remote sensing

MDPI

Article

Infrared and Visible Image Object Detection via Focused Feature Enhancement and Cascaded Semantic Extension

Xiaowu Xiao [1,*], Bo Wang [1], Lingjuan Miao [1], Linhao Li [1], Zhiqiang Zhou [1], Jinlei Ma [2] and Dandan Dong [3]

[1] School of Automation, Beijing Institute of Technology, Beijing 100081, China; wangbo@bit.edu.cn (B.W.); miaolingjuan@bit.edu.cn (L.M.); lilinhao@bit.edu.cn (L.L.); zhzhzhou@bit.edu.cn (Z.Z.)
[2] China Helicopter Research and Development Institute, Tianjin 300300, China; majl027@avic.com
[3] College of Petroleum, China University of Petroleum, Karamay 834000, China; ddd@cupk.edu.cn
* Correspondence: 3120170438@bit.edu.cn

Abstract: Infrared and visible images (multi-sensor or multi-band images) have many complementary features which can effectively boost the performance of object detection. Recently, convolutional neural networks (CNNs) have seen frequent use to perform object detection in multi-band images. However, it is very difficult for CNNs to extract complementary features from infrared and visible images. In order to solve this problem, a difference maximum loss function is proposed in this paper. The loss function can guide the learning directions of two base CNNs and maximize the difference between features from the two base CNNs, so as to extract complementary and diverse features. In addition, we design a focused feature-enhancement module to make features in the shallow convolutional layer more significant. In this way, the detection performance of small objects can be effectively improved while not increasing the computational cost in the testing stage. Furthermore, since the actual receptive field is usually much smaller than the theoretical receptive field, the deep convolutional layer would not have sufficient semantic features for accurate detection of large objects. To overcome this drawback, a cascaded semantic extension module is added to the deep layer. Through simple multi-branch convolutional layers and dilated convolutions with different dilation rates, the cascaded semantic extension module can effectively enlarge the actual receptive field and increase the detection accuracy of large objects. We compare our detection network with five other state-of-the-art infrared and visible image object detection networks. Qualitative and quantitative experimental results prove the superiority of the proposed detection network.

Keywords: infrared and visible image object detection; convolutional neural network; difference maximum loss function; focused feature enhancement module; cascaded semantic extension module

Citation: Xiao, X.; Wang, B.; Miao, L.; Li, L.; Zhou, Z.; Ma, J.; Dong, D. Infrared and Visible Image Object Detection via Focused Feature Enhancement and Cascaded Semantic Extension. *Remote Sens.* **2021**, *13*, 2538. https://doi.org/10.3390/rs13132538

Academic Editors: Stefano Mattoccia, Piotr Kaniewski and Mateusz Pasternak

Received: 6 June 2021
Accepted: 23 June 2021
Published: 29 June 2021

1. Introduction

Infrared sensors can perform target detection in almost all weather conditions, and are not affected by night, occlusion, or fog. However, infrared images are usually lacking detailed information of the scene and objects within it. By contrast, visible images contain more detail features and are more convenient for visual perception. When it comes to bad weather, night, or occlusion, visible sensors usually tend to lose interesting objects. Combined infrared and visible image object detection (also known as multi-band/multi-sensor image object detection) aims to fuse the advantages of both infrared and visible images, producing more accurate object detection results for scene perception and intelligent decision-making.

In recent years, the convolutional neural network (CNN) has become the most effective way to implement object detection for natural images [1–5]. Inspired by this, more and more scholars have used CNNs to perform infrared and visible image object detection [6–10], in which the infrared image and the visible image have the same resolution and are already registered. Infrared and visible image object detection methods based on CNNs generally include two stages: a feature extraction stage and an object

detection stage. For the object detection in the natural image, only one base CNN (e.g., VGG16 [11], ResNet [12], DenseNet [13]) is used to extract convolutional features. Unlike the detection in natural images, infrared and visible image object detection usually needs two base CNNs to respectively extract infrared features and visible features. The extracted features are then combined in one convolutional layer (usually the last layer), or multiple convolutional layers. In the object detection stage, combined features are utilized to classify and locate interesting objects. The combined features from only one convolutional layer usually face difficulties in detecting multi-scale objects in various scenes. In contrast, the combined features from multiple convolutional layers are more suitable for the detection of multi-scale objects.

In feature extraction stage, since the inputs of the multi-band detection network comprise two images (i.e., an infrared image and a visible image), two base CNNs are added to the whole detection network. The two CNNs are usually designed to be the same in many detection methods [8,9,14]. However, the features contained in the infrared image and the visible image are complementary and quite different. It would be very difficult for two identical CNNs to extract those diverse infrared and visible features. To solve this problem, we design a difference maximum loss function to extract diverse features. The designed loss function would punish similar features and reward different features in order to maximize the diversity and complementarity of the infrared features and the visible features in the extracted features.

In the object detection stage, the extracted features from multiple convolutional layers are usually used to detect multi-scale objects [15,16]. The convolutional layers are usually divided into two groups, that is, shallow layers and deep layers. The shallow layer is at the front of the CNN and has a relatively high resolution, while the deep layer is at the back of CNN and the resolution of the deep layer is relatively high. The features from relatively shallow convolutional layers are responsible for the detection of small objects. Large objects are recognized and located by features from deeper convolutional layers. However, directly using features from shallow or deep layers to implement object detection may result in some drawbacks.

For the detection of small objects, although the shallow detail features are used, these features are still relatively rough and not significant for some small objects. In this case, the shallow layers may not produce good detection results for these small objects. In this paper, we design a focused feature enhancement module to strengthen the shallow convolutional features of small objects, so as to make the detection of small objects easier. The designed module is achieved via the supervised training of semantic segmentation. The segmentation labels can be automatically generated on the basis of ground-truth detection labels (bounding box). In addition, the focused feature enhancement module is only added to the training stage, and not used in the testing stage. Hence, our designed module can effectively improve the detection accuracy of small objects without increasing the testing time.

On the other hand, large objects are usually detected and recognized by using the features from deeper convolutional layers. This is because deeper layers have a larger receptive field and thus contain more semantic and structural information, which is good for the detection of large objects. However, according to practical experience, the actual receptive field is usually smaller than the theoretical receptive field. In this case, convolutional features from deeper layers would not completely cover some large objects, resulting in the decrease of detection accuracy for large objects. In this paper, we propose a cascaded semantic extension module to enlarge the receptive field, in order to improve the detection performance for large objects. The proposed module utilizes multi-scale convolutional kernels and dilated convolutions, and can easily be integrated into original convolutional neural networks.

The rest of this paper is organized as follows. In Section 2, we describe the details of the proposed infrared and visible image object detection algorithm. In Section 3, experimental

results and comparisons are given to verify the superiority of the proposed detection method. Some network analysis is given in Section 5. We conclude this paper in Section 6.

2. The Proposed Detection Network

Figure 1 shows the pipeline of the proposed infrared and visible image object detection network. The proposed detection network adopts the architecture of the classical detection method Faster R-CNN [1]. The two base networks (i.e., CNN1 and CNN2) both use the architecture of feature pyramid network (FPN) [15] (as shown in Figure 2) to effectively utilize multiple convolutional layers. The 50 convolutional layers in the ResNet-50 model [12] are used for CNN1 and CNN2. The resolutions of the second stage, the third stage, the fourth stage, the fifth stage are 1/4, 1/8, 1/16, and 1/32, respectively.

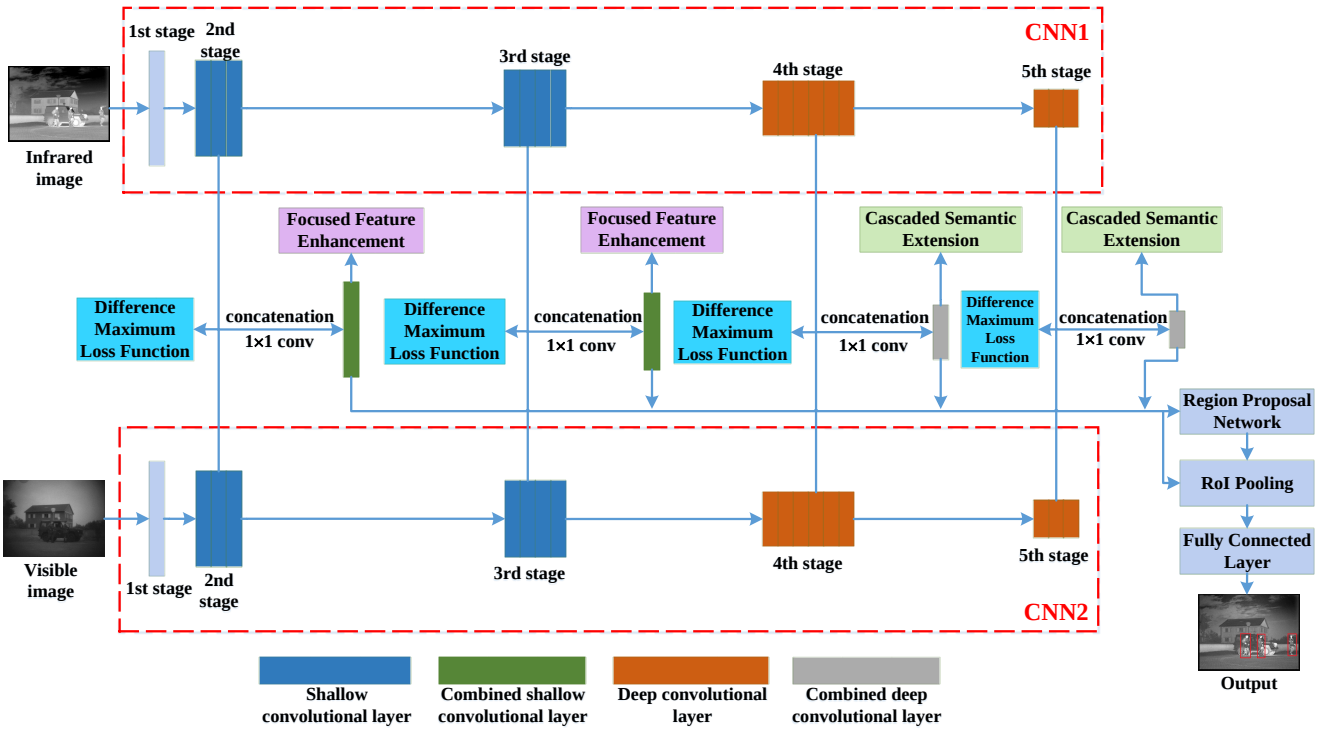


Figure 1. The pipeline of the proposed infrared and visible image object detection network.

Firstly, CNN1 and CNN2 are used to extract the multi-scale features of the infrared and the visible images from multiple convolutional layers. We design the difference maximum loss function to guide the learning directions of the two base networks in order to extract more complementary and diverse multi-band features. Then, infrared features and visible features in each stage (except for the first stage) are respectively combined via concatenation by the channel and a 1×1 convolution. Since the concatenation doubles the number of channels, a 1×1 convolution is added to reshape the channels to the original number.

We define the second stage and the third stage as shallow convolutional layers, and the fourth stage and the fifth stage as deep convolutional layers. The shallow/deep convolutional layers from the infrared image and the shallow/deep convolutional layers from the visible image are combined shallow/deep convolutional layers. We propose a focused feature enhancement module to enhance the shallow features of small objects. In this way, the detection performance of small objects can be effectively improved while keeping the computational cost unchanged. In addition, we design a cascaded semantic extension module to enlarge the receptive field of the deep convolutional layer. The large receptive field is able to increase the detection accuracy of large objects. With the

proposed focused feature enhancement module and cascaded semantic extension module, the detection network can more accurately detect small and large objects.

The other components, such as the region proposal network, RoI pooling, and fully connected layer, are the same as those in Faster R-CNN [1]. More details can be found in [1].

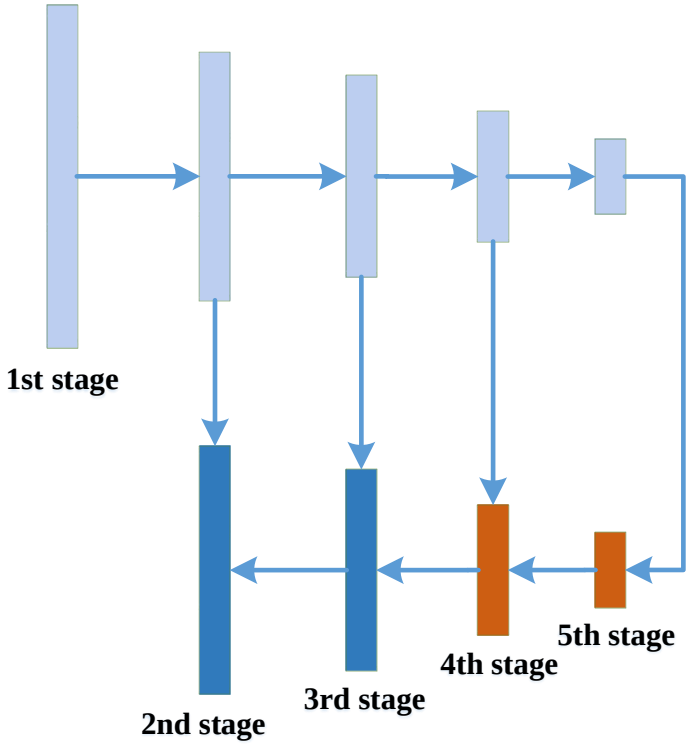


Figure 2. The architecture of the feature pyramid network (FPN).

2.1. Difference Maximum Loss

In recent years, convolutional neural networks (CNNs) have shown great advantages in object detection for natural images [17–19]. This encouraged researchers in related fields to detect and recognize multi-band (infrared and visible) images using CNNs. For infrared and visible image object detection methods based on CNNs, they usually design two base CNNs to respectively extract infrared features and visible features. The two base CNNs can be the same or different.

On the one hand, when two base CNNs are the same, although two CNNs are used to respectively extract features from two images, two networks may learn in the same direction in the training process. In this situation, the extracted features using two of the same networks would not be distinct and complementary. These features are not able to represent the respective advantages of the infrared image and the visible image, resulting in the reduction of the detection accuracy. On the other hand, when the two base CNNs are different, extracted features are usually distinct and complementary. However, the extracted features are complementary only on one or several levels (i.e., not both), because a network with a given structure can only extract one type or several types of features. There are many complementary features on other levels in infrared and visible images. Hence, it is also difficult for two different base CNNs to effectively extract the complementary features.

Although we do not know how many complementary features exist in the infrared image and the visible image, we can be sure that the complementary features must be distinct and varied, because only by combining distinct features can the object detection accuracy be improved. Based on this finding, in this paper we propose a difference maximum loss function. The loss function can guide two base CNNs in learning in different directions in the training process, in order to extract complementary features on

more levels. As shown in Figure 1, the input of the difference maximum loss function is set as the last convolutional layer of each stage (except the first stage). By judging the similarity of two convolutional layers, the loss function guides the learning directions of the two base CNNs. Since the loss function is used to extract complementary features, the structures of the two base CNNs are set to be the same, that is, 50 convolutional layers in the ResNet-50 model [12]. We take advantage of Kullback–Leibler (KL) divergence [20] to define the difference maximum loss function:

$$L_d(p_1, p_2) = 1 - \frac{1}{N} \sum_{p_1 \in E_1, p_2 \in E_2} p_1 \log \frac{p_1}{p_2}, \quad (1)$$

where E_1 denotes features from the last convolutional layer of each stage in CNN1, E_2 denotes features from the last convolutional layer of each stage in CNN2, p_1 is the intensity value in each position of E_1, and p_2 is the intensity value in each position of E_2. p_1 and p_2 are computed via softmax function. N denotes the number of features from E_1 or E_2.

The second term in Equation (1) is KL divergence. When CNN1 and CNN2 are learning in different directions, the gap between p_1 and p_2 becomes large, resulting in the enlargement of the KL divergence. In this case, the loss function L_d becomes small, which implies that the learning directions of the two networks can meet the requirements of extracting complementary features. When CNN1 and CNN2 are learning in the same direction, the gap between p_1 and p_2 becomes small, resulting in the decrease of KL divergence. Then, the loss function becomes large, implying that the learning directions of the two networks are incorrect. A large loss function would guide the two networks to learn in different directions in subsequent iterations. Through continuous iterations in the training process, features from the two networks can be diverse and complementary on multiple levels.

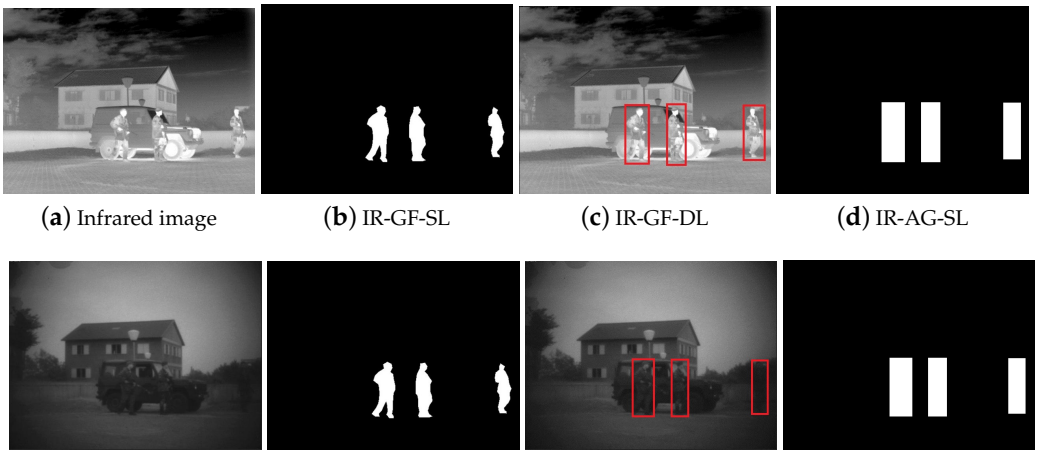
2.2. Focused Feature Enhancement

Since R-CNN [21], SPP-Net [22], Fast R-CNN [23], and Faster R-CNN [1] have been used to perform object detection, researchers have found that these CNN-based detection algorithms can produce significantly higher classification and location accuracy than conventional detection algorithms like the Viola–Jones detector [24,25], HOG detector [26], deformable part-based models [27–31], and so on. However, these early CNN-based methods are not able to produce satisfactory detection results for small objects. This is mainly because detection methods like Faster R-CNN only utilize one convolutional layer to perform object detection, and the used convolutional features are rough and sparse, which is bad for the detection of small objects but good for the detection of large objects. To solve this problem, multiple convolutional layers are employed to improve the small-object detection performance in later CNN-based detection methods, including SSD [2], FPN [15], YOLOv4 [32], and so on.

Generally, multiple convolutional layers are usually divided into two groups: shallow layers and deep layers. The deep layer is at the back of CNN and has relatively low resolution. The deep layer contains more semantic features, which are good for the detection of large objects. The early detection methods, such as Fast R-CNN [23] and Faster R-CNN [1], take advantage of the deep convolutional layer to produce more accurate detection results than conventional detection methods. However, the deep layer lacks detail information, which is usually required for the detection of small objects. Hence, it is very difficult for the early detection methods to accurately detect small objects. On the other hand, the shallow layer is in the front of the CNN. The resolution of the shallow layer is relatively large. Thus, the shallow layer contains more detail features and fewer semantic features, which is beneficial for the detection of small objects. Based on the above characteristics, deep layers and shallow layers (i.e., multiple convolutional layers) are used at the same time to perform multi-scale object detection [33–37]. This strategy can produce higher detection accuracy for both small objects and large objects.

However, directly utilizing the shallow layer and the deep layer may present some shortcomings. When it comes to the detection of small objects using shallow convolutional layers, in order to make shallow layers contain more semantic features, the resolution of the initial shallow layer is usually set as 1/4 of the resolution of the input image [2,4]. In this situation, the shallow layer would contain very few features for some small objects, and these features would be rough and not significant. In addition, since some small objects are relatively vague and indistinguishable, the shallow layer may lose features for these small objects. Although shallow layers are more suited to the detection of small objects than deep layers, it is still difficult for shallow layers to accurately detect some small objects.

In order to solve this problem, in this paper, we propose a focused feature enhancement module to strengthen the convolutional features of small objects in shallow layers. A common way to strengthen convolutional features is to stacking many convolutional layers. Although this can be straightforward and effective, it is time- and resource-consuming. To overcome this difficulty, we introduce semantic segmentation to achieve focused feature enhancement. As shown in Figure 1, semantic segmentation is added to shallow convolutional layers (i.e., the second and third stages) of the detection network. Figure 3a,e shows the infrared image and the visible image, respectively. Figure 3b,f shows the ground-truth segmentation labels of subfigures (a) and (e). In the segmentation label, pixels in white regions denote positive samples (i.e., 1), and pixels in black regions denote negative samples (i.e., 0).



(**a**) Infrared image (**b**) IR-GF-SL (**c**) IR-GF-DL (**d**) IR-AG-SL

(**e**) Visible image (**f**) VI-GF-SL (**g**) VI-GF-DL (**h**) VI-AG-SL

Figure 3. (**a**) The infrared image; (**b**) the ground-truth segmentation label of (**a**), abbreviated IR-GF-SL; (**c**) the ground-truth detection label of (**a**), abbreviated IR-GF-DL; (**d**) the automatically generated segmentation label based on bounding boxes in (**c**), abbreviated IR-AG-SL. (**e**) The visible image; (**f**) the ground-truth segmentation label of (**e**), abbreviated VI-GF-SL; (**g**) the ground-truth detection label of (**e**), abbreviated VI-GF-DL; (**h**) the automatically generated segmentation label based on bounding boxes in (**g**), abbreviated VI-AG-SL.

In the training process, as shown in Figure 4, the last convolutional layer of the shallow layer outputs the segmentation result. The resolution of the last layer of the second stage is $w_2 \times h_2 \times c_2$ (width $\times$ height $\times$ channel). The last layer is then computed with a 3×3 convolution to produce the segmentation result, whose resolution is $w_2 \times h_2 \times 2$, where 2 denotes the number of the sample category (i.e., positive sample and negative sample). The softmax function is used to resize the feature values of the segmentation result to [0,1]. Finally, we use a cross-entropy loss function to compute the gap between the segmentation result and the ground-truth segmentation label. Based on the gap, the detection network guides shallow layers to be focused on enhancing the features of small objects. This process is also suitable for the third stage. Through the supervised training of semantic segmentation, the features of small objects in the shallow layers can be effectively enhanced in order to improve the small-object detection performance.

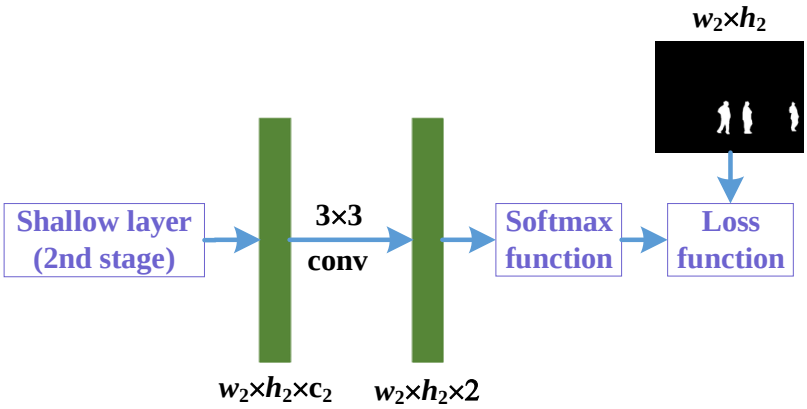


Figure 4. The focused feature enhancement for the second stage.

Since semantic segmentation is used, we need to manually annotate the ground-truth segmentation label for each training image using annotation tool LabelMe (https://github.com/CSAILVision/LabelMeAnnotationTool, accessed on 20 April 2019). However, this will consume too much time and effort. Figure 3c,g shows the ground-truth detection labels of subfigures (a) and (e), respectively. For the relief of the burden, we automatically generate segmentation labels (Figure 3d,h) based on the ground-truth detection labels (Figure 3c,g). This saves significant time and effort. From Figure 3b,d, we can see that the generated labels cover a relatively larger region than the ground-truth labels. Hence, the generated labels (Figure 3d,h) can also be focused on strengthening the features of small objects. The cross-entropy loss function used in semantic segmentation is defined as

$$L_f(h, p, q) = -\lambda \sum_{p \in I_+} \log h_p - \sum_{q \in I_-} \log h_q, \tag{2}$$

where I_+ and I_- denote the positive sample set and the negative sample set, respectively. h_p is the probability that pixel p is classified as a positive sample. h_q denotes the probability that pixel q is classified as a negative sample. h_p and h_q are computed with softmax function. For class balancing, we introduce the weight λ. λ is defined as $\frac{|I_-|}{|I_+|}$, where $|I_-|$ and $|I_+|$ are the number of negative and positive samples, respectively.

The proposed focused feature enhancement module based on semantic segmentation is only added to the training stage. In the testing stage, the shallow layers do not output segmentation results. In this way, the proposed module can effectively increase the detection rate of small objects without increasing the testing time.

2.3. Cascaded Semantic Extension

Compared with the shallow layer, the deep layer is at the back of CNN, and has a relatively lower resolution and larger receptive field. The deep layer contains more semantic features and structure information, which can contribute to a more accurate detection of large objects. The receptive field is defined as the region in the input image that the pixel in the convolutional layer can affect. The deeper the convolutional layer, the larger the receptive field. The larger receptive field can make the pixel in the convolutional layer affect a greater range, and contain more deep features. The reason that the deep convolutional layer is usually taken advantage of to detect large objects is that the deeper layer contributes a larger receptive field.

However, the actual receptive field only occupies a fraction of the theoretical receptive field [38–41]. The actual receptive field has a Gaussian distribution, and pixels at the center of a receptive field have a much larger impact on the output, and the impact of surrounding pixels generally decays quickly. Under this circumstance, convolutional features from the deep layer would not completely cover some large objects, and some important information would be left out when making the prediction. Therefore, this probably induces some decrease of detection accuracy and robustness.

A simple and natural way of enlarging the actual receptive field is to increase the number of convolutional layers. Unfortunately, this would lead to a high computational cost and limit the efficiency of the detection network. In this paper, we propose a cascaded semantic extension module to effectively enlarge the receptive field while still keeping the computational cost under control. As shown in Figure 1, the proposed cascaded semantic extension module is cascaded with deep convolutional layers (i.e., the fourth stage and the fifth stage) of the detection network. The structure of the proposed module is shown in Figure 5. We can see that the proposed module mainly makes use of a multi-branch convolutional layer with different kernel sizes, and each convolution is followed by a dilated convolution with a corresponding dilation rate. The outputs of the three branches are concatenated and then reshaped with a 1×1 convolution to produce the final enhanced features.

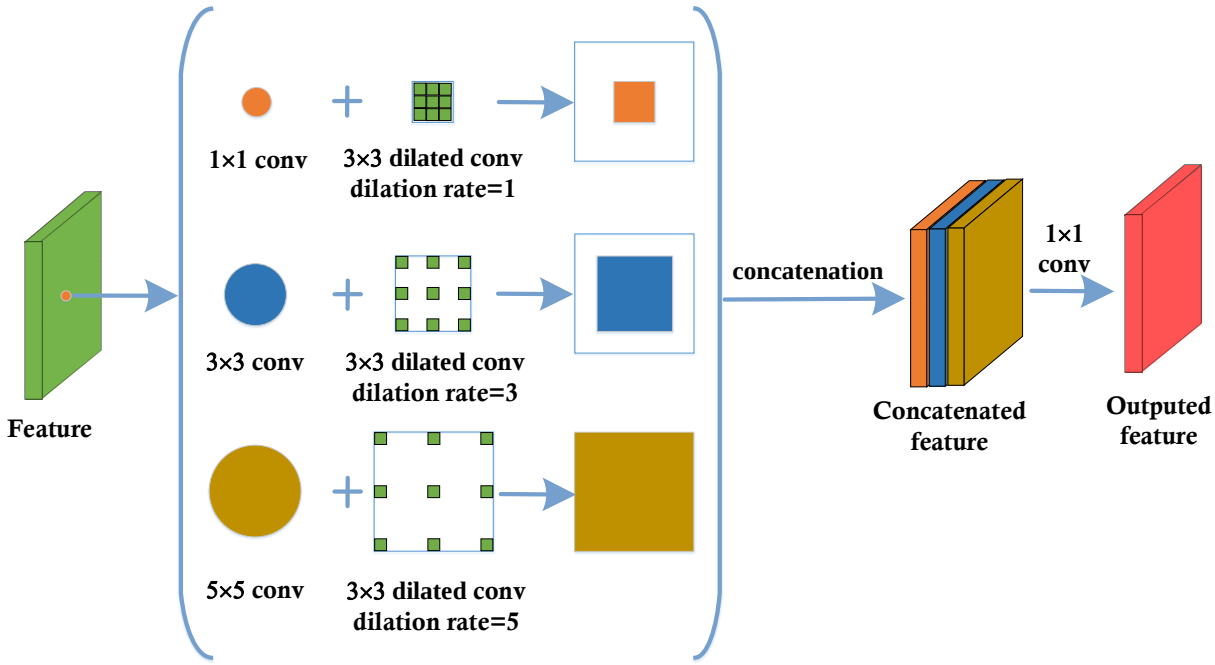


Figure 5. The structure of the proposed cascaded semantic extension module, in which 'conv' denotes convolution.

As shown in Figure 5, in the proposed module, we first design a three-branch convolutional layer (i.e., 1×1 convolution, 3×3 convolution, 5×5 convolution). The 1×1 and 3×3 convolutions are responsible for extracting relatively small-scale features, and the aim of the 5×5 convolution is to extract large-scale features. Through the three-branch convolutional layer, the deep features can be enhanced and the receptive field can be enlarged. To further enlarge the receptive field, we introduce dilated convolution following the three-branch convolutional layer. Dilated convolution [42–44], also known as atrous convolution [41,45,46], aims to generate a feature map with higher resolution, capturing information in a larger area with more context while minimally increasing the computation cost.

We set a 3×3 dilated convolution with dilation rate 1 followed by 1×1 convolution, a 3×3 dilated convolution with dilation rate 3 followed by 3×3 convolution, and a 3×3 dilated convolution with dilation rate 5 followed by 5×5 convolution. The larger the dilation rate, the larger the receptive field. The dilation rate is set as the same as the front convolutional kernel size in order to make different branches focus on enhancing the features with particular sizes. Then, we concatenate the outputs of the three branches by the channel. Finally, a 1×1 convolution is used to reshape the concatenated features to the original size.

In order to effectively enlarge the receptive field, we stack three cascaded semantic extension modules following each deep layer. In this way, the actual receptive field can be significantly enlarged and the deep features can also be enhanced. Therefore, the detection performance of large objects can be effectively boosted. Besides, since the proposed module only contains a three-branch convolutional layer, the increase of the computational cost can be very minor.

2.4. End-to-End Training

The proposed detection method adopts the detection architecture of Faster R-CNN, and we define the loss function of Faster R-CNN as L_{Faster}. In this paper, the loss functions of the newly proposed difference maximum loss and focused feature enhancement module are L_d (Equation (1)) and L_f (Equation (2)), respectively. Thus, the loss function of the whole detection network is defined as

$$L = L_{Faster} + L_d + L_f, \tag{3}$$

The base networks CNN1 and CNN2 are initialized with ResNet-50 pre-trained weights for ImageNet classification [12]. The stochastic gradient descent (SGD) optimizer [47] is used to optimize the network parameters. The detection network is trained end-to-end, which means that the proposed detection network can directly output the detection results based on the input images, without any other operations. We use an NVIDIA GTX 1080 Ti GPU to train and test the detection network. The weights of the network are updated with a learning rate of 10^{-4} for the first 50k iterations, and 10^{-5} for the next 50k iterations. The momentum, weight decay, and batch size are set as 0.9, 0.0005, and 2, respectively. The code of the proposed detection network is implemented based on PyTorch [48].

3. Experiments

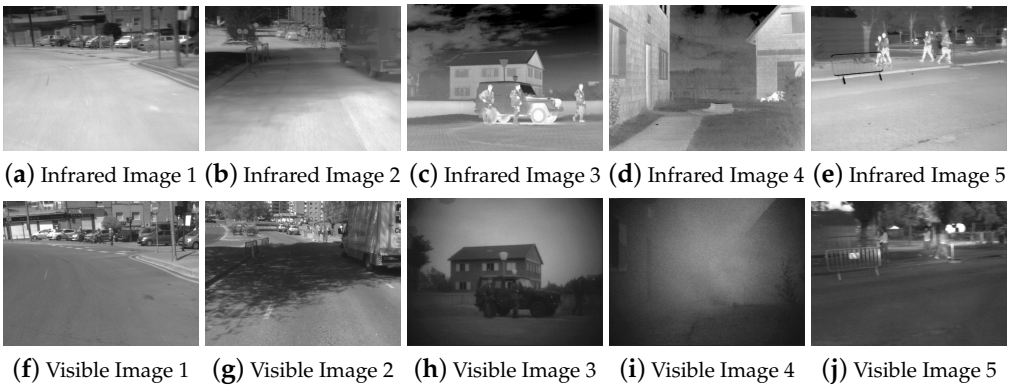
3.1. Infrared and Visible Image Dataset

In this paper, the used infrared and visible image dataset is collected from [49,50]. Each pair of infrared and visible images are collected from aligned infrared and visible cameras, and each pair of images were already registered. Both the infrared and visible images are single-channel gray images. The used infrared images are far-infrared images. Table 1 lists the composition of the infrared and visible image dataset. We can see that the dataset contains a total of 3318 pairs of infrared and visible images, in which 1641 pairs of images are in the daytime and 1677 are in the night. We randomly select 668 pairs of infrared and visible images as the testing images, and the remaining 2650 pairs of infrared and visible images as the training images. In the 668 pairs of testing images, 352 pairs of images are in the daytime and 316 are in the night. In the 2650 pairs of training images, 1289 pairs of images are in the daytime, and 1361 are in the night. The image sizes include 640×471 (width $\times$ height) and 640×480, and we resize all infrared and visible images into 640×480. Data augmentation is introduced to avoid the over-fitting of the detection network. We use two augmentation strategies, that is, horizontal flip and Gaussian blur with standard deviation of 2, to increase the number of training images. Through data augmentation, we get 7950 pairs of infrared and visible images for the training of the detection network.

Table 1. The composition of the infrared and visible image dataset.

	Training Images	Testing Images	Sum
Daytime	1289	352	1641
Night	1361	316	1677
Sum	2650	668	3318

In the infrared and visible image dataset, there are several different object categories, including person, car, tree, building, and so on. The images of people in the person category include people that are still, walking, running, and carrying various things. In this paper, we only detect one category, that is, person. Figure 6 shows some examples in the infrared and visible image dataset. The images in the first row are visible images, and the images in the second row are infrared images. The first two columns show the images in the daytime, and the last three columns show the images in the night. We can see that although the objects in the visible images contain more detail information, the visible images are easily affected by low brightness (see Figure 6h,j), smoke (see Figure 6i), and noise (see Figure 6h). On the other hand, the contrast of the infrared objects is relatively high (see Figure 6c–e), while detail features in infrared objects are missing (see Figure 6c–e). Besides, from the first two columns, we can see that in the daytime, visible images may have better visual effects than infrared images.



(**a**) Infrared Image 1 (**b**) Infrared Image 2 (**c**) Infrared Image 3 (**d**) Infrared Image 4 (**e**) Infrared Image 5

(**f**) Visible Image 1 (**g**) Visible Image 2 (**h**) Visible Image 3 (**i**) Visible Image 4 (**j**) Visible Image 5

Figure 6. (**a**–**j**) Some examples of infrared and visible images from the used image dataset. The images in the first row are infrared images, and the images in the second row are visible images.

3.2. Method Comparison

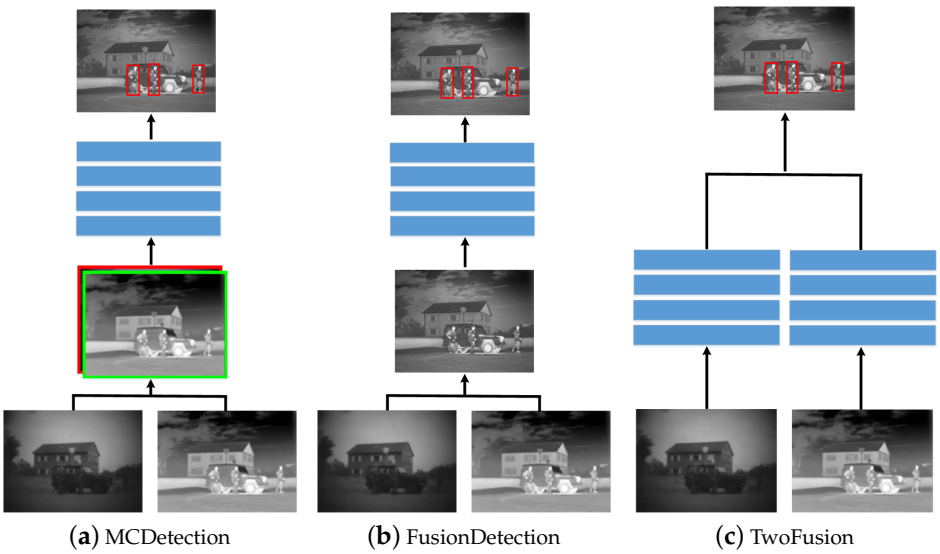
The proposed detection method is compared with five other infrared and visible image detection methods, including MCDetection [51], FusionDetection [1,52], TwoFusion [53], TripleFusion [8], and IAF R-CNN [9]. The first two detection methods use pixel-level fusion, and the last three methods and our detection method use feature-level fusion. All six methods are respectively trained and tested with the same training images and testing images, and the common network parameters of the six methods are identical.

MCDetection: As shown in Figure 7a, MCDetection [51] first combines a single-channel infrared image and a single-channel visible image into a two-channel pseudo-color image. Then, the two-channel image is used as the input of the detection network Fast R-CNN. Fast R-CNN finally outputs the detection results for the infrared image and the visible image. Since the other five detection networks are designed based on Faster R-CNN, we change the Fast R-CNN in MCDetection to Faster R-CNN. The infrared features and the visible features are fused at the pixel level, and thus MCDetection uses one base CNN to extract infrared and visible features.

FusionDetection: As shown in Figure 7b, FusionDetection first utilizes the multi-band image fusion method HMSD [52] to fuse a single-channel infrared image and a single-channel visible image into a single-channel fused image. Note that other state-of-the-art multi-band image fusion methods [54–58] can also be used for FusionDetection. Then, the fused image is detected with the detection network Faster R-CNN [1]. For the detection method FusionDetection, infrared features and visible features are also fused at the pixel level, and thus one base CNN is used to extract multi-band features.

TwoFusion: As shown in Figure 7c, TwoFusion [53] uses two base CNNs to respectively extract infrared features and visible features, and then the extracted features are fused for the subsequent recognition and location of the objects of interest. The detection architecture of Faster R-CNN is adopted in TwoFusion, and the two base CNNs have the same structure.

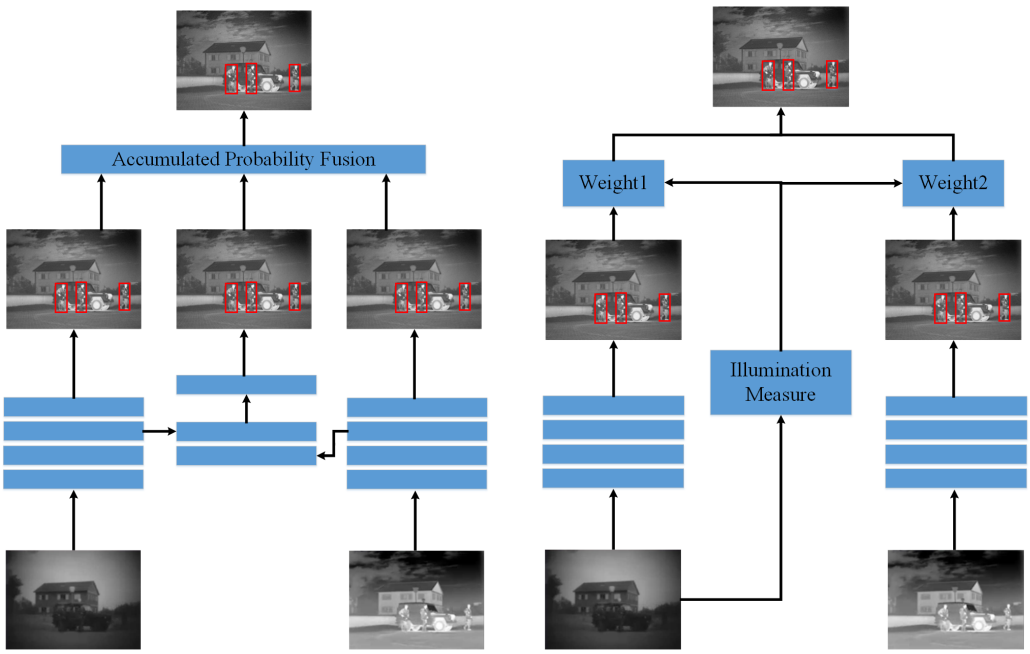
TripleFusion: As shown in Figure 8a, TripleFusion [8] proposes a three-branch detection architecture, and takes advantage of infrared features, visible features, and their fused features to respectively perform classification and regression of the region proposal. Then, the output results of the three branches are combined with an accumulated probability fusion layer to produce more accurate detection results. Note that TripleFusion also uses two of the same base CNNs.



(**a**) MCDetection (**b**) FusionDetection (**c**) TwoFusion

Figure 7. (**a**–**c**) The pipelines for MCDetection, FusionDetection, and TwoFusion.

IAF R-CNN: As shown in Figure 8b, IAF R-CNN [9] discovers that the detection performance is correlated with illumination conditions. Therefore, IAF R-CNN first uses two detection networks (Faster R-CNN) to respectively produce detection results for the infrared image and the visible image. In this process, the fused features from the two networks are used to generate region proposals. Then, an illumination-aware network is introduced to measure the illumination of the visible image. According to the measured illumination value, the detection results from the two detection networks are adaptively merged to obtain the final detection outputs.



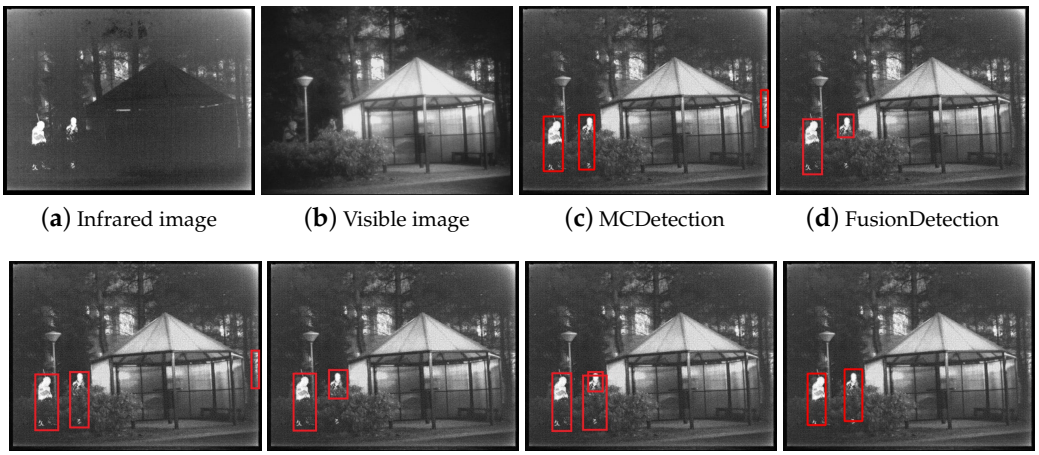
(**a**) TripleFusion (**b**) IAF R-CNN

Figure 8. (**a**,**b**) The pipelines for TripleFusion and IAF R-CNN.

3.3. Comparison

Figure 9 shows the infrared and visible image detection results from the six different detection methods. In order to display the results more clearly, the detection results are shown on the fused images of the infrared and visible images, and the fusion method HMSD [52] is used to produce the fused images. Since MCDetection and FusionDetection simply combine an infrared image and a visible image into a fused image, the detection network cannot effectively extract complementary features from only a fused image. Therefore, they produce inaccurate location and classification results (Figure 9c,d). TwoFusion uses two of the same base CNNs to extract complementary and diverse infrared and visible features, and it is very difficult for this strategy to achieve the desired goal. Without complementary features, the detection network TwoFusion gives incorrect detection results (Figure 9e). Although TripleFusion and IAF R-CNN design more complex network structures, the detection results are still unsatisfactory (Figure 9f,g). Thanks to the carefully designed difference maximum loss function, focused feature enhancement module, and cascaded semantic extension module, our detection network gives more accurate detection results.

Figure 10 shows another comparison example. We can see that objects in the images are crowded and not easily distinguished, and some irrelevant objects are very similar to interesting objects to be detected. In this situation, the other five detection methods give inaccurate location or classification results (Figure 10c–g), while the proposed method outputs satisfactory detection results (Figure 10h). Figure 11 shows a similar example to Figure 10. Some objects in Figure 11 are very small, and some lighting can confuse detection networks. From Figure 11c–e, we can see that MCDetection, FusionDetection, and TwoFusion give quite inaccurate location and classification results. The detection results of TripleFusion and IAF R-CNN can also be improved (see Figure 11f,g). In contrast, the output of our detection network is more accurate.

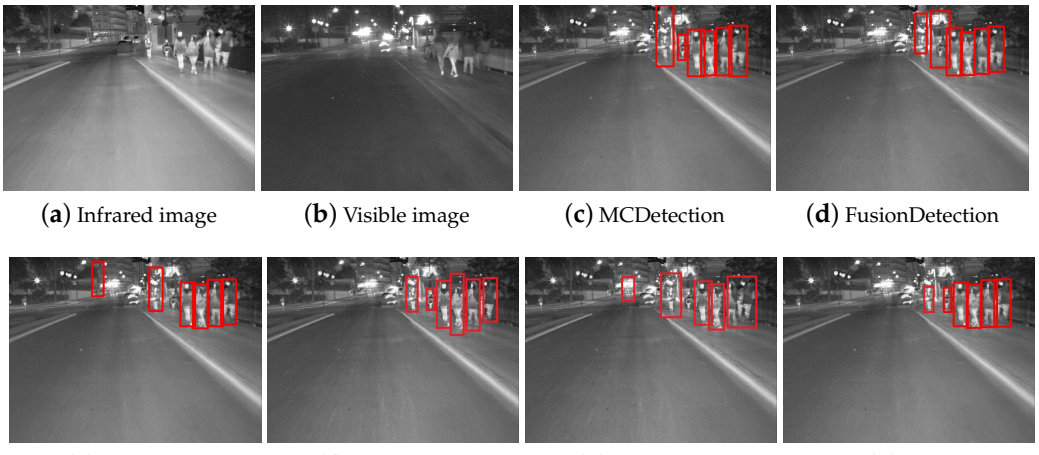


(**a**) Infrared image (**b**) Visible image (**c**) MCDetection (**d**) FusionDetection

(**e**) TwoFusion (**f**) TripleFusion (**g**) IAF R-CNN (**h**) Ours

Figure 9. (**a**–**h**) The infrared and visible image detection results from six different detection methods. In order to better display the results, the detection results are shown on the fused images.

We use mean average precision (mAP) to quantitatively evaluate the detection performance of the six detection networks. Table 2 lists the mAPs of the different detection methods on the testing set. The first column shows the six detection methods, the second column shows the mAPs of the testing images in the daytime, the third column shows the mAPs of the testing images in the night, and the fourth column shows the mAPs of all testing images. MCDetection and FusionDetection only use one base CNN to extract diverse multi-band features, resulting in lower mAP compared with the other detection networks. Since TwoFusion uses the same two CNNs, its detection accuracy is also relatively low. Although TripleFusion and IAF R-CNN introduce well-designed network structures, their mAPs are still lower than that of our detection network.



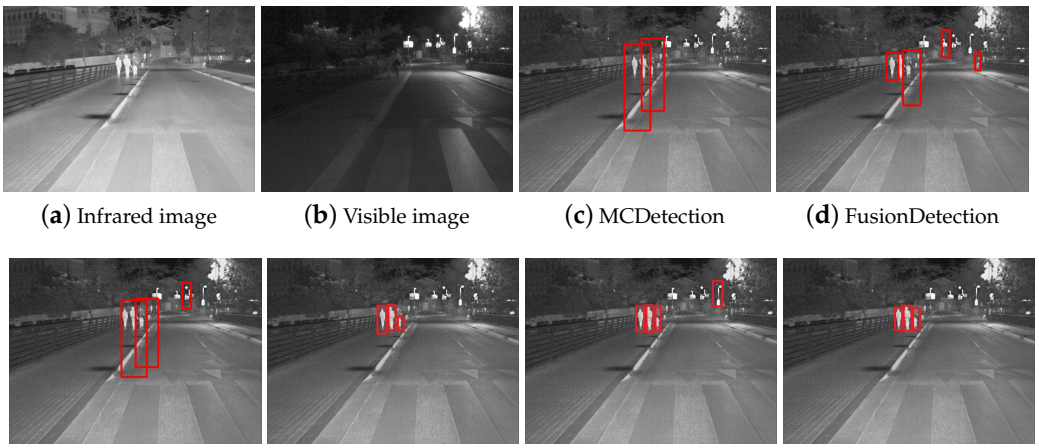
(**a**) Infrared image (**b**) Visible image (**c**) MCDetection (**d**) FusionDetection

(**e**) TwoFusion (**f**) TripleFusion (**g**) IAF R-CNN (**h**) Ours

Figure 10. (**a**–**h**) The infrared and visible image detection results from six different detection methods. In order to better display the results, the detection results are shown on the fused images.

Table 2. The mAP for the different detection methods on the testing set.

Method	Daytime	Night	ALL
MCDetection	74.8%	73.9%	74.3%
FusionDetection	74.4%	75.5%	75.1%
TwoFusion	78.8%	77.5%	78.2%
TripleFusion	81.1%	80.1%	80.5%
IAF R-CNN	80.9%	81.8%	81.3%
Ours	84.3%	83.2%	83.7%



(**a**) Infrared image (**b**) Visible image (**c**) MCDetection (**d**) FusionDetection

(**e**)TwoFusion (**f**) TripleFusion (**g**) IAF R-CNN (**h**) Ours

Figure 11. (**a**–**h**) The infrared and visible image detection results from six different detection methods. In order to better display the results, the detection results are shown on the fused images.

4. Individual Results

Figure 12 gives some detection results from our detection network. The first column shows infrared images, the second column shows visible images, and the third column shows detection results. From the first three rows, we can see that our method is able to accurately recognize low-contrast objects. Owning to the focused feature enhancement module, the proposed network outputs good detection results for small objects (see the fourth row in Figure 12). Thanks to our proposed cascaded semantic extension module, the large object in the fifth row is accurately located. In the daytime (see the last row), our method still produces good detection performance.

(**a**) Infrared image (**b**) Visible image (**c**) Ours

Figure 12. (**a**–**c**) Some detection results from our detection network. In order to better display the results, the detection results are shown on the fused images. The images in the first three rows have very low contrast; the objects in the fourth row are relatively small; the object in the fifth row is relatively large; and the images in the last row are in the daytime.

5. Discussion

A. The effectiveness of the proposed difference maximum loss, focused feature enhancement, and cascaded semantic extension. In this paper, we mainly propose three novel modules (i.e., difference maximum loss function, focused feature enhancement module, and cascaded semantic extension module). In order to demonstrate the effectiveness of the three proposed modules, massive experiments are implemented and some results are listed in Table 3. × denotes that the module is not used in the detection network, and ✓ denotes that the module is used in the detection network. All testing images are used to produce the detection accuracy (mAP). We can see that without the three proposed modules, the accuracy of the detection network decreases to 78.9%. Using only the difference maximum loss function, the detection accuracy is 80.7%. Using only the focused feature enhancement module, the detection accuracy is 80.1%. Using only the cascaded semantic extension module, the detection accuracy is 81.6%. Hence, all three modules can improve the detection performance, and the cascaded semantic extension module has the largest benefit. When using two modules for the detection network, the detection accuracy can be further improved. When using all three modules, the detection accuracy reaches a maximum of 83.7%.

Table 3. The effectiveness of difference maximum loss, focused feature enhancement, and cascaded semantic extension.

Difference Maximum Loss	Focused Feature Enhancement	Cascaded Semantic Extension	mAP
×	×	×	78.9%
✓	×	×	80.7%
×	✓	×	80.1%
×	×	✓	81.6%
×	✓	✓	82.5%
✓	×	✓	82.8%
✓	✓	×	82.1%
✓	✓	✓	83.7%

B. The effectiveness for small objects and large objects. In order to demonstrate the detection effectiveness of small objects and large objects, we define objects with size smaller than 48 × 48 in testing images as small objects, and objects with sizes larger than 96 × 96 as large objects. Our method produces a mAP of 76.5% for small objects, while without the focused feature enhancement module the mAP for small objects drops to 74.2%. For large objects, the mAP from our method is 88.6%. Without the cascaded semantic extension module, the mAP for large objects drops to 87.1%.

C. The first stage in CNN1 and CNN2. In the proposed detection network, the first stage in CNN1 and CNN2 is not used for the difference maximum loss function and the focused feature enhancement module (see Figure 1). There are two reasons for this. Firstly, in many other state-of-the-art object detection networks [19,59], the first stage of the base CNN is not used for various designed modules. Secondly, the resolution of the first stage is one-half the resolution of the input image. Therefore, semantic features in the first stage are very few, and features in the first stage are usually edges and gradients. In this case, those features would be useless for the detection of small objects. Edges and gradients have very small differences between two base CNNs, and thus the difference maximum loss function also does not use the first stage. Table 4 lists the detection accuracy with and without the first stage. We can see that it would be good for the designed detection network not to use the first stage.

Table 4. The mAP with and without the first stage for difference maximum loss and focused feature enhancement.

With/Without	mAP
With the first stage	83.5%
Without the first stage	83.7%

D. Automatically generated segmentation labels for the focused feature enhancement. In the focused feature enhancement module, we use automatically generated segmentation labels instead of ground-truth segmentation labels to save time and effort (see Figure 3). Although the automatically generated segmentation labels are relatively inaccurate compared with the ground-truth segmentation labels, the automatically generated segmentation labels can meet the requirements of the focused feature enhancement module. The generated labels can make the enhancement module effectively concentrate on strengthening the features of small objects. Table 5 lists the detection accuracy based on the ground-truth segmentation labels and the automatically generated segmentation labels. We can see that our used strategy (i.e., automatically generated segmentation labels) can produce satisfactory detection results while effectively saving time and effort for annotation.

Table 5. The mAP for the ground-truth segmentation labels and the automatically generated segmentation labels.

Segmentation Labels	mAP
Ground-truth	83.8%
Automatically generated	83.7%

6. Conclusions

In this paper, a novel infrared and visible image object detection network is proposed. First, we design a difference maximum loss function to guide the learning directions of the two base CNNs. In this way, the extracted multi-band features from the two base CNNs can be complementary on more levels and diverse, which is beneficial for the multi-band object detection. Secondly, the proposed focused feature enhancement module is added to the shallow convolutional layer to improve the small-object detection performance. The proposed module is only employed in the training process without increasing the testing time. Finally, in order to enlarge the receptive field of the deep convolutional layer and increase the large-object detection accuracy, a cascaded semantic extension module is introduced. This module can be easily integrated into the detection network while minimally affecting the computational cost. Experimental results demonstrate that the proposed detection network can achieve superior performance compared with many other state-of-the-art detection methods. Further research will include infrared and visible image fusion and semantic segmentation for the infrared and visible images.

Author Contributions: Conceptualization, X.X.; methodology, X.X., B.W. and J.M.; software, X.X.; validation, X.X., B.W.; investigation, B.W. and L.M.; resources, L.L. and L.M.; data curation, X.X.; writing—original draft preparation, X.X., Z.Z. and J.M.; writing—review and editing, X.X., L.M. and J.M.; visualization, B.W., L.L. and D.D.; supervision, B.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137. [CrossRef]
2. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016; pp. 21–37.
3. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2999–3007.
4. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
5. Wang, J.; Chen, K.; Yang, S.; Loy, C.C.; Lin, D. Region Proposal by Guided Anchoring. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019.
6. Alejandro, G.; Fang, Z.; Yainuvis, S.; Joan, S.; David, V.; Xu, J.; Antonio, L. Pedestrian Detection at Day/Night Time with Visible and FIR Cameras: A Comparison. *Sensors* **2016**, *16*, 820.
7. Konig, D.; Adam, M.; Jarvers, C.; Layher, G.; Teutsch, M. Fully Convolutional Region Proposal Networks for Multispectral Person Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Honolulu, HI, USA, 22–25 July 2017.
8. Park, K.; Kim, S.; Sohn, K. Unified multi-spectral pedestrian detection based on probabilistic fusion networks. *Pattern Recognit. J. Pattern Recognit. Soc.* **2018**, *80*, 143–155. [CrossRef]
9. 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]
10. Shopovska, I.; Jovanov, L.; Philips, W. Deep Visible and Thermal Image Fusion for Enhanced Pedestrian Visibility. *Sensors* **2019**, *19*, 3727. [CrossRef]
11. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
12. He, K.; Zhang, X.; Ren, S.; Jian, S. 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.
13. Huang, G.; Liu, Z.; Laurens, V.D.M.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
14. Hou, Y.L.; Song, Y.; Hao, X.; Shen, Y.; Qian, M. Multispectral pedestrian detection based on deep convolutional neural networks. In Proceedings of the 2017 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC), Macau, China, 21–24 August 2018.
15. Lin, T.Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
16. Liu, S.; Huang, D.; Wang, Y. Receptive Field Block Net for Accurate and Fast Object Detection. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018.
17. Kaiming, H.; Georgia, G.; Piotr, D.; Ross, G. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
18. Cao, J.; Cholakkal, H.; Anwer, R.M.; Khan, F.S.; Shao, L. D2Det: Towards High Quality Object Detection and Instance Segmentation. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020.
19. Dong, Z.; Li, G.; Liao, Y.; Wang, F.; Ren, P.; Qian, C. CentripetalNet: Pursuing High-quality Keypoint Pairs for Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020
20. Rubner, Y.; Puzicha, J.; Tomasi, C.; Buhmann, J.M. Empirical Evaluation of Dissimilarity Measures for Color and Texture. *Comput. Vis. Image Underst.* **2001**, *84*, 25–43. [CrossRef]
21. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014.
22. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *37*, 1904–1916. [CrossRef]
23. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
24. Viola, P. Rapid Object Detection using a Boosted Cascade of Simple Features. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Kauai, HI, USA, 8–14 December 2001.
25. Viola, P.; Jones, M.J. Robust Real-Time Face Detection. *Int. J. Comput. Vis.* **2004**, *57*, 137–154. [CrossRef]
26. Dalal, N. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–26 June 2005.
27. Felzenszwalb, P.F.; Mcallester, D.A.; Ramanan, D. A discriminatively trained, multiscale, deformable part model. In Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008.
28. Felzenszwalb, P.F.; Girshick, R.B.; Mcallester, D.A. Cascade object detection with deformable part models. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010.
29. Felzenszwalb, P.F.; Girshick, R.B.; McAllester, D.; Ramanan, D. Object Detection with Discriminatively Trained Part-Based Models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 1627–1645. [CrossRef] [PubMed]

30. Girshick, R.B.; Felzenszwalb, P.F.; Mcallester, D. Object detection with grammar models. *Adv. Neural Inf. Process. Syst.* **2011**, *24*, 442–450.
31. Girshick, R.B. *From Rigid Templates to Grammars: Object Detection with Structured Models*; University of Chicago: Chicago, IL, USA, 2012.
32. Bochkovskiy, A.; Wang, C.Y.; Liao, H. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
33. Zhou, P.; Geng, C.; Transmission. Scale-Transferrable Object Detection. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.
34. Singh, B.; Davis, L.S. An Analysis of Scale Invariance in Object Detection—SNIP. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.
35. Singh, B.; Najibi, M.; Davis, L.S. SNIPER: Efficient Multi-Scale Training. *arXiv* **2018**, arXiv:1805.09300.
36. Zhu, C.; He, Y.; Savvides, M. Feature Selective Anchor-Free Module for Single-Shot Object Detection. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019.
37. Hong, S.; Roh, B.; Kim, K.H.; Cheon, Y.; Park, M. Pvanet: Lightweight deep neural networks for real-time object detection. *arXiv* **2016**, arXiv:1611.08588.
38. Glorot, X.; Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. *J. Mach. Learn. Res.* **2010**, *9*, 249–256.
39. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.
40. Luo, W.; Li, Y.; Urtasun, R.; Zemel, R. Understanding the Effective Receptive Field in Deep Convolutional Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
41. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 834–848. [CrossRef]
42. Yu, F.; Koltun, V. Multi-Scale Context Aggregation by Dilated Convolutions. *arXiv* **2015**, arXiv:1511.07122.
43. Zhang, X.; Zou, Y.; Wei, S. Dilated convolution neural network with LeakyReLU for environmental sound classification. In Proceedings of the 2017 22nd International Conference on Digital Signal Processing (DSP), London, UK, 23–25 August 2017.
44. Qiao, Z.; Cui, Z.; Niu, X.; Geng, S.; Yu, Q. Image Segmentation with Pyramid Dilated Convolution Based on ResNet and U-Net. In *International Conference on Neural Information Processing*; Springer: Cham, Switzerland, 2017.
45. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. In Proceedings of the Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014.
46. Chen, J.; Wang, C.; Tong, Y. AtICNet: Semantic segmentation with atrous spatial pyramid pooling in image cascade network. *EURASIP J. Wirel. Commun. Netw.* **2019**, *2019*, 1–7. [CrossRef]
47. Krizhevsky, A.; Sutskever, I.; Hinton, G. ImageNet Classification with Deep Convolutional Neural Networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]
48. Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; Lerer, A. Automatic Differentiation in Pytorch. 2017. Available online: https://openreview.net/forum?id=BJJsrmfCZ (accessed on 14 March 2019).
49. Toet, A.; Hogervorst, M.A.; Pinkus, A.R. The TRICLOBS Dynamic Multi-Band Image Data Set for the Development and Evaluation of Image Fusion Methods. *PLoS ONE* **2016**, *11*, e0165016. [CrossRef]
50. CVC14 Dataset. Available online: http://adas.cvc.uab.es/elektra/enigma-portfolio/cvc-14-visible-fir-day-night-pedestrian-sequence-dataset (accessed on 25 July 2019).
51. Liu, S.; Liu, Z. Multi-Channel CNN-based Object Detection for Enhanced Situation Awareness. In Proceedings of the Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
52. Zhou, Z.; Wang, B.; Li, S.; Dong, M. Perceptual fusion of infrared and visible images through a hybrid multi-scale decomposition with Gaussian and bilateral filters. *Inf. Fusion* **2016**, *30*, 15–26. [CrossRef]
53. Liu, J.; Zhang, S.; Wang, S.; Metaxas, D.N. Multispectral Deep Neural Networks for Pedestrian Detection. *arXiv* **2016**, arXiv:1611.02644.
54. Li, S.; Kang, X.; Hu, J. Image Fusion With Guided Filtering. *IEEE Trans. Image Process.* **2013**, *22*, 2864–2875.
55. Adu, J.; Gan, J.; Wang, Y.; Huang, J. Image fusion based on nonsubsampled contourlet transform for infrared and visible light image. *Infrared Phys. Technol.* **2013**, *61*, 94–100. [CrossRef]
56. Liu, Y.; Liu, S.; Wang, Z. A general framework for image fusion based on multi-scale transform and sparse representation. *Inf. Fusion* **2015**, *24*, 147–164. [CrossRef]
57. Zhang, Q.; Maldague, X. An adaptive fusion approach for infrared and visible images based on NSCT and compressed sensing. *Infrared Phys. Technol.* **2016**, *74*, 11–20. [CrossRef]
58. Ma, J.; Zhou, Z.; Wang, B.; Zong, H. Infrared and visible image fusion based on visual saliency map and weighted least square optimization. *Infrared Phys. Technol.* **2017**, *82*, 8–17. [CrossRef]
59. Aslam, A. Object Detection for Unseen Domains while Reducing Response Time using Knowledge Transfer in Multimedia Event Processing. In Proceedings of the ICMR 20 International Conference on Multimedia Retrieval, Dublin, Ireland, 8–11 June 2020.