IEEE Open Journal of Intelligent Transportation Systems

Received XX Month, XXXX; revised XX Month, XXXX; accepted XX Month, XXXX; Date of publication XX Month, XXXX; date of current version XX Month, XXXX.

# Multimodal Learning for Traffic Risk Prediction: Combining Aerial Imagery with Contextual Data

**Hanlin Tian*, Yuxiang Feng *, Mohammed Quddus*, Yiannis Demiris†(Senior Member, IEEE), AND Panagiotis Angeloudis* (Member, IEEE)**

[1]Centre for Transport Engineering and Modelling, Department of Civil and Environmental Engineering, Imperial College London, UK
[2]Personal Robotics Laboratory, Department of Electrical and Electronic Engineering, Imperial College London, UK

CORRESPONDING AUTHOR: Hanlin Tian (e-mail: h.tian22@imperial.ac.uk).

**ABSTRACT** Accurately predicting traffic risks at urban intersections is essential for improving road safety. While traditional models use data sources like road traffic conditions, geometry, and signals, they often miss the spatial interactions between road networks and buildings. This study introduces a multimodal deep learning framework that integrates aerial imagery, building footprint data, and traffic flow information to improve traffic risk prediction and better capture these complex relationships.

By leveraging datasets from OpenStreetMap, the UK Traffic Count, and high-resolution aerial imagery, our approach creates a comprehensive representation of the urban environment, capturing intricate spatial relationships between road networks, surrounding structures, and traffic conditions. Using DeepLabV3+ and UNet++ as baseline models, we demonstrate that combining building and traffic data enhances prediction accuracy compared to models relying solely on visual data.

Our results show that the DeepLabV3+ model, when incorporating both building and traffic data, achieves the highest Intersection over Union (IoU) score of 0.4052 and the lowest Root Mean Square Error (RMSE) of 0.0907. These findings underscore the effectiveness of a multimodal approach in traffic risk assessment, offering a more precise tool for urban planning and traffic management interventions. The code and data used in this study are available at https://github.com/zachtian/Multimodal-Learning-for-Traffic-Risk-Prediction.

**INDEX TERMS** Geographic information systems, Traffic crash prediction, Computer Vision

## I. INTRODUCTION

ROAD traffic accidents continue to pose a significant global public health challenge, resulting in approximately 1.19 million fatalities and 20 to 50 million non-fatal injuries annually, along with substantial economic costs [1]. According to the World Health Organization (WHO), these accidents account for nearly 3% of the Gross Domestic Product (GDP) for most countries [1]. The causes of traffic accidents are complex and multifaceted, encompassing factors such as road conditions, traffic volume, weather, driver behaviour, and vehicle characteristics [2]–[6]. Predicting accident-prone areas is particularly challenging due to the dynamic interplay of these variables. Intersections, where vehicles, pedestrians, and cyclists frequently interact, are especially prone to collisions [7].

Traditional methods for traffic accident prediction often integrate data from multiple sources, including historical accident records, traffic volume, weather conditions, and road infrastructure information [8], [9]. While these methods provide valuable insights, they often rely on static analyses that may not fully capture the complex dynamics of urban environments. Additionally, many existing models do not consider the micro-level impacts of built environment features, such as building footprints, which can significantly influence accident risk by affecting visibility, traffic flow, and driver behaviour. Recent advancements, such as the work by He et al. [10], have leveraged satellite imagery and GPS trajectories to create high-resolution traffic accident risk maps. However, models that effectively integrate detailed spatial data, such as the layout and density of surrounding structures, remain underexplored. This study addresses these gaps by introducing a multimodal deep-learning framework that enhances traffic risk prediction by integrating aerial imagery, building footprint data, and traffic flow information. Additionally, this model offers a valuable solution for developing countries where historical accident data may be

limited or incomplete. By relying on more accessible data sources, such as satellite imagery and traffic flow data, our framework enables accurate accident risk prediction even in areas with sparse or unavailable accident records.

To address these gaps, we propose a novel multimodal learning framework that utilizes advanced models such as DeepLabV3+ and UNet++ while incorporating additional data layers, including satellite images, building information, and traffic flow data, to improve traffic risk prediction accuracy. Unlike the baseline versions of these models, which rely solely on aerial imagery, our approach integrates supplementary information on building structures and traffic flow. This fusion of multiple data types enables a more comprehensive analysis of traffic risk factors, potentially revealing patterns that remain hidden when using single-modality data. Figure 1 illustrates the task, where satellite images (left) are used with other data, such as building footprint and traffic flow, to predict traffic risk (right). The heatmaps show risk intensity, with brighter areas indicating higher risk.

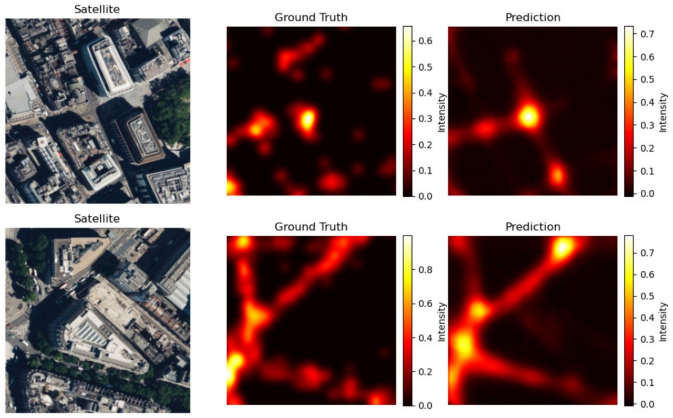


**FIGURE 1. Comparison of satellite images, ground truth, and predicted traffic risk heatmaps.**

Our approach aims to overcome the limitations of traditional models by dynamically incorporating diverse data sources, providing a more adaptive and precise method for traffic risk prediction.

The main contributions of our work are as follows:

- We propose a novel multimodal learning framework for traffic accident risk prediction that integrates diverse data sources, including aerial imagery, building footprints, and traffic flow data. This approach provides a comprehensive and context-aware understanding of the factors influencing traffic risk at intersections.
- We investigate how junction visibility, influenced by surrounding buildings and structures, affects accident risk, showing its potential for real-world applications in traffic management and urban planning by providing actionable insights for targeted safety interventions.
- Our approach achieves notable performance improvements over baseline models such as DeepLabV3+ and U-Net++ by incorporating these diverse data inputs, resulting in enhanced predictive accuracy for identifying high-risk zones at intersections.

Our paper is structured as follows:

In Section II, a review of studies on intersection-related accidents, contributing factors, the use of Geographic Information Systems (GIS) in accident analysis, and deep learning approaches for risk prediction is provided. Section III describes the data sources, collection methods, and preprocessing techniques for model training and evaluation. Section IV explains the methodologies, including the baseline models (DeepLabV3+ and UNet++), data fusion strategies, and evaluation metrics. Section V presents the experimental results, and model performance, and discusses the influence of different factors on accident risk. Section VI summarises the key findings, implications, and potential directions for future research.

## II. RELATED WORK

Research on traffic accident risks has extensively explored factors contributing to intersection-related accidents, the use of GIS for spatial analysis, and the application of deep learning techniques. This section reviews the key developments in these areas, highlighting the gaps our work addresses.

### *A. Intersection-Related Accidents and Contributing Factors*

Intersections merit spacial attention within the context of road safety analysis due to the high frequency of conflicts between road users, making them hotspots for accidents [11]. The complexity of interactions at these junctions, often exacerbated by obstructed sightlines, increases the likelihood of collisions [12]. Notably, intersections account for a substantial portion of pedestrian accidents, as highlighted by the National Highway Traffic Safety Administration [13].

Previous studies have investigated the factors contributing to intersection-related accidents, focusing on both human and environmental aspects. Human factors, such as driver behaviour and pedestrian activity, have been examined in relation to accident rates [2], [3]. However, these studies often rely on self-reported data, which can introduce biases.

Environmental factors, including road type, traffic volume, and intersection design, have been analysed using various quantitative methods [4], [14], [15]. While these studies provide valuable insights, they often overlook the impact of geographical and contextual factors that are less tangible but equally important in understanding accident risks at intersections.

### *B. Utilising Geographic Information Systems (GIS) in Accident Analysis*

Traditional accident studies often rely on macro-level datasets that cover extensive geographical areas, focusing on

accident frequency and severity over time. While many studies concentrate on specific regions, fewer have approached accident data analysis on a broader, macro-scale [16]–[20]. However, such analyses may encounter issues related to spatial autocorrelation, which challenges the assumption of independent observations across regions [21]. Therefore, incorporating spatial considerations is crucial in these investigations.

Geographic Information Systems (GIS) tools are increasingly utilised in road safety research for their ability to handle and analyse geospatial data effectively. Two primary applications of GIS in this field include the identification of accident hotspots and the geocoding of accidents for spatial analysis. Hotspot identification methods commonly used include kernel density estimation, nearest neighbour distances, and spatial indices like Moran's I. For instance, [22] employed the SANET toolkit to identify hotspot locations, while [23] used the Getis-Ord Gi* statistic to examine spatial autocorrelation in accident data.

The second key application of GIS, geocoding, involves the precise positioning of accident data for statistical analysis. Studies such as [17] have visualised geocoded road and accident datasets to link accidents with specific road segments. Similarly, [24] integrated traffic accident data with road network information to accurately locate accident occurrences. [25] organised road casualty data spatially in England, associating it with various land-use types within electoral districts.

### C. Deep Learning Approaches for Accident Risk Prediction

In recent years, deep learning has emerged as a powerful tool for predicting traffic accident risks. One of the most used deep learning architectures employed in this domain is the Convolutional Neural Network (CNN). CNNs are adept at processing grid-like data structures, such as images or spatial data, making them ideal for analyzing traffic data embedded within geographic contexts. For instance, CNNs have been used to analyse satellite imagery, road network layouts, and traffic volume maps to predict accident hotspots with high accuracy [10].

Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, have also been utilised to capture temporal dependencies in traffic data [26]. Another advanced technique is the use of Graph Neural Networks (GNNs), which are designed to handle data represented as graphs, such as road networks [8], [27].

Deep learning models for traffic accident risk prediction have the potential to integrate multiple data sources, including historical accident records, traffic volume, weather conditions, and road infrastructure information [28]. Lin et al. [8] proposed a novel variable selection method based on a frequent pattern tree to enhance real-time traffic accident risk prediction, while Ren et al. [9] introduced a deep learning framework that integrates heterogeneous data for citywide accident risk prediction.

In summary, while previous studies have contributed to our understanding of traffic accident risks at intersections, there has been limited exploration of integrating detailed spatial and dynamic traffic data at a micro-level. Our research addresses this gap by developing a deep learning framework that incorporates multiple data modalities, including building footprints, traffic volume, and road network data, to predict accident risks with higher accuracy and detail.

## III. DATA COLLECTION AND PREPARATION

This section outlines the data sources, extraction methods, and preprocessing techniques used to prepare these datasets for effective modelling and analysis.

### A. Data Sources

In this study, five primary datasets are used to analyse factors influencing road accidents at intersections: OpenStreetMap [29] for detailed road network and intersection information, Mapbox Satellite [30] for high-resolution aerial images, the Road Safety Data [31] for comprehensive historical accident records, the UK Traffic Count dataset [32] for traffic volume statistics, and the Building Geographic Dataset from OpenStreetMap, which provides architectural data for building dimensions surrounding road networks. Table 1 summarises the key elements extracted from each dataset.

#### 1) OpenStreetMap Data, Building Geographic Dataset, and Mapbox Satellite Imagery

OpenStreetMap (OSM) [29] serves as an extensive, collaboratively maintained mapping resource that provides free access to detailed geographic information worldwide. For our research, OSM offered data on the structural layout of road networks, including the precise geolocations of roads, classifications of road types, and intersection configurations. The Building Geographic Dataset, also derived from OSM, provided detailed spatial data on the built environment surrounding road intersections, including the coordinates, footprint areas, and heights of buildings [33]. Understanding the spatial distribution and dimensions of buildings is essential, as these structures can influence driver visibility and manoeuvrability at intersections, thus impacting accident likelihood and severity. An illustration of the Building Geographic Data is shown in Fig. 2.

To complement OSM data, we utilised Mapbox Satellite imagery [30] to obtain high-resolution aerial views of the study area. These satellite images provided essential visual context that enhanced our understanding of the surrounding environment's influence on traffic patterns and accident occurrences. Satellite imagery carries additional layers of information about road conditions, such as the number of lanes, the presence of road shoulders, pavement quality, and pedestrian crossings, which are not always available in vector

**FIGURE 2. Visual representation of the Building Geographic Data (red).**

data [10]. This contextual information enriched our analysis, helping us capture crucial details like lane markings, traffic density, and pedestrian movement. We used Folium [34], a Python library, to integrate and visualise the OSM data and Mapbox Satellite imagery.

#### 2) UK Traffic Count: Annual Average Daily Flow (AADF)

The UK Traffic Count dataset [32] offers Annual Average Daily Flow (AADF) statistics, which represent the estimated average number of vehicles traversing specific road segments each day. This dataset is important for understanding traffic volume and flow patterns, fundamental factors influencing the probability and frequency of accidents at intersections [35].

#### 3) Road Safety Data

We utilised the Department for Transport's Road Safety Data [31], covering the period from 1979 to mid-2023. This dataset includes comprehensive details such as the location (latitude/longitude), date, time, severity, vehicle types, and road conditions of accidents in the UK. For the purpose of this study, we exclusively used the geographic location (latitude and longitude) of each accident to focus on spatial risk patterns, particularly at intersections.

**TABLE 1. Summary of Datasets Used in the Study**

| Dataset | Key Elements |
|---|---|
| OpenStreetMap | Road networks, intersection details |
| Mapbox Satellite | High-resolution aerial images |
| Building Geographic Dataset | Building footprints |
| UK Road Accident Dataset | Accident locations, severity |
| UK Traffic Count | Annual Average Daily Flow statistics |

### *B. Data Processing*

We employed a multi-step data processing approach to extract meaningful features from the raw datasets. We focused on preparing the data for neural network training and generating detailed risk maps for traffic accident prediction. The process involved creating spatial representations of accident risk and traffic volume.

The first step in our analysis involved the extraction of all intersection points from the OpenStreetMap database, with the aim of identifying key areas of interest within the road network. For each intersection, a square bounding box was defined and centred. This bounding box had a fixed size of 0.002 degrees in both latitude and longitude, roughly equivalent to 222 meters in London. This specific size was selected to ensure that most accidents were captured without overlapping with nearby intersections, while still leaving sufficient space for data augmentation and image cropping during model training.

Due to differences in scale between $x$ (longitude) and $y$ (latitude) coordinates, we applied a correction factor to the y-coordinates to account for the Earth's curvature and projection distortions. This adjustment ensures that the bounding box maintains the correct proportions across different latitudes, compensating for the fact that distances between points of longitude decrease as you move away from the equator. The adjusted y-coordinates were scaled using the following equation:

$$y_{\text{scaled}} = \left( \frac{y - \min(y)}{\max(y) - \min(y)} \right) \times \left( \frac{\max(x) - \min(x)}{\cos(\text{latitude})} \right), \tag{1}$$

where $y_{\text{scaled}}$ is the adjusted y-coordinate, and latitude represents the center latitude of the bounding box.

#### 1) Accident Heatmap Creation

To represent accident risk, we created heatmaps by rasterizing accident points within a fixed bounding box around each intersection. Accident counts were mapped onto a uniform grid, and a log transformation was applied to manage the Poisson-like distribution of the accident data:

$$\text{Heatmap}_{\log}(x, y) = \log(1 + \text{Heatmap}(x, y)), \tag{2}$$

where $\text{Heatmap}(x, y)$ represents the raw accident count at each grid position $(x, y)$. The log transformation mitigates variance and normalizes the distribution, ensuring a more balanced representation of accident density. Gaussian smoothing with a parameter of $\sigma = 15$ was applied to produce a realistic spatial visualization of accident risk. Through empirical testing, we found that $\sigma = 15$ offers an optimal balance between visual clarity and smoothness, highlighting high-risk areas without excessive blurring of localized details (Fig. 3).

#### 2) Traffic Volume Rasterization

We calculated traffic volume as a key feature to evaluate the influence on accident risk. Traffic counts from the UK Traffic Count dataset were assigned to each road segment by finding the nearest traffic count point. These counts

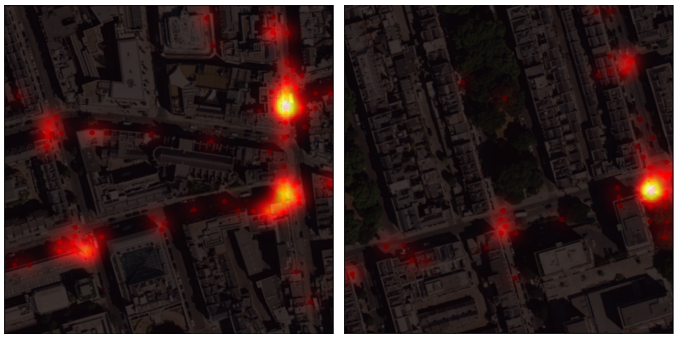

FIGURE 3. Visual representation of ground truth accident density using heatmaps, highlighting high-risk areas at urban intersections.

were normalised to create a traffic volume raster for each intersection, representing the intensity of traffic flow in the area, as depicted in Fig. 4.

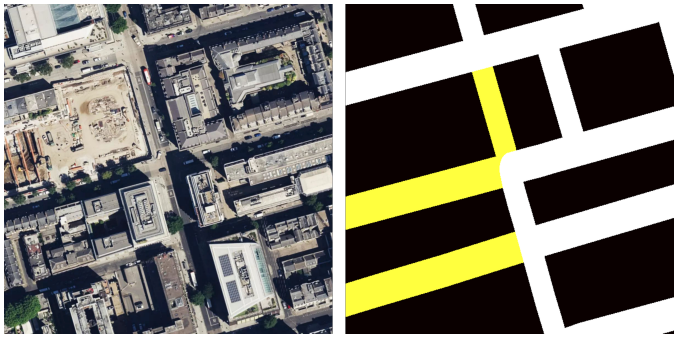

FIGURE 4. Comparison of GIS data (left) with corresponding traffic volume data (right), illustrating traffic flow on road segments.

### C. Experiment Setup

The experiment was conducted using data from London, UK, focusing on an area centred near Imperial College London at the coordinates (51.50212, -0.19123). This urban environment was chosen due to its high traffic volume, diverse intersection layouts, and varying road types, making it suitable for a comprehensive analysis of traffic risks. We collected geographical data within a 3000-meter radius around this central point for the training dataset. To ensure that the model generalizes well to new urban environments and does not overfit to a specific region, we partitioned the data into distinct training and testing sets, based on geographic separation.

The test set was defined by shifting the sampling area 4000 meters to the east of the initial radius, thus eliminating any overlap with the training data. This geographical split ensures a robust evaluation of the model's ability to perform in unseen urban contexts. This sampling method resulted in 5096 samples for training and 600 samples for testing, providing a broad dataset for evaluating model performance across different traffic scenarios and intersection types.

## IV. METHODOLOGY

In this section, we describe the methodologies used to develop and evaluate our proposed multimodal traffic risk prediction model. This includes the baseline models, data fusion strategies, and evaluation metrics.

### A. Baseline Models: DeepLabV3+ and UNet++

To establish a benchmark for our multimodal traffic risk prediction model, we employ two well-known segmentation models: DeepLabV3+ [36] and UNet++ [37].

#### 1) DeepLabV3+

DeepLabV3+ is a state-of-the-art semantic segmentation model that extends the DeepLabV3 architecture by integrating an encoder-decoder structure. The encoder employs dilated convolutions, which are effective in capturing contextual information at various scales without compromising spatial resolution, making it suitable for dense prediction tasks. The decoder module further enhances segmentation results by recovering spatial details lost during the encoding phase. This design allows DeepLabV3+ to extract meaningful features from high-resolution aerial imagery, which is crucial for applications such as road extraction and land cover mapping [36].

#### 2) UNet++

UNet++ is an advanced variant of the U-Net architecture, designed to improve segmentation accuracy by reducing the semantic gap between encoder and decoder feature maps. It introduces dense skip connections and nested convolutional blocks, which enhance gradient flow and enable the model to learn finer spatial details. This architecture is adept at capturing multi-scale contextual information, making it particularly effective for segmenting complex urban environments from aerial images [37].

Both DeepLabV3+ and UNet++ serve as baseline models that use aerial images as input to predict traffic risk. These models provide a reference point for evaluating the performance improvements achieved by our proposed multimodal approach.

### B. Multimodal Data Integration

The extracted building and traffic data are combined with aerial images by stacking them into multi-channel inputs. Specifically, the building data is represented as a 2D grid that aligns with the aerial imagery, where each tile contains the total floor area for the corresponding location. Similarly, traffic data is transformed into a grid format, where each tile represents the traffic volume for a particular area. These data layers are concatenated with the aerial image channels, resulting in a unified input consisting of the original 3-channel aerial imagery plus additional channels for building

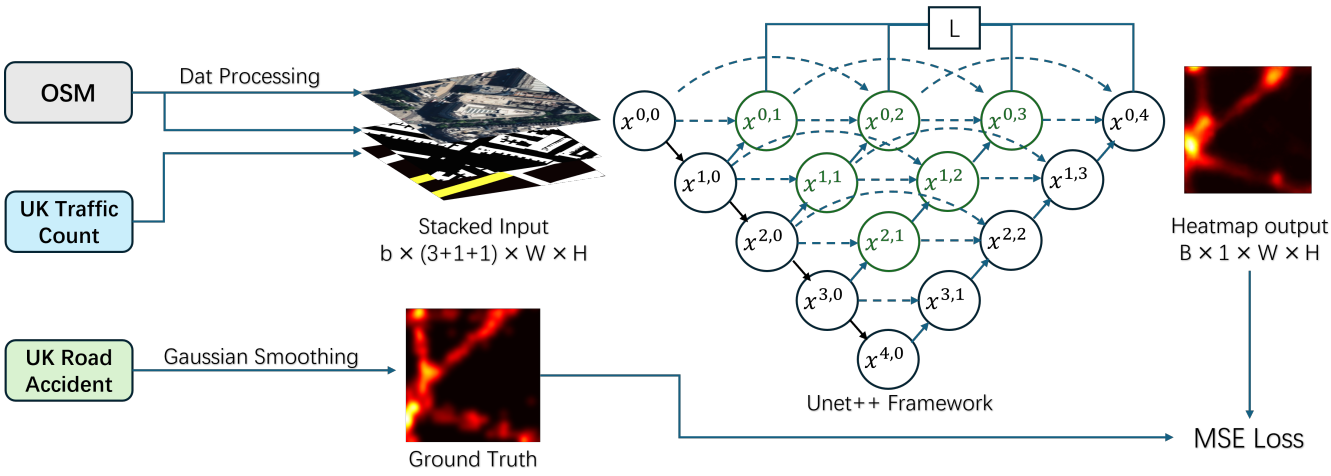


**FIGURE 5. Overview of the proposed framework for traffic risk prediction. The framework integrates multiple data sources, including OpenStreetMap (OSM) data, UK Traffic Count data, and UK Road Accident data, to create a comprehensive representation of the urban environment. Aerial imagery and building footprint data from OSM are combined with traffic flow information to form the input for the deep learning model.**

and traffic information. The general framework design is shown in Fig. 5.

Modifications were made to the input layer of both the DeepLabV3+ and UNet++ models to accommodate the multimodal input data. Traditionally, these models use a 3-channel input corresponding to the RGB channels of aerial images. However, in our proposed multimodal framework, we extended the input layer to accept 5 channels.

### C. Evaluation Metrics

To ensure a comprehensive evaluation of the model's effectiveness in both identifying high-risk areas and accurately predicting accident risk, we follow the approach of He et al. [10] by dividing the evaluation into two distinct tasks: classification and regression.

For the classification task, our objective is to accurately identify high-risk zones within the predicted heatmaps. To achieve this, we utilise the Intersection over Union (IoU) metric, a standard measure of overlap between the predicted high-risk regions and the actual high-risk regions. We define a high-risk zone as any area where the predicted or actual risk score exceeds the 90th percentile. Both the predicted and ground truth heatmaps are thresholded at this 90% level to classify zones as high risk.

The IoU is computed as:

$$\text{IoU} = \frac{|P_{hr} \cap G_{hr}|}{|P_{hr} \cup G_{hr}|} \tag{3}$$

where:

- $P_{hr}$ denotes the set of predicted high-risk zones.
- $G_{hr}$ denotes the set of ground truth high-risk zones.
- $|P_{hr} \cap G_{hr}|$ represents the cardinality of the intersection of these sets (i.e., the number of pixels correctly identified as high risk).
- $|P_{hr} \cup G_{hr}|$ represents the cardinality of the union of these sets (i.e., the total number of pixels classified as high risk in either set).

This metric provides a robust measure of the model's accuracy in identifying high-risk zones, as it penalises both false positives (incorrectly identified high-risk areas) and false negatives (missed high-risk areas).

For the **regression task**, where the objective is to predict the risk value in each tile and time window, we use the **Mean Square Error (MSE)** to assess the model's accuracy. The MSE captures the discrepancy between the predicted and actual risk values (normalized between 0 and 1). The **MSE** is defined as:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(r_i - \hat{r}_i)^2 \tag{4}$$

where:

- $n$ is the total number of tiles.
- $r_i$ is the actual normalised risk value (between 0 and 1) in tile $i$.
- $\hat{r}_i$ is the predicted normalised risk value (between 0 and 1) in tile $i$.

The **Root Mean Square Error (RMSE)**, derived from MSE, provides a measure of the model's overall predictive performance across all tiles. RMSE is calculated as:

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(r_i - \hat{r}_i)^2} \tag{5}$$

## V. RESULTS AND DISCUSSION

In this section, we present the results of our experiments, comparing the performance of the proposed multimodal

approach with baseline models. We analyse the impact of integrating different data modalities on the model's ability to predict traffic risk accurately and discuss the implications of these findings for urban traffic management.

### A. Training Details

The model was trained to minimize the Mean Squared Error (MSE) loss, effectively capturing the discrepancy between the predicted and actual accident heatmaps. We used the Adam optimizer [38] with an initial learning rate of $1 \times 10^{-4}$, leveraging its adaptive learning rate capabilities for improved convergence. To enhance the model's robustness, data augmentation techniques, such as random rotation, flipping, and cropping, were applied during training to diversify the input data.

Each model was trained for 300 epochs, and the best-performing model in terms of RMSE was selected for comparison. The experiments were conducted on a system equipped with an Intel(R) Xeon(R) w5-2455X processor, an NVIDIA A6000 graphics card, and 256GB of RAM.

### B. Baseline Comparison

We compare our proposed multimodal approach to two widely recognised baseline models: DeepLabV3Plus [36] and UNetPlusPlus [37]. For this study, the baselines were tested with three different input configurations:

- **Aerial Images Only**: Using just the aerial images to predict traffic risk.
- **Aerial Images with Building Data**: Combining aerial images with building footprint data to analyse the influence of surrounding structures on accident risks.
- **Aerial Images with Building and Traffic Data**: Integrating aerial images with both building footprint and traffic flow data.

The performance of each model configuration is evaluated using two key metrics: Root Mean Square Error (RMSE) and Intersection over Union (IOU).

**TABLE 2. Performance comparison of different models for traffic risk prediction. RMSE (↓) indicates the Root Mean Square Error, where lower values are better, and IOU (↑) represents the Intersection over Union, where higher values are preferred.**

| Model Name | RMSE (↓) | IOU (↑) |
|---|---|---|
| Unet++ | .1018 | .3281 |
| Unet++ w/ building | .0944 | .3717 |
| Unet++ w/ building and traffic | **.0900** | .4035 |
| DeepLabV3+ | .0999 | .3406 |
| DeepLabV3+ w/ building | .0939 | .3767 |
| DeepLabV3+ w/ building and traffic | .0907 | **.4052** |

### C. Results Analysis

The results, presented in Table 2, highlight the impact of incorporating additional data modalities on the performance of both baseline models.

Both the **UNet++** and **DeepLabV3+** models exhibit a consistent trend where incorporating additional data modalities leads to lower RMSE values and higher IOU scores. This trend confirms that the integration of diverse data types improves the model's ability to predict traffic risks accurately and localise high-risk zones.

For the **UNet++** model, the RMSE decreases from 0.1018 to 0.0900 when both building and traffic data are integrated, indicating greater accuracy in predicting the number of accidents. Similarly, the IOU metric increases from 0.3281 to 0.4035, demonstrating enhanced capability in accurately identifying high-risk zones.

The **DeepLabV3+** model also shows improved performance with the inclusion of building and traffic data. The RMSE reduces from 0.0999 to 0.0907, and the IOU increases from 0.3406 to 0.4052, further validating the effectiveness of using multimodal data. Notably, the DeepLabV3+ model with both building and traffic data achieves the highest IOU of 0.4052, outperforming all other configurations in identifying high-risk areas.

As shown in Fig. 6, both models exhibit clear improvements in RMSE and IOU metrics as additional data modalities are incorporated. The bar and line plots provide a visual representation of these performance gains.

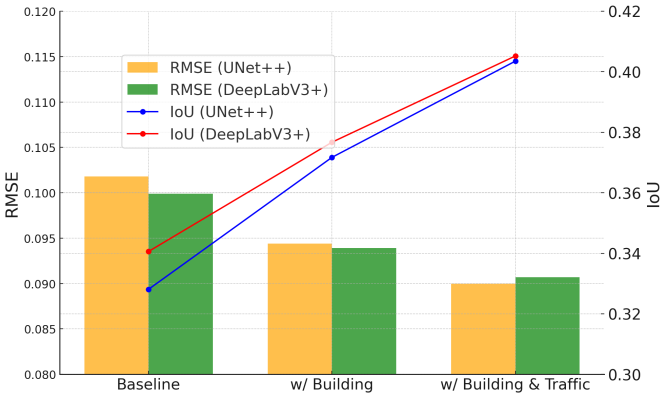


**FIGURE 6. Performance comparison of UNet++ and DeepLabV3+ models with different input configurations. The plot shows RMSE (bar plot) and IoU (line plot) metrics for each model configuration: baseline (aerial images only), with building data, and with both building and traffic data.**

### D. Heatmap Outputs Analysis

The heatmap outputs in Fig. 7 illustrate the predicted traffic risk for several urban intersections, with each column representing a different model configuration. The baseline DeepLabV3+ model, shown in the forth column, captures some high-risk zones but with considerable dispersion and false positives due to its reliance on aerial imagery alone.

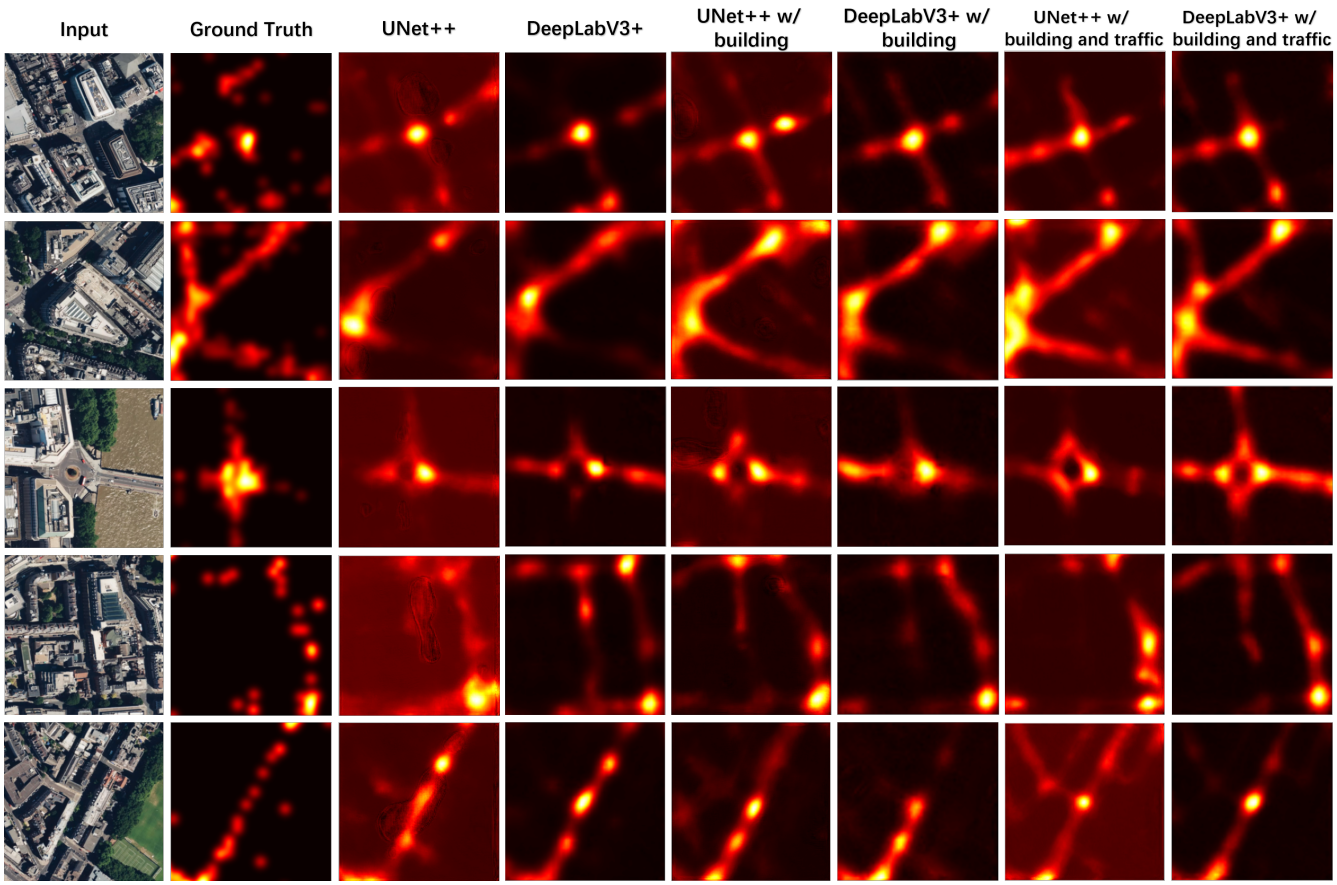


**FIGURE 7. Heatmap outputs of the predicted traffic risk for various urban intersections, with the leftmost column showing the original aerial images and subsequent columns representing different model configurations.**

In contrast, the DeepLabV3+ model with multimodal data exhibits notably improved performance by identifying more concentrated and accurate high-risk zones. It better reflects the real-world complexity of traffic risk factors.

Similarly, the UNet++ baseline model, shown in the third column, demonstrates the ability to detect accident hotspots, but its predictions are less specific and contain several inaccuracies. When enhanced with multimodal data, it also shows improved performance, with more distinct and well-defined high-risk areas. However, the multimodal DeepLabV3+ model outperforms the multimodal UNet++ model, providing sharper and more focused risk predictions.

These findings suggest that integrating diverse data sources like building footprints and traffic flow boosts the predictive power of both models and enables more accurate identification of high-risk zones, offering insights for targeted urban traffic management interventions.

## VI. CONCLUSIONS AND FUTURE WORK

This study demonstrates the effectiveness of a multimodal approach for predicting traffic accident risks at urban intersections by integrating aerial imagery, building footprints, and traffic flow data. Our experiments with the DeepLabV3+ and UNet++ models show that this combination of diverse data sources leads to notable improvements in model performance. The inclusion of building and traffic data alongside aerial images results in lower Root Mean Square Error (RMSE) and higher Intersection over Union (IOU) scores, indicating more accurate predictions and better identification of high-risk areas. These results underscore the value of a multimodal data fusion strategy in enhancing urban traffic safety assessments.

Future work could focus on developing specialized modules to independently process each data modality—such as aerial images, building data, and traffic flow—enabling the model to extract unique features from each source more effectively. Incorporating advanced techniques like cross-attention mechanisms could further enhance the model's ability to learn complex relationships between different data types, providing a deeper understanding of spatial and contextual factors contributing to traffic risks. Moreover, future research could explore the integration of additional data sources that influence accident risks, such as junction geometry, traffic signal timings, and right-turn traffic volumes.

Additionally, extending the model's application to various urban environments could help test its robustness and

adaptability across cities with different traffic conditions, road networks, and socio-economic contexts. This approach could also be highly beneficial for developing countries where historical accident data may be limited or unavailable. The multimodal framework provides a data-driven alternative for predicting accident risks using readily available satellite imagery, road infrastructure, and traffic data, making it a valuable tool for traffic safety improvements in regions with constrained resources.

## REFERENCES


[1] World Health Organization, "Road traffic injuries," 2022. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

[2] J. Lee, M. Abdel-Aty, and K. Choi, "Analysis of residence characteristics of at-fault drivers in traffic crashes," *Safety science*, vol. 68, pp. 6–13, 2014.

[3] M. Tomoda, H. Uno, S. Hashimoto, S. Yoshiki, and T. Ujihara, "Analysis on the impact of traffic safety measures on children's gaze behavior and their safety awareness at residential road intersections in japan," *Safety science*, vol. 150, p. 105706, 2022.

[4] H. Huang, H. C. Chin, and M. M. Haque, "Severity of driver injury and vehicle damage in traffic crashes at intersections: a bayesian hierarchical analysis," *Accident Analysis & Prevention*, vol. 40, no. 1, pp. 45–54, 2008.

[5] S. Useche, L. Montoro, F. Alonso, and O. Oviedo-Trespalacios, "Infrastructural and human factors affecting safety outcomes of cyclists," *Sustainability*, vol. 10, no. 2, p. 299, 2018.

[6] R. Elvik, A. Høye, T. Vaa, and M. Sørensen, *The handbook of road safety measures*. Emerald Group Publishing Limited, 2009.

[7] A. Laureshyn, Å. Svensson, and C. Hydén, "Evaluation of traffic safety, based on micro-level behavioural data: Theoretical framework and first implementation," *Accident Analysis & Prevention*, vol. 42, no. 6, pp. 1637–1646, 2010.

[8] L. Lin, Q. Wang, and A. W. Sadek, "A novel variable selection method based on frequent pattern tree for real-time traffic accident risk prediction," *Transportation Research Part C: Emerging Technologies*, vol. 55, pp. 444–459, 2015.

[9] H. Ren, Y. Song, J. Wang, Y. Hu, and J. Lei, "A deep learning approach to the citywide traffic accident risk prediction," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2018, pp. 3346–3351.

[10] S. He, M. A. Sadeghi, S. Chawla, M. Alizadeh, H. Balakrishnan, and S. Madden, "Inferring high-resolution traffic accident risk maps based on satellite imagery and gps trajectories," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 977–11 985.

[11] S. Lefèvre, C. Laugier, and J. Ibañez-Guzmàn, "Intention-aware risk estimation for general traffic situations, and application to intersection safety," Ph.D. dissertation, INRIA, 2013.

[12] S. S. C. Congress, A. J. Puppala, A. Banerjee, and U. D. Patil, "Identifying hazardous obstructions within an intersection using unmanned aerial data analysis," *International journal of transportation science and technology*, vol. 10, no. 1, pp. 34–48, 2021.

[13] National Highway Traffic Safety Administration, "Traffic safety facts," 2021. [Online]. Available: https://www.nhtsa.gov/

[14] K. Dixon, C. Monsere, R. Avelar, J. S. Barnett, P. Escobar, and S. M. Kothuri, "Improved safety performance functions for signalized intersections," 2015.

[15] L. Eboli, C. Forciniti, and G. Mazzulla, "Factors influencing accident severity: an analysis by road accident type," *Transportation research procedia*, vol. 47, pp. 449–456, 2020.

[16] S. V. Gomes, "The influence of the infrastructure characteristics in urban road accidents occurrence," *Accident Analysis & Prevention*, vol. 60, pp. 289–297, 2013.

[17] L. Ma, X. Yan, W. Qiao *et al.*, "A quasi-poisson approach on modeling accident hazard index for urban road segments," *Discrete dynamics in nature and society*, vol. 2014, 2014.

[18] L. Hu, X. Wu, J. Huang, Y. Peng, and W. Liu, "Investigation of clusters and injuries in pedestrian crashes using gis in changsha, china," *Safety science*, vol. 127, p. 104710, 2020.

[19] Z. Chen and W. D. Fan, "A multinomial logit model of pedestrian-vehicle crash severity in north carolina," *International journal of transportation science and technology*, vol. 8, no. 1, pp. 43–52, 2019.

[20] H. Blayney, H. Tian, H. Scott, N. Goldbeck, C. Stetson, and P. Angeloudis, "Bezier everywhere all at once: Learning drivable lanes as bezier graphs," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15 365–15 374.

[21] M. R. Dale and M.-J. Fortin, "Spatial autocorrelation and statistical tests in ecology," *Ecoscience*, vol. 9, no. 2, pp. 162–167, 2002.

[22] A. K. Al-Aamri, G. Hornby, L.-C. Zhang, A. A. Al-Maniri, and S. S. Padmadas, "Mapping road traffic crash hotspots using gis-based methods: A case study of muscat governorate in the sultanate of oman," *Spatial Statistics*, vol. 42, p. 100458, 2021.

[23] Z. Cheng, Z. Zu, and J. Lu, "Traffic crash evolution characteristic analysis and spatiotemporal hotspot identification of urban road intersections," *Sustainability*, vol. 11, no. 1, p. 160, 2018.

[24] B. P. Loo, "Validating crash locations for quantitative spatial analysis: a gis-based approach," *Accident Analysis & Prevention*, vol. 38, no. 5, pp. 879–886, 2006.

[25] R. B. Noland and M. A. Quddus, "A spatially disaggregate analysis of road casualties in england," *Accident Analysis & Prevention*, vol. 36, no. 6, pp. 973–984, 2004.

[26] P. Li, M. Abdel-Aty, and J. Yuan, "Real-time crash risk prediction on arterials based on lstm-cnn," *Accident Analysis & Prevention*, vol. 135, p. 105371, 2020.

[27] B. Huang and B. Hooi, "Traffic accident prediction using graph neural networks: New datasets and the travel model," *Traffic*, vol. 27, no. 29, p. 31, 2022.

[28] Y. Zhang, X. Dong, L. Shang, D. Zhang, and D. Wang, "A multi-modal graph neural network approach to traffic risk forecasting in smart urban sensing," in *2020 17th Annual IEEE international conference on sensing, communication, and networking (SECON)*. IEEE, 2020, pp. 1–9.

[29] M. Haklay and P. Weber, "Openstreetmap: User-generated street maps," *IEEE Pervasive computing*, vol. 7, no. 4, pp. 12–18, 2008.

[30] Mapbox, "Mapbox satellite imagery," https://www.mapbox.com/imagery.

[31] Department for Transport. (2022) Reported road casualties in great britain. [Online]. Available: https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data

[32] ——. (2022) Road traffic statistics. [Online]. Available: https://roadtraffic.dft.gov.uk/downloads

[33] F. Biljecki, Y. S. Chow, and K. Lee, "Quality of crowdsourced geospatial building information: A global assessment of openstreetmap attributes," *Building and Environment*, vol. 237, p. 110295, 2023.

[34] Filipe, "python-visualization/folium: v0.17.0," Jun 2024. [Online]. Available: https://zenodo.org/record/11840616

[35] A. Glushkov, V. Shepelev, A. Vorobyev, V. Mavrin, A. Marusin, and S. Evtykov, "Analysis of the intersection throughput at changes in the traffic flow structure," *Transportation Research Procedia*, vol. 57, pp. 192–199, 2021.

[36] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[37] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE transactions on medical imaging*, vol. 39, no. 6, pp. 1856–1867, 2019.

[38] D. P. Kingma, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

## BIOGRAPHIES

**Hanlin Tian** is a postgraduate researcher at the Centre for Transport Engineering and Modelling, Imperial College London. He received a BEng in Computer Science from Shandong University and an MSc in Computer Engineering from New York University. His main research interests include computer vision and autonomous vehicles.

**Yuxiang Feng** is a Research Associate and Lab Manager at the Centre for Transport Engineering and Modelling, Imperial College London. He received a BEng in Mechanical Engineering from Tongji University and an MSc in Mechatronics and PhD in Automotive Engineering from the University of Bath. His main research interests include environment perception, sensor fusion and artificial intelligence for robotics and autonomous vehicles.

**Mohammed Quddus** received the B.Sc. degree in civil engineering from Bangladesh University of Engineering and Technology in 1998, the master's degree in transportation engineering from the National University of Singapore in 2001, and the Ph.D. degree from Imperial College London in 2006. He joined the School of Architecture, Building and Civil Engineering, Loughborough University, U.K., in 2006, as a Lecturer, where he was a Professor of intelligent transport systems (ITS) in 2013. In 2021, he moved to Imperial College London as the Chair Professor of ITS. He has authored over 200 technical papers in international refereed journals and conference proceedings. His research interests include connected and autonomous vehicles, AI, and statistical modeling. He is an Associate Editor of Transportation Research—C: Emerging Technologies.

**Yiannis Demiris (SM'03)** received the B.Sc. (Hons.) degree in artificial intelligence and computer science and the Ph.D. degree in intelligent robotics from the Department of Artificial Intelligence, University of Edinburgh, Edinburgh, U.K., in 1994 and 1999, respectively. He is a Professor with the Department of Electrical and Electronic Engineering, Imperial College London, London, U.K., where he is the Royal Academy of Engineering Chair in Emerging Technologies, and the Head of the Personal Robotics Laboratory. His current research interests include human-robot interaction, machine learning, user modeling, and assistive robotics. Prof. Demiris is a Fellow of the Institution of Engineering and Technology (IET), and the British Computer Society (BCS).

**Panagiotis Angeloudis** is Reader and Head of the Transport Systems and Logistics Laboratory (TSL), based in the Centre for Transport Studies (CTS) at Imperial College London. Before establishing TSL, Panagiotis held a JSPS Research Fellowship at Kyoto University. He previously obtained a PhD in Transportation at Imperial College London and spent periods as a research analyst at DP World and the United Nations in Geneva. His research focuses on the study of networks, optimisation methods and multi-agent systems, as well as their applications in autonomous transport systems, urban infrastructure and logistics.