Thermal Infrared Salient Human Detection Model Combined with Thermal Features in Airport Terminal

*YU Yuecheng**, *LIU Chang*, *WANG Chuan*, *SHI Jinlong*

School of Computer, Jiangsu University of Science and Technology, Zhenjiang 212100, P.R. China

(Received 27 September 2021; revised 11 February 2022; accepted 18 April 2022)

Abstract: Target detection in low light background is one of the main tasks of night patrol robots for airport terminal. However, if some algorithms can run on a robot platform with limited computing resources, it is difficult for these algorithms to ensure the detection accuracy of human body in the airport terminal. A novel thermal infrared salient human detection model combined with thermal features called TFSHD is proposed. The TFSHD model is still based on U-Net, but the decoder module structure and model lightweight have been redesigned. In order to improve the detection accuracy of the algorithm in complex scenes, a fusion module composed of thermal branch and saliency branch is added to the decoder of the TFSHD model. Furthermore, a predictive loss function that is more sensitive to high temperature regions of the image is designed. Additionally, for the sake of reducing the computing resource requirements of the algorithm, a model lightweight scheme that includes simplifying the encoder network structure and controlling the number of decoder channels is adopted. The experimental results on four data sets show that the proposed method can not only ensure high detection accuracy and robustness of the algorithm, but also meet the needs of real-time detection of patrol robots with detection speed above 40 f/s.

Key words: thermal infrared image; human body detection; saliency; thermal features; lightweight model

CLC number: TN391 **Document code:** A **Article ID:** 1005-1120(2022)04-0434-16

0 Introduction

With the vigorous development of civil aviation, airport terminal safety patrol has gradually become one of the important work contents to ensure airport safety. The existing safety patrol mode of airport terminal is mainly manual patrol. At the same time, it is often necessary to supplement the central control room with information technology means such as patrol personnel positioning and video monitoring. In this way, airport managers must face some problems, including the increase of labor cost and labor intensity. In addition, due to the reason that patrol inspection in the airport terminal is mainly performed by people, the staff's sense of responsibility has become one of the key factors affecting the patrol inspection effect. If the staff are distracted in the process of patrol inspection, it is very easy to cause potential safety hazards in the airport terminal. In recent years, patrol robots has been widely used in many fields, including power patrol, hazy weather detection[1] and intruder detection[2]. Therefore, it has become an inevitable trend of the development of intelligent security technology to use robots to carry out patrol inspection in airport terminal.

In order to meet the needs of night patrol of airport terminal, patrol robots should have the ability to accurately identify pedestrians in low light or even no-light environment. Therefore, it is necessary to select the thermal infrared camera as the monitoring camera of the patrol robot. Essentially, the core of robot patrol is the process of human body detection on the image captured by the camera. Traditional human body detection algorithms mostly re-

*Corresponding author, E-mail address: zhjyuyuecheng@163.com.

How to cite this article: YU Yuecheng, LIU Chang, WANG Chuan, et al. Thermal infrared salient human detection model combined with thermal features in airport terminal[J]. Transactions of Nanjing University of Aeronautics and Astronautics, 2022, 39(4): 434-449.

http://dx.doi.org/10.16356/j.1005-1120.2022.04.006

ly on manual features, such as histograms of oriented gradients (HOG)[3], integral channel features (ICF)[4] and deformable part model (DPM)[5]. Certain effects can be achieved by using such methods in visible light scenes. However, it is often difficult to achieve good results if it is directly used in the thermal infrared images environment. The reason is that compared with visible light images, thermal infrared images have many disadvantages, including lack of texture features, blurred visual effects, low resolution and low signal-to-noise ratio.

In recent years, some researchers have proposed to extend the saliency detection method to human body detection in thermal infrared scenes. According to the features of low contrast and high noise between target and background in infrared images, an associated saliency based visual attention model was proposed[6]. In this method, the associative saliency generated from region saliency and edge contrast is used to improve the accuracy and robustness of infrared target segmentation. Similarly, based on the visual attention mechanism, an infrared image-based saliency extraction algorithm is also proposed[7]. Using pedestrian brightness and appearance characteristics, a pedestrian detection method in infrared images is implemented by significant propagation between design domains[8].

However, only the shallow features of the image are used in the above methods. In order to complete the detection task, the above traditional methods need to design manual features for a class of targets in a specific scene and extract valid features from the image by using manual features. Although features can be quickly extracted from images manually, these features fail to cope with the misrecognition caused by other factors such as changes in human pose and occlusion in complex scenes. In recent years, deep neural network models have been proposed for target detection. Deep neural networks perform end-to-end learning through multi-layer neural networks, and can directly use training samples to deeply mine the potential features of data. For this reason, the feature representation of different human postures in complex scenes can be obtained by self-learning using deep neural network models, which effectively avoids the shortcomings of traditional manual design features.

Human body detection methods based on deep neural network generally include two types, namely target detection and image segmentation. For target detection methods, commonly used models mainly include R-CNN[9] and YOLO[10]. This type of method regards target detection as a region detection problem, and ensures high accuracy of the algorithm by performing two tasks of classification and positioning at the same time. When the YOLOv3 model[11] is directly applied to the thermal infrared scene, the detection and positioning of the human body in the thermal infrared images can be realized. However, there are obvious defects in the target detection model. On one hand, it is prone to miss the detection behavior, on the other hand, it reguires high computing resources. For the patrol task of the terminal building, the human detection task can only be achieved by the limited computing resources of the patrol robots, and the target missed detection is not allowed. It can be seen that the target detection algorithm represented by YOLOv3 is not suitable for the actual needs of terminal patrol task.

Image segmentation model is another widely used human body detection method, in which pixels are its detection units. When it is applied to a thermal infrared scene, the edge pixels may be incorrectly detected, and the segmented target may be incomplete. Nevertheless, most areas of suspicious targets can still be correctly detected by image segmentation, which will not affect the recognition of suspicious targets. Therefore, it is still possible to effectively avoid the occurrence of missed detection. In addition, the requirements of the image segmentation model for computing resources are much lower than that of the target detection model. Existing research work shows that CNN can be used to design pixel wise classifier in thermal infrared images[12]. As a fully connected neural network, U-Net model[13] can be used for rapid and accurate detection of human targets in thermal infrared images. In fact, as an image segmentation method, the saliency detection model based on the U-Net network not only can detect most saliency human body objects in

different application scenarios, but also can be adapted to patrol robot platforms with limited computing resources. This provides a feasible technical framework for human body detection during airport terminal inspections.

However, when using image segmentation method to detect human body in thermal infrared images, there are still some objective factors that affect target recognition in the airport terminal. First of all, pedestrians in the terminal have various postures such as standing, walking, sitting, and squatting. At the same time, fixed objects including seats, beams and pillars in the airport terminal will also form a partial occlusion of pedestrian targets. Secondly, when using thermal infrared cameras to obtain target images, thermal sources such as light sources and display screens in the airport terminal will also be imaged in the thermal infrared images. Furthermore, when the patrol robot detects the surrounding environment from a horizontal perspective, the human body will show different scales in the image due to the different distances from the camera.

Therefore, in order to reduce the occurrence of missed detection and cope with the adverse factors such as multi-posture, multi-scale, local occlusion and thermaling source interference, we propose a novel thermal infrared image saliency detection algorithm based on U-Net model in this paper. We still adopt U-net as the architecture of our method. But its encoder is replaced by a VGG network[14], and the fusion module is added to the decoder. In this way, after the overlay convolution operation, the saliency decoder feature map contains both the thermal features and the saliency features of the detection target. Furthermore, when designing the loss function of the final prediction map, the weight of the pixels in the high temperature area of the thermal map increases. As a result, the algorithm becomes more sensitive to high temperature areas in the image to reduce the adverse effects of various interference factors in human body detection of the airport terminal. Finally, by simplifying the VGG network structure and controlling the number of decoder channels, the model realizes lightweight improvement and can be better adapted to the patrol robots with limited computing resources.

Our contributions can be summarized as follows:

(1) We adopt VGG network as the encoder and improve the decoder mechanism of U-Net network. Consequently, the adaptability of the model to the night scene of the airport terminal is improved by the effective fusion of the thermal features and saliency features of the detection target.

(2) We design a learning method that uses the saliency map and the thermal map to train the saliency branch and the thermal branch in the decoder respectively. Furthermore, by redesigning the loss function of the final prediction map, the accuracy of the algorithm for human body target detection is improved.

(3) By simplifying the VGG network structure and controlling the number of decoder channels, the complexity of our model is reduced, and as a result, the demand for computing resources of our algorithm is also reduced.

1 Related Work

1.1 Thermal features in thermal infrared images

Because the temperature of the human body is usually higher than that of surrounding objects[15], thermal features become one of the most efficient features to characterize the human body in thermal infrared images. In addition, the thermal features can be easily extracted, so they are widely used in various algorithms of human body detection. The main factors affecting the gray value of the object in thermal infrared images are temperature and radiation[16], which has nothing to do with the lighting conditions. When the gray value of the object is larger, it means that the temperature of the object is relatively higher. Therefore, extracting thermal features in thermal infrared images based on gray values has become one of the most important methods in thermal infrared image detection tasks. However, the objects with obvious thermal features in the image are sometimes not only the human body, but al-

so various devices such as light sources and display screens. These interfering factors have brought challenges to human detection tasks.

In order to cope with the above-mentioned unfavorable factors, the combination of thermal features and other human body features has become one of the methods of human body detection. Fernández-Caballero et al.[15] proposed a thermal-infrared pedestrian ROI extraction method by fusing thermal features and dynamic information. Zheng et al.[17] proposed an infrared human body detection method based on saliency propagation, in which both the thermal and appearance features of the human body are used. In addition, strengthening and highlighting the human body area in the infrared image is another effective method to overcome external interference factors. Mi et al.[18] proposed a method to highlight the human body part in the thermal infrared images by enhancing the thermal contrast between the human body and the background. From the perspective of the highlighting thermal feature distribution and gradient features, Lu et al.[19] proposed a saliency detection method for infrared images based on the contrast and the distribution.

Inspired by the above methods, and taking into account the fact that the body temperature of the human body is often higher than the ambient temperature, we use the thermal map extracting from the thermal infrared images to improve the robustness of algorithm to external interference.

1.2 Saliency detection based on U-Net network

Saliency detection is a task to segment the most visually distinctive object or region from the image. Early saliency detection methods mostly use manual features that rely on image contrast. However, such methods are not applicable when the contrast between the target and the background is lower. In recent years, the wide application of the deep neural network breaks the limitation of insufficient accuracy of manual features in low contrast images. Among them, the U-Net network[13] is one of the most popular methods. It can perform pixel-level segmentation on the input image according to the encoder-decoder architecture. For the U-Net network, the features of the same size in the encoder and decoder are merged by superposition. The process of fusion of these features plays an important role in combining contextual information, making the U-Net network still has good results when facing images with insufficient contrast. As a result, the image output of the network also has higher quality than that of earlier models.

Nevertheless, when the U-Net network is used for complex target detection, there will still be the defect of the poor edge pixel effect. Luo et al.[20] designed a network structure that can integrate local and global features to improve the performance of significant regions by penalizing the loss of boundary errors. Similarly, BASNet is a network model based on U-Net. With the help of the hybrid loss function of pixel-level, patch-level and map-level, a boundary-aware salient object detection method is realized[21]. Considering the relationship between contour and saliency, Zhou et al.[22] extended the decoder into two branches consisting of a saliency branch and a contour branch. In this way, the detection effect of edge pixels can be improved by learning the association between the saliency map and the contour map. In addition, introducing context feature information into the model is also one of the effective ways to improve the accuracy of the algorithm[23-24]. For such methods, attention mechanisms are usually used to design methods that can select and fuse multi-level contextual information. It can be seen that when designing a saliency detection method based on the U-net network, it is often necessary to improve the original network architecture according to the actual situation of the application scenario.

Inspired by the above work, we will improve the architecture of the U-Net network. Then the thermal features contained in the thermal map will be used to improve the segmentation performance of the salient human body. Furthermore, by lightening the network model, the algorithm can efficiently run in the patrol robot system with limited computing resources.

2 Salient Human Detection Model Combined with Thermal Features

2.1 Framework

Based on the fact that the human body temperature is often higher than the temperature of the surrounding environment, salient human detection model combined with thermal features (TFSHD) is proposed in this paper. The framework of our method is demonstrated in Fig.1, where E_i represents the encoder feature map, A_i the feature map output of the embedded module, D_i the original decoder feature map, and S_i the saliency decoder feature map. The fusion module includes the thermal branch $H^{(i)}$ and saliency branch $S^{(i)}$, respectively. In addition, S_0 indicates the final output image obtained after up-sampling all the salient decoder feature maps. Similar to the traditional U-Net network, our neural network model is still an encoder-decoder architecture. However, in order to effectively make use of the thermal feature information, VGG network is adopted as the encoder, and the decoder of the original U-Net network is expanded. The expanded decoder consists of three modules, namely the original decoder module, the saliency decoder module and the fusion module. Among them, the fusion module contains two parts, i.e. the thermal branch $H^{(i)}$ and the saliency branch $S^{(i)}$.

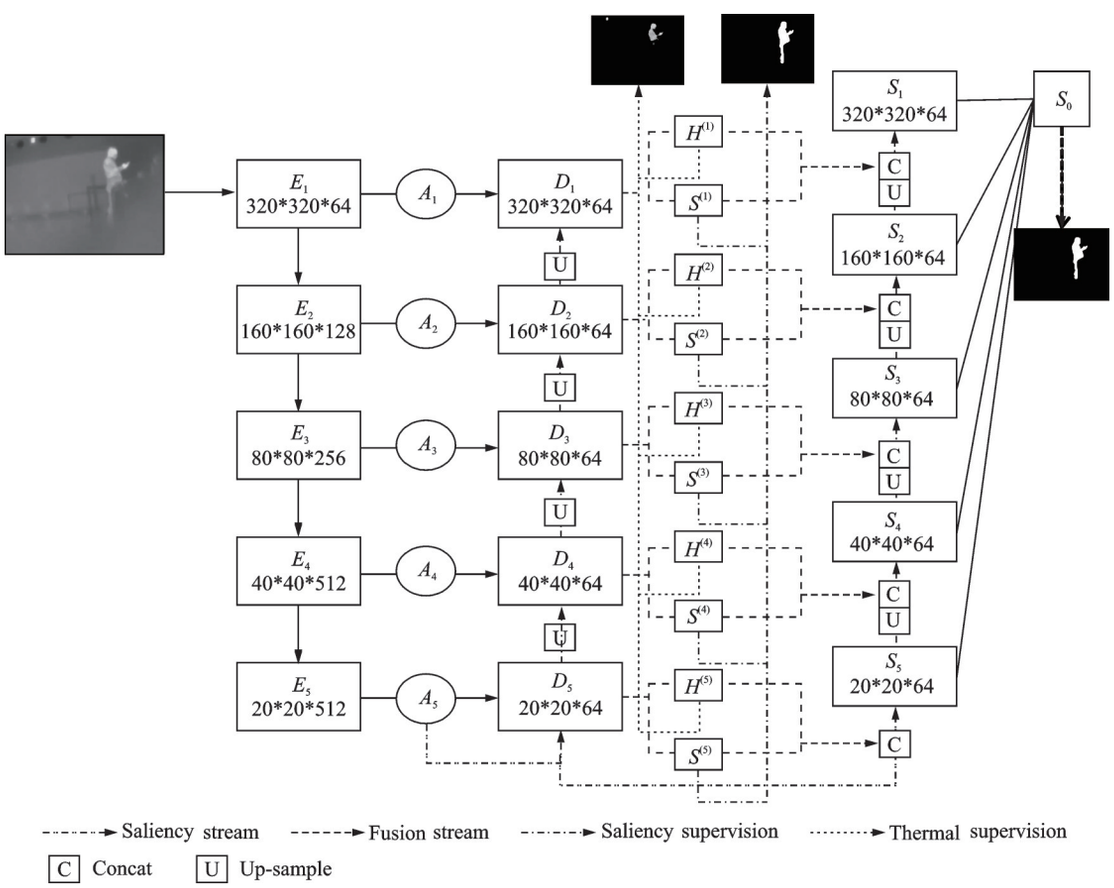

Fig.1 Framework of TFSHD

2.2 Extraction of thermal maps

As described in Section 2.1, the information of the thermal feature is beneficial to significantly improve the detection performance of the human body. For this reason, the widely used thermal map is used as the original representation of thermal features in our paper. For thermal infrared cameras, temperature and radiation are the main factors forming thermal infrared images. Correspondingly, the temperature information in the thermal infrared images is mainly reflected in the gray value of the pixel. The larger the gray value, the higher the temper-

ature of the pixel. In the airport terminal at night, the temperature of the human body is relatively high compared with the surrounding environment. It can be inferred that the area with a higher gray value in thermal infrared images is likely to be the target area of the human body. Therefore, as shown in Fig.2, we obtain the thermal map by segmenting the high gray value region from the input image.

Fig.2 Input images on the left and extracted thermal maps on the right

However, there are not only pedestrians, but also other high temperature sources in the robot working scene. In addition, the gray value of human targets in thermal infrared images will also be affected due to factors such as clothing, distance from the camera, etc. Thus, the region with medium or high gray value in the thermal infrared image is selected as the thermal map. Similar to the method in Ref.[15], the segmentation method based on gray threshold is used to segment objects and regions with thermal features in the image.

$$\theta_{TA}=\frac{5}{4}(\bar{I}+\sigma_{\mathrm{I}}) \tag{1}$$

where the gray threshold θ_{TA} is calculated from the standard deviation σ_{I} and the average value $\bar{I}$ of the input image I.

2.3 Encoder module

In the U-Net network, the main function of the encoder module is to extract feature maps of different scales from the images. We choose VGG neural network as the feature extraction network of the encoder module. When the VGG network is used for feature extraction, the extracted feature maps of different scales will show a progressive state. This provides a variety of deep and shallow information for the later feature integration. In addition, VGG neural network is an encoder module in TFSHD model, whose purpose is to extract the feature map of the input image. Therefore, in order to make the model lightweight, the full-connection layer, which is the last part of the VGG network, is discarded. In fact, only the first five different scales of encoder feature maps extracted by the VGG neural network is used in our method. As shown in Fig.1, the feature map of the encoder extracted by the VGG network is represented by E_i, where $i=1,2,\cdots,5$.

It should be noted that although there are more lightweight feature extraction networks including Shufflenet[25], this type of network is usually more suitable for small neural networks. If we directly select such a lighter-weight network for feature extraction, when the decoder module of the network model is up-sampling, it will often cause serious distortion because there is not enough context information in the feature map.

2.4 Lightweight operation

The computing resources of patrol robots are very limited, but it has high requirements for the real-time performance of the target detection. For this purpose, when designing the TFSHD model, the process of model lightweight must be considered. Model lightweight is an effective way to reduce the computing resource requirements of the algorithm and improve the operating efficiency of the algorithm. By analyzing the structure of the TFSHD model, we can see that there are two main ways to realize the purpose of model lightweight. One way is the simplified processing of the VGG encoder network described in Section 2.3, and another feasible way is to reduce the number of channels of the feature map.

When using the deep network to learn the feature map, it is necessary to perform the operation of changing the number of channels. For the traditional U-Net network, the change of the channel number is generally realized by convolution operation. But

the convolution operation is very time-consuming and more parameters will be generated. Similar to interactive two-stream decoder (ITSD) model[22], the convolution operation in U-Net network will be abandoned and replaced by embedded module. As shown in Fig.1, A_i is the output the encoder feature map E_i passing through the embedded module and the channel number of A_i related to E_i is changed. Thus, in the process of obtaining feature map A_i, the model parameters and the amount of calculation can be reduced, and the purpose of model lightweight can also be achieved. The operation performed by the embedded module is shown as follows

$$A_i = \underset{j\in[0,m-1]}{\text{collect}}\left(\max_{k\in\left[0,\frac{n}{m}-1\right]} E_i^{j\times\frac{n}{m}+k}\right) \tag{2}$$

where A_i represents the corresponding feature map after the feature map E_i is operated by the embedded module. Compared with E_i, the number of channels in A_i is changed. As the index value of the encoder feature map, the value range of i is [1, 5]. The term, indicated as $E_i^{j\times\frac{n}{m}+k}$, represents a channel of the encoder feature map E_i, where $j\times\frac{n}{m}+k$ is the number of channels. It should be noted that j and k are the integers, and n and m represent the number of input and output channels, respectively. Note that n and m must be divided with no remainder. As shown in Eq.(2), in order to change the number of channels, the embedded module is essentially realized by gathering the maximum value of each group of channels, where the original number of channels for each group is n/m.

2.5 Decoder module

In order to effectively use the thermal feature information of the detection target to improve the recognition performance of the salient human body, the decoder of the traditional U-Net network is expanded in the TFSHD model. The expanded decoder contains three modules, i.e., the original decoder module, the saliency decoder module and the fusion module. The main purpose of designing the fusion module is to fuse the thermal features and saliency features extracted from the original decoder feature map, so as to generate a saliency decoder feature map which is more sensitive to high temperature region.

The original decoder module in Fig.1 consists of a series of up-sampling, concat operations, convolutional layers and activation functions. As shown in Eq.(3), the original decoder feature map D_i is obtained by gradually fusing features of various scales, where $i=1,2,\cdots,5$.

$$D_i = T_{D_i}(\text{cat}(A_i, \text{up}(D_{i+1}))) \tag{3}$$

where T indicates that the learning method used in the training process is supervised learning, and its subscript indicates the corresponding module. The symbol of cat indicates the concat operation and the symbol of up the up-sampling operation. Specifically, D_i is obtained by superimposing A_i with the up-sampled D_{i+1}, where A_i is the feature map output of the embedded module and D_{i+1} the original decoder feature map of the upper layer of D_i. It should be noted that A_i is obtained from the feature map E_i of the same layer of encoder, but it is necessary to change the number of channels in the embedded module to obtain A_i. When i is equal to 5, D_{i+1} in Eq.(3) is equal to A_5, and no up-sampling operation is performed. In addition, in order to improve the efficiency of the algorithm, the upsampling method adopted by the original decoder module is bilinear interpolation.

In order to improve the robustness of the TFSHD model to suspicious human targets with partial occlusion, multi-scale and multi-posture, the feature map of the saliency decoder that is more sensitive to the high temperature region of the thermal infrared images will be trained. Thus, a fusion module is designed in this paper. With the help of the module, the thermal and saliency feature information contained in the original decoder feature map D_i is extracted. Furthermore, the two types of information are fused and applied to the learning of the saliency decoder feature map S_i. However, if the two kinds of feature information are fused directly and then used to learn the saliency decoder feature map, the saliency map of the final output of the neural network still cannot meet the needs of practical application.

Based on the above considerations, a fusion module as shown in Fig.3 is designed in this paper. Firstly, we learn the thermal feature information and saliency feature information contained in the original decoder feature maps D_i of five different scales. Secondly, we use these two types of information to construct the thermal branch $H^{(i)}$ and the saliency branch $S^{(i)}$, respectively. Finally, we use a supervised learning method to train the parameters of these two branch structures separately. Specifically, as shown in Eq.(4), the saliency human body is obtained by manual labeling as training data, and the saliency branch is trained by supervised learning. Similarly, as shown in Eq.(5), the thermal map is obtained from the input image as training data, and the thermal branch is trained also by supervised learning. In Eqs.(4, 5), the symbol T indicates that the learning method used in the training process is supervised learning, and its subscript represents the corresponding branch.

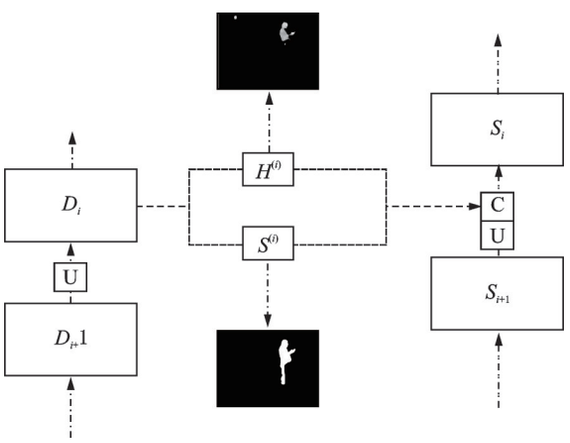

Fig.3 Fusion module

$$S^{(i)} = T_{S^{(i)}}(D_i) \tag{4}$$

$$H^{(i)} = T_{H^{(i)}}(D_i) \tag{5}$$

As shown in Eq.(6), after training the thermal branch $H^{(i)}$ and the saliency branch $S^{(i)}$, S_{i+1}, $H^{(i)}$ and $S^{(i)}$ will be adopted together in the saliency decoder module to generate the saliency decoder feature map S_i. In Eq.(6), T still indicates the method of supervised learning and its subscript the corresponding module. In addition, "cat" represents the concat operation, "cp" the same method of changing the number of channels as the embedded module, and "up" the up-sampling operation. Eq.(6) is as follows

$$S_i = T_{S_i}(\mathrm{cat}(\mathrm{cp}(\mathrm{up}(S_{i+1})), \mathrm{cp}(\mathrm{cat}(H^{(i)}, S^{(i)})))) \tag{6}$$

where the saliency decoder feature map S_i is obtained by fusing the up-sampled S_{i+1} and the two branchs, namely $H^{(i)}$ and $S^{(i)}$, where the channel number of $H^{(i)}$ and $S^{(i)}$ are changed, respectively, and S_{i+1} is the decoder feature map of the upper layer of S_i. Similar to the learning of the feature map of the original encoder, when i is equal to 5, S_{i+1} in Eq.(6) is equal to D_5, and no up-sampling operation is performed.

Since the two branches of $H^{(i)}$ and $S^{(i)}$ strengthen the thermal information and saliency information, the feature map S_i learned by the saliency decoder module also strengthens the thermal and saliency features in the thermal infrared images accordingly. When these feature maps are restored to the final output image, it will contribute to the effect of the human body segmentation in the high temperature area.

2.6 Output module

As shown in Eq.(7), the final output of the model is represented by S_0. It can be obtained by integrating five saliency decoder feature maps S_i (i= 1, 2, …, 5) of different scales. Specifically, five S_i of different scales are respectively up-sampled to obtain the same scale as the input image, and then they are superimposed and fused to obtain S_0.

$$S_0 = T_{S_0}(\mathrm{cat}(\mathrm{up}(S_i)) \tag{7}$$

In addition, in order to realize the model lightweight, a mixture up-sampling method, namely bilinear interpolation method and nearest neighbor method, is adopted in Eq.(7).

2.7 Loss function

In the process of the saliency map detection of the human body by TFSHD, three prediction outputs are involved, i.e., the thermal map branch output, the saliency map branch output in the fusion module, and the final output of the TFSHD model. As shown in Eqs.(8—10), the prediction map is calculated from the feature map of each branch.

$$P_i^H = \mathrm{cp}(H^{(i)}) \tag{8}$$

$$P_i^S = \mathrm{cp}(S^{(i)}) \tag{9}$$

$$P^{S_0} = \mathrm{cp}(S_0) \tag{10}$$

where $i=1,2,\cdots,5$. Thus, by changing the number of feature map channels, a prediction map with a channel number of 1 is obtained. Obviously, P_i^H and P_i^S are the prediction images corresponding to the two branches, denoted as $H^{(i)}$ and $S^{(i)}$. P^{S_0} represents the final output prediction map.

For the loss function in the two branch structures, we choose binary crossentropy loss, which is widely used in segmentation tasks[26]. Then it can be used to calculate the loss between the real image and the corresponding predicted image. As the loss functions corresponding to thermal branch $H^{(i)}$ and saliency branch $S^{(i)}$, $L^{H^{(i)}}$ and $L^{S^{(i)}}$ are defined in Eqs. (11,12), respectively.

$$L^{H^{(i)}} = -\frac{1}{n}\sum_{m=1}^{n}\Big[G_m^{H^{(i)}}\ln(P_m^{H^{(i)}}) + (1-G_m^{H^{(i)}})\cdot \ln(1-P_m^{H^{(i)}})\Big] \tag{11}$$

$$L^{S^{(i)}} = -\frac{1}{n}\sum_{m=1}^{n}\Big[G_m^{S^{(i)}}\ln(P_m^{S^{(i)}}) + (1-G_m^{S^{(i)}})\cdot \ln(1-P_m^{S^{(i)}})\Big] \tag{12}$$

where n represents the total number of pixels and m the index value of pixels. In addition, the symbol G denotes a real image, the superscript represents the image name on the corresponding branch, and the subscript represents the location of the current pixel. Similarly, the symbol P denotes a predicted image, the superscript represents the name of the predicted image on the corresponding branch, and the subscript represents the location of the current pixel. It should be noted that the values of $G_m^{H^{(i)}}$ and $G_m^{S^{(i)}}$ are both 0 or 1, and the values of $P_m^{H^{(i)}}$ and $P_m^{S^{(i)}}$ are in the interval $[0,1]$.

Similar to the definition of $L^{H^{(i)}}$ and $L^{S^{(i)}}$, the definition of the loss function of the final output graph S_0 is shown in Eq.(13). However, in order to make the final output map S_0 of the TFSHD model pay more attention to the pixels in the high thermal region, the term denoted as $1+G_m^{H^{(1)}}$ is added to the loss function L^{S_0}. This means that the pixels in the high thermal region of the thermal map will have a higher weight. In addition, when calculating the prediction loss of the final output map S_0, only the information of the real image $G_m^{S^{(1)}}$ of the first branch on the saliency branch is used. For the same reason, when using the thermal information to weight L^{S_0}, only the real image of the first branch in the thermal branch is used.

$$L^{S_0} = -\frac{1}{n}\sum_{m=1}^{n}[\,G_m^{S^{(1)}}\ln(P_m^{S_0}) + (1-G_m^{S^{(1)}})\cdot \ln(1-P_m^{S_0})\,](1+G_m^{H^{(1)}}) \tag{13}$$

Thus, the total loss function L of the TFSHD model can be defined as the weighted sum of the above three loss functions. The definition of L is shown as follows

$$L = w_{S_0}L^{S_0} + \sum_{i=1}^{5} w_i[\,L^{H^{(i)}} + L^{S^{(i)}}] \tag{14}$$

where w_{S_0} is the weight of L^{S_0} and w_i the weight of the total loss on the two branches of $H^{(i)}$ and $S^{(i)}$. The subscript i corresponds to the subscript of the prediction map. When the size of the prediction map is closer to the size of the input image, the corresponding weight is usually set to a larger value.

3 Results and Analysis

The GPU model used in training the model is RTX 2080 Ti, and the code of the model is built based on the pytorch framework. For model training, the stochastic gradient descent (SGD) method is adopted. Correspondingly, the learning rate is set to 0.001, and the batch value is set to 6.

3.1 Dataset

In order to verify the efficiency and effectiveness of the model, we use four test data sets, among which OSU[27], KAIST[28] and FLIR are three public data sets, and ATH is the actual data set collected by ourselves in the airport terminal. For the OSU data set, some sequence images from irw01 to irw06 in OSU are selected. These images mainly focus on the occlusion of pedestrian targets, including the partial occlusion of the body by the object in the hands of the pedestrian and the occlusion of the body when two pedestrians meet. The data sets are taken by a fixed camera, so the background does not change. Consequently, there is a high contrast between pedestrians and the background.

Both KAIST and FLIR data sets are image data collected on the streets using an in-vehicle camera. These image data often show different imaging effects due to different camera parameters. In addition, pedestrians at different distances from the camera will also show different scales in the image. Considering the actual working scenario of the airport patrol robots, we focus on the different scales of the human body, biking objects and crowds in the data sets.

The ATH data set is an actual data set we collected in the real scene of the airport terminal. Similar to the two data sets of KAIST and FLIR, the ATH data set also faces the problem of image scale caused by the different distances between the human body and the camera. However, unlike other standard data sets, ATH data set contain samples from scenarios such as departure gates, seat areas, and windows that are unique to airport terminals. In addition, the human body in the data set includes a variety of postures such as standing, squatting, and bending. Moreover, the data set also contains multiple occlusion scenarios such as mutual occlusion by human bodies, or occlusion by seats, columns, and clothing. Compared with the standard test data set, the ATH data set is more realistic and the scene is more complicated.

3.2 Evaluation metrics

In order to evaluate the performance of the model, we select the commonly used evaluation indicators F-measure and mIOU in the segmentation model. In general, the higher the F-measure and mIOU values, the better the performance of the model. As shown in Eq.(15), F-measure defines the average value of the weighted harmony of the precision and recall of the saliency human image segmentation.

$$F_\beta\text{-measure}=\frac{(1+\beta^2)\times \text{Precision}\times \text{Recall}}{\beta^2\times \text{Precision}+\text{Recall}} \tag{15}$$

For patrol robots, it is more important to correctly detecte a suspicious target than to correctly detect all areas of the suspicious target's body. This means that when using F-measure to measure model performance, the accuracy of the saliency human image segmentation should take up a larger proportion. As recommended in Ref.[29], the value of β^2 is set to 0.3. The definitions of Precision and Recall are shown in Eq.(16).

$$\begin{cases}\text{Precision}=\dfrac{\text{TP}}{\text{TP}+\text{FP}}\\ \text{Recall}=\dfrac{\text{TP}}{\text{TP}+\text{FN}}\end{cases} \tag{16}$$

where TP represents the number of positive samples that are correctly predicted, FP and FN represent the number of negative samples detected as positive samples and the number of positive samples predicted as negative samples, respectively.

mIOU defined in Eq.(17) represents the intersection ratio of the positive sample in the real image and the positive sample in the predicted image.

$$\text{mIOU}=\frac{\text{TP}}{\text{TP}+\text{FP}+\text{FN}} \tag{17}$$

3.3 Experimental results and analysis

To verify the validity of the model, the model of TFSHD is compared with five saliency methods including Hsaliency[30], Amulet[31], BASNet[21], CPD[32], and SRM[33]. Among these five models, Hsaliency belongs to the traditional saliency detection method, and the other four methods are based on the saliency detection method of the deep neural network model. Table 1 shows the experimental results of the above six methods on four data sets including OSU, KAIST, FLIR and ATH. Table 2 shows the FPS (Frames per second) values of the methods on two different devices, namely RTX 2080 Ti and GTX 1060.

As shown in Table 1, as a traditional saliency detection method, the experimental effect of the Hsaliency method on four data sets is significantly lower than that of the other five methods. According to the experimental results of the Hsaliency method itself on four data sets, the overall performance of this method on the OSU data set is the best. This shows that the traditional saliency method has a certain effect on the data set with the strong contrast between the target and the background. In contrast, the TFSHD model and the other four models have

Table 1 Experimental results on data sets of OSU, KAIST, FLIR and ATH

Method	OSU		KAIST		FLIR		ATH	
	F-measure	mIOU	*F*-measure	mIOU	*F*-measure	mIOU	*F*-measure	mIOU
TFSHD	0.932	0.840	0.733	0.560	0.790	0.697	0.821	0.727
BASNet	0.928	0.854	0.712	0.604	0.741	0.680	0.805	0.705
CPD	0.719	0.444	0.516	0.281	0.563	0.364	0.728	0.491
Hsaliency	0.216	0.179	0.055	0.034	0.020	0.016	0.264	0.134
SRM	0.835	0.692	0.657	0.452	0.663	0.529	0.789	0.653
Amulet	0.914	0.836	0.769	0.558	0.762	0.648	0.807	0.704

better adaptability to the human body detection in various complex scenes due to the use of a deep network structure.

Table 2 FPS values of methods on two different devices

Method	RTX 2080 Ti	GTX 1060
TFSHD	41	12
BASNet	32.5	9.6
CPD	45.7	14.7
Hsaliency	1.5	0.8
SRM	28	8.6
Amulet	17	5.4

Throughout the experimental results of these five deep learning methods, the detection speed of the TFSHD method is only lower than that of CPD method, but significantly higher than that of the other four methods. From the values of *F*-measure in Table 1, the TFSHD method is higher than other methods on three other data sets except that it is slightly lower on KAIST data set than Amulet method. Similarly, from the values of mIOU in Table 1, the TFSHD method is still optimal on the FLIR and ATH data sets, but only slightly lower than the BASNet method on the other two data sets. In general, the performance of Amulet method is close to that of the TFSHD method on these four data sets, and both of them have high accuracy and robustness. However, as shown in Table 2, the Amulet method requires high computational power, which is difficult to meet requirements of real-time detection for patrol robots. In addition, although the CPD method is slightly better than our method in terms of the detection speed, its overall detection accuracy is the worst among these five depth models. Therefore, the experimental performance of the TFSHD method on four data sets has some advantages, especially on ATH data set.

In Fig.4 we show some visualization examples of the six experimental methods on four data sets. For the OSU data set, almost all models can achieve good detection results due to its single background and high contrast. Due to the influence of outdoor radiation on imaging, the final imaging effect of long-distance targets in the KAIST data set is very poor. This leads to the missed detection of distant small-scale targets in the detection process of CPD and BASNet methods. The FLIR data set has a high average gray value of the images due to the differences in camera parameters, which makes it difficult for the Hsaliency method to extract effective features from it. The ATH data set is collected from the airport and contains various complex situations such as multi-scale, multi-posture, and partial occlusion. These factors directly lead to the failure of those models with poor robustness to obtain good results. From another point of view, since ATH data set shows the real state of pedestrians in airport terminals more clearly, the actual application effect of each model in airport terminals is also more clearly reflected by the running results of different models on ATH data set. Based on Table 1 and Fig.4, the TFSHD method has the best experimental results on the ATH data set.

It can be seen from Table 2 and Fig.4 that the BASNet method has both better robustness and faster running speed because it pays attention to the model lightweight and the segmentation effect of target edge pixels at the same time. However, the

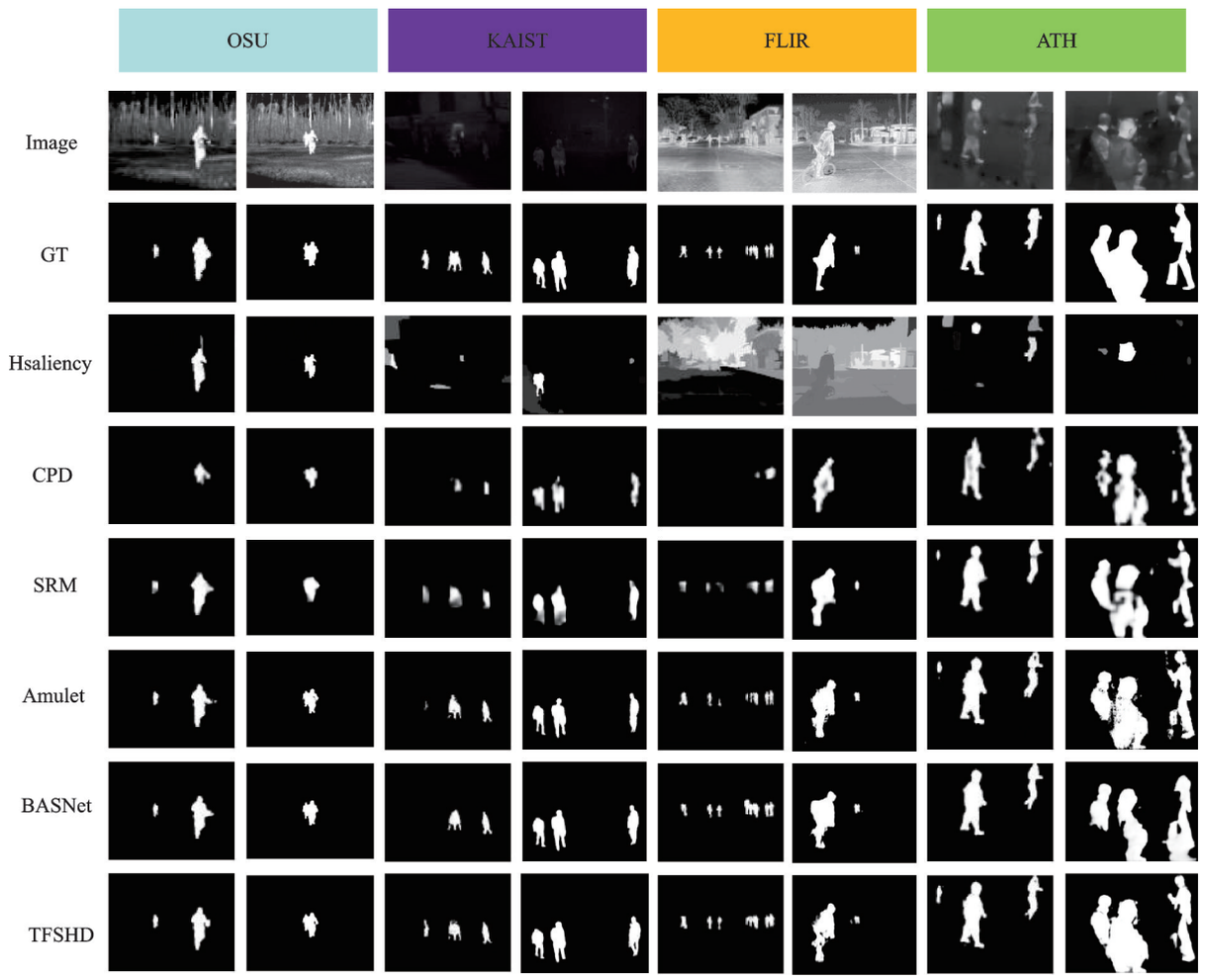

Fig.4 Visual comparison of TFSHD model with other saliency models on four data sets

BASNet model often misses detection when dealing with small-scale targets in the distance. In comparison, the TFSHD model can pay more attention to the pixels in the high temperature area, which ensures that the model still has a better effect even when facing small-scale targets. Additionally, due to the lightweight design of the TFSHD model, this allows the model to use fewer parameters and computing power to complete the calculation of the model.

As shown in Table 2, the TFSHD method can obtain detection results at a speed above 40 f/s on the RTX 2080 Ti device, and obtain detection results at a speed above 12 f/s on a GTX 1060 device with limited computing power. The detection speed has been able to meet the real-time detection task under the thermal imaging camera.

Fig. 5 shows the segmented effect of the TFSHD method on human body targets in airport ticket gates, rest areas, passages and other scenes. As can be seen from Fig. 5, the TFSHD method has achieved good segmentation results for multi-posture targets such as bending, sitting, and standing in the above scenes, as well as multi-scale targets such as small scales and large scales in the distance. When the human body is occluded by objects such as ticket gates, seats, etc., the target can still accurately segmented by the TFSHD method.

Fig.6 shows the effect of the proposed method on the human body target segmentation when there is interference from other thermal sources. Among the three application scenarios, the last column is the night terminal scenario. In addition to the human body, these scenes also contain other thermal sources, such as lights, screens and vehicles. In fact, as mentioned above, the TFSHD method improves the sensitivity of the model to the thermal features

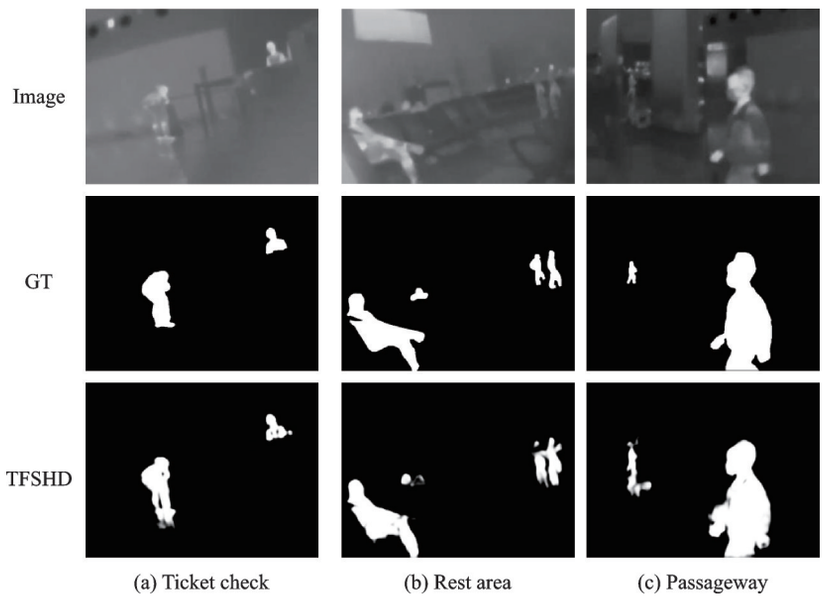

Fig.5 Visual images of TFSHD model in various scenarios of airport terminal

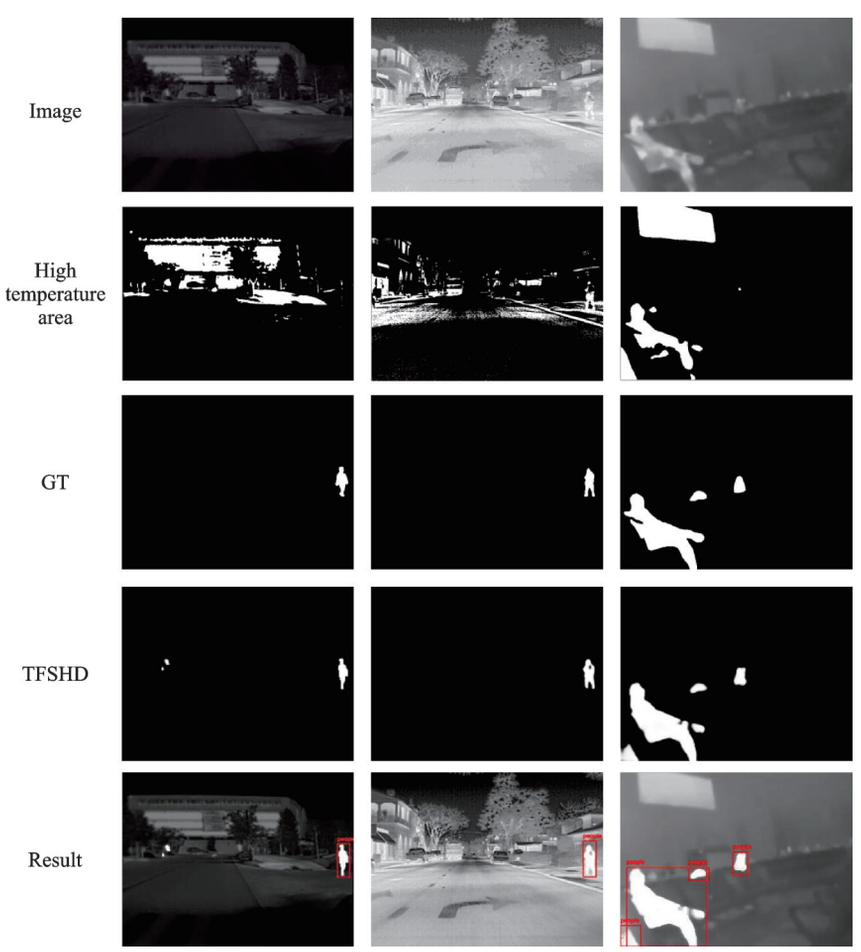

Fig.6 Detection results of TFSHD model in scenarios with multiple thermal sources

of the human body region and optimizes the detection performance of the model for salient human bodies by means of the thermal map of the regions with medium and high gray value in the infrared image. In this way, thermal sources other than the human body are regarded as interference factors. As shown in Fig.6, after the human body region segmentation of the image is completed, the location,

the center and the size of the detection target can also be obtained through the region division, and can be displayed to the user in a visual way.

4 Conclusions

We propose a novel salient human detection model called TFSHD. The proposed TFSHD is based on the traditional U-Net network with an encoder-decoder architecture. However, in order to optimize the detection performance of the salient human body by using the thermal information in the image, the decoder in TFSHD is composed of three components, i.e., the original decoder module, the saliency decoder module and the fusion module. With the help of the optimization of the model architecture, thermal features in the image are used for model parameter training and the learning of salient feature maps. The experimental results on four data sets show that the proposed method is superior to the other five methods in terms of prediction accuracy and robustness. The experimental results on the actual terminal data set ATH further show that the proposed method can effectively generate salient human bodies in multi-posture, multi-scale and partial occlusion situations, and efficiently complete the human body detection in different scenarios in the airport terminal. Additionally, a series of model lightweight design is adopted in our paper. Thus the detection results can be obtained at a speed above 40 f/s, which has been able to meet the requirements of real-time detection for patrol robots.

References

[1] LI H, DENG Y M, XU X B, et al. Eagle-vision-based object detection method for UAV formation in hazy weather[J]. Transactions of Nanjing University of Aeronautics & Astronautics, 2020, 37(4): 517-527.

[2] ZHANG Z Y, CAO Y F, ZHONG P Y, et al. An edge-boxes-based intruder detection algorithm for UAV sense and avoid system[J]. Transactions of Nanjing University of Aeronautics and Astronautics, 2019, 36(2): 253-263.

[3] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]//Proceedings of IEEE Computer Society Conference on Computer Vision & Pattern Recognition. [S.l.]: IEEE, 2005.

[4] DOLLÁR P, TU Z, PERONA P, et al. Integral channel features[C]//Proceedings of British Machine Vision Conference, BMVC 2009. London, UK: DBLP, 2009.

[5] FELZENSZWALB P F, MCALLESTER D A, RAMANAN D. A discriminatively trained, multiscale, deformable part model[C]//Proceedings of 2008 IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2008.

[6] LEI Z, ZHANG Y, WEI W, et al. An associative saliency segmentation method for infrared targets[C]//Proceedings of IEEE International Conference on Image Processing. [S.l.]: IEEE, 2014.

[7] QIN S, LAN W, HUA C, et al. Infrared image saliency detection based on human vision and information theory[C]//Proceedings of 2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI). [S.l.]: IEEE, 2016.

[8] ZHENG Y, ZHOU F, LI L, et al. Propagation based saliency detection for infrared pedestrian images[C]//Proceedings of IEEE International Conference on Image. [S.l.]: IEEE, 2017: 1527-1531.

[9] 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.

[10] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016.

[11] GENG S. Infrared image pedestrian target detection based on Yolov3 and migration learning[EB/OL]. (2020-12-21). https://arxiv.org/abs/2012.11185.

[12] PARK J, CHEN J, YONG K C, et al. CNN-based person detection using infrared images for night-time intrusion warning systems[J]. Sensors (Basel, Switzerland), 2020, 20(1): 1-15.

[13] RONNEBERGER O, FISCHER P, BROX T. U-Net: Convolutional networks for biomedical image segmentation[C]//Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention. [S.l.]: Springer, 2015: 234-241.

[14] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. Computer Science, 2014, 79(7): 4639-4659.

[15] FERNÁNDEZ-CABALLERO A, LÓPEZ M, SER-

RANO-CUERDA J. Thermal-infrared pedestrian ROI extraction through thermal and motion information fusion[J]. Sensors (Basel, Switzerland), 2014, 14(4): 6666-6676.

[16] GOUBET E, KATZ J, PORIKLI F. Pedestrian tracking using thermal infrared imaging[C]//Proceedings of Infrared Technology and Applications XXXII. Kissimmee, FL, US: International Society for Optics and Photonics, 2006.

[17] ZHENG Y, ZHOU F, LI L, et al. Mutual guidance-based saliency propagation for infrared pedestrian images[J]. IEEE Access, 2019, 7: 113355-113371.

[18] MI R J, KWAK J Y, SON J E, et al. Fast pedestrian detection using a night vision system for safety driving[C]//Proceedings of 2014 11th International Conference on Computer Graphics, Imaging and Visualization (CGIV). [S.l.]: IEEE, 2014.

[19] LU L, YU Z, ZHOU F. Contrast and distribution based saliency detection in infrared images[C]//Proceedings of IEEE International Workshop on Multimedia Signal Processing. [S.l.]: IEEE, 2015.

[20] LUO Z, MISHRA A, ACHKAR A, et al. Non-local deep features for salient object detection[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). [S.l.]: IEEE, 2017.

[21] QIN X, ZHANG Z, HUANG C, et al. BASNet: Boundary-aware salient object detection[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). [S.l.]: IEEE, 2019.

[22] ZHOU H, XIE X, LAI J H, et al. Interactive two-stream decoder for accurate and fast saliency detection[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). [S.l.]: IEEE, 2020.

[23] LIU N, HAN J, YANG M H. PiCANet: Learning pixel-wise contextual attention for saliency detection[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). [S.l.]: IEEE, 2018.

[24] ZHANG X, WANG T, QI J, et al. Progressive attention guided recurrent network for salient object detection[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2018.

[25] ZHANG X, ZHOU X, LIN M, et al. ShuffleNet: An extremely efficient convolutional neural network for mobile devices[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). [S.l.]: IEEE, 2018.

[26] BOER P, KROESE D P, MANNOR S, et al. A tutorial on the cross-entropy method[J]. Annals of Operations Research, 2005, 134(1): 19-67.

[27] MIEZIANKO R. Terravic research infrared database[EB/OL].[2020-10-18]. http://vcipl-okstate.org/pbvs/ bench/.

[28] HWANG S, PARK J, KIM N, et al. Multispectral pedestrian detection: Benchmark dataset and baseline[C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). [S.l.]: IEEE, 2013.

[29] ACHANTA R, HEMAMI S, ESTRADA F, et al. Frequency-tuned salient region detection[C]//Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, FL, USA: IEEE, 2009: 1597-1604.

[30] YAN Q, LI X, SHI J, et al. Hierarchical saliency detection[C]//Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Portland, OR, USA: IEEE, 2013.

[31] ZHANG P, WANG D, LU H, et al. Amulet: Aggregating multi-level convolutional features for salient object detection[C]//Proceedings of IEEE Computer Society. [S.l.]: IEEE Computer Society, 2017.

[32] WU Z, SU L, HUANG Q. Cascaded partial decoder for fast and accurate salient object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2019.

[33] WANG T, BORJI A, ZHANG L, et al. A stagewise refinement model for detecting salient objects in images[C]//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). [S.l.]: IEEE, 2017.

Acknowledgements This work was supported in part by the National Key Research and Development Program of China (No. 2018YFC0309104), and the Construction System Science and Technology Project of Jiangsu Province (No.2021JH03).

Author Dr. **YU Yuecheng** received the M.S. degree in Computer Science from Jiangsu University of Science and Technology in 2005 and Ph.D. degree in Computer Science from Nanjing University of Aeronautics and Astronautics in 2012, respectively. He joined in Jiangsu University of Science and Technology in June 2012, where he is an assistant professor of School of Computer. His research is focused on machine learning, computer vision, data mining and relevant fields.

Author contribution: Dr. **YU Yuecheng** designed the study and the network model structure, and wrote the manuscript. Mr. **LIU Chang** designed the lightweight operation, the experiment scheme. Mr. **WANG Chuan** contributed to data and model components for the model. Prof. **SHI Jinlong** contributed to the discussion and background of the study. All authors commented on the manuscript draft and approved the submission.

Competing interests The authors declare no competing interests.

(Production Editor: XU Chengting)

融合热特征的机场航站楼热成像显著人体检测模型

於跃成，刘　畅，汪　川，史金龙

(江苏科技大学计算机学院，镇江 212100，中国)

摘要：弱光背景下的目标检测是航站楼夜间巡检机器人的主要任务之一。然而，那些能够在计算资源有限的机器人平台运行的算法往往难以确保航站楼中人体目标的检测精度。为此，本文提出了一种融合热特征的显著人体检测模型。该模型仍然以U-Net神经网络作为基本架构，但是在解码器模块结构和模型轻量化方面重新进行了设计。一方面，在模型的解码器部分增加了由热特征分支和显著特征分支构成的融合模块，进而设计对图像高温区域更为敏感的预测损失函数，以提升算法在复杂场景下的检测精度；另一方面，通过精简编码器网络结构和控制解码器通道数的方式对模型进行了轻量化改进，以降低算法对计算资源的需求。4个数据集上的实验结果表明，本文方法既能确保较高的检测精度和很好的算法鲁棒性，又能以40 f/s以上的检测速度满足巡检机器人实时检测的需要。

关键词：热成像图像；人体检测；显著性；热特征图；轻量化模型