applied sciences

MDPI

Article

# Pedestrian Detection at Night in Infrared Images Using an Attention-Guided Encoder-Decoder Convolutional Neural Network

**Yunfan Chen * and Hyunchul Shin ***

Division of Electrical Engineering, Hanyang University, Ansan 426-791, Korea

* Correspondence: chenyunfan@hanyang.ac.kr (Y.C.); shin@hanyang.ac.kr (H.S.); Tel.: +82-31-400-4083 (Y.C.); +82-31-400-5176 (H.S.)

Received: 9 December 2019; Accepted: 19 January 2020; Published: 23 January 2020


**Abstract:** Pedestrian-related accidents are much more likely to occur during nighttime when visible (VI) cameras are much less effective. Unlike VI cameras, infrared (IR) cameras can work in total darkness. However, IR images have several drawbacks, such as low-resolution, noise, and thermal energy characteristics that can differ depending on the weather. To overcome these drawbacks, we propose an IR camera system to identify pedestrians at night that uses a novel attention-guided encoder-decoder convolutional neural network (AED-CNN). In AED-CNN, encoder-decoder modules are introduced to generate multi-scale features, in which new skip connection blocks are incorporated into the decoder to combine the feature maps from the encoder and decoder module. This new architecture increases context information which is helpful for extracting discriminative features from low-resolution and noisy IR images. Furthermore, we propose an attention module to re-weight the multi-scale features generated by the encoder-decoder module. The attention mechanism effectively highlights pedestrians while eliminating background interference, which helps to detect pedestrians under various weather conditions. Empirical experiments on two challenging datasets fully demonstrate that our method shows superior performance. Our approach significantly improves the precision of the state-of-the-art method by 5.1% and 23.78% on the Keimyung University (KMU) and Computer Vision Center (CVC)-09 pedestrian dataset, respectively.

**Keywords:** infrared pedestrian detection; encoder-decoder; attention; convolutional neural network

## 1. Introduction

Pedestrian detection has attracted considerable attention from researchers in computer vision. Although many studies in pedestrian detection areas have been reported during the past decade [1–8], most of them are confined to detecting pedestrians during daytime using visible (VI) cameras. However, the performance of VI cameras depends on good illumination conditions and can be affected when illumination is poor. Recently, multispectral detectors that employ a fusion of infrared (IR) and VI cameras have been developed to achieve robust and reliable pedestrian detection in adverse illumination circumstances [9–13]. However, the multispectral detectors cannot work well at nighttime since the VI sensor only works when there is a substantial amount of visual information in the environment. Furthermore, the multispectral detectors only support fully aligned images. On a dark night, IR-based pedestrian detectors can effectively replace VI- and multispectral-based detectors, because the IR sensors do not require external light but mainly rely on the radiant temperature of the object. IR pedestrian detection has a wide range of applications, such as patrols, video surveillance, and rescues at night.

The main challenges for robust IR pedestrian detection can be classified into two main types. First, IR images have some adverse properties, such as their noisy nature, low-resolution, and no visual

detailed information. These adverse properties make the discriminative feature extraction of an object very difficult, and thus affect the detection performance. Second, IR images are susceptible to weather conditions since IR cameras detect the difference in the temperature of the environment. For instance, pedestrians look brighter than the background in cold weather, while their brightness looks similar to the background in hot weather.

Traditional IR pedestrian detection methods require manually designed features to describe IR objects, which are not conducive to extracting informative features from unstable IR images since IR sensors are susceptible to changeable weather conditions. In recent years, new technologies have been reported with very promising results. Deep learning has enabled progress on object detection using deep convolutional neural networks (CNNs) due to the capacity to generate semantic features via learning from raw pixels. This approach has superior discriminative ability to recognize targets with various shapes from a complex background [14–20]. Therefore, it is very natural to apply the CNN to IR pedestrian detection. A CNN-based method is an effective tool for IR object detection since it can handle variations on images affected by environmental changes, as long as these effects are widely present in the dataset. Therefore, our goal is to design an effective CNN framework specialized for IR pedestrian detection at night time that can achieve the best performance regardless of low-resolution or season changes.

In this research, a new attention-guided encoder-decoder convolutional neural network for accurate IR pedestrian detection at night, called AED-CNN, is developed. The AED-CNN mainly contains an encoder-decoder module and an attention module. The encoder-decoder structure efficiently captures the long-range context information while keeping the spatial information of IR objects, which is helpful for extracting discriminative features from low-resolution and noisy IR images. Based on the encoder-decoder modules, we further propose an attention module to overcome the variation problem due to changeable thermal energy in IR images under different weather conditions.

The following are the main contributions of this work.

- First, novel encoder-decoder modules are proposed to generate multi-scale features. We add an additional decoder module at the end of the single shot multibox detector (SSD) [16] architecture (encoder module) to form an encoder-decoder module, in which a new skip connection block is incorporated into each layer of the decoder to integrate the feature maps from the encoder and decoder modules. The proposed encoder-decoder modules effectively enrich the feature maps via integrating the high-level semantically strong features with low-resolution and low-level detailed features with high-resolution. This method is effective to extract discriminative features even from low-resolution and noisy IR images.
- Second, we propose an attention module that re-weights the multi-scale features generated from the encoder-decoder module. By adding the attention mechanism, the network selectively emphasizes useful information and suppresses ineffective information while re-weighting the features from the encoder-decoder modules. The attention module significantly eliminates background interference while highlighting pedestrians so that there is a boost in the detection performance of the IR pedestrian detector, even when the brightness is similar among the pedestrians and backgrounds.
- Finally, experimental results on two challenging datasets demonstrate that our AED-CNN shows the best performance. Our approach outperforms in detection precision by 5.1% and 23.78% on the Keimyung University (KMU) [21] (the KMU pedestrian detection database [21] is downloaded from: https://cvpr.kmu.ac.kr/ for academic use) and Computer Vision Center (CVC)-09 [22] (the CVC-09 far infrared (FIR) sequence pedestrian dataset [22] (available online, 28 April 2016) is download from: http://adas.cvc.uab.es/elektra/enigma-portfolio/item-1/) pedestrian datasets, respectively, when compared with the state-of-the-art oriented center-symmetric local binary (OCS LBP) + cascade random forest (CaRF) [21] method.

The remaining part of this paper is arranged in the following manners. Section 2 briefly introduces the previous related works. Section 3 explains the proposed AED-CNN in detail. Experimental

results and analysis are presented in Section 4. Finally, Section 5 summarizes our work and describes future works.

## 2. Background

### 2.1. Infrared (IR) Pedestrian Detection

In the past decades, IR pedestrian detection has attracted much interest from many researchers. Hotspot + SVM [23] acquiesced that a pedestrian's body looks brighter than the background. It generated pedestrian candidates by searching hotspot regions, then a support vector machine (SVM) was applied to conduct the detection stage. On the basis of hotspot features, Ko et al. [24] added analysis of face and shoulder parts in candidate hotspot regions and exploited a random forest classifier to classify the candidates. In [25], a dual-threshold segmentation method was developed to generate regions of interest (ROIs) via detecting hotspot regions. Then the Haar-like and histogram-of-oriented-gradients (HOG) features were extracted for classification. However, the aforementioned hotspot-based methods only show reasonable performance when the thermal energy difference between pedestrians and background is distinguishable. It generates considerable missing instances during hot weather or when the pedestrians wear well-insulated clothing. In [26], a stereo system formed by combining two IR cameras was introduced, in which hotspot detection, edge detection, and disparity calculation were adopted to produce candidate regions. Then, the morphology and thermal features of the pedestrian's head were used to validate candidate regions. In [27], a feature-based region growth with a high-intensity seed method is proposed for segmenting pedestrians, in which vertical deviation-based morphological closing is combined to compensate for distortion caused by clothing. Zhao et al. [28] first proposed a contour precision saliency map for detecting ROI. Then the shape distribution histogram feature was acquired by calculating the distances between the random points on the thinned contour map in the ROI. Finally, a modified sparse representation classification (MSRC) method was developed to detect IR pedestrians. OCS-LBP + CaRF [21] exploited the oriented center-symmetric local binary pattern (OCS-LBP) features to describe the IR pedestrians, and proposed a cascade random forest (CaRF) to classify pedestrians. Biswas et al. [29] proposed a local steering kernel (LSK) to reduce intrinsic noise in IR images and combined an image similarity kernel with SVM to train the LSK tensor. Recently, Heo et al. [30] applied a CNN, named 'you only look once version 2' (YOLOv2) [20] to IR pedestrian detection. They proposed a handcrafted adaptive Boolean-map-based saliency (ABMS) kernel to infuse with YOLOv2. Cao et al. [31] presented an automatic region proposal network by designing a new loss function and adding a segmentation task, for IR pedestrian detection. This method does not consider the factors that IR images are susceptible to environmental changes, and thus the results are not ideal. In the latest work [32], a CNN-based IR person detector based on residual network (ResNet) [33] and atrous spatial pyramid pooling (SPP-net) [34] was proposed. The developed IR person detector was evaluated by using infrared closed-circuit television (CCTV) images.

### 2.2. Convolutional Neural Network (CNN)-Based Object Detection

With the rapid growth of CNN technologies recently, a large variety of CNN-based approaches have facilitated object detection to a new stage. CNN-based object detectors are generally grouped into two main categoreis. The first categore is termed as two-stage approaches, including fast region-based CNN (Fast R-CNN) [14] and Faster R-CNN [15]. The two-stage methods include two parts. The first part aims to generate a sparse series of candidate object proposals, and the second part refines the candidate proposals to determine the accurate objects locations and its corresponding class label. Although the aforementioned two-stage detectors achieve desirable accuracy, the computational speed is slow. By contrast with two-stage detectors, one-stage methods such as SSD [16] and YOLO [20,35] avoid generating region proposals and resampling features by encapsulating all operations in a single network, which performs better in speed than the two-stage detectors. In CNN-based methods, the high-level layers which contain sufficient semantic features lack the detailed spatial features of the

objects because of the striding operations in pooling and convolutional layers. This makes it hard to extract discriminative features. Consequently, it is difficult to identify target locations. On the contrary, the features from the low-level layers which contain enough spatial information but lack the semantic information are not robust enough to the challenges of appearance variation and occlusions. To address these issues, some networks [18,36] try to observe and utilize the pyramidal features to a large extent by building an encoder-decoder architecture with lateral connections. These networks show dramatic improvements in accuracy when compared with conventional detectors. However, effective methods to combine the information from the encoder and decoder modules for good performance still need to be explored. Motivated by the aforementioned works, we introduce a novel encoder-decoder module that takes an IR image as an input to the encoder module to capture richer semantic information and adopts a decoder module to progressively recover the spatial information. In our proposed encoder-decoder module, a new skip-connection block is used to fuse the information from the encoder and decoder module. The encoder-decoder structure can efficiently capture the long-range context information while keeping the spatial information of IR objects. This owes to the expressiveness of the encoder-decoder layers, which are originated from the combinatorial nature of Rectified Linear Unit (ReLU) for decomposition and reconstruction.

### 2.3. Attention Mechanism

Attention mechanism is widely utilized in CNNs to solve different computer vision tasks, like video classification [37] and object detection [38]. Wang et al. [37] applied the attention mechanism to non-local filter operation, which is effective in performing classification tasks in videos through computing the feature relationships in different positions. DeepSaliency [38] exploited the attention mechanism to learn the spatial relationships between salient features, which shows promising results for object detection and segmentation tasks. The above works are to investigate to model spatial correlations. In contrast, another method, squeeze-and-excitation network (SENet) [39], is to model the correlative dependence between channels of CNN features to perform feature recalibration. It utilizes global information to learn layout or emphasis of the scene and to obtain the context information which guides visual processing to task-related image regions. Inspired by SENet [39], we find that we can recalibrate the features in the decoder aiming at guiding the detector to pay more attention to pedestrian locations. Specifically, we propose an attention module that adopts SE blocks to learn attention vectors from the feature maps in the encoder. Then the learned attention vectors are utilized to scale the feature maps from the decoder to enhance informative features and to suppress less useful features. The attention module indicates the probability that each pixel belongs to a pedestrian region. It has a high discriminative capability to recognize a human target, because background interference is reduced, and pedestrians are highlighted.

In this research, a novel attention-guided encoder-decoder convolutional neural network (AED-CNN) is developed to detect pedestrians in IR images. To better describe pedestrians in low-resolution and noisy IR images, the encoder-decoder module is adopted, in which a skip connection block is integrated to effectively combine the features from the encoder and decoder for increasing additional context information. Furthermore, the attention module is proposed to highlight pedestrians while eliminating background interference, which can help to detect pedestrians under different weather conditions.

## 3. Proposed Attention-Guided Encoder-Decoder Convolutional Neural Network (AED-CNN)

The proposed AED-CNN mainly contains an encoder-decoder module and an attention module. The encoder-decoder module takes an IR image as an input to the encoder module to capture rich semantic information and adopts a decoder module to progressively recover the spatial information. The attention module learns attention vectors from the feature maps in the encoder, which are used to scale the feature maps from the decoder. The attention module effectively enhances informative

features to highlight pedestrians and suppress less useful features to exclude background interference. In the following sections, we give detailed explanations.

### *3.1. Overview of the Proposed AED-CNN Architecture*

Figure 1 presents an overview of our AED-CNN architecture. The input image is processed by the encoder module. The encoder gradually reduces the resolution of the feature map for conducting a multi-scale search of bounding boxes, as in the SSD [16]. Then, the feature maps in the decoder are up-scaled via deconvolutional layers and then combined with the corresponding feature maps of the same resolution in the encoder through skip connection blocks. The multi-scale features generated by the encoder-decoder module are re-weighted through the attention net. Finally, these re-weighted features are sent to the prediction stage for classifying pedestrians and regressing bounding boxes. The architectures of skip connection block and the attention net are presented in Figures 2 and 3, respectively.

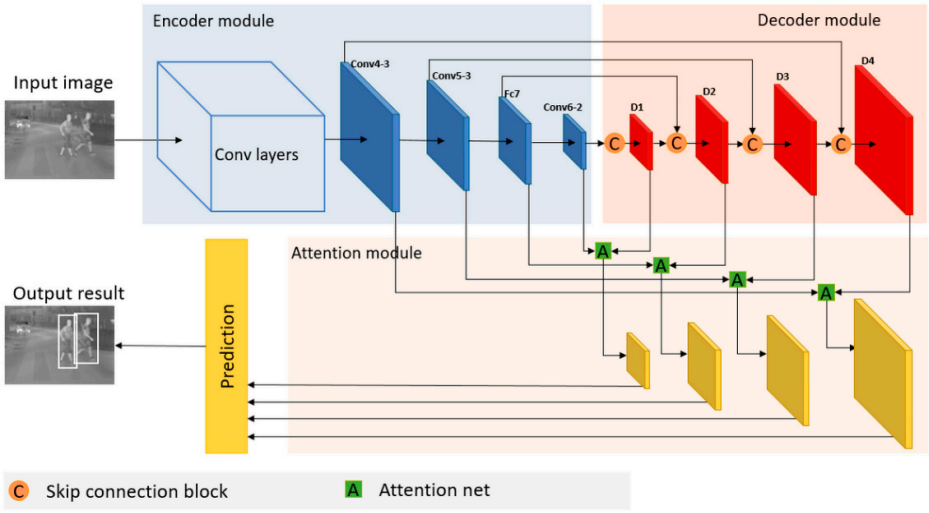



**Figure 1.** Architecture of the proposed attention-guided encoder-decoder convolutional neural network (AED-CNN).

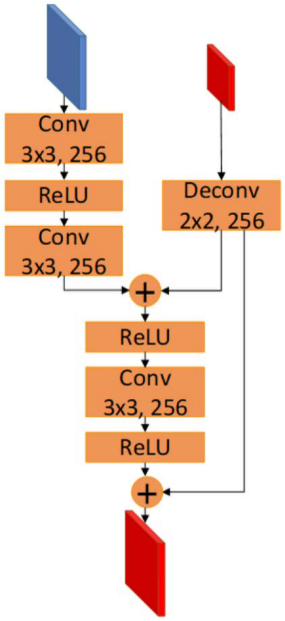



**Figure 2.** Skip connection block.

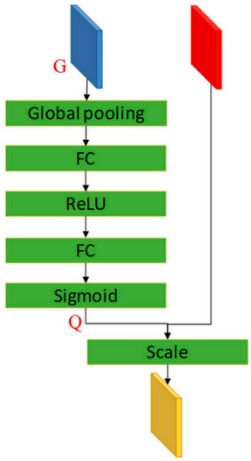



**Figure 3.** Attention net.

### *3.2. Encoder Module*

As shown in Figure 1, using a similar method as in SSD [16], the encoder module is built on top of a "base" network (a feedforward convolutional network that contains several convolution layers and pooling layers) which is followed by a series of progressively smaller convolution layers to constitute a cascade of feature maps with a gradually increasing field of view and decreasing spatial resolution. The VGG-16 [40] is chosen as our "base" network since it has good detection performance as well as fast detection speed. Following [19], two additional convolution layers (conv6-1 and conv6-2) are added after fully connected layer (Fc7) of the truncated VGG-16 to obtain more semantic information at a high-level and to detect pedestrians at multi-scales. However, it is hard for SSD to classify the pedestrians in IR images of low-resolution and with noisy characteristics, owing to the weak semantic information on the shallow features. Inspired by [18], we added a decoder module in the proposed architecture to improve the quality of IR images. The decoder module can increase more context information that enriches semantic features extracted from IR images. The decoder module is described in the next section.

### *3.3. Decoder Module*

In the encoder module, the low-level layers contain high spatial resolution but lack enough semantic features. By contrast, the high-level layers have rich semantic features but low spatial resolution. It is difficult to extract discriminative features from the feature layers of encoder module to detect pedestrians in low-resolution IR images. In [18], the deconvolutional network has been proposed to refine the traditional feedforward convolutional network, which shows that the deconvolutional module equalizes the representability of each feature map of a sophisticated network and makes the network more efficient. Therefore, to include more high-level context in detection, we add a decoder module at the end of the encoder to effectively make an encoder-decoder hourglass network structure. The decoder module consists of a series of deconvolution layers with successively increasing resolution of feature map layers, namely deconvolution layer 1 (D1) to deconvolution layer 4 (D4), as shown by the red blocks in Figure 1. In [18], five deconvolution layers are used from the top-most feature maps, in which all fine details are missing. Moreover, the additional deconvolution layers result in more computational cost, making it impractical for real-time applications because of the large inference time. To overcome this, we make our decoder shallower and only four layers are used to maintain a higher detection speed.

#### Skip Connection Block

As mentioned above, the encoder-decoder module is exploited to obtain feature maps with rich context information via combining the high-level semantic information with the low-level detailed

information. To this end, a new skip connection block, as shown in Figure 2, is introduced to fuse the feature maps from the decoder and the corresponding feature maps from the encoder. In order to share the structure of skip connection block, the deep layers are symmetrically connected with these shallow layers. In other words, the deep features have the same down sampling factor. Specifically, feature maps in the decoder are up sampled and then merged with the feature maps in the encoder. Figure 2 shows the proposed skip connection block. First, the feature map in the decoder is up sampled by a $2 \times 2$ deconvolution layer to match the size of the feature map in the encoder. The feature map in the encoder is followed by a $3 \times 3$ convolution layer, a ReLU layer, and a $3 \times 3$ convolution layer. We merge them by using the element-wise sum. Then, the summed layer passes through a block (ReLU, $3 \times 3$ convolution, ReLU), which is useful for discernable feature extraction as proven in [19], and then summed with the deconvolution layer, which skips the block. Note that the Conv6-2 directly goes through the first skip connected block to generate D1 without combination with the feature map in the encoder module. The skip connection is beneficial to back-propagation, which can effectively solve the problem of performance degradation of deep convolutional neural networks under extremely deep conditions, as shown in [33]. It speeds up the training of a deeper network and thus makes the learning easier. In addition, the skip connection helps traverse the information in deep neural networks to boost the classification performance.

### 3.4. Attention Module

The attention module plays an important role in AED-CNN to re-weight the multi-scale features. As shown in the lower part of Figure 1, the attention module contains four attention nets which take the feature maps of the same scale in the encoder and decoder as inputs, and then outputs four re-weighted feature maps in multi-scales for the final pedestrian prediction part. Under different weather conditions, the difference in thermal energy between pedestrian and background will change. In hot weather, pedestrians' brightness looks similar to the background, which makes it difficult for pedestrians to be distinguished. The success of SENet [39] shows the effectiveness of the attention mechanism on images classification. SENet proposed a channel-wise attention mechanism that helps to selectively enhance useful features and suppress ineffective features by performing feature recalibration. For IR pedestrian detection, we need to find the most effective features to highlight IR pedestrian regions. Therefore, we propose attention net to recalibrate the multi-scale features generated by the decoder module, in which large weights are assigned to channels that show a high response to IR objects. The key idea of the attention net is to learn the feature weights according to the loss, with the objective of maximizing the weights of effective feature maps and minimizing the weights of less effective feature maps. The attention mechanism filters out some background details according to re-weighted features and focuses more on the foreground regions, which helps to generate effective features for pedestrian detection.

Figure 3 shows the architecture of our attention net. The left side of Figure 3 takes the feature map in the encoder as the input guidance $G$. Then the guidance $G$ passes through one global average pooling layer and two fully connected (FC) layers (The two FC layers are followed by a ReLU layer and a sigmoid layer, respectively) to learn the mapping function $F$ used to generate the channel-wise attention vector Q:

$$\mathrm{Q} = F\left(G^T\right) \tag{1}$$

The generated attention vector Q is used to scale the layer of the decoder. We apply the same attention net to four groups of feature maps of the same scale from the encoder and decoder. Finally, four multi-scale re-weighted layers are generated, which will be used for the final pedestrian prediction task. Figure 4 shows a visualization of feature maps with and without the attention module. One can see that the pedestrian regions are highlighted while the backgrounds are suppressed, in the feature maps from the attention module.

The attention module specifically models the interdependencies between the convolutional channels that effectively boost the representational capability for various samples. After applying the attention module, more useful features are emphasized while less informative ones are restrained,

via re-weighting the sample-reliant features. In other words, the pedestrians are highlighted, while the background interferences are suppressed. This ensures the performance of the detector when the thermal energy of pedestrians is not distinguishable due to weather changes. The attention module is not only easy to implement but also realizes remarkable improvements at little extra computational cost.

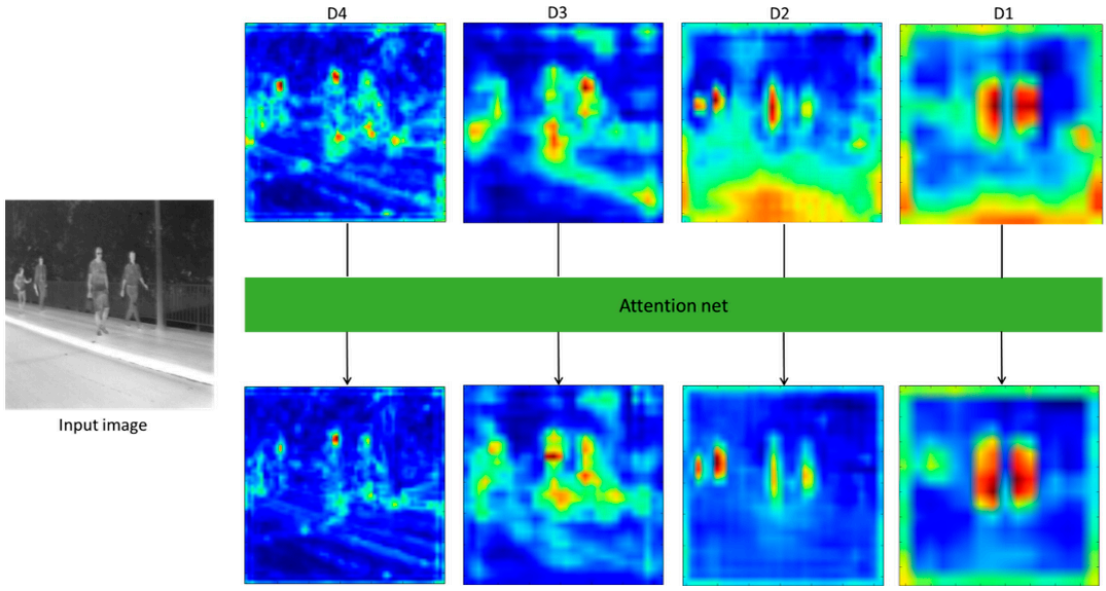



**Figure 4.** Visualization of feature maps from decoder module (**top**) and visualization of feature maps from the attention module (**bottom**).

### *3.5. Training*

#### 3.5.1. Matching and Hard Negative Mining

In the training stage, the anchor boxes are used to localize objects in multiple scales, which need to be matched with the ground truth bounding boxes. The correspondence between each ground truth bounding box A to the anchor box B is determined by the jaccard overlap [41]. The value of the jaccard overlap is defined as:

$$\mathrm{J}(\mathrm{A},\mathrm{B}) = \frac{\mathrm{area}(\mathrm{A} \cap \mathrm{B})}{\mathrm{area}(\mathrm{A} \cup \mathrm{B})} \tag{2}$$

Each ground truth bounding box is first matched to the anchor box with the best $\mathrm{J}(\mathrm{A},\mathrm{B})$. Then the remaining anchor boxes are matched to any ground truth with $\mathrm{J}(\mathrm{A},\mathrm{B}) > 0.5$. This strategy is beneficial to predict multiple bounding boxes with high scores for overlapped objects. After matching, the majority of samples are negative. Similar to SSD [16], we select the negative samples with the top loss values from the non-matched anchor boxes to set the ratio between positive and negative samples as 1:3.

#### 3.5.2. Loss Function

Equation (3) shows our overall loss function, which is a weighted summation of the two branches, one is the confidence loss (*conf*) of the softmax classifier, and the other is the localization loss (*loc*) of the bounding box regression.

$$L(\{p_i\},\{t_i\}) = \frac{1}{N}\left(\left(L_{conf}\left(p_i, p_i^*\right)\right) + \lambda L_{loc}\left(t_i, t_i^*\right)\right) \tag{3}$$

where $p_i^*$ is the ground truth label of an anchor $i$ in a mini-batch, and the value of $p_i$ is the probability of an anchor $i$ being a pedestrian. The ground truth location of the anchor $i$ is denoted by $t_i^*$ and

the predicted bounding box location of the anchor $i$ is represented by $t_i^*$. $N$ denotes the number of positive anchors. Notably, if $N = 0$, the overall loss $L = 0$. Currently, we set weight term $\lambda$ to 1 via cross-validation. The classification loss $L_{conf}$ is the cross-entropy loss over two classes (pedestrian vs. non-pedestrian). As in Fast R-CNN [14], the smooth $L_1$ loss is adopted as our regression loss $L_{loc}$.

$$L_{loc}\left(t_i, t_i^*\right) = \sum\nolimits_{j\in\{x,y,w,h\}} \text{smooth}_{L_1}\left(t_j, t_j^*\right) \tag{4}$$

where $\text{smooth}_{L_1}$ is defined as:

$$\text{smooth}_{L_1}(x) = \begin{cases} 0.5x^2 & if\ |x| < 1 \\ |x| - 0.5 & otherwise, \end{cases} \tag{5}$$

#### 3.5.3. Optimization

The "base" network VGG-16 in our AED-CNN is pretrained on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) dataset [42]. The Xavier approach [43] is applied to initialize the parameters of two additionally convolutional layers (conv6-1 and conv6-2). In training, the default batch size is set to 11. Then, the entire network is fine-tuned by using a stochastic gradient descent method with the momentum 0.9 and weight decay 0.0005. To prevent gradient explosion in early iterations, we first run 10 k (where k = 1000) iterations using a learning rate of 0.00005. Then, we reset the learning rate to 0.001 for the next 70 k iterations. After the completion of 70 k iterations, the learning rate is reduced by 10 times after every 20 k iterations. Learning stops after 120 k iterations.

## 4. Experimental Results

### *4.1. Datasets and Processing Platform*

Even though there are extensive color video/image datasets for pedestrian detection, only a small number of thermal infrared pedestrian datasets are available. In this study, we evaluate our proposed AED-CNN on the widely used KMU [21] and CVC-09 [22] pedestrian datasets.

The KMU was captured by a far infrared (FIR) camera from moving vehicles (at 20 to 30 km/h) during the summer and winter nights for pedestrian detection. The training data contains 4474 positive images and 3405 negative images. The positive images include pedestrians with various sizes and postures. The negative images were generated through random cropping from the background. The testing data contains 5045 images of the same size of 640 × 480, including pedestrians with various activities under different weather conditions, such as walking down the sidewalk or crossing the road in hot summer as well as in cold winter.

The CVC-09 FIR sequence pedestrian dataset was collected using a FIR camera mounted on a car roof during summer days. This dataset is more challenging since it comprises pedestrians with varying moving speeds, a variety of motions, all types of poses, changeable thermal energy, and partial or full occlusions at night. The CVC-09 contains 5309 positive images and 2115 negative images for training, 5763 images for testing, with the same size of 640 × 480.

For both KMU and CVC-09 datasets, the pedestrian images of different scales are combined for training. The experiments are performed on a standard computer under Ubuntu 14.04 with a core i7-6850k 3.6 GHz central processing unit (CPU) and 64 GB random-access memory. For the graphics processing unit (GPU), we used three NVIDIA Titan X for training and a single Nvidia Titan X for testing. Our codes are built on Caffe which requires a compute unified device architecture (CUDA) and CUDA deep neural network library (cuDNN).

### *4.2. Evaluation Metric*

We validate the detection performance by employing precision-recall curves, which are generally applied to the evaluation of human detection performance. To verify the detection results, the predicted

bounding boxes generated by the detector are compared with the ground-truth bounding boxes and checked as either true positive (TP), false positive (FP), or false negative (FN), via measuring the overlap between the bounding boxes. TP indicates the quantity of properly detected pedestrians, FP indicates the number of mistakenly detected pedestrians by the detector, and FN indicates the number of pedestrians that are not detected by the detector. As the detection criteria, the detected bounding box is considered to be TP if the overlap ratio between the detected bounding box and the ground-truth bounding box exceeds 50%. TP/(TP + FP) calculates the precision and TP/(TP + FN) computes the recall. The average precision (AP) describes the tendency of the precision-recall curve and is calculated by averaging the precision at several evenly spaced recall levels via changing the threshold of the detection scores. In our experiments, the AP is obtained by averaging the precision values at 11 evenly spaced recall levels between 0 and 1.

### *4.3. Comparison of the Detection Performance on the Keimyung University (KMU) Dataset*

We compare the performance of AED-CNN with a set of five well-known methods, including histogram-of-oriented-gradients (HOG) + support vector machine (SVM) [44], Hotspot + SVM [23], OCS-LBP + CaRF [21], YOLOv2 [20], and YOLOv2 + ABMS [30].

As shown in Figure 5, our AED-CNN clearly shows the best performance when compared to other methods and achieves the highest AP of 97.5%, which significantly exceeds the two recent optimal results, YOLOv2 + ABMS [30] by 17%, and OCS-LBP + CaRF [21] by 5.1%, respectively. HOG + SVM [44] shows the worst results because the HOG feature is not suitable to characterize pedestrians in IR images. Although the hotspot regions of pedestrians can be detected in IR images, Hotspot + SVM [23] is still ineffective when the temperature of pedestrians is similar to the background, for example during a summer night. With the same limitations as Hotspot + SVM, OCS-LBP + RF [21] and YOLOv2 [20] showed worse performance than ours. ABMS is proposed by YOLOv2 + ABMS [30] to pre-process the IR images for enhancing pedestrians during hot weather. However, the YOLOv2 network cannot detect small-sized pedestrians in low-resolution IR images. These comparative results prove that our AED-CNN shows superior robustness compared to other methods under various circumstances.

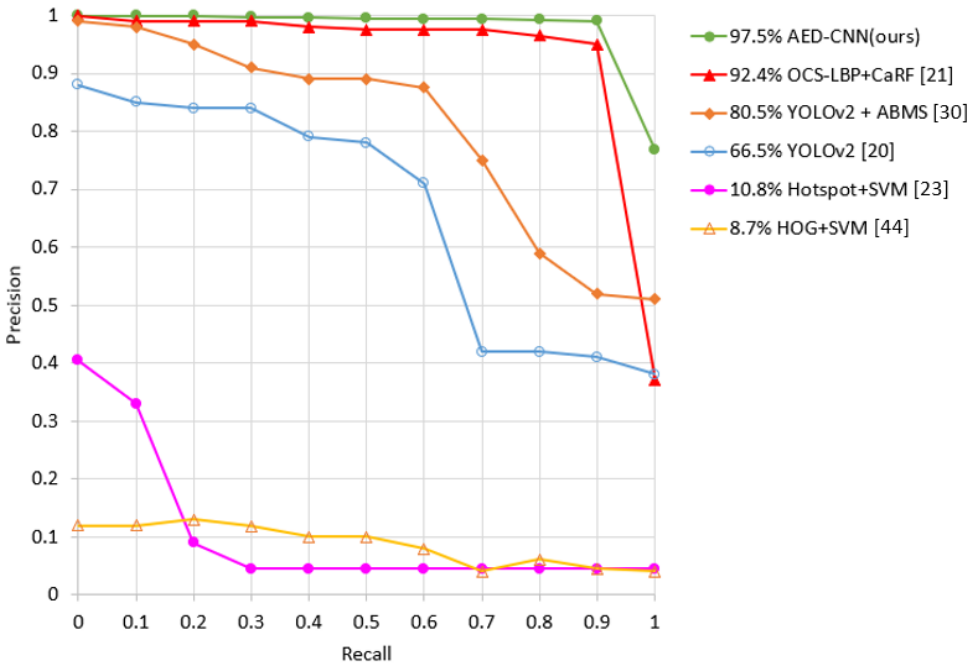



**Figure 5.** Performance comparison of precision versus recall using the Keimyung University (KMU) dataset.

Figure 6 presents some detection results on the KMU test set using our method. It is clear that our method successfully detects pedestrians with various scale instances during a hot summer night. From this observation, we can conclude that our method is effective at detecting pedestrians in IR images and is robust under different weather conditions.

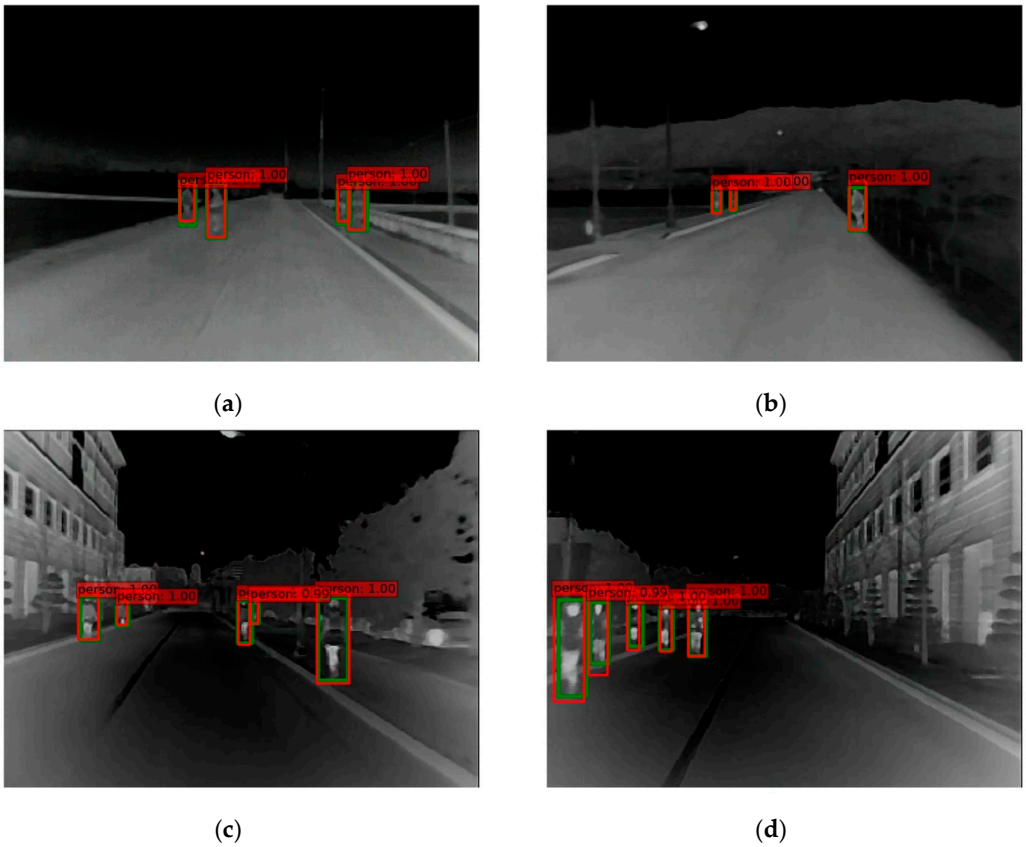



**Figure 6.** Some detection results on the KMU test dataset. The green bounding boxes denote the ground truth, and the red bounding boxes show the detection results. (**a–d**) are visualization of detection results on the image set selected from the dataset.

### *4.4. Comparison of the Detection Performance on the CVC-09 Dataset*

The detection performance is also evaluated by using the CVC-09 dataset. The precision-recall curves of the proposed AED-CNN and other well-known approaches are shown in Figure 7. A similar tendency can be observed in the results using the KMU dataset: the AP gap is quite large, 87.68% of ours versus 63.9% of the state-of-the-art OCS-LBP + CaRF [21]. Figure 7 reveals that our approach significantly outperforms all other approaches. The results are meaningful because the CVC09 dataset includes numerous pedestrian images with thermal energy levels similar to the background levels. Furthermore, the resolution is low, and some images are occluded. The performance comparison in Figure 7 demonstrates that our AED-CNN is obviously better than other state-of-the-art methods in detecting pedestrians.

Example detection results of our method on the CVC09 test set are displayed in Figure 8. In Figure 8a,b, AED-CNN can detect all the pedestrians. Figure 8c shows that a heavily occluded pedestrian was missed. In Figure 8d, three heavily occluded pedestrians were missed. From these observations, we think that it is necessary to improve the performance of AED-CNN to detect pedestrians with heavy occlusions in the future, even though AED-CNN shows the best performance when compared with other state-of-the-art methods.

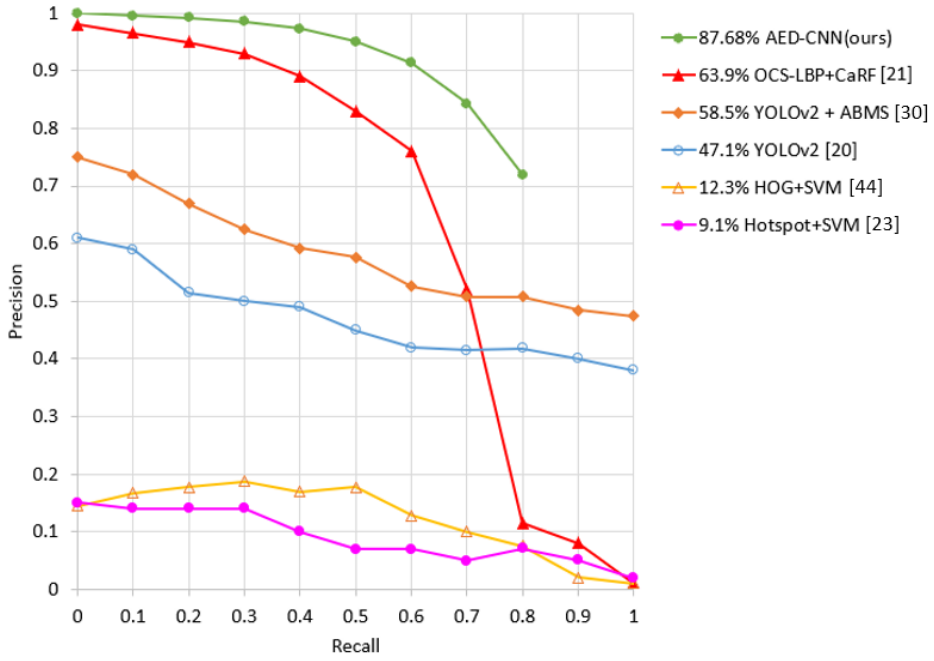
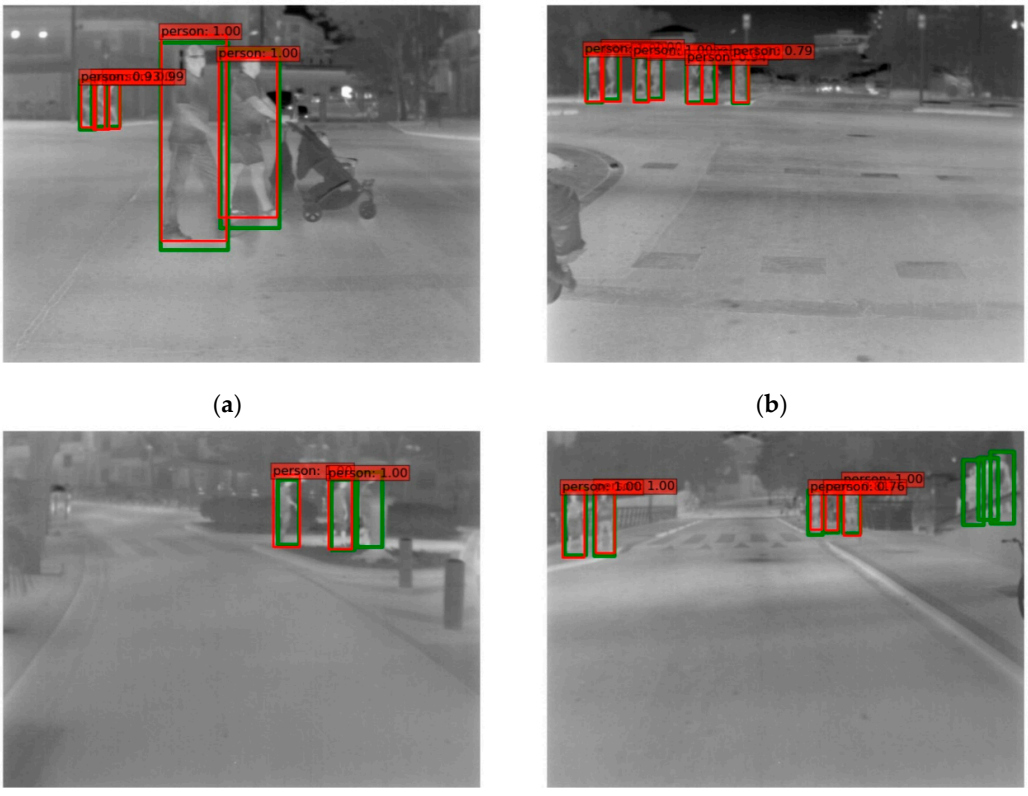



**Figure 7.** Performance comparison of precision versus recall by using the Computer Vision Center (CVC)09 dataset.




(**a**) (**b**)

(**c**) (**d**)

**Figure 8.** Four detection results of the CVC09 test dataset. The green bounding boxes denote the ground truth, and the red bounding boxes show the detection results. (**a–d**) are visualization of detection results on the image set selected from the dataset.

### *4.5. Comparison of the Computational Speed*

The computational speed and AP of the proposed AED-CNN with other reported methods are compared in Table 1. We measure the running times of all the methods using the same machine. Although YOLOv2 [20] and YOLOv2 + ABMS [30] have the fastest computational speed (0.02 s/f), their precision is worse than our approach (over 10%). The computational speed of our method is 0.03s per image, which is very competitive compared to other approaches.

**Table 1.** Comparison of the computation times and average precisions (APs) for the Keimyung University (KMU) and Computer Vision Center (CVC)09 test sets.

| Methods | KMU Test Set | | CVC09 Test Set | |
|---|---|---|---|---|
| | Speed (s/f) | Precision | Speed (s/f) | Precision |
| Histogram-of-oriented-gradients (HOG) + Support vector machine (SVM) [44] | 0.09 | 8.7% | 0.09 | 12.3% |
| Hotspot + SVM [23] | 0.08 | 10.8% | 0.08 | 9.1% |
| Oriented center-symmetric local binary (OCS-LBP) + Cascade random forest (CaRF) [21] | 0.06 | 92.4% | 0.06 | 63.9% |
| You only look once version 2 (YOLOv2) [20] | 0.02 | 66.5% | 0.02 | 47.1% |
| YOLOv2 + Adaptive Boolean-map-based saliency (ABMS) [30] | 0.02 | 80.5% | 0.02 | 58.5% |
| Attention-guided encoder-decoder convolutional neural network (AED-CNN) (ours) | 0.03 | 97.5% | 0.03 | 87.68% |

### *4.6. Ablation Experiments*

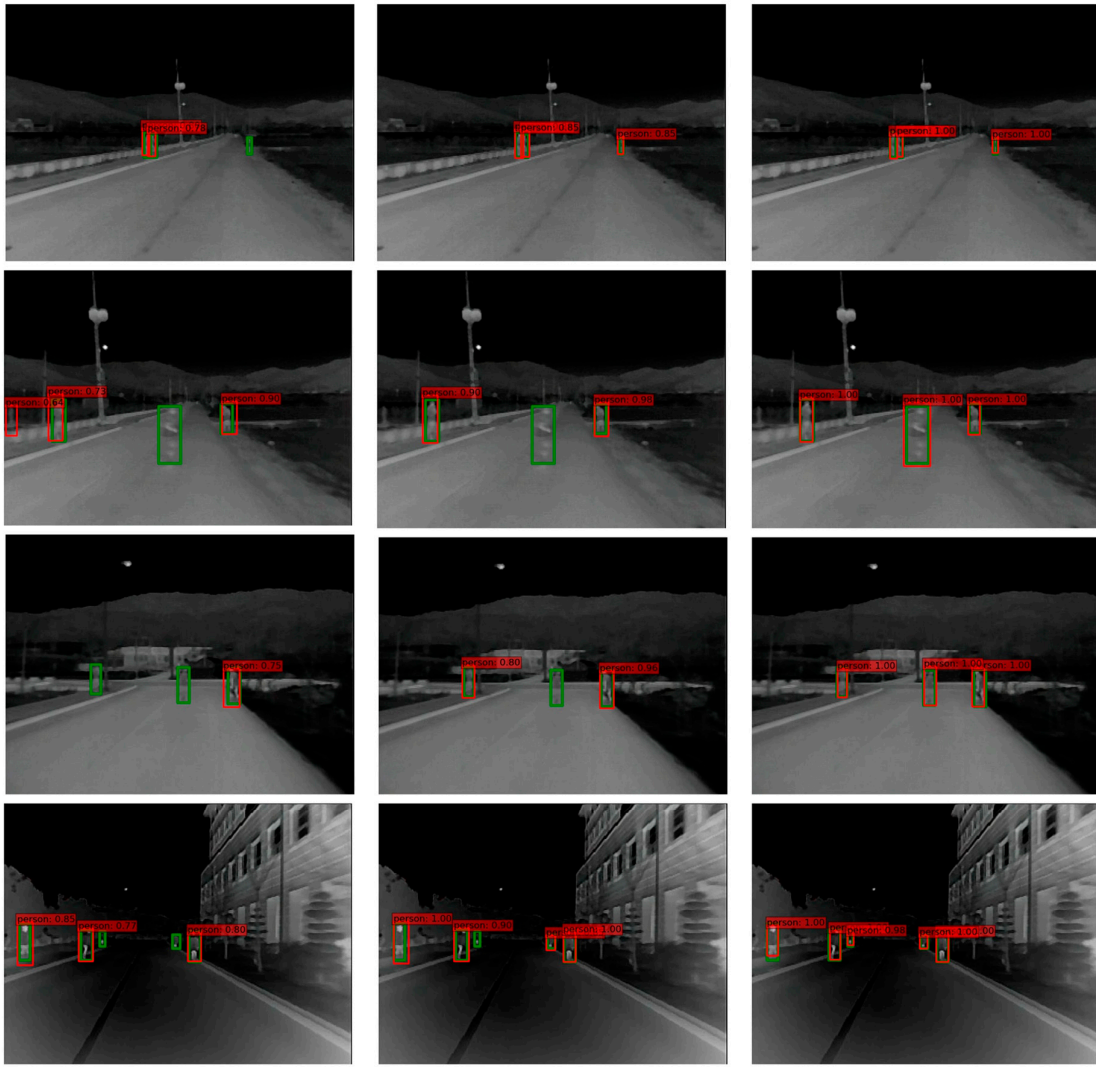
The ablation experiments are typically used to remove the main proposed components of the developed method gradually to evaluate the effectiveness of each component on detection performance. To verify the effectiveness of the encoder-decoder module and attention module, ablation experiments have been done using the KMU dataset. The initial simulation is the original SSD [16], which is the baseline detector. Then the decoder module is added to make an improved version, which is the encoder-decoder module. Finally, the attention module is added to form our AED-CNN. The average precision value of each module is given in Table 2. The baseline detector SSD [16] achieved 92.04% AP, while the encoder-decoder module and attention module have significantly improved the detection accuracies, of 95.36% and 97.50%, respectively.

**Table 2.** Results of the ablation experiments on KMU test set.

| Methods | Average Precision on KMU Test Set |
|---|---|
| Single shot multibox detector (SSD) [16] | 92.04% |
| Encoder-decoder module | 95.36% |
| Encoder-decoder module + attention module (AED-CNN) | 97.50% |

#### 4.6.1. Evaluation of the Encoder-Decoder Module

As shown in Table 2, the encoder-decoder module significantly outperforms the original SSD [16], by improving the AP by 3.32%. The SSD [16] built the pyramid starting from high up in the network without using low-level feature maps. However, the low-level feature maps contain high spatial resolutions and rich detailed features, which is essential for pedestrian detection in low-resolution IR images. Our method proposed a new decoder module to effectively combine the high-resolution low-level detailed features with the low-resolution high-level semantic features, which can increase context information that helps to boost the pedestrian detection rate in low-resolution IR images. Figure 9a,b illustrate visual comparisons of the original SSD detection results versus the encoder-decoder detection results. It is clear that the SSD produces more false positives and false negatives.



(**a**) (**b**) (**c**)

**Figure 9.** Visual comparisons of detection results on four example images among SSD, encoder-decoder module, and encoder-decoder + attention (AED-CNN). The green bounding boxes denote the ground truth. The red bounding boxes show the detection results. (**a**) SSD [16], (**b**) encoder-decoder module (**c**) encoder-decoder + attention (AED-CNN).

#### 4.6.2. Evaluation of the Attention Module

By incorporating our proposed attention module with the encoder-decoder module, the detection AP is further improved by 2.14%, from 95.36% to 97.05%, as shown in Table 2. This comparison demonstrates that the proposed attention module is useful to detect pedestrians in IR images with changeable thermal energy by highlighting pedestrian regions while reducing background interference. The visual comparison of the encoder-decoder detection results and those of the AED-CNN detection is provided in Figure 9b,c. One can see that the encoder-decoder module without the attention module fails to detect pedestrians, when the brightness values of pedestrians are close to those of the background. However, the AED-CNN is able to successfully detect all pedestrians in these examples even when the thermal energy difference between pedestrians and background is indistinguishable.

### *4.7. Discussion*

The objective of this research is to solve two main challenges of IR pedestrian detection. One is that IR images have some adverse properties, such as a noisy nature and low-resolution. The

other is that IR images are susceptible to weather. To overcome these two problems, we propose an attention-guided encoder-decoder convolutional neural network. Motivated by the recent success of CNNs in object detection, we design a novel encoder-decoder module based on the well-known SSD approach [16]. The proposed encoder-decoder module learns effective features from low-resolution IR images. Inspired by the attention mechanism applied in CNN [39], we further propose an attention module to highlight IR pedestrians even when thermal energy is indistinguishable between pedestrians and the background.

#### 4.7.1. Explanation of Results

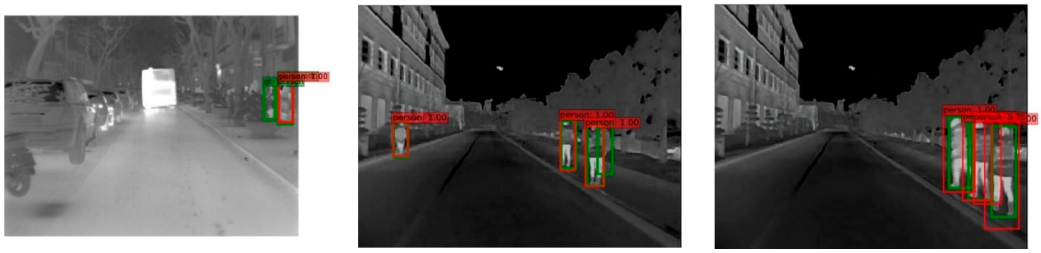
Empirical experiments described in Section 4.6 well validate the effectiveness of the proposed encoder-decoder and attention modules. Each of these modules in AED-CNN can bring a significant improvement in the detection accuracy. In Sections 4.3 and 4.4, we compare our method with several published state-of-the-art methods, on two challenging datasets. From the experimental results, we found our proposed CNN-based architecture significantly outperforms hand-crafted methods. The proposed method outperforms the hand-crafted method by 5.1% and 23.78% in AP on the KMU and CVC-09 pedestrian dataset, respectively. This result inspires us to explore in depth how to develop more effective CNN architectures to further improve the performance of IR pedestrian detection. As far as I know, there are only few research works which apply CNN to IR pedestrian detection.

Another research direction we can further explore is how to improve the confidence score of detected pedestrians in adverse environments. One can compare the result visualization of Figures 6 and 8, to easily find that the confidence score in Figure 8 is lower. The confidence score of some detected pedestrians in Figure 8b,d are about 0.7. The reason is that the difference in brightness between pedestrians and the background in the CVC-09 data is lower than the KMU data. It is necessary to explore how to further emphasize pedestrian regions. Some recent works apply CNN for saliency detection in images [45], which shows promising results. Saliency detection aims to learn the significant difference between different groups. Saliency detection helps to reduce the complicity of the background and detect the foreground objects. For the task of IR pedestrian detection, the pedestrian and background are two groups, we can apply saliency detection algorithms to learn specific features to find pixels that belong to saliency regions (pedestrian regions) [45]. Therefore, how to incorporate saliency detection into IR pedestrian detection can be a future research problem.

#### 4.7.2. Limitation of the Proposed Method

Despite having achieved state-of-the-art detection accuracy and real-time detection speed, our approach does not perform well in some cases, as shown in Figure 10. The failure cases are caused by occlusions. In Figure 10a, one person occluded by a tree is not detected. In Figure 10b, one person in crowed is not detected. In Figure 10c, a false detection is generated among the crowd. Thus, occlusions may degrade detection performance, which is a common limitation with any methods.

It is necessary to solve the challenges of detecting occluded pedestrians. This common problem has already been explored [46,47]. We can regard pedestrians as a combination of different body parts. For each body part, CNN learns features and generate a score, respectively. If the body part is occluded, the score will be lower, otherwise, the score will be higher. During training, the features of each body part will be integrated. The integrated feature will be used for pedestrian classification and localization. The issues to solve in future are how many parts of a body should be used and how to effectively integrate the features of each body part, with low resolution IR images. In addition, we can take advantage of supervised learning to optimize the occlusion problem from the loss function perspective. We can consider designing the loss function to force proposals away from the second largest ground-truth bounding boxes that overlap with it. The objective is to make the proposal close to real objects and away from the false objects, thereby reducing the false detection rate due to occlusions.



(**a**) (**b**) (**c**)

**Figure 10.** Failure cases of our proposed infrared (IR) pedestrian detection method. The green bounding boxes denote the ground truth. The red bounding boxes show the detection results. (**a–c**) are visualization of failure cases on the image set selected from the KMU and CVC09 datasets.

## 5. Conclusions

In this paper, an effective attention-guided encoder-decoder convolutional neural network (AED-CNN) is developed for pedestrian detection at night using IR images with various road scenes and weather conditions. The encoder-decoder module is used for generating multi-scale feature maps, in which the skip-connection block is proposed to combine the features from the encoder and decoder layers. This combination increases context information that is helpful to extract discriminative features even in low-resolution and noisy IR images. Furthermore, an attention module is devised to highlight pedestrians while eliminating background interference, which is effective to detect pedestrians even when the differences in thermal energy between pedestrians and background is poorly distinguishable.

Experimental results on the two widely used datasets fully verify that our method significantly outperforms other state-of-the-art approaches. The proposed method achieves 97.50% AP on the well-known KMU pedestrian dataset, which is 5.1% better in detection accuracy and two times faster in detection speed than the published state-of-the-art method. Furthermore, our method achieves an AP of 87.68% which is a 23.78% improvement over the previous best performance on the challenging CVC-09 pedestrian dataset. Moreover, our AED-CNN achieves a real-time detection speed of about 0.03 s/f. In future works, solutions to detect occluded IR pedestrians will be explored.

**Author Contributions:** Y.C. developed the idea, implemented the experiments, and wrote the manuscript. H.S. supervised the research and performed revisions and improvements. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by National Research Foundation of Korea (2017-R1D1A1B04-031040).

**Acknowledgments:** This work was supported by Basic Research Project in Science and Engineering through the Ministry of Education of the Republic of Korea and National Research Foundation of Korea (National Research Foundation of Korea 2017-R1D1A1B04-031040).

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Dollár, P.; Appel, R.; Belongie, S.; Perona, P. Fast feature pyramids for object detection. *IEEE Trans. Pattern Anal.* **2014**, *36*, 1532–1545. [CrossRef]
2. Tian, Y.; Luo, P.; Wang, X.; Tang, X. Deep learning strong parts for pedestrian detection. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1904–1912.
3. Cai, Z.; Fan, Q.; Feris, R.S.; Vasconcelos, N. A unified multi-scale deep convolutional neural network for fast object detection. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 354–370.
4. Xiao, F.; Liu, B.; Li, R. Pedestrian object detection with fusion of visual attention mechanism and semantic computation. *Multimed. Tools Appl.* **2019**, 1–15. [CrossRef]
5. Brazil, G.; Yin, X.; Liu, X. Illuminating pedestrians via simultaneous detection & segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4950–4959.

6. Guo, Z.; Liao, W.; Xiao, Y.; Veelaert, P.; Philips, W. An occlusion-robust feature selection framework in pedestrian detection. *Sensors* **2018**, *18*, 2272. [CrossRef]
7. Li, J.; Liang, X.; Shen, S.; Xu, T.; Feng, J.; Yan, S. Scale-aware fast r-cnn for pedestrian detection. *IEEE Trans. Multimed.* **2017**, *20*, 985–996. [CrossRef]
8. Gu, J.; Lan, C.; Chen, W.; Han, H. Joint pedestrian and body part detection via semantic relationship learning. *Appl. Sci.* **2019**, *9*, 752. [CrossRef]
9. Hwang, S.; Park, J.; Kim, N.; Choi, Y.; So Kweon, I. Multispectral pedestrian detection: Benchmark dataset and baseline. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1037–1045.
10. Liu, J.; Zhang, S.; Wang, S.; Metaxas, D.N. Multispectral deep neural networks for pedestrian detection. In Proceedings of the British Machine Vision Conference, York, UK, 19–22 September 2016; pp. 1–13.
11. König, D.; Adam, M.; Jarvers, C.; Layher, G.; Neumann, H.; Teutsch, M. Fully convolutional region proposal networks for multispectral person detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 243–250.
12. Chen, Y.; Xie, H.; Shin, H. Multi-layer fusion techniques using a CNN for multispectral pedestrian detection. *IET Comput. Vis.* **2018**, *12*, 1179–1187. [CrossRef]
13. Li, C.; Song, D.; Tong, R.; Tang, M. Illumination-aware faster r-cnn for robust multispectral pedestrian detection. *Pattern Recognit.* **2019**, *85*, 161–171. [CrossRef]
14. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
15. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 91–99.
16. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Cham, Switzerland, 2016; pp. 21–37.
17. Dai, J.; Li, Y.; He, K.; Sun, J. R-fcn: Object detection via region-based fully convolutional networks. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 379–387.
18. Fu, C.Y.; Liu, W.; Ranga, A.; Tyagi, A.; Berg, A.C. Dssd: Deconvolutional single shot detector. *arXiv* **2017**, arXiv:1701.06659.
19. Zhang, S.; Wen, L.; Bian, X.; Lei, Z.; Li, S.Z. Single-shot refinement neural network for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4203–4212.
20. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
21. Jeong, M.; Ko, B.C.; Nam, J.Y. Early detection of sudden pedestrian crossing for safe driving during summer nights. *IEEE Trans. Circ. Syst. Video* **2016**, *27*, 1368–1380. Available online: https://cvpr.kmu.ac.kr/ (accessed on 28 April 2016). [CrossRef]
22. CVC-09 FIR Sequence Pedestrian Dataset. Available online: http://adas.cvc.uab.es/elektra/enigma-portfolio/item-1/ (accessed on 28 April 2016).
23. Xu, F.; Liu, X.; Fujimura, K. Pedestrian detection and tracking with night vision. *IEEE Trans. Intell. Trans. Syst.* **2005**, *6*, 63–71. [CrossRef]
24. Ko, B.; Kim, D.; Nam, J. Detecting humans using luminance saliency in thermal images. *Opt. Lett.* **2012**, *37*, 4350–4352. [CrossRef] [PubMed]
25. Ge, J.; Luo, Y.; Tei, G. Real-time pedestrian detection and tracking at nighttime for driver-assistance systems. *IEEE Trans. Intell. Trans. Syst.* **2009**, *10*, 283–298.
26. Bertozzi, M.; Broggi, A.; Caraffi, C.; Del Rose, M.; Felisa, M.; Vezzoni, G. Pedestrian detection by means of far-infrared stereo vision. *Comput. Vis. Image Underst.* **2007**, *106*, 194–204. [CrossRef]
27. O'Malley, R.; Jones, E.; Glavin, M. Detection of pedestrians in far-infrared automotive night vision using region-growing and clothing distortion compensation. *Infrared Phys. Technol.* **2010**, *53*, 439–449. [CrossRef]

28. Zhao, X.; He, Z.; Zhang, S.; Liang, D. Robust pedestrian detection in thermal infrared imagery using a shape distribution histogram feature and modified sparse representation classification. *Pattern Recognit.* **2015**, *48*, 1947–1960. [CrossRef]
29. Biswas, S.K.; Milanfar, P. Linear support tensor machine with LSK channels: Pedestrian detection in thermal infrared images. *IEEE Trans. Image Process.* **2017**, *26*, 4229–4242. [CrossRef]
30. Heo, D.; Lee, E.; Ko, B.C. Pedestrian detection at night using deep neural networks and saliency maps. *Electron. Imaging* **2018**, *17*, 1–9. [CrossRef]
31. Cao, Z.; Yang, H.; Zhao, J.; Pan, X.; Zhang, L.; Liu, Z. A new region proposal network for far-infrared pedestrian detection. *IEEE Access* **2019**, *7*, 135023–135030. [CrossRef]
32. Park, J.; Chen, J.; Cho, Y.K.; Kang, D.Y.; Son, B.J. CNN-based person detection using infrared images for night-time intrusion warning systems. *Sensors* **2020**, *20*, 34. [CrossRef]
33. 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.
34. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal.* **2015**, *37*, 1904–1916. [CrossRef]
35. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767(1804).
36. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
37. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803.
38. Li, X.; Zhao, L.; Wei, L.; Yang, M.H.; Wu, F.; Zhuang, Y.; Ling, H.; Wang, J. Deepsaliency: Multi-task deep neural network model for salient object detection. *IEEE Trans. Image Process.* **2016**, *25*, 3919–3930. [CrossRef]
39. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
40. Simonyan, K.; Zisserman, A. Very deep convolutional networks for largescale image recognition. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015; pp. 1–14.
41. Erhan, D.; Szegedy, C.; Toshev, A.; Anguelov, D. Scalable object detection using deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 2147–2154.
42. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
43. Glorot, X.; Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, Chia Laguna, Sardinia, Italy, 13–15 May 2010; pp. 249–256.
44. Xu, Y.; Xu, D.; Lin, S.; Han, T.X.; Cao, X.; Li, X. Detection of sudden pedestrian crossings for driving assistance systems. *IEEE Trans. Syst. Man Cybern. B* **2012**, *42*, 729–739.
45. Wang, W.; Lai, Q.; Fu, H.; Shen, J.; Ling, H. Salient object detection in the deep learning era: An in-depth survey. *arXiv* **2019**, arXiv:1904.09146.
46. Zhang, S.; Wen, L.; Bian, X.; Lei, Z.; Li, S.Z. Occlusion-aware r-cnn: Detecting pedestrians in a crowd. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 637–653.
47. Wang, X.; Xiao, T.; Jiang, Y.; Shao, S.; Sun, J.; Shen, C. Repulsion loss: Detecting pedestrians in a crowd. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7774–7783.