

WatchoutPed: A dataset and model for Vulnerable Pedestrian Anticipation in surveillance videos

Je-Seok Ham ^{a,b},¹, Dae Hoe Kim ^a,¹, Jinyoung Moon ^{a,c},^{*}

^a Electronics and Telecommunications Research Institute, 218 Gajeong-ro, Yuseong-gu, Daejeon, 34129, Republic of Korea

^b Korea Advanced Institute of Science and Technology (KAIST), 291 Daehak-ro, Yuseong-gu, Daejeon, 34141, Republic of Korea

^c University of Science and Technology (UST), 217 Gajeong-ro, Yuseong-gu, Daejeon, 34113, Republic of Korea

ARTICLE INFO

Keywords:

Pedestrian situation anticipation
Deep learning
Recurrent learning
Attention model
Multi-modal feature fusion

ABSTRACT

This paper addresses the crucial challenge of Vulnerable Pedestrian Anticipation (VPA) in urban environments, utilizing surveillance video data to enhance pedestrian safety. VPA is crucial for identifying pedestrians in potentially dangerous situations, such as walking alongside roads or crossing crosswalks, where the risk of vehicular collisions is elevated. To advance research in this field, we introduce two primary components: the WatchoutPed dataset and the Vulnerable Pedestrian Anticipation Network (VPANet), a baseline network especially designed for VPA. The WatchoutPed dataset has been meticulously enriched with extensive annotations through an innovative auto-labeling technique that integrates ground region analysis with pedestrian state estimation, thus providing a solid foundation for VPA research. Complementing this, the VPANet is engineered to process visual and non-visual inputs extracted from past frames in surveillance footage, enabling it to predict the future state of pedestrians as either safe or unsafe. Tested on the WatchoutPed dataset, VPANet achieves an impressive 89% accuracy, outperforming current methods. Furthermore, we demonstrate the effectiveness of our auto-labeling approach. Notably, the accuracy of VPANet, when trained with the auto-generated annotations from the WatchoutPed, closely parallels that achieved with human-verified annotations, with a negligible variance of less than 1%. The broader implications of our work are significant for the development of smart urban safety infrastructures. Integrating these insights into intelligent crosswalk systems could greatly enhance the monitoring of pedestrian activity near crosswalks, enabling the timely alerting of drivers to the presence of vulnerable pedestrians, and thereby proactively preventing potential vehicular accidents.

1. Introduction

Enhancing pedestrian safety in intelligent cars and traffic systems necessitates the prediction and anticipation of both the current and future states of pedestrians, particularly those vulnerable to vehicular collisions. Recent methodological advancements have predominantly focused on pedestrian detection in crowded conditions, as evidenced by numerous studies [1–9]. These methods have been rigorously trained and evaluated using established pedestrian datasets such as Caltech pedestrians [10], CityPersons [11] and CrowdHuman [12]. Moreover, there has been an increasing focus on developing datasets tailored for pedestrian trajectory prediction [13–15] and pedestrian crossing intention estimation, utilizing videos from onboard vehicle cameras. As a result, methods for pedestrian crossing intention estimation have actively been proposed, employing these specific datasets [16–27].

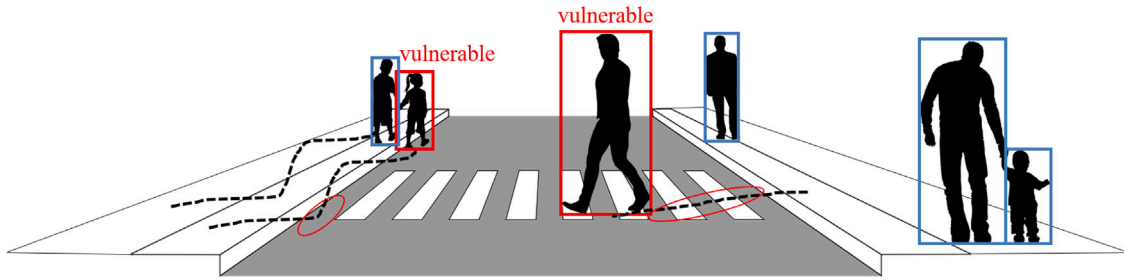
In intelligent traffic systems, the emphasis is progressively placed on predicting vulnerable pedestrians, as illustrated in Fig. 1(a), going

beyond merely estimating pedestrian crossing intentions to bolster pedestrian safety. Traditional methods for estimating crossing intentions, depicted in Fig. 1(b)-(1), determine if pedestrians will commence crossing after a predefined time span, t seconds, from the start of observation. Post initiation of crossing, these methods discontinue their assessment, as they are principally designed for driver assistance systems or autonomous vehicles that preempt crossings prior to time t . Following the onset of a crossing action of a pedestrian, drivers, with or without the aid of autonomous driving technology, proactively participate in collision prevention with the pedestrian. Nonetheless, for intelligent traffic systems managing traffic signals or issuing driver alerts, it is crucial to maintain continuous surveillance for the full extent of any potential hazard to vulnerable pedestrians, as indicated in Fig. 1(b). Thus, it becomes imperative to observe not just the beginning but also the end of the crossing action. Additionally, the anticipation

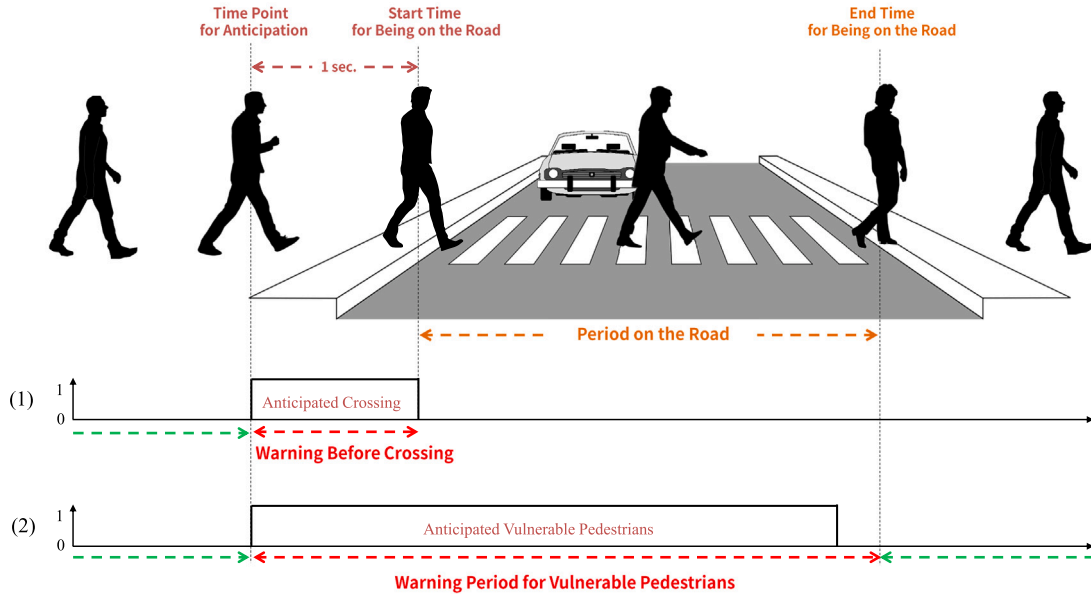
* Corresponding author at: Electronics and Telecommunications Research Institute, 218 Gajeong-ro, Yuseong-gu, Daejeon, 34129, Republic of Korea.

E-mail addresses: jsham@etri.re.kr (J.-S. Ham), dhkim19@etri.re.kr (D.H. Kim), jymoon@etri.re.kr (J. Moon).

¹ These authors contributed equally.



(a) Concept of vulnerable pedestrians



(b) Comparisons between (1) existing methods for crossing intention estimation and (2) our proposed method for anticipating vulnerable pedestrians

Fig. 1. Concept of vulnerable pedestrians and comparisons between existing methods for crossing intention estimation and our proposed method for anticipating vulnerable pedestrians.

of pedestrian vulnerability is paramount, not only during their transit across roads but also while walking adjacent to them, including at designated crosswalks. Contrary to prevalent perceptions of safety in these areas, statistical data substantiates a considerable hazard of pedestrian–vehicle collisions [28].

We posit that pedestrians exhibit vulnerability while navigating alongside roads or crossing at crosswalks due to the imminent risk of vehicular collisions, as illustrated in Fig. 1. Stemming from this premise, we characterize a pedestrian’s unsafe state as one wherein they are exposed to potential hazards on the road or at crosswalks. An anticipatory method for vulnerable pedestrians, depicted in Fig. 1(b)-(2), predicts the likelihood of pedestrians being on the road, encompassing those at crosswalks, within a one-second timeframe. Once pedestrians transition to the sidewalk, their status is reassessed to that of safe pedestrians. In crossing scenarios, the predictive analysis extends from the commencement to the termination of the crossing motion, during which pedestrians are deemed to be in an unsafe state.

In pursuit of advancing pedestrian safety, we introduce a dataset alongside a baseline model dedicated to predicting the vulnerability of pedestrians in precarious states as captured in surveillance footage. To assemble our WatchoutPed dataset, we have implemented an automated labeling method that combines ground region analysis with pedestrian state estimation. Concurrently, we introduce the Vulnerable Pedestrian Anticipation Network (VPANet) as a foundational model. This model processes both visual and non-visual data from historical

past frames within surveillance videos to predict the state of pedestrians, classifying them as either safe or unsafe. The VPANet demonstrates superior performance, surpassing existing state-of-the-art models when trained on our WatchoutPed dataset. Furthermore, the VPANet, when trained on the WatchoutPed dataset with auto-generated annotations, exhibits performance on par with that of the VPANet trained on human-verified annotations, with a negligible accuracy differential of less than 1%.

Our key contributions can be summarized as follows:

- **Comprehensive Dataset and Baseline Model:** We introduce the WatchoutPed dataset, a robust and richly annotated dataset specifically designed for anticipating vulnerable pedestrians. This dataset, paired with our baseline model VPANet, facilitates accurate predictions of pedestrian states in surveillance footage. The extensive annotations and the novel auto-labeling technique make this dataset a substantial resource for advancing research in pedestrian safety.
- **Vulnerable Pedestrian Anticipation Network (VPANet):** We propose VPANet, a model specifically designed to predict the future states of pedestrians using both visual and non-visual data from historical surveillance footage. VPANet demonstrates superior performance, achieving 89% accuracy, which surpasses current state-of-the-art methods. The model effectively processes

multiple types of input data, including local context, global context, and bounding box information, to provide accurate and timely predictions.

- **Preliminary Evaluation of Automatically Generated Annotations:** We conducted initial experiments comparing models trained on automatically generated annotations with those trained on human-verified annotations. These preliminary results indicate that our auto-labeling technique can generate annotations with a performance variance of less than 1% compared to human-verified data. While these findings are promising, they suggest that the auto-labeling method has the potential to serve as an efficient alternative to human annotators for generating large-scale annotations, though further validation is needed to confirm its effectiveness across diverse settings.

This study introduces the first benchmark dataset and baseline model for Vulnerable Pedestrian Anticipation (VPA) in surveillance videos, establishing a foundation for future safety applications. Building on this work, our future efforts aim to enhance pedestrian safety by deploying this method and dataset in real-world environments, specifically through collaborations with local governments utilizing CCTV surveillance. Additionally, we plan to expand the dataset and refine the model to ensure effective performance in diverse urban settings.

2. Related work

2.1. Pedestrian detection

Extensive research has been conducted to improve pedestrian detection in real-world environments. In [9], Fore-Background Contrast Attention (FBCA) was proposed, which improves pedestrian detection under low-light conditions by incorporating background information into the channel attention mechanism. While some research has focused on reducing the high costs of extensive annotations, a new fine-tuning method was proposed in [7] for unsupervised pedestrian detection that does not require source and target data. Additionally, the Scene-adaptive Pseudo Annotation (SaPA) approach was proposed in [8], which utilizes both annotated source data and unannotated target data to generate high-quality pseudo annotations, thereby improving pedestrian detection in semi-supervised scenarios.

Among these general pedestrian detection advancements, certain studies have concentrated on refining the detection of vulnerable pedestrians to enhance pedestrian safety. In [3], single-shot detector [1] with a MobileNet [2] backbone was used for the rapid detection of pedestrians and vehicles in surveillance videos, particularly around crosswalks. Another significant contribution by Mehtab and Yan [5] involved employing CSPNet [4] for detecting pedestrians and cyclists within the KITTI dataset [29], recorded from dashboard cameras. However, a common limitation of these previous methods in detecting vulnerable pedestrians is their assumption that all individuals in the scene are in unsafe situations. Consequently, they primarily focus on identifying bounding boxes of persons without incorporating contextual scene information.

2.2. Pedestrian crossing intention estimation

Many studies have focused on predicting pedestrian crossing intentions. Most of these studies have utilized datasets recorded from vehicle-mounted cameras, and predictive models based on neural networks have been developed for each dataset.

The pioneering dataset for predicting pedestrian crossing intentions, the Joint Attention in Autonomous Driving (JAAD), was first introduced by Rasouli et al. [30]. This dataset goes beyond mere pedestrian detection and bounding box annotation by incorporating behavioral and contextual information, thereby offering a more comprehensive understanding of pedestrian intentions. Subsequent research utilizing

this dataset has included the extraction of pose information [31,32] and the implementation of multi-task approaches [33–35] to improve prediction accuracy. However, the utility of this dataset is constrained by the limited information regarding ego vehicles and its relatively modest volume.

Addressing the limitations of the earlier dataset, Rasouli et al. [36] introduced the Pedestrian Intention Estimation (PIE), a large-scale dataset distinct for its annotations of pedestrian crossing intentions. Various methods leveraging the PIE dataset for predicting pedestrian intentions have emerged, often involving modifications and expansions in network architecture. Rasouli et al. [16] implemented Stacked RNNs, while others have utilized camera-acquired information [37,38], employed graph convolutional networks [39], and further developed and refined the models [17–20]. Additionally, previous works [21–27] have performed comparative analyses using both JAAD and PIE datasets. The Stanford-TRI Intent Prediction (STIP) dataset [40] is noteworthy for including videos captured from three different camera angles. citepliu2020spatiotemporal employed graph convolution techniques to predict pedestrian intentions. However, a notable distinction exists in video characteristics recorded by onboard vehicle cameras compared to those captured by surveillance videos. Consequently, models trained with datasets designed for pedestrian crossing intention, such as PIE and JAAD, are not directly transferable for use in pedestrian safety monitoring systems that rely on surveillance footage.

Furthermore, other pedestrian-related datasets have been introduced. One such dataset, Trajectory Inference using Targeted Action prior Network (TITAN), introduced by Malla et al. [41], is distinguished by its inclusion of diverse behavioral labels for both vehicles and pedestrians. Additionally, [42] unveiled the Euro-PVI dataset, which encompasses a wide array of data on pedestrians and bicyclists. Predictive analyses within this dataset were conducted using a joint- β -cVAE approach. This trend towards sophisticated prediction models is mirrored in numerous other datasets, reflecting a growing emphasis on the nuanced understanding of pedestrian behaviors in diverse traffic situations.

While there have been endeavors to develop models specifically for surveillance videos [43–45], these models often exhibit lower predictive accuracy. This paper delineates our unique contribution to the field, distinct from existing research, by introducing an innovative approach to surveillance dataset construction. Our focus is squarely on anticipating pedestrian vulnerabilities, thereby offering a novel perspective on pedestrian safety research and intelligent surveillance systems.

2.3. Research gaps and motivation

Although various studies have addressed pedestrian detection and crossing intention prediction, most existing datasets and models are based on videos recorded from vehicle-mounted cameras, such as the JAAD and PIE datasets. These datasets focus on pedestrian crossing intentions and general pedestrian detection but lack the specific context and characteristics of surveillance videos, where pedestrians are observed from a fixed viewpoint. Moreover, many pedestrian detection models assume that all observed individuals are in a vulnerable situation without contextually differentiating safe and potentially vulnerable states. These gaps increase demand for a specialized approach that anticipates pedestrian vulnerabilities using surveillance videos with contextual annotations. Our work addresses these gaps by introducing the WatchoutPed dataset, specifically designed for vulnerable pedestrian anticipation in surveillance videos, and the VPANet model, which integrates visual and non-visual data to predict pedestrian states with high accuracy. By doing so, our contributions provide a robust foundation for advancing pedestrian safety monitoring in urban environments.

Table 1
Properties of the proposed WatchoutPed dataset.

	Property
Number of video clips	180
FPS (Frames per Second)	10
Length of each clip	30 s
Total number of frames	54K
Total number of instances	98,502
Safe class	43,363
Unsafe class	55,139

3. Dataset

To construct a dataset for anticipating vulnerable pedestrians, we utilize the videos publicized in [46]. These videos, recorded using surveillance cameras in school zones, are part of the dataset in [46]. This AIHub dataset [46] comprises video clips, each lasting 30 s, captured in Ultra High Definition (UHD) resolution (3840×2160 pixels) at a frame rate of 10 FPS. The dataset provides clip-level annotations highlighting dangerous pedestrian behavior, such as jaywalking, falling on the road, and walking in the driveway. In addition, the dataset provides frame-level annotations, encompassing bounding boxes for various objects such as vehicles, pedestrians, traffic lights, and cycles.

From the video clips, we selected 180 clips that distinctly capture pedestrians walking and crossing on roads or sidewalks. Table 1 shows the overall properties of our WatchoutPed dataset used in this study. For the purpose of model training, bounding boxes and corresponding ground truth labels for each pedestrian (categorized as either in a safe or unsafe state) were generated for every frame using our auto-labeling method, which is described in the following subsections. Please note that for the experiments, human-verified labels were also established for the experiments, with details provided in Section 5.5. During training, automatically generated labels were used, except in the ablation study in Section 5.5, which evaluates the effectiveness of these automatically generated labels compared to those refined by human intervention. For accurate evaluation of the trained models, human-verified labels are used during testing.

3.1. Auto-labeling method

Annotations, encompassing bounding boxes in the image and corresponding state labels of pedestrians for each frame, are essential for training models aimed at anticipating vulnerable pedestrians. However, the manual annotation process is time-consuming and labor-intensive. To address this challenge, we introduce an auto-labeling method that combines ground region estimation with pedestrian state estimation.

3.1.1. Ground region estimation

To determine whether each pedestrian is in a safe or unsafe state, it is necessary to know where the pedestrian is standing. Semantic segmentation results in the vicinity of the pedestrian can be used for this purpose. However, accurately recognizing ground regions in a single-camera setup poses significant challenges, primarily due to occlusions by moving objects like pedestrians, cyclists, and vehicles. To mitigate the occlusion, we adopt ground region estimation [47], focusing on the characteristic that background elements remain static over time in fixed surveillance cameras.

To utilize the temporal consistency of the background, the probability of ground-related classes of the semantic segmentation is averaged over frames as follows:

$$\mathbf{G}_i(\mathbf{x}) = \operatorname{argmax}_c \frac{\sum_{k=0}^i \mathbf{S}_{c,k}(\mathbf{x}) \circ \mathbf{M}_k(\mathbf{x})}{\sum_{k=0}^i \mathbf{M}_k(\mathbf{x})}, \quad (1)$$

where \mathbf{G}_i is the ground region map at the i th frame, $\mathbf{S}_{c,k}(\mathbf{x}) \in \mathbb{R}^{\mathbf{X} \times \mathbf{C}}$ is the probability map of each class c for every pixel \mathbf{x} obtained from the

semantic segmentation, $\mathbf{M}_k(\mathbf{x})$ is the binary mask that indicates whether pixel \mathbf{x} is associated with ground-related classes. Specifically, $\mathbf{M}_k(\mathbf{x})$ is set to 1 if the class with the highest probability at $\mathbf{S}_{c,k}(\mathbf{x})$ is one of the predefined ground-related classes, and 0 otherwise. This mask is used to filter out irrelevant pixels not related to the ground-related classes. \circ is the element-wise multiplication.

Please note that for semantic segmentation, we employ the Mask2Former [48] with a Swin-L backbone, which has been pre-trained on the Mapillary Vistas dataset [49]. For the ground-related classes, we specifically focus on the following classes that are relevant to ground elements and are frequently observed in the dataset [46] used in this paper: [curb, curb cut, road, sidewalk, lane marking (crosswalk), lane marking (general), and terrain].

3.1.2. Pedestrian state estimation

The pedestrian state can be determined by where the pedestrian stands in the road environment. Specifically, we assume that a pedestrian is in an unsafe state if they are positioned on a road or crosswalk. Conversely, their state is considered safe if they are situated on a sidewalk.

In a surveillance camera setup, the region around a pedestrian's feet is typically encompassed by the ground region they are standing on, due to the high positioning of the cameras. Based on this observation, the pedestrian state estimation is devised to determine whether the pedestrian is in a safe or unsafe state. Initially, pedestrian tracks are extracted using ByteTrack [50], trained on the CrowdHuman [12] and MOT20 [51] datasets. Following this, the location of each pedestrian is determined based on the predominant ground region class within the bottom 10% of their bounding box in the ground region map. If a pedestrian is positioned on the sidewalk, curb, curb cut, or terrain, their corresponding bounding box is labeled as 'safe'. Conversely, if they are located on lane markings or the road, the assigned label is 'unsafe'.

Through this process, a total of 98,502 instances were generated, encompassing 43,363 safe instances and 55,139 unsafe instances. It is important to note that any instances with a duration shorter than three seconds, including both observation and anticipation periods, were excluded from the dataset. Please note that any instances shorter than 30 frames are ignored, as the model requires 20 observation frames and predicts 10 frames after which will be described in the following Section 4.1.

Fig. 2 illustrates examples of both ground region estimation and pedestrian state estimation. The figure demonstrates the ability to generate the ground region map even in the presence of occlusions caused by moving objects. It also shows the pedestrian state estimation could generate reasonable annotations.

4. VPANet: Vulnerable Pedestrian Anticipation Network

4.1. Observation length and anticipation time

In this paper, we propose a Vulnerable Pedestrian Anticipation network (VPANet), which is designed to predict vulnerable pedestrians who are likely to be in an unsafe state on the road or crosswalk, thereby aiming to enhance pedestrian safety. The VPANet is tailored to improve the accuracy of forecasting pedestrian vulnerability based on their historical movement patterns. Our objective is to determine whether a pedestrian is in a vulnerable situation regarding safety in the future frame, based on their past movements. In our analysis, we define three critical temporal points, illustrated in Fig. 3: the experimental start frame, the current frame, and the prediction frame. The experimental start frame marks the commencement of tracking a pedestrian's trajectory. The current frame signifies the end of the observation period and the beginning of the predictive analysis. The duration from the experimental start frame to the current frame, known as the observation length (m), is the time window where we gather

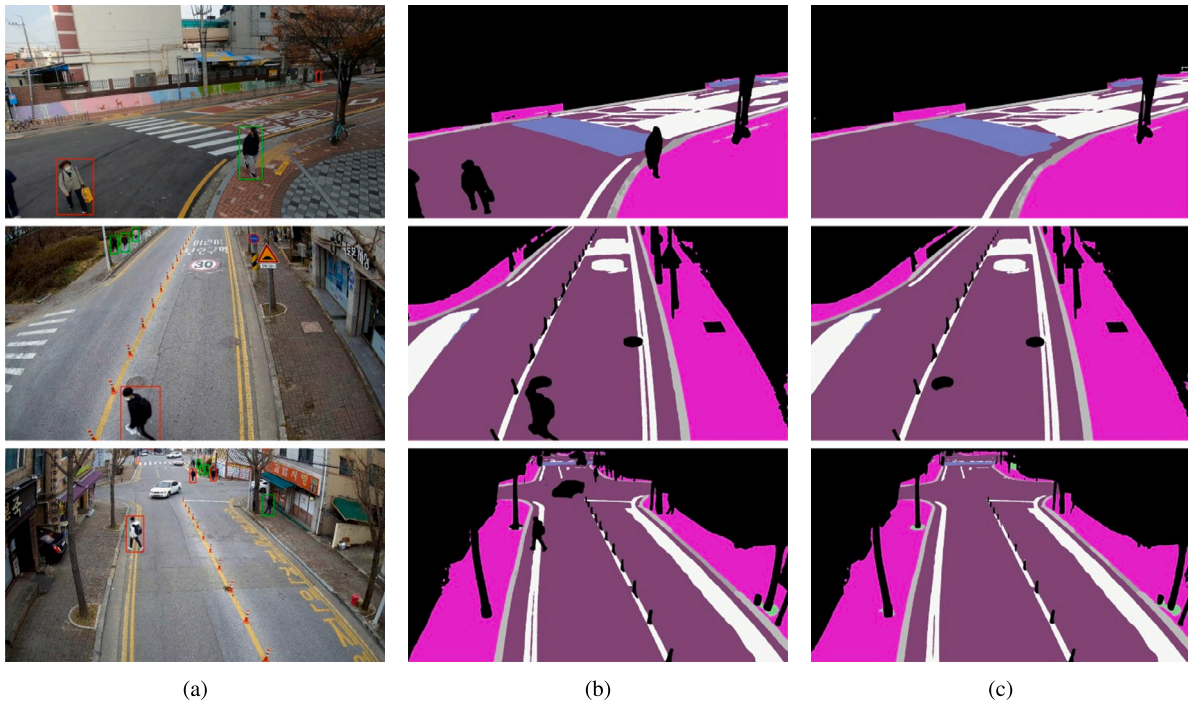


Fig. 2. Results of the auto-labeling method (a) input image with generated annotations (i.e., the red boxes indicate pedestrians in unsafe situations, while the green boxes indicate pedestrians in safe situations), (b) segmentation map, (c) ground region map.

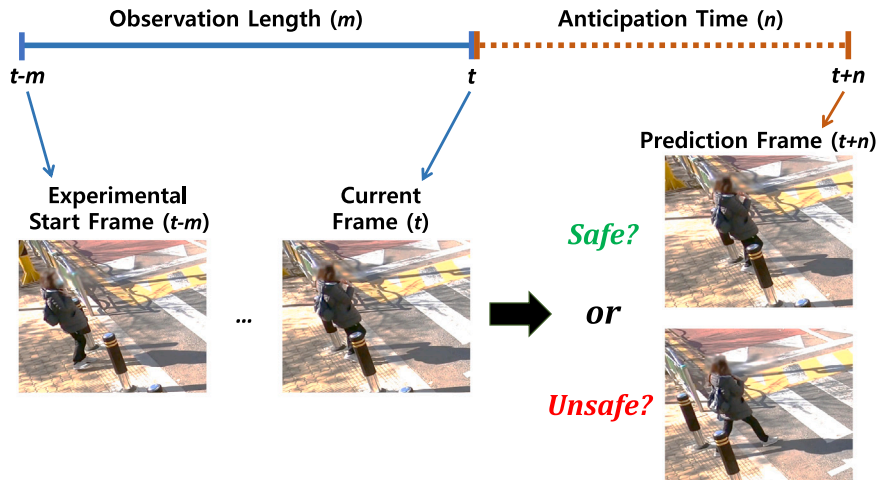


Fig. 3. Observation Length and Anticipation Time The observation length (m) and anticipation time (n) are defined, observing from frame $(t-m)$ to (t) . The frame at the start of observation is designated as the "Experimental Start Frame $(t-m)$ ", and the frame where observation ends and prediction begins is termed the "Current Frame (t) ". Following this, the prediction process starts, and the frame where anticipation concludes is labeled the "Prediction Frame $(t+n)$ ". This setup is used to anticipate whether a pedestrian is in a safe or unsafe situation, and the results of this prediction are then outputted.

data on the pedestrian's prior behavior; in our experiments, this is set to 20 frames as $(x_{t-19}, x_{t-18}, x_{t-17}, \dots, x_t)|t$, where x_t is the observation at frame t . Beyond the current frame, we proceed to the prediction phase, which concludes at the prediction frame. This frame represents the future point where we assess the pedestrian's risk of encountering unsafe conditions. The anticipation time in our model is the duration between the current frame and the prediction frame, which is a critical interval in our predictive analysis. The anticipation time is set to 10 frames to predict the pedestrian crossing intention 10 frames later as $(y_{t+1}, y_{t+2}, y_{t+3}, \dots, y_{t+10})|t$, where y_t is the prediction at frame t .

4.2. Problem definition

The proposed model uses observed information included in past frames from surveillance videos, including both visual and non-visual inputs. Depending on the type of input, the network is divided into two branches: visual and non-visual branches. In the visual branch, three types of visual inputs for each pedestrian are used: local context, global context, and local surround. Conversely, the non-visual branch solely utilizes one type of input, which is the bounding box. Subsequently, the output from each branch goes through an attention mechanism, ultimately aiming to predict pedestrians who are likely to be in an unsafe state. The overall structure is described in Fig. 4.

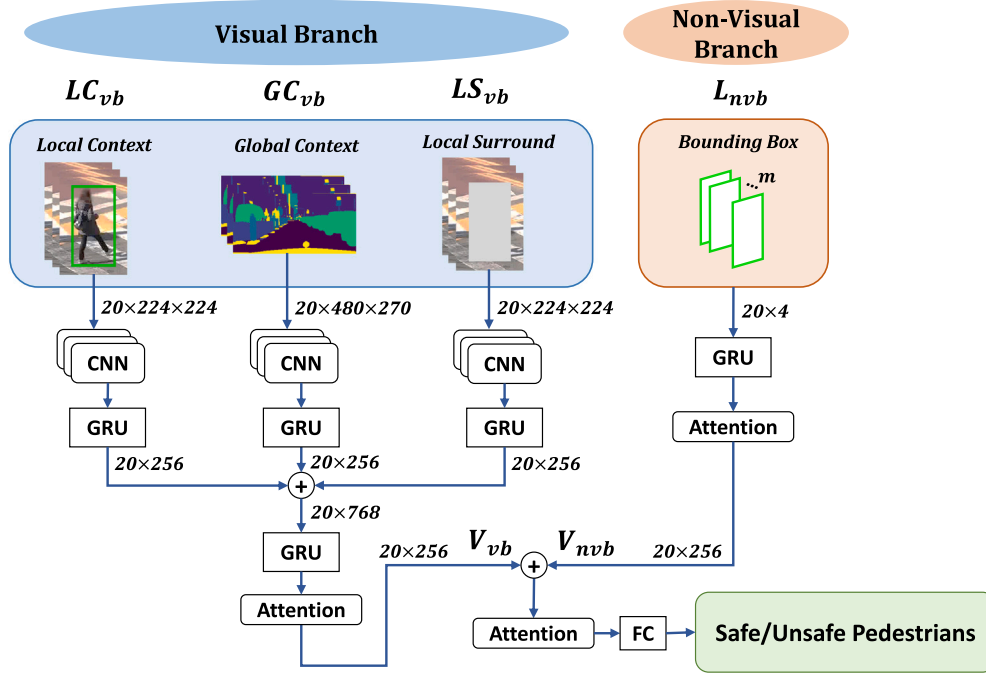


Fig. 4. Architecture of our proposed vulnerable pedestrian anticipation network (VPANet) VPANet is composed of four inputs: bounding box (L_{nvb}), local context (LC_{vb}), global context (GC_{vb}), local surround (LS_{vb}). In particular, the three visual inputs are integrated into a visual branch, each passing through a CNN and GRU being processed by an attention mechanism, resulting in the vector V_{vb} . The non-visual input, represented by the bounding box, independently passes through a GRU and an attention module, yielding the vector V_{nvb} , which is also concatenated with the output of the visual branch, V_{vb} . Ultimately, this integration extracts a 'safe' or 'unsafe' state.

4.3. Input acquisition

4.3.1. Visual branch

In the visual branch, three types of inputs are employed: local context, global context, and local surround. Depending on each input type, visual image features are extracted by a CNN backbone for each pedestrian observed in frames from videos. The CNN layers extract spatial features, allowing the model to capture key information about the pedestrian's behavior and the overall scene flow. We used the VGG19 [52] as the CNN backbone, which has been pre-trained on the ImageNet dataset, in conjunction with a max pooling layer. These features are subsequently fed into a gate recurrent unit (GRU) network, an RNN-based encoder. It is known that GRU, which is a simpler structure than LSTM, enhances efficiency and thus results in quicker training. The GRU layers capture temporal intent from sequential frames, playing a crucial role in tracking changes in the pedestrian's state and predicting their future movements. In this context, the GRU module comprises 256 hidden states. Following this, the outputs of the three input features are concatenated and fed into another GRU. The final output from this latter GRU is then passed through an attention mechanism to yield the vector V_{vb} , where "vb" abbreviates "visual branch", distinguishing the input features between the visual branch ("vb") and non-visual branch ("nvb").

The *local context* feature LC_{vb} is determined as:

$$LC_{vb} = \{lc_i^{t-m}, lc_i^{t-m+1}, \dots, lc_i^t\}, \quad (2)$$

where lc_i represents the local context image. In the equation for this input feature, i represents the ID of each individual pedestrian within an image frame. The ID i is sequentially assigned to each pedestrian appearing in a video and is used for numbering purposes to track past feature information. The variable t denotes the current frame, while m represents the observation length. Therefore, lc_i^{t-m} refers to the local context image from the frame that is m frames before the current frame t . Consequently, lc_i^{t-m+1} corresponds to the image from the next frame, lc_i^{t-m+2} represents the image from the following frame, and so on, up

to the final image at the current frame t , lc_i^t . These sequential past local context images are extracted and used as input to the model. Such notation interpretation is also consistently applied to other visual input feature equations.

The behavior of the pedestrian is profoundly influenced by interactions with surrounding traffic participants. The surrounding region contains various information that can greatly impact pedestrian behavior, such as traffic lights, intersections, crosswalks, and roads. Therefore, to precisely predict and anticipate the state of pedestrians walking on the road, it is essential to thoroughly consider the visual information of the target pedestrians and the objects they interact with in their surroundings. For this purpose, the input feature, lc_i , extracted the area around the target pedestrian by 1.5 times the size of the 2D bounding box coordinates and resized them to 224×224 pixels as RGB images. This cropped image is then processed through a pre-trained VGG19 network through a max pooling layer with 14×14 kernel for feature extraction. The output dimensions become $(m, 512)$, where 512 denotes the number of channels.

The *global context* feature GC_{vb} is determined as:

$$GC_{vb} = \{gc_i^{t-m}, gc_i^{t-m+1}, \dots, gc_i^t\}, \quad (3)$$

where gc_i denotes the semantic segmentation values. This value accurately distinguishes each object, providing detailed information that surpasses simple object recognition, indicating which sections are roads and which are pedestrians. Moreover, the results from this segmentation include valuable information that aids in understanding interactions within the roadway environment. The semantic map is extracted using the DeepLabV3 model pre-trained on the Cityscapes dataset [11] and encompasses both the global scene and road information. In the end, the features are outputted through 512 channels to match the dimensions of other input features.

The *local surround* feature LS_{vb} is determined as:

$$LS_{vb} = \{ls_i^{t-m}, ls_i^{t-m+1}, \dots, ls_i^t\}, \quad (4)$$

where t denotes the latest temporal frame. As previously mentioned in the *local context* section, the area surrounding the pedestrian predominantly contains information from traffic agents interacting with the pedestrian, which plays a substantial role in anticipation. Therefore, this input feature only utilizes the image area surrounding the pedestrian. l_{s_i} is extracted the area around the target pedestrian by 1.5 times the size of the 2D bounding box coordinates, but the region within the pedestrian bounding box grayed out, thereby emphasizing only the pedestrian's surrounding area. The input layer for *local surround* produces an output vector of dimensions $(m, 512)$, in conjunction with a 14×14 kernel max pooling layer.

4.3.2. Non-visual branch

As a non-visual input, the bounding box input is used. The *bounding box* feature L_{nvb} is determined as:

$$L_{nvb} = \{l_i^{t-m}, l_i^{t-m+1}, \dots, l_i^t\}, \quad (5)$$

where “*nvb*” is used as an abbreviation for “non-visual branch”. Similar to the visual input features, i represents the pedestrian ID sequentially assigned to each pedestrian appearing in the video, t denotes the current frame, and m represents the observation length. Therefore, l_i^{t-m} refers to the pedestrian's bounding box coordinates at frame $t-m$, l_i^{t-m+1} corresponds to the bounding box coordinates at frame $t-m+1$, and so on, until extracting the bounding box coordinates at the current frame t , denoted as l_i^t . These sequential coordinates are extracted and used as input. $l_i = [x_1, y_1, x_2, y_2] \in \mathbb{R}^4$ represents a 2D bounding box, which is composed of coordinates determined by the top-left $([x_1, y_1])$ and bottom-right $([x_2, y_2])$ of each pedestrian. The bounding box provides the exact location of the pedestrian. The bounding box consists of four coordinates, thus having a dimension of $m \times 4$.

4.4. Recurrent block

To capture the temporal dynamics of the input features, we employed a Gated Recurrent Unit (GRU) [53], a type of recurrent neural network that offers a simpler architecture compared to the Long Short Term Memory (LSTM) layer [54,55]. The GRU model operates through a series of gated mechanisms that control the flow of information. The equations governing the k th level of the GRU operation are as follows,

$$r_k^t = \sigma(W_k^{xr} x_k^t + h_k^{t-1} W_k^{hr} + c_k^r), \quad (6)$$

$$z_k^t = \sigma(W_k^{xz} x_k^t + h_k^{t-1} W_k^{hz} + c_k^z), \quad (7)$$

$$\hat{h}_k^t = \tanh(W_k^{xh} x_k^t + W_k^h (r_k^t \otimes h_k^{t-1}) + c_k), \quad (8)$$

$$h_k^t = (1 - z_k^t) \otimes \hat{h}_k^t + z_k^t \otimes h_k^{t-1}, \quad (9)$$

where σ denotes the sigmoid function, r_k^t is the reset gate. W represents the weight matrices, x_k^t is the input vector at time t , h_k^{t-1} is the previous hidden state, and c_k^r , c_k^z , c_k are the bias term. The update gate z_k^t is a vector, which takes as input the current input x_k^t , the previous hidden state h_k^{t-1} . The candidate hidden state \hat{h}_k^t is calculated using a hyperbolic tangent function. The reset gate r_k^t with h_k^{t-1} is multiplied, and the last hidden state is h_k^t , which combines the candidate hidden state and the previous hidden state to produce the current hidden state. Specifically, the update gate and reset gate effectively retain important information while filtering out unnecessary data, allowing information from past frames to be appropriately reflected in predictions for future frames.

4.5. Attention layer

The attention mechanism [56] is employed to emphasize relevant segments of the input data, improving the model's ability to interpret the features more effectively. The attention mechanism is designed to emphasize significant information in each frame, helping the model focus on essential elements for accurate prediction rather than on irrelevant features. By dynamically adjusting the importance of each frame in a sequence, the mechanism assigns higher weights to frames where the pedestrian's behavior changes, ensuring that this information is well-represented in the model's predictions. The attention output vector $\beta_{attention}$ is the result of the hyperbolic tangent function applied to the combined form of the attention-weighted hidden state's sum and the final hidden state of the encoder, all processed by the weight matrix W_{att} as below:

$$\beta_{attention} = \tanh(W_{att} [\sum_{weighted} ; h_f]), \quad (10)$$

where $\sum_{weighted}$ is the aggregated context after applying attention weights, and h_f represents the encoder's last hidden state. The aggregate context $\sum_{weighted}$ represents a composite of all previous hidden states from the encoder, each scaled by its corresponding attention weights is given by:

$$\sum_{weighted} = \sum_s \alpha_t h_{s_t}, \quad (11)$$

where h_{s_t} denotes a specific hidden state from the encoder's output sequence, and the attention weights α_t signify the relevance of each state. The attention weights α_t are calculated as follows:

$$\alpha_t = \frac{\exp(\text{score}(h_f, \tilde{h}_t))}{\sum_{s=1}^{T'} \exp(\text{score}(h_f, \tilde{h}_s))}, \quad (12)$$

The last hidden state and each previous hidden state are computed by taking the dot product of the transformed previous hidden state with the weight matrix P_{weight} :

$$\text{score}(h_f, \tilde{h}) = h_f^T P_{weight} \tilde{h}, \quad (13)$$

where P_{weight} is a trainable parameter that learns to weigh the relevance of each hidden state in context to the final hidden state. The output vector $\beta_{attention}$ captures the distilled information extracted from the input sequence, highlighting the model's capacity for dynamic feature analysis.

5. Experimental results

5.1. Evaluation metrics

To evaluate and compare the performance of models designed for predicting pedestrians in vulnerable situations, we employed five representative metrics: Accuracy (ACC), Precision, Recall, F1 score, and Area Under the Curve (AUC). These metrics are the most widely used in the research field for predicting pedestrian behavior.

Accuracy is a metric indicating how well the model has anticipated pedestrians being in vulnerable situations. In other words, it compares the binary values predicted by the model with the pedestrian state, expressing the proportion of correct predictions out of the total number of samples. The definition is as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \quad (14)$$

TP indicates the number of positive samples that were correctly forecasted to be positive, TN indicates the number of negative samples that were accurately forecasted to be negative, FP indicates the number of negative samples that were incorrectly forecasted to be positive, FN indicates the number of positive samples that were wrongly forecasted to be negative.

Precision refers to the ratio of samples that are predicted as positive by the model and are actually positive. The definition is as follows:

$$\text{Precision} = \frac{TP}{TP + FP}, \quad (15)$$

Recall, also known as sensitivity, indicates the ratio of samples that are actually positive among those that the model has predicted as positive. The definition is as follows:

$$\text{Recall} = \frac{TP}{TP + FN}, \quad (16)$$

F1 score represents the harmonic mean of precision and recall, taking a value between 0 and 1, with closer to 1 indicating better performance. It is particularly crucial for evaluating performance on imbalanced datasets. The definition is as follows:

$$\text{F1 score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}, \quad (17)$$

AUC [57] stands for Area Under the Curve and represents the area under the Receiver Operating Characteristics (ROC) curve. The ROC curve is frequently used to assess the performance of a model in classification problems. This curve plots the True Positive Rate (TPR) on the X -axis and the False Positive Rate (FPR) on the Y -axis. The AUC ranges between 0 and 1, with values closer to 1 indicating a better model. The definition is as follows:

$$\text{AUC} = \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n \mathbf{1}(s_i > s_j), \quad (18)$$

where m refers to the number of positive samples, n is the number of negative samples, s_i is the score of the i th positive sample, and s_j is the score of the j th negative sample. $\mathbf{1}$ is the indicator function, which returns 1 if the condition inside is true, and 0 otherwise. This formula examines all pairs of positive and negative samples and calculates the probability that the predicted score of a positive sample is higher than that of a negative sample.

5.2. Implementation details

In this study, the proposed model was implemented and trained on AMD EPYC 7513 32-core processors along with 8 Nvidia RTX A6000 GPUs, all operating within a TensorFlow framework. For the visual branch of our model, we employed GRUs configured with 256 hidden units. Additionally, for the training process of VPANet, we utilized the RMSProp optimizer, setting the learning rate at 5×10^{-5} , running for a total of 50 epochs on the WatchoutPed dataset. In an effort to curb the overfitting in the visual data, a dropout rate of 0.5 was applied following the attention block. The learning rate and dropout rate were determined through a grid search, which indicated the best configurations for training stability and model performance. The chosen values provided the best generalization, ensuring effective learning, and minimizing loss. We also introduced an L2 regularization with 0.0001 to the last fully connected layer, which helped further reduce overfitting. In the experiments, clips were divided into two subsets; 140 for training and 40 for testing. Accordingly, instances were assigned to their corresponding subsets. To prevent randomness in the experiments, each result was evaluated based on the average of five repeated trials.

5.3. Performance comparison

In our study, a comparative performance analysis was conducted on the WatchoutPed dataset, pitting our proposed model against six established benchmark models in pedestrian behavior prediction, as detailed in Table 2. The WatchoutPed dataset consists of videos recorded by CCTV, and VPANet is a model specifically developed for this dataset. However, since there were no models previously suitable for surveillance tasks, comparisons were made with existing pedestrian behavior prediction benchmark models. These benchmark models

include: MultiRNN [58], SingleRNN [17], Stacked with Multilevel Fusion RNN (SFRNN) [16], Pedestrian Crossing Prediction with Attention (PCPA) [18], Multi-Stream Network for Pedestrian Crossing Intention Prediction (MCIP) [26], and Crossing Intention Prediction Network based on Feature Fusion Modules (CIPF) [20]. Each model is distinguished by its unique visual encoder for image extraction and the integration of varied input feature types, offering a comprehensive assessment of their performance in comparison to our proposed solution.

Existing benchmark models generally conducted experiments on the PIE [36] or JAAD [59] datasets, which were recorded from within an ego-vehicle. Although the WatchoutPed dataset proposed in this paper is captured by CCTV, it is structurally the same as the PIE and JAAD datasets. Like the PIE-JAAD datasets, it is divided into an *annotations file* and an *annotations attributes file*. The *annotations file* contains mappings for each frame, specifying the pedestrian's bounding box coordinates, pedestrian id, and whether each of the nine behaviors is performed. The *annotations attributes file* records automatically calculated values for each frame, such as the 'experimental start frame', 'critical frame', and 'crossing frame', along with information on the pedestrian's gender and id. WatchoutPed dataset can also be directly applied to PIE-JAAD-specific models without additional adjustments, allowing straightforward performance comparisons across models. Additionally, each benchmark model utilizes up to six different features. In Table 2, the features used by each model can be found. 'L' represents Bounding Box coordinates, 'LC' stands for Local Context, and 'GC' for Global Context, while 'LS' denotes Local Surround; these input features are also utilized in our VPANet and are described in section 4.3. Although not used in VPANet, there are two additional input features utilized in existing benchmark models: 'SC' and 'LB'. 'SC' is an abbreviation for Scene Context, which refers to the entire image within a frame, including pedestrians, the ego-vehicle, and all agents participating in road traffic. 'LB' stands for Local Box, which is an image cropped to the pedestrian bounding box area; combining this image with the Local Surround image forms the Local Context image area.

The MultiRNN presents the advantage of outputting prediction uncertainties through Bayesian modeling, providing a certain level of understanding regarding incorrect forecasts in ambiguous situations. However, there is somewhat of a limitation in accuracy due to the sparse information utilized as inputs. Compared to our model, the MultiRNN model registers lower values by 9.49%, 9.83%, and 5.93% in terms of accuracy, AUC, and F1 score, respectively.

The SFRNN is a stacked RNN structure that relies on various levels of input features related to pedestrian crossing action, sequentially passing through GRU layers from bottom to top. Each modality is concatenated with the hidden state of the GRU from the previous level, serving as the input for the GRU in the next level. The hidden state of the GRU in the last level then becomes the final output prediction. Compared to our model, SFRNN demonstrates lower performance, falling behind by 5.21% in accuracy, 8.71% in AUC, and 3.11% in F1 score.

The SingleRNN employs an encoder-decoder structure utilizing RNN, where the last hidden state passes through a fully connected layer. This methodology is predominantly reliant on bounding boxes, lacking the integration of contextual information. As a result, it presents limitations in accurately capturing the movement and intricate details associated with pedestrian behavior. In comparison to our model, SingleRNN exhibits lower performance, with our model outperforming it by 9.49% in accuracy, 19.24% in AUC, and 5.93% in F1 score.

The PCPA stands out as a pioneering model, incorporating an attention weight vector into its design. Notably, this model utilizes a 3D convolutional branch for encoding local context as part of visual information. Following this, the 3D convolutional features are flattened and then passed through a fully connected layer. Our model has achieved significantly higher performance, outpacing PCPA by 14.14% in accuracy, 12.01% in AUC, and 9.17% in F1 score.

Table 2

Comparison of Model Anticipation Performance on the WatchoutPed Dataset. {*L*: Bounding Box, *LC*: Local Context, *GC*: Global Context, *SC*: Scene Context, *LB*: Local Box, *LS*: Local Surround}.

Model	Visual Encoder	Features	ACC	AUC	F1	Precision	Recall
MultiRNN [58]	VGG16	<i>L, LB</i>	0.81	0.77	0.88	0.92	0.84
SFRNN [16]	VGG16	<i>L, LB, LS</i>	0.84	0.78	0.90	0.92	0.87
SingleRNN [17]	VGG16	<i>L, LB, LS</i>	0.81	0.71	0.88	0.89	0.88
PCPA [18]	VGG19+C3D	<i>L, LC</i>	0.78	0.76	0.85	0.92	0.79
MCIP [26]	VGG19	<i>L, LC, GC</i>	0.87	0.84	0.92	0.95	0.89
CIPF [20]	VGG19+C3D	<i>L, LC, GC, SC, LB, LS</i>	0.85	0.77	0.91	0.91	0.90
VPANet(Ours)	VGG19	<i>L, LC, GC, LS</i>	0.89	0.85	0.93	0.95	0.91

The MCIP processes input features through separate non-visual and visual modules, uniquely incorporating a semantic segmentation map, which previous models do not employ. The model applies a multi-stream encoding technique, sequentially handling each input, enhancing its comprehension of the entire scene through the global context input. Despite its innovative approach, when compared to our model, MCIP has recorded lower performance, lagging behind by 2.07% in accuracy, 1.05% in AUC, and 1.53% in F1 score.

The CIPF skillfully integrates a diverse range of input features to achieve optimal results. The input data is systematically distributed across three modules, each designed to handle different types of features. Among these, the module responsible for extracting features of image inputs is divided into two sub-modules: one uses VGG19, while the other utilizes Convolutional 3D (C3D). Despite the extensive use of input features, our proposed model has achieved higher results, showing improvements of 4.6% in accuracy, 10.0% in AUC, and 2.7% in the F1 score.

Our proposed model, VPANet, utilizes four types of inputs, which are processed through separate visual and non-visual branches. For image feature extraction, we employ VGG19. This model has achieved the highest accuracy in predicting vulnerable pedestrians, reaching an anticipation accuracy of 89%, surpassing other benchmark models in comparison.

5.4. Qualitative examples

Fig. 5 illustrates the results of predicting the pedestrian state at $t + 1$ seconds, based on observations from the past 2 s from the current time (t seconds). In this visualization, red bounding box borders signify pedestrians forecasted to be in an unsafe state, whereas green borders indicate those predicted to be safe. The internal color of the bounding box represents the actual ground truth. In Fig. 5, the first and fourth rows depict scenarios where pedestrians, initially marked red for being in an unsafe state at time t , are predicted to shift to a safe state (indicated by the green border) one second later. This prediction aligns with the ground truth observed at $t + 1$ seconds, where these pedestrians indeed transition into a safe state. Conversely, in the second and third rows, pedestrians marked green for being in a safe state at time t are anticipated to move into an unsafe state (shown with the red border) one second later. This forecast is corroborated at $t + 1$ seconds, as these pedestrians are observed to be in an unsafe situation, demonstrating the predictive accuracy of the model.

Fig. 6 presents the qualitative results of an ablation study on VPANet, which consists of a Visual Branch and a Non-Visual Branch. This study examines the results when each branch is excluded. The first, second, and third columns show the predicted safety status of pedestrians one second into the future, as indicated by the bounding box borders around each pedestrian. A red box border predicts that the pedestrian will be in an unsafe situation one second later, while a green box border predicts safety. The inside of each pedestrian bounding box indicates whether the pedestrian's current state is safe or unsafe. The first column shows predictions made using only the visual branch inputs, the second column using only the non-visual branch inputs, and the third column shows the results when both branches of VPANet are utilized. Both the first and second rows demonstrate that predictions

solely based on the visual or non-visual branch inputs differ from the ground truth (GT). Correct predictions are only made when both branches are used together. Particularly in the second row, using each branch separately results in predictions that are opposite to the GT for both pedestrians shown. In the third row, even using only the visual branch inputs accurately predicts an unsafe situation for a pedestrian one second later. The fourth row shows that using only the non-visual branch inputs correctly predicts a safe situation for a pedestrian one second later. Thus, the highest prediction accuracy is achieved when both branches are appropriately utilized.

Fig. 7 shows the results of applying the proposed auto-labeling method and VPANet to other CCTV. These videos were captured by surveillance cameras installed and operated by a Korean local government. Our VPANet predicts pedestrian safety status without additional training and adjustments. In the first row, the model consistently and accurately predicts the pedestrian moving from the sidewalk to the road as an unsafe pedestrian (red box). The second row shows the prediction results for multiple pedestrians. As the pedestrians gradually move from the road to the sidewalk, the model gradually shifts its predictions from unsafe (red box) to safe (green box). Once they fully reach the sidewalk, the model accurately predicts all pedestrians as safe. The third row highlights the model's ability to accurately predict a pedestrian moving from the road to the sidewalk as safe (green box). This example is particularly noteworthy as it shows the model's effectiveness even in challenging conditions where snow makes it difficult to distinguish between the road and the sidewalk.

5.5. Feasibility of the proposed auto-labeling method

To evaluate the feasibility of the proposed auto-labeling method, we also establish human-verified labels. This verification process involved six annotators who precisely reviewed the labels for all video clips. For this task, the bounding box and the associated pedestrian state class produced by the auto-labeling method were imported to CVAT [60], enabling the annotators to verify each pedestrian state class by observing the videos within CVAT. Note that for the human verification process, an extra attribute was added to each bounding box. This attribute comprised five categories: unchanged, unsafe-start, unsafe-end, safe-start, and safe-end, as shown in Fig. 8. If an incorrect class was assigned any period, annotators marked the exact start and end frames of the pedestrian's state class using these additional attributes.

The human-verified dataset was then generated by importing these additional attributes from CVAT. Out of the 98,502 instances in the dataset, labels for 3,800 instances were revised for accuracy. Specifically, 3,482 instances were initially auto-labeled as unsafe, and 318 instances labeled as safe were corrected by the annotators. In our experiments, the proposed models were trained using both the automatically generated and the human-verified labels in the training set. While performances were evaluated on a testing set with human-verified labels for accurate evaluation.

Table 3 presents the experimental results of predicting vulnerable pedestrians in an unsafe state using both the automatically generated labels and the human-verified labels for training. While the AUC and precision performances remained consistent between both labels, the accuracy, F1-score, and recall were only 1% lower for the automatically

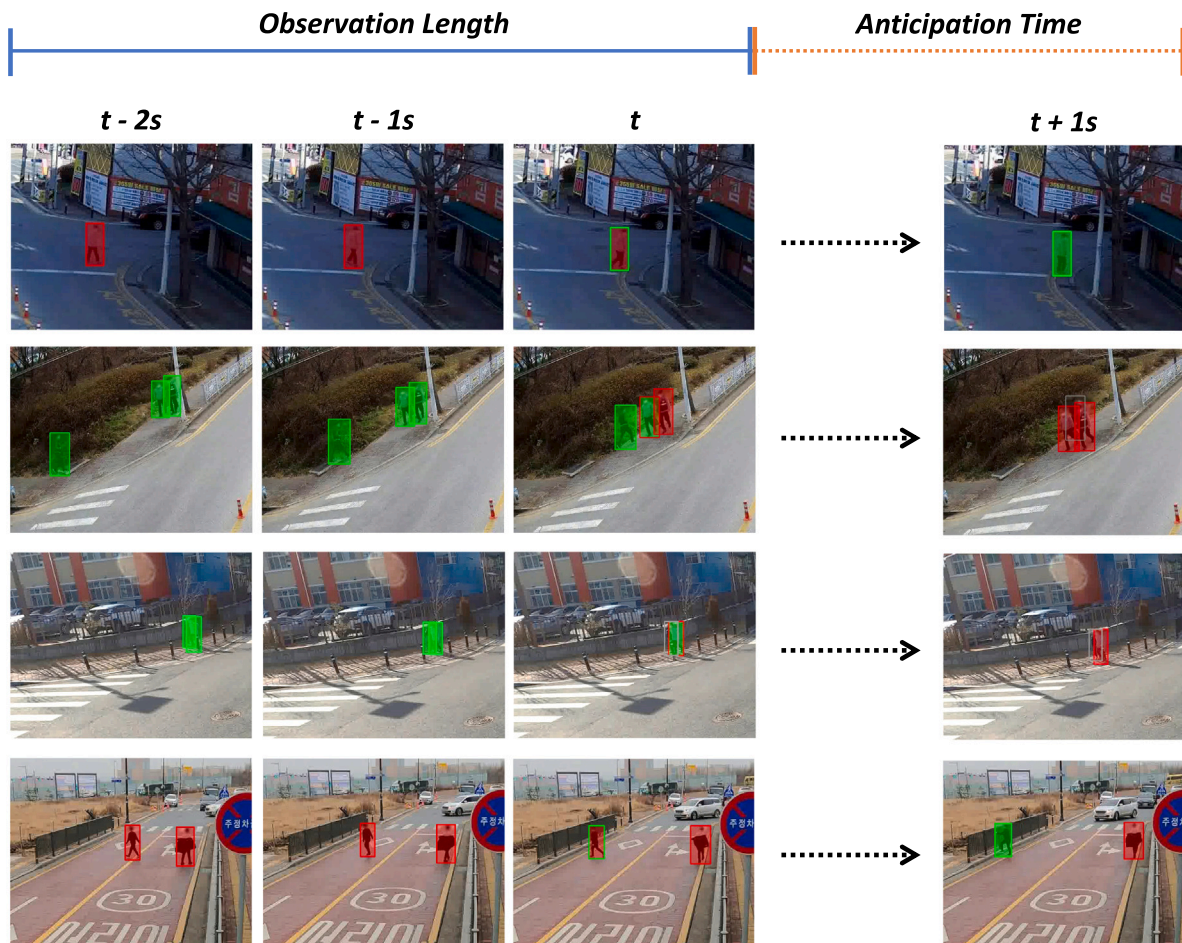


Fig. 5. Qualitative Examples. The color inside the bounding box represents the ground truth, and the color of the border represents the predicted value. Green represents safe pedestrians, while red represents unsafe pedestrians.

Table 3
Comparison of Effects Depending on Training Labels.

Training labels	ACC	AUC	F1	Precision	Recall
Automatically generated	0.88	0.85	0.92	0.95	0.90
Human-verified	0.89	0.85	0.93	0.95	0.91

generated labels than the human-verified version of the training set. These results indicate that some errors exist on the automatically generated labels did not significantly affect the anticipation performances. This suggests that the large-scale dataset created using our proposed auto-labeling technique can achieve performance near state-of-the-art results, without the need for exhaustive human verification.

5.6. Impact of the training set on performance

To explore how the predictive performance of our models on the WatchoutPed dataset is influenced by varying the size of the training set, we carried out a series of systematic experiments, the results of which are outlined in Table 4. We methodically increased the size of the training set, beginning with a baseline of 60 instances and incrementally adding in sets of 20, progressing through 80, 100, 120, and up to a maximum of 140 instances. During these experiments, the size of the test set was kept constant to ensure uniformity in performance comparison across different training set sizes. This approach provided valuable insights into the scalability and efficiency of our models in relation to the volume of training data.

Additionally, to ensure consistent representation of behavior types, each training set from the WatchoutPed dataset was structured to contain a fixed proportion of the nine distinct pedestrian behaviors. We defined the proportion of each action as follows to construct a balanced dataset.: driveway walk (11%), fall down (9%), fight (4%), jaywalk (16%), put umbrella (8%), ride cycle (17%), ride motorcycle (11%), ride kick (15%), and suddenly appear (9%). The allocation was designed so that behaviors directly related to pedestrian dynamics, such as walking on roads and sidewalks, constituted over 10% of the dataset. In contrast, behaviors with a more indirect relation to pedestrian safety were limited to less than 10%. When trained with only 60 clips, the VPANet recorded its lowest performance across all metrics, achieving an accuracy of 78%. However, as the training set size increased, there was a notable improvement in performance for all metrics, excluding AUC. Specifically, the anticipation accuracy showed a steady upward trend, ultimately reaching a peak of 88% with a training set of 140 clips. There was a minor decrease in AUC by 0.01 when the training set expanded from 100 to 120 clips, but this metric rebounded to 0.85 with the inclusion of 140 clips. These findings underscore that augmenting the training set size, particularly with data generated via our auto-labeling technique, leads to enhanced anticipation performance. Nevertheless, it is crucial to consider that expanding the training set beyond 140 clips could disproportionately dominate the test set, potentially leading to overfitting and a consequent dip in performance. Thus, maintaining an optimal training-to-testing set ratio is essential to avoid overfitting and ensure robust model performance.

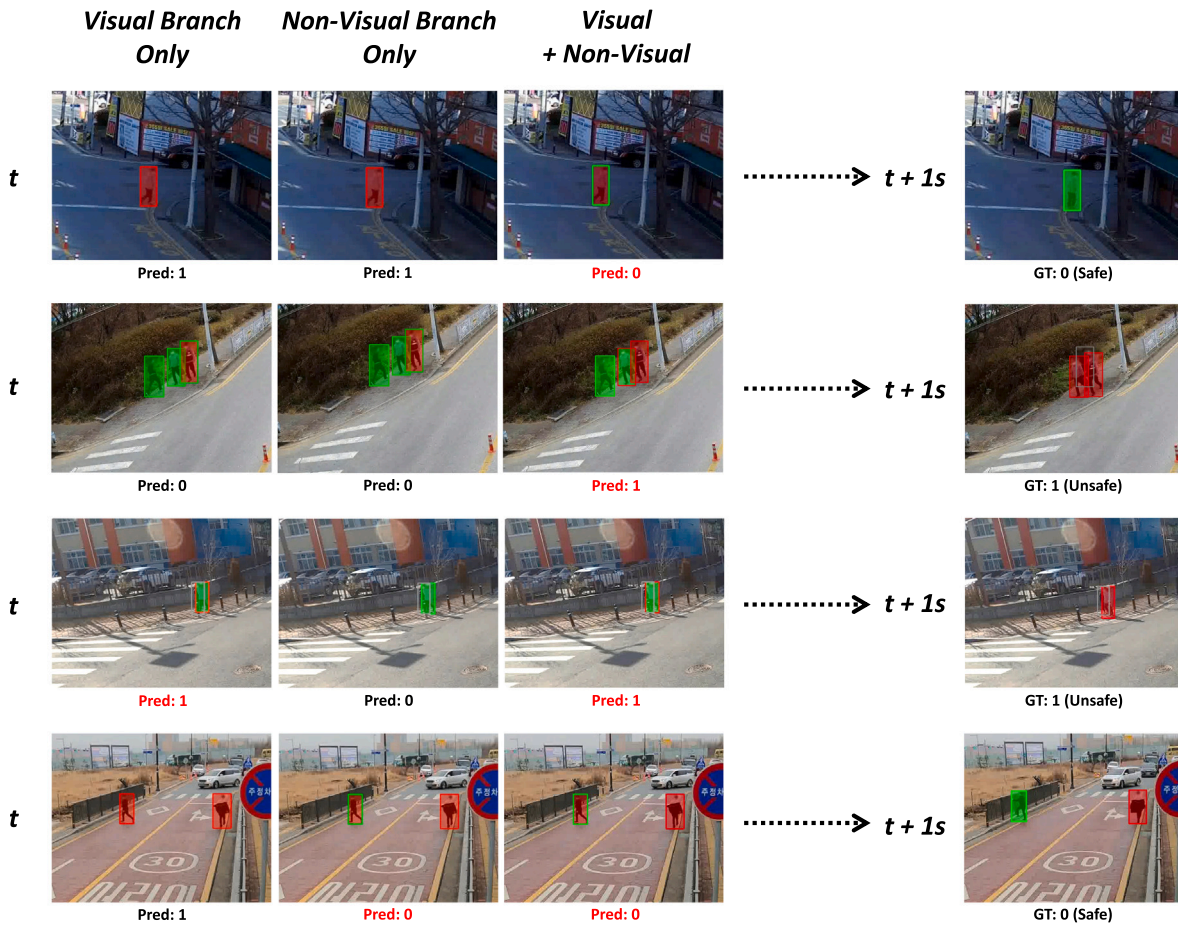


Fig. 6. Qualitative examples from an ablation study based on each branch. The first column shows the results predicted using only the visual branch, the second column using only the non-visual branch, and the third column shows the predictions using both branches of VPANet one second later. The border of each pedestrian bounding box represents the predicted value, with red indicating that the pedestrian is predicted to be in an unsafe situation one second later, and green indicating that the pedestrian is predicted to be in a safe situation. The inside of each pedestrian bounding box represents the current state of the pedestrian.

Table 4
Performance Comparison with Increasing the Number of Clips for Training.

# of clips	ACC	AUC	F1	Precision	Recall
60	0.78	0.76	0.85	0.92	0.79
80	0.79	0.80	0.86	0.94	0.79
100	0.83	0.85	0.89	0.96	0.82
120	0.86	0.84	0.91	0.95	0.87
140	0.88	0.85	0.92	0.95	0.90

5.7. Ablation study on input features

To evaluate the impact of each input feature on our proposed VPANet, we conducted an ablation study by removing each feature and observing the resultant performance changes. The results, as presented in Table 5, clearly indicate that every input feature contributes positively to the model’s accuracy. This finding validates the efficacy of our feature fusion process, confirming the importance of each feature in enhancing the overall predictive performance of the model.

Firstly, removing the global context derived from object segmentation on the road resulted in the most significant performance decline. The performance comparison is as follows: anticipation accuracy dropped from 0.88 to 0.83, AUC from 0.85 to 0.81, and F1-score from 0.92 to 0.89 (where the first number represents the performance with all four input features included, and the second number shows the performance without the global context input). This highlights the critical role of semantic information from the global context, which provides

insights into pedestrian behavior, vehicular dynamics, and road conditions, in enhancing the model’s predictive capability. Subsequently, excluding the local context, which includes not only the pedestrian but also the surrounding image, led to the second-largest performance decline. Compared to the full input performance of 0.88, anticipation accuracy fell to 0.84 when the local context was excluded. The AUC remained the same at 0.85, while the F1-score decreased from 0.92 to 0.89. Since the local context captures crucial elements such as nearby vehicles, and crosswalks, which can affect pedestrian movements, it is evident that excluding it would engender a substantial performance decrement. Interestingly, in the absence of pedestrian bounding box coordinates, there was only a slight decline in anticipation accuracy, dropping from 0.88 to 0.86. This indicates that just the coordinates do not have the necessary information to predict the state of a pedestrian, therefore coordinates alone are insufficient to determine the pedestrian’s state. Finally, when considering only the area surrounding the pedestrian without the pedestrian (local surround), this led to the smallest decline, with accuracy dropping slightly from 0.88 to 0.87, while the AUC and F1-score remained unchanged. This minimal change indicates that it is difficult to predict the state of a pedestrian using only the surrounding information, without the presence of the pedestrian.

6. Conclusion

This paper addresses the critical issue of pedestrian safety in intelligent traffic systems by focusing on predicting vulnerable pedestrians at risk of vehicular collisions in an unsafe state. In contrast to most current

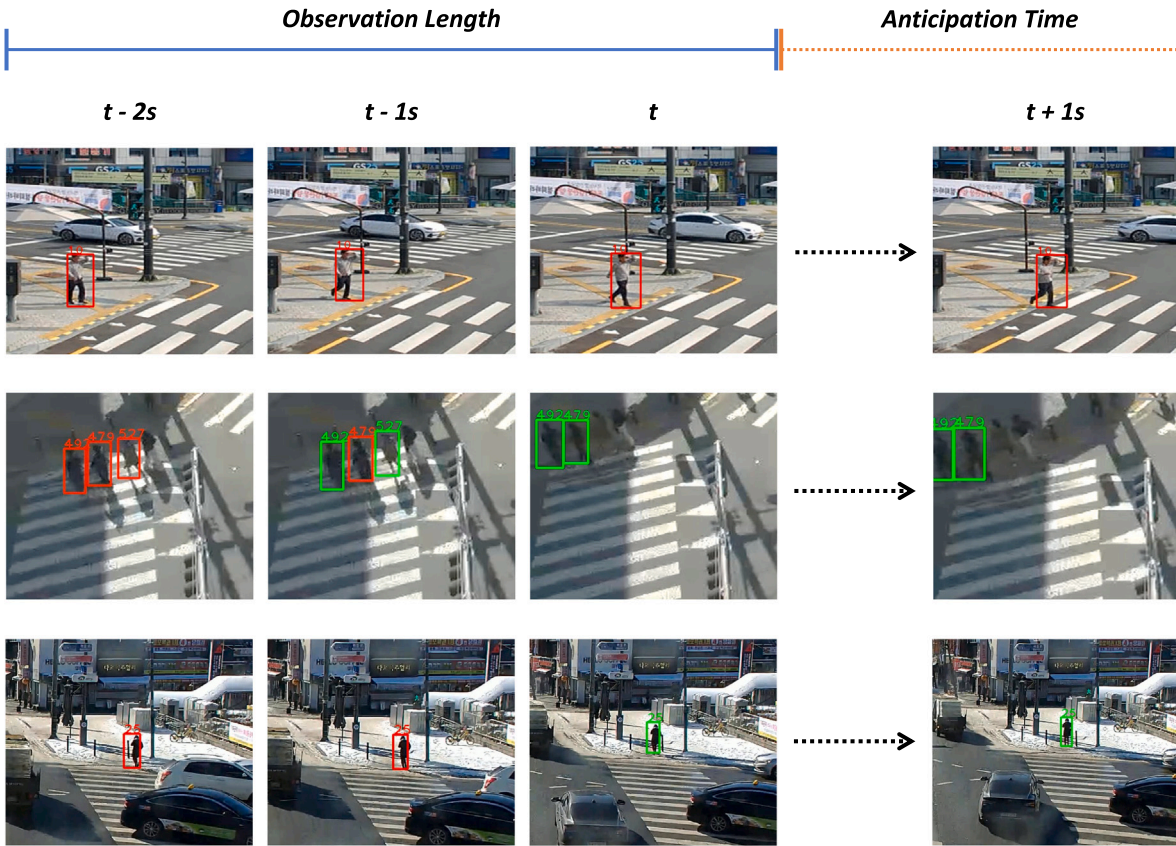


Fig. 7. Qualitative Examples for Generalization. These results include predictions on other surveillance videos. The green box indicates a safe pedestrian, while the red box indicates an unsafe pedestrian.



Fig. 8. Example of human verification using CVAT.

Table 5
Comparison of the Effects of the Four Input Features on VPANet Performance.

Local Context	Global Context	Local Surround	Bounding Box	ACC	AUC	F1	Precision	Recall
✓	✓	✓	✓	0.88	0.85	0.92	0.95	0.90
–	✓	✓	✓	0.84-0.04	0.85	0.89-0.03	0.97+0.02	0.83-0.07
✓	–	✓	✓	0.83-0.05	0.81-0.04	0.89-0.03	0.94-0.01	0.85-0.05
✓	✓	–	✓	0.87-0.01	0.85	0.92	0.95	0.88-0.02
✓	✓	✓	–	0.86-0.02	0.85	0.91-0.01	0.95	0.87-0.03

methodologies limited to pedestrian detection, tracking, and crossing intention estimation, our study pioneers the task of anticipating pedestrian vulnerability, especially when pedestrians are on the road. Our notable contributions encompass the development of the WatchoutPed dataset via an innovative auto-labeling method and the formulation of the Vulnerable Pedestrian Anticipation Network (VPANet). VPANet utilizes both visual and non-visual information from prior surveillance footage to determine the impending state of pedestrians, distinguishing between safety and risk. Demonstrating superior performance, VPANet surpasses state-of-the-art pedestrian intention prediction methods by achieving an accuracy of 98% on the WatchoutPed dataset. Further validating our methodology, models trained with automatically generated annotations by our labeling technique show performance parity with human-verified annotations, with a marginal discrepancy of less than 1%. This attests to the potential of our auto-labeling approach in generating additional annotations for model refinement in target-specific surveillance videos.

Our research contributes to the field of knowledge-based systems by providing a robust framework for intelligent decision support and data-driven optimization in urban safety. We envision that our WatchoutPed dataset and the VPANet model will be instrumental for ongoing research and enhancements in this field, particularly in applications related to intelligent urban safety infrastructures. As a future work, we aim to apply our approach to CCTV systems in real-world settings through collaboration with local government, integrating alert mechanisms to enhance pedestrian safety. Additionally, we seek to further improve the adaptability and effectiveness of our VPANet in diverse urban environments. These advancements will help develop a more robust and practical system for pedestrian safety applications.

CRedit authorship contribution statement

Je-Seok Ham: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Data curation, Conceptualization. **Dae Hoe Kim:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Data curation. **Jinyoung Moon:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Funding acquisition, Conceptualization.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT in order to improve language and readability. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Jinyoung Moon reports financial support was provided by Institute for Information Communication Technology Planning and Evaluation. Jinyoung Moon reports financial support was provided by Korea Ministry of Science and ICT. Dae Hoe Kim, Je-Seok Ham, and Jinyoung Moon has patent #18/584703 pending to ETRI. Je-Seok Ham, Dae Hoe Kim, and Jinyoung Moon has patent #18/592497 pending to ETRI. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by IITP grant funded by the Korea government (MSIT) (No. RS-2020-II200004, Development of Previsional Intelligence based on Long-term Visual Memory Network). This work used videos included in datasets from 'The Open AI Dataset Project (AI-Hub, S. Korea)'. All video data information can be accessed through 'AI-Hub (www.aihub.or.kr)'.

Data availability

The authors do not have permission to share data.

References

- [1] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.C. Berg, SSD: Single shot MultiBox detector, in: *Proceedings of the European Conference on Computer Vision, ECCV, Springer International Publishing, 2016*, pp. 21–37.
- [2] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, MobileNets: Efficient convolutional neural networks for mobile vision applications, 2017, arXiv preprint arXiv:1704.04861.
- [3] H. Song, I.K. Choi, M.S. Ko, J. Bae, S. Kwak, J. Yoo, Vulnerable pedestrian detection and tracking using deep learning, in: *International Conference on Electronics, Information, and Communication, ICEIC, 2018*, pp. 1–2.
- [4] C.Y. Wang, H.Y.M. Liao, Y.H. Wu, P.Y. Chen, J.W. Hsieh, I.H. Yeh, Cspnet: A new backbone that can enhance learning capability of CNN, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2020*.
- [5] S. Mehtab, W.Q. Yan, Flexible neural network for fast and accurate road scene perception, *Multimedia Tools Appl.* 81 (5) (2022) 7169–7181.
- [6] W. Mo, W. Zhang, H. Wei, R. Cao, Y. Ke, Y. Luo, PVDet: Towards pedestrian and vehicle detection on gigapixel-level images, *Eng. Appl. Artif. Intell.* 118 (2023) 105705, <http://dx.doi.org/10.1016/j.engappai.2022.105705>, URL <https://www.sciencedirect.com/science/article/pii/S0952197622006959>.
- [7] Z.R. Tang, R. Hu, Y. Chen, Z.H. Sun, M. Li, Multi-expert learning for fusion of pedestrian detection bounding box, *Knowl.-Based Syst.* 241 (2022) 108254, <http://dx.doi.org/10.1016/j.knosys.2022.108254>.
- [8] W. Wu, Q. Jiao, H.-S. Wong, G. Li, S. Wu, Learning scene-adaptive pseudo annotations for pedestrian detection in semi-supervised scenarios, *Knowl.-Based Syst.* 243 (2022) 108439, <http://dx.doi.org/10.1016/j.knosys.2022.108439>.
- [9] H. Yao, Y. Zhang, H. Jian, L. Zhang, R. Cheng, Nighttime pedestrian detection based on fore-background contrast learning, *Knowl.-Based Syst.* 275 (2023) 110719, <http://dx.doi.org/10.1016/j.knosys.2023.110719>.
- [10] P. Dollar, C. Wojek, B. Schiele, P. Perona, Pedestrian detection: A benchmark, in: *CVPR, 2009*, pp. 304–311.
- [11] S. Zhang, R. Benenson, B. Schiele, CityPersons: A diverse dataset for pedestrian detection, in: *CVPR, 2017*, pp. 4457–4465.
- [12] S. Shao, Z. Zhao, B. Li, T. Xiao, G. Yu, X. Zhang, J. Sun, CrowdHuman: A benchmark for detecting human in a crowd, 2018, arXiv preprint arXiv:1805.00123.
- [13] Y. Wang, Z. Guo, C. Xu, J. Lin, A multimodal stepwise-coordinating framework for pedestrian trajectory prediction, *Knowl.-Based Syst.* (2024) 112038.
- [14] H. Zhou, X. Yang, M. Fan, H. Huang, D. Ren, H. Xia, Static-dynamic global graph representation for pedestrian trajectory prediction, *Knowl.-Based Syst.* 277 (2023) 110775.
- [15] F. Gao, W. Huang, L. Weng, Y. Zhang, SIF-TF: A scene-interaction fusion transformer for trajectory prediction, *Knowl.-Based Syst.* 294 (2024) 111744.
- [16] A. Rasouli, I. Kotseruba, J.K. Tsotsos, Pedestrian action anticipation using contextual feature fusion in stacked rnns, 2020, arXiv preprint arXiv:2005.06582.
- [17] I. Kotseruba, A. Rasouli, J.K. Tsotsos, Do they want to cross? understanding pedestrian intention for behavior prediction, in: *2020 IEEE Intelligent Vehicles Symposium, IV, IEEE, 2020*, pp. 1688–1693.
- [18] I. Kotseruba, A. Rasouli, J.K. Tsotsos, Benchmark for evaluating pedestrian action prediction, in: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, WACV, 2021*, pp. 1258–1268.
- [19] D. Yang, H. Zhang, E. Yurtsever, K.A. Redmill, Ü. Özgüner, Predicting pedestrian crossing intention with feature fusion and spatio-temporal attention, *IEEE Trans. Intell. Veh.* 7 (2) (2022) 221–230.
- [20] J.-S. Ham, D.H. Kim, N. Jung, J. Moon, CIPF: Crossing intention prediction network based on feature fusion modules for improving pedestrian safety, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2023*, pp. 3665–3674.
- [21] K. Kim, Y.K. Lee, H. Ahn, S. Hahn, S. Oh, Pedestrian intention prediction for autonomous driving using a multiple stakeholder perspective model, in: *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, IEEE, 2020*, pp. 7957–7962.
- [22] S. Zhao, H. Li, Q. Ke, L. Liu, R. Zhang, Action-vit: Pedestrian intent prediction in traffic scenes, *IEEE Signal Process. Lett.* 29 (2021) 324–328.
- [23] Y. Yao, E. Atkins, M.J. Roberson, R. Vasudevan, X. Du, Coupling intent and action for pedestrian crossing behavior prediction, 2021, arXiv preprint arXiv:2105.04133.
- [24] N. Osman, E. Cancelli, G. Camporese, P. Coscia, L. Ballan, Early pedestrian intent prediction via features estimation, in: *2022 IEEE International Conference on Image Processing, ICIP, IEEE, 2022*, pp. 3446–3450.
- [25] A. Rasouli, T. Yau, M. Rohani, J. Luo, Multi-modal hybrid architecture for pedestrian action prediction. arXiv 2020. arXiv preprint arXiv:2012.00514.

- [26] J.-S. Ham, K. Bae, J. Moon, MCIP: Multi-stream network for pedestrian crossing intention prediction, in: *Computer Vision—ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part I, 2022*, pp. 663–679.
- [27] M. Dong, Pedestrian cross forecasting with hybrid feature fusion, in: *International Conference on Learning Representations 2023 Workshop on Scene Representations for Autonomous Driving*, 2023.
- [28] V. Venkatraman, C.M. Richard, K. Magee, K. Johnson, Countermeasures that work: A highway safety countermeasures guide for State Highway Safety Offices, 10th edition, Technical Report, National Highway Traffic Safety Administration(NHTSA), 2021.
- [29] A. Geiger, P. Lenz, R. Urtasun, Are we ready for autonomous driving? The KITTI vision benchmark suite, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2012*, pp. 3354–3361.
- [30] A. Rasouli, I. Kotseruba, 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, 2017*, pp. 206–213.
- [31] Z. Fang, A.M. López, Is the pedestrian going to cross? answering by 2d pose estimation, in: *2018 IEEE Intelligent Vehicles Symposium, IV, IEEE, 2018*, pp. 1271–1276.
- [32] J. Lorenzo, I. Parra, F. Wirth, C. Stiller, D.F. Llorca, M.A. Sotelo, Rnn-based pedestrian crossing prediction using activity and pose-related features, in: *2020 IEEE Intelligent Vehicles Symposium, IV, IEEE, 2020*, pp. 1801–1806.
- [33] H. Razali, T. Mordan, A. Alahi, Pedestrian intention prediction: A convolutional bottom-up multi-task approach, *Transp. Res. Part C: Emerg. Technol.* 130 (2021) 103259.
- [34] A. Singh, U. Suddamalla, Multi-input fusion for practical pedestrian intention prediction, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021*, pp. 2304–2311.
- [35] S.A. Bouhsain, S. Saadatnejad, A. Alahi, Pedestrian intention prediction: A multi-task perspective, 2020, arXiv preprint arXiv:2010.10270.
- [36] A. Rasouli, I. Kotseruba, T. Kunic, J.K. Tsotsos, Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019*, pp. 6262–6271.
- [37] W.M. Alvarez, F.M. Moreno, O. Sipele, N. Smirnov, C. Olaverri-Monreal, Autonomous driving: Framework for pedestrian intention estimation in a real world scenario, in: *2020 IEEE Intelligent Vehicles Symposium, IV, IEEE, 2020*, pp. 39–44.
- [38] L. Neumann, A. Vedaldi, Pedestrian and ego-vehicle trajectory prediction from monocular camera, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021*, pp. 10204–10212.
- [39] T. Chen, R. Tian, Z. Ding, Visual reasoning using graph convolutional networks for predicting pedestrian crossing intention, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021*, pp. 3103–3109.
- [40] B. Liu, E. Adeli, Z. Cao, K.-H. Lee, A. Sheno, A. Gaidon, J.C. Nibbles, Spatiotemporal relationship reasoning for pedestrian intent prediction, *IEEE Robot. Autom. Lett.* 5 (2) (2020) 3485–3492.
- [41] S. Malla, B. Dariush, C. Choi, Titan: Future forecast using action priors, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020*, pp. 11186–11196.
- [42] A. Bhattacharyya, D.O. Reino, M. Fritz, B. Schiele, Euro-pvi: Pedestrian vehicle interactions in dense urban centers, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021*, pp. 6408–6417.
- [43] W. Zhou, Y. Liu, L. Zhao, S. Xu, C. Wang, Pedestrian crossing intention prediction from surveillance videos for over-the-horizon safety warning, *IEEE Trans. Intell. Transp. Syst.* (2023).
- [44] C. Chan, Prediction of pedestrian crossing behavior based on surveillance video, *Sensors* 22 (2022).
- [45] H. Shin, K.-I. Na, J. Chang, T. Uhm, Multimodal layer surveillance map based on anomaly detection using multi-agents for smart city security, *ETRI J.* 44 (2) (2022) 183–193.
- [46] AIHUB2021, Video dataset of children’s dangerous behavior while walking on the road in school zones, 2021, [Online]. Available: <https://aihub.or.kr/aihubdata/data/view.do?dataSetSn=169>.
- [47] D. Kim, J. Moon, PedRiskNet: Classifying pedestrian situations in surveillance videos for pedestrian safety monitoring in smart cities, in: *IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS, 2024*.
- [48] B. Cheng, I. Misra, A.G. Schwing, A. Kirillov, R. Girdhar, Masked-attention mask transformer for universal image segmentation, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2022*, pp. 1280–1289, <http://dx.doi.org/10.1109/CVPR52688.2022.00135>.
- [49] G. Neuhold, T. Ollmann, S.R. Bulò, P. Kotschieder, The mapillary vistas dataset for semantic understanding of street scenes, in: *IEEE/CVF International Conference on Computer Vision, ICCV, 2017*, pp. 5000–5009, <http://dx.doi.org/10.1109/ICCV.2017.534>.
- [50] Y. Zhang, P. Sun, Y. Jiang, D. Yu, F. Weng, Z. Yuan, P. Luo, W. Liu, X. Wang, ByteTrack: Multi-object tracking by associating every detection box, in: *ECCV, 2022*.
- [51] P. Dendorfer, H. Rezatofighi, A. Milan, J. Shi, D. Cremers, I. Reid, S. Roth, K. Schindler, L. Leal-Taixé, MOT20: A benchmark for multi object tracking in crowded scenes, 2020, arXiv preprint arXiv:2003.09003.
- [52] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.
- [53] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder–decoder for statistical machine translation, in: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014*, pp. 1724–1734, <http://dx.doi.org/10.3115/v1/D14-1179>.
- [54] S. Zhang, M. Abdel-Aty, J. Yuan, P. Li, Prediction of pedestrian crossing intentions at intersections based on long short-term memory recurrent neural network, *Transp. Res. Rec.* 2674 (4) (2020) 57–65.
- [55] R. Quan, L. Zhu, Y. Wu, Y. Yang, Holistic LSTM for pedestrian trajectory prediction, *IEEE Trans. Image Process.* 30 (2021) 3229–3239.
- [56] T. Luong, H. Pham, C.D. Manning, Effective approaches to attention-based neural machine translation, in: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015*, pp. 1412–1421, <http://dx.doi.org/10.18653/v1/D15-1166>.
- [57] T. Calders, S. Jaroszewicz, Efficient AUC optimization for classification, in: *Knowledge Discovery in Databases: PKDD 2007, 2007*, pp. 42–53, http://dx.doi.org/10.1007/978-3-540-74976-9_8.
- [58] A. Bhattacharyya, M. Fritz, B. Schiele, Long-term on-board prediction of people in traffic scenes under uncertainty, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2018*, pp. 4194–4202.
- [59] A. Rasouli, I. Kotseruba, J.K. Tsotsos, Are they going to cross? A benchmark dataset and baseline for pedestrian crosswalk behavior, 2017, pp. 206–213.
- [60] CVAT, Computer vision annotation tool (CVAT), 2018, [Online]. Available: <https://github.com/opencv/cvat>.