IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

Detection of Texting While Walking in Occluded Environment Using Variational Autoencoder for Safe Mobile Robot Navigation

Hayato Terao[1], Jiaxu Wu[2], Qi An[1] and Atsushi Yamashita[1]

***Abstract*—As autonomous mobile robots begin to populate public spaces, it is becoming increasingly important for robots to accurately distinguish pedestrians and navigate safely to avoid collisions. Texting while walking is a common but hazardous behavior among pedestrians that poses significant challenges for robot navigation systems. While several studies have addressed the detection of text walkers, many have overlooked the impact of occlusions, a very common phenomenon where parts of pedestrians are obscured from sensor's view. This study proposes a machine learning method that distinguishes text walkers from other pedestrians in video data. The proposed method processes each video frame to extract body keypoints, encodes the keypoints into a latent space, and classifies pedestrian activities into three categories: normal walking, texting while walking, and other activities. A variational autoencoder is incorporated to enhance the system's robustness under various occlusion scenarios. Performance tests in real-world environments identified potential areas for improvement, particularly in distinguishing pedestrian activities with similar body postures. However, ablation studies demonstrated that the proposed system performs reliably across different occlusion scenarios.**

***Index Terms*—Autonomous Vehicle Navigation, Robot Safety, Computer Vision for Transportation, Object Detection, Segmentation and Categorization.**

I. Introduction

THE integration of autonomous mobile robots into our daily lives is gradually increasing. As robots and humans begin to coexist in spaces such as restaurants, offices, and public streets, the likelihood of interactions between them rises significantly. Ensuring safety in these interactions is paramount, and a critical component of this is the design of robot navigation systems that can effectively avoid collisions with pedestrians.

Amongst various pedestrian behaviors, texting while walking, referred to as "text walkers", presents a particularly heightened risk of collisions (Fig. 1). This is due to the diminished awareness of the surroundings, caused by the fixation on their phone screens [1]. Text walkers not only increase the risk of direct collisions, but also cause erratic movements of nearby pedestrians, who may make sudden and unpredictable changes in direction to avoid them [2]. This behavior poses a unique challenge for mobile robots, making text walkers one of the most difficult pedestrian behaviors to manage. Although numerous studies have been conducted on robot navigation systems, many still struggle to match human accuracy in detecting pedestrian movements. This shortfall is primarily due to insufficient consideration of specific pedestrian activities [3]. Consequently, the accurate detection of text walkers is a crucial first step toward developing safe and reliable navigation systems for autonomous mobile robots.

Fig. 1: Mobile robot meets a text walker.

One of the primary challenges in detecting text walkers is occlusion, where a portion of pedestrian's body is obscured from robot's view due to the presence of other objects, making it difficult to extract meaningful features from the acquired data. For instance, when the hands or arms of a text walker is occluded, even humans may struggle to accurately determine whether the individual is texting. In our previous work [4], we addressed this issue by leveraging a variational autoencoder (VAE) to infer the occluded body keypoints based on the visible ones. While this method demonstrated effectiveness in controlled experimental settings, its performance has yet to be evaluated in real-world environments. In this study, we extend our previous work by assessing the updated method on pedestrian data collected under real-world environments.

The subsequent sections of this paper are organized as follows: Section 2 reviews the related work in the field, providing a foundation for understanding the context of this study. Section 3 describes the proposed method. Section 4 explains the datasets utilized in this study, and Section 5 presents and discusses the experimental results. Finally, Section 6 concludes

Manuscript received: November 29, 2024; Revised April 21, 2025; Accepted June 6, 2025.

This paper was recommended for publication by Editor Ashis Banerjee upon evaluation of the Associate Editor and Reviewers' comments. This work was not supported by any organization.

1: Hayato Terao, Qi An, and Atsushi Yamashita are with the Department of Human and Engineered Environmental Studies, Graduate School of Frontier Sciences, The University of Tokyo, Japan. {terao, qi, yamashita}@robot.t.u-tokyo.ac.jp

2: Jiaxu Wu is with the Department of Precision Engineering, Graduate School of Engineering, The University of Tokyo, Japan. wujiaxu@robot.t.u-tokyo.ac.jp

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

the paper and outlines directions for future work.

II. Related Work

Previous studies on the detection of text walkers can be categorized into two primary approaches: in-situ sensing and remote sensing. In-situ sensing involves gathering data from sensors that are in close proximity to the subject, such as body-worn accelerometers or data acquired directly from pedestrians' smartphones [5]. In contrast, remote sensing methods utilize sensors like cameras [6], [7] and LiDAR [8] to monitor pedestrians from a distance. While in-situ sensing methods are inherently unaffected by occlusions, they have significant impracticality, as they require pedestrians to wear specific sensors or provide access to their smartphone data. Due to these constraints, remote sensing methods are considered a more practical and scalable solution for widespread use in real-world environments.

In a LiDAR-based approach, Wu et al. [8] utilized a LiDAR sensor to analyze the 3D point cloud data of pedestrians to distinguish text walkers from other pedestrians. Camerabased methods such as those by Kumamoto and Yamada [6], and Rangesh et al. [7], pre-processed an RGB image to extract the 2D coordinates of pedestrians' body keypoints, which were then analyzed to detect text walkers. Although the LiDAR-based method benefits from being unaffected by varying light conditions, it struggles with lower detection accuracy compared to camera-based methods, making the latter a more viable option.

Camera-based methods that rely solely on analyzing body postures can struggle to accurately classify pedestrian activities with similar postures [6]. For example, text walkers typically exhibit a distinctive posture: bent arm, hand held in front of the chest, and a head facing down to look at the phone. However, other pedestrian activities that share these characteristics can easily be misclassified as text walkers, and conversely, text walkers can be misclassified as engaging in other activities.

To address this challenge, Rangesh et al. [7] extended the analysis by incorporating pedestrian gaze information and detecting hand-held objects in addition to body postures. While this approach improved classification accuracy, the issue of occlusions remains unresolved in camera-based methods. Images of occluded pedestrians are often excluded from datasets or not explicitly addressed. Furthermore, gaze estimation and hand-held object detection become problematic when pedestrians are too distant from the camera, resulting in low image resolution. Similar accuracy challenges are expected in situations involving occlusion, where detecting hand-held objects and accurately estimating gaze becomes significantly more difficult. Our previous study [4] addressed the challenge of occlusion by utilizing sequential data rather than relying on singleframe analysis, and by leveraging a VAE to handle occluded data. This approach proved successful in classifying occluded pedestrians under controlled experimental conditions. However, its effectiveness in real-world environments remained unexplored.

III. Proposed Method

The primary challenge in handling occluded pedestrians is the loss of critical information. When certain body parts are obscured, it becomes challenging to extract and analyze relevant features, often leading to incorrect pedestrian activity classification. To overcome this challenge, a machine-learning based method incorporating the following three innovations are proposed: the use of sequential data, the application of a VAE, and the integration of full-body keypoints.

First, by using sequential data instead of single-frame input, the proposed method enhances robustness against temporary occlusions, where occlusions occur in only certain frames. In this approach, each frame of a video is preprocessed by a pose estimation module, and the detected body keypoints are used as input. This strategy of extracting and utilizing body keypoints, rather than directly analyzing video images, is particularly advantageous for real-time applications, such as robot navigation systems [9].

Second, a VAE [10] is employed, along with a customdesigned loss function, to effectively manage permanent occlusions; situations where occlusions persist across all frames. VAEs are well-regarded for their ability to extract meaningful information from corrupted or noisy data. By combining this capability with a custom loss function, depicted in Eq. (5), the proposed method can extract relevant features from the input data, even in the presence of occlusions.

Finally, the use of full-body keypoints allows the proposed method to capture more data relevant to identifying text walkers. Unlike regular pedestrians, text walkers exhibit distinct gait patterns, such as slower walking speeds and shorter stride lengths [11]. Previous camera-based methods [4]–[7] have overlooked this by focusing solely on upperbody keypoints. By incorporating the analysis of lowerbody keypoints, the proposed method is expected to reduce misclassification of activities that share similar upper-body postures.

A. Overview

The proposed method's overview is depicted in Fig. 2. The system consists of two main components: the Pre-trained VAE and the Occlusion Handling Module (OHM). The Pretrained VAE is responsible for learning the mapping of unoccluded body keypoints into a latent space. It serves as a supervisor during the training of the OHM, which processes sequences of potentially occluded body keypoints, encodes them into the latent space, and classifies the pedestrian activities.

B. Pre-trained VAE

VAEs are a type of deep learning model that has gained popularity for its robustness in feature extraction from incomplete or noisy data, and for regenerating a noise-free version of the input data. It consists of an encoder, a latent space, and a decoder. The encoder maps the input data to parameters of a probability distribution over the latent space. The decoder then reconstructs the encoded data back to the input data. Because a VAE learns a distribution over the latent space, they can capture the underlying structure of the data even

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

when some parts of the input are missing. For instance, even if some body keypoints are undetected and missing, the VAE can use the information from the other detected keypoints to generate a complete and coherent representation of the pedestrian's body keypoints. This capability makes VAEs particularly useful for the application of detecting pedestrians in real-world environments where occlusions are common.

The Pre-trained VAE is designed to encode complete body keypoints coordinates into a latent space. Its role is to provide a ground truth data when training the encoder of the OHM and is only used in the offline training phase. Since the quality of the Pre-trained VAE is crucial for the success of the proposed method, highly accurate body keypoints coordinates are required for training this component.

The training of the Pre-trained VAE is conducted using the Evidence Lower Bound (ELBO) loss [10], which consists of the reconstruction loss Lrecon and the Kullback-Leibler (KL) divergence L_{kl}. L_{recon} is defined as the L2-norm between the original input data x and the reconstructed data $\hat{x}$ as shown in Eq. (1):

$$L_{recon} = ||x - \hat{x}||_2^2. \tag{1}$$

KL divergence measures how closely the learned probabilistic distribution in the latent space matches the standard distribution $N(0,I)$. The calculation of KL loss is shown in Eq. (2):

$$L_{kl} = -\beta(1 + \log\sigma^2 - \mu^2 - \sigma^2), \tag{2}$$

where β is the KL coefficient [12], a hyper-parameter used to adjust the balance between L_{recon} and L_{kl} when computing the ELBO loss. Finally, the ELBO loss is computed as shown in Eq. (3):

$$L_{ELBO} = L_{recon} + L_{kl}. \tag{3}$$

C. Occlusion Handing Module

The OHM takes a sequence of potentially occluded body keypoints as input, collected at 5 fps for 1 second. The input is passed through a long short-term memory (LSTM) network [13] followed by a fully connected layer. The motivation behind this structure is that the LSTM layers can predict the current body keypoints of the pedestrian from the given sequence of data, while the fully connected layer encodes this data into a latent vector. While LSTM layers may be sufficient for handling temporal occlusions, where occlusions only occur in some of the input frames, they may not be adequate for handling persistent occlusions where the coordinates of certain body keypoints are unavailable throughout the entire input sequence.

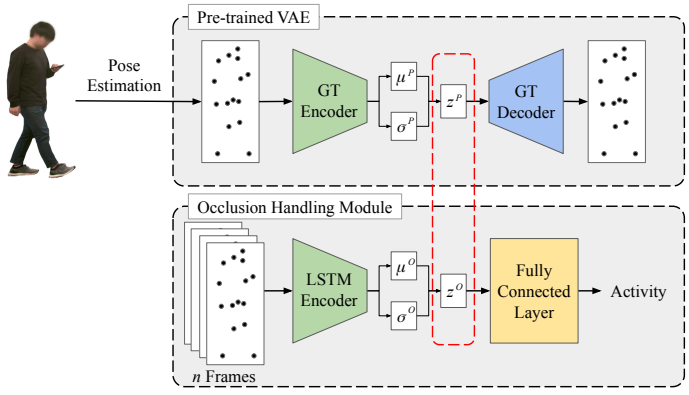

Fig. 2: Overview of the Proposed Method. The method is consisted of Pre-trained VAE and Occlusion Handling Module.

This is where the supervision by the Pre-trained VAE comes into play. The idea is to train the encoder of the OHM to denoise the input data and extract meaningful features from it. This is accomplished through the following processes:

- 1) Extract the latent representation of "clean data" using GT encoder of the Pre-trained VAE
- 2) Artificially add occlusions to the "clean data" to generate "occluded data"
- 3) Pass a sequence of "occluded data" into the LSTM encoder of the OHM
- 4) Compute the loss of the LSTM encoder by comparing the latent representations of the "clean data" and the "occluded data"

By training the LSTM encoder in this manner, it learns to extract robust features even from noisy or incomplete data. An encoding loss Lenc is introduced as the loss function for training the LSTM encoder as shown in Eq. (4):

$$L_{enc} = ||z^P - z^O||_2^2, \tag{4}$$

where z^P and z^O are the latent vectors generated by the encoders of the Pre-trained VAE and the OHM, respectively. The aim of the loss function is to let the LSTM encoder learn to map the input data into a similar region of the latent space, regardless of occlusion patterns. The latent vector z^O is then passed into a network of fully connected layers, where the pedestrian activity is classified. This classification task is trained using the categorical crossentropy loss L_{cat}, which quantifies the discrepancy between the predicted and actual activity labels assigned to the pedestrian. The LSTM encoder and the fully connected layers network are trained simultaneously in a unified stage via a custom-designed loss function. This approach ensures that the model's optimization is intricately aligned with accurately determining pedestrian activities while benefiting from the supervision of the Pre-trained VAE. The loss function L_{ohm}, tailored to train the OHM, is presented in Eq. (5), where wenc is an arbitrary coefficient taking the value between 0.0 and 1.0.

$$L_{ohm} = w_{enc} \cdot L_{enc} + (1 - w_{enc}) \cdot L_{cat}. \tag{5}$$

The precise adjustment of wenc is crucial for the model to benefit from the supervision of the Pre-trained VAE while ensuring its ability to accurately classify pedestrian activities. For instance, if w_{enc} is set to 0.0, the supervision by the Pretrained VAE will not be in effect during the training process, and the model focuses solely on classification accuracy. In contrast, if wenc is set to 1.0, the model will be trained to learn only the mapping of input data into a latent space based on the Pre-trained VAE, disregarding the categorization of pedestrian activities since L_{cat} will not influence L_{ohm}.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

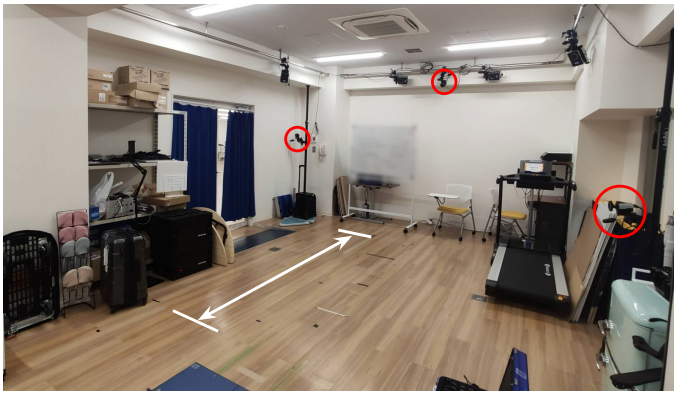

Fig. 3: Experimental setup of Theia Markerless. The participant's trajectory is shown as white arrow, and cameras used for motion capture is circled in red. Additional cameras are located outside of the image.

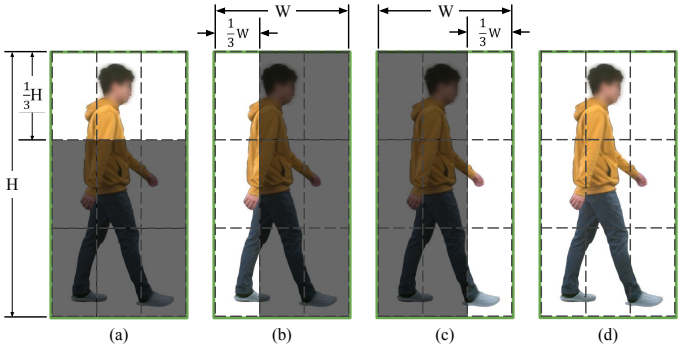

Fig. 4: Examples of occlusion patterns considered in this research. The shaded area represents the occluded regions. (a) Only top portion of bounding box is visible. (b) Only left portion of bounding box is visible. (c) Only right of bounding box portion is visible. (d) The bounding box is fully visible. (The experiment were conducted under ethical approval from the research ethics committee of the Graduate School of Engineering, University of Tokyo.)

IV. Dataset Generation

A. Overview

Two datasets were created to train and test the proposed method. The Body Keypoints Dataset was utilized for both training and testing the system, while the Video Dataset was primarily used to evaluate the system's performance in real-world environments.

B. Body Keypoint Dataset

The Body Keypoints Dataset consists of 2D coordinates of pedestrian's body keypoints. The data were acquired using Theia Markerless [14], a marker-less motion capture system, to ensure precise and accurate locations of the body keypoints.

The experimental setup for generating the Body Keypoints Dataset is illustrated in Fig. 3. Three individuals, each experienced in texting while walking, took part in the data collection. The participants were asked to traverse a room while being recorded with multiple cameras arranged in the room. For texting while walking scenario, they were instructed to perform familiar activities such as checking social media and messaging friends. The videos were subsequently processed by Theia Markerless to extract the 3D coordinates of body keypoints (head, neck, shoulders, elbows, hands, pelvis, hip joints, knees, and heels).

The 3D coordinates of the body keypoints were converted into 2D coordinates to simulate views from cameras positioned at various angles. This view projection was achieved by applying translation and rotation matrices to the 3D coordinates. Finally, the 2D coordinates were normalized to fit within a square bounding box of unit height and width, following the approach outlined in previous work by Kumamoto and Yamada [6]. The resulting dataset comprises "clean data" of accurate pedestrian body keypoints, captured from a total of eight different view angles. This data was used to train the Pre-trained VAE, which requires ground-truth coordinates of the body keypoints. In addition, the "clean data" were further processed to simulate the effects of occlusions on body keypoint detection. The four most common occlusion patterns [15], depicted in Fig. 4, were replicated by masking the data of the occluded region. Both the clean and occluded data were then used to train the OHM.

C. Video Dataset

The Video Dataset consists of RGB videos of pedestrians, recorded at 30 fps using a RealSense D455 RGB-D camera [16]. Two datasets were collected: one from a relatively controlled environment and one from environments close to the real world (Fig. 5). In controlled environment dataset, the videos were captured by the camera mounted on top of a four-wheeled mobile robot under various conditions, including scenarios with or without static obstacles, and with single or multiple pedestrians. These four different situations help replicate the effects of occlusions encountered in real-world environments, such as those caused by static objects or other pedestrians (Fig. 5 top). In the real-world environment dataset, the camera was mounted on a four-legged robot navigating inside crowds by remote control. Vibration and rotation of the robot body made the captured images blurrier. We recruited 11 participants and created diverse experimental scenarios in hallways, lobbies, and outdoor areas, with 4 to 6 individuals randomly selected per trial to simulate natural crowds (Fig. 5 bottom). Participants were instructed to perform a range of daily activities, including texting while walking, as well as several activities with postures visually similar to texting, such as holding a coffee cup, bottle, or umbrella, or carrying a handbag, eco-bag, or clothes (Fig. 6). These actions were intentionally chosen to reflect realistic sources of ambiguity that the detector may face in public spaces. In addition, different light conditions and occlusion patterns in various scenes also made the detection task more challenging compared to the controlled environment, as they cause more incomplete or erroneous body keypoint estimations.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

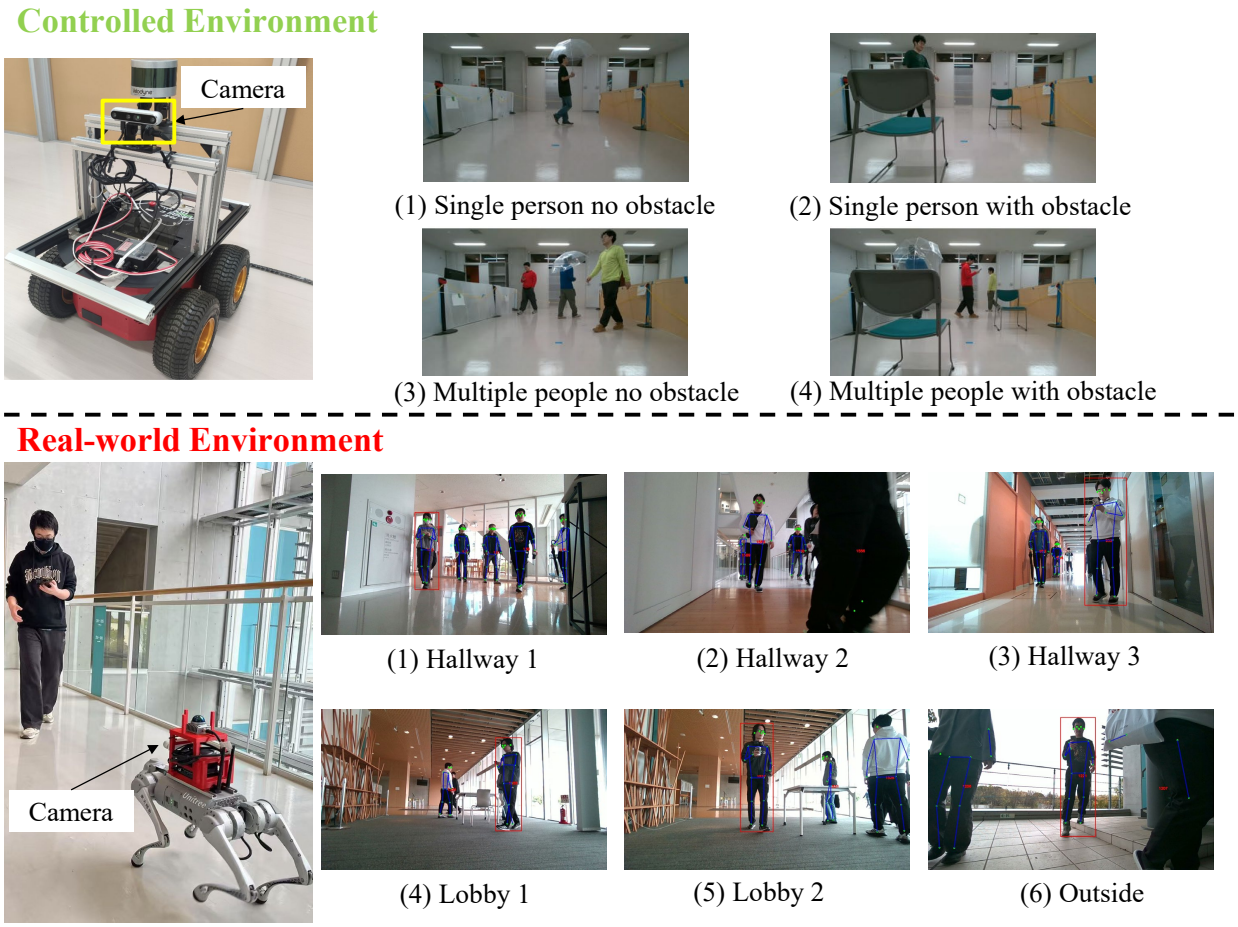

Fig. 5: Video datasets collected under different environments and conditions.

Fig. 6: Human activities in real-world environment dataset.

V. Experiment and Results

A. Hyper-Parameter Tuning and Model Structures

The hyper-parameters of the proposed model, including the number of neurons, latent dimension, and the encoding weight w_{enc}, were optimized using Bayesian optimization. Unlike traditional methods such as experience-based adjustments or brute-force search, Bayesian optimization offers a more efficient and sophisticated approach to hyper-parameter tuning [17]. It systematically explores the hyper-parameter space by building a probabilistic model, increasing the likelihood of finding optimal hyper-parameters within a shorter time frame. merged with the training subset to create a larger dataset for final model training. The test subset was reserved for the final evaluation of the model's performance. As a result of the hyper-parameter tuning with Optuna, the latent dimension was determined to be 4, and the encoding weight w_{enc} was set to 0.2.

The Body Keypoints Dataset was split into training (80%), validation (10%), and test (10%) subsets. During the Bayesian optimization process, the validation subset was used to fine-tune the hyper-parameters. Once the optimal hyper-parameters were identified, the validation subset was

B. Ablation Study on Body Keypoints Dataset

An ablation study was conducted to assess the contribution of each key component of the proposed system: utilizing Pretrained VAE and leveraging sequential data, to the overall result. In addition to the proposed model, three other models with some modifications to the proposed method were trained and tested on the same dataset. A state-of-the-art method [6]

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

TABLE I: Result of ablation study for each occlusion pattern. The best F1 score for each occlusion pattern is highlighted in bold.

Model	F1 scores				
	Fully Visible	Top Vis.	Left Vis.	Right Vis.	Mean
(a)	0.970	0.946	0.891	0.902	**0.927**
(b)	0.953	0.928	**0.905**	**0.909**	0.924
(c)	0.945	0.926	0.784	0.810	0.866
(d)	0.955	0.912	0.768	0.814	0.862
(e)	**0.973**	**0.951**	0.652	0.759	0.834

was also trained and tested for comparison. The F1 scores of the models were computed as follows:

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{6}$$

Table I summarizes the F1 score for each of the 4 occlusion patterns, with each column representing the occlusion patterns in Fig. 4. The models are represented as follows in the table:

- (a) The proposed method
- (b) OHM without Pre-trained VAE
- (c) OHM with Pre-trained VAE using single-frame input
- (d) OHM without Pre-trained VAE using single-frame input
- (e) State-of-the-art method

The influence of the Pre-trained VAE is evident when comparing models (a) and (b). The use of the Pre-trained VAE positively impacted the classification of pedestrian activities for fully and top-visible patterns, improving the F1 score by approximately 0.15. However, it did not show an effect on the remaining patterns. Since the Pre-trained VAE was trained on fully visible patterns only, this result suggests that model (a) may have overfitted to fully visible keypoints.

The influence of sequential input can be observed by comparing models (a) and (c). The use of sequential input resulted in an increase in F1 score across all occlusion patterns, ranging from approximately 0.02 to 0.1. This underscores the importance of leveraging temporal information in text walker detection.

Lastly, although model (e) performed the best for fully and top-visible patterns, it performed the worst for the left and right visible patterns, indicating a lack of robustness against occlusions from the sides. In contrast, model (a) achieved consistently high F1 scores across all occlusion patterns. The proposed method outperformed the state-of-theart method in terms of the mean F1 score, and demonstrated high robustness against different occlusion patterns.

C. Performance Test on Video Dataset

To test the performance of the proposed system in real-world environments, the proposed system was tested using the Video Dataset. You Only Look Once (YOLO) [18] was used to detect the pedestrian pose from the videos. Different from 3D motion capture, YOLO may result in incomplete or erroneous pose estimations due to motion blur, bad light conditions, and occlusions. Keypoints with detection confidence lower than 0.5 were treated as a miss detection. The pose detection from YOLO is given in COCO format [19], which differs from the format used in the Body Keypoints Dataset. As such, some keypoints of YOLO detection results were converted to match the format of the Body Keypoints Dataset following the conversion methods shown in Table II. After converting the results of YOLO pose detection to an appropriate format and normalizing it, the data were then passed to the proposed system to evaluate its performance. We integrated the pose estimation, format translation, and text walker detection on a notebook PC with an Intel Corei9-9980HK CPU and a Nvidia RTX 2080 Mobile GPU. The process time for pose estimation was approximately $0.01[s]$ on average, and the format translation and text walker detection took $0.004[s]$ on average.

The F1 score for the controlled environment dataset is summarized in Table III. The results indicate areas for improvement in the classification task.

The F1 score for the real-world environment dataset is summarized in Table IV. Besides the overall F1 score, we also evaluated F1 score under different patterns of incomplete body keypoint estimations. For each sample, the most dominant pattern in the pose sequence was used. The results indicate that when only one arm (involving hand, elbow, and shoulder keypoints) was detected in most of the frames, the performance of text walker detection degraded. The F1 score on left arm-detected samples was significantly higher than that of right arm-detected samples. This may be due to the fact that participants tended to operate their cell phones with their left hand. The overall F1 score obtained in real-world environment was lower than the score obtained in the controlled environment. Since the distribution of the patterns in incomplete pose estimation was found to be similar between the controlled environment and the real-world environment, the reason is considered to be erroneous pose estimation due to motion blur or poor lighting conditions, and false positive detections due to human activities with postures visually similar to those of text walkers.

The false positive rate against other human activities in the real-world dataset is summarized in Table V. Activities that require humans to keep their arms in front of their chest, such as holding a cup of coffee or an umbrella, were most likely to induce false positive detections.

The qualitative evaluation results are shown in Fig. 7. The results show that the proposed system could detect text walkers using one or both hands. The proposed system could also detect text walkers with different view angles from the robot and have the ability to detect text walkers with partial observation (Fig. 7 (c)). On the other hand, in Fig. 7 (d)-(f), normal pedestrians with poses similar to text walkers induced false positive detections, which indicates that sometimes human body poses are not enough for classification between text walkers and normal pedestrians. Moreover, Fig. 7 (g) and (h) show failed classifications due to erroneous pose estimations for the positions of human arms. This suggests the model's strong reliance on the location and angles of the hands and arms when classifying pedestrian activities. Since erroneous pose estimation is inevitable for some input RGB images, such as those with extremely bright areas (g) and (h) and those with heavy motion blur, estimating human poses

using both RGB-D images [19] may be a promising way to address the current bottleneck. Fig. 7 (i) shows false positive detection on the human with a black t-shirt due to partial observability. Constrained by robot heights and the field of view (FOV) of cameras, this problem is also inevitable in real applications. One way to solve this problem is to use fish-eye or spherical cameras, which provide larger FOVs. Since a real-world dataset of text walker has already been collected using the spherical camera in our previous study [8], we will extend the proposed system to the spherical camera in future work.

TABLE II: Conversion of Body Keypoints from COCO to Body Keypoints Dataset Format.

Custom Format	**COCO Keypoint**	**Conversion Method**
Head	Nose	Directly use
Neck	Left Shoulder, Right Shoulder	Average of Left Shoulder and Right Shoulder
Shoulders	Left Shoulder, Right Shoulder	Directly use
Elbows	Left Elbow, Right Elbow	Directly use
Hands	Left Wrist, Right Wrist	Directly use
Pelvis	Left Hip, Right Hip	Average of Left Hip and Right Hip
Hips	Left Hip, Right Hip	Directly use
Knees	Left Knee, Right Knee	Directly use
Feet	Left Ankle, Right Ankle	Directly use

TABLE III: Performance of proposed system on controlled environment dataset.

Activity	F1 scores			
	Single Ped. No Obs.	Single Ped. with Obs.	Multi Ped. No Obs.	Multi Ped. with Obss
Normal	0.827	0.788	0.716	0.727
Text	0.518	0.307	0.448	0.409

TABLE IV: Performance of proposed system on real-world environment dataset.

Detected keypoints	# samples	Percentage	F1 score
One arm (R)	20636	9.89%	0.114
One arm (L)	10679	5.12%	0.195
Two arms	20636	9.89%	0.370
Head-Neck-One arm(R)	8952	4.29%	0.058
Head-Neck-One arm(L)	8689	4.16%	0.206
Head-Neck-Two arms	126781	60.74%	0.378
Overall	208725	100%	0.350

VI. Conclusion

In this paper, a novel machine learning-based method for a robust detection of text walkers under occluded environments was proposed and evaluated. The method was unique in three ways: it used a Pre-trained VAE for feature extraction from occluded data and leveraged sequential full-body keypoints over single-frame upper-body keypoints as the input to capture more information.

TABLE V: False Positive Rate (FPR) Against Other Activities.

Activities	# samples	Percentage	FPR
Handbag	5831	13.97%	0.355
Umbrella	5374	12.87%	0.560
Bottle	4117	9.86%	0.520
Clothes	4221	10.11%	0.383
Coffee	3741	8.96%	0.607
Eco-bag	3737	8.95%	0.462
Cellphone (purely holding)	1192	2.86%	0.325
Cellphone (calling)	812	1.95%	0.522
Nothing in hand	4820	11.55%	0.370

An ablation study on the Body Keypoints Dataset demonstrated the effectiveness of these key components. The proposed method also outperformed the state-of-the-art, showing robustness in various occlusion scenarios. However, performance tests on the Video Dataset highlighted areas for improvement, such as distinguishing between activities with similar body postures.

Future work includes expanding the dataset and adding additional features to the system. The Video Dataset contained data that were not present in the Body Keypoints Dataset, such as normal pedestrians walking with crossed arms and many activities that have postures similar to text walkers. False positive detection could be reduced by adding those challenging negative samples to the training and performing a Contrastive Learning with Hard Negative Mining [20]. Moreover, adding depth information to the detection system could help obtain more accurate human pose estimation and potentially improve performance.

References

[1] F. Obayashi and K. Kozuka, "Sight Property at the Time 'Texting While Walking' by the Gaze Measurement, and Its Influence to Walking," IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences (IEICE Trans. Fundam. Electron. Commun. Comput. Sci.), vol. J100-A, no. 9, pp. 338–345, 2017.

[2] H. Murakami, C. Feliciani, Y. Nishiyama and K. Nishinari, "Mutual Anticipation Can Contribute to Self-Organization in Human Crowds," Science Advances (Sci. Adv.), vol. 7, no. 12, p. eabe7758, 2021.

[3] D. Ridel et al., "A Literature Review on the Prediction of Pedestrian Behavior in Urban Scenarios," Proceedings of the IEEE International Conference on Intelligent Transportation Systems (IEEE ITSC), pp. 3105–3112, 2018.

[4] H. Terao et al., "Detection of Texting While Walking in Occluded Scenarios Using Variational Autoencoder," Proceedings of the IEEE/SICE International Symposium on System Integration (SII), pp. 768–773, 2024.

[5] A. Shikishima, K. Nakamura and T. Wada, "Detection of Texting While Walking Using Smartphone Sensors," Proceedings of the International Conference on Intelligent Transportation Systems Telecommunications (ITST), pp. 1–6, 2018.

[6] K. Kumamoto and K. Yamada, "Detecting Pedestrian Interaction with Smartphones Based on Body Keypoints," Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (IEEE SMC), pp. 3261–3266, 2018.

[7] A. Rangesh and M. M. Trivedi, "Recognizing Phone-Based Activities of Pedestrians," IEEE Transactions on Intelligent Vehicles (IEEE Trans. Intell. Veh.), vol. 3, no. 2, pp. 218–227, 2018.

[8] J. Wu et al., "Smartphone Zombie Detection from LiDAR Point Cloud for Mobile Robot Safety," IEEE Robotics and Automation Letters (IEEE RA-L), vol. 5, no. 2, pp. 2256–2263, 2020.

[9] N. Ma et al., "A Survey of Human Action Recognition and Posture Prediction," Tsinghua Science and Technology (Tsinghua Sci. Technol.), vol. 27, no. 6, pp. 973–1001, 2022.

[10] D. P. Kingma and M. Welling, "Auto-encoding Variational Bayes," International Conference on Learning Representations (ICLR), 2014.

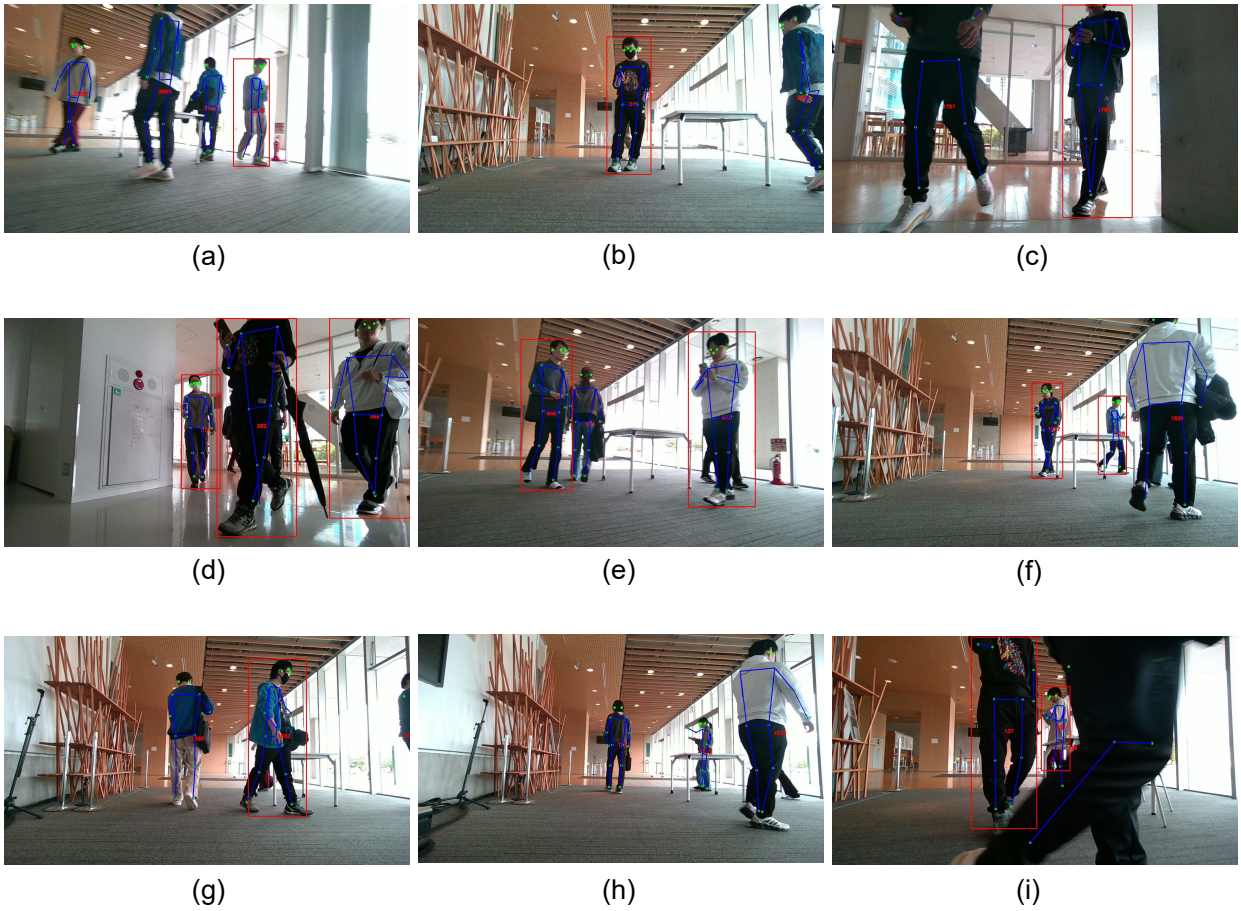

Fig. 7: Qualitative evaluation results in the real-world dataset. Detected body keypoints are shown as green dots. Skeletons are shown as blue lines. Text walker detection results are shown as red bounding boxes. (a)-(c): true positive detections. (d)-(f): false positive detections due to activities with similar posture. (g) and (h): false positive detection and false negative detection due to erroneous pose estimations. (i): false positive detection on the person with a black t-shirt due to the incomplete pose estimation result.

[11] E. Lamberg and L. M. Muratori, "Cell Phones Change the Way We Walk," Gait and Posture (Gait Posture), vol. 35, no. 4, pp. 688–690, 2012.
[12] I. Higgins et al., "Beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework," Proceedings of the International Conference on Learning Representations (ICLR), 2017.
[13] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation (Neural Comput.), vol. 9, no. 8, pp. 1735–1780, 1997.
[14] Theia Markerless, [Online]. Available: https://www.theiamarkerless.ca/
[15] P. Dollar et al., "Pedestrian Detection: An Evaluation of the State of the Art," IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI), vol. 34, no. 4, pp. 743–761, 2012.
[16] Intel RealSense D455, [Online]. Available: https://www.intel.com/content/www/us/en/products/sku/205847/intel-realsense-depth-camera-d455/specifications.html
[17] J. Snoek et al., "Practical Bayesian Optimization of Machine Learning Algorithms," Advances in Neural Information Processing Systems (NeurIPS), vol. 25, 2012.
[18] G. Jocher et al., "YOLOv8," Ultralytics, 2023. [Online]. Available: https://github.com/ultralytics/ultralytics
[19] C. Zimmermann, T. Welschehold, C. Dornhege, W. Burgard and T. Brox, "3D Human Pose Estimation in RGBD Images for Robotic Task Learning," Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), pp. 1986-1992, 2018.
[20] J. Robinson, C. Y. Chuang, S. Sra, and S. Jegelka, "Contrastive learning with hard negative samples," Proceedings of the International Conference on Learning Representations (ICLR), 2021.