International Journal of Innovative
Computing, Information and Control
Volume **20**, Number **6**, December **2024**

ICIC International ©2024 ISSN 1349-4198
pp. 1573–1585

IMPROVING A FASTER R-CNN MODEL FOR VEHICLE DETECTION AND HUMAN ACTION RECOGNITION AT NIGHT VIA INFRARED THERMAL IMAGING USING TRANSFER LEARNING

Yaru Liu, Kai Matsui, Yoichi Kageyama*, Hikaru Shirai and Chikako Ishizawa

Graduate School of Engineering Science
Akita University
1-1 Tegata Gakuen-machi, Akita-shi, Akita 010-8502, Japan
d8522008@s.akita-u.ac.jp; kmatsui@gmail.com; {shirai; ishizawa}@ie.akita-u.ac.jp
*Corresponding author: kageyama@ie.akita-u.ac.jp

Received February 2024; revised June 2024

Abstract. *At present, the older population is the fastest-growing segment of the driving population, which has led to higher rates of traffic accidents. Data on the number of casualties in accidents involving pedestrians and motor vehicles during the day and at night indicate that the proportion of fatalities is significantly higher at night. Consequently, focusing on the traffic safety of the elderly, reducing the occurrence of nighttime traffic accidents, and promoting sustainability are crucial for Japan, which faces the challenge of becoming a "super-aging" society. Thus, we propose a system to support the safety and security of pedestrians and drivers using infrared thermal imaging data at night. In previous studies, we developed methods to detect pedestrian actions using a novel convolutional neural network (CNN)-based model, specifically VGG16. In this study, we propose improvements to an existing detection method using an improved Faster R-CNN model to detect vehicles and recognize human actions in real time at night. We acquired new video data demonstrating multiple human actions related to distant target objects captured by the infrared thermal camera. These data can be used to investigate vehicle detection and action recognition in scenes involving multiple humans using transfer learning. We experimentally evaluated the performance of our method in terms of the detection accuracy, and the results indicate that our proposed method achieved a mean average precision of 0.97 in detecting actions in scenes with multiple people positioned far from the camera. It exhibited superior accuracy compared to conventional methods.*

Keywords: Vehicle detection, Human action recognition, Infrared image, Nighttime, Faster R-CNN

1. **Introduction.** Population aging has become a notable and common demographic phenomenon in most countries. In Japan, the number of people aged 65 years and older has increased rapidly in recent years, with the elderly population accounting for 29.1% of the total population by 2023, which is the highest proportion in history [1]. Japan has one of the most rapidly aging populations in the world. According to statistics from the United Nations and other international organizations, Japan consistently ranks highest in terms of the aging population. In addition, life expectancy in Japan is among the highest globally, with women living an average of over 87 years and men over 81 years. These factors exacerbate the aging population issue, making Japan a significant reference point for global research and strategies to address aging [2].

The rapid development of road traffic systems in modern society, coupled with an aging population, has made traffic safety a particularly prominent issue in Japan as a

DOI: 10.24507/ijicic.20.06.1573

country with a substantial proportion of elderly people. As people age, their physiological functions decline, leading to deteriorating vision and hearing, and slower reaction times. These changes affect their ability to make timely judgments regarding road conditions and traffic situations. Elderly drivers often have diminished driving skills and physical capabilities compared to younger drivers, making them more prone to operational errors in complex traffic conditions. Moreover, older adults may struggle to adapt to new traffic regulations and technologies, posing additional traffic safety risks. Therefore, as the aging population continues to grow, increasing focus is placed on addressing the traffic safety needs of older adults. According to the Japanese National Police Agency, drivers aged 75 years and older caused 460 fatal road accidents in Japan in 2018. Moreover, the proportion of such accidents increased from 8.7% to 14.8% over the past 10 years [3]. In particular, in the intricate conditions of nighttime settings, the increased likelihood of fatal accidents is two to four times higher than that in daytime scenarios [4]. In addition, the diminished levels of ambient light in nighttime road environments have emerged as a significant contributing factor in accidents involving pedestrians, cyclists, and other low-contrast obstacles. Reduced visibility is more likely to be a leading cause than driver fatigue or alcohol consumption [5]. The high population density leads to frequent road congestion, particularly in major cities such as Tokyo. A significant amount of vehicle and pedestrian traffic exists even at night. Elderly drivers are more likely to be involved in traffic accidents compared to other age groups, especially in complex nighttime traffic environments. In addition, the walking abilities of elderly individuals decline, making them more prone to falling or failing to avoid oncoming vehicles when crossing the road [1]. Therefore, a need exists for high-precision detection of human behavior and traffic environments, such as the real-time detection of vehicles and pedestrians and recognition of pedestrian actions. By recognizing the actions of vehicles and pedestrians, the safety of road users can be effectively addressed, reducing nighttime traffic accidents and preventing incidents. This approach is crucial for the sustainable development of societies that face the challenges of a "super-aging" population, such as Japan, and is vital for the development of intelligent transportation systems globally [6].

Methods for reducing nighttime vehicle accidents and fatalities have been investigated in previous studies. For instance, researchers introduced a new approach using the Hough transform to extract speed limit signs and identify them through template matching. The experimental results validated the high sensitivity of this method in detecting changes in the status of nighttime speed limit signs [7,8], making it applicable to accidents resulting from speed violations at nighttime. However, to reduce traffic accidents significantly, detecting not only the behaviors of pedestrians on the road but also those of moving vehicles is imperative. Therefore, the researchers proposed a VGG16-based detection method for human action analysis in various environments under varying conditions (illumination and temperature) at nighttime [9]. Although high accuracy has been observed in pedestrian detection in previous studies, some limitations remain. For example, considering image data that are obtained under various weather conditions, such as rain or snow, is essential for practical use. In addition, achieving distant target recognition and real-time detection is crucial for enhancing pedestrian safety. A comprehensive investigation of nighttime human action has not yet been conducted.

In recent decades, deep-learning models have achieved significant success in reducing vehicle traffic accidents and deaths due to nighttime driving environments. In particular, deep-learning models employing convolutional neural networks (CNNs) have proven to be powerful tools that do not require manual feature extraction [10]. For example, Zhang et al. developed a dual-anchor region-based CNN (R-CNN) to capture paired body and head parts and detect humans within a crowd [11]. Farid et al. used the concept of transfer

learning by fine-tuning the weights of the pretrained YOLOv5 architecture, which has outperformed several traditional vehicle detection methods [12]. However, most of these studies utilized visible-light images to train the models, neglecting the complexities of road environments such as lighting changes, insufficient imaging light, shadows, background noise, and object occlusion. Consequently, these approaches have low detection performance for images containing ambiguous or noisy features, especially in nighttime driving environments [13,14]. Farooq et al. proposed the simple, fast, and efficient SIFR-CNN framework for detecting nighttime pedestrian actions. However, this approach does not effectively reduce noise in nighttime images and lacks specific evaluations for pedestrian detection at varying distances [15]. Therefore, we considered these issues in our study by using a thermal infrared camera to capture nighttime thermal infrared images to reduce noise. In addition, we systematically improved methods for long-range multitarget detection and real-time recognition.

The principles of vehicle detection and human action classification are similar to those for object detection. Iftikhar et al. surveyed deep-learning approaches for pedestrian detection in autonomous vehicles. Furthermore, it was verified that Faster R-CNN achieved better results in the evaluation metrics of low-attribute images, with an accuracy as high as 0.919 [16]. Arora et al. proposed a method for detecting moving vehicles using a fast region-based convolutional neural network with an overall average accuracy of 94.35% at nighttime [17]. Therefore, in this study, we used an improved detection method with a Faster R-CNN model based on the ResNet-50 network architectural approach to detect vehicles and recognize human actions in real time at nighttime. Furthermore, we acquired new video data showing multiple human actions for target objects far (15 m, 20 m, 25 m, and 35 m) from the infrared thermal camera to investigate vehicle detection and action recognition in scenes with multiple humans using transfer learning.

The methods and results of this study are applicable to vehicle detection and pedestrian action recognition in urban road traffic environments. The tested thermal infrared imaging technology and enhanced models exhibit robust detection capabilities, which are effective across various temperature and distance conditions. These advancements offer the potential to enhance nighttime traffic safety significantly, reducing the occurrence of nighttime traffic accidents.

In summary, the main contributions of this paper are summarized as follows.

1) An outdoor environment nighttime database was created using an infrared thermal imaging camera. The established datasets consist of 15,184 images capturing vehicle and human action patterns (standing, squatting, bending, and walking) at distances ranging from 15 m to 35 m.
2) We propose an improved Faster R-CNN model based on the ResNet-50 architecture using transfer learning. We also suggest adjusting the confusion region anchor box sizes for each object, specifically 64×64, 128×128, 256×256, and 512×512 pixels, to address significant distances (15 m, 20 m, 25 m and 35 m) from the camera.
3) In a complex nighttime environment, the proposed approach exhibited effectiveness in vehicle detection and human action recognition, achieving a mean average precision (mAP) of 0.97. Notably, the approach performed well at significant distances in the range of 25 m to 35 m from the camera, surpassing CNN (VGG16), Multi-task Faster R-CNN [18] and YOLOv5 [19-21] with improvements of 53.50%, 33.72%, and 15.33%, respectively.

2. Experiment and Data Used.

2.1. Camera.

In contrast to images that are captured during the day, lighting conditions in outdoor environments at night can vary significantly owing to factors such as the presence of streetlights and vehicle headlights. In this study, an infrared thermal imaging camera (manufactured by D-Eyes Co., Ltd., "ultra-high sensitivity + far-infrared" 2-in-1 camera WCAM001-AU, 640×480 pixels) [22] was used to acquire video data in an outdoor environment at night. This camera is equipped with two sensors: an ultra-high-sensitivity camera and a far-infrared camera, allowing the simultaneous independent recording of images. The external view of the equipment is shown in Figure 1(a).

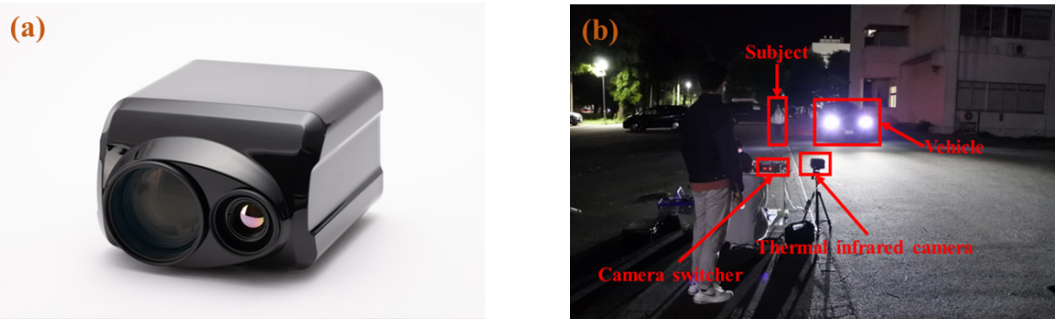


FIGURE 1. Equipment and data-acquisition environment: (a) External view of the infrared thermal imaging camera (standard type); (b) data-acquisition environment

Figure 1(b) shows the experimental environment. Owing to the difficulty of recruiting elderly volunteers in a school setting, most participants in our study were students in their 20s. This practical decision was made to facilitate the collection of necessary data to validate the effectiveness of our methods. Thus, although the background of our research highlights traffic safety issues arising from an aging population, the initial experiments were conducted with younger participants. The intention was to ensure the efficacy of our approach with this demographic before applying it to older individuals in future studies. Although the participants did not include elderly individuals, the insights gained from these experiments are expected to provide a foundation for subsequent research targeting the elderly population, who are more directly impacted by the traffic safety issues discussed.

The data used in this investigation were acquired in accordance with ethical regulations concerning human studies at Akita University, Japan.

2.2. Data acquisition.

2.2.1. *Data used in the pretrained model.* In the proposed method, transfer learning was adopted using a pretrained model. To obtain the data used in the pretrained model, we enrolled four participants of East Asian descent, including two males and two females in their 20s (males: 22 years old, 27 years old; females: 21 years old, 22 years old). To consider pedestrian actions, data were acquired at night with the participants standing, squatting, bending, and walking individually. Standing, squatting, and bending actions were captured from camera positions in front, behind, to the left, and to the right of the participants from a range of 10.0 m. Walking actions were recorded with the participants 5 m to 15.0 m from the camera from front and back views. Figure 2 shows an example of the pretrained model data.

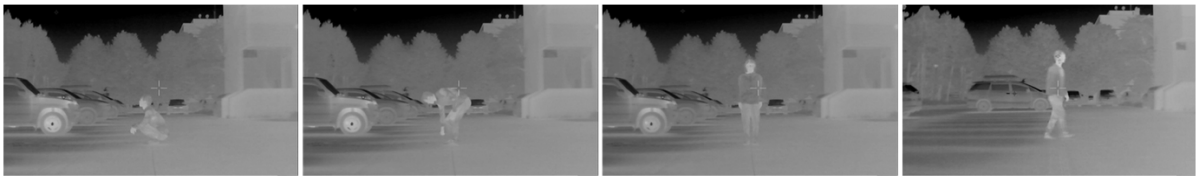


FIGURE 2. Sample of pretrained model data

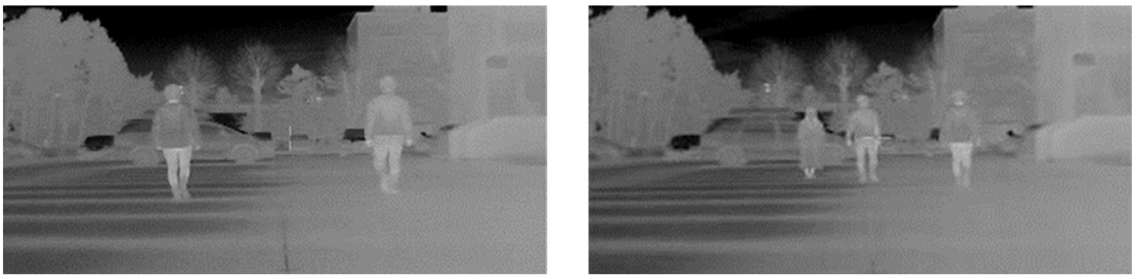


FIGURE 3. Sample of additional model data

2.2.2. *Additional data in improved model.* Figure 3 shows an example of the additional data. To analyze the actions performed by multiple humans at night, we acquired new video data on human actions in scenes with more than one person using an infrared thermal camera as additional data. The vehicle motion and two to three subjects were captured (two males and two females in their 20s; males: 22 years old, 27 years old; females: 21 years old, 22 years old). The process of additional data collection was as follows.

1) At 15 m from the camera, the action patterns (squatting and bending) of the two subjects were captured.
2) At 15 m from the camera, the action patterns (squatting and bending) of the three subjects were captured.
3) At 25 m from the camera, the action patterns (squatting and bending) of the two subjects were captured.
4) At 25 m from the camera, the action patterns (squatting and bending) of the three subjects were captured.
5) After completion of the actions in 1) and 2), the action patterns of the subjects walking (from 15 m to 20 m from the camera) and standing (including vehicle movement) were captured.
6) After completion of the actions in 3) and 4), the action patterns of the subjects walking (from 20 m to 25 m from the camera) and standing (including vehicle movement) were captured.

2.2.3. *Further distance image data.* We used additional image data captured at greater distances to validate the effectiveness of our method for further object detection. The data acquisition environment is shown in Figure 4. The data collection involved four subjects performing four different patterns at a range of 25 m to 35 m from the camera. The subjects included Subject I (male, squatting), Subject II (male, bending), Subject III (male, squatting), and Subject IV (female, bending). All subjects were of East Asian descent and in their 20s (males: 22 years old, 23 years old, 27 years old; female: 22 years old). The following motion patterns were captured:

1) The action patterns of two subjects (Subject I: squatting, Subject II: bending);

FIGURE 4. Further distance image data

2) The action patterns of three subjects (Subject I: squatting, Subject II: bending, and Subject III: squatting);
3) The action patterns of four subjects (Subject I: squatting, Subject II: bending, Subject III: squatting, and Subject IV: bending);
4) Following the completion of the motions described in patterns 1), 2), and 3), the subjects were instructed to walk from 35 m to 25 m from the camera and stand upright.

3. Analysis Method.

3.1. **Overview of the analysis method.** A flowchart of the proposed method is presented in Figure 5. The thermal infrared camera used in this study displays a temperature bar in the captured images. Therefore, it was necessary to reduce the noise from the images in the deep-learning model. First, the temperature bar regions in the original image were removed. Subsequently, the LabelImg tool [23] was used to label the human actions and vehicles to construct an annotated dataset. In addition, a Faster R-CNN model was implemented using the ResNet-50 [24] backbone as the underlying architecture. This model was pretrained on a large dataset (Section 2.2.1) and used for transfer learning, which involved fixing the weights of the pretrained model to utilize them for feature extraction. Finally, the sizes of the anchor boxes produced by the Faster R-CNN model were adjusted and the transfer learning model was trained using additional data. We then evaluated the accuracy of the proposed method for detecting vehicles and recognizing human actions (standing and walking) using the obtained model.

3.2. **Dataset processing in pretrained model.** The dataset of infrared images used for the pretrained model included 6,079 images and was manually labeled using LabelImg. An annotated dataset was created from these labeled images and divided into sets of 4,257 and 1,822 images for training and validation, respectively. The additional data comprised 1,651 images, which were similarly divided into sets of 1,320 and 331 images for training and validation, respectively.

3.3. **Improving Faster R-CNN model.** To generate feature maps from the input infrared images, we used the ResNet-50 network as the backbone of the Faster R-CNN model to extract image features. The proposed method accommodates the detection of

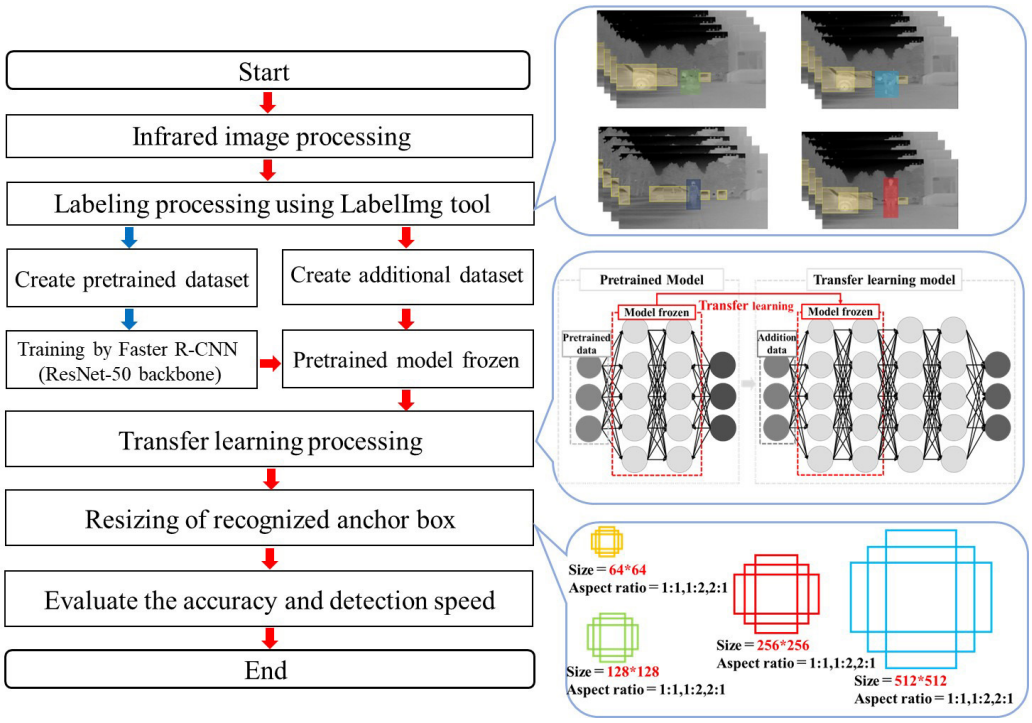


FIGURE 5. Flowchart of the analysis method

new objects at varying distances (15 m, 20 m, and 25 m) from the camera. We propose adjusting the recognition anchor box sizes for each object with confusion regions of 64 × 64, 128 × 128, 256 × 256, and 512 × 512 pixels in the infrared images. The training model was frozen to reduce the learning time, and the number of training epochs was set to 50 for both the frozen and unfrozen training.

3.4. **Transfer learning in the improved Faster R-CNN model.** Transfer learning can improve the training speed while maintaining the vector values and weights of a pretrained model. We applied this approach to transfering the learning of a model trained using our previous method. This process is illustrated in Figure 6. The pretrained model was trained using the large dataset described in Section 2.2.1. We used the vector values and weights from the pretrained model as a feature extractor to retrain our additional data, as described in Section 2.2.2 using the improved Faster R-CNN model architecture. During this process, the weight of the pretrained model was frozen and training was performed for 50 epochs.

4. Experiment.

4.1. **Implementation.** The proposed model was trained for 100 epochs using an Intel Core i7-9700k 3.6 GHz processor with an NVIDIA GeForce RTX 2080 Ti 27GB GPU, and the Python programming language and PyTorch learning framework were used. The average success rates for the best number of training iterations were calculated using the improved Faster R-CNN model. For transfer learning, we used a large dataset (6,079 thermal infrared images) as training data to calculate the weight vectors in the pretrained Faster R-CNN model. Subsequently, we used the weight vectors in the proposed improved Faster R-CNN model and trained the improved Faster R-CNN model using additional data (Section 2.2.2). To evaluate the trained Faster R-CNN model, a testing dataset of 1,904 images was constructed using additional data on five patterns: actions with multiple humans standing, squatting, bending, and walking, as well as vehicle motion. The results

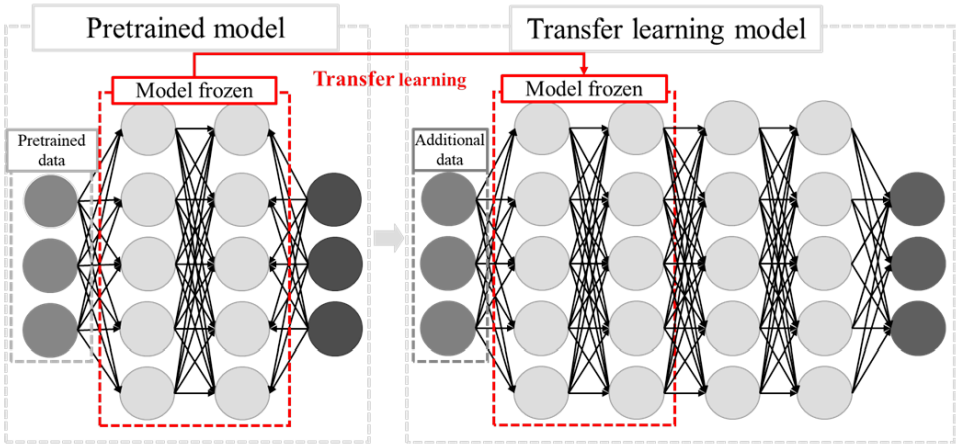


FIGURE 6. Transfer learning in Faster R-CNN model

of the improved Faster R-CNN model were assessed in terms of the F1, recall, precision, AP, and mAP and were compared with the results of the normal Faster R-CNN model [25]. In addition, to validate the effectiveness of our method for object detection further, we used the additional further distance image data (Section 2.2.3) and employed several object detection and recognition model methods including CNN (VGG16), Multi-task Faster R-CNN, and YOLOv5 as baselines for comparison with our approach.

4.2. **Comparison results with normal Faster R-CNN.** To evaluate the proposed approach, we compared it with a normal Faster R-CNN model using additional data. Each model was evaluated using a test dataset. Tables 1 and 2 present the results of the evaluation.

Based on the patterns presented in Table 1, the normal Faster R-CNN model recognized the standing (AP: 0.18), walking (AP: 0.07), and vehicle (AP: 0.51) patterns relatively poorly and was ineffective in identifying actions in scenes with multiple people. Conversely, as indicated in Table 2, the AP was improved by approximately 0.77, 0.90, and 0.47 in

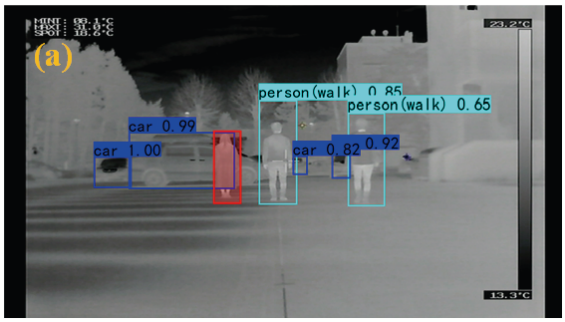
TABLE 1. Evaluation results for normal Faster R-CNN

Pattern	F1	Recall	Precision	AP
Standing	0.37	0.38	0.36	**0.18**
Squatting	0.88	0.98	0.79	**0.88**
Bending	0.86	0.95	0.77	**0.92**
Walking	0.20	0.15	0.27	**0.07**
Vehicle	0.54	0.69	0.43	**0.51**
mAP				**0.51**

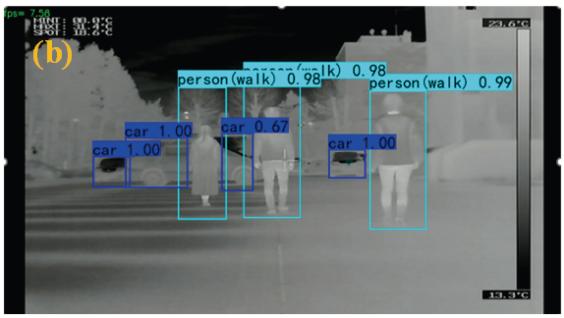
TABLE 2. Evaluation results for proposed approach

Pattern	F1	Recall	Precision	AP
Standing	0.95	0.96	0.93	**0.95**
Squatting	0.98	0.99	0.96	**0.99**
Bending	0.98	0.98	0.97	**0.98**
Walking	0.96	0.97	0.95	**0.97**
Vehicle	0.98	0.98	0.97	**0.98**
mAP				**0.97**

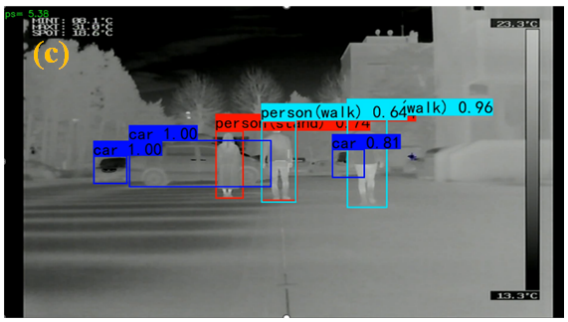
recognizing standing and walking actions and detecting vehicles, respectively, using the proposed approach. Furthermore, the proposed method exhibited higher accuracy than the normal Faster R-CNN model, with an mAP of 0.97. Figure 7 shows the detection results of the normal Faster R-CNN model and the proposed approach. Owing to the complex background features in the images and the significant distance from the camera, the intermediate layers extracted erroneous image features, as shown in Figure 7(a). It was difficult for the normal Faster R-CNN model to recognize the human object 20 m from the camera, the obtained detection boxes exhibited numerous outlying areas, and the model failed to detect vehicles and humans at a significant distance (15 m, 20 m, and 25 m) from the camera accurately, as shown in Figures 7(a) and 7(b). However, as can be observed from Figures 7(c) and 7(d), the proposed approach detected humans and vehicles effectively in accurate detection boxes at a significant distance from the camera, accurately detected objects and further recognized their action 20 m from the camera (Figure 7(c)) and accurately detected human and vehicle areas in the detection boxes (Figure 7(d)). This demonstrates that the proposed method was effective in identifying actions in scenes with multiple people and could detect human actions and vehicles with an mAP of 0.97 by using the feature values obtained in the transfer learning and confusion region recognition anchor boxes generated by the improved Faster R-CNN model.



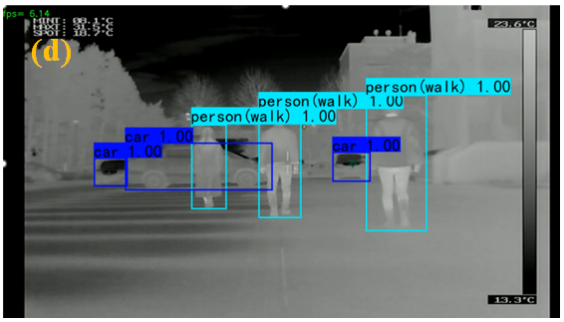
(a) False detection of humans (20 m)



(b) False detection of vehicles



(c) Accurate detection of humans (20 m)



(d) Accurate detection of humans and vehicles

FIGURE 7. Example of detection results by the normal Faster R-CNN model and proposed approach

4.3. **Results of comparison with CNN (VGG16), Multi-task Faster R-CNN, and YOLOv5.** To evaluate the proposed approach at varying distances, we compared it with several models, including CNN (VGG16), Multi-task Faster R-CNN, and YOLOv5, as shown in Table 3.

We provided annotations and labeled vehicle and human actions (standing, squatting, bending, and walking) in 15,184 infrared images. Furthermore, 3,272 images were used as

Table 3. Average detection precision results for each approach

Avg. (%)	Range (m)														
	[15, 20]					[20, 25]					[25, 35]				
	Stand	Squat	Bend	Walk	Veh.	Stand	Squat	Bend	Walk	Veh.	Stand	Squat	Bend	Walk	Veh.
CNN (VGG16)	94.41	87.64	88.53	93.15	90.92	85.43	83.16	80.51	85.15	83.54	40.28	44.21	40.21	44.50	58.33
Multi-task Faster R-CNN	90.85	96.52	98.12	92.96	93.18	84.63	87.90	90.80	86.69	87.32	52.22	54.31	58.45	50.78	65.29
YOLOv5	90.17	98.44	99.10	98.95	99.47	95.73	90.63	88.15	96.61	94.93	71.36	78.37	70.39	83.54	83.68
Proposed approach	94.89	99.78	98.52	95.52	94.24	97.70	99.24	98.73	98.78	99.64	96.53	99.13	99.46	99.92	99.01

Table 4. Average detection precision results for each range

Avg. (%)	Range (m)		
	[15, 20]	[20, 25]	[25, 35]
CNN (VGG16)	90.93	83.56	45.51
Multi-task Faster R-CNN	94.32	87.47	65.29
YOLOv5	97.23	93.21	83.68
Proposed approach	96.59	98.81	99.01

the test dataset to evaluate each detection and recognition model. The test dataset was divided according to distances in the ranges of 15 m to 20 m, 20 m to 25 m, and 25 m to 35 m. Table 3 presents the average detection precision results for each approach across different patterns and ranges, whereas Table 4 shows the average detection results in each range.

Although YOLOv5 exhibited higher precision than our method in the 15 m to 20 m range, our proposed method showed slight improvements over CNN (VGG16) and Multi-task Faster R-CNN, with average detection precision increases of 5.66% and 2.27%, respectively. For the 20 m to 25 m and 25 m to 35 m ranges, the proposed method achieved average precisions of 98.81% and 99.01% for each pattern. In the 20 m to 25 m range, our method yielded improvements over CNN (VGG16), Multi-task Faster R-CNN, and YOLOv5, with precision increases of 15.25%, 11.34%, and 5.6%, respectively. Furthermore, for the 25 m to 35 m range, our method exhibited significant improvements, with increases of 53.50%, 33.72%, and 15.33% over the respective approaches. These results indicate that the proposed method outperforms CNN (VGG16), Multi-task Faster R-CNN, and YOLOv5 in thermal infrared images. It is effective for detecting vehicle and human actions at nighttime across various distances, specifically in the ranges of 15 m to 20 m and 20 m to 25 m. Notably, it shows significant improvement in detecting objects at further distances, particularly in the 25 m to 35 m range.

5. **Conclusions.** One of the primary challenges of this study was achieving the precise detection of vehicles and human actions at significant distances, particularly under nighttime conditions and at further distances. Existing technologies such as CNN (VGG16), Multi-task Faster R-CNN, and YOLOv5 have exhibited limitations in accurately detecting objects at these extended distances in complex nighttime road traffic environments. To address these issues, we utilized an infrared thermal imaging camera to create a comprehensive nighttime database. This dataset comprises 15,184 thermal infrared images of various vehicle and human action patterns within distances ranging from 15 m to 35 m, providing a robust foundation for training and testing our models. Furthermore, we

proposed an enhanced Faster R-CNN model based on the ResNet-50 architecture by leveraging transfer learning. We also introduced a method to adjust the recognition anchor box sizes for each object within the confusion regions. The aim was to distinguish between vehicle and human action states automatically in outdoor scenes at night, in real time, and at varying distances from the camera. The results of experiments conducted for each pattern indicated that the proposed method was effective for vehicle detection and human action recognition, achieving an mAP of 0.97, thereby demonstrating its effectiveness in vehicle detection and human action recognition in complex nighttime environments. Importantly, we observed significant improvements in the detection accuracy, particularly at distances of 25 m to 35 m, where our method outperformed CNN (VGG16), Multi-task Faster R-CNN, and YOLOv5 by 53.50%, 33.72%, and 15.33%, respectively. It also exhibited superior accuracy compared to conventional methods. These results highlight the capability of our proposed method in addressing specific challenges posed by nighttime environments, including the detection of objects at significant distances and accurate action recognition. By elucidating these challenges and improvements, we aim to contribute valuable insights to the advancement of computer vision applications in low-visibility and long-range distance scenarios. In future research, we plan to investigate methods that consider the distances between the camera and the subject as well as between different subjects to further improve the accuracy and performance of the system.

Acknowledgment. This study was supported by a Grant-in-Aid for Scientific and Technological Research from the Suzuki Foundation.

REFERENCES

[1] Statistics Bureau of Japan, *Population Estimates: Result of the Population Estimates*, https://www.stat.go.jp/english/data/jinsui/tsuki/index.html, Accessed on November 22, 2023.

[2] L. Grinin, A. Grinin and A. Korotayev, Global aging: An integral problem of the future. How to turn a problem into a development driver?, in *Reconsidering the Limits to Growth. World-Systems Evolution and Global Futures*, V. Sadovnichy, A. Akaev, I. Ilyin, S. Malkov, L. Grinin and A. Korotayev (eds.), Cham, Springer International Publishing, 2023.

[3] S. Liu, T. Yamamoto, E. Yao and T. Nakamura, Examining public transport usage by older adults with smart card data: A longitudinal study in Japan, *Journal of Transport Geography*, vol.93, 103046, 2021.

[4] J. A. Kimlin, A. A. Black and J. M. Wood, Nighttime driving in older adults: Effects of glare and association with mesopic visual function, *Investigative Ophthalmology & Visual Science*, vol.58, no.5, pp.2796-2803, 2017.

[5] J. M. Wood, Nighttime driving: Visual, lighting and visibility challenges, *Ophthalmic and Physiological Optics*, vol.40, no.2, pp.187-201, 2017.

[6] F. F. Paschaline, R. A. Prastita and E. Mega, Japan's aging society: A challenge to Japan's diversity and social inclusion, *Transformasi Global*, vol.10, no.1, pp.20-34, 2023.

[7] Y. Kageyama, K. Suzuki, C. Ishizawa and T. Suzuki, Extraction and recognition of speed limit signs in night-scene videos, *Journal of the Institute of Industrial Applications Engineers*, vol.6, no.1, pp.29-33, 2018.

[8] T. Suzuki, Y. Kageyama and C. Ishizawa, Recognition method for speed limit signs and its applicability in recognition of vehicle entry prohibition signs at night, *IEEJ Transactions on Electrical and Electronic Engineering*, vol.15, no.10, pp.1448-1456, 2020.

[9] Y. Liu, K. Matsui, Y. Kageyama, H. Shirai and C. Ishizawa, A CNN-based method for human action analysis using nighttime infrared images, *International Journal of Innovative Computing, Information and Control*, vol.19, no.6, pp.1861-1875, 2023.

[10] Y. H. Liu, Feature extraction and image recognition with convolutional neural networks, *Journal of Physics: Conference Series*, vol.1087, 2018.

[11] K. Zhang, F. Xiong, P. Sun, L. Hu, B. Li and G. Yu, Double anchor R-CNN for human detection in a crowd, *arXiv Preprint*, arXiv: 1909.09998, 2019.

[12] A. Farid, F. Hussain, K. Khan, M. Shahzad, U. Khan and Z. Mahmood, A fast and accurate real-time vehicle detection method using deep learning for unconstrained environments, *Applied Sciences*, vol.13, no.5, 3059, 2023.
[13] H. Qu, L. Zhang, X. Wu, X. He, X. Hu and X. Wen, Multiscale object detection in infrared streetscape images based on deep learning and instance level data augmentation, *Applied Sciences*, vol.9, no.3, 565, 2019.
[14] M. Hasan, S. Ullah, M. J. Khan and K. Khurshid, Comparative analysis of SVM, ANN and CNN for classifying vegetation species using hyperspectral thermal infrared data, *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, no.42, pp.1861-1868, 2019.
[15] M. S. Farooq, H. Khalid, A. Arooj, T. Umer, A. B. Asghar, J. Rasheed, R. M. Shubair and A. Yahyaoui, A conceptual multi-layer framework for the detection of nighttime pedestrian in autonomous vehicles using deep reinforcement learning, *Entropy*, vol.25, no.1, 135, 2023.
[16] S. Iftikhar, Z. Zhang, M. Asim, A. Muthanna, A. Koucheryavy and A. A. A. El-Latif, Deep learning-based pedestrian detection in autonomous vehicles: Substantial issues and challenges, *Electronics*, vol.11, no.21, 3551, 2022.
[17] N. Arora, Y. Kumar, R. Karkra and M. Kumar, Automatic vehicle detection system in different environment conditions using Fast R-CNN, *Multimedia Tools and Applications*, vol.81, no.13, pp.18715-18735, 2022.
[18] X. Dai, J. Hu, H. Zhang, A. Shitu, C. Luo, A. Osman, S. Sfarra and Y. Duan, Multi-task Faster R-CNN for nighttime pedestrian detection and distance estimation, *Infrared Physics & Technology*, vol.115, 103694, 2021.
[19] Ultralytics, *Ultralytics YOLOv5*, https://github.com/ultralytics/yolov5, Accessed on June 10, 2024.
[20] A. Bochkovskiy, C. Y. Wang and H. Y. M Liao, YOLOv4: Optimal speed and accuracy of object detection, *arXiv Preprint*, arXiv: 2004.10934, 2020.
[21] Y. Zhang, Z. Guo, J. Wu, Y. Tian, H. Tang and X. Guo, Real-time vehicle detection based on improved YOLO v5, *Sustainability*, vol.14, no.19, 12274, 2022.
[22] *D-Eyes: Products*, https://d-eyes.co.jp/products, Accessed on November 30, 2023.
[23] *LabelImg*, https://github.com/heartexlabs/labelImg, Accessed on December 11, 2023.
[24] K. He, X. Zhang, S. Ren and J. Sun, Deep residual learning for image recognition, *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.770-778, 2016.
[25] S. Ren, K. He, R. Girshick and J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, *Advances in Neural Information Processing Systems*, vol.28, 2015.

Author Biography

Yaru Liu received the B.E. degree in Computer Science and Engineering from Qingdao Agricultural University, China, in 2019, and he received the M.E. degree in Computer Science and Engineering from Akita University, Japan, in 2022. He is now enrolled in a doctoral program with the Graduate School of Engineering Science in Akita University. His research interests include human sensing and image processing.

Kai Matsui received the B.E. and M.E. degrees in Computer Science and Engineering and the Dr. Eng. degree from Akita University, Japan, in 2017, 2019, and 2022, respectively. He is currently working as as an engineer at SUZIKI MOTOR CORPORATION. His research interests include remote sensing and image processing.

Yoichi Kageyama received the B.E. and M.E. degrees in Computer Science and Engineering and the Dr. Eng. degree from Akita University, Japan, in 1995, 1997, and 2001, respectively. He joined Akita University as a Research Associate in 1997. He became an Assistant Professor in 2001 and an Associate Professor in 2004. He is now a Professor with the Department of Mathematical Science and Electrical-Electronic-Computer Engineering, Graduate School of Engineering Science. His research interests include human sensing, remote sensing, and image processing.

Hikaru Shirai received the B.E., M.E., and Dr. Eng. degrees in Computer Science and Engineering from Akita University, Japan, in 2011, 2013, and 2017, respectively. He joined Akita Electronics Systems Co., Ltd. in 2013. He joined Ricoh IT Solutions Co., Ltd. in 2017. He joined Akita University as a Technical Staff in 2019. He became an Assistant Professor in 2020. He is now a Lecturer with the Department of Mathematical Science and Electrical-Electronic-Computer Engineering, Graduate School of Engineering Science. His research interests include remote sensing and image processing.

Chikako Ishizawa received the B.E. degree in Chemical Engineering for Resources from Akita University, Japan, in 1992, and joined FUJIFILM Software Co., Ltd. She joined Akita University in 1995. She received a Dr. Eng. degree from Akita University in 2012. She is now a Professor with the Department of Mathematical Science and Electrical-Electronic-Computer Engineering, Graduate School of Engineering Science. Her research interests include visual information processing and log analysis.