

Article

CNN-Based Person Detection Using Infrared Images for Night-Time Intrusion Warning Systems

Jisoo Park ¹, Jingdao Chen ¹, Yong K. Cho ¹, Dae Y. Kang ² and Byung J. Son ^{3,*}

¹ School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA; jpark711@gatech.edu (J.P.); jchen490@gatech.edu (J.C.); yong.cho@ce.gatech.edu (Y.K.C.)

² College of Sports and Leisure Studies, Yonsei University, Seoul 03722, Korea; ctokang@lskorea.org

³ Department of International Civil & Plant Engineering, Konyang University, Nonsan 32992, Korea

* Correspondence: strustar@konyang.ac.kr

Received: 12 November 2019; Accepted: 17 December 2019; Published: 19 December 2019



Abstract: Night-time surveillance is important for safety and security purposes. For this reason, several studies have attempted to automatically detect people intruding into restricted areas by using infrared cameras. However, detecting people from infrared CCTV (closed-circuit television) is challenging because they are usually installed in overhead locations and people only occupy small regions in the resulting image. Therefore, this study proposes an accurate and efficient method for detecting people in infrared CCTV images during the night-time. For this purpose, three different infrared image datasets were constructed; two obtained from an infrared CCTV installed on a public beach and another obtained from a forward looking infrared (FLIR) camera installed on a pedestrian bridge. Moreover, a convolution neural network (CNN)-based pixel-wise classifier for fine-grained person detection was implemented. The detection performance of the proposed method was compared against five conventional detection methods. The results demonstrate that the proposed CNN-based human detection approach outperforms conventional detection approaches in all datasets. Especially, the proposed method maintained F1 scores of above 80% in object-level detection for all datasets. By improving the performance of human detection from infrared images, we expect that this research will contribute to the safety and security of public areas during night-time.

Keywords: lifesaving; human detection; deep learning; CNN; infrared image

1. Introduction

Night-time intrusion warning systems are some of the most important applications of IoT (Internet of Things) technology. Night-time intruders into dangerous areas face a high risk of injury or death due to the lack of safety personnel during the night and limited visibility. Having an automated warning system to deter intruders is crucial for both the safety of the general public and the security of restricted areas. One of the most important applications for night-time intrusion warning is drowning prevention. According to the World Health Organization (WHO), 372,000 people die due to drowning accidents every year [1]. Although most of the drowning accidents happen during daytime, the night-time intrusion warning system for drowning prevention is important because during the night-time, lifeguards might be absent, and rescue operations are more difficult [2]. As part of ongoing efforts to prevent these drowning accidents, several intelligent surveillance systems have been developed using closed-circuit television (CCTV) cameras paired with computer vision and machine learning technologies [2–4]. These systems automatically detect people swimming between the established boundaries in CCTV images and issue an alert to the supervisor whenever the system finds someone trespassing across the safety lines [4]. However, most systems operate with images

taken in daylight, whereas the performance for detecting people in the water during the night-time is considerably worse.

To detect people more effectively at night, several studies have adopted machine learning-based approaches using specific features extracted from infrared camera images. These machine learning-based human detection approaches can be classified according to the type of feature descriptors and classifiers that are applied. For example, histograms of oriented gradients (HOG) [5,6], scale-invariant feature transform (SIFT)-like oriented features [7], and intensity distribution-based inertia features (INERTIA) [8] are widely used for feature extraction, and support vector machines (SVM) [9–11], or AdaBoost [11] are frequently used as classifiers. In addition, the Haar-Cascade classifier, which is a well-known object detection method based on Haar-like features with boosted cascade classifiers, is also adopted for human detection in infrared images in several studies [12–14]. If proper features and feature extraction techniques are selected, classical machine learning approaches usually show satisfactory detection performance. On the other hand, these approaches suffer from low detection performance in images containing ambiguous or noisy features [15–17].

To address this problem, deep learning approaches based on convolution neural networks (CNN) have recently been presented for person detection in infrared images [18–21]. These deep learning-based approaches can enhance the object detection performance in infrared images since the increased model complexity leads to better learning capacity. It has been found in the literature that CNN-based human detection approaches show better detection rates compared to SVM-based classification approaches [16,22]. The detection performance of deep learning-based approaches depends on having the right training data; thus, constructing an appropriate dataset for the specific situation of detecting swimmers at night-time is critical. Several studies have developed infrared image datasets such as KAIST [23], CVC-14 [24], and LSI [25] for human detection during night-time. However, since these infrared image datasets aim to detect pedestrians, they consist of images captured from cameras installed on a ground vehicle. The people in these images have a significantly different appearance compared to that of CCTV installed on the beach for safety purposes in which people are represented by only a few pixels (i.e., part of the body above water). Therefore, in order to achieve the purpose of detecting people in CCTV images for long-range surveillance during night-time, there is a need for more specific datasets and a more fine-grained object detection method.

This research proposes an accurate and efficient method for detecting people in infrared CCTV installed at overhead locations during the night-time. For this purpose, a CNN-based pixel-wise classifier for fine-grained person detection was proposed. This study acquired three different datasets based on images taken with infrared cameras from an overhead point-of-view. In addition, this paper also proposes a modified CNN model in which the three input channels are taken from an (i) original infrared image, (ii) difference image from the previous frame, (iii) background subtraction mask. To verify the proposed methods, a comparative analysis with four different human detection methods is conducted. The main contributions of this research are

- (i) to introduce a fine-grained person detector that is 10–30% more accurate compared to baseline methods;
- (ii) to demonstrate the ability to perform long-range object detection under challenging conditions;
- (iii) to validate the ability of the trained detector to generalize across different datasets.

2. Methodology

2.1. Composition of Datasets

This research used three infrared image datasets, named “university”, “beach”, and “shore”, for training and testing the person detection algorithms. The “university” dataset was taken from on top of a pedestrian bridge at a university campus using an infrared (FLIR) camera. The “beach” and “shore” datasets were taken from a rooftop of a building next to a public beach in Korea using an infrared CCTV. All datasets were collected at night-time and labeled at the pixel level. Table 1

shows the details of the datasets whereas Figures 1–3 depict sample original and labeled images in all datasets. All datasets were split into 80% training data and 20% test data. The split was carried out while ensuring that the images in the training data and test data occur at different time segments in the original video.

The “university” dataset was used to benchmark the performance of night-time human detection from an overhead viewpoint in brightly lit urban areas. The dataset was taken on a university campus with many different light sources that pose a challenge to intensity-based detection algorithms. On the other hand, the “beach” and “shore” datasets were used to practically evaluate human detection for night-time drowning prevention. Figure 4 shows the layout of the night-time scene captured by the beach CCTV camera. The scene consists of a public beach bordered by safety buoys with several groups of people scattered throughout. The task of interest was to automatically detect any persons swimming in the sea area, which is prohibited during night-time. This is a challenging task because the area of interest only constitutes a small region in the image since the camera is located far from the scene. Other challenging factors include noise from the waves and irregular lighting from the surrounding buildings. To partially overcome these problems, the original image was cropped to contain only the area of interest which is the sea area close to the beach. For this reason, only the people that intruded into the water area were detected, whereas the people on the beach were not detected.

Table 1. Details of each dataset.

Dataset	University	Beach	Shore
Number of images	120	160	540
Resolution	1364 × 1024	1920 × 1080	1920 × 1080
Duration	120 s	40 s	180 s
Sampling rate	1 Hz	4 Hz	3 Hz

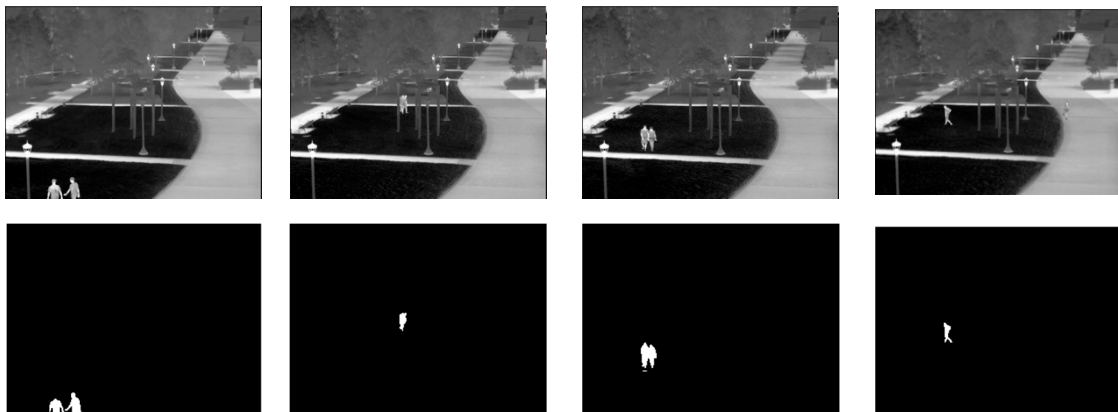


Figure 1. Example images and ground truth labels for the university dataset.

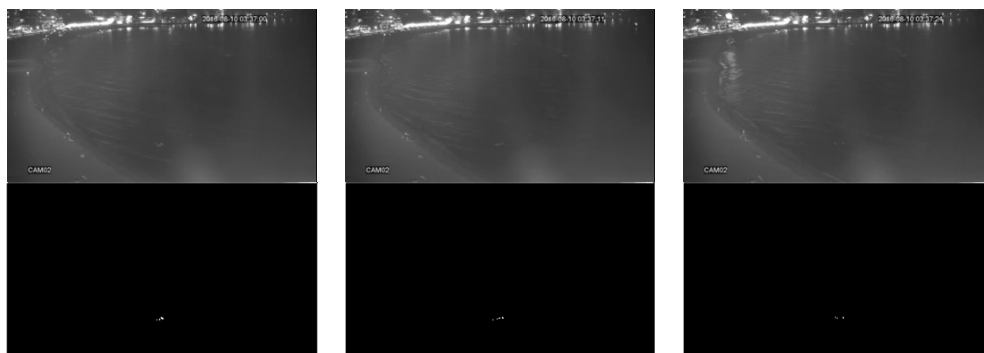


Figure 2. Example images and ground truth labels for the beach dataset.

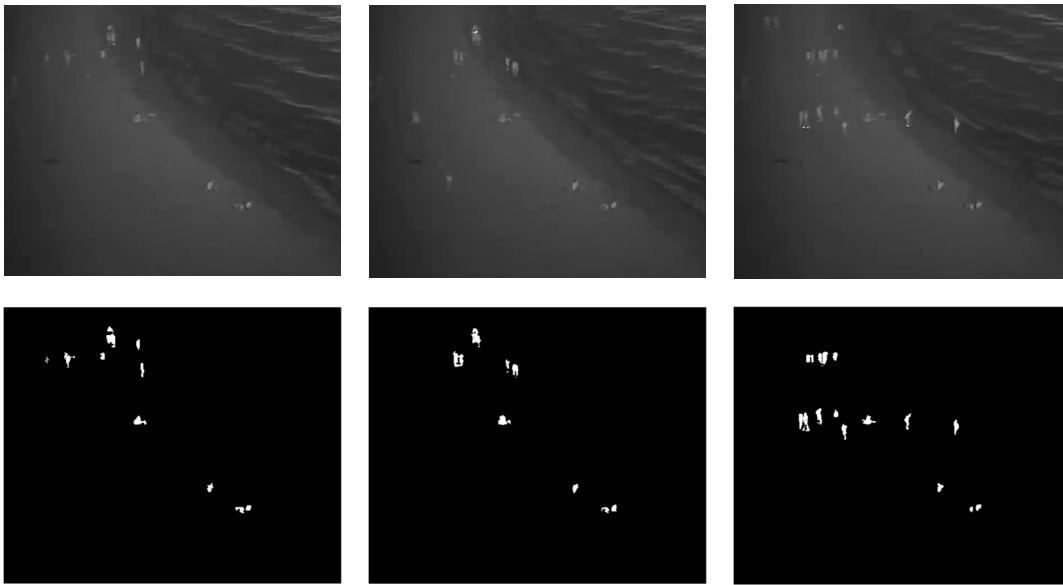


Figure 3. Example images and ground truth labels for the shore dataset.

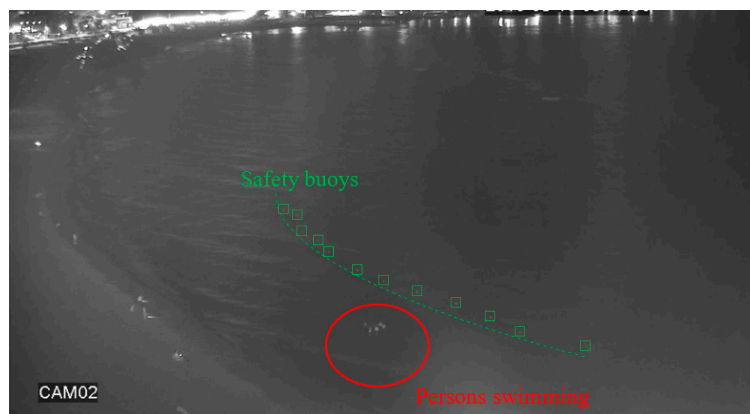


Figure 4. Visualization of the area of interest for surveillance in the beach dataset.

2.2. Human Detection from Infrared Images

This research implemented six different methods of detecting people in infrared images including our proposed method and five baseline methods: (i) simple thresholding, (ii) adaptive thresholding, (iii) background subtraction, and (iv) K-means clustering and (v) Convolutional Neural Networks (CNN). The detection task was formulated as binary classification at the pixel level, i.e., each pixel is classified as either part of a person or not. This allows for a fine-grained detection where the exact object outline can be determined compared to bounding boxes which are inexact and only give a rough approximation of the object [26].

2.2.1. Human Detection Using Conventional Computer Vision

The simplest method of detecting humans in infrared images is through thresholding. The technique of simple thresholding applies a constant threshold to the pixel intensity to determine pixels that correspond to humans. Since humans usually have a higher temperature than the surroundings, the corresponding pixels in infrared images will have higher intensities. Compared to simple thresholding, adaptive thresholding does not use a fixed threshold but a varying threshold that is determined by the Gaussian-weighted sum of the local intensity values. This allows the thresholding procedure to be more robust to changes in lighting conditions.

This study also considers the background subtraction method based on the fact that humans usually move relative to the surrounding environment. With background subtraction, the human can be distinguished from a sequence of infrared images as pixels that are changing over time in contrast to pixels that remain constant. This method is applicable in this research because the CCTV camera is installed with a fixed location and viewing angle; thus, any substantial change in pixel values can be attributed to moving objects. The background subtraction method implemented is based on the method of [27], where Gaussian mixture models are used to represent pixel intensity variation.

In addition, the K-means clustering method, which is commonly used to segment regions of an image that have a certain range of intensity values [28–30], was considered. The intuition is that the foreground objects such as humans will have a certain distribution of pixel values that is distinct from that of background pixels. Thus, the K-means algorithm was used to determine the mean of each distribution (i.e., foreground vs. background) so that foreground and background pixels can be separated.

2.2.2. Human Detection Using Deep Neural Networks

Deep neural networks have been shown in the literature to demonstrate state-of-the-art performance in various image processing tasks [31–33]. Even though infrared images may have a different appearance compared to visual images, similar techniques can still be applied. This study used CNN designed for the task of semantic image segmentation. The architecture has a base network of ResNet [31] and top layers with atrous convolutions [34]. The network takes in inputs of $N \times N \times 3$ images and outputs $N \times N$ pixel-wise labels. The source code implemented in this paper has been made publicly available on GitHub (https://github.com/jingdao/IR_detection).

Figure 5 shows the detailed network architecture. The three input channels were taken from an (i) original infrared image, (ii) difference image from the previous frame, (iii) background subtraction mask. This is so that the network is able to capture temporal information as well when making predictions. The network was trained on the binary classification task using the cross-entropy loss. The output of the network is a 2D array of confidence scores for each pixel. The array is filtered with a detection threshold of 0.99 to create a binary mask. The high level of the detection threshold is used to ensure that noise in the detection results is adequately filtered out. To determine the actual number of persons, the connected component algorithm was used to merge neighboring pixels together in the binary mask to form individual objects. The network was trained for a total of 150 epochs with a learning rate of 0.001 and a batch size of 5.

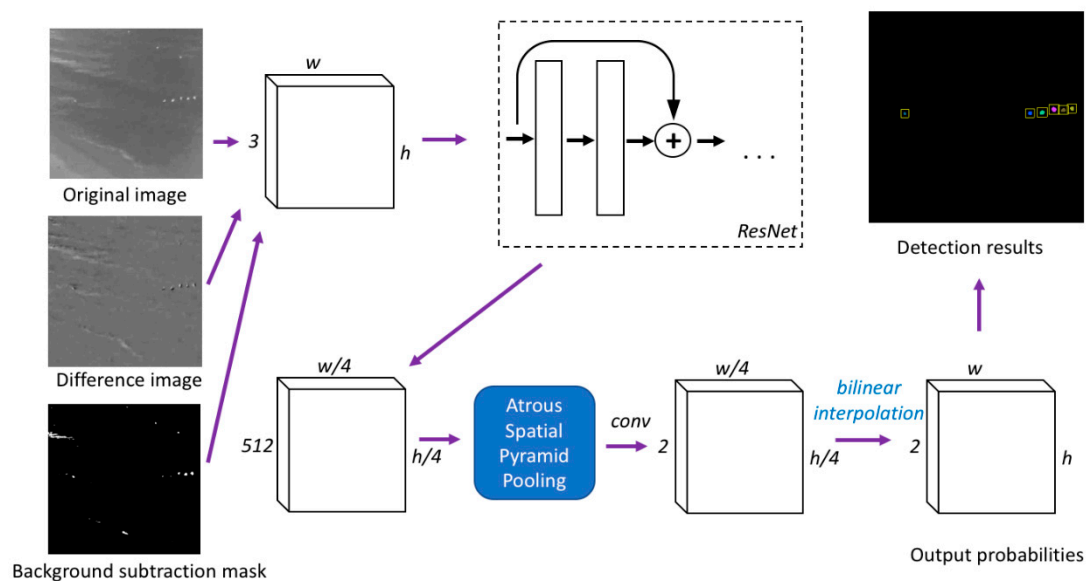


Figure 5. The network architecture of the proposed person detection method.

Two network models were trained for each dataset. In the rest of this paper, “university*”, “beach*”, and “shore*” refer to the scenario where all three input channels (i.e., containing temporal information) were used whereas “university”, “beach”, and “shore” refer to the scenario where only the first input channel was used (i.e., no temporal information). Figure 6 shows a graph of the training loss for each dataset for the first 20 epochs of training. The training progress shows a similar trend for all datasets, where the loss function rapidly decreases during the first few epochs and then gradually plateaus afterward. The graph also shows that the training loss is generally higher for the “university” dataset compared to the “beach” and “shore” datasets. This may be because the “university” dataset has a wider range of pixel intensity values since it is in a lighted urban scene.

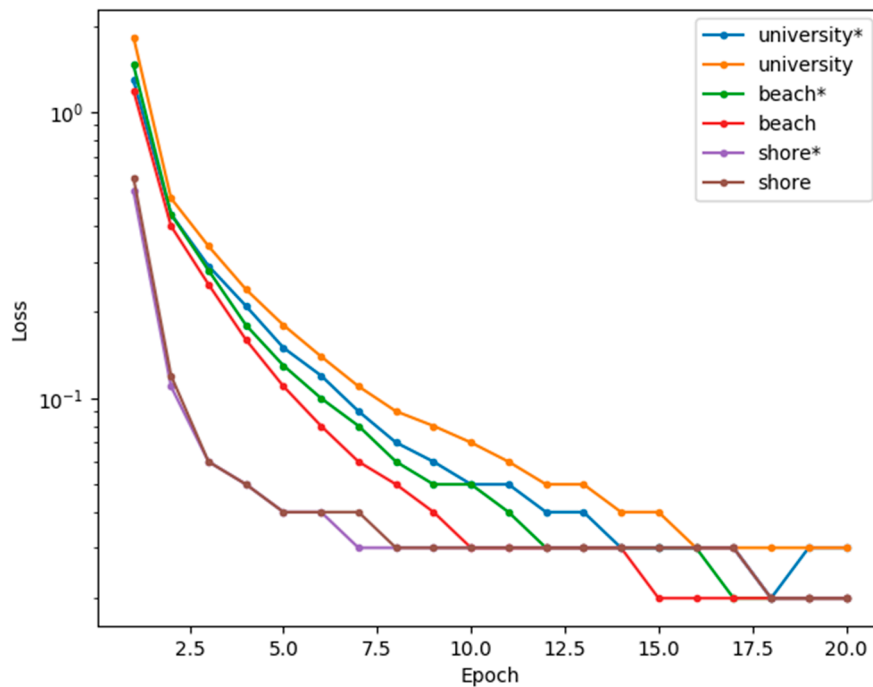


Figure 6. Loss function (logarithmic scale) of the first 20 epochs during training for each dataset. (*) indicates that temporal input is used.

There are several challenges in implementing a segmentation network for our particular dataset compared to conventional visual images. First, the pixels that correspond to positive detections only occupy small regions in the image since the images are taken from an overhead view. To overcome this problem, a modified version of ResNet was used with a much smaller output stride while reducing the feature map depth to keep the memory consumption at a reasonable level. This allows the network to make fine-grained predictions in terms of the pixel neighborhood. Additionally, since the label distribution is severely skewed (each image has around 99% background pixels and only 1% foreground pixels), a weighted cross-entropy loss function was used where the loss from background pixels and foreground pixels were separately weighted based on their frequencies.

3. Experimental Results

The performance of the proposed human detection method in infrared images was evaluated through the “university”, “beach”, and “shore” datasets. The evaluation metrics are (i) precision, (ii) recall, and (iii) F1-score, measured at both the pixel level and the object level. The proposed method was evaluated under two configurations: one where only the first input channel was used (i.e., no temporal information) and one where all three input channels (i.e., containing temporal information) were used (indicated by *). The proposed method was compared against five baseline methods which were (i) simple thresholding, (ii) adaptive thresholding, (iii) background subtraction,

and (iv) K-means clustering (v) baseline CNN. The baseline CNN is a semantic segmentation CNN adapted directly from [34] without any of our modifications such as output stride tuning or weighted loss function.

Figures 7–9 show a visualization of the detection results compared to the baseline methods whereas Tables 2–7 show the numerical results. Figures 7–9 show that our method obtains much better detection results, almost approaching the ground truth, compared to the baseline methods. Our method is able to correctly detect humans while ignoring background objects such as lamp posts and waves. This is confirmed by the numerical results in Tables 2–7, where our method obtains the best pixel-level and object-level F1-scores in all datasets. In the “university” dataset, the use of temporal information as CNN input is able to improve the detection performance whereas, in the “beach” and “shore” datasets, the temporal information was not as effective. This is because the “beach” and “shore” datasets are highly dynamic due to moving waves, thus the detector could become overly sensitive to random motions of the waves and cause false positives.

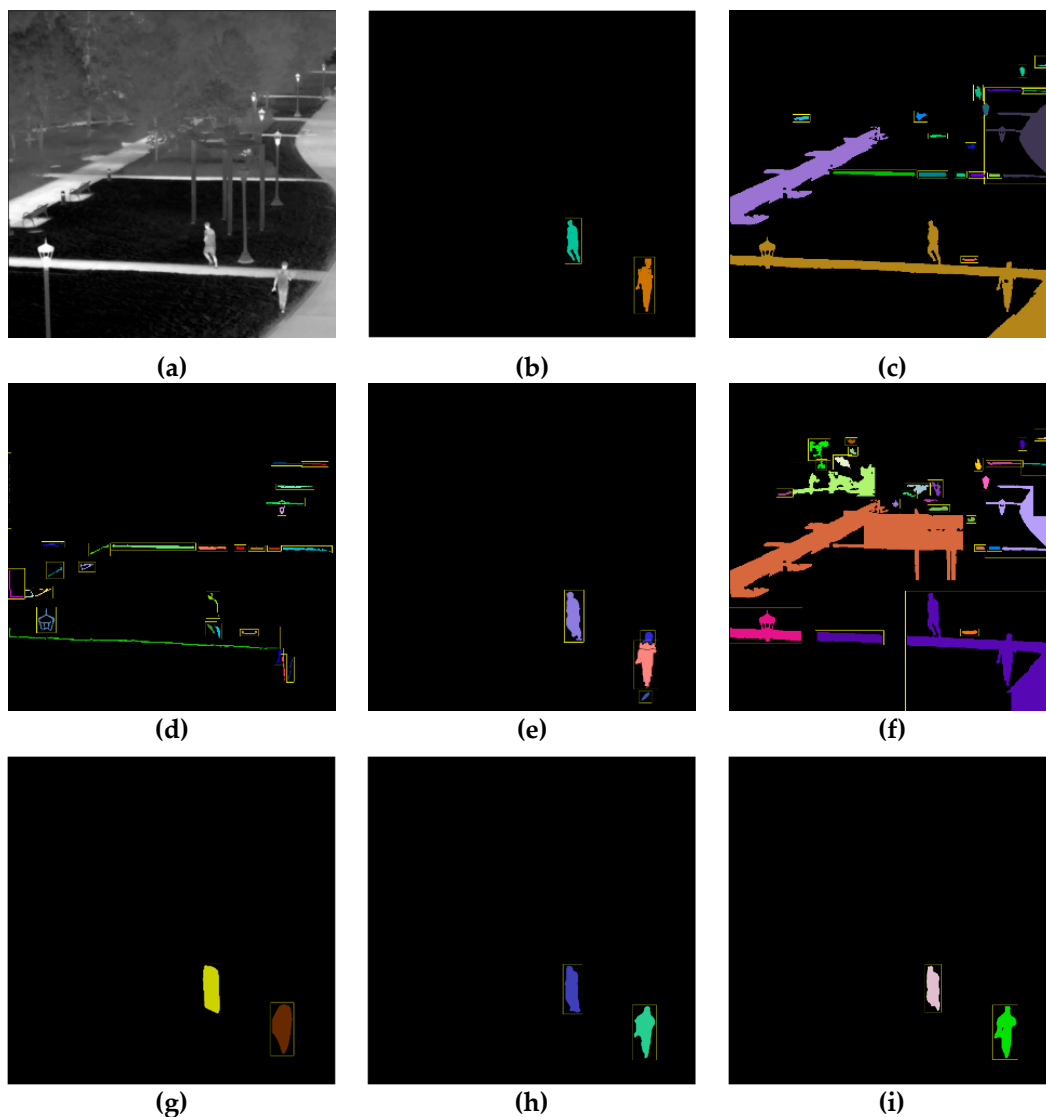


Figure 7. Detection results for the university dataset: (a) original infrared image, (b) ground truth, (c) simple thresholding, (d) adaptive thresholding, (e) background subtraction, (f) K-means clustering, (g) baseline convolution neural network (CNN), (h) our method, and (i) our method*.

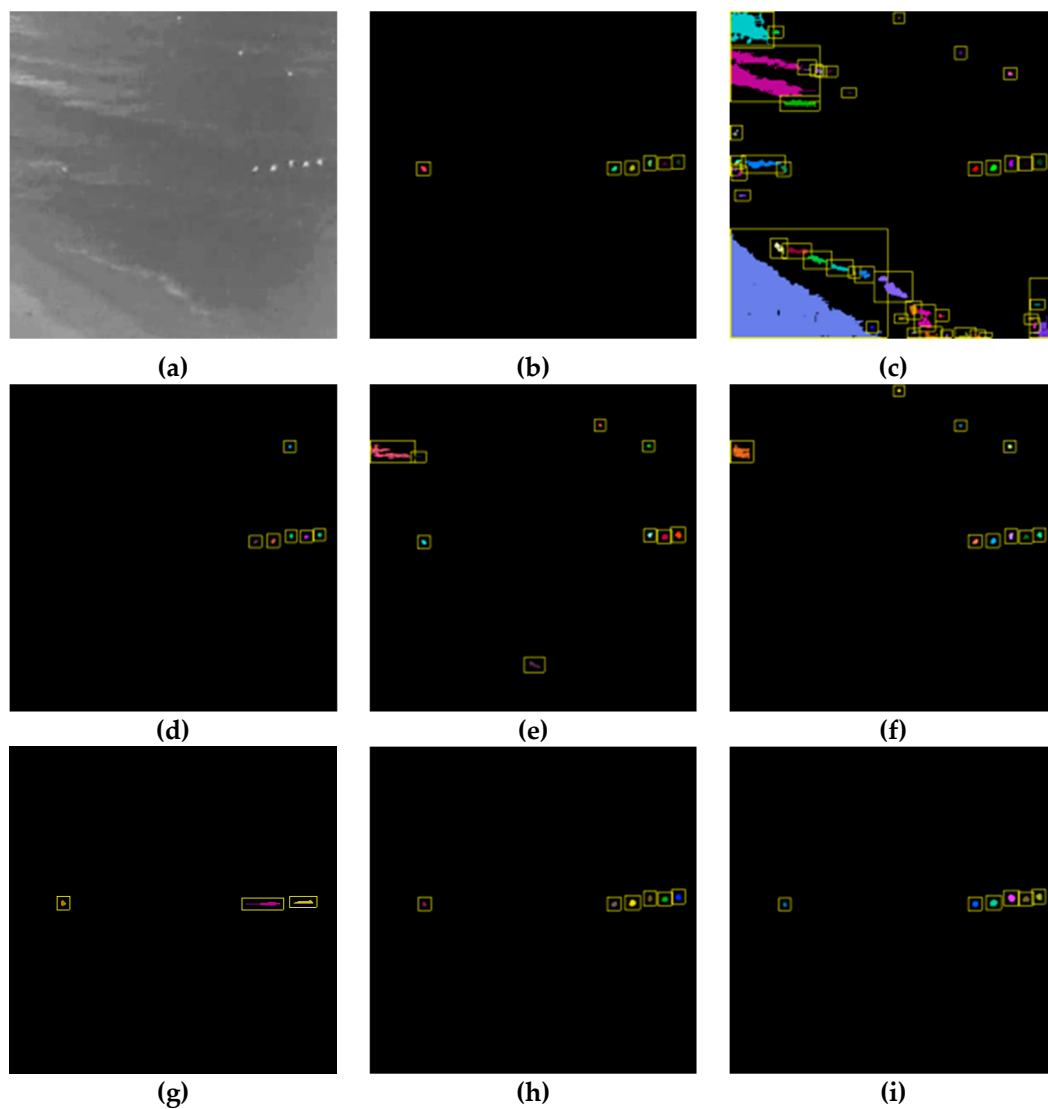


Figure 8. Detection results for the beach dataset: (a) original infrared image, (b) ground truth, (c) simple thresholding, (d) adaptive thresholding, (e) background subtraction, (f) K-means clustering, (g) baseline CNN, (h) our method, and (i) our method*.

Table 2. Pixel-level detection score for the university dataset.

Method	Precision	Recall	F1-Score
Simple thresholding	0.040	0.871	0.076
Adaptive thresholding	0.083	0.312	0.131
Background subtraction	0.534	0.951	0.684
K-means clustering	0.037	0.966	0.071
Baseline CNN	0.503	0.901	0.646
Our method	0.726	0.919	0.811
Our method*	0.759	0.948	0.843

(*) indicates that temporal input is used.

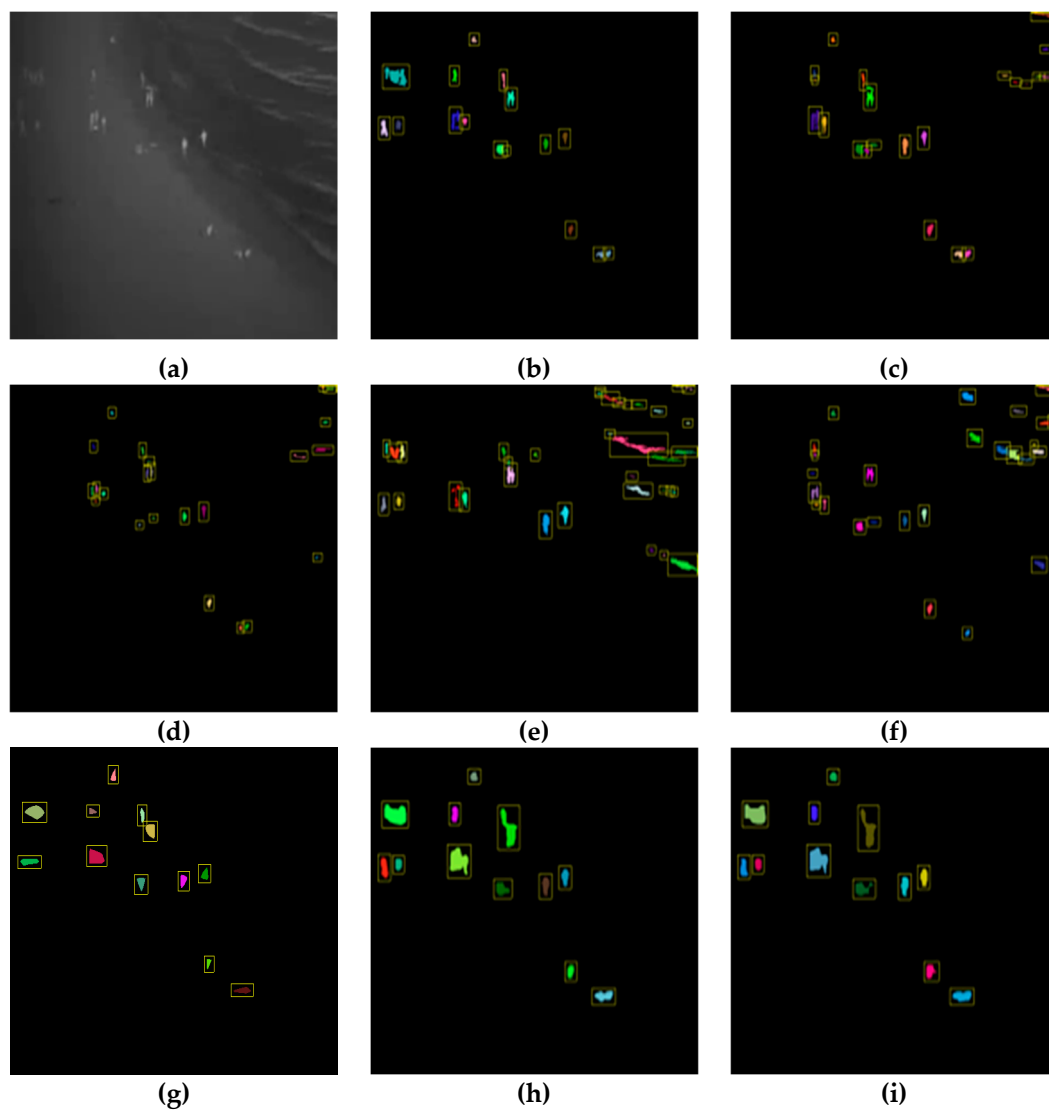


Figure 9. Detection results for the shore dataset: (a) original infrared image, (b) ground truth, (c) simple thresholding, (d) adaptive thresholding, (e) background subtraction, (f) K-means clustering, (g) baseline CNN, (h) our method, and (i) our method*.

Table 3. Object-level detection score for the university dataset.

Method	Precision	Recall	F1-Score
Simple thresholding	0.084	0.909	0.153
Adaptive thresholding	0.068	0.932	0.128
Background subtraction	0.506	1.000	0.672
K-means clustering	0.066	0.886	0.122
Baseline CNN	0.976	0.909	0.941
Our method	1.000	0.909	0.952
Our method*	1.000	0.932	0.965

(*) indicates that temporal input is used.

Table 4. Pixel-level detection score for the beach dataset.

Method	Precision	Recall	F1-Score
Simple thresholding	0.013	0.971	0.025
Adaptive thresholding	0.802	0.601	0.687
Background subtraction	0.214	0.590	0.314
K-means clustering	0.687	0.725	0.705
Baseline CNN	0.510	0.471	0.490
Our method	0.750	0.796	0.772
Our method*	0.645	0.760	0.698

(*) indicates that temporal input is used.

Table 5. Object-level detection score for the beach dataset.

Method	Precision	Recall	F1-Score
Simple thresholding	0.131	0.953	0.230
Adaptive thresholding	0.769	0.836	0.801
Background subtraction	0.313	0.661	0.425
K-means clustering	0.622	0.684	0.652
Baseline CNN	0.971	0.585	0.730
Our method	1.000	0.877	0.935
Our method*	0.961	0.860	0.907

(*) indicates that temporal input is used.

Table 6. Pixel-level detection score for the shore dataset.

Method	Precision	Recall	F1-Score
Simple thresholding	0.383	0.801	0.518
Adaptive thresholding	0.254	0.320	0.283
Background subtraction	0.293	0.551	0.383
K-means clustering	0.302	0.649	0.412
Baseline CNN	0.478	0.569	0.520
Our method	0.711	0.914	0.800
Our method*	0.652	0.933	0.768

(*) indicates that temporal input is used.

Table 7. Object-level detection score for the shore dataset.

Method	Precision	Recall	F1-Score
Simple thresholding	0.459	0.834	0.593
Adaptive thresholding	0.425	0.788	0.552
Background subtraction	0.269	0.532	0.357
K-means clustering	0.349	0.819	0.490
Baseline CNN	0.974	0.569	0.718
Our method	0.977	0.813	0.888
Our method*	0.948	0.741	0.832

(*) indicates that temporal input is used.

Detecting swimming people in the “beach” dataset is challenging because swimmers only appear as small white dots. The conventional detection methods such as thresholding or background subtraction do not work well due to the safety buoys and moving waves that reflect the background light and also appear as small white dots in the IR images. As shown in Figure 8, these confounding factors can cause multiple false positives for simpler detection methods. On the other hand, the proposed CNN

approach is able to take into account higher-level features such as shape and intensity variations that help to distinguish swimmers. Tables 4 and 5 demonstrate that our proposed method outperforms simpler methods such as background subtraction in the task of human detection.

When compared to the baseline CNN approach, our proposed approach is still able to outperform it in all three datasets according to the results in Tables 2–7. The proposed approach has similar object-level detection scores but much higher pixel-level detection scores compared to the baseline CNN. This effect is apparent in Figures 7g and 8g, where the baseline CNN method was only able to identify the coarse outline of people. In contrast, our method was able to label specific pixels accurately. This suggests that our proposed improvements to the CNN training such as weighted loss function and output stride tuning help to improve the network’s ability to capture finer details at the pixel level compared to the baseline CNN.

Figure 10 shows the detection results over time for the beach dataset. The detection results show that the proposed method is able to successfully detect people in the water even though they are constantly moving to different positions in the scene.

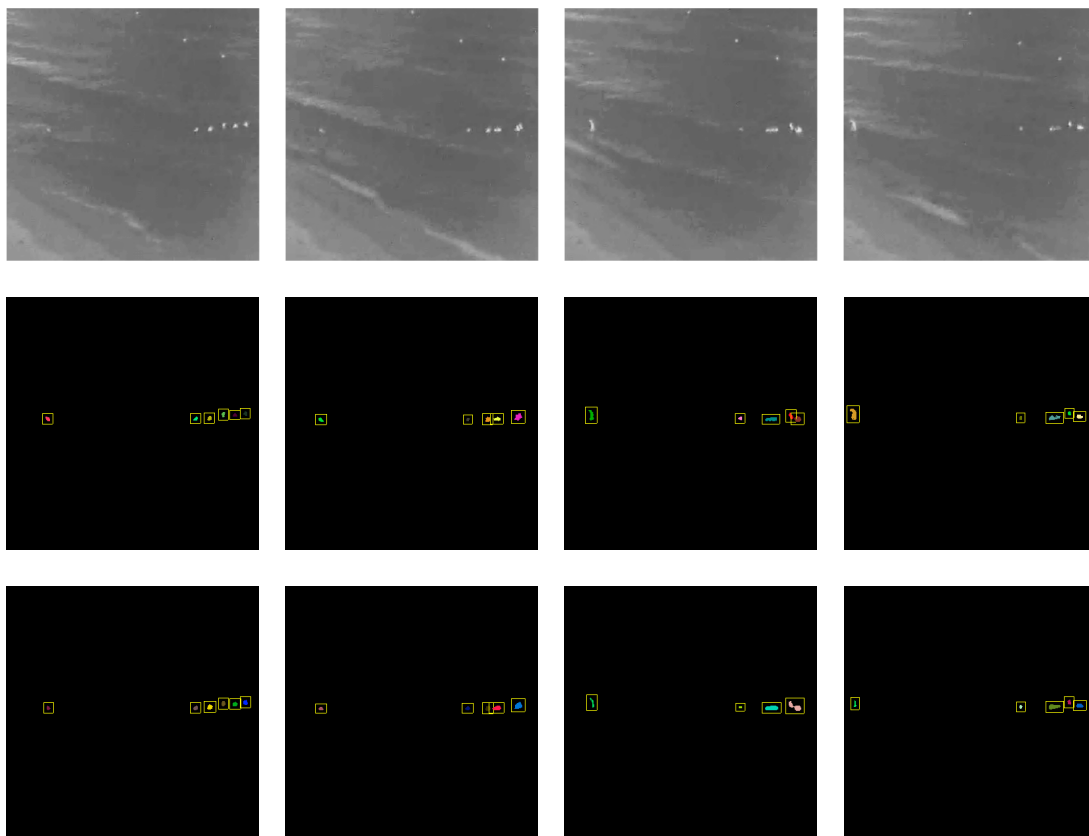


Figure 10. Detection results over time for the beach dataset. The top row shows the original infrared image; the middle row shows the ground truth; the bottom row shows the detection results based on our proposed method.

Figure 11 shows the precision-recall curve for pixel-wise detections in the “beach”, “shore”, and “university” datasets. The curve is constructed by applying different detection thresholds to the pixel-wise CNN confidence output. As the detection threshold increases, the precision values increase whereas the recall values decrease. To maximize the F1-score, a final detection threshold of 0.99 was used.

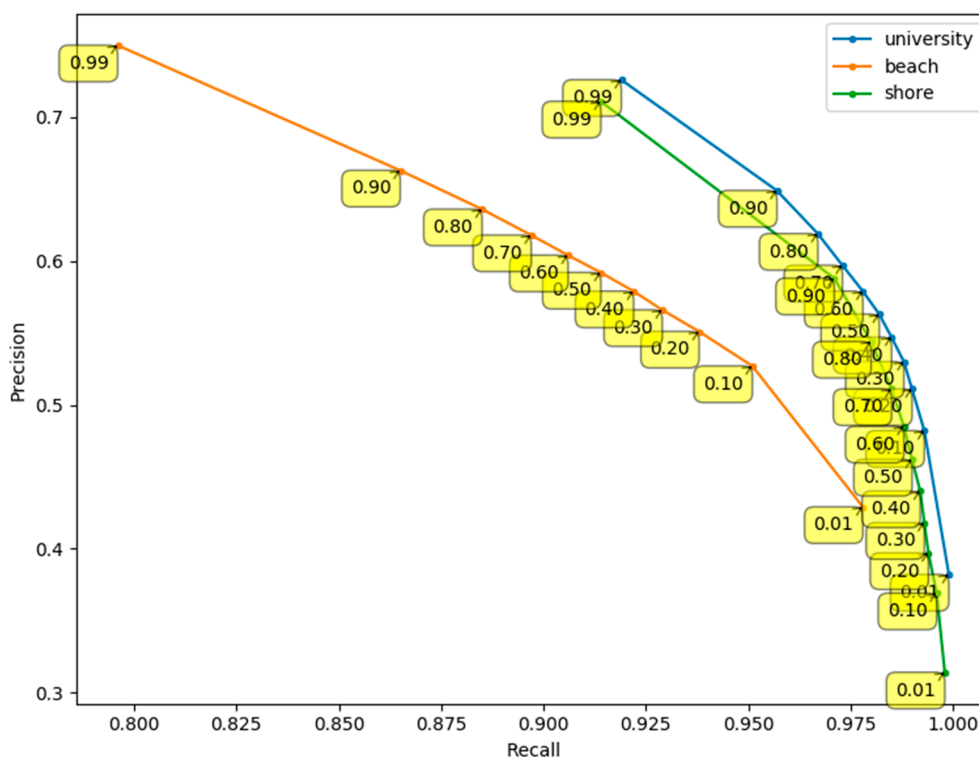


Figure 11. Precision-recall curve of pixel-wise detection results. Yellow boxes indicate the detection threshold at each point on the curve.

To determine whether the detector trained on detecting humans in one dataset is able to generalize and detect humans in another dataset, this research conducted a cross-comparison study where the detector was trained and tested on different datasets. This is a challenging task since the appearance, object sizes, and background scenery between the university and beach datasets are significantly different and some of the features learned might not generalize. Table 8 shows the pixel-level and object-level F1 scores with different train/test configurations. The F1 scores are lower when using a test dataset that is different from the training dataset than when they are the same; however, the results still outperform most of the baseline methods (refer to Tables 2–5). This suggests that the human detector has some ability to generalize across datasets, even though the objects in the training and test sets are dissimilar. The use of temporal information as CNN input has inconsistent results, i.e., the performance improves under certain configurations but not others. This suggests that the network is sensitive to temporal changes and may overfit to training data with noisy movements in the background.

Table 8. Cross-comparison of F1-scores with different train/test configurations. Each cell shows the pixel-level F1 score on the left and the object-level F1-score on the right.

Train Dataset \ Test Dataset	University	Beach
University	0.811 / 0.952	0.436 / 0.749
University*	0.843 / 0.965	0.278 / 0.439
Beach	0.294 / 0.383	0.772 / 0.935
Beach*	0.089 / 0.354	0.698 / 0.907
University + Beach	0.779 / 0.901	0.435 / 0.702
University + Beach*	0.618 / 0.782	0.604 / 0.825

(*) indicates that temporal input is used.

Table 9 shows the computation time per frame for each method, measured with an Intel Xeon E3-1246 (3.5 GHz) CPU and an NVIDIA GTX 1080 GPU. In general, our method requires more computation time compared to the baseline methods, other than K-means clustering. This is due to the high computational demands of using a deep neural network. However, our method is still able to process the images at a high frame rate (i.e., 6–11 fps), which is sufficient for timely safety warning.

Table 9. Computation time per frame for each method.

Method	Dataset #1 Processing Time (s)	Dataset #2 Processing Time (s)	Dataset #3 Processing Time (s)
Simple thresholding	0.001	<0.001	<0.001
Adaptive thresholding	0.008	<0.001	0.003
Background subtraction	0.007	0.001	0.007
K-means clustering	1.947	0.327	2.36
Baseline CNN	0.148	0.102	0.068
Our method	0.126	0.084	0.061
Our method*	0.156	0.088	0.070

(*) indicates that temporal input is used.

4. Conclusions

This study aimed to use infrared cameras to automatically detect people intruding into hazardous areas during night-time to help prevent accidents. To solve this problem, three infrared image datasets for training and testing a person detection algorithm were created. Next, this research developed a CNN-based human detection approach that can perform pixel-wise segmentation and make fine-grained predictions in terms of the object neighborhood. To verify the detection performance of the proposed method, comparative analyses were conducted with four baseline object detection methods. Experimental results demonstrate that the proposed CNN-based human detection approach outperforms conventional detection approaches in all datasets. Especially, the proposed method maintained F1 scores of above 90% in object-level detection for all datasets. Moreover, the computational speed of the proposed method was adequate for real-time safety warning purposes.

This study is expected to contribute to the safety and security of public areas during night-time by improving the performance of human detection from infrared images. However, since the datasets were obtained at specific times and locations, the proposed method may lack generality when applied to different datasets. Therefore, future research will be focused on constructing datasets with infrared images acquired from a wider variety of viewing angles and locations. In addition, more advanced methods for incorporating temporal image information such as Recurrent Neural Networks (RNN) will be explored.

Author Contributions: The work in this study is the collaborative development of all authors. J.P. and Y.K.C. contributed to conceptualization. J.P. and J.C. made contributions to developing methodology and writing the paper. Y.K.C. and B.J.S. undertook project administration and reviewed the paper. D.Y.K. contributed to data curation and funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Ministry of Land, Infrastructure, and Transport (MOLIT) of Korea Agency for Infrastructure Technology Advancement (KAIA), grant number (18CTAP-C144787-01), and the National Research Foundation (NRF) of Korea, grant number NRF-2017R1D1A03028210.

Acknowledgments: This research was supported by the Ministry of Land, Infrastructure, and Transport (MOLIT) of Korea Agency for Infrastructure Technology Advancement (KAIA) and National Research Foundation (NRF) of Korea. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of MOLIT and NRF. Their financial support is gratefully acknowledged.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. World Health Organization (WHO). *Global Report on Drowning: Preventing a Leading Killer*; WHO: Geneva, Switzerland, 2015; pp. 1–59.
2. Kang, D.Y.; Hong, S.; Lee, J. An introduction of AELIS (Aquatic environment life-saving integrated system). *J. Coast. Res.* **2017**, *79*, 11–15. [[CrossRef](#)]
3. Alam, M.; Ferreira, J.; Mumtaz, S.; Jan, M.A.; Rebelo, R.; Fonseca, J.A. Smart cameras are making our beaches safer: A 5G-envisioned distributed architecture for safe, connected coastal areas. *IEEE Veh. Technol. Mag.* **2017**, *12*, 50–59. [[CrossRef](#)]
4. Lee, J.; Park, J.; Kim, I.; Kang, D.Y. Application of vision-based safety warning system to haeundae beach, Korea. *J. Coast. Res.* **2019**, *91*, 216–220. [[CrossRef](#)]
5. Suard, F.; Rakotomamonjy, A.; Bensrhair, A.; Broggi, A. Pedestrian detection using infrared images and histograms of oriented gradients. In Proceedings of the 2006 IEEE Intelligent Vehicle Symposium, Tokyo, Japan, 13–15 June 2006; pp. 206–212. [[CrossRef](#)]
6. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–25 June 2005; pp. 886–893. [[CrossRef](#)]
7. Mikolajczyk, K.; Schmid, C.; Zisserman, A. Human detection based on a probabilistic assembly of robust part detectors. In Proceedings of the 8th European Conference on Computer Vision, Prague, Czech Republic, 11–14 May 2004; pp. 69–82.
8. Fang, Y.; Yamada, K.; Ninomiya, Y.; Horn, B.K.P.; Masaki, I. A shape-independent method for pedestrian detection with far-infrared images. *IEEE Trans. Veh. Technol.* **2004**, *53*, 1679–1697. [[CrossRef](#)]
9. Ragh, H.K.; Asari, V.K.; Aspiras, T.H. Human detection in infrared imagery using intensity distribution, gradient and texture features. In Proceedings of the Mobile Multimedia/Image Processing, Security, and Applications 2018, Orlando, FL, USA, 16–17 April 2018. [[CrossRef](#)]
10. Xu, F.; Liu, X.; FujiMura, K. Pedestrian detection and tracking with night vision. *IEEE Trans. Intell. Transp. Syst.* **2005**, *6*, 63–71. [[CrossRef](#)]
11. Zhang, L.; Wu, B.; Nevatia, R. Pedestrian detection in infrared images based on local shape features. In Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 17–22 June 2007; pp. 1–8. [[CrossRef](#)]
12. Setjo, C.H.; Achmad, B.; Djoko, F. Thermal image human detection using Haar-cascade classifier. In Proceedings of the 7th International Annual Engineering Seminar (InAES), Yogyakarta, Indonesia, 1–2 August 2017; pp. 1–6. [[CrossRef](#)]
13. Neves, A.; Ribeiro, R. Algorithms for face detection on infrared thermal images. *Int. J. Adv. Softw.* **2018**, *10*, 499–512.
14. Guennouni, S.; Ahaitouf, A.; Mansouri, A. A comparative study of multiple object detection using haar-like feature selection and local binary patterns in several platforms. *Model. Simul. Eng.* **2015**, *2015*, 8. [[CrossRef](#)]
15. Qu, H.; Zhang, L.; Wu, X.; He, X.; Hu, X.; Wen, X. Multiscale object detection in infrared streetscape images based on deep learning and instance level data augmentation. *Appl. Sci.* **2019**, *9*, 565. [[CrossRef](#)]
16. Yardimci, O.; Ayyıldız, B.Ç. Comparison of SVM and CNN classification methods for infrared target recognition. In Proceedings of the Automatic Target Recognition XXVIII, Orlando, FL, USA, 15–19 April 2018. [[CrossRef](#)]
17. Hasan, M.; Ullah, S.; Khan, M.J.; Khurshid, K. Comparative analysis of SVM, ANN and CNN for classifying vegetation species using hyperspectral thermal infrared data. In Proceedings of the International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences, Enschede, The Netherlands, 10–14 June 2019; pp. 1861–1868. [[CrossRef](#)]
18. Kim, J.H.; Hong, H.G.; Park, K.R. Convolutional neural network-based human detection in nighttime images using visible light camera sensors. *Sensors* **2017**, *17*, 1065. [[CrossRef](#)]
19. Herrmann, C.; Ruf, M.; Beyerer, J. CNN-based thermal infrared person detection by domain adaptation. In Proceedings of the Autonomous Systems: Sensors, Vehicles, Security and the Internet of Everything, Orlando, FL, USA, 15–19 April 2018. [[CrossRef](#)]

20. Khalid, B.; Khan, A.M.; Akram, M.U.; Batool, S. Person detection by fusion of visible and thermal images using convolutional neural network. In Proceedings of the 2019 2nd International Conference on Communication, Computing and Digital systems (C-CODE), Islamabad, Pakistan, 6–7 March 2019; pp. 143–148. [[CrossRef](#)]
21. Liu, J.; Zhang, S.; Wang, S.; Metaxas, D.N. Multispectral deep neural networks for pedestrian detection. *arXiv* **2016**, arXiv:1611.02644v1.
22. D’Acremont, A.; Fablet, R.; Baussard, A.; Quin, G. CNN-based target recognition and identification for infrared imaging in defense systems. *Sensors* **2019**, *19*, 2040. [[CrossRef](#)] [[PubMed](#)]
23. Hwang, S.; Park, J.; Kim, N.; Choi, Y.; Kweon, I.S. Multispectral pedestrian detection: Benchmark dataset and baseline. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1037–1045. [[CrossRef](#)]
24. González, A.; Fang, Z.; Socarras, Y.; Serrat, J.; Vázquez, D.; Xu, J.; López, A.M. Pedestrian detection at day/night time with visible and FIR cameras: A comparison. *Sensors* **2016**, *16*, 820. [[CrossRef](#)] [[PubMed](#)]
25. Khellal, A.; Ma, H.; Fei, Q. Pedestrian classification and detection in far infrared images. In Proceedings of the International Conference on Intelligent Robotics and Applications, Portsmouth, UK, 24–27 August 2015; pp. 511–522.
26. Lin, T.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Ramanan, D.; Zitnick, C.L.; Dollar, P. Microsoft COCO: Common objects in context. In Proceedings of the European conference on computer vision, Zurich, Switzerland, 6–12 September 2015.
27. Zivkovic, Z.; van der Heijden, F. Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern Recognit. Lett.* **2006**, *27*, 773–780. [[CrossRef](#)]
28. Wu, M.; Lin, C.; Chang, C. Brain tumor detection using color-based K-means clustering segmentation. In Proceedings of the Third International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP 2007), Kaohsiung, Taiwan, 26–28 November 2007; pp. 245–250. [[CrossRef](#)]
29. Dhanachandra, N.; Manglem, K.; Chanu, Y.J. Image segmentation using K-means clustering algorithm and subtractive clustering algorithm. *Procedia Comput. Sci.* **2015**, *54*, 764–771. [[CrossRef](#)]
30. Chen, J.; Fang, Y.; Cho, Y.K. Real-time 3D crane workspace update using a hybrid visualization approach. *J. Comput. Civ. Eng.* **2017**, *31*, 04017049. [[CrossRef](#)]
31. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [[CrossRef](#)]
32. Feng, C.; Liu, M.Y.; Kao, C.C.; Lee, T.Y. Deep active learning for civil infrastructure defect detection and classification. In Proceedings of the ASCE International Workshop on Computing in Civil Engineering 2017, Washington, DC, USA, 25–27 June 2017.
33. Cao, Q.D.; Choe, Y. Building damage annotation on post-hurricane satellite imagery based on convolutional neural networks. *arXiv* **2018**, arXiv:1807.01688.
34. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.

