Decision-level fusion detection for infrared and visible spectra based on deep learning

Tang Cong[1,2,3], Ling Yongshun[1,2,3], Yang Hua[1,2,3], Yang Xing[1,2,3], Lu Yuan[1,2,3]

(1. College of Electronic Countermeasure Institute, National University of Defense Technology, Hefei 230037, China;
2. State Key Laboratory of Pulsed Power Laser Technology, Hefei 230037, China;
3. Key Laboratory of Infrared and Low Temperature Plasma of Anhui Province, Hefei 230037, China)

Abstract: A fusion detection methodology for infrared and visible spectra was presented based on deep learning. First, a parameter transfer model for deep learning models was proposed. Then a pretraining model for infrared object detection was extracted from a visible object detection model based on deep learning and was fine-tuned on a collected infrared image dataset to obtain an infrared object detection model based on deep learning. On this basis, a decision-level fusion model for infrared and visible detection based on deep learning was established, and the model design, image registration and decision-level fusion processes were discussed in detail. Finally, an experiment comparing single-band detection and dual-band fusion detection during the daytime and nighttime was presented. Qualitatively, compared with the results of single-band detection, the confidences and bounding boxes achieved through dual-band fusion detection are superior, owing to the utility of their complementary information. Quantitatively, in the daytime, the mAP of dual-band fusion detection is 86.0% and is higher than those of infrared detection and visible detection by 9.9% and 5.3%, respectively; at nighttime, the mAP of dual-band fusion detection is 89.4% and is higher by 3.1% and 14.4%, respectively. The experimental results show that the dual-band fusion detection method proposed in this paper shows better performance and stronger robustness than the single-band object detection methods do, thus verifying the effectiveness of the proposed method.

Key words: object detection; decision-level fusion; dual band; deep learning

CLC number: TP391.4 **Document code:** A **DOI:** 10.3788/IRLA201948.0626001

基于深度学习的红外与可见光决策级融合检测

唐 聪[1,2,3],凌永顺[1,2,3],杨 华[1,2,3],杨 星[1,2,3],路 远[1,2,3]

(1. 国防科技大学电子对抗学院,安徽 合肥 230037;
2. 脉冲功率激光技术国家重点实验室,安徽 合肥 230037;
3. 红外与低温等离子体安徽省重点实验室,安徽 合肥 230037)

摘 要: 提出了一种基于深度学习的红外与可见光决策级融合检测方法。首先,提出了一种介于深度

收稿日期:2018-09-20; **修订日期:**2018-10-17

基金项目:国家自然科学基金(61405248,61503394);安徽省自然科学基金(1708085MF137)

作者简介:唐聪(1989-),男,博士生,主要从事计算机视觉、深度学习、模式识别等方面的研究。Email:tangcong_eei@163.com

导师简介:凌永顺(1937-),男,中国工程院院士,教授,博士生导师,主要从事光电工程等方面的研究。Email:lys@126.com

学习模型之间的参数传递模型，进而从基于深度学习的可见光物体检测模型上抽取了用于红外物体检测的预训练模型，并在课题组实地采集的红外数据集上进行 fine-tuning，从而得到基于深度学习的红外物体检测模型。在此基础上，提出了一种基于深度学习的红外与可见光决策级融合检测模型，并对模型设计、图像配准、决策级融合过程进行了详细地阐述。最后，进行了白天和傍晚条件下基于深度学习的单波段检测实验和双波段融合检测实验。定性分析上，由于波段之间的信息互补性，相比于单波段物体检测，双波段融合物体检测在检测结果上具有更高的置信度和更精确的物体框；定量分析上，白天时，双波段融合检测的 mAP 为 86.0%，相比于红外检测和可见光检测分别提高了 9.9%和 5.3%；傍晚时，双波段融合检测的 mAP 为 89.4%，相比于红外检测和可见光检测分别提高了 3.1%和 14.4%。实验结果表明：基于深度学习的双波段融合检测方法相比于单波段检测方法具有更好的检测性能和更强的鲁棒性，同时也验证了所提出方法的有效性。

关键词：物体检测； 决策级融合； 双波段； 深度学习

0 Introduction

Object detection is a significant focus of research in the field of computer vision[1-2], with applications in driverless cars, robotics, video surveillance and pedestrian detection[3-5]. In object detection, the utilization of multisensor data or information fusion can yield object detection results that cannot be achieved with a single sensor and can improve object detection performance[6]. In traditional fusion methods for object detection, fusion detection is commonly performed on the basis of infrared and visible spectra, making full use of the complementarity of visible and infrared images. At present, fusion detection for infrared and visible spectra relies primarily on traditional methods, such as multiresolution fusion[7], edge feature fusion[8] and Dempster-Shafer (DS) evidence theory[9]. However, few studies have investigated fusion detection for infrared and visible spectra based on deep learning. Moreover, at present, research on deep-learning-based object detection is primarily focused on the visible spectrum, while little research has been done on the infrared spectrum.

There are three methods of fusion detection for infrared and visible spectra: pixel-level fusion detection, feature-level fusion detection and decision-level fusion detection[10]. In pixel-level fusion detection, the images must be fused at the pixel level prior to object detection; in feature-level fusion detection, feature extraction must be performed before feature fusion, and object detection is then performed based on the fused feature vectors; in decision-level fusion detection, different sensors are independently used for object detection, and the detection results are then fused. Because deep learning models are data-driven models, when either of the first two fusion detection methods is adopted, a massive amount of image data is needed for model training. At present, visible datasets can be obtained from public sources; however, infrared datasets are difficult to obtain. Furthermore, the visible and infrared datasets should be acquired from the same scenes for the training of models for pixel-level and feature-level fusion detection, which makes it even harder to obtain suitable training data. By contrast, when decision-level fusion detection is employed, the visible spectrum detection model can be extracted from current models with no further additional training, and the infrared spectrum detection model can be obtained through fine-tuning[11] based on the visible spectrum detection model, thus significantly reducing the required amount of training data. Erhan et al. have conducted extensive simulations with the existing algorithms and have found that pretrained networks can learn qualitatively different features and perform better than traditionally trained networks do[12]. Fine-tuning is a crucial stage

in refining models to adapt them to specific tasks and datasets[13]. Therefore, we adopt the decision-level fusion approach and apply pretraining and fine-tuning to carry out research on fusion detection for infrared and visible spectra based on deep learning in this paper.

1 Infrared object detection model based on deep learning

At present, most object detection models based on deep learning are designed for visible detection and are not applicable for the detection of objects in infrared images or video sequences. Therefore, the design and training of an infrared object detection model based on deep learning must first be implemented in preparation for fusion detection.

The main difference between infrared object detection and visible object detection is that they are performed in different optical bands. However, they are both methods of object detection based on image data. Here, the results of applying a visible object detection model to a set of infrared pedestrian images are shown in Fig.1.

Fig.1 Infrared object detection with a visible object detection model (a), (b) and (c) represent three scenes

When the visible detection model is applied for object detection in the infrared images, the pedestrians in Fig.1(a) and Fig.1(b) can be successfully detected, demonstrating that a visible object detection model can achieve detection in infrared images to a certain extent. The reason for this capability may be that compared with visible images, infrared images have similar contour features, which serve as effective appearance features[14-15] for object detection. However, the classification scores of the pedestrians in Fig.1(a) and Fig.1 (b) are not very high, meaning that the visible object detection model has only a weak capability for infrared object detection. Moreover, one person is erroneously classified as a dog in Fig.1(b), while no person is detected in Fig.1(c); instead, in this last case, the model merely outputs a result classifying the entire image into the train class, with a score of 0.2.

Hence, the visible object detection model can achieve infrared object detection to some degree, but its detection accuracy is poor. Therefore, this experiment offers a theoretical and experimental basis for the possibility that a visible object detection model could be fine-tuned to obtain an infrared object detection model.

1.1 Pretraining model for infrared object detection based on deep learning

In general, pretraining models for visible object detection are established on the basis of models trained on the ImageNet dataset[16] or the Pascal VOC dataset[17]; one example is the Single Shot Detector (SSD) model[18]. When the SSD model(VGG VOC0712_ SSD_300 ×300_ iter_240000 . caffemodel) is directly adopted as the pretraining model for the development of our new model, because the classes in the classification layer of the new model are mismatched with those in the SSD model, the parameters of the classification convolution layers in the SSD model are unsuitable to be transferred to the new model. In this paper, it is assumed that the research objects of

interest for infrared detection are typical objects found at road intersections, including bicycles, buses, cars, motorbikes and people. Because the VOC dataset already contains bicycles, buses, cars, motorbikes and people, the relevant feature extraction weights of the model for these five classes and for background class (six object classes in all) can be extracted from the VOC-trained models. Meanwhile, as the object location is only related to the four coordinate value information, and is independent of the number of classes, it can still inherit the visible model to locate the object by parameter sharing. Therefore, the effective method of parameter transfer is essential to achieve a perfect pretraining model, which will rapidly complete the training and more easily acquire good detection performance. In addition, the classification layer of the model must be redesigned and combined with the convolution layers to obtain the pretraining model.

1.1.1 Parameter transfer model

In the design of the infrared pretraining model, the initial model is first initialized according to the parameters of the model structure. To make full use of the object detection ability of the visible model (VOC-trained SSD model), the parameters of the layers in the visible model are transferred to the convolution layer corresponding to the infrared model, which realizes the parameter sharing. The parameters transferred between the convolution layers include convolution kernel parameters and bias parameters, which are considered matrices. The parameter matrices of the visible model and infrared model are W_1 and W_2 respectively. W_1 represents a four-dimensional matrix of $(M\times r)\times c\times k\times k$, and W_2 represents a four-dimensional matrix of $(M\times r)\times c\times k\times k$, where, M and N represent the number of object classes of the visible model and the infrared model respectively, r represents the index of the default bounding boxes, c represents the channel number of the convolution layers, and k is the size of the convolution kernel. The biasing matrices of the visible model and infrared model are b_1 and b_2 respectively, and they are both one-dimensional matrices of $(4\times r)\times 1$.

In the parameter transfer, the relationship between W_1 and W_2 is

$$W_2[j+N\times r]=W_1[i+M\times r,\ \cdots] \tag{1}$$

In addition, the relationship between b_1 and b_2 is

$$b_2[j+N\times r]=b_1W_1[i+M\times r] \tag{2}$$

where, i and j represent the class index of visible model and infrared model respectively.

1.1.2 Implementation of parameter transfer

The pretraining infrared model designed in this paper is based on the SSD model structure. The convolution layers of which need to implement the parameter transfer: conv4_3_norm_mbox_conf, fc7_mbox_conf, conv6_2_mbox_conf, conv7_2_mbox_conf, conv8_2_mbox_conf, conv9_2_mbox _conf. The layers of conv4_3_norm_mbox_conf, conv8_2_ mbox_conf and conv9_2_mbox_conf have four anchors in the feature map, which means that the value of r can equal 0, 1, 2 and 3, and the anchor index set $R=\{0, 1, 2, 3\}$. The layers of conv4_3_norm_mbox_conf, conv8_2_mbox_conf and conv9_2_mbox_conf have six anchors in the feature map, which means that the value of r can be 0, 1, 2, 3, 4 and 5, and the anchor index set $R=\{0, 1, 2, 3, 4, 5\}$. The object classes chosen are bicycle, bus, car, motorbike, people and background, whose indexes in the classification layers of the visible model are 2, 6, 7, 14, 15 and 0, respectively, and whose indexes in the classification layers of the infrared model are 1, 2, 3, 4, 5 and 0, respectively. Therefore, the class index set I of the visible model is equal to {2, 6, 7, 14, 15, 0} and the class index set J of the infrared model is equal to {1, 2, 3, 4, 5, 0}. The parameter transfer process between two convolution layers of the two models is shown in algorithm 1.

By parameter transfer, the parameter sharing of the object location and recognition between pretraining infrared model and visible model has been realized, which allows the pretraining infrared model to have a certain capability for object detection.

Algorithm 1: PT (Parameter transfer)

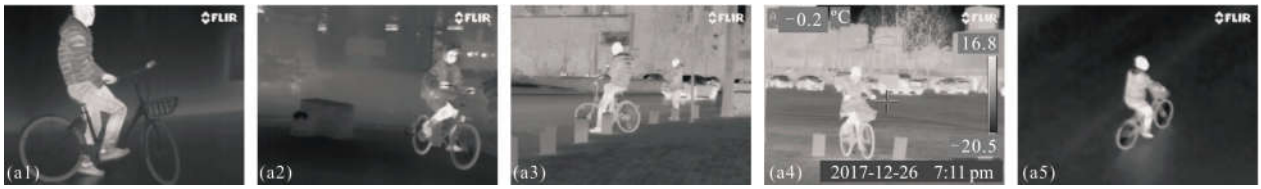
Input: Object class number M, object class number N, class index set I, class index set J, anchor index set R, convolution kernel parameters matrix W_1, and biasing parameters matrix b_1
Output: Convolution kernel parameters matrix W_2 and bias parameters matrix b_2
Initialize the convolution kernel parameters matrix and bias parameters matrix b_2 to all-zero matrices, $W_2=O$ $((M\times s)\times c\times k\times k)$, $b_2=O((4\times r)\times 1)$
Foreach r in R do
    Foreach $[i,j]$ in $\{[I_0,J_0],[I_1,J_1],\cdots\}$ do
        $W_2[j+N\times r,\cdots]=W_1[i+M\times r,\cdots]$
        $b_2[j+N\times r]=b_1[i+M\times r]$

1.2 Model training on an infrared image dataset

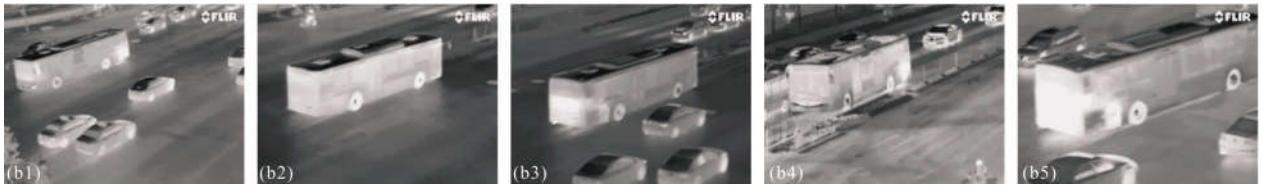
Once the pretraining model has been obtained, model training must be implemented on the collected infrared dataset. Our research team collected infrared images of objects in the above five classes on campus, in plazas, at road intersections and at pedestrian crosswalks, where many bicycles, buses, cars, motorbikes and pedestrians can be found. In total, we collected 1 800 infrared images, including 334 bicycle images, 205 bus images, 430 car images, 339 motorbike images and 492 pedestrian images. The infrared camera used was a FLIR T630.

1.2.1 Collection of the Infrared Image Dataset

Examples from the collected infrared dataset are shown in Fig.2.

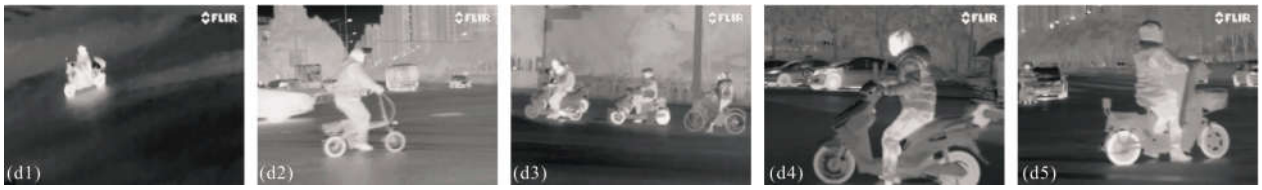


(a) Bicycle



(b) Bus

(c) Car



(d) Motorbike

(e) Person

Fig.2 Infrared dataset (examples)

1.2.2 Model training

After 3 000 training iterations, the training loss on the infrared dataset was approximately 2.5. The curve of the training loss versus the number of iterations is shown in Fig.3.

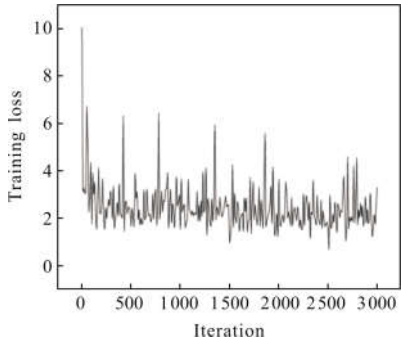


Fig.3 Curve of the training loss versus the number of iterations

During the training process, the accuracy was tested once every 500 iterations, and the curve of the test accuracy versus the number of iterations is shown in Fig.4.

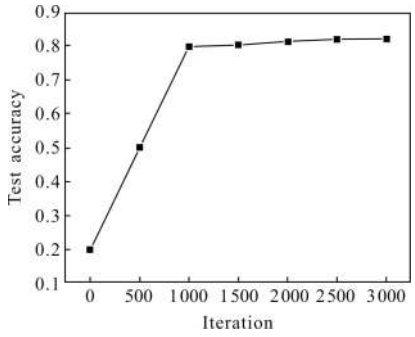


Fig.4 Curve of the test accuracy versus the number of iterations

The infrared object detection model obtained via fine-tuning was applied to the whole infrared test dataset for object detection. The average precision (AP) values for the bicycle, bus, car, motorbike and person classes and the mean average precision (mAP)[19-20] are shown in Tab.1.

Tab.1 mAP for the five classes on the infrared test datasets

mAP	AP				
	Bicycle	Bus	Car	Motorbike	Person
82.3%	78.8%	89.6%	86.4%	80.6%	75.8%

When the AP values are considered, it is seen that the infrared object detection model achieves the best detection performances for buses, cars, and motorbikes, with all of the APs for these classes being above 80%; this is due to the larger scale and better information response of these objects at all levels of the feature maps in the deep-learning-based detection model, which help to improve the precision and recall of object detection.

1.3 Detection performance comparison between the infrared and visible models on the datasets in different bands

The infrared object detection model obtained after fine-tuning still has some detection capability when applied to visible images. Here, the performances of the infrared and visible object detection models are compared on both the infrared and visible test datasets. The infrared test dataset is the test dataset from the abovementioned infrared dataset, while the visible test dataset was obtained by extracting the images of objects in the five classes of interest (bicycle, bus, car, motorbike and person) from the VOC test dataset. The results of the infrared and visible object detection models are compared in Tab.2.

Tab.2 Comparison of the results of infrared and visible object detection

Test dataset	mAP	
	Visible model	Infrared model
Infrared dataset	19.9%	82.3%
Visible dataset	78.8%	71.4%

The results in Tab.2 show that the mAP achieved on the infrared dataset by the infrared object detection model is much higher than that of the visible object detection model; meanwhile, the mAP on the visible test dataset is 71.4% and is just lower 7.4% than that of the visible object detection model, indicating that the model retains a certain capability for object detection in visible images.

2 Decision-level fusion model for infrared and visible detection based on deep learning

To enhance the object detection performance, the infrared and visible object detection models are combined to achieve dual-band object detection. Because the results of deep-learning-based object detection depend on dataset training, a massive amount of fused data is needed to retrain the model when pixel-level or feature-level fusion is adopted. For this reason, the decision-level fusion approach is adopted to merge the results of infrared and visible object detection.

2.1 Model design

According to the previous findings, although the infrared model after fine-tuning shows good detection performance on visible images, its performance is still weaker than that of the visible model. Hence, a good solution is to apply the infrared model for object detection in infrared images and the visible model for object detection in visible images and then perform decision-level fusion on the results. The decision-level fusion model is shown in Fig.5.

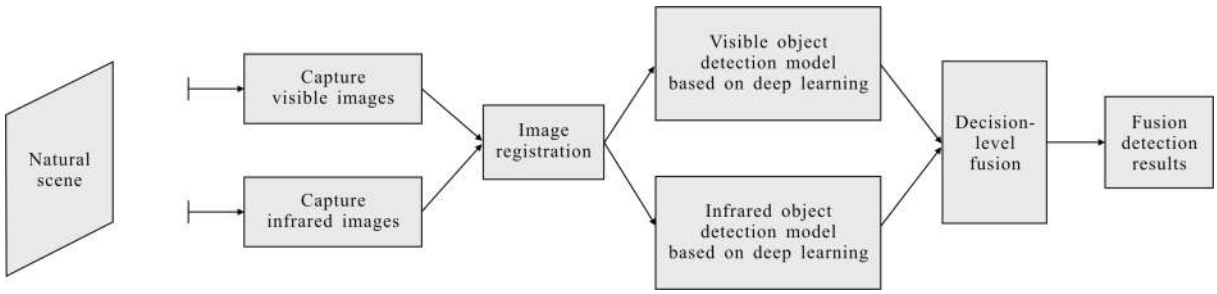


Fig.5 Decision-level fusion model for infrared and visible object detection based on deep learning

2.2 Image registration

Image registration is the process of overlaying two or more images of the same scene acquired from different viewpoints and/or by different sensors [21]. Consider an infrared camera and a visible camera that are used to capture images. Suppose that they are located side by side to ensure that their optical axes are parallel and that the distance between the centers of the camera lenses is 12 cm. In this case, for a scene that is located one hundred meters from the camera, the parallax between the cameras is negligible, and the only differences are those that result from the translation, rotation and scaling between the two images captured by the two cameras[22]. Thus, the differences between the images can be eliminated via affine transformation. The affine transformation between the benchmark image coordinates and the input image coordinates can be expressed as

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} e \\ f \end{bmatrix} \tag{3}$$

Here, (x,y) represents the benchmark image coordinates, and (x,y) represents the input image coordinates. The affine transformation between the two coordinate systems is

$$\begin{bmatrix} x_1' & x_2' & \cdots & x_n' \\ y_1' & y_2' & \cdots & y_n' \end{bmatrix} = \begin{bmatrix} a & b & e \\ c & d & f \end{bmatrix} \begin{bmatrix} x_1 & x_2 & \cdots & x_n \\ y_1 & y_2 & \cdots & y_n \\ 1 & 1 & \cdots & 1 \end{bmatrix} \tag{4}$$

In the above formula, $\begin{bmatrix} a & b & e \\ c & d & f \end{bmatrix}$ is an unknown coefficient matrix, whose elements can be calculated from the correspondence relationships between more than three pairs of pixels in the input and benchmark images.

With the notations $A = \begin{bmatrix} x_1' & x_2' & \cdots & x_n' \\ y_1' & y_2' & \cdots & y_n' \end{bmatrix}$, $B = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \\ y_1 & y_2 & \cdots & y_n \\ 1 & 1 & \cdots & 1 \end{bmatrix}$, and $P = \begin{bmatrix} a & b & e \\ c & d & f \end{bmatrix}$, the formula can be expressed as

$$A = PB \tag{5}$$

When both sides of the equation are simultaneously multiplied by B^{T}, the formula becomes

$$AB^{T}=PBB^{T} \tag{6}$$

When the pairs of pixels in the input and benchmark images do not form a straight line, BB^{T} is an invertible matrix, meaning that $(BB^{T})^{-1}$ exists. By simultaneously multiplying both sides of the equation by $(BB^{T})^{-1}$ and applying some transformations, P can be expressed as

$$P=AB^{T}(BB^{T})^{-1} \tag{7}$$

This formula shows how the coefficient matrix can be obtained from the coordinate information of more than three pairs of pixels in the input and benchmark images that do not form a straight line.

If the relative positions of the infrared and visible cameras remain unchanged, then the coefficient matrix will not change. Thus, an input image can be transformed into a new image with the same coordinate system as that of the benchmark image by taking the product of the input image and the coefficient matrix. Then, registration with the benchmark image can be achieved via the necessary interpolation in the transformed image.

2.3 Decision-level fusion

Decision-level fusion is performed on the basis of dual-band region merging, which removes object boxes with high overlap rates from the detection results of the visible and infrared models while preserving representative object boxes with high confidences as the final detection results, as shown in Fig.6.

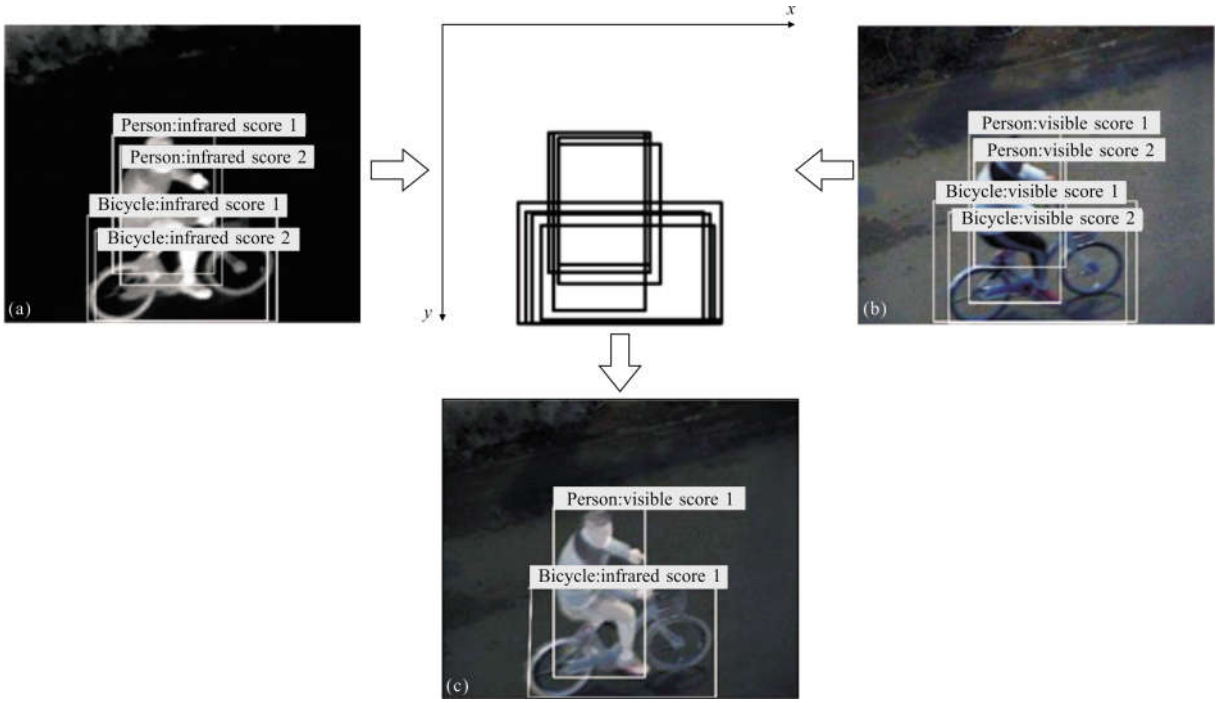


Fig.6 Process of decision-level fusion (a) visualization of the results of infrared detection; (b) visualization of the results of visible detection; (c) visualization of the results of dual-band region merging

(1) As illustrated by the examples of single-object detection shown in Fig.6(a) and Fig.6(b), the infrared model and the visible model each generate a certain quantity of candidate boxes with different confidences. All corresponding boxes are collected into a candidate box group, and the object box with the highest confidence is selected.

(2) The Intersection-over-Union (IoU) is calculated between the box with the highest confidence and each other box in the candidate box group to assess whether this value exceeds the suppression threshold. If the IoU is greater than the

suppression threshold, this means that the current box is suppressed by the box with the highest confidence and should be abandoned. At the end of each iteration of this operation, the object boxes whose IoUs do not reach the suppression threshold and the box with the highest confidence remain.

(3) The object box with the highest confidence is selected from among the boxes whose IoUs did not reach the suppression threshold in the last iteration, and step (2) is then repeated until the candidate box set is empty. Thus, all suppressed object boxes are rejected from the set of candidate object boxes, and the remaining object boxes are treated as the results of the fusion of the infrared and visible models, as shown in Fig.6(c).

The detail of dual-band region merging is shown in algorithm 2.

Algorithm 2: DRM (Dual-band region merging)

Input: Candidate bounding box set B_I of infrared model detection, B_v of visible model detection
Output: Region-merged object box set B_{DRM} of dual-band detection
Classify the candidate bounding boxes according to the label of them in the infrared and visible detection results and obtain the dual-band bounding box set of each class B_D^k, where k is the class index.
Obtain the set of dual-band bounding boxes B_D, $B_D=\{B_D^k\}$.
Foreach k in B_D do
  Obtain the set of the indexes I according to the order of candidate bounding boxes sorted by the score
  Initialize the region-merged object box set B_{DRM}^k, the index set P of B_{DRM}^k, overlap threshold $O_{threshold}$, $P=\phi$, $B_{DRM}^k=\phi$, $O_{threshold}=t(t>1)$
  While I is not null do
    Obtain the first index i of I: $i=I[0]$
    Append the last index i to P: $P\cup i$
    Foreach index n in I do
      $j=I[n]$
      Calculate the overlap $O(B_k(i),(B_k(j))$ using the theory of IoU
      If $O(B_k(i),(B_k(j))>O_{threshold}$ do
        Append the index n to the set S: $S=S\cup n$
    Remove the suppressing bounding box set S in I: $I=I/S$
Foreach p in P do
    Extract the object bounding boxes B_{DRM}^k: $B_{DRM}^k=B_{DRM}^k\cup B_D^k[p]$
Obtain the Region-merged object box set B_{DRM}: $B_{DRM}=\{B_{DRM}^k\}$

2.4 Theoretical analysis on fusion detection performance

The mAP is the average AP across all classes, and the calculation of the AP is directly related to the area under the precision-recall curve; therefore, the mAP value is also directly related to the area under the precision-recall curve. The formulas for calculating the recall and precision are as follows:

$$R=\frac{N_{TP}}{N_{TP}+N_{FN}}=\frac{N_{TP}}{N} \tag{8}$$

$$P=\frac{N_{TP}}{N_{TP}+N_{FP}}=\frac{1}{1+\frac{N_{FP}}{N_{TP}}} \tag{9}$$

where TP denotes the true positives and represents the detection results that are correctly classified as belonging to the ground-truth class and whose IoUs are higher than the IoU threshold, FP denotes the false positives and represents the detection results that are incorrectly classified as belonging to the ground-truth class and whose IoUs are higher than the IoU threshold, FN denotes the false negatives and represents the detection results that are incorrectly classified as belonging to other classes or whose IoUs are lower than the IoU threshold, and N denotes the number of ground-truth objects in the scene.

To compare the precision-recall curves of single-band detection and fusion detection, the precision values corresponding to a given recall are considered. According to equations (8) and (9), when the R values for detection are the same, the N_{TP} values will also be the same; therefore, the value of P is determined only by N_{FP}, and the smaller N_{FP} is, the larger P will be. Consequently, when fewer objects have IoUs below the IoU threshold, better detection results will be obtained. According to the definition of N_{FP}, N_{FP} represents the number of incorrect detection results. Generally, because fusion detection integrates the detection results from two bands, fusion detection will yield fewer incorrect detection results than single-band detection does, which means that fusion detection is expected to achieve a higher mAP value and better performance than those of single-band detection.

3 Fusion detection experiments, results and analysis

Visible images reflect the reflection characteristics of objects and infrared images reflect their radiation characteristics. Generally, the imaging contents acquired according to the reflection characteristic and radiation characteristic of an object and its background are not completely consistent; therefore, visible images and infrared images exhibit complementarity. Infrared object detection is more advantageous at night, while visible object detection behaves better in the daytime, especially for objects without obvious infrared features. Thus, the decision-level fusion of the information from the visible and infrared bands will increase the probability of successful object detection.

We collected images of pedestrians and cars at an intersection in Hefei city with an infrared camera and a visible camera. The time periods for collection were 13:00 –15:00 (daytime) and 18:00 –19:00 (nighttime). We collected 105 pairs of infrared and visible images during both daytime and nighttime.

Firstly, infrared images and visible images need registration operation, the results of which are shown in Fig.7 and Fig.8.

(a) Infrared image (b) Visible image (c) Visible image after registration

Fig.7 Image registration of infrared and visible bands (daytime)

(a) Infrared image (b) Visible image (c) Visible image after registration

Fig.8 Image registration of infrared and visible bands (nighttime)

From Fig.7 and Fig.8, it can be seen that there are black block regions in the registered visible image whose contents are not exactly the same with the infrared images. Therefore, region capture needs to be implement to reconcile the infrared images and visible images. For the convenience of capturing, the infrared images and visible images are all added by 0.5 weight, as shown in Fig.9 (a). After the common region of infrared image and registered visible image is captured, a pair of dual band images with the same region are obtained, which are shown in Fig.9(b) and Fig.9(c).

(a) Fusion image before capturing

(b) Infrared image

(c) Visible image

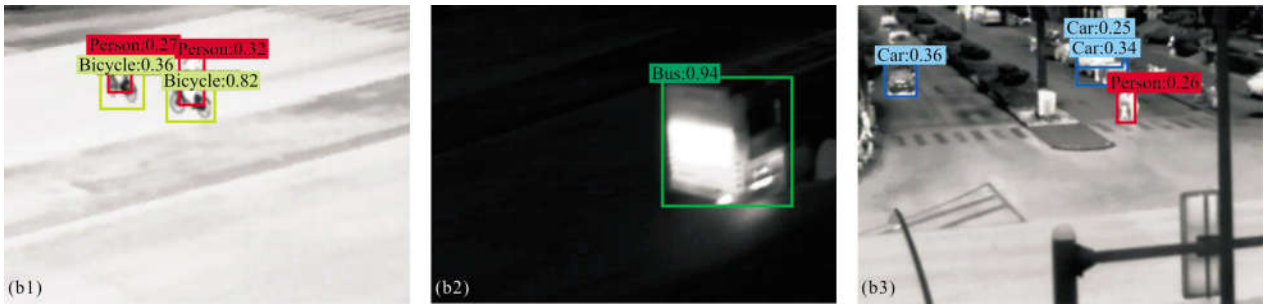
Fig.9 Capturing of infrared and visible images

3.1 Qualitative analysis

After the collection of the dual-band dataset of infrared and visible spectra, the infrared and visible object detection models were applied to detect objects in the infrared and visible images, respectively. Meanwhile, the decision-level fusion results between the two bands are implemented. Examples of the results are visualized in Fig.8 and Fig.10.

Figure 10 and Fig.11 demonstrate the following phenomena: (1) Infrared object detection and visible object detection are complementary, and this complementarity is fully exploited in the results of dual-band fusion detection; (2) Compared with the results of single-band detection, the confidences and

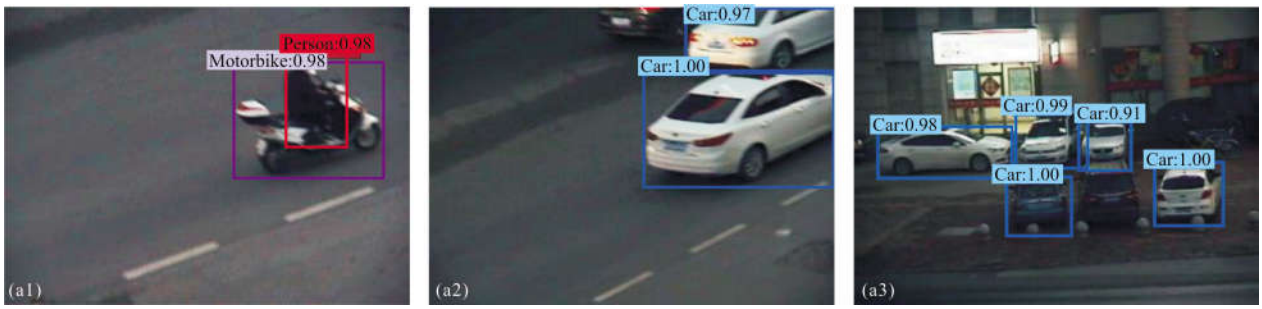
(a) Visible object detection



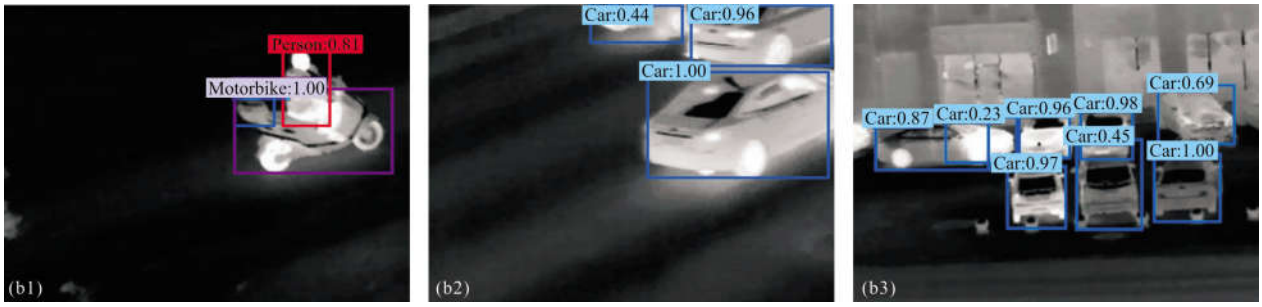
(b) Infrared object detection

(c) Fusion detection

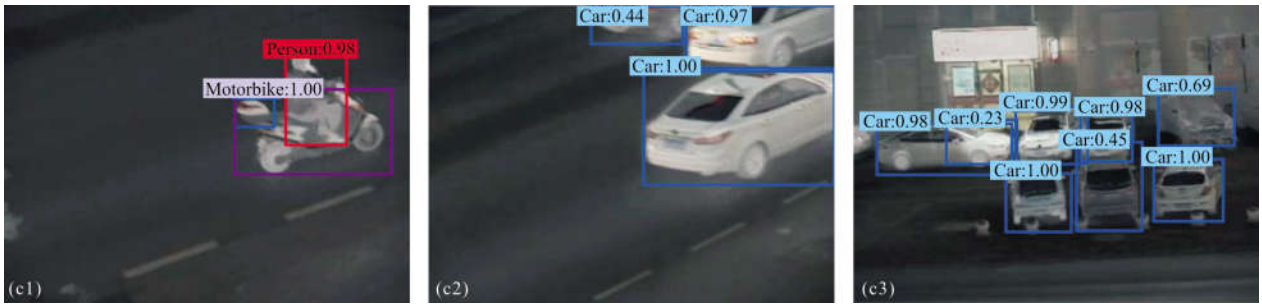
Fig.10 Visualization of the results of single-band and dual-band detection (daytime)



(a) Visible object detection



(b) Infrared object detection



(c) Fusion detection

Fig.11 Visualization of the results of single-band and dual-band detection (nighttime)

bounding boxes achieved through dual-band fusion detection are superior. These will allow dual-band fusion to detect more objects and locate more accurate than infrared and visible models, which increases the true positives in the detection results. According to equation (8) and (9), the precision-recall curve of dual-band fusion will more ideal than those of infrared and visible models, which allows the AP of dual-band fusion to be higher than those of infrared and visible models. Thus, dual-band fusion detection has obvious advantages over single-band detection.

3.2 Quantitative analysis

Table 3 and Tab.4 present the object detection results and the corresponding mAP evaluations for the two different time periods.

Tab.3 Comparison of single-band and dual-band object detection (daytime)

Detection	mAP	AP				
		Bicycle	Bus	Car	Motorbike	Person
Infrared model	76.1%	89.3%	72.5%	79.2%	73.0%	66.7%
Visible model	80.7%	84.3%	88.0%	90.2%	65.9%	75.1%
Fusion model	86.0%	92.0%	88.2%	91.0%	77.1%	81.5%

Tab.4 Comparison of single-band and dual-band object detection (nighttime)

Detection	mAP	AP				
		Bicycle	Bus	Car	Motorbike	Person
Infrared model	86.8%	92.0%	99.7%	76.9%	88.6%	76.7%
	86.3%	92.0%	99.7%	76.9%	–	76.7%
Visible model	68.9%	82.7%	91.5%	64.7%	44.5%	61.2%
	75.0%	82.7%	91.5%	64.7%	–	61.2%
Fusion model	87.1%	95.7%	99.8%	81.2%	77.9%	80.7%
	89.4%	95.7%	99.8%	81.2%	–	80.7%

For daytime detection, both the infrared detection model and the visible detection model perform well. In the daytime experiment, the mAPs of the infrared and visible detection models are 76.1% and 80.7%, respectively, while the mAP of the fusion detection model is 86.0%. And the results show that although the mAP of the visible detection model is higher than that of the infrared detection model, the APs of the visible detection model are not higher than those of the infrared detection model in all the classes. In this case, the advantages of fusion detection compared to single-band detection are apparent. For nighttime detection, because the infrared detection model is not affected by darkness, its mAP is 86.8%. However, the mAP of the visible detection model is only 68.9% due to the dimness, blurring, and noise in the visible images. When we apply the fusion detection model, because of the poor detection performance of the visible object detection model for motorbikes, the mAP is 87.1%, which is nearly the same as that of the infrared detection model. In contrast to the motorbike results, the APs of the visible model for the nighttime detection of bicycles, buses, cars, and people are also lower than those of the infrared model, but they are all over 60.0%, which indicates that the detection performance remains acceptable. When motorbike detection is not included, the mAPs of the infrared and visible detection models are 87.1% and 75.0%, respectively, while the mAP of the fusion detection model is 89.4%. Under these conditions, the performance of the fusion detection model is still better than that of the infrared object detection model, making it the best of the three. And the results show that although the APs of the visible detection model are lower than those of the infrared detection model in all the classes, the mAP of fusion model is the highest in the three detection models. The reason is that the visible model can detect some objects which can´t detect by the infrared model, which is helpful to increase the true positives in the fusion results. However, if the detection performance of single model is too poor, much more false negatives will be obtained and the performance of fusion detection will become worse, such as the motorbike in Tab.4.

Therefore, the detection performance of fusion model is determined to some extent by those of infrared and visible models. As long as the detection performance of the single-band model is not too poor, dual-band decision-level fusion detection has an obviously beneficial effect.

4 Conclusions

In this paper, we proposed a fusion detection methodology for infrared and visible spectra based on deep learning. The visible spectrum detection model can be extracted from current models with no further additional training, and the infrared spectrum detection model can be obtained through fine-tuning based on the visible spectrum detection model, thus significantly reducing the required amount of training data.

As few researches on deep-learning-based object detection has been done on the infrared spectrum, we need to establish an infrared object detection model based on deep learning. First, we proposed a parameter transfer model for deep learning models. Then a pretraining model for infrared object detection based on deep learning was first obtained by extracting the shared weights of the five specified object classes and background class from a visible

object detection model based on deep learning, and redesigning the classification layer of the model. Then the model was fine-tuned on a collected infrared image dataset with the five object classes to obtain an infrared object detection model based on deep learning. The fine-tuned infrared model demonstrated a good object detection performance in infrared images and even retained a certain capability for object detection in visible images. However, the detection performance of the fine-tuned infrared detection model was poorer than that of the visible detection model on some scenes, especially under good lighting conditions. To take advantage of their complementary information, a fusion detection model for infrared and visible spectra based on deep learning was established. Experimental results demonstrate that the dual-band decision-level fusion detection model proposed in this paper offers better performance and stronger robustness than the single-band detection models do. Therefore, the approach proposed in this paper offers an excellent solution for improving the dual-band fusion detection capabilities of deep learning models.

In addition, the applicability of the proposed approach is not restricted to two bands; the model can be applied for the combination of detection results from multiple bands using deep learning models. Thus, the proposed approach can enable the full use of the complementarity of multiple bands in object detection and further improve the performance and robustness of deep learning models.

References:

[1] Erhan D, Szegedy C, Toshev A, et al. Scalable object detection using deep neural networks [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014: 2147-2154.

[2] Luo Haibo, Xu Lingyun, Hui Bin, et al. Status and prospect of target tracking based on deep learning [J]. ***Infrared and Laser Engineering***, 2017, 46(5): 0502002. (in Chinese)

[3] Wei P, Ball J E, Anderson D T. Fusion of an ensemble of augmented image detectors for robust object detection [J]. ***Sensors***, 2018, 18(3): 894.

[4] Jeong Y N, Son S R, Jeong E H, et al. An integrated self-diagnosis system for an autonomous vehicle based on an IoT gateway and deep learning [J]. ***Applied Sciences***, 2018, 7: 1164.

[5] Tian Y, Luo P, Wang X, et al. Deep learning strong parts for pedestrian detection[C]//IEEE International Conference on Computer Vision, 2015: 1904-1912.

[6] Hall D L, Llinas J. An introduction to multisensor data fusion[C]//Proceedings of the IEEE, 1997, 85(1): 6-23.

[7] Petrovic V S, Xydeas C S. Gradient-based multiresolution image fusion [J]. ***IEEE Transactions on Image Processing***, 2004, 13(2): 228-237.

[8] Davis J W, Sharma V. Background-subtraction using contour-based fusion of thermal and visible imagery [J]. ***Computer Vision and Image Understanding***, 2007, 106(2): 162-182,.

[9] Zeng D, Xu J, Xu G. Data fusion for traffic incident detection using DS evidence theory with probabilistic SVMs [J]. ***Journal of Computers***, 2008, 3(10): 36-43.

[10] Fendri E, Boukhriss R R, Hammami M. Fusion of thermal infrared and visible spectra for robust moving object detection [J]. ***Pattern Analysis & Applications***, 2017, 20(10): 1-20.

[11] Guo Y, Liu Y, Oerlemans A, et al. Deep learning for visual understanding: A review[J]. ***Neurocomputing***, 2016, 187(C): 27-48.

[12] Erhan D, Bengio Y, Courville A, et al. Why does unsupervised pre-training help deep learning?[J]. ***Journal of Machine Learning Research***, 2010, 11(3): 625-660.

[13] He K, Sun J. Convolutional neural networks at constrained time cost [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 5353-5360.

[14] Shotton J, Blake A, Cipolla R. Contour-based learning for object detection [C]//Proceedings of the IEEE Conference on Computer Vision, 2005, 1: 503-510.

[15] Shen W, Wang X, Wang Y, et al. Deepcontour: A deep convolutional feature learned by positive -sharing loss for contour detection [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 3982-3991.

[16] Russakovsky O, Deng J, Su H, et al. Image net large scale visual recognition challenge [J]. ***International Journal of Computer Vision***, 2014, 115(3): 211-252.

[17] Vicente S, Carreira J, Agapito L, et al. Reconstructing PASCAL VOC [C]//Proceedings of the IEEE Conference on

Computer Vision and Pattern Recognition, 2014: 41-48.

[18] Liu W, Anguelov D, Erhan D, et al. SSD: Single shot multibox detector [C]//Proceedings of European Conference on Computer Vision, 2016: 21-37.

[19] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. ***IEEE Transactions on Pattern Analysis & Machine Intelligence***, 2017, 39(6): 1137-1149.

[20] Tang Cong, Ling Yongshun, Zheng Kedong, et al. Object detection method of multi-view SSD based on deep learning [J]. ***Infrared and Laser Engineering***, 2018, 47(1): 0126003. (in Chinese)

[21] Zitová B, Flusser J. Image registration methods: a survey[J]. ***Image & Vision Computing***, 2003, 21(11): 977-1000.

[22] Heather J P, Smith M I. Multimodal image registration with applications to image fusion [C]//Proceedings of the IEEE International Conference on Information Fusion, 2005: 372-379.