI was used as a node feature, reflecting the relative sizes of different objects in the driver's vision. For the *Road* node's features, we adopted the average proportional size of all detected road objects' bounding boxes relative to the SVI, hence, capturing how much of the image frame is occupied by the road and related objects, such as vehicles, road signs, and traffic lights. Such a proportional measurement of *Road* node serves as a proxy for assessing the road's visibility and its interaction with other urban elements within the image. In line with Liu et al. (2023c), we also incorporated a data augmentation step by embedding sequential information from SVI on the roads, expanding the feature set of each node, as illustrated in Figure 4. Such an approach not only benefits model training by enhancing data dimensionality and nonlinearity but also enables the model to capture temporal variations in urban mobility.

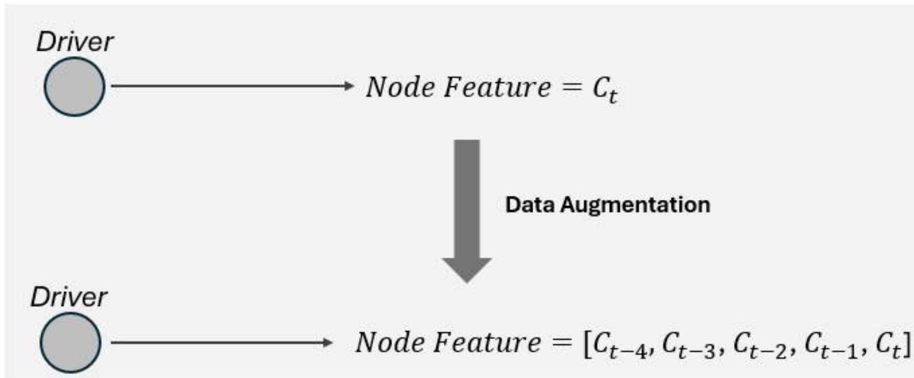

Figure 4: An example of the data augmentation step. Taking a node in a graph with the label *Driver* as an example, the node feature for it contains visual complexities of the five previous time steps.

3.3. GAT-LSTM

In this paper, we introduce a Graph Attention Network (Veliˇckovi´c et al., 2018)-Long Short-Term Memory (Graves and Graves, 2012) (GAT-LSTM) model, representing an advanced hybrid neural network architecture, to predict urban traffic accident frequency, as shown in the GAT-LSTM part of Figure 1. As a type of graph convolutional network, GATs represent nodes as numeric vectors and aim to learn a feature function across the nodes that effectively incorporates the graph structure. Each node (urban objects in our paper) *i* in the graph (i.e., the constructed human-centered graph) is associated with a feature vector $\mathbf{h}_i$. The core idea behind the GAT is to compute the hidden representations of each node by attending to its neighbors, selectively emphasizing information from more 'important' nodes, which is achieved through a set of learnable parameters and the attention mechanism. The attention mechanism in a GAT involves a shared attention function α that computes attention coefficients indicating the importance of node *j*'s features to node *i*. These coefficients are computed as follows:

$$e_{ij} = \mathrm{LeakyReLU} a^T [\mathrm{W}h_i || \mathrm{W}h_j] \tag{3}$$

where **W** is a weight matrix applied to every node, and **a** is a weight vector that transposes the concatenated features of nodes *i* and *j*. The use of LeakyReLU ensures non-linearity in the attention mechanism. The coefficients are then normalized across all choices of *j* using the softmax function (Bridle, 1989), ensuring that they sum to one and thus can be interpreted as probabilities:

$$a_{ij} = \frac{\exp(e_{ij})}{\sum_{k \in \mathcal{N}(i)} \exp(e_{ik})} \tag{4}$$

where $\mathrm{N}(i)$ denotes the neighbors of node *i* in the graph. Such a normalization step is crucial as it makes the model's output invariant to the size of the neighborhood. The final step in the GAT layer is to compute the new features for each node based on these normalized attention coefficients. The output feature of each node, $\mathbf{h}'_i$, is obtained by a weighted combination of the transformed features of its neighbors, summed together with the node's own transformed features:

$$h_i' = \sigma(\sum_{j\in\mathcal{N}(i)} a_{i,j}\mathrm{W}h_j) \tag{5}$$

where σ is a non-linear activation function of ReLU (Fukushima, 1969). This update rule allows each node to dynamically choose which of its neighbors contribute to the next layer's features, making GATs highly adaptive and powerful for extracting complex patterns from graph data. The use of a GAT in our study approximates the relative importance of visual elements based on humancentered graphs, studying which urban objects would lead to a high risk of traffic accidents on roads.

The second component of the model, an LSTM network, addresses the temporal aspect of traffic data. LSTMs are a type of recurrent neural network (RNN) optimized to capture long-term dependencies in time series data, avoiding the vanishing gradient problem (Hochreiter, 1998) common in traditional RNNs. As shown in Figure 1, at each time step t, the output of the GAT, $\mathbf{h}'_{it}$, serves as the input feature sequence, $\mathbf{x}_t$, for the LSTM. h and c are the hidden states and cell states of LSTM cells updated by a series of activation functions (details regarding those functions can be referred to Graves and Graves (2012)). Owing to practical considerations such as the uneven distribution of SVI on the roads, for example, primary roads typically featuring a higher concentration of SVI, we implemented an additional configuration for the LSTM. Such a modification enables the model to handle input sequences of varying lengths, thereby obviating the need to pad them to a uniform length. Consequently, the traditional LSTM model was transformed into a 'dynamic' LSTM. This dynamic nature was enabled by constructing a computation graph tailored to each input sequence. In essence, this means that the graph structure can adapt to accommodate the length of the data it processes in each batch or instance. Such an adaptation not only enhances the model's efficiency but also its ability to process SVI data related to driving experiences, where input sequence lengths can vary significantly, thereby more accurately reflecting the variable nature of data encountered in practical scenarios.

By integrating LSTM units, the GAT-LSTM model can track temporal changes, such as the evolution of sequential variations in urban environments captured by SVI. We apply such a GAT-LSTM model for a regression task for predicting the number of car accidents on the roads, denoted as $Y^{(t)}$ in Figure 1.

3.4. Two-step XAI

After obtaining predictions from the GAT-LSTM model mentioned above, we introduce two novel XAI methods: the LSTM attention mechanism approach and the perturbation- and graph-based XAI method. Such a two-step XAI method involves creating a framework that first identifies critical time steps via LSTM's attention mechanism and then elucidates node-level influences within those critical time steps using a perturbation-based method for the GAT, as shown in the last part of Figure 1. Such a dual-layered explanation strategy provides both temporal and spatial insights into the drivers of predictions, which is particularly useful in our accident frequency prediction task, where both temporal dynamics and network structures play key roles.

The first step involves modifying the LSTM to incorporate an attention mechanism that highlights which time steps are most influential in its predictions, achieved by adding a soft attention layer (Yuan et al., 2020) over the LSTM outputs before they are used for the final prediction. An intuitive understanding of the mechanism can be seen in Figure 1, where the attention mechanism was added to the LSTM part of the model by calculating attention scores *AC* for each hidden state h_t, reflecting its potential importance in the model's prediction:

$$AC_t = v^T \tanh(W_a h_t + b_a) \tag{6}$$

where W_a, b_a and v are trainable parameters, with v typically being a context vector that helps capture the relevance of each time step. Then, a softmax normalization step was applied to the attention scores calculated, resulting in a distribution of weights β_t that sum to one. Each weight quantifies the relative importance of its corresponding time step:

$$\beta_t = \frac{\exp(AC_t)}{\sum_{k=1}^{T} \exp(AC_k)} \tag{7}$$

By selecting the highest β_t, we identified the most influential time step of the driving activity, which contributed to different numbers of traffic accidents.

After identifying the most influential time step using the LSTM attention mechanism, we selected the corresponding graph to conduct our graph-based XAI analysis. This phase of the analysis utilized a perturbation-based approach to elucidate the functioning of the GAT. This method entails methodical modifications (perturbations) of the graph's nodes or edges and monitoring the resultant changes in output, thus pinpointing the critical elements of the graph. For this study, given our objective to explore the nexus between driving experiences and the likelihood of car accidents, our attention was particularly drawn to the explanation of individual nodes, specifically the node labeled as *Driver*. Hence, we systematically removed each Dist-YOLOv5-identified node connected to the *Driver* in the graph, introduced each modification into the GAT-LSTM predictive model, and observed the predicted outcomes, denoted by $Y^{predicted}$. By comparing these values with the original predictions, Y, we can identify which nodes exerted the greatest impact on predictive accuracy. Further explanation is provided in the Results section.

3.5. Model Implementation

We developed our proposed framework using the Python programming language, specifically utilizing the PyTorch library (Paszke et al., 2019) and the PyTorch Geometric library (Fey and Lenssen, 2019) for its implementation. The model in question, a GAT-LSTM, was fully trained in 200 epochs, each epoch representing a complete pass through the dataset, which follows a standard train-test data split (70% data as the training data, 10% as the validation data and 20%

as the test dataset) through the cloud-based Google Colab[2] platform. We employed the Adam optimization algorithm (Kingma and Ba, 2014) with a learning rate of 0.001 to refine the model's weights.

In line with standard practice for regression tasks, we adopted mean squared error (MSE) as the loss function, given its widespread use and robustness in quantifying the average squared difference between predicted and observed values. Model parameters were initialized randomly, and training was conducted using a batch size of 16 to balance computational efficiency with stable optimization. For performance evaluation, we used the Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and the coefficient of determination (R^2) to assess predictive accuracy.

4. Study Area

In this study, we have selected Denver, Colorado, as our case study area. Denver is particularly well-suited as a focus for traffic accident analysis due to its considerable traffic volume, rapidly expanding population, and high urban density (Goetz and Boschmann, 2018). These elements are correlated with a diverse range of driving conditions that can influence accident rates, leading to an increase in traffic accidents in recent decades. The city offers a diverse range of road types, from bustling urban streets to challenging mountainous routes, and provides extensive, accessible traffic and accident datasets. Furthermore, Denver is actively engaged in traffic management strategies, notably its Vision Zero initiative (City and County of Denver, 2022), which aims to eliminate all traffic-related deaths and serious injuries by 2030. These factors not only enrich the data available for analysis but also underscore the city's significance as a model for research aimed at improving road safety and informing urban planning policies.

4.1. Traffic Data Collection

We gathered data on traffic crashes occurring on urban roads in Denver, spanning more than 10 years and covering all seasons, from 1st January 2013 to 16th February 2024. This data was obtained from the City and County of Denver's open data portal[3], a publicly accessible resource providing detailed datasets on various civic metrics. The raw data, initially presented as geospatial points, was meticulously processed to map each accident to specific road segments, enabling a structured analysis of traffic incidents across different thoroughfares. To facilitate a clearer understanding of accident distribution, we developed a visual representation of the compiled data. The resulting map, which illustrates the spatial distribution of traffic accidents, is displayed in the left part of Figure 5. The map delineates these high-risk zones, showing a clear concentration of incidents along I-25, a critical artery that carries a significant volume of vehicular traffic daily. Additionally, the areas surrounding the city center, which are in close

[2] https://colab.research.google.com/
[3] https://www.denvergov.org/opendata/dataset/city-and-county-of-denver-trafficaccidents

proximity to the I-25, also show a heightened frequency of traffic accidents. Such a pattern suggests that the central urban layout, combined with heavy traffic, contributes to the increased risk of collisions.

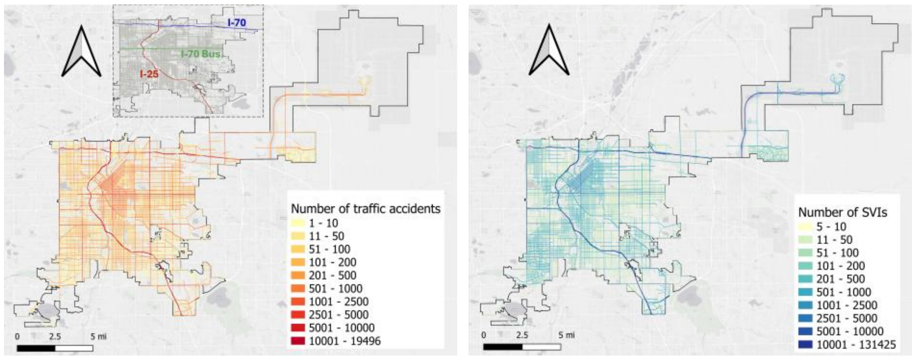

Figure 5: Number of the traffic accidents and SVI in Denver. Map sources: Esri, HERE, Garmin, INCREMENTP, © OpenStreetMap contributors, and the GIS user community.

4.2. Street View Images Collection

For the SVI utilized as input for this study, we acquired data from Mapillary, covering the period from 1st January 2018 to 31st December 2023 in the study area. This extensive dataset comprises over one million images, placing it in the same order of magnitude as commercial SVI sources, thereby offering a comprehensive visual resource. Mapillary SVI comes with rich metadata, including geolocation, timestamps, camera orientation, identifiers linking sequential images, camera settings, and image quality metrics (e.g., resolution), as shown in Figure 2. Notably, although SVI collected were not directly associated with the timing of traffic accidents, the inclusion of timestamps and sequence identifiers enables the exploration of temporally organized SVI as a sequential representation of urban environments, as highlighted by Zhang et al. (2023a), where such sequences act as powerful proxies for urban understanding and urban safety plan outcomes (Hu et al., 2023b). Therefore, SVI images are not temporally aligned with individual traffic accidents. Rather, we use temporally ordered sequences of SVI to capture the evolution of street-level environments over time, while accident frequency is treated as an aggregate measure over the 2013–2024 period. This approach is designed to identify persistent visual features that are structurally associated with higher accident risks at the road level.

A visualization detailing the distribution patterns of these collected SVI is presented on the right of Figure 5. The two maps in Figure 5 reveal a distinct spatial-temporal correlation; major roads such as the I-25 and central urban areas exhibit both a high incidence of traffic accidents and a substantial number of SVI. Interestingly, the visualization also highlights that residential areas, particularly in the northwest and southwest of Denver, show a high density of SVI despite lower traffic accident statistics. This discrepancy suggests varying traffic dynamics across different regions of the city, underscoring the potential of temporally organized SVI to uncover nuanced patterns in urban environments beyond immediate accident correlations.

4.3. Data Cleaning

A 3-step data-cleaning process was employed to ensure data integrity and usability. The first step addresses the inconsistency between the locations of recorded traffic accidents and the availability of SVI. Specifically, not all roads with documented traffic accidents have corresponding SVI values, particularly for narrower or less frequently used pathways, such as alleys and side streets. These areas, although prone to accidents, often lack visual documentation from the drivers' perspectives, likely due to their limited access or lower traffic importance. We excluded these roads from our study area to ensure that our analysis is based solely on data-rich environments where both accident records and SVI are available, thereby enhancing the precision of our findings and the effectiveness of subsequent recommendations.

The second step in our data-cleaning process was to address the uneven distribution of SVI across the Denver urban area. As illustrated in Figure 5, primary thoroughfares such as the I-25 can amass upwards of 13,425 SVI, whereas less frequented roads may have only a handful. To achieve a more uniform distribution of data and prevent biases arising from uneven data availability, previous studies (Liang et al., 2023; Li et al., 2022) have employed methods to select a fixed quota of SVI per spatial unit. However, this could pose issues if the spatial units under examination differ substantially in size. For example, primary roads tend to be longer than residential or business district streets, and a limited set of images may not adequately represent the varied driving experiences found along these extended routes.

To overcome this imbalance, we refined our method by segmenting each road into 50-metre intervals, selecting the five most recent SVI for each segment, organized in a time-evolving manner based on their timestamps to leverage the sequential nature of SVI as a powerful tool for capturing temporal and spatial variations in urban environments (Zhang et al., 2023a). This strategy ensures a more representative visual sampling across different road lengths and types. For roads shorter than 50 meters, we also collected the five most recent SVI, provided they were available. Roads lacking a minimum of five SVIs were excluded from our dataset, ultimately narrowing the study's scope to 3,429 roads out of 8,806. We acknowledge that such a filtering process inevitably introduces a selection bias by favoring roads with more comprehensive SVI coverage, which typically corresponds to higher-traffic and more accessible routes. Consequently, residential streets and low-volume roads are underrepresented in the final analysis. Nevertheless, since major roads account for a disproportionately high share of traffic volume and accident occurrences, focusing on these segments still provides meaningful insights into systemic spatial risks.

The third phase was specifically engineered to mitigate the uncertainties associated with using SVI data, as highlighted by Biljecki et al. (2023). Crowdsourced data, while widely accessible, is often marred by inconsistencies in quality and uniformity. This variability stems from the broad spectrum of contributors and the heterogeneous methodologies they utilize for data collection. As detailed in the *Method* section, the Shannon Entropy approach was employed to exclude images likely depicting roads or walls. In this study, we excluded panoramic SVI (i.e., those that provide a 360-degree view) to more accurately reflect the realistic field of view available to drivers. To enhance data quality further, we applied the method outlined by Hou and Biljecki (2022) to eliminate blurry SVI and adopted a pre-trained, deep learning-based technique from Hou et al. (2024) for detecting SVI orientations (e.g., facing front or side), ensuring the inclusion of only front-facing views, thus better simulating a driver's perspective. Note that, for each image eliminated, another SVI that was the most temporally recent, non-blurry, and front-facing was selected from the remaining available images for the respective stretch of road. As a result of such a step, our collected SVI dataset consists of 204,675 geotagged photographs, organized in sequential order for each road.

5. Results

The findings from the proposed study are delineated in two distinct sections. The first section assesses the performance of the GAT-LSTM predictive model by comparing it against a series of baselines and conducting ablation studies. The second section conducted a sensitivity analysis to assess the robustness of our human-centered graph construction against inaccuracies in distance estimation. The third section undertakes a comprehensive exploration into what urban characteristics on the roads within drivers' visual fields are relevant to traffic accidents, utilizing the innovative two-step XAI methodology proposed in this research.

5.1. Pattern Analysis of Urban Traffic Accidents

We established three baseline models to evaluate alongside our proposed GAT-LSTM model, which utilizes the proposed human-centered graphs. These baselines include a conventional machine learning algorithm, Random Forest (RF) (Breiman, 2001), and two advanced graph-based neural network approaches: a Temporal Graph Convolutional Network (T-GCN) (Zhao et al., 2019) and a GraphSAGE-LSTM (Liu et al., 2023c). It is pertinent to note that the GraphSAGE-LSTM model is the only baseline that can serve as a direct comparator to our method, as it can be adapted to handle data with identical structures. In contrast, the other two models—while representing comparisons with current state-of-the-art techniques—require a series of modifications to their input formats to enable direct comparison. The setup for these baseline models is described as follows:

- RF: it is widely recognized as one of the foremost machine learning methodologies, and it has been extensively applied within the realm of traffic analysis (Hruthik et al., 2023; Shahid et al., 2021; Cheng et al., 2019; Lana et al., 2018; Liu and Wu, 2017), including numerous studies focused on accident forecasting (Cai et al., 2022; Hamad et al., 2020; Lee et al., 2019; Dogru and Subasi, 2018). Unlike our proposed GAT-LSTM model, which adeptly captures urban objects using SVI obtained from SVI camera perspectives, RF requires a distinct data-processing approach for input preparation. In this research, we employed a conventional SVI processing method that extracts semantic classes for each road segment, as outlined in the Study Area section. We then computed the visual proportions of these objects using the same pre-trained Mask2Former model used in our primary method. The extracted data, featuring semantic classes and their averaged visual proportions, was used as input features for the RF model. This model was configured in accordance with established protocols from previous studies (Liu et al., 2023c), with 200 estimators, a maximum depth of 220, and the default maximum number of features. We structured the RF model as a regression model and utilized the same metrics of MAE, MAPE, and R^2 to assess its performance.

- T-GCN: an advanced neural network framework designed to manage data that is both graph-structured and temporally varying. Such a model was initially developed with a focus on traffic prediction, making it an appropriate benchmark for comparing our proposed framework. As detailed by Zhao et al. (2019), we also conceptualized urban roads as nodes within a graph, with edges representing the connections between these roads. For node features, we utilized semantic classes extracted from SVI collected per road segment using the Mask2Former, which was organized in a temporal manner. However, it is crucial to note that the roads in Denver, as described in the Study Area section, exhibit variations in the number of SVI available. Such a discrepancy necessitates a feature engineering process to ensure that all nodes in the graph are represented with features of consistent dimensions. In our experiments, we employed padding techniques to standardize the input sizes for all nodes (Dwarampudi and Reddy, 2019). Additionally, we re-engineered the graph convolutional component of the T-GCN into a regression model to better suit our analytical needs.

- GraphSAGE-LSTM: The GraphSAGE-LSTM model, as developed in the framework outlined by Liu et al. (2023c), serves as a pertinent comparator to our own framework due to its ability to process input data in a similar format. In the original implementation of this model, human-centered graphs were constructed using a binary approach to define connectivity; connections between urban objects and the viewpoints of cameras were either absent (denoted by 0) or present (denoted by 1) in the adjacency matrix, without estimating actual distances. Hence, to compare with our framework, we modified the adjacency matrix introduced in the Humancentered Graph section that aligns with the specifications set forth by Liu et al. (2023c). Because of their different adjacency matrices, our comparison focuses on the implications of omitting visual distances from the model, allowing us to examine how the inclusion or exclusion of such vision-based spatial reasoning affects the model's predictive performance in urban environments.

In addition to the baseline comparisons, we set up two ablation studies, which omit parts of the GAT-LSTM network:

- GAT-only: For each SVI per road, we used the GAT to create a graphlevel embedding (numeric vectors) (Cai et al., 2018). These embeddings encapsulate key features of the graph nodes by encoding both semantic and structural details of the SVI. After creating individual embeddings for the SVI of each road, we merged these into a cohesive street imagery embedding for that specific road. Such a comprehensive embedding was subsequently used as input to a simple RF classifier for regression. The purpose of this approach was to conduct an ablation study to assess the impact of the sequential temporal information inherent in the SVI on the model's performance. By omitting the LSTM component of the GATLSTM, the model loses its original ability to capture temporal dependencies in driving activities.

- LSTM-only: This ablation study investigated the importance of the humancentered graph for the model's performance. By removing the GAT component, the LSTM model takes the semantic segmentation output as input for the regression task.

The outcomes of this study are detailed in Table 1, which presents a thorough comparison of the previously delineated baseline models. Each model utilizes a distinct combination of features to evaluate its performance on the SVI data collected during this research. Among these, the GAT-LSTM model stands out, consistently outperforming across all assessed metrics. This model, which integrates a sophisticated mix of features within its hybrid structure (combining

Table 1: Results comparisons. '✓' represents features that the model incorporates. Results were reported as the mean values, with upper and lower limits provided, after running each model 10 times. Average Computational Time is the average computational time per prediction for the models on the test dataset.

Features	GAT-LSTM	Random Forest	T-GCN	GraphSAGE-LSTM	GAT-only	LSTM-only
Semantic Segmentation	✓	✓	✓	✓	✓	✓
Distance Estimation	✓				✓	
Visual Complexity	✓			✓	✓	
Temporal Dependency of SVI	✓		✓	✓		✓
MAE (lower ~ upper limits)	**0.67** (0.64 ~ 0.70)	3.31 (3.27 ~ 3.35)	0.85 (0.82 ~ 0.88)	0.73 (0.70 ~ 0.76)	1.83 (1.80 ~ 1.86)	1.36 (1.33 ~ 1.39)
MAPE (%, lower ~ upper limits)	**6.79** (6.75 ~ 6.83)	26.61 (25.51 ~ 26.67)	9.72 (9.68 ~ 9.76)	8.92 (8.67 ~ 8.96)	18.65 (17.98 ~ 18.70)	16.87 (15.87 ~ 16.91)
R^2(lower ~ upper limits)	**0.80** (0.77 ~ 0.82)	0.33 (0.31 ~ 0.35)	0.71 (0.69 ~ 0.73)	0.73 (0.71 ~ 0.75)	0.47 (0.45 ~ 0.49)	0.56 (0.54 ~ 0.58)
Average Computational Time (seconds)	**0.92**	0.49	0.61	0.89	0.70	0.56

GAT and LSTM), achieves the lowest MAE and MAPE and the highest R^2. Such results underscore its capability to capture both the spatial and temporal dimensions of the SVI dataset effectively. Although, as indicated in the table, the GAT-LSTM model is the most computationally intensive due to its complex architecture, its predictive accuracy across all three assessments suggests a trade-off between computational cost and predictive performance. It is important to re-emphasize that this research does not aim to provide a real-time traffic accident prediction model; therefore, although we include computational time as an important metric for evaluating the model, how to improve the model efficiency is not within the primary scope of this research.

Conversely, the RF model, which represents a more traditional machine learning approach, is limited to utilizing only the semantic segmentation outputs from the SVI. This constraint leads to the poorest performance among the tested models, reaffirming the need for advanced modeling techniques to simulate urban residents' experiences accurately (Liu et al., 2023c; Cai et al., 2022; Abdelrahman et al., 2022). On the other hand, the T-GCN model, which is designed to incorporate temporal variations in driving experiences through sequential SVI at an urban scale, shows commendable results in predicting traffic accident frequency, pointing to the critical role of comprehensive visual modeling in accurately capturing spatial objects on urban roads.

The GraphSAGE-LSTM model, a direct competitor to our proposed GATLSTM model, secures the second-best performance. Such a baseline comparison highlights the effectiveness of incorporating estimated distances into the humancentered graphs, enhancing the GAT-LSTM capability to capture a nuanced understanding of the urban environment and improving its prediction accuracy on traffic accident frequency (Liu et al., 2023a; Cai et al., 2022; Ning et al., 2022; Guan et al., 2022). Moreover, both GAT and GraphSAGE are variants of graph convolutional neural networks, despite adopting different convolutional approaches (Wu et al., 2020). One of the distinctive differences between GAT and GraphSAGE is the attention mechanism implemented in the GAT component. Therefore, the results also indicate the usefulness of the attention mechanism in capturing critical urban objects from SVI that pose high risk of urban traffic accidents.

In stark contrast, simpler models, such as GAT-only and LSTM-only, which rely on fewer features, demonstrate inferior performance. Intriguingly, models based solely on the LSTM component outperform those based solely on GAT. This finding highlights the importance of temporal dependencies in SVI, suggesting that they are more effective at mimicking real-world driving conditions and, consequently, at enhancing the precision of traffic accident predictions (Sainju and Jiang, 2020; Zhang et al., 2020). The evidence clearly indicates a pivotal need to integrate both spatial and temporal data to create more accurate and reliable models for simulating urban experiences.

5.2. Sensitivity Analysis: Impact of Distance Estimation Error

To assess the robustness of our human-centered graph construction against inaccuracies in distance estimation, we conducted a sensitivity analysis by introducing controlled perturbations to the distance values estimated by DistYOLOv5. Such an analysis aims to evaluate the sensitivity of the GAT-LSTM model's predictive performance to noise in the edge weights, which represent object-to-camera distances.

As described in Section 3.2, edge weights in the human-centered graph are computed as the inverse of the estimated object distances. To simulate uncertainty in these distance estimates, we applied additive Gaussian noise to the raw distances before computing edge weights. Specifically, for each distance d, we generated perturbed distances $d' = d + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$. We tested three noise levels with $\sigma = 0.5, 1.0$, and 2.0, corresponding to low, moderate, and high noise. Each perturbed graph set was used to retrain the GAT-LSTM model under the same training-validation-test split as the original experiment. All other settings, including node features and model hyperparameters, remained unchanged.

Table 2: Sensitivity analysis of GAT-LSTM performance under different levels of noise added to distance estimations.

Noise Level	**MAE**	**MAPE (%)**	$\boldsymbol{R^2}$
No noise (baseline)	0.67	6.79	0.80
Low noise (σ = 0.5)	0.70	7.01	0.79
Moderate noise (σ = 1.0)	0.72	7.26	0.78
High noise (σ = 2.0)	0.89	9.34	0.69

Table 2 presents a comparative performance analysis across different noise levels. The results suggest that the GAT-LSTM model is robust to moderate levels of distance estimation error. At low and moderate noise levels, the changes in MAE, MAPE, and R^2 are within 3–5% of the baseline. Performance degradation becomes more pronounced only under high noise conditions. Hence, these findings indicate that, although precise distance measurements are not achieved, the relative spatial configuration captured by Dist-YOLOv5 is sufficient for the model to learn meaningful associations between visual structure and accident frequency. The model's reliance on approximate spatial reasoning, rather than metrically accurate input, supports the feasibility of using lightweight monocular vision approaches for urban-scale risk modeling.

5.3. XAI interpretation

To interpret the model performance of GAT-LSTM, we categorized the roads in Denver into two distinct groups based on traffic accident data from 2013 to 2024. Roads experiencing, on average, more than five accidents per month over this period were designated as *High Risk*, while those with five or fewer were labeled *Low Risk*. As a result, 95.53% of the roads in Denver were considered as *Low Risk*, and 4.47% of roads were identified as *High Risk*, as visualized in Figure 6. It is important to acknowledge that there is no universally accepted criterion for classifying roads as "high risk" or "low risk"; the thresholds used here are specific to our study and not intended to establish general standards for road safety. Therefore, this binary classification of roads is introduced solely for interpretive purposes within the XAI framework. It does not influence model training or prediction. The threshold of five accidents per month is a heuristic that ensures the 'high risk' category captures roads with consistently elevated incident rates over the 10-year span, facilitating clearer interpretability in model explanation.

To intuitively understand how the XAI model was adopted in this paper, we chose a section of U.S. Road I-70 Bus, a west-east road with a length of 8.7 kilometers that functions as a primary route for commuters in and out of the city, as illustrated in Figure 6. Such a road experiences heavy urban traffic, with the number of traffic accidents recorded over the 10 years at 5,551. As shown in Figure 7, with the GAT-LSTM predicting 5,937 accidents on such a road, we began to interpret the model by analyzing which time steps of the SVI contributed to this prediction, utilizing the attention mechanism implemented in the LSTM component of the model. As demonstrated in the figure, after determining that the time step at $t + 1$ had the most significant impact on model performance, we applied a perturbation-based method to the humancentered graph at this specific time step, which involved singly removing each edge connected to the *Driver* node and reintroducing the resulting graphs to the GAT-LSTM for prediction. By comparing the outputs from the GAT-LSTM on various perturbed graphs, we found that the *Truck*, visually estimated by Dist-YOLOv5 to be 2.22 meters away, exerted the greatest impact, potentially increasing the risk of traffic accidents on I-70 Bus.

To quantitatively study how such an explanation contributes to our understanding of the correlation between urban objects and traffic accidents, we iterated the process for those roads marked as *High Risk* in Denver and investigated which leading urban objects in vision may lead to a higher chance of traffic accidents on the road, and the results are visualized in Figure 8. As the figure indicates, the presence of cars and trucks on these roads is a primary factor contributing to traffic accidents, which is a result that aligns with expectations (Vanlaar and Yannis, 2006; Dadashova et al., 2016; Khorashadi et al., 2005). Interestingly, in the city center of Denver, the visibility of pedestrians correlates strongly with high accident risks. Such a pattern, on the one hand, indicates a non-vehicle activity mode of the urban residents clustered in the city center, which aligned with Denver city's recent efforts (e.g., Pedestrian Program

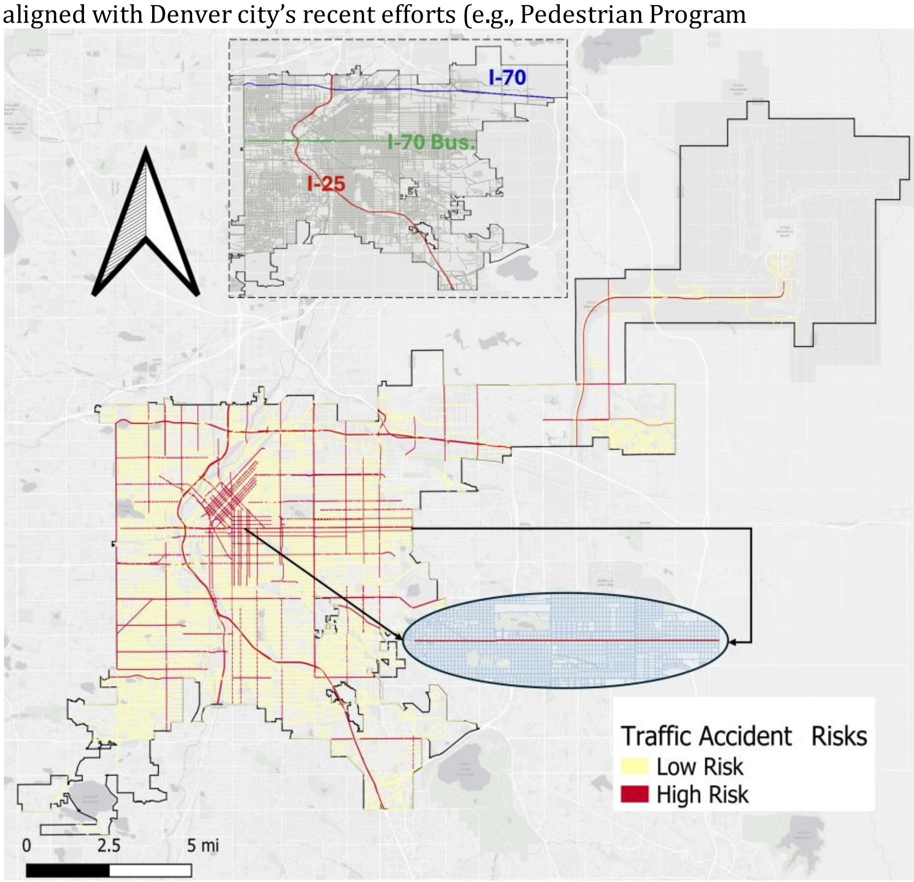

Figure 6: The distribution of *High Risk* and *Low Risk* roads in Denver, and the selected road for the demonstration. Map sources: Esri, HERE, Garmin, INCREMENTP, © OpenStreetMap contributors, and the GIS user community.

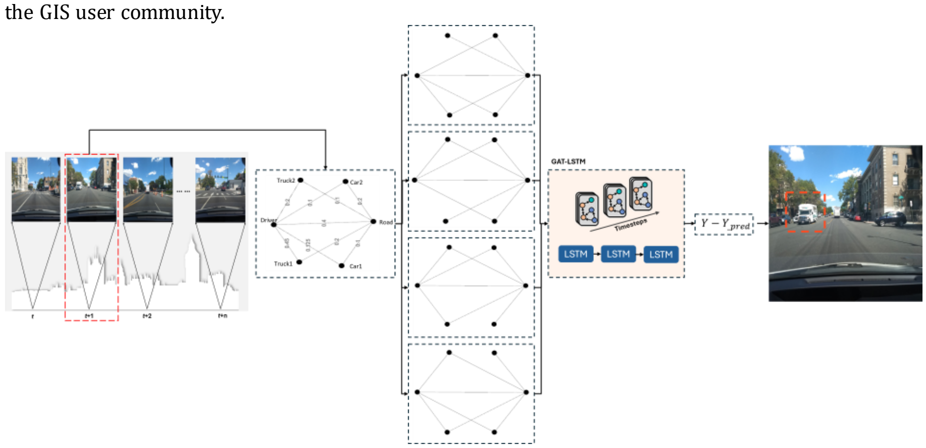

Figure 7: An illustration of the two-step XAI for the analysis of U.S. Road I-70 Bus. Image contains SVI collected from Mapillary, under the Creative Commons (CC) licenses.

(City and County of Denver, 2018)) in promoting walking as a primary mobility mode in the city center; on the other hand, the correlation between such a high risk of traffic accidents and pedestrians suggests that despite Denver's substantial investment in the Vision Zero initiative (City and County of Denver, 2022), which aims to decrease traffic accidents by reducing vehicular traffic, there remains considerable progress to be made towards achieving the goal of zero traffic incidents (Ferenchak, 2023).

6. Discussion and Conclusion

This paper presents a human-centered GeoAI framework to investigate the association between urban street-level environments, as captured by crowdsourced SVI, and traffic accident frequency. Central to this framework is the integration of a GAT-LSTM model, which combines graph-based spatial reasoning with temporal sequence modeling. This hybrid architecture enables the model to capture both non-linear relationships among urban objects and the temporal evolution of visual environments observed through sequential SVI. Importantly, the model's predictive capability does not rely on temporal alignment with individual accident events. Rather, it leverages temporally ordered imagery to identify persistent or slowly evolving visual patterns that characterize high-risk urban environments. The GAT component assigns adaptive weights to different urban objects, enabling the model to attend to more influential features within each scene, whereas the LSTM component captures temporal dependencies among these features, reflecting the continuity of urban exposure. Together, these components provide a robust mechanism for modeling structural risk embedded in the built environment.

Building upon prior work on human-centered GeoAI (Liu et al., 2023c), this study advances graph construction by incorporating distance-based spatial ap-

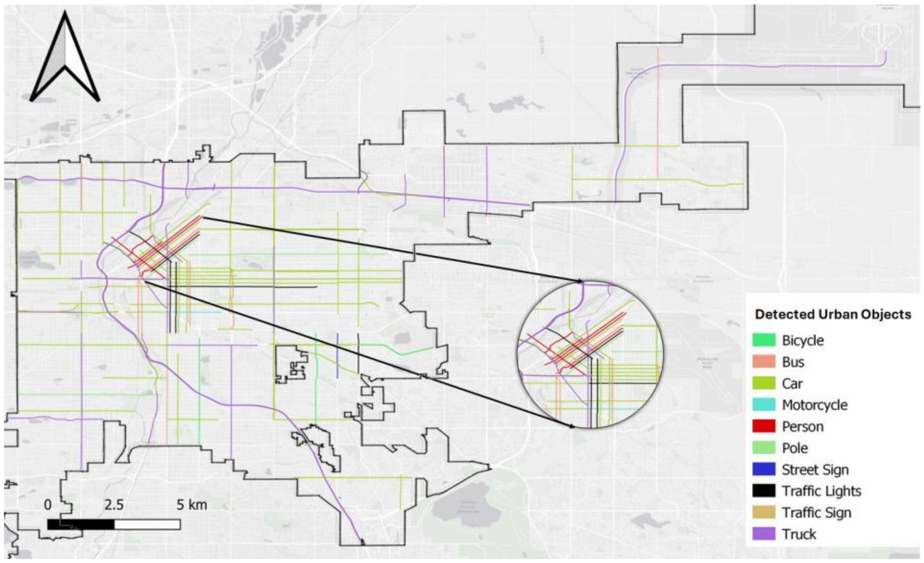

Figure 8: A quantitative understanding of the XAI results in identifying which urban objects from the drivers' viewpoints may correlate to a high risk of urban traffic accidents. Map sources: Esri, HERE, Garmin, INCREMENTP, © OpenStreetMap contributors, and the GIS user community.

proximation between the camera viewpoint and detected urban objects. This design aims to approximate perceptual salience, whereby objects closer to the driver are more likely to influence attention and decision-making. The temporally varying graph structure further enables the representation of dynamic yet structured urban environments. Nevertheless, it is important to emphasize that the proposed framework is not designed to model real-time accident occurrence driven by short-term factors such as weather, traffic flow, or driver behavior. Instead, it focuses on identifying structural associations between stable environmental characteristics—such as visual complexity, object density, and spatial configuration—and long-term accident frequency. As such, the model should be interpreted as capturing persistent exposure to spatial risk rather than causal mechanisms of individual accident events.

From a planning perspective, the proposed framework offers practical value by translating complex model outputs into interpretable insights through the two-step explainable AI (XAI) approach. By identifying both the most influential temporal contexts and the specific urban objects contributing to higher predicted accident frequencies, the framework enables a more granular understanding of spatial risk factors. For instance, environments characterized by high visual complexity or dense clustering of roadside elements may indicate increased cognitive load or reduced visibility for drivers. Such findings can inform targeted interventions, including simplifying signage, optimizing roadside vegetation placement, and redesigning street layouts to improve sightliness and reduce visual clutter. Crucially, the perturbation-based XAI approach allows practitioners to move beyond aggregate correlations and identify context-specific contributors to risk, supporting more precise and evidence-based urban design strategies. In this sense, the framework provides a bridge between data-driven modeling and actionable planning insights.

The use of crowdsourced SVI represents another key contribution of this study. By relying on platforms such as Mapillary, the framework demonstrates the feasibility of utilizing volunteered geographic information as a cost-effective alternative to commercial data sources. This aligns with the broader paradigm of "humans as sensors" (Goodchild, 2007), in which citizen-contributed data enrich urban analytics with diverse, spatially extensive observations. While such data sources inherently introduce variability in quality and coverage, our results indicate that, when combined with appropriate data cleaning and model design, they can support reliable predictive modeling. Furthermore, although the distance estimates derived from Dist-YOLOv5 are approximate due to the limitations of monocular vision, they effectively capture relative spatial relationships between objects. As demonstrated by the sensitivity analysis, the model remains robust to perturbations of these estimates, suggesting that relative proximity—rather than absolute accuracy—is sufficient to represent perceptual structure in the graph.

The temporal design of the framework also supports analysis beyond specific observation periods. By modeling sequences of SVI as representations of stable or gradually evolving environmental conditions, the approach captures enduring characteristics of the urban landscape rather than transient events. This indicates that the learned relationships are not confined to a single time period but can support analysis across different temporal contexts. Consequently, the framework is particularly well-suited for informing long-term planning and policy interventions aimed at improving traffic safety, rather than for real-time prediction or operational traffic management.

Despite these contributions, several limitations should be acknowledged. First, the spatial distribution of available SVI introduces sampling bias, with well-mapped roads being overrepresented relative to smaller residential streets. While the framework is designed to operate under realistic data availability constraints, future work should incorporate targeted data collection or augmentation strategies to improve representation across diverse road typologies. Second, although the distance estimation approach is sufficient for capturing relative spatial relationships, integrating more advanced depth estimation techniques may further enhance spatial accuracy. Third, the current study is limited to a single case study city; extending the framework to multiple cities with varying urban forms and traffic dynamics would enable a more comprehensive assessment of its transferability. Finally, while this study focuses on structural environmental factors, integrating additional data sources, such as traffic volume or weather conditions, could further enrich the modeling of urban safety dynamics.

Future research will build upon these directions. In particular, advances in multi-source urban sensing and hierarchical semantic graph modeling offer promising opportunities to extend the current framework beyond driver-centric representations towards more comprehensive urban functional modeling.

Additionally, incorporating uncertainty quantification methods, such as Bayesian graph neural networks (Liu et al., 2025b) or prediction intervals, would allow for a more robust assessment of model reliability and support risk-informed decision-making. Improvements in computational efficiency and the integration of streaming data may also enable real-time or near-real-time applications, expanding the framework's utility in dynamic traffic management contexts.

In conclusion, this study demonstrates that a human-centered, graph-based GeoAI approach can effectively uncover meaningful associations between streetlevel environmental characteristics and traffic accident frequency. By combining crowdsourced SVI, spatial-temporal graph modeling, and explainable AI, the proposed framework not only advances methodological development in urban analytics but also provides actionable insights for improving traffic safety through built environment interventions.

Data Availability Statement

Source code is available on GitHub at https://github.com/PengyuanLiu1993/SensingUrbanTraffic.

Disclosure Statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

M. Abdel-Aty, J. Ugan, and Z. Islam, "Exploring the Influence of Drivers' Visual Surroundings on Speeding Behavior", *Accident Analysis & Prevention*, 198 (2024), 107479.

M. M. Abdelrahman, A. Chong, and C. Miller, "Personal Thermal Comfort Models Using Digital Twins: Preference Prediction with Bim-Extracted Spatial–Temporal Proximity Data from Build2vec", *Building and Environment*, 207 (2022), 108532.

M. Batty, "Cities as Complex Systems: Scaling, Interaction, Networks, Dynamics and Urban Morphologies.", (2009).

F. Biljecki, "GeoAI for Urban Sensing", in , *Handbook of Geospatial Artificial Intelligence*, (CRC Press, 2023).

F. Biljecki and K. Ito, "Street View Imagery in Urban Analytics and GIS: A Review", *Landscape and Urban Planning*, 215 (2021), 104217.

F. Biljecki, T. Zhao, X. Liang, and Y. Hou, "Sensitivity of Measuring the Urban Form and Greenery Using Street-Level Imagery: A Comparative Study of Approaches and Visual Perspectives", *International Journal of Applied Earth Observation and Geoinformation*, 122 (2023), 103385.

L. Breiman, "Random Forests", *Machine learning*, 45 (2001), 5–32.

J. Bridle, "Training Stochastic Model Recognition Algorithms as Networks Can Lead to Maximum Mutual Information Estimation of Parameters", *Advances in neural information processing systems*, 2 (1989).

H. Cai, V. W. Zheng, and K. C.-C. Chang, "A Comprehensive Survey of Graph Embedding: Problems, Techniques, and Applications", *IEEE transactions on knowledge and data engineering*, 30: 9 (2018), 1616–1637.

Q. Cai, M. Abdel-Aty, O. Zheng, and Y. Wu, "Applying Machine Learning and Google Street View to Explore Effects of Drivers' Visual Environment on Traffic Safety", *Transportation research part C: emerging technologies*, 135 (2022), 103541.

L. Cheng, S. Chu, W. Zong, S. Li, J. Wu, and M. Li, "Use of Tencent Street View Imagery for Visual Perception of Streets", *ISPRS International Journal of Geo-Information*, 6: 9 (2017), 265.

L. Cheng, X. Chen, J. De Vos, X. Lai, and F. Witlox, "Applying a Random Forest Method Approach to Model Travel Mode Choice Behavior", *Travel behaviour and society*, 14 (2019), 1–10.

B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "MaskedAttention Mask Transformer for Universal Image Segmentation,", paper presented at Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (New Orleans, 2022).

City and County of Denver, Predestrain Program, (2018), https://www.denv ergov.org/Government/Agencies-Departments-Offices/Agencies-Dep artments-Offices-Directory/Department-of-Transportation-and-Inf rastructure/Programs-Services/Pedestrians Accessed 2024-04-26.

City and County of Denver, 2022 Vision Zero Action Plan, (2022), https: //www.denvergov.org/files/assets/public/vision-zero/documents/ denver-vision-zero-action-plan.pdf Accessed 2024-04-18.

P. Cui, M. Abdel-Aty, X. Yang, C. Wang, and Y. Yuan, "Quantifying Spatial Inequities in Traffic Injury Rates Through the Integration of Urban Road Network Measures and Social Vulnerability", *Accident Analysis & Prevention*, 211 (2025), 107916.

B. Dadashova, B. Arenas-Ram´ırez, J. Mira-McWilliams, and F. AparicioIzquierdo, "Methodological Development for Selection of Significant Predictors Explaining Fatal Road Accidents", *Accident Analysis & Prevention*, 90 (2016), 82–94.

K. Debattista, T. Bashford-Rogers, C. Harvey, B. Waterfield, and A. Chalmers, "Subjective Evaluation of High-Fidelity Virtual Environments for Driving Simulations", *IEEE Transactions on Human-Machine Systems*, 48: 1 (2017), 30–40.

N. Dogru and A. Subasi, “Traffic Accident Detection Using Random Forest Classifier,”, paper presented at 2018 15th learning and technology conference (L&T) (Jeddah, 2018).

R. Du, C. Chen, B. Yang, N. Lu, X. Guan, and X. Shen, “Effective Urban Traffic Monitoring by Vehicular Sensor Networks”, *IEEE transactions on Vehicular Technology*, 64: 1 (2014), 273–286.

M. Dwarampudi and N. Reddy, “Effects of Padding on Lstms and Cnns”, *arXiv preprint arXiv:1903.07288*, (2019).

J. Edquist, C. M. Rudin-Brown, and M. G. Lenn´e, “The Effects of on-Street Parking and Road Environment Visual Complexity on Travel Speed and Reaction Time”, *Accident Analysis & Prevention*, 45 (2012), 759–765.

V. Faure, R. Lobjois, and N. Benguigui, “The Effects of Driving Environment Complexity and Dual Tasking on Drivers’ Mental Workload and Eye Blink Behavior”, *Transportation research part F: traffic psychology and behaviour*, 40 (2016), 78–90.

N. N. Ferenchak, “US Vision Zero Cities: Modal Fatality Trends and Strategy Effectiveness”, *Transportation letters*, 15: 8 (2023), 957–968.

M. Fey and J. E. Lenssen, “Fast Graph Representation Learning with PyTorch Geometric,”, paper presented at ICLR Workshop on Representation Learning on Graphs and Manifolds (New Orleans, 2019).

P. Florio, T. Leduc, Y. Sutter, and R. Br´emond, “Visual Complexity of Urban Streetscapes: Human vs Computer Vision”, *Machine Vision and Applications*, 35: 1 (2023), 7.

K. Fukushima, “Visual Feature Extraction by a Multilayered Network of Analog Threshold Elements”, *IEEE Transactions on Systems Science and Cybernetics*, 5: 4 (1969), 322–333.

P. Gallo and H. Kettani, “On Privacy Issues with Google Street View”, *SDL Rev.*, 65 (2020), 608.

H. Ghahremannezhad, H. Shi, and C. Liu, “Object Detection in Traffic Videos: A Survey”, *IEEE Transactions on Intelligent Transportation Systems*, (2023).

A. R. Goetz and E. E. Boschmann, *Metropolitan Denver: Growth and Change in the Mile High City*, (University of Pennsylvania Press, 2018).

M. F. Goodchild, “Citizens as Sensors: The World of Volunteered Geography”, *GeoJournal*, 69 (2007), 211–221.

A. Graves and A. Graves, “Long Short-Term Memory”, *Supervised sequence labelling with recurrent neural networks*, (2012), 37–45.

F. Guan, Z. Fang, L. Wang, X. Zhang, H. Zhong, and H. Huang, “Modelling People’s Perceived Scene Complexity of Real-World Environments Using StreetView

Panoramas and Open Geodata", *ISPRS Journal of Photogrammetry and Remote Sensing*, 186 (2022), 315–331.

K. Hamad, R. Al-Ruzouq, W. Zeiada, S. Abu Dabous, and M. A. Khalil, "Predicting Incident Duration Using Random Forests", *Transportmetrica A: transport science*, 16: 3 (2020), 1269–1293.

M. Haraguchi, A. Nishino, A. Kodaka, M. Allaire, U. Lall, L. Kuei-Hsien, K. Onda, K. Tsubouchi, and N. Kohtake, "Human Mobility Data and Analysis for Urban Resilience: A Systematic Review", *Environment and Planning B: Urban Analytics and City Science*, 49: 5 (2022), 1507–1535.

M. Helbich, M. Danish, S. Labib, and B. Ricker, "To Use or Not to Use Proprietary Street View Images in (health and Place) Research? That Is the Question", *Health & Place*, 87 (2024), 103244.

B. L. Hills, "Vision, Visibility, and Perception in Driving", *Perception*, 9: 2 (1980), 183–216.

S. Hochreiter, "The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions", *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6: 02 (1998), 107–116.

Y. Hou and F. Biljecki, "A Comprehensive Framework for Evaluating the Quality of Street View Imagery", *International Journal of Applied Earth Observation and Geoinformation*, 115 (2022), 103094.

Y. Hou, M. Quintana, M. Khomiakov, W. Yap, J. Ouyang, K. Ito, Z. Wang, T. Zhao, and F. Biljecki, "Global Streetscapes—a Comprehensive Dataset of 10 Million Street-Level Images Across 688 Cities for Urban Science and Analytics", *ISPRS Journal of Photogrammetry and Remote Sensing*, 215 (2024), 216–238.

G. Hruthik, R. Tharun, S. C. S. V. Prasad, S. S. Swain, S. Srithar, and S. Aravinth, "Efficient Road Structural Design and Traffic Accident Analysis Using Supervised Learning Algorithms,", paper presented at 2023 International Conference on Computer Communication and Informatics (ICCCI) (Coimbatore, 2023).

C. Hu, S. Jia, F. Zhang, and X. Li, "A Saliency-Guided Street View Image Inpainting Framework for Efficient Last-Meters Wayfinding", *ISPRS Journal of Photogrammetry and Remote Sensing*, 195 (2023), 365–379.

S. Hu, H. Xing, W. Luo, L. Wu, Y. Xu, W. Huang, W. Liu, and T. Li, "Uncovering the Association Between Traffic Crashes and Street-Level BuiltEnvironment Features Using Street View Images", *International Journal of Geographical Information Science*, 37: 11 (2023), 2367–2391.

M. R. Ibrahim, J. Haworth, and T. Cheng, "Understanding Cities with Machine Eyes: A Review of Deep Computer Vision in Urban Analytics", *Cities*, 96 (2020), 102481.

A. Khorashadi, D. Niemeier, V. Shankar, and F. Mannering, "Differences in Rural and Urban Driver-Injury Severities in Accidents Involving Large-Trucks: An Exploratory Analysis", *Accident Analysis & Prevention*, 37: 5 (2005), 910–921.

D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization", *arXiv preprint arXiv:1412.6980*, (2014).

A.-L. Koehler, I. Koch, and S. Ladwig, "Investigating the Role of Visual and Corresponding Auditory Stimuli in Driving-Related Speed Perception", *Transportation Research Part F: Traffic Psychology and Behaviour*, 110 (2025), 1–14.

I. Lana, J. Del Ser, M. Velez, and E. I. Vlahogianni, "Road Traffic Forecasting: Recent Advances and New Challenges", *IEEE Intelligent Transportation Systems Magazine*, 10: 2 (2018), 93–109.

S. Law, R. Hasegawa, B. Paige, C. Russell, and A. Elliott, "Explaining Holistic Image Regressors and Classifiers in Urban Analytics with Plausible Counterfactuals", *International Journal of Geographical Information Science*, 37: 12 (2023), 2575–2596.

J. Lee, T. Yoon, S. Kwon, and J. Lee, "Model Evaluation for Forecasting Traffic Accident Severity in Rainy Seasons Using Machine Learning Algorithms: Seoul City Study", *Applied Sciences*, 10: 1 (2019), 129.

B. Lei, P. Liu, X. Liang, Y. Yan, and F. Biljecki, "Developing the Urban Comfort Index: Advancing Liveability Analytics with a Multidimensional Approach and Explainable Artificial Intelligence", *Sustainable Cities and Society*, 120 (2025), 106121.

Z. Li, "Extracting Spatial Effects from Machine Learning Model Using Local Interpretation Method: An Example of SHAP and XGBoost", *Computers, Environment and Urban Systems*, 96 (2022), 101845.

Z. Li, "Leveraging Explainable Artificial Intelligence and Big Trip Data to Understand Factors Influencing Willingness to Ridesharing", *Travel Behaviour and Society*, 31 (2023), 284–294.

W. Li and Z. Luo, "Research on Traffic Accident Risk Prediction Method Based on Spatial and Visual Semantics", *ISPRS International Journal of Geo-Information*, 12: 12 (2023), 496.

Y. Li, L. Peng, C. Wu, and J. Zhang, "Street View Imagery (SVI) in the Built Environment: A Theoretical and Systematic Review", *Buildings*, 12: 8 (2022), 1167.

X. Liang, T. Zhao, and F. Biljecki, "Revealing Spatio-Temporal Evolution of Urban Visual Environments with Street View Imagery", *Landscape and Urban Planning*, 237 (2023), 104802.

T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Doll´ar, and C. L. Zitnick, "Microsoft Coco: Common Objects in Context,", paper presented at Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13 (2014).

P. Liu and F. Biljecki, "A Review of Spatially-Explicit GeoAI Applications in Urban Geography", *International Journal of Applied Earth Observation and Geoinformation*, 112 (2022), 102936.

P. Liu and S. De Sabbata, "A Graph-Based Semi-Supervised Approach to Classification Learning in Digital Geographies", *Computers, Environment and Urban Systems*, 86 (2021), 101583.

Y. Liu and H. Wu, "Prediction of Road Traffic Congestion Based on Random Forest,", paper presented at 2017 10th International Symposium on Computational Intelligence and Design (ISCID) (Hangzhou, 2017).

D. Liu, Y. Jiang, R. Wang, and Y. Lu, "Establishing a Citywide Street Tree Inventory with Street View Images and Computer Vision Techniques", *Computers, Environment and Urban Systems*, 100 (2023), 101924.

L. Liu, Z. Gao, P. Luo, W. Duan, M. Hu, M. R. R. Mohd Arif Zainol, and M. H. Zawawi, "The Influence of Visual Landscapes on Road Traffic Safety: An Assessment Using Remote Sensing and Deep Learning", *Remote Sensing*, 15: 18 (2023), 4437.

P. Liu, T. Zhao, J. Luo, B. Lei, M. Frei, C. Miller, and F. Biljecki, "Towards Human-Centric Digital Twins: Leveraging Computer Vision and Graph Models to Predict Outdoor Comfort", *Sustainable Cities and Society*, 93 (2023), 104480.

Y. Liu, M. Chen, M. Wang, J. Huang, F. Thomas, K. Rahimi, and M. Mamouei, "An Interpretable Machine Learning Framework for Measuring Urban Perceptions from Panoramic Street View Images", *Iscience*, 26: 3 (2023).

P. Liu, Y. Zhang, and F. Biljecki, "Explainable Spatially Explicit Geospatial Artificial Intelligence in Urban Analytics", *Environment and Planning B: Urban Analytics and City Science*, (2024), 23998083231204689.

J. Liu, K. Shen, X. Liu, and C. Wu, "Unravelling the Multiple Effects of Multilevel Neighborhood Characteristics on Traffic Crash Risk from a Spatiotemporal Heterogeneity Perspective", *Travel Behaviour and Society*, 40 (2025), 101044.

P. Liu, Y. Chen, X. Liang, H. Li, F. Biljecki, and R. Stouffs, "A Graph Neural Network for Small-Area Estimation: Integrating Spatial Regularisation, Heterogeneous Spatial Units, and Bayesian Inference", *International Journal of Geographical Information Science*, (2025), 1–39.

P. Liu, B. Lei, W. Huang, F. Biljecki, Y. Wang, S. Li, and R. Stouffs, "Sensing Climate Justice: A Multi-Hyper Graph Approach for Classifying Urban Heat and Flood Vulnerability Through Street View Imagery", *Sustainable Cities and Society*, 118 (2025), 106016.

G. Lyons and K. Chatterjee, “A Human Perspective on the Daily Commute:Costs, Benefits and Trade-Offs”, *Transport reviews*, 28: 2 (2008), 181–198.

H. Ma, Y. Zhang, P. Liu, F. Zhang, and P. Zhu, “How Does Spatial Struc-ture Affect Psychological Restoration? a Method Based on Graph Neural Networks and Street View Imagery”, *Landscape and Urban Planning*, 251 (2024), 105171.

C. Miao, X. Chen, and C. Zhang, “Perceived Built Environment and NonMotorist Crashes: An Exploration with Street View Imagery”, *Journal of Transport Geography*, 128 (2025), 104363.

G. Neuhold, T. Ollmann, S. Rota Bulo, and P. Kontschieder, “The Mapil-lary Vistas Dataset for Semantic Understanding of Street Scenes,”, paper presented at Proceedings of the IEEE international conference on computer vision (Venice, 2017).

H. Ning, Z. Li, X. Ye, S. Wang, W. Wang, and X. Huang, “Exploring the VerticalDimension of Street View Image Based on Deep Learning: A Case Study on Lowest Floor Elevation Estimation”, *International Journal of Geographical Information Science*, 36: 7 (2022), 1317–1342.

I. I. Otkovi´c, A. Deluka-Tiblja ˇs, and S. Surdonja, “Validation of the Calibra-ˇ tion Methodology of the Micro-Simulation Traffic Model”, *Transportation Research Procedia*, 45 (2020), 684–691.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., “Pytorch: An Imperative Style, High-Performance Deep Learning Library”, *Advances in neural information processing systems*, 32 (2019).

O. Prasolenko, O. Lobashov, and A. Galkin, “The Human Factor in Road Traffic City”, *International Journal of Automation, Control and Intelligent Systems*, 1: 3 (2015), 77–84.

J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection,”, paper presented at Proceedings of the IEEE conference on computer vision and pattern recognition (Las Vegas, 2016).

Z. Rezaei and H. Ebrahimpour-Komleh, “Prediction and Detection of Car Accidents in Video by Deep Learning,”, paper presented at 2021 5th International Conference on Pattern Recognition and Image Analysis (IPRIA) (2021).

C. M. Rudin-Brown, J. Edquist, and M. G. Lenn´e, “Effects of Driving Experience and Sensation-Seeking on Drivers’ Adaptation to Road Environment Complexity”, *Safety science*, 62 (2014), 121–129.

A. M. Sainju and Z. Jiang, “Mapping Road Safety Features from Streetview Imagery: A Deep Learning Approach”, *ACM Transactions on Data Science*, 1: 3 (2020), 1–20.

N. Shahid, M. A. Shah, A. Khan, C. Maple, and G. Jeon, "Towards Greener Smart Cities and Road Traffic Forecasting Using Air Pollution Data", *Sustainable Cities and Society*, 72 (2021), 103062.

C. E. Shannon, "A Mathematical Theory of Communication", *ACM SIGMOBILE mobile computing and communications review*, 5: 1 (2001), 3–55.

A. W. Siddiqui, S. A. Raza, M. A. Elahi, K. S. Minhas, and F. M. Butt, "Temporal Impacts of Road Safety Interventions: A Structural-Shifts-Based Method for Road Accident Mortality Analysis", *Accident Analysis & Prevention*, 174 (2022), 106767.

T. Tanprasert, C. Siripanpornchana, N. Surasvadi, and S. Thajchayapong, "Recognizing Traffic Black Spots from Street View Images Using EnvironmentAware Image Processing and Neural Network", *IEEE Access*, 8 (2020), 121469–121478.

H. Tapiro, T. Oron-Gilad, and Y. Parmet, "Pedestrian Distraction: The Effects of Road Environment Complexity and Age on Pedestrian's Visual Attention and Crossing Behavior", *Journal of safety research*, 72 (2020), 101–109.

G. Underwood, "Visual Attention and the Transition from Novice to Advanced Driver", *Ergonomics*, 50: 8 (2007), 1235–1249.

M. Vajgl, P. Hurtik, and T. Nejezchleba, "Dist-YOLO: Fast Object Detection with Distance Estimation", *Applied Sciences*, 12: 3 (2022), 1354.

W. Vanlaar and G. Yannis, "Perception of Road Accident Causes", *Accident Analysis & Prevention*, 38: 1 (2006), 155–161.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L . Kaiser, and I. Polosukhin, "Attention Is All You Need", *Advances in neural information processing systems*, 30 (2017).

P. Veliˇckovi´c, G. Cucurull, A. Casanova, A. Romero, P. Li`o, and Y. Bengio, "Graph Attention Networks,", paper presented at International Conference on Learning Representations (Vancouver, 2018).

S. Wang, X. Huang, P. Liu, M. Zhang, F. Biljecki, T. Hu, X. Fu, L. Liu, X. Liu, R. Wang, et al., "Mapping the Landscape and Roadmap of Geospatial Artificial Intelligence (GeoAI) in Quantitative Human Geography: An Extensive Systematic Review", *International Journal of Applied Earth Observation and Geoinformation*, 128 (2024), 103734.

P. Wisutwattanasak, S. Jomnonkwao, C. Se, T. Champahom, and V. Ratanavaraha, "Correlated Random Parameters Model with Heterogeneity in Means for Analysis of Factors Affecting the Perceived Value of Road Accidents and Travel Time", *Accident Analysis & Prevention*, 183 (2023), 106992.

Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and S. Y. Philip, "A Comprehensive Survey on Graph Neural Networks", *IEEE transactions on neural networks and learning systems*, 32: 1 (2020), 4–24.

J. Xu, Q. Xiong, Y. Jing, L. Xing, R. An, Z. Tong, Y. Liu, and Y. Liu, "Understanding the Nonlinear Effects of the Street Canyon Characteristics on Human Perceptions with Street View Images", *Ecological Indicators*, 154 (2023), 110756.

C. Yang, L. Jiang, W. Dai, D. Peng, K. Deng, M. Zhao, X. Huang, and X. Chen, "An Operational Framework for Reconstructing Lane-Level Road Maps Using Open Access Data", *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, (2023).

W. Yap, J.-H. Chang, and F. Biljecki, "Incorporating Networks in Semantic Understanding of Streetscapes: Contextualising Active Mobility Decisions", *Environment and Planning B: Urban Analytics and City Science*, 50: 6 (2023), 1416–1437.

W. Yap, R. Stouffs, and F. Biljecki, "Urbanity: Automated Modelling andAnalysis of Multidimensional Networks in Cities", *npj Urban Sustainability*, 3: 45 (2023).

X. Ye, S. Li, W. Gong, X. Li, X. Li, B. Dadashova, W. Li, J. Du, and D. Wu,"Street View Imagery in Traffic Crash and Road Safety Analysis: A Review", *Applied Spatial Analysis and Policy*, 18: 2 (2025), 50.

X. Yuan, L. Li, Y. A. Shardt, Y. Wang, and C. Yang, "Deep Learning with Spatiotemporal Attention-Based LSTM for Industrial Soft Sensor Model Development", *IEEE Transactions on Industrial Electronics*, 68: 5 (2020), 4404–4414.

H. Yue, "Investigating the Influence of Streetscape Environmental Characteristics on Pedestrian Crashes at Intersections Using Street View Images and Explainable Machine Learning", *Accident Analysis & Prevention*, 205 (2024), 107693.

T. Zhang and X. Huang, "Monitoring of Urban Impervious Surfaces Using Time Series of High-Resolution Remote Sensing Images in Rapidly Urbanized Areas: A Case Study of Shenzhen", *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11: 8 (2018), 2692–2708.

M. Zhang, H. Fu, Y. Li, and S. Chen, "Understanding Urban Dynamics from Massive Mobile Traffic Data", *IEEE Transactions on Big Data*, 5: 2 (2017), 266–278.

F. Zhang, B. Zhou, L. Liu, Y. Liu, H. H. Fung, H. Lin, and C. Ratti, "Measuring Human Perceptions of a Large-Scale Urban Region Using Machine Learning", *Landscape and Urban Planning*, 180 (2018), 148–160.

C. Zhang, W. Ding, G. Peng, F. Fu, and W. Wang, "Street View Text Recognition with Deep Learning for Urban Scene Understanding in Intelligent Transportation

Systems”, *IEEE Transactions on Intelligent Transportation Systems*, 22: 7 (2020), 4727–4743.

Z. Zhang, Y. Zhang, T. He, and R. Xiao, “Urban Vitality and Its Influencing Factors: Comparative Analysis Based on Taxi Trajectory Data”, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15 (2022), 5102–5114.

Y. Zhang, P. Liu, and F. Biljecki, “Knowledge and Topology: A Two Layer Spatially Dependent Graph Neural Networks to Identify Urban Functions with Time-Series Street View Image”, *ISPRS Journal of Photogrammetry and Remote Sensing*, 198 (2023), 153–168.

Y. Zhang, T. Zhao, S. Gao, and M. Raubal, “Incorporating Multimodal Context Information into Traffic Speed Forecasting Through Graph Deep Learning”, *International Journal of Geographical Information Science*, 37: 9 (2023), 1909–1935.

J. Zhang, Y. Li, T. Fukuda, and B. Wang, “Urban Safety Perception Assessments via Integrating Multimodal Large Language Models with Street View Images”, *Cities*, 165 (2025), 106122.

L. Zhao, Y. Song, C. Zhang, Y. Liu, P. Wang, T. Lin, M. Deng, and H. Li, “T-Gcn: A Temporal Graph Convolutional Network for Traffic Prediction”, *IEEE transactions on intelligent transportation systems*, 21: 9 (2019), 3848– 3858.

J. Zheng, J. Wang, Z. Lai, C. Wang, and H. Zhang, “A Deep Spatiotemporal Network for Forecasting the Risk of Traffic Accidents in Low-Risk Regions”, *Neural Computing and Applications*, 35: 7 (2023), 5207–5220.

Z. Zhou, Y. Wang, X. Xie, L. Chen, and C. Zhu, “Foresee Urban Sparse Traffic Accidents: A Spatiotemporal Multi-Granularity Perspective”, *IEEE Transactions on Knowledge and Data Engineering*, 34: 8 (2020), 3786–3799.

H. Zhou, J. Wang, K. Wilson, M. Widener, D. Y. Wu, and E. Xu, “Using Street View Imagery and Localized Crowdsourcing Survey to Model Perceived Safety of the Visual Built Environment by Gender”, *International Journal of Applied Earth Observation and Geoinformation*, 139 (2025), 104421.