IET Intelligent Transport Systems

WILEY

The Institution of Engineering and Technology

ORIGINAL RESEARCH OPEN ACCESS

Enriched Pedestrian Crossing Prediction Using Carla Synthetic Data

Mohsen Azarmi[1] | Mahdi Rezaei[1] | He Wang[2]

[1]Institute for Transport Studies, University of Leeds, Leeds, UK | [2]AI Centre, Department of Computer Science, University College London, London, UK

Correspondence: Mahdi Rezaei (m.rezaei@leeds.ac.uk)

Received: 18 October 2024 | **Revised:** 13 September 2025 | **Accepted:** 7 October 2025

ABSTRACT

Pedestrian crossing prediction, which involves anticipating whether a pedestrian will cross the street or not, is a crucial function in autonomous driving systems. This is also a safety requirement for the interaction of highly automated vehicles and pedestrians. The endeavours in this research domain heavily rely on processing videos captured by the frontal cameras of autonomous vehicles using advanced computer vision techniques and deep learning methods. While recent studies focus on the model architecture for crossing prediction by utilising pre-trained visual feature extractors, they often encounter challenges stemming from inaccurate input features such as pedestrian body pose and/or scene semantic information. In this study, we aim to enhance pose estimation and semantic segmentation algorithms by using synthetic data augmentation (SDA) and domain randomisation (DR) techniques. SDA allows for automatic annotations through predefined agents and objects in a simulated urban environment. However, it creates a domain gap between synthetic and real-world data. To tackle this, we introduce a DR technique to generate synthetic data mimicking various weather and ambient illumination conditions. We evaluated two training strategies on six algorithms for both pose estimation and semantic segmentation algorithms, and ultimately, we target four deep learning architectures for crossing prediction, including convolutional, recurrent, graph, and transformer neural networks. The proposed technique improves the extraction of pedestrian body pose and categorical semantic information, which in turn enhances the state-of-the-art. This results in effective feature selection as the input for the PIP task, improving prediction accuracy by 3.2%, 4.2%, and 6.3% to reach 87.6%, 92.2%, and 73.6% against the JAAD, PIE, and FU-PIP datasets, respectively. The study indicates that using a simulated environment with structural randomised properties can enhance the resilience of the pedestrian crossing prediction to variations in the input data.

1 | Introduction

Pedestrian safety is a central concern in the development of autonomous vehicles (AVs), particularly at road crossings where close interaction between vehicles and pedestrians occurs. To ensure safe navigation, AVs must not only detect pedestrians but also anticipate whether they intend to cross the street. Accurate prediction of pedestrian crossing behaviour allows AVs to adjust speed and trajectory proactively, reducing abrupt manoeuvres, improving passenger comfort, and minimising collision risk [1].

In recent years, pedestrian crossing prediction has become a key research focus within intelligent transportation and computer vision. State-of-the-art approaches leverage deep learning and data-driven methods, building on datasets such as JAAD [2], PIE [3], STIP [4], and FU-PIP [5]. Despite significant progress, pedestrian crossing prediction remains a particularly difficult task for several reasons. **First**, pedestrian behaviour is inherently complex and influenced by multiple external and social factors, including traffic signals, nearby vehicles, and interactions with other pedestrians. **Second**, short observation windows,

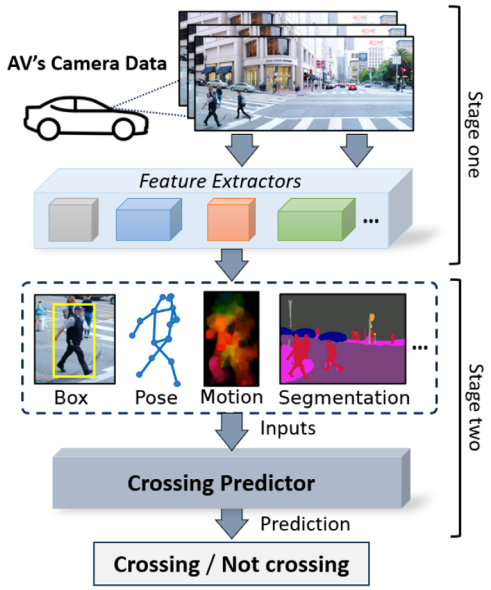

FIGURE 1 | An overview of the pedestrian crossing prediction framework. Both Stage one and two are trained separately.

occlusions, and variations in body posture or orientation make it challenging to infer intention reliably. **Third**, annotated datasets often lack sufficient diversity, limiting the ability of models to generalise across diverse urban environments. Together, these challenges make crossing prediction a complex task that requires models to capture both visual patterns and contextual decision-making cues.

Most crossing prediction models address these challenges through a two-stage design. In the first stage, pretrained feature extraction methods are applied to generate structured representations of visual input, such as pedestrian location, body pose, and semantic context [5]. In the second stage, these extracted features are passed to a predictive classifier, often a convolutional, recurrent, or transformer-based network, that estimates the likelihood of crossing [1, 6]. As illustrated in Figure 1, the feature extraction stage is typically trained independently, often using models pretrained on generic datasets such as COCO [7] for pose or Cityscapes [8] for semantic segmentation. Once trained, these extractors are used in a fixed (non-trainable) manner during the training of the crossing predictor, which functions as a binary classifier distinguishing between 'crossing' and 'not crossing'.

A notable limitation of this approach is that errors in feature extraction–caused, for instance, by slight inaccuracies in body pose estimation or scene segmentation—propagate into the downstream classifier, directly degrading crossing prediction accuracy. Yet, relatively little research has examined how improving the quality of feature extractors could enhance the robustness of crossing intention prediction.

Synthetic data provides one potential avenue to improve feature extraction by offering scalable, controllable, and automatically annotated samples. However, its use in crossing prediction has historically been limited due to several challenges. First, synthetic datasets often exhibit a domain gap relative to real-world data: differences in texture, illumination, and behavioural realism reduce transferability. Second, while high-fidelity simulations can generate photorealistic images, they are computationally expensive and still fail to capture subtle behavioural cues such as hesitation or gaze. Third, simulated pedestrians usually follow deterministic or rule-based scripts rather than exhibiting rich, context-dependent decision-making of real pedestrians. Consequently, models trained directly on synthetic data often perform poorly when deployed in real-world scenarios.

In this paper, we address these challenges by focusing on the feature extraction stage of pedestrian crossing prediction and proposing methods to improve its accuracy and resilience. We generate synthetic pedestrian crossing scenarios in the Carla simulator, enabling large-scale, automatically annotated data creation. To mitigate the domain gap, we adopt domain randomisation (DR)[9], systematically varying environmental factors such as weather, lighting, and traffic infrastructure. Additionally, we employ synthetic data augmentation (SDA)[10] to enrich data variability. Together, these techniques enable more robust training of body pose estimation (BPE) and object semantic segmentation (OSS) algorithms, providing higher-quality inputs for downstream crossing prediction models and thereby enhancing overall prediction performance.

The main contributions of this study are as follows:

- Generating diverse synthetic pedestrian crossing scenarios using Carla to augment real-world datasets with automatically annotated agents and objects;
- Applying domain randomisation (DR) to systematically vary environmental conditions (lighting, weather, traffic infrastructures), thereby bridging the domain gap between synthetic and real data;
- Evaluating two training strategies, hybrid (synthetic + real) and fine-tuned hybrid (synthetic pretraining followed by real fine-tuning), across three BPE and three OSS algorithms;
- Assessing the downstream impact of enriched feature extractors on six state-of-the-art crossing prediction models across the JAAD, PIE, and FU-PIP datasets.

The remainder of this paper is organised as follows. Section 2 provides a comprehensive background on pedestrian crossing prediction studies, highlighting the feature extractors employed in existing models. Section 3 details our proposed methodology for synthetic data generation, incorporating randomised parameters to develop candidate feature extractors. Section 4 presents the experimental results of applying enhanced feature extractors to candidate models for pedestrian crossing prediction. Finally, Section 5 concludes the paper with a summary of key findings and implications.

2 | Background

Understanding and predicting pedestrian behaviour is a long-standing challenge in autonomous driving research, where data-driven approaches have shown great promise [2, 3, 5]. We first review recent crossing prediction methods based on computer vision and machine learning, highlighting their strengths as well

as their inherent challenges. We then turn to data enhancement and domain adaptation techniques that have been developed to address the constraints of purely data-driven models and the limitations of existing real-world datasets.

2.1 | Pedestrian Crossing Prediction

In recent years, there has been a significant increase in the number of studies on pedestrian crossing prediction thanks to deep neural networks (DNNs), which demonstrate significant ability in learning complex patterns from multimodal data, such as video sequences, sensor information, and contextual cues. Unlike traditional methods such as probabilistic trajectory matching [11] and conditional random fields [12], which often struggle to model temporal dependencies and high-level features, DNNs are particularly effective in understanding pedestrian behaviour in dynamic environments. Researchers have been leveraging architectures like convolutional neural networks (CNNs) [13, 14], recurrent neural networks (RNNs) [2, 15], and graph convolutional networks (GCNs) [16, 17], as well as more recent advancements in attention mechanisms [18] and transformer models [19, 20], to enhance crossing prediction estimation accuracy.

While supervised deep learning methods such as DNNs have shown strong performance compare to unsupervised methods in pedestrian crossing prediction [21], they depend on high-quality, task-specific annotated data to effectively learn relevant representations [22]. Although several datasets such as JAAD [2], PIE [3], STIP [4], and FU-PIP [5] are available, many of them offer limited annotation diversity (e.g. lacking dense pose labels or fine-grained semantic segmentation), which restricts the training of fully end-to-end models that depend on rich visual features. As illustrated in Figure 1, recent crossing prediction models have two processing stages: feature extraction and crossing classification. Feature extraction consists of a group of computer vision algorithms, such as pedestrian detection [23], body pose estimation (BPE) [24], and object semantic segmentation (OSS) [25], which generate structured representations of the raw data by focusing on task-relevant features like pedestrian location, body posture, and semantic maps of surrounding objects. These extracted features are then fed into a crossing classifier head to predict the pedestrian's crossing action based on the spatial and temporal features identified during the feature extraction phase.

Studies have focused on selecting various features of pedestrians and/or the environment, which are often extracted through pre-trained algorithms or sometimes fine-tuned models, depending on the model design [26–28], and primarily contribute to designing the feature fusion and architecture of the crossing classifier head. In terms of feature selection, key elements such as the pedestrian's location and size using the pedestrian bounding box information [29], walking trajectory [30], body pose [31–34], and environmental features [35–38] have been studied. There has also been a study on feature importance for pedestrian crossing prediction [39] which experiments with input features that could potentially contribute more to enhancing prediction performance. They have shown that the pedestrian body pose feature has the least contribution, as it is influenced by the accuracy of the pose estimation algorithm.

2.2 | Data Enhancement and Domain Adaptation Techniques

The performance of pedestrian crossing prediction models is closely tied to the quality and precision of their input features, as even small inaccuracies in extracted features (as shown in Figure 1) can propagate through the pipeline, undermining downstream performance. However, it remains unclear to what extent improving the reliability of body pose estimation (BPE) and object semantic segmentation (OSS) as the main feature extractors of crossing prediction models can enhance overall prediction performance. To address this open question, prior research in related domains has explored various strategies for strengthening feature extractors by data enhancement and domain adaptation techniques.

Data enhancement techniques aim to improve the richness, diversity, and quality of training data, thereby supporting more robust and generalisable feature extractors. Standard approaches include data augmentation strategies such as geometric transformations (e.g. rotations, flipping, cropping) and photometric changes (e.g. brightness, contrast, noise) [40]. These methods increase variability in the training distribution and help prevent overfitting to narrow visual contexts. More advanced forms of augmentation include synthetic data augmentation (SDA) [10], in which new training samples are generated or existing ones are modified using simulation environments. For instance, pedestrian images can be programmatically altered to introduce background changes, occlusions, pose variations, or weather effects [41]. Simulation platforms like Carla [42] enable the creation of richly annotated, diverse synthetic datasets that include depth, segmentation, and pose information, useful for tasks like BPE and OSS. These synthetic enhancements are particularly valuable in scenarios where annotated real-world data is limited, expensive, or ethically constrained [43, 44].

Domain adaptation techniques aim to address distribution shifts between the source (e.g. synthetic) and target (e.g. real-world) domains. These shifts may arise due to differences in sensor modalities, lighting, weather, environments, or data collection conditions [45]. Among adaptation techniques, domain randomisation (DR) [9] is a widely adopted method to effectively align the synthetic and real-world data domains. It systematically introduces random variations into the synthetic data, such as changes in lighting, texture, or object positioning, helping models become more robust to variations when deployed in the real world. By exposing models to a wide range of simulated conditions, DR helps bridge the gap between synthetic and real-world data, enhancing the model's ability to handle unpredictable scenarios [46].

Domain randomisation has been applied in various research fields. For example, it has been used to enable DNNs to consider traffic context during vehicle detection [47]. Yue et al. [48], proposes a method using simulations to perform semantic segmentation in self-driving scenes without relying on data from the target domain. Hagelskjaer et al. [49] present a method for pose estimation that optimises configuration parameters using only synthetic data. Pasanisi et al. [50] investigated the feasibility of fine-tuning an object detection model in real industrial scenarios using a fully synthetic dataset.

Together, DR and SDA represent powerful complementary strategies for improving the generalisability of vision-based models [51]. While existing works [41, 43, 44] have leveraged synthetic data to enhance individual feature extractors, such as BPE and OSS, for detection and tracking tasks (e.g. MOTSynth [52]), the direct impact of domain-randomised synthetic data on the downstream crossing prediction task remains underexplored.

3 | Methodology

We aim to improve OSS and BPE modules via domain randomisation (DR) and synthetic data augmentation (SDA) and directly evaluates how it affects downstream pedestrian crossing prediction models performance in real-world urban crossing scenarios for autonomous driving systems. Figure 2 illustrates our methodology for enriching OSS and BPE algorithms. This section details steps 1 to 3, focusing on DR and data generation, followed by the candidate algorithms investigated for enhancement.

3.1 | Domain Randomisation

We propose a domain randomisation strategy to generate structurally consistent synthetic data with controlled variation in visual appearance. Specifically, we create multiple instances of each driving scenario that preserve fixed structural elements (e.g. road layout, pedestrian posture, and object positions) while randomising weather and lighting conditions (see the last three rows in Figure 3). This design aims to encourage BPE and OSS models to learn invariant representations of structural cues (such as pedestrian pose and road geometry), while disregarding irrelevant visual factors (e.g. rain streaks, sun glare, or shadow artefacts).

To formalise this, let us consider θ as a set of parameters that describe the global environmental conditions of the simulation, including illumination (e.g. day, night, sunset), weather (e.g. clear, cloudy, light or heavy rain), and surface conditions (e.g. dry, wet). For each training episode, values of θ are sampled from a predefined distribution $P(\theta)$, ensuring diverse but structurally consistent scenes. The synthetic dataset $\mathcal{D}' = \{x_1, x_2, \ldots, x_n\}$ is then generated by simulating the same task τ (e.g. OSS or BPE) under different instantiations of θ where each x_i is an input image sampled from the simulator and paired with a label y_i^τ.

The objective is to learn the set of parameters ϕ_τ for the task τ that minimises the expected loss across the distribution of environments, expressed as:

$$\phi_\tau^* = \arg\min_{\phi_\tau} \mathbb{E}_{\theta\sim P(\theta)}[\mathcal{L}_\tau(f_\theta^\tau(x_i, \phi_\tau), y_i^\tau)] \quad \forall\tau \in \{\text{OSS, BPE}\} \quad (1)$$

where ϕ_τ^* denotes the task-specific model trained under environment θ, and y_i^τ denotes the ground-truth label (segmentation mask for OSS, joint coordinates for BPE). Crucially, the ground-truth labels remain invariant across different θ values; for instance, the same pedestrian skeleton or segmentation mask applies whether the scene is rendered in sunlight, rain, or at night. This forces the model to focus on invariant structural cues, such as road geometry, pedestrian posture, or object layout, rather than overfitting to appearance-specific artefacts.

In the case of OSS, $y_i^{\text{OSS}} \in \mathbb{R}^{H\times W\times C}$ denotes the segmentation mask for the ith pedestrian sample, where H and W are the mask height and width (matching the input image size), and C is the number of semantic classes (e.g. $C = 19$ for Cityscapes [8]: road, sidewalk, building, car etc.). For BPE, $y_i^{\text{BPE}} \in \mathbb{R}^{J\times 2}$ contains the 2D coordinates (x, y) of the J body joints for the ith pedestrian (e.g. $J = 17$ for COCO [7]).

In this way, domain randomisation differs fundamentally from conventional image-level augmentations (e.g. flipping, cropping, or brightness adjustments). Instead of perturbing individual pixels, DR systematically alters the entire simulated environment, thereby encouraging robustness to broad contextual variations while preserving semantic consistency. Figure 3 illustrates this process, showing identical crossing scenarios generated under diverse weather and illumination conditions.

We also observed in preliminary experiments that simply generating additional synthetic samples under fixed conditions did not improve real-world performance. The key improvement stems from systematically varying environmental factors (θ), such as lighting and weather, which encourage models to learn invariant structural cues (e.g. road geometry, body pose) rather than overfitting to fixed visual appearances.

3.2 | Synthetic Data Generation

We propose a synthetic data generation pipeline using the Carla simulator to create diverse and automatically annotated pedestrian crossing scenarios. The synthetic data are generated to be used exclusively to train OSS and BPE models, which serve as spatial feature extractors in downstream pedestrian crossing prediction. Since these models operate at the frame level and do not capture temporal dependencies, temporally realistic behaviours are unnecessary. Hence, we acknowledge that our simulated agents lack the humanised intention behaviours of real pedestrians, and the downstream prediction model is aimed to be trained on real-world datasets to capture temporal dynamics

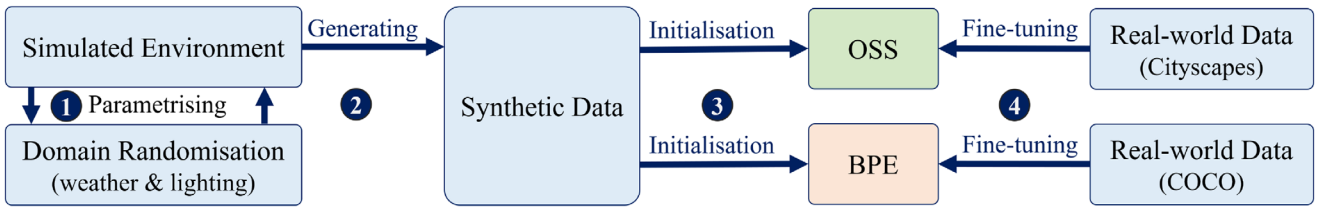

FIGURE 2 | The structure of the proposed technique including initialisation and fine-tuning of the object semantic segmentation (OSS) and body pose estimation (BPE) algorithms.

FIGURE 3 | A sample simulated pedestrian crossing scenario, designed under various lighting and adverse weather conditions.

and intention cues, such as walking styles and social interactions, from sequences of extracted features.

3.2.1 | Simulation Environment

Carla [42] is a widely used open-source driving simulator that enables the creation of custom driving scenarios in diverse urban environments. While hyper-realistic simulations [53–55] can provide photorealistic visual fidelity, they also introduce more complex and detailed visual effects that may act as confounding variables during training, especially for tasks where structure is more relevant than appearance. In our case, Carla offers an effective trade-off: it supports sufficient visual diversity for domain randomisation, while maintaining runtime efficiency, reproducibility, and precise control over structural configuration. This makes it particularly suitable for our objective of training BPE and OSS models as robust feature extractors under varied but controlled visual conditions.

3.2.2 | Scenarios

In this study, we generated 36 driving scenarios simulating diverse pedestrian crossing events. An ego vehicle, equipped with a front-facing RGB camera mounted with zero pitch, yaw, or roll, that is oriented straight ahead to place the horizon line near the centre of the image (Figure 3), was navigated using Carla's autopilot to ensure consistent vehicle behaviour and collision avoidance. The ego vehicle approaches a pedestrian from a uniformly random distance between 20 and 30 m, and travels at a speed randomly selected between 15 and 30 mph. These ranges were selected to reflect typical mid-range urban driving conditions and interaction distances found in real datasets such as PIE and JAAD. Prior studies have shown that crossing predictive models perform with higher stability and accuracy when the longitudinal distance between the pedestrian and the ego vehicle is approximately less than 25 m [56]. The simulation environment includes 36 distinct locations: 16 intersections, 10 straight roads, 6 curved roads, and 4 roundabouts, covering a range of typical urban layouts (see the first row in Figure 3).

A total of 48 pedestrian avatars, which are all the available avatars in the simulator, were included to capture variability in age, body shape, and clothing style (see the second row in Figure 3). To achieve consistent, reproducible, and visually diverse pedestrian movements across scenarios, pedestrian agents were programmed to follow a manually defined sequence of 3D spatial coordinates (i.e. waypoints), enabling precise control over their crossing trajectories, directions, walking speeds, and waiting times. This approach overrides Carla's default behaviour, which enforces legal crossings and vehicle-aware pauses before crossing, and allows more flexible configurations. Therefore, in each scenario, the crossing direction (left-to-right or right-to-left), waiting time before crossing (uniformly sampled from 0 to 4 s), and walking speed (walking or running) were also

randomly assigned. These variations were introduced intentionally to diversify visual and motion patterns for data enrichment purposes, rather than to simulate realistic pedestrian decision-making processes.

3.2.3 | Randomised Parameters (θ)

Each scenario for each pedestrian at each location is repeated 21 times to involve three different ambient illumination conditions (during the night, noon, and sunset), and six different weather conditions (clear, cloudy, heavy rain, moderate rain, light rain, damp overcast, and wet). This is shown in Figure 3 in the last three rows.

The weather and lighting conditions used in the synthetic data generation process were selected to reflect common urban driving scenarios, while introducing controlled visual diversity for domain randomisation. Specifically, we included various weather conditions, along with daytime and nighttime lighting, based on their presence in real-world datasets such as PIE, JAAD, and FU-PIP, and their ability to simulate visually distinct environments.

3.2.4 | Data Collection

Each scenario was recorded for approximately 10 s with 2 fps until the pedestrian finished the crossing or the vehicle passed the pedestrian. This provides sufficient spatial sampling for training frame-level OSS and BPE models, while allowing more effective variability in appearance and motion through DR augmentation. This design choice was validated during preliminary experiments, where increasing the frame rate did not yield significant performance gains but substantially increased training time and reduced DR effectiveness.

For each frame, with a resolution of 1920×1080, a body pose of the pedestrian is automatically generated by the Carla simulator, which provides access to ground-truth 2D joint coordinates of the pedestrian rendered in the scene. The full skeleton includes 67 points representing the location of joints, however, we ignore the hands and foots' keypoints and just extract 17 points of body joints. The high-quality mask labels for pedestrians, roads, pavements, cars, trucks, buses, buildings, sky, traffic signs, traffic lights, and poles are also automatically extracted from the semantic segmentation sensor (shown in Figure 3 rows three).

Ultimately, 648 scenarios (36×21) have been generated consisting of 15,120 ($756 \times 10 \times 2$) annotated frames for semantic information, and 725,760 ($15{,}120 \times 48$) frames of pedestrians plus the keypoints representing their body joints.

3.3 | Model Investigation

We evaluate six state-of-the-art models, three in the area of object semantic segmentation (OSS) and three in body pose estimation (BPE), to test our hypothesis on domain randomisation and synthetic data augmentation techniques. These models have been chosen for their architectural diversity and performance in their respective tasks.

The OSS models were trained on the Cityscapes dataset [8], with mean intersection over union (mIoU) as the evaluation metric. IoU is defined as $\frac{TP}{(TP+FP+FN)}$, where TP, FP, and FN are the true positive, false positive, and false negative pixels for each class, respectively. The candidate OSS models are as follows:

- EfficientPS [57] features a semantic head built upon an optimised encoder and a two-way feature pyramid network, which adaptively fuses fine and contextual features for improved segmentation.
- ViT-Adapter [58] uses vision transformers, this model learns dense visual representations from large-scale data, which are then used for fine-grained semantic predictions.
- InternImage [59] employs deformable convolution as its core operator and adaptive spatial feature aggregation to enlarge the receptive field and improve segmentation accuracy.

The BPE models were trained on the COCO pose datasets [7], with average precision (AP) based on object keypoint similarity (OKS) as the evaluation metric. The following candidate models were selected for their distinct architectural approaches:

- HRNet [60] integrates multi-level fusion to produce detailed, high-resolution keypoint representations for human pose estimation.
- YOLO-Pose [61] uses an end-to-end approach to unify object detection and human pose estimation at lower computational complexity.
- RTMPose [62] utilises parameter optimisation to perform real-time multi-person pose estimation.

To further evaluate these models, we employed two distinct training strategies: 1) Hybrid data training strategy involves combining the synthetic and real-world data and training the model only once; 2) Fine-tuned hybrid strategy entails training each model on the synthetic data, and subsequently fine-tuning it using real-world data.

To choose the feature extractors used downstream, we adopt the following criteria: (C1) accuracy on real-world benchmarks after DR/SDA enrichment, (C2) generalisation to an external dataset with diverse illumination/weather, (C3) training stability and convergence efficiency under DR/SDA, and (C4) inference practicality (throughput/memory) for integration into crossing predictors.

4 | Experiment

In this section, we retrain the object semantic segmentation (OSS) and body pose estimation (BPE) algorithms on the generated synthetic data, aiming to achieve the highest performance improvements. We then evaluate the effectiveness of various augmentation strategies applied during training. Finally, we assess the impact of using these enriched baseline models on the overall performance metrics of state-of-the-art (SOTA) pedestrian crossing crossing prediction methods.

4.1 | Experimental Setup

All experiments were conducted on a workstation equipped with an Intel Core i9-13900 CPU, 64 GB RAM, and an NVIDIA RTX A6000 GPU, using TensorRT in half-precision floating-point (FP16) format.

Cityscapes dataset [8] was used to train and validate the OSS models. It contains 5000 finely annotated urban street images split into 2975 for training, 500 for validation, and 1525 for testing. We standardise our OSS benchmarking on Cityscapes for comparability across candidates, including EfficientPS [57], ViT-Adapter [58], and InternImage [59].

COCO keypoint dataset [7] was used for BPE model training and validation following the COCO 2017 keypoints protocol. The dataset comprises ∼57,000 train images with keypoint annotations (totalling ∼150k person instances), 5000 validation images, and 40,670 images test images. This setup aligns with the training/evaluation protocols used by our BPE candidates HRNet [60], YOLO-Pose [61], and RTMPose [62].

PIE dataset [3] contains about 6 h of driving video (≈1.6M frames) with 1,842 pedestrian instances annotated for crossing intention. We follow the official split of 1484 pedestrians for training and 358 for testing.

JAAD dataset [2] provides 2785 video clips (≈346,000 frames) with 686 pedestrian crossing instances. Following prior work, we adopt an 80/20 split with 2420 clips for training and 365 clips for testing.

FU-PIP dataset [5], a subset of the Waymo Open Dataset [63], consists of 1,082 annotated frames with semantic masks, crossing labels, and 2D body keypoints (13 joints), covering 541 pedestrian instances. FU-PIP (Frontal Urban-PIP) is derived from the Urban-PIP collection, which spans diverse operational conditions (daylight and nighttime) and multiple weather regimes (sunny, cloudy, rainy, foggy), captured in varied urban/rural scenes and intersections [5]. While the source dataset includes a multi-camera, multi-sensor platform (front-left, front, front-right cameras; LiDAR, radar, IMU), we use only the front camera frames for evaluation to ensure compatibility with single-camera baselines [5]. As no official train/val/test split is provided, FU-PIP is used exclusively for evaluation in our experiments.

Finally, our domain-randomised synthetic dataset (Section 3.2.4) contains 15,120 annotated semantic frames and 725,760 pose-labelled frames across 648 simulated crossing scenarios. These were used to initialise and fine-tune the OSS and BPE models prior to evaluation on real-world datasets.

4.2 | Training of Candidate SOTA Models

We use default optimiser selection and momentum configuration for each baseline as provided in their source code repository. The learning rate (η), the total number of epochs (ξ), and the batch size (β) are mentioned in Table 1. In all experiments, the training process was stopped once the model had reached a stable convergence with no further improvements. The epoch number at which convergence was observed has been denoted by ξ^* in Table 1.

The results for OSS models in Table 1 indicate that all three models achieved high accuracy on the Cityscapes dataset. InternImage achieved the highest accuracy, with an mIoU score of 86.9%. ViT-Adapter achieved an mIoU score of 85.1%, and EfficientPS achieved an mIoU score of 84.7%

In the comparison of BPE models using COCO, the AP_R values for the RTMPose model after fine-tuning reached the highest accuracy (80.3%), followed by YOLO-Pose (73.8%) and HRNet (72.1%). Considering the AP_R values on a hybrid dataset of both synthetic and real-world data, RTMPose achieved the highest accuracy again (76.1%), followed by YOLO-Pose (72.7%) and HRNet (71.4%). This indicates that RTMPose is the most suitable model for the BPE task when the model is trained by adapting the fine-tuned hybrid strategy.

4.3 | Augmentation Strategy

Two training strategies, hybrid training (synthetic + real) and fine-tuned hybrid training (synthetic pretraining followed by real fine-tuning), are evaluated in Table 1 across the candidate BPE and OSS models. As shown in Table 1, models initialised with synthetic data exhibit rapid early improvements (owing to the relative simplicity of synthetic samples); however, their performance on real-world data remains poor, underscoring the persistent domain gap between synthetic and real distributions.

Notably, when synthetic data were generated without weather or lighting variations, performance on real-world datasets remained limited despite using a comparable number of training samples. By contrast, domain-randomised synthetic data consistently improved convergence and robustness. This confirms that the reported gains are not merely due to the presence of synthetic data, but specifically attributable to environmental variability introduced through domain randomisation.

To explore the effect of the synthetic-to-real data ratio, we conducted additional experiments varying the proportion of synthetic data used during training. Quantitatively, we found that a rough ratio of 5:1 (synthetic:real) for OSS and 3:1 for BPE, during initial training and fine-tuning phases, respectively, provides the best trade-off between convergence speed and accuracy. Ratios with significantly less synthetic data reduce the benefits of domain randomisation, while excessively high synthetic data proportions without fine-tuning lead to degraded real-world performance due to domain gaps.

In the fine-tuned hybrid training approach, we observed faster convergence and improvement in performance, suggesting that domain-randomised synthetic data helps models learn structural representations more efficiently. For instance, InternImage reached 86.9% mIoU on real data samples in only 1803 epochs (992 train + 811 fine-tune), compared to 86.1% in 2606 epochs using hybrid training. Similarly, RTMPose achieved 80.3% AP in 305 epochs, outperforming its hybrid training counterpart (76.1% in 323 epochs). This shows overall faster convergence time and training stability.

TABLE 1 | Performance comparison of selected SOTA models, under two training strategies: Fine-tuned hybrid (the first row for each model), and Hybrid training (the second row for each model). The second and third columns from left are the baseline models' performance.

OSS										
Model (η, ξ, β)	**mIoU$_R$**	**mIoU$_S$**	**Train**	**ξ^***	**mIoU$_R$**	**mIoU$_S$**	**Fine-tune**	**ξ^***	**mIoU$_R$**	**mIoU$_S$**
EfficientPS *(2×10^{-2}, 200, 16)*	84.2%	63.52%	Synthetic	133	62.3%	87.4%	Cityscapes	94	84.7%	80.8%
			Hybrid training (synthetic + Cityscapes)					164	83.1%	87.0%
ViT-Adapter *(2×10^{-5}, 3k, 16)*	85.2%	61.5%	Synthetic	943	66.4 %	88.2%	Cityscapes	880	85.1%	81.9%
			Hybrid training (synthetic + Cityscapes)					2711	83.7%	85.1%
InternImage *(2×10^{-5}, 3k, 16)*	86.1%	60.7%	Synthetic	992	66.8%	88.7%	Cityscapes	811	86.9%	83.2%
			Hybrid training (synthetic + Cityscapes)					2606	82.2%	87.3%
BPE										
Model (η, ξ, β)	**AP$_R$**	**AP$_S$**	**Train**	**ξ^***	**AP$_R$**	**AP$_S$**	**Fine-tune**	**ξ^***	**AP$_R$**	**AP$_S$**
HRNet *(1×10^{-3}, 140, 32)*	72.3%	59.2%	Synthetic	77	51.2%	89.1%	COCO	65	72.1%	79.8%
			Hybrid training (synthetic + COCO)					109	71.4%	82.0%
YOLO-Pose *(1×10^{-3}, 150, 40)*	74.7%	55.8%	Synthetic	88	52.4%	86.3%	COCO	61	73.8%	77.0%
			Hybrid training (synthetic + COCO)					118	72.7%	80.0%
RTMPose *(4×10^{-3}, 420, 64)*	75.3%	54.2%	Synthetic	172	56.4%	90.1%	COCO	133	80.3%	76.9%
			Hybrid training (synthetic + COCO)					323	76.1%	74.9%

Notations: mIoU and AP with subscripts R/S for real/synthetic test sets, η: Learning rate, ξ: Number of epochs, β: Batch size, ξ^*: Converged epoch number.

Moreover, initialisation on synthetic data can encourage models to learn features that are invariant to randomisation (e.g. road layout and posture). Correspondingly, the model can learn to focus on the important features of the input data that are relevant to the task at hand, rather than relying on specific patterns or properties that may be affected by weather and lighting conditions. On the other hand, when adopting the hybrid training strategy, synthetic data is regarded as a form of data augmentation. As a result, models often achieve moderate but limited generalisation on real data, even after numerous epochs. This is due to the simplicity of the synthetic data, which may reduce the models' ability to learn the complexity of real data. This emphasises that utilising our structurally generated data is preferable for initialising models since augmentation usually leads to decreased performance compared to the baseline performance.

4.4 | Generalisation

Since FU-PIP spans diverse illumination (day/night) and weather (sunny, cloudy, rainy, foggy) conditions, improvements observed on FU-PIP provide an external check on the source of gains.

Table 2 reports the base performance of models (the original real-data model) and the performance of enriched models (fine-tuned hybrid training) using 5:1 (OSS) and 3:1 (BPE) synthetic:real ratios, across the FU-PIP dataset. As shown, the enriched models consistently outperform their baseline counterparts across both OSS and BPE tasks. For OSS, improvements are observed in all models, with InternImage achieving the highest absolute gain (80.3% to 81.2% mIoU). For BPE, similar trends are seen, where RTMPose improves from 69% to 69.6% AP, and YOLO-Pose shows a modest but consistent increase from 67.6% to 68.5% AP.

TABLE 2 | Performance comparison of algorithms on FU-PIP (Waymo) for OSS and BPE under base and enriched settings.

OSS (mIoU %)			BPE (AP %)		
Model	Base	Enriched	Model	Base	Enriched
EfficientPS	77.5	**79.1**	HRNet	67.2	**68.4**
ViT-Adapter	78.1	**78.7**	YOLO-Pose	67.6	**68.5**
InternImage	80.3	**81.2**	RTMPose	69.0	**69.6**

Although the margins are relatively small, these results demonstrate that domain-randomised synthetic data improves model robustness and generalisation, particularly in complex urban scenarios. This supports our hypothesis that training feature extractors under structurally consistent but visually varied conditions helps models learn to ignore irrelevant appearance changes while preserving scene structure, a key requirement for downstream pedestrian crossing prediction.

4.5 | Model Selection

Guided by the criteria in Section 3.3, we select InternImage (OSS) and RTMPose (BPE) as downstream feature extractors. For (C1), InternImage attains the highest Cityscapes mIoU post-enrichment and RTMPose the highest COCO AP (Table 1); for (C2), both yield the strongest scores on the external FU-PIP benchmark spanning day/night and multiple weather regimes (Table 2), indicating robustness beyond dataset-specific effects; for (C3), fine-tuned hybrid training achieves faster, stable convergence (InternImage: 992+811 vs. 2606 total epochs; RTMPose: 305 vs. 323; Section 4.2, Table 1); and for (C4), all OSS candidates

TABLE 3 | Performance comparison between SOTA and enriched SOTA models on PIE, JAAD, and FU-PIP datasets.

Model	Input features	PIE			JAAD			FU-PIP		
		Acc	AUC	F1	Acc	AUC	F1	Acc	AUC	F1
SFGRU (2020)	P, L, B, M	0.871	0.854	0.783	0.832	0.794	0.630	0.652	0.651	0.553
Enriched SFGRU	**P**$^+$, L, B, M	**0.878**	**0.865**	**0.784**	**0.844**	**0.797**	**0.642**	**0.665**	**0.659**	**0.565**
CAPformer (2021)	P, L, B, M	0.897	0.860	0.805	0.834	0.825	0.634	0.641	0.598	0.554
Enriched CAPformer	**P**$^+$, L, B, M	**0.907**	**0.867**	**0.816**	**0.844**	**0.839**	**0.646**	**0.652**	**0.615**	**0.562**
PCIP (2022)	S, P, L, B, M	0.897	0.866	0.807	0.836	0.826	0.633	0.634	0.611	0.514
Enriched PCIP	**S**$^+$, **P**$^+$, L, B, M	**0.916**	**0.885**	**0.809**	**0.858**	**0.869**	**0.657**	**0.668**	**0.648**	**0.547**
GraphPlus (2022)	S, P, L, M	0.895	0.904	0.818	0.868	0.854	0.748	0.639	0.608	0.572
Enriched GraphPlus	**S**$^+$, **P**$^+$, L, M	**0.914**	**0.913**	**0.857**	**0.878**	**0.867**	**0.764**	**0.660**	**0.630**	**0.591**
LGCF (2023)	S, P, L, B, O	0.795	0.774	0.686	0.836	0.843	0.729	0.644	0.645	0.677
Enriched LGCF	**S**$^+$, **P**$^+$, L, B, O	**0.815**	**0.803**	**0.715**	**0.869**	**0.860**	**0.742**	**0.666**	**0.671**	**0.693**
PIP-Net (2025)	S, P, L, B, O, D	0.917	0.895	0.846	0.848	0.834	0.735	0.732	0.709	0.690
Enriched PIP-Net	**S**$^+$, **P**$^+$, L, B, O, D	**0.922**	**0.915**	**0.857**	**0.876**	**0.869**	**0.757**	**0.736**	**0.712**	**0.703**

Features: **S**: Semantic segmentation, **B**: Bounding box, **L**: Local context, **M**: Ego-vehicle speed, **O**: Optical flow, **D**: Depth information, **P**: body Pose. $^+$: Enriched feature.

fit our memory budget at β=16, while among BPE models RTMPose sustains the largest training batch on our A6000 GPU (β=64 vs. 32 for HRNet and 40 for YOLO-Pose), reflecting a favourable throughput/memory trade-off without sacrificing accuracy. Consequently, we adopt InternImage and RTMPose to produce the enriched semantic and pose features (**S**$^+$, **P**$^+$) used by the downstream crossing predictors (Table 3), with qualitative examples illustrated in Figure 4.

4.6 | Impact on Crossing Prediction Models

We integrated enhanced feature extractors into a modular framework, allowing us to experiment with six different crossing prediciton models, including stack fused gated recurrent unit (SFGRU) [15], crossing action prediction via transformers (CAPformer) [19], pedestrian crossing intention prediction using 3D convolutions (PCIP) [18], graph convolutional neural networks (GraphPlus) [17], local and global contextual fusion (LGCF) [64], and pedestrian crossing prediction network (PIP-Net) [5].

4.6.1 | Implementation Details

In the implementation of the crossing prediction models, each pedestrian instance is represented by six observation sequences, where each sequence consists of 16 consecutive data frames (including features like bounding boxes, local context, body pose etc.) extracted from the pedestrian's full observation time. The sequences for each pedestrian instance have 40% (6 frames) overlap to capture temporal continuity and ensure smoother transitions between segments while preserving contextual information across time. Each sequence is associated with a time-to-event (TTE) value, starting from 0 to 60, which indicates the temporal distance (in frames) from the critical moment. The critical moment is defined as either the first frame in which a crossing pedestrian steps onto the street or the last observable frame for a non-crossing pedestrian.

To provide a consistent evaluation, we compute performance by averaging predictions across all TTEs for each pedestrian instance. This differs from some original implementations, which often report performance at a specific TTE (typically the closest to the crossing moment), where the prediction task can be easier due to richer contextual cues. Our approach on using six observation sequences ensures that model evaluation reflects performance across the full available temporal range for pedestrian instances across both datasets.

All listed predictive models were retrained from scratch using their publicly available source codes. We modified the input features by replacing the original pose and/or semantic segmentation inputs with those extracted from our enriched versions of InternImage as an OSS and RTMPose as a BPE.

Pedestrian body pose estimation is known to degrade in performance under complex conditions, particularly when pedestrians are partially occluded or located in crowded scenes. To address these challenges, RTMPose incorporates a comprehensive set of data augmentation techniques during training, including mosaic augmentation, colour jittering, random geometric transformations, and MixUp [65]. These augmentations enhance the model's robustness by exposing it to a wider variety of appearance conditions. Moreover, our downstream evaluation datasets naturally contain numerous occlusion scenarios, such as pedestrians partially obscured by other individuals or vehicles, thereby offering a realistic assessment of performance under such conditions. This design choice ensures that our evaluation setup accounts for the practical limitations of pose estimation models in real-world environments.

The training configurations, including learning rate, number of epochs, batch size, and optimiser, initialisation random seed,

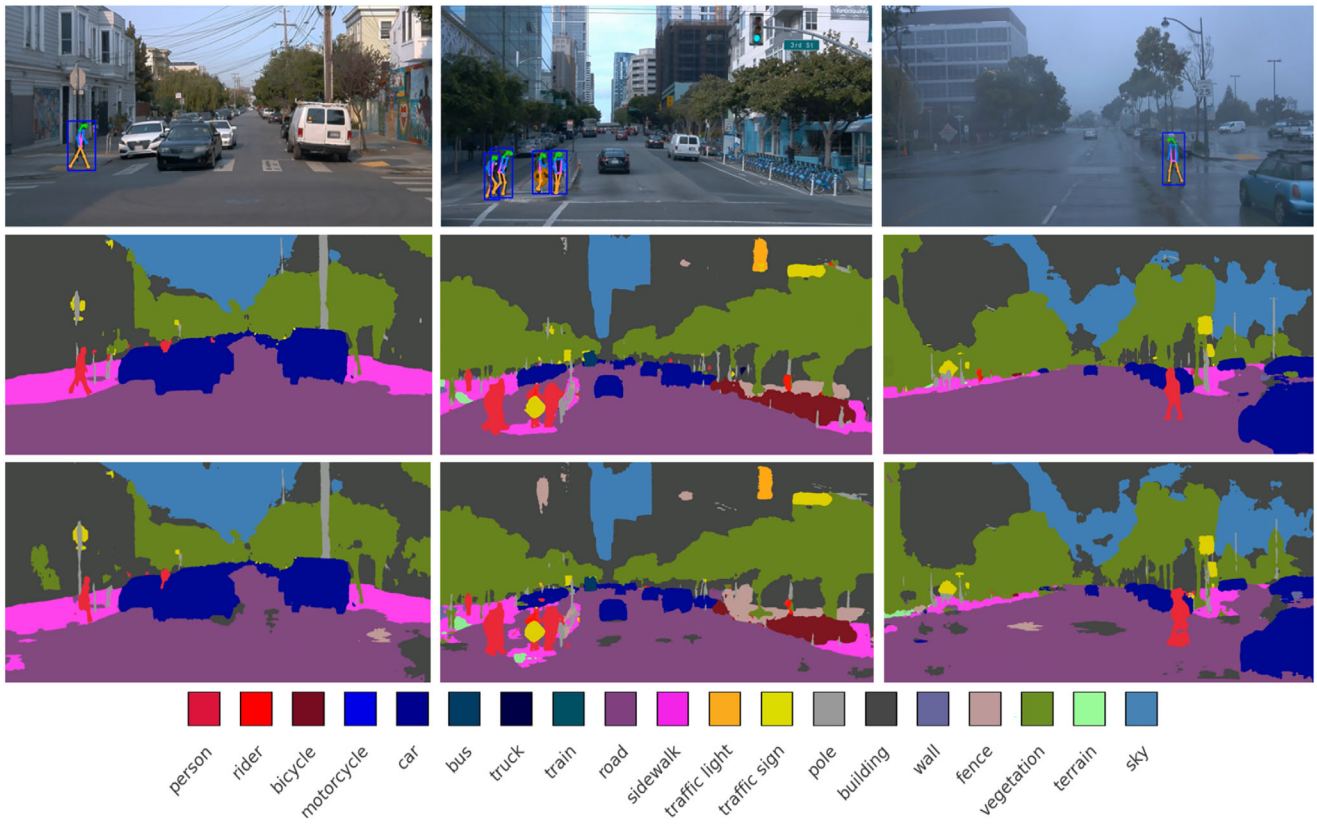

FIGURE 4 | Results of BPE via enhanced RTMPose and OSS via InternImage (second row) models on traffic scenes with single and multiple pedestrians in sunny, cloudy, and foggy weather. The third row represents the OSS results from InternImage before enrichment

were kept consistent with the default or reported settings in each model's original implementation. For all models, training was governed by an early stopping strategy based on validation set performance, as predefined in their original code.

Each model was independently trained and evaluated on the PIE [3] and JAAD [2] datasets using their standard train/test splits. For the FU-PIP dataset, as the number of pedestrian instances (541) was insufficient for training, we evaluated the models trained on PIE directly on FU-PIP, as recommended by the authors of FU-PIP [5].

The training time for OSS models ranged from 7 to 12 h, while BPE models took approximately 5 to 7 h, depending on the architecture and training setup. The fine-tuned hybrid approach can reduce this time by approximately 30 and 170 min for RTMPose and InternImage, respectively. The pedestrian crossing prediction models, using extracted features, required an additional 4 to 6 h. These durations confirm the computational efficiency of our pipeline and its feasibility for practical applications.

4.6.2 | Results

The evaluation metrics used are crossing intention accuracy (Acc), area under the curve (AUC), and F1 scores. The Acc measures how well the model correctly identifies whether a pedestrian intends to cross the street. The AUC is used to evaluate the model's capability to differentiate between the 'crossing' and 'not crossing' classes. A high AUC score signifies the model's adeptness in prioritising instances that are more likely to involve crossing. The F1 score is the harmonic mean of precision and recall. A high F1 score means the model effectively reduces both false positives (incorrectly predicting a pedestrian will cross) and false negatives (failing to predict a crossing estimation), thereby enhancing pedestrian safety by minimising these types of errors.

Table 3 demonstrates the comparison between the original models and boosted models on the two specific datasets of pedestrian crossing prediction. The results show that providing more accurate input features, enhanced through domain randomisation and synthetic data augmentation, can boost the performance of pedestrian crossing prediction models. The enriched models consistently outperformed their baselines across all performance metrics (accuracy, AUC, and F1 score) on both the PIE and JAAD datasets, validating the impact of improved feature extraction on crossing prediction accuracy. Moreover, the consistent boosts of enriched models on FU-PIP support that environmental diversity induced by DR underpins the observed generalisation improvements.

While the observed improvements from enrichment are modest in some cases, this variability can be attributed to several factors. First, certain predictive models already demonstrate strong baseline performance, which may limit the extent of measurable gains, particularly given the complexity of pedestrian crossing prediction. This task often requires more than visual cues alone and may benefit from additional input modalities such as

TABLE 4 | Confidence delta (conf Δ) for each model across datasets. Each cell reports max/average confidence change between consecutive predictions. Lower values indicate more stable predictions across different TTEs.

Model	PIE	JAAD	FU-PIP
SFGRU	0.10 / 0.04	0.17 / 0.07	0.13 / 0.09
Enriched SFGRU	**0.09** / **0.03**	**0.14** / 0.07	0.10 / 0.07
CAPformer	0.12 / 0.05	0.18 / **0.08**	0.15 / **0.06**
Enriched CAPformer	0.12 / 0.04	0.15 / 0.07	0.12 / 0.05
PCIP	0.14 / 0.08	0.20 / 0.09	0.16 / 0.07
Enriched PCIP	0.14 / 0.07	0.17 / **0.08**	0.13 / **0.06**
GraphPlus	0.16 / 0.07	0.21 / 0.14	0.18 / 0.11
Enriched GraphPlus	0.16 / 0.06	0.20 / 0.12	0.15 / 0.10
LGCF	0.17 / 0.07	0.23 / 0.10	0.19 / 0.08
Enriched LGCF	0.16 / 0.06	0.22 / 0.09	0.16 / 0.07
PIP-Net	0.13 / 0.06	0.19 / 0.09	0.17 / 0.07
Enriched PIP-Net	0.13 / 0.05	0.18 / **0.08**	0.14 / 0.07

multi-sensor fusion or traffic-aware features (e.g. traffic light state or contextual information about designated vs. non-designated crossing environments). In this context, architectures that rely more heavily on spatial detail, such as PPCI and PIP-Net, tend to benefit more from enriched modules like BPE and OSS, which provide fine-grained structural cues. Furthermore, it is important to emphasise that even marginal improvements can be meaningful in the context of pedestrian intention estimation, where decisions are safety-critical and prediction robustness plays a vital role in real-world deployment.

To further investigate the impact of enriched modules, we evaluate the stability of model confidence across TTEs using confidence delta (confΔ). For each pedestrian instance, we compute the change in model confidence for a given class between two consecutive sequences and report both the maximum and average delta, defined as:

$$\text{confΔ} = \frac{1}{n-1}\sum_{i=1}^{n-1}\left|\text{conf}_i^c - \text{conf}_{i+1}^c\right|, \tag{2}$$

where n is the number of sequences for a given pedestrian instance, and conf_i^c is the model's predicted confidence for class c at sequence i. Lower values of confΔ indicate more stable and consistent predictions over time. Table 4 presents the average and maximum confidence deltas across models and datasets.

5 | Conclusion

The study indicates that using a simulated environment with structural randomised properties can enhance the resilience of the pedestrian crossing prediction to variations in the input data. The study aimed at improving the precision and accuracy of semantic information and posture feature extraction algorithms which successfully affected the accuracy of predicting the crossing of pedestrians, as well. Furthermore, the evaluation results revealed a performance gap between different training modes based on real and synthetic data for both semantic segmentation and body pose estimation. We show using synthetic data as an augmentation technique may not necessarily lead to a sensible enhancement of existing semantic segmentation algorithms.

We suggested domain randomisation technique by generating a parameter-randomised synthetic dataset to bridge the domain gap between synthetic and real-world based training models. The final outcomes of our experiments (as per Table 1) proved that all evaluated SOTA models have been improved (with no exception) against all performance metrics.

We utilised Carla simulation to generate a wide range of synthetic environmental conditions including diversity of weather and ambient light. The study found that when the models were trained with synthetic data only, they showed poor performance on real-world data; however, when the models were adapted with subsequent real-world data training and tuning, a higher generalisability and performance were achieved compared to the original model. The experimental results indicate improvements in the accuracy, AUC and F1 scores for both JAAD and PIE datasets, showing the promising potential of using synthetic data for improving the prediction of pedestrian crossing in AVs.

In summary, our findings demonstrate that the observed improvements are not merely the result of incorporating synthetic data, but are primarily driven by the structured diversity introduced through domain randomisation. By systematically varying weather and illumination, the models are better prepared to handle unseen real-world conditions, thereby enhancing both robustness and generalisation in pedestrian crossing prediction.

Author Contributions

Mohsen Azarmi: conceptualisation, formal analysis, investigation, methodology, software, visualisation, writing - original draft, writing - review & editing. **Mahdi Rezaei:** conceptualisation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, supervision, validation, visualisation, writing - review & editing. **He Wang:** conceptualisation, formal analysis, methodology, supervision, validation, writing - review & editing.

Acknowledgements

The authors would like to thank all partners within the Hi-Drive project for their cooperation and valuable contribution. This research has received funding from the European Union's Horizon 2020 research and innovation programme, under grant Agreement No 101006664. The article reflects only the author's view and neither the European Commission nor CINEA is responsible for any use that may be made of the information this document contains.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The synthetic data generation code is openly available in GitHub at https://github.com/azarmi/SyntheticCrossing.

References

1. N. Sharma, C. Dhiman, and S. Indu, "Pedestrian Intention Prediction for Autonomous Vehicles: A Comprehensive Survey," *Neurocomputing* 508 (2022): 120–152.

2. A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Are They Going to Cross? A Benchmark Dataset and Baseline for Pedestrian Crosswalk Behavior," in *Proceedings of the IEEE International Conference on Computer Vision Workshops* (IEEE, 2017), 206–213.

3. A. Rasouli, I. Kotseruba, T. Kunic, and J. K. Tsotsos, "Pie: A Large-Scale Dataset and Models for Pedestrian Intention Estimation and Trajectory Prediction," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (IEEE, 2019), 6261–6270.

4. B. Liu, E. Adeli, Z. Cao, et al., "Spatiotemporal Relationship Reasoning for Pedestrian Intent Prediction," *IEEE Robotics and Automation Letters* 5, no. 2 (2020): 3485–3492.

5. M. Azarmi, M. Rezaei, and H. Wang, "PIP-Net: Pedestrian Intention Prediction in the Wild," *IEEE Transactions on Intelligent Transportation Systems* 26, no. 7 (2025): 9824–9837.

6. J. Fang, F. Wang, J. Xue, and T.-S. Chua, "Behavioral Intention Prediction in Driving Scenes: A Survey," *IEEE Transactions on Intelligent Transportation Systems* 25, no. 8 (2024): 8334–8355.

7. T.-Y. Lin, M. Maire, S. Belongie, et al., "Microsoft COCO: Common Objects in Context," in *13th European Conference on Computer Vision–ECCV 2014* (Springer, 2014), 740–755.

8. M. Cordts, M. Omran, S. Ramos, et al., "The Cityscapes Dataset for Semantic Urban Scene Understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2016), 3213–3223.

9. S. Z. Valtchev and J. Wu, "Domain Randomization for Neural Network Classification," *Journal of Big Data* 8, no. 1 (2021): 94.

10. N. Jaipuria, X. Zhang, R. Bhasin, et al., "Deflating Dataset Bias Using Synthetic Data Augmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (IEEE, 2020), 772–773.

11. C. G. Keller and D. M. Gavrila, "Will the Pedestrian Cross? A Study on Pedestrian Path Prediction," *IEEE Transactions on Intelligent Transportation Systems* 15, no. 2 (2013): 494–506.

12. A. T. Schulz and R. Stiefelhagen, "Pedestrian Intention Recognition Using Latent-Dynamic Conditional Random Fields," in *2015 IEEE Intelligent Vehicles Symposium (IV)* (IEEE, 2015), 622–627.

13. X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, "Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting," in *NIPS'15: Proceedings of the 29th International Conference on Neural Information Processing Systems* (ACM, 2015), 802–810.

14. J. Gesnouin, S. Pechberti, B. Stanciulcscu, and F. Moutarde, "TrouspiNet: Spatio-Temporal Attention on Parallel Atrous Convolutions and U-GRUs for Skeletal Pedestrian Crossing Prediction," in *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)* (IEEE, 2021), 01–07.

15. A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Pedestrian Action Anticipation Using Contextual Feature Fusion in Stacked RNNs," arXiv:2005.06582 (2020).

16. T. Chen, R. Tian, and Z. Ding, "Visual Reasoning Using Graph Convolutional Networks for Predicting Pedestrian Crossing Intention," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (IEEE, 2021), 3103–3109.

17. P. R. G. Cadena, Y. Qian, C. Wang, and M. Yang, "Pedestrian Graph+: A Fast Pedestrian Crossing Prediction Model Based on Graph Convolutional Networks," *IEEE Transactions on Intelligent Transportation Systems* 23, no. 11 (2022): 21050–21061.

18. D. Yang, H. Zhang, E. Yurtsever, K. A. Redmill, and Ü. Özgüner, "Predicting Pedestrian Crossing Intention With Feature Fusion and Spatio-Temporal Attention," *IEEE Transactions on Intelligent Vehicles* 7, no. 2 (2022): 221–230.

19. J. Lorenzo, I. Parra, R. Izquierdo, A. L. Ballardini, Á. Hernández-Saz, D. F. Llorca, and M. Á. Sotelo, "Capformer: Pedestrian Crossing Action Prediction Using Transformer," *Sensors* 21, no. 17 (2021): 5694.

20. A. Bhattacharjee and S. L. Waslander, "AugTrEP: Scene and Occlusion-Aware Pedestrian Crossing Intention Prediction,"*Proceedings of the 21st Conference on Robots and Vision* (CRV, 2024), 1–8.

21. S. Scaccia, F. Pro, and I. Amerini, "Unsupervised Pedestrian Intention Estimation Through Deep Neural Embeddings and Spatio-Temporal Graph Convolutional Networks," *Pattern Analysis and Applications* 28, no. 2 (2025): 108.

22. J. Gesnouin, S. Pechberti, B. Stanciulescu, and F. Moutarde, "Assessing Cross-Dataset Generalization of Pedestrian Crossing Predictors," in *2022 IEEE Intelligent Vehicles Symposium (IV)* (IEEE, 2022), 419–426.

23. M. Rezaei, M. Azarmi, and F. M. P. Mir, "3D-Net: Monocular 3D Object Recognition for Traffic Monitoring," *Expert Systems with Applications* 227 (2023): 120253.

24. S. Dubey and M. Dixit, "A Comprehensive Survey on Human Pose Estimation Approaches," *Multimedia Systems* 29, no. 1 (2023): 167–195.

25. M. Gao, F. Zheng, J. J. Yu, C. Shan, G. Ding, and J. Han, "Deep Learning for Video Object Segmentation: A Review," *Artificial Intelligence Review* 56, no. 1 (2023): 457–531.

26. J. Li, X. Shi, F. Chen, et al., "Pedestrian Crossing Action Recognition and Trajectory Prediction With 3D Human Keypoints," in *2023 IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, 2023), 1463–1470.

27. Y. Zhou, G. Tan, R. Zhong, Y. Li, and C. Gou, "PIT: Progressive Interaction Transformer for Pedestrian Crossing Intention Prediction," *IEEE Transactions on Intelligent Transportation Systems* 24, no. 12 (2023): 14 213–14 225.

28. Z. Zhang, R. Tian, and Z. Ding, "TrEP: Transformer-Based Evidential Prediction for Pedestrian Intention With Uncertainty," in *Proceedings of the AAAI Conference on Artificial Intelligence* 37, no. 3 (2023), 3534–3542.

29. L. Achaji, J. Moreau, T. Fouqueray, F. Aioun, and F. Charpillet, "Is Attention to Bounding Boxes all you Need for Pedestrian Action Prediction?" in *2022 IEEE Intelligent Vehicles Symposium (IV)* (IEEE, 2022), 895–902.

30. A. Rasouli, M. Rohani, and J. Luo, "Bifold and Semantic Reasoning for Pedestrian Behavior Prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (IEEE, 2021), 15600–15610.

31. R. Quintero, I. Parra, J. Lorenzo, D. Fernández-Llorca, and M. Sotelo, "Pedestrian Intention Recognition by Means of a Hidden Markov Model and Body Language," in *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)* (IEEE, 2017), 1–7.

32. R. Q. Mínguez, I. P. Alonso, D. Fernández-Llorca, and M. A. Sotelo, "Pedestrian Path, Pose, and Intention Prediction Through Gaussian Process Dynamical Models and Pedestrian Activity Recognition," *IEEE Transactions on Intelligent Transportation Systems* 20, no. 5 (2018): 1803–1814.

33. J. Gesnouin, S. Pechberti, G. Bresson, B. Stanciulescu, and F. Moutarde, "Predicting Intentions of Pedestrians From 2D Skeletal Pose Sequences With a Representation-Focused Multi-Branch Deep Learning Network," *Algorithms* 13, no. 12 (2020): 331.

34. F. Piccoli, R. Balakrishnan, M. J. Perez, et al., "FuSSI-Net: Fusion of Spatio-Temporal Skeletons for Intention Prediction Network," in *2020 54th Asilomar Conference on Signals, Systems, and Computers* (IEEE, 2020), 68–72.

35. F. Schneemann and P. Heinemann, "Context-Based Detection of Pedestrian Crossing Intention for Autonomous Driving in Urban

Environments," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (IEEE, 2016), 2243–2248.

36. S. Neogi, M. Hoy, K. Dang, H. Yu, and J. Dauwels, "Context Model for Pedestrian Intention Prediction Using Factored Latent-Dynamic Conditional Random Fields," *IEEE Transactions on Intelligent Transportation Systems* 22, no. 11 (2020): 6821–6832.

37. B. Yang, W. Zhan, P. Wang, C. Chan, Y. Cai, and N. Wang, "Crossing or not? Context-Based Recognition of Pedestrian Crossing Intention in the Urban Environment," *IEEE Transactions on Intelligent Transportation Systems* 23, no. 6 (2021): 5338–5349.

38. D. Yang, H. Zhang, E. Yurtsever, K. A. Redmill, and U. Ozguner, "Predicting Pedestrian Crossing Intention With Feature Fusion and Spatio-Temporal Attention," *IEEE Transactions on Intelligent Vehicles* 7 (2021): 221–230.

39. M. Azarmi, M. Rezaei, H. Wang, and A. Arabian, "Feature Importance in Pedestrian Intention Prediction: A Context-Aware Review," *arXiv:2409.07645* (2024).

40. C. Shorten and T. M. Khoshgoftaar, "A Survey on Image Data Augmentation for Deep Learning," *Journal of Big Data* 6, no. 1 (2019): 1–48.

41. A. Mumuni, F. Mumuni, and N. K. Gerrar, "A Survey of Synthetic Data Augmentation Methods in Machine Vision," *Machine Intelligence Research* 21, no. 5 (2024): 831–869.

42. A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An Open Urban Driving Simulator," in *Conference on Robot Learning* (PMLR, 2017), 1–16.

43. D. Peng, Y. Lei, L. Liu, P. Zhang, and J. Liu, "Global and Local Texture Randomization for Synthetic-to-Real Semantic Segmentation," *IEEE Transactions on Image Processing* 30 (2021): 6594–6608.

44. C. Li and G. H. Lee, "From Synthetic to Real: Unsupervised Domain Adaptation for Animal Pose Estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2021), 1482–1491.

45. S. Zhu and Y. Tian, "Shape Robustness in Style Enhanced Cross Domain Semantic Segmentation," *Pattern Recognition* 135 (2023): 109143.

46. J. Tremblay, A. Prakash, D. Acuna, et al., "Training Deep Networks With Synthetic Data: Bridging the Reality Gap by Domain Randomization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (IEEE, 2018), 969–977.

47. A. Prakash, S. Boochoon, M. Brophy, et al., "Structured Domain Randomization: Bridging the Reality Gap by Context-Aware Synthetic Data," in *2019 International Conference on Robotics and Automation (ICRA)* (IEEE, 2019), 7249–7255.

48. X. Yue, Y. Zhang, S. Zhao, A. Sangiovanni-Vincentelli, K. Keutzer, and B. Gong, "Domain Randomization and Pyramid Consistency: Simulation-to-Real Generalization Without Accessing Target Domain Data," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (IEEE, 2019), 2100–2110.

49. F. Hagelskjær and A. G. Buch, "ParaPose: Parameter and Domain Randomization Optimization for Pose Estimation Using Synthetic Data," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (IEEE, 2022), 6788–6795.

50. D. Pasanisi, E. Rota, M. Ermidoro, and L. Fasanotti, "On Domain Randomization for Object Detection in Real Industrial Scenarios Using Synthetic Images," *Procedia Computer Science* 217 (2023): 816–825.

51. X. Hu, S. Li, T. Huang, B. Tang, R. Huai, and L. Chen, "How Simulation Helps Autonomous Driving: A Survey of Sim2Real, Digital Twins, and Parallel Intelligence," *IEEE Transactions on Intelligent Vehicles* 9, no. 1 (2023): 593–612.

52. M. Fabbri, G. Braso, G. Maugeri, et al., "Motsynth: How Can Synthetic Data Help Pedestrian Detection and Tracking?" in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (IEEE, 2021), 10829–10839.

53. W. Wu, Y. Zhao, M. Z. Shou, H. Zhou, and C. Shen, "DiffuMask: Synthesizing Images With Pixel-Level Annotations for Semantic Segmentation Using Diffusion Models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (IEEE, 2023), 1206–1217.

54. C. Lindström, G. Hess, A. Lilja, M. Fatemi, L. Hammarstrand, C. Petersson, and L. Svensson, "Are NeRFs Ready for Autonomous Driving? Towards Closing the Real-to-Simulation Gap," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2024), 4461–4471.

55. M.-Y. Shen, C.-C. Hsu, H.-Y. Hou, et al., "DriveEnv-NeEF: Exploration of a NERF-Based Autonomous Driving Environment for Real-World Performance Validation,"*arXiv:2403.15791* (2024).

56. J. Ma and W. Rong, "Pedestrian Crossing Intention Prediction Method Based on Multi-Feature Fusion," *World Electric Vehicle Journal* 13, no. 8 (2022): 158.

57. R. Mohan and A. Valada, "EfficientPS: Efficient Panoptic Segmentation," *International Journal of Computer Vision* 129, no. 5 (2021): 1551–1579.

58. Z. Chen, Y. Duan, W. Wang, J. He, T. Lu, J. Dai, and Y. Qiao, "Vision Transformer Adapter for Dense Predictions," *arXiv:2205.08534* (2022).

59. W. Wang, J. Dai, Z. Chen, et al., "InternImage: Exploring Large-Scale Vision Foundation Models With Deformable Convolutions," *arXiv:2211.05778* (2022).

60. Z. Geng, K. Sun, B. Xiao, Z. Zhang, and J. Wang, "Bottom-up Human Pose Estimation via Disentangled Keypoint Regression," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2021), 14 671–14 681.

61. D. Maji, S. Nagori, M. Mathew, and D. Poddar, "YOLO-Pose: Enhancing YOLO for Multi Person Pose Estimation Using Object Keypoint Similarity Loss," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2022), 2637–2646.

62. T. Jiang, P. Lu, L. Zhang, et al., "RTMPose: Real-Time Multi-Person Pose Estimation Based on MMPose," *arXiv:2303.07399* (2023).

63. P. Sun, H. Kretzschmar, X. Dotiwalla, et al., "Scalability in Perception for Autonomous Driving: Waymo Open Dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2020), 2446–2454.

64. M. Azarmi, M. Rezaei, T. Hussain, and C. Qian, "Local and Global Contextual Features Fusion for Pedestrian Intention Prediction," in *International Conference on Artificial Intelligence and Smart Vehicles* (Springer, 2023), 1–13.

65. H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond Empirical Risk Minimization,"*arXiv:1710.09412* (2017).