

Received September 15, 2020, accepted September 27, 2020, date of publication October 5, 2020, date of current version October 16, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3028656

Far-Infrared Object Segmentation Focus on Transmission of Overall Semantic Information

YING ZANG^{1,2,3}, BO YU⁴, LONGJIAO YU⁵, DONGSHENG YANG^{1,2}, AND QINGSHAN LIU³

¹School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 100049, China

²School of Computer Application Technology, Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Shenyang 110168, China

³School of Information Engineering, Huzhou University, Huzhou 313000, China

⁴School of Artificial Intelligence, Jilin University, Changchun 130012, China

⁵School of Information Science and Technology, Dalian Maritime University, Dalian 116026, China

Corresponding authors: Dongsheng Yang (dtsyang@sict.ac.cn) and Qingshan Liu (02726@zjhu.edu.cn)


This work was supported in part by the Major National Science and Technology Projects under Grant 2019ZX04014001-004.

ABSTRACT In the vision task of a self-driving system, the use of visible light images to segment an object often loses its functionality at night or in harsh weather. The far-infrared image shows different pixel values according to the thermal radiation quantity of the object itself, so it can be adapted to perform well at night and in harsh weather conditions. However, at the same time, it has insufficient texture features, blurred object boundaries and temperature inversion, which has a great impact on the segmentation task of traditional algorithms. In response to the above problems, this article proposes a far-infrared object segmentation algorithm using deep learning. In the current popular encoding-decoding structure, multi-scale pooling layers are used to obtain receptive fields of different sizes. This is used to solve the effects caused by the blurring of infrared objects. The feature enhancement module is designed for the multi-receptive field feature map, which can filter out the most versatile and highly semantic feature channels to reduce the effect of temperature inversion on segmentation. The obtained high semantic feature map is guided into the decoding structure and is fused with the features obtained by the encoder and the decoder. This allows richer information to be obtained between different feature maps. Finally, we also release a new low-resolution far-infrared segmentation dataset. Experiments are performed on three datasets, and the segmentation result of the mIoU(mean Intersection over Union) reaches 70.59%, 30.98% and 60.67%. A large number of experiments confirm the effectiveness and robustness of the network in far-infrared images and verify that the dataset released in this article has strong reference significance.

INDEX TERMS Self-driving system, far-infrared image, multi-receptive field, feature enhancement.

I. INTRODUCTION

As the main direction of intelligent development in the global automotive and transportation field, self-driving technology with unique advantages plays an essential role in human travel, showing its infinite value. With the advent of the intelligent era, the industrial revolution represented by the self-driving system and assisted driving systems is facing enormous challenges while developing rapidly. In terms of ensuring pedestrian safety, the use of computer vision technology to detect and locate pedestrians has become increasingly mature. However, in recent years, self-driving vehicles have caused many accidents and have even caused deaths. These incidents warn us of the importance of producing

The associate editor coordinating the review of this manuscript and approving it for publication was Ye Duan .

self-driving cars that can find pedestrians or other objects quickly and effectively, which is currently the most urgent and practical problem.

Many self-driving cars now use visible light cameras to classify objects that appear in front of them. At night or in adverse weather conditions, it is difficult for visible light cameras to obtain high-quality image data due to the influence of lighting conditions and other factors. This has a substantial impact on object discrimination during driving. Because of the above problems, the use of far-infrared images can effectively improve the detection accuracy at night and in harsh environments. The far-infrared camera filters out most of the other light by adding different polarizers to the lens. By receiving far-infrared rays from natural objects, sensors obtain images after processing. The imaging quality depends on the amount of infrared radiation from the object itself.

It is not affected by other environmental conditions, such as lighting conditions, and it can aid in the early identification of pedestrians and other objects in the driving area. The imaging quality helps the driver expand the blind field of vision and increases the trustworthiness of the self-driving system. Far-infrared technology is popular in this field due to its unique characteristics. Figure 1 shows the performance of different camera sensors in the same dark environment.



FIGURE 1. (a) Is the visible light image, (b) is the near-infrared image and (c) is the far-infrared image.

There has been some work on the object detection of far-infrared images [1]–[5]. However, object detection can only provide the bounding box of the object, whereas pixel-level object segmentation is more important for automatic driving. It can be used to distinguish the drivable area and the detailed contour of the object, thereby providing more rigorous conditions for the judgement and decision of the self-driving system. Using traditional image processing algorithms to segment objects is very common in visible light images. It is regrettable that there are few texture feature details of far-infrared images, and it is easy for the object temperature to be reversed. At the same time, far-infrared equipment is expensive, so the collected images tend to have a low resolution, which brings about the problem of blur. These problems make traditional methods perform poorly in far-infrared image segmentation tasks. In recent years, deep learning technology has been used in the field of visible light image segmentation. We believe that deep learning method can also be used for far-infrared images [7], [16]. However, the current popular semantic segmentation method for deep learning is not completely suitable for far-infrared images. We need to optimize the characteristics and difficulties of far-infrared images. Therefore, our goal is to design a neural network algorithm with high robustness that can address the above complex situations that appear in far-infrared images.

In this article, we develop a multi-receptive field and high semantic guidance far-infrared image object segmentation network model (FSGNet). The following are the main contributions of this work.

(1) In the current popular encoding-decoding structure, multi-scale pooling layers are used to obtain receptive fields of different sizes. It has good adaptability to objects of different sizes and solves the sudden change in size caused by the blurring of infrared objects. The feature enhancement module is designed for the multi-receptive field feature map, and the most versatile and expressive feature channels are selected to reduce the effect of object temperature reversal on segmentation. The screened multi-receptive field and high

semantic feature map are guided to the decoding structure and are then merged with the features obtained by the encoder and the conventional up-sampling operation. This reduces the shortcomings caused by the lack of texture features in the far-infrared image.

(2) We release a new dataset for far-infrared image segmentation. Unlike the other two datasets used in this article, our dataset uses a low-resolution far-infrared lens, which can help the algorithm test the robustness under different resolution conditions and can make the algorithm more applicable. At present, it consists of 1,000 low-resolution images, and the initially released dataset has pixel-level annotations for pedestrians. It also has object detection labels for pedestrians. To the best of our knowledge, this is the first detection and segmentation far-infrared dataset made with low-resolution images. It will be available online for free academic usage.

(3) We test the algorithm in this article on three datasets, and the segmentation results of mIoU reach 70.59%, 30.98% and 60.67%.

The rest of this article will be arranged as follows. In the second section, we will present and compare the related work. In the third section, we will introduce the algorithm and dataset proposed in this article. In the fourth section, we will show some experimental data and images and analysis the experimental results. Finally, we will summarize the work of this article and present the possibilities for future work.

II. RELATED WORK

Semantic segmentation is one of the key problems in the field of computer vision. From the macro point of view, semantic segmentation is a high-level task, which realizes the complete understanding of the scene. As a core computer vision problem, the scene understanding is important because more and more applications infer knowledge from images. Some applications include autopilot, human-machine interaction, virtual reality, and so on. In recent years, many semantic segmentation problems are being solved by deep structure with the popularity of deep learning. The most common one is convolution neural network. It is much more accurate than other methods.

A. FAR-INFRARED IMAGE SEGMENTATION

In past work, most scholars have used traditional image processing algorithms to segment objects in far-infrared images. Deckers *et al.* [6] use low-resolution infrared array sensors to segment pedestrians. In the initial stage, they use sensor fusion to collect data. After azimuth correction, the position difference obtained by different sensors is used as a condition to judge whether a pedestrian is present. Simultaneously, Gaussian fuzzy and conditional random fields are used to improve the segmentation accuracy of the algorithm. Brehar *et al.* [7] performed a lot of image enhancement techniques on the far infrared data set. And use ERFNet to segment the processed far-infrared image, and achieve good segmentation results. However, they only applied the methods in visible light image segmentation to far-infrared

images, and did not propose corresponding improvements for far-infrared images. Liu and Zhuang [8] suggests that it is difficult to pre-define filters based on prior knowledge, so they first use fast generalized fuzzy C-means to reconstruct the image and then analysis the clustering center and use an adaptive threshold segmentation algorithm to obtain the final mask image. Wang *et al.* [9] propose a three-step method for pedestrian segmentation. First, a pavement horizontal area estimation algorithm is designed to detect possible locations, and the vertical projection method is also used. Second, the width of each pedestrian's bounding box is very important for the segmentation algorithm. Finally, they use morphological operations to solve the problem of noise in the results. Zhou *et al.* [10] first use a neighborhood-based multi-projection algorithm to generate possible pedestrian positions in a global model. Then, they build a local model based on the pedestrian's head. Finally, the global model and the head model are merged to segment pedestrians. Qiao *et al.* [11] propose a new pedestrian segmentation method based on level sets. They use transformed convolution kernels to generate soft labels, which are used to enhance the edges of the object. An adaptive weight function is used to adjust the unevenness of the image. Based on the adaptively adjusted edges, the model stops shrinking at the boundary. Olmeda *et al.* [12] first searches for possible object areas through the properties of local proportional phase symmetry. Then, the generated high-probability area is sent to the classifier for object judgment. This method filters out many background areas, so the efficiency is higher than that of other methods. Bai *et al.* [13] suggests that the symmetric information of the object is useful, so the geometric symmetry information of the object is generated based on the Markov random field theory, and then the pedestrian segmentation is completed using the fuzzy C-means clustering method. Piniarski and Pawłowski [14] propose a region expansion algorithm with double thresholds, which can effectively reduce the possible object regions, thereby improving the segmentation accuracy. Jin *et al.* [15] also suggests that in the case of a loss of texture features, the object's contour symmetry information is very important, so the ellipse symmetry area is used to construct the object pixel probability map. With symmetric information and probability graphs, a new function based on fuzzy clustering is constructed. At the same time, the local similarity information based on regions is used to effectively control the influence of noise and unevenness. He *et al.* [16] presents an improved PCNN model. Firstly, the weight matrix of the feeding input field is designed by the anisotropic Gaussian kernels (ANGKs), in order to suppress the infrared noise effectively. Secondly, the normalized spectral residual saliency is introduced as linking coefficient to enhance the edges and structural characteristics of segmented pedestrians remarkably. Finally, the improved dynamic threshold based on the average gray values of the iterative segmentation is employed to simplify the original PCNN model. Experiments on the infrared pedestrian image database have been built by their laboratory. However, their algorithm was not verified

on the standard data set, and the segmentation object was relatively single, and there was no large-scale verification on multi-object data.

B. SEGMENTATION METHOD COMBINING INFRARED AND VISIBLE LIGHT IMAGES

Visible light images can provide useful information, so some scholars fuse visible light images and infrared images to perform image segmentation. Sun *et al.* [17] design a neural network that can extract the features of visible light images and infrared images. Then, the two features are fused and decoded. Finally, the object segmentation result is obtained. Shivakumar *et al.* [18] propose a quick calibration method for the alignment of visible light images and far-infrared images and then design a convolutional neural network to extract the features of the above two images. It is worth noting that this network combines infrared images by using visible light images independently. The method proposed in [19] generates many spots on two types of images, and the spots correspond to pedestrians. After the image is geometrically calibrated, the homograph between the spots is calculated according to the parallax of each spot pair. Finally, a multilayer tracking algorithm and background model are used to complete the segmentation. The authors in [20] propose an effective method using infrared images as the fourth channel of visible light images. This combination gives the visible light images more feature details, so that the segmentation accuracy is improved.

C. VISIBLE LIGHT IMAGE SEGMENTATION METHOD USING DEEP LEARNING

Convolutional neural networks initially achieved great success in the field of object detection. In 2015, the proposal of a fully convolutional network (FCN) pioneered the application of convolution in the field of image segmentation [21]. FCNs replace all the last fully connected layers of the network with convolutional layers so that a 2-dimensional feature map can be obtained. It then accesses Softmax to obtain the classification information of each pixel, thereby solving the problem of segmentation. In 2015, the SegNet model was published by Badrinarayanan *et al.* [22]. Based on the semantic segmentation task of FCN, an encoder-decoder symmetric structure was built to achieve end-to-end pixel-level image segmentation. The keys to SegNet's semantic segmentation network are down-sampling and up-sampling. During the up-sampling process, the coordinate position of the largest pixel recorded by the down-sampling is used. Zhao *et al.* [23] suggests that many problems occur because FCN cannot effectively handle the relationship between scenes and global information. This article proposes a pyramid scene analysis network PSPNet that can obtain global scenes. It can fuse the appropriate global features, fuse local and global information together, and finally embed in the FCN-based prediction framework. At the same time, an optimization strategy for moderately supervised losses is proposed, which performs well on multiple datasets. Compared to FCN, PSPNet, etc.,

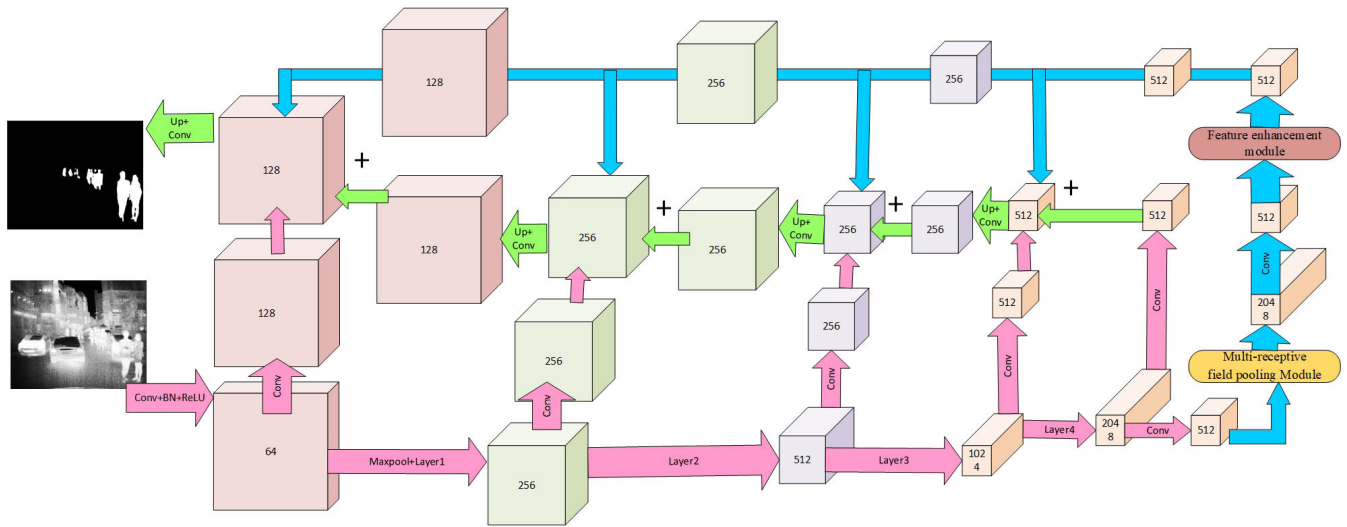


FIGURE 2. The overall structure of FSGNet.

UNet [26] has a total of 4 up-sampling processes and uses a skip-connect structure with the same step size. It abandons the design of training and backward propagation directly on high-level semantic features to ensure that the recovered feature map incorporates more semantic information. At the same time, the feature maps of different scales are fused so that multi-scale prediction and training can be performed. By up-sampling 4 times, the edge and other information of the restored image become more refined.

III. MATERIALS AND METHODS

A. THE OVERALL STRUCTURE OF THE NETWORK

Figure 2 provides the global structure of the FSGNet. In the figure, we clearly indicate the three streams of the data. In the first data stream, we encode the image on ResNet50. The blue part of the overall structure of FSGNet represented the ‘Transmission’. The pink stream part of the Figure 2 is the feature extraction part of the network. It is also called the coding process. Through the sub network, we can get the features of the image. From the graph, the different scale feature maps can be obtained for the network. The green flow part in the middle of the graph indicates that the feature images of different scales are up-sampled and restored to the original image size, which is the decoding process. The blue stream sections on the top and right are the ‘Transmission’ section. The sub network performs a series of operations on the last layer of the coding part, such as multi-receptive field pooling, feature enhancement and high semantic guidance. The purpose is to increase the recognition ability and feature expression ability of the network. Finally, the feature representation is transmitted to each feature map with different scales. Therefore, this article uses the ‘Transmission’ to express the network meaning of this part.

After the input image undergoes multi-layer convolution, the resolution continues to decrease, and the number of channels continues to increase. In the second data stream, the resulting feature map of the last layer is

dimension-reduced and up-sampled to the same size as the previous layer. After feature fusion, the up-sampling operation is continued until the final prediction layer. The last data stream is the highlight of this article. This includes a multi-receptive field pooling structure, feature enhancement module and high semantic guidance structure. Next, we will focus on the third data stream. Table 1 is the network architecture in tabular form as well that explains the size of feature maps with layer input-output filters of each convolutional layer.

TABLE 1. Network architecture for explains the size of feature maps.

input	640×480×3
Conv+BN+Relu	320×240×64
Layer1	160×120×256
Layer2	80×60×512
Layer3	40×30×1024
Layer4	40×30×2048
Multi-receptive field module	40×30×2048
Feature enhance module	40×30×512
Up-sample1+fusion	40×30×512
Up-sample2+fusion	80×60×256
Up-sample3+fusion	160×120×256
Up-sample4+fusion	320×240×128
Segmentation result	640×480×21(2)

B. MULTI-RECEPTIVE FIELD POOLING

Figure 3 shows the specific structure of multi-receptive field pooling. The algorithm first extracts the last layer of the basic network features and uses a 1×1 convolution kernel to reduce the dimension. The input is reduced from 2048 channels to 512 channels. This step aims to reduce the operating burden of the network when analyzing multi-receptive fields and uses refined features to extract objects of different sizes. After the simplified feature map is obtained, an adaptive pooling kernel is used to capture different

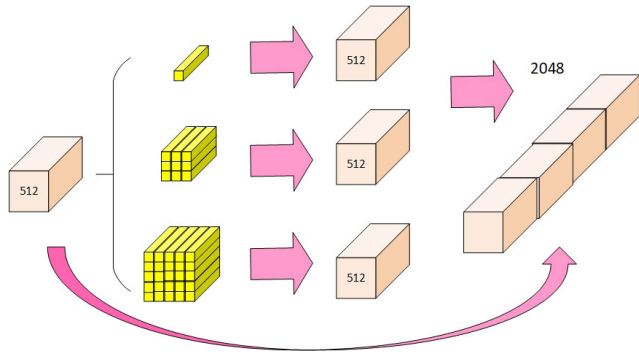


FIGURE 3. The structure of multi-receptive field pooling module. The input features will be effectively combined with the features from different receptive fields.

object sizes. The receptive field is the size of the visual range that can be perceived. Through pooling layers of different sizes, we can obtain a variety of visual ranges. This structure based on pyramid pooling has achieved great success in visible light images [23]. However, we find that the multi-receptive field is more effective for far-infrared images. He *et al.* [24], [25] tried to use the multi-receptive field model to solve the far-infrared and visible image super-resolution task. Super-resolution mainly emphasizes the details of the image, which can better restore the details of the image. Therefore, literature [24] designs the multi-receptive field on the high-resolution detailed feature map, literature [25] integrating these hierarchical features can generate better mappings on recovering high-fidelity details at different scales. However, our image segmentation mainly emphasizes the image overall characteristics. If we pay too much attention to details, there will be many misjudgments. Features of far-infrared images are mainly concentrated on some large overall structures, and far-infrared devices are not good at capturing detailed features. Far-infrared image segmentation pays more attention to semantic features such as contour and gray value. Therefore, we only add multiple receptive fields to the feature map of the last layer. According to the characteristics of far-infrared images, we use global pooling to get multiple receptive fields, and act on the final output feature map to obtain features of different scales semantically. The design of this study differs from that of others in the existing literature and is detailed below.

For far-infrared images, due to the expensive acquisition equipment, most of the currently available dataset resolution is 640×480 , while the thermal resolution of the dataset released in this article is lower, only 160×140 . The lower resolution makes the edges of pedestrians and other objects appear blurred in the image. At the same time, due to the characteristics of the far-infrared image, the value of the object pixel is only related to the heat radiation. This causes some background areas or unrelated categories to be easily judged as the current category. However, this kind of error is very easy for humans to judge because humans can obtain different comparison references according to different visual ranges. Assuming that there is a blurry pure white area, based

on the available information, we cannot distinguish whether it belongs to the pedestrian category or another category. However, if we zoom in on the visual range, we can see other parts, including the head and limbs. Then, we can easily judge that it is part of the pedestrian object area. Figure 4 shows the different visual ranges. This is particularly important for far-infrared images without texture information and blurred edges. Due to the low resolution of the far-infrared image, the last layer of the feature map has a small size, so we use three pooling layers with sizes of 1, 3, and 5. Through combining them with the last layer of features before processing, we finally obtain a 2048-dimensional feature vector.



FIGURE 4. The importance of different visual ranges in understanding object categories. It is difficult to distinguish the category of the object with a small visual range. As the visual range gradually becomes larger, the category of the object can be accurately determined based on the contour and other information.

C. FEATURE ENHANCEMENT MODULE

Multi-receptive fields can obtain multi-scale objects that solve the problems caused by low resolution and blurring. Another difficult problem is the situation of temperature inversion. Under normal circumstances, the temperature of pedestrians will be higher than the temperature of the background, so the pixel grey value will also be higher. However, temperature reversal often occurs. During the summer, long-term sun exposure can cause the road temperature to rise sharply. During this time, the temperature of the surrounding objects will be higher than the temperature of the pedestrians, which will cause the contrast of the pixel grey value to decrease. In winter, pedestrians wear warm clothing, such as down jackets, with only their heads exposed. Therefore, the temperature of the body will also be low in most cases, reducing the contrast compared to the background and other objects. Figure 5 shows the situation that has been mentioned. These problems are very simple for human judgment. Human vision is not only based on the pixel grey value but also based on contour and position information. For the designed algorithm, we need to strengthen the network's ability to extract common features of far-infrared image objects and establish effective connections between the different features.

In this article, the global pooling layer is used to obtain the most response part of the feature map. The two fully

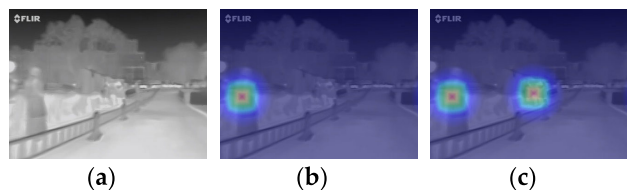


FIGURE 5. The situation in (a) is that the pedestrian objects are hidden in the background. Networks which do not have the selection capabilities of enhanced features can only focus on more obvious feature objects, as shown in (b). Thanks to the strong expression ability of the feature enhancement module, it can effectively find the objects hidden in the background, as shown in (c).

connected layers are mainly used to select channels with representative far-infrared features, and train weight values that can respond to key position information. Finally, the weights with screening ability are assigned to these feature maps by multiplication, and the feature values with important expressive ability are enhanced to complete the task of selecting important features. The filtered features can effectively distinguish the importance of the foreground and background features, pay more attention to the position where the object contour may appear, and ignore some pixels with noisy in brightness values, thereby improving the accuracy of segmentation.

The model is shown in Figure 6. Based on literature [22], we introduce the feature enhancement module into our far-infrared image segmentation task. This module has a low computational cost and is easy to implement. On this basis, we modify the designed network. First, the obtained 2048-dimensional multi-receptive field vector is reduced to 512 channels using a 1×1 convolution kernel. Then, the feature map is subjected to global max-pooling, and the result is converted into a 1-dimensional vector. It is entered into the 128-dimensional fully connected layer. After the Relu activation function, it is connected to the 512-dimensional fully

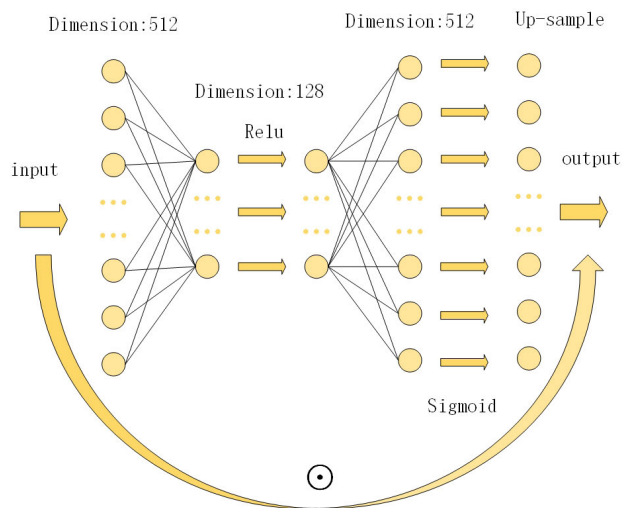


FIGURE 6. The structure of feature enhancement module. The effective channels that are screened out will be multiplied by the input features, thereby enhancing the ability to express key features.

connected layer and through the Sigmoid activation function. This step can effectively filter out more representative and versatile feature channels, such as contours and positions, and effectively reduces the impact of grayscale color inversion. Finally, unlike the design in literature [27], we directly up-sample the obtained expressive channels and multiply them into the original input. Since the original structure is embedded in the basic feature extraction network, it is presented as a residual structure. The design presented in this article uses the original structure, as opposed to the residual structure, for the screening of general features, so as to exploit its outstanding selection ability.

D. SEMANTIC GUIDANCE STRUCTURE

Currently, popular semantic segmentation networks all use encoding-decoding architectures [22], [26]. The basic structure of the method in this article is also designed in this way, but using only the classic encoding-decoding architecture does not perform well with far-infrared images. Because far-infrared images lack considerable texture and color information, if only one up-sampled data stream is used, the semantic extraction ability of the network will be weakened. In the last two sections, we discussed the design of our novel far-infrared image modules; next, we will discuss their integration into the basic network structure.

The commonly used pyramid pooling structure directly up-samples to the size of the input image after extracting multiple features, which has little effect on the classification of visible light images with rich texture information. However, the far-infrared image has almost no texture information. We only rely on the edge and brightness information for classification and judgement. Therefore, after extracting the multi-receptive field features and enhancing the module, the structure we designed is gradually up-sampled to fully integrate the features of each layer. Once the features are fully integrated, our structure continues up-sampling. In each layer, we use the strategy of adding corresponding positions to avoid an increase in channel numbers. The guidance structure is shown by the blue data flow in Figure 2. Based on the original up-sampling structure, each layer has three different features for fusion. They are the down-sampling detail features, conventional up-sampling semantic features, and multi-receptive field's high semantic features. This design can better ease the fusion process between features and avoids the 'semantic gap' caused by the large span of space.

E. SUPPLEMENTARY EXPLANATION OF THE NETWORK STRUCTURE

This article uses a stochastic gradient descent optimizer (SGD). SGD performs gradient updates on each sample during training. For large datasets, there may be similar samples. If global samples are used, then there will be redundancy in the calculation of gradients. SGD is updated only once each time, it has no redundancy, it is faster, and it can enrich the sample. Although it contains a certain degree of randomness, from the perspective of our expectations, it is



FIGURE 7. Some images and labeled samples of the LFPD dataset.

equal to the correct derivative. We chose this optimizer to verify the network’s ability to extract features and reduce the impact of other functions on the network. Equation (1) is used to calculate the gradient of the current parameter where, ω_t is the weight value, f is the derivative operation, and g_t is the derivative result. (2) is used to calculate the gradient of the current drop, α is the learning rate, and θ_t is a decreasing value. (3) is used to calculate the updated parameter, and ω_{t+1} is the new weight value.

$$g_t = \nabla f(\omega_t) \tag{1}$$

$$\theta_t = \alpha \cdot g_t \tag{2}$$

$$\omega_{t+1} = \omega_t - \theta_t \tag{3}$$

This article uses the NLL Loss cost function. Its input is a log probability vector and an object label. While it does not calculate the log probability, it is suitable for the case where the last layer of the network is Log_Softmax. For the traditional Softmax, the calculation equation is:

$$\text{Softmax}(x_i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)} \tag{4}$$

x_i is the current category, x_i is the all-category, and \exp is the exponential operation. Log_Softmax performs a logarithmic operation on the calculation result of Softmax. The calculation equation is as follows, where a_k^L is the output of Softmax, the value of y_k is 0 or 1. It is 1 when the training data are in this category; otherwise, it is 0. k is the real class number of the training sample.

$$J(W, b, a^L, y) = - \sum_k y_k \ln a_k^L \tag{5}$$

Due to the low resolution of the far-infrared image after five down-samplings, the final feature map becomes very small. Therefore, we use dilated convolution in the last convolution operation instead of down-sampling, so that a receptive field of the same size can be obtained without changing the resolution. Dilated convolution uses a hole between

each filter point to expand the receptive field so that the 3×3 convolution kernel obtains 5×5 convolution kernel semantic information with the same number of parameters and computational complexity. It replaces the traditional down-sampling operation.

F. LOW-RESOLUTION FAR-IR INFRARED DATASET

To conduct multi-resolution research on far-infrared images and verify the robustness of the algorithm, this article releases the low-resolution far-infrared pedestrian dataset (LFPD). We collected the driving scene by installing a FLIR One Pro3 far-infrared camera on the rear-view mirror inside a car. It can measure temperatures between -20°C and 400°C . The images were collected from the video data; each video is 30 minutes, and there are 30 videos in total. We carefully selected the images that have some useable objects, and a total of 1000 images were obtained to make the LFPD dataset. All videos were taken on the streets at Dalian Maritime University, Dalian, China, in May 2018. Videos were taken under generally clear-skies and rainy conditions at both day and night. The dataset folder form is the same as the public dataset VOC [28]. The labelImg tool is used to label the images and automatically generate an.xml file. At the same time, we use Labelme to annotate images at the pixel level. The thermal resolution is 160×140 , and the image size is 640×480 . Currently, only pedestrian categories are used for object detection and semantic segmentation. We will continue to provide other categories of images and annotation information in the future. At present, the dataset is publicly available. You can send an email to the corresponding author of this article to obtain the download address. Some marked sample results are shown in Figure 7.

IV. EXPERIMENTAL RESULTS

A. DATASET

In the experiments in this article, in addition to using the low-resolution far-infrared images we released for verification,

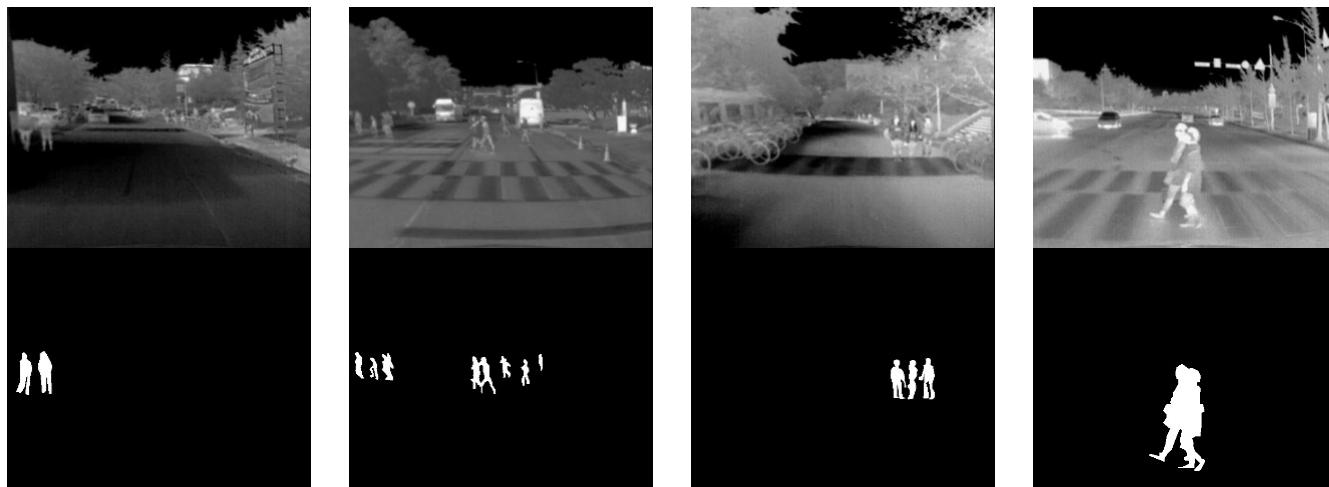
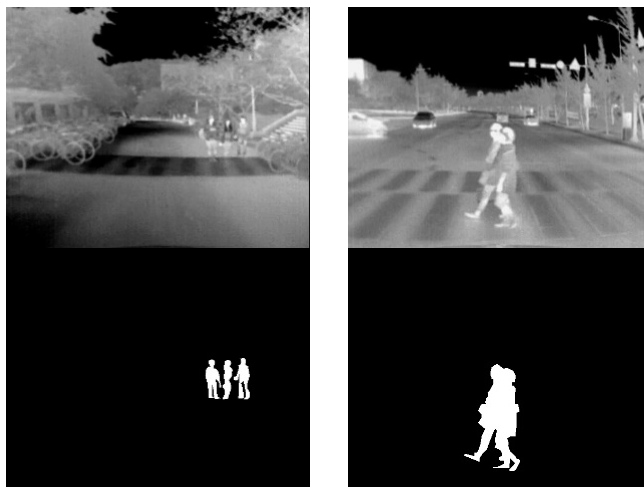


FIGURE 8. Some images and labeled samples of the KAIST dataset.

we also used far-infrared segmentation datasets currently published by other scholars. Now, we briefly introduce them.

The KAIST Multispectral Pedestrian Dataset consists of 95k color-thermal pairs (640×480 , 20 Hz) taken from a vehicle [29]. The developed imaging hardware consists of a color camera, a thermal camera, and a beam splitter to capture aligned multispectral (RGB + far-infrared) images. With this hardware, various regular traffic scenes during the day and night were captured to consider the changes in light conditions. In this article, the far-infrared data is selected for experimental analysis, and the corresponding segmentation labels disclosed by other scholars are used. Based on the annotation information of the original data set, they organized and annotated the KAIST far-infrared data set for semantic segmentation, and published the segmentation data set in the paper [31]. They only marked the pedestrian category at the pixel level. Some marked sample results are shown in the 2nd line of Figure 8. Its address is as follows. (Address: <https://information-fusion-lab-umass.github.io/Salient-Pedestrian-Detection/>).

The ‘Segment Objects in Day And night (SODA)’ dataset was released by Li *et al.* (Address: <https://drive.google.com/drive/folders/1ZF2vDk9j69kP5U0zcp-liOBk-atWcw-5?usp=sharing>). There are 2168 annotated images and 5000 pseudo far-infrared images collected by the FLIR thermal camera. Among them, the resolution of the real far-infrared images is 640×480 , which was collected from scenes of daily life. In our experiments, only the real far-infrared images in this dataset were used, and the pseudo far-infrared images were eliminated. The dataset contains a total of 20 categories, such as person, car, tree, road, and bicycle. This can not only verify the segmentation performance of the network for far-infrared pedestrians in this article, but also verify the segmentation performance of the proposed method for other categories of far-infrared objects. The Labelme tool was used to annotate the images at the pixel level. Through training, pixel values are labelled from 1 to 20. The visual color of the label is the same as that of the VOC segmentation



dataset. The dataset is divided into the training data and test data using a 1:1 ratio, so the training data have 1168 images and the test data have 1000 images. Some examples are shown in Figure 9.

B. EXPERIMENTAL CONFIGURATION

This article uses a Nvidia GTX1080 with an 11 G memory GPU for the experiment. The CPU model is Intel i7-7700HQ, the running memory is 8 GB, and the experimental environment is Ubuntu16.04 + Cuda9.0 + Cudnn7.0.5 + Pytorch0.4.1. The experimental comparison methods are all open source code (<https://github.com/bodokaiser/piwise>). The initial learning rate of the network is 0.0001, the optimizer uses SGD, the batch size is 2, and each algorithm is trained with 100 epochs.

In the training process, the first step is to experiment with the original UNet structure. On this basis, the parameter of the dilated rate of 1 is used to obtain the semantic guidance structure. Then, it is up-sampled layer by layer and merged with the feature map obtained by UNet before. Multi-receptive field and feature enhancement module are based on semantic guidance structure. In the second step, the adaptive pooling layer is used to obtain 1×1 , 3×3 and 5×5 feature maps after dilated operation. Then get the multi-receptive field features. After up-sampling, fusion them with the original features in the channel dimension. Finally, the feature enhancement module acts on the fused feature map. If there are no multi-receptive field features, directly apply it to the feature map after dilated convolution. The number of head and tail neurons in this module is 512, and the middle is 128. In the process of training and testing, verify them one by one according to the above method, and get the final segmentation result.

C. EVALUATION METRIC

In this article, we use precision, recall, intersection-over-union (IoU), f-score and mean absolute error (MAE) [30] to evaluate the pedestrian segmentation results.

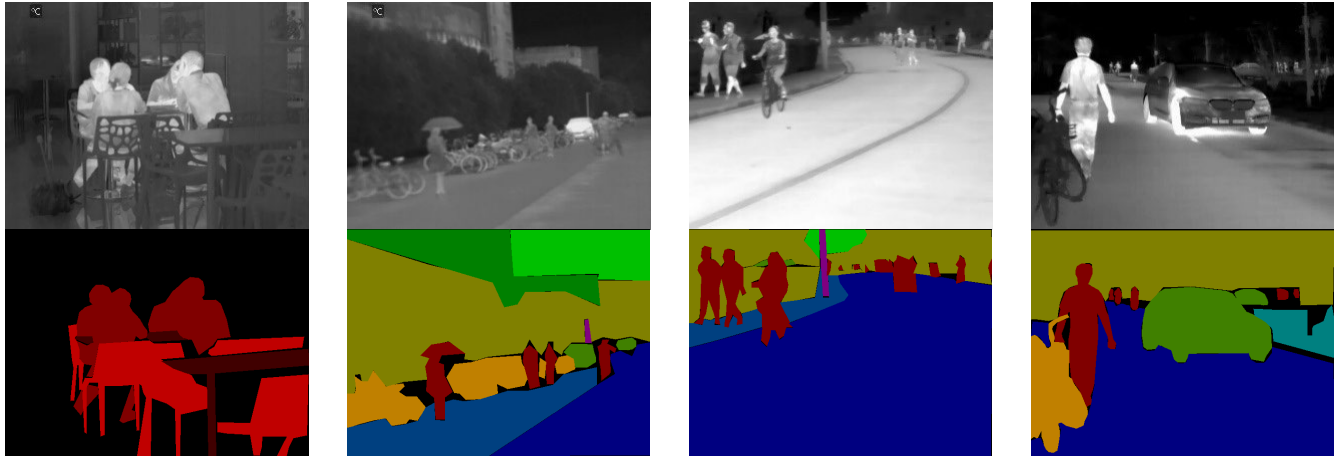


FIGURE 9. Some images and labeled samples of the SODA dataset.

For each pixel, if it is a positive example, the correct classification is recorded as TP. Otherwise, it will be marked as FN. When it is a negative case, the proper label will be recorded as TN, and if it is incorrectly classified as a positive case, it is recorded as FP. Therefore, the precision and recall can be calculated using Equations (6) and (7) based on the above four indicators.

IoU is a concept widely used in object segmentation and detection. The purpose is to calculate the overlap ratio of the generated true pixel value and labelled pixel value. We can also say it is a ratio of their intersection to their union. The best case is a complete overlap, where the rate is 1. Let the mask pixel point obtained by the program be W and the labelled pixel point be E ; it is calculated by Equation (8). In the multi-category test, the IoU of all categories is averaged to obtain the mIoU.

To comprehensively evaluate the quality of the segmented images, the f-score index is used, which can measure the degree of influence between the precision and recall. In this article, β^2 is 1. It is defined as Equation (9).

The MAE represents the average of the absolute errors, and the absolute differences of the corresponding position pixels are summed and then averaged. It can effectively reflect the actual situation of the prediction error. After normalization to $[0, 1]$, let S and Z represent the predicted segmented image and the label image, respectively. The MAE can be calculated by Equation (10).

In addition, for multi-category test datasets, we only use the mIoU and pixel accuracy (PA) indicators for evaluation. In Equation (11), k represents the number of categories, p_{ii} represents the number of pixels predicted correctly, and p_{ij} represents the total number of pixels.

$$P(\text{Precision}) = \frac{TP}{TP + FP} \tag{6}$$

$$R(\text{Recall}) = \frac{TP}{TP + FN} \tag{7}$$

$$\text{IoU} = \frac{\text{area}(W) \cap \text{area}(E)}{\text{area}(W) \cup \text{area}(E)} \tag{8}$$

$$F_\beta = \frac{(1 + \beta^2)\text{Precision} \times \text{Recall}}{\beta^2\text{Precision} + \text{Recall}} \tag{9}$$

$$\text{MAE} = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W |S(i, j) - Z(i, j)| \tag{10}$$

$$\text{PA} = \frac{\sum_{i=0}^k p_{ii}}{\sum_{i=0}^k \sum_{j=0}^k p_{ij}} \tag{11}$$

D. COMPARISON EXPERIMENT OF SEGMENTATION

Figure 10 shows the training situation of different algorithms on three datasets. With the increase in epoch, their loss becomes small, and the FSGNet achieves a low loss level. The LFPD, KAIST and SODA datasets all reach a loss level of approximately 0.001. Compared with other algorithms, our algorithm has a good convergence effect. The curve proves the rationality of the network structure.

In Figure 11, the test results on the two datasets that use five evaluation metrics are shown by curves. Because these two datasets only have pedestrian categories, we can use the above metrics to evaluate them. In LFPD, with increasing test image number, the indicators are relatively stable and have a slight downward trend. This shows that the network structure of this article can be well adapted to low-resolution images and has a stable test ability compared with that of the other algorithms. However, in the KAIST dataset, the test results are relatively low at the beginning. With the increase in epoch, the indicator value gradually increases and tends to be stable. After analysis the test images, we found that most of the images arranged in the front did not contain pedestrian objects or they contained small objects. Small objects affect the detection ability of the network. Compared with other networks, FSGNet obtains a higher detection performance from the beginning, which proves that it has good adaptability to the object scale.

Table 2 and Table 3 also prove the above analysis. The FSGNet achieves the best performance when compared with the other algorithms. The IoU reaches 70.59% in LFPD and 30.98% in KAIST. Table 4 shows the test results for the

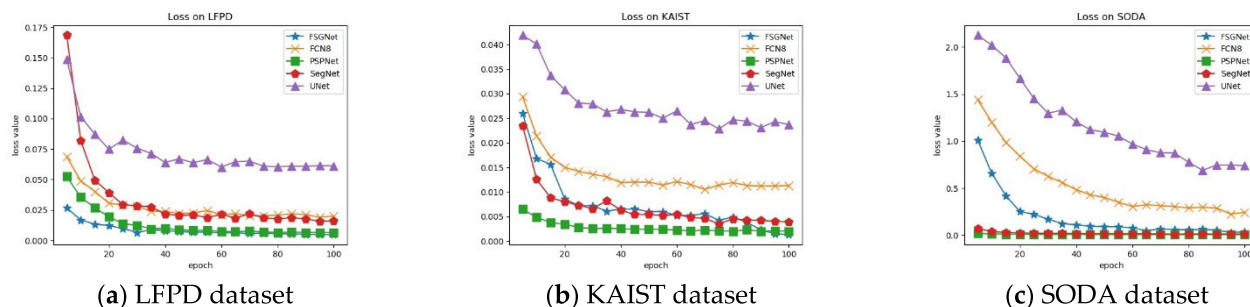


FIGURE 10. The performance of different algorithm training in three datasets.

TABLE 2. The average test results of different algorithms in the LFPD dataset.

Unit: %	Precision	Recall	IoU	F-score	MAE
UNet [26]	67.40	63.18	45.59	59.54	2.61
FCN8 [21]	76.99	71.68	60.85	72.36	1.44
PSPNet [23]	80.04	75.31	63.14	74.65	1.57
SegNet [22]	81.80	74.69	64.55	75.48	1.31
FSGNet ^a	83.30	80.75	70.59	80.58	1.03

^aThis work

TABLE 3. The average test results of different algorithms in the KAIST dataset.

Unit: %	Precision	Recall	IoU	F-score	MAE
UNet [26]	40.74	10.87	9.61	15.50	0.71
FCN8 [21]	54.23	27.03	23.36	32.75	0.52
PSPNet [23]	49.84	19.36	17.42	24.68	0.57
SegNet [22]	52.22	33.93	26.93	37.31	0.54
FSGNet ^a	62.55	35.52	30.98	42.06	0.46

^aThis work

TABLE 4. The average test results of different algorithms in the SODA dataset (Use IoU and PA).

Unit: %	UNet [26]	FCN8 [21]	PSPNet [23]	SegNet [22]	FSGNet ^a
person	37.98	56.24	36.40	61.04	68.16
building	42.94	55.58	45.15	60.89	68.21
tree	58.26	63.62	58.49	67.12	72.49
road	69.35	74.80	73.64	76.04	81.64
pole	1.02	29.68	12.95	30.05	49.35
grass	28.30	35.49	36.21	42.43	51.45
door	0.00	22.93	15.17	0.00	45.12
table	0.00	0.00	14.25	0.00	27.01
chair	1.59	30.98	15.68	37.25	46.16
car	35.73	65.04	50.43	68.65	77.07
bicycle	15.27	43.95	30.17	44.93	56.29
lamp	0.00	0.00	10.08	0.00	78.63
monitor	24.61	43.17	37.09	15.87	58.56
Traffic Cone	0.17	49.33	37.41	54.61	60.00
trash can	0.00	37.64	42.79	11.42	58.18
animal	0.00	47.31	30.85	31.13	68.43
fence	6.09	38.10	35.44	34.32	54.39
sky	67.67	69.29	62.21	74.03	78.92
river	24.92	53.05	57.44	59.75	78.32
sidewalk	36.02	39.15	33.63	49.34	60.78
mIoU	29.45	46.35	36.11	44.12	60.67
PA	60.31	70.23	66.44	73.91	81.31

^aThis work

SODA dataset. Because it is a multi-category dataset, we use mIoU and PA to prove the segmentation ability of our network. FSGNet achieves the highest score in all categories, and no zero values are found. The mIoU and PA are 60.67%

and 81.31%, respectively. The data effectively illustrate that FSGNet is not only be suitable for pedestrian segmentation of far-infrared images but is also suitable for multi-category segmentation of far-infrared images.

There are some segmentation images by different algorithms in the three datasets. In Figure 12, FSGNet can adapt well to the shortcomings of insufficient texture features in low-resolution images. Especially in the case where pedestrians are blocked by clothing, the FSGNet can completely divide the pedestrian area without many holes. In Figure 13, the pedestrian object of the test image is small, and the environment is complex, but our network still segments the pedestrians very well. It is worth noting that due to the rough manual annotation of the dataset, the model achieves relatively low results in the final test. In fact, the final segmentation result is much finer than the marked image. This also shows that our network has a good generalization ability. Figure 14 shows the test result in the multi-category dataset. It can be seen from the figure that FSGNet is very prominent in its ability to segment image details, and it has good adaptability for different scales and brightness.

From the above three tables, it can be seen that the methods proposed in this article have improved detection accuracy compared to UNet, FCN8, PSPNet, and SegNet. We chose these methods for comparison because in related work, we have introduced the role of these work in visible light images. But how do they perform in far infrared images? Is our model more suitable for far infrared image segmentation? With these questions, this article chooses these algorithms for comparison. In the evaluation of pedestrian segmentation, the precision, recall, IoU, Fscore and MAE indicators have good advantages. In the multi-category data set, the mIoU and PA indicators also have strong advantages, and each category in the data set can also obtain the highest segmentation accuracy. This shows that the network we designed can segment the object in the far infrared image well.

Figure 12-14 shows the test results of different algorithms on three data sets. It can be seen from the figure that the compared several algorithms have a greater loss of pedestrian details. Especially the parts that are occluded by clothing. Due to the different brightness values of pedestrians, the analysis performance of the features is reduced. Our method

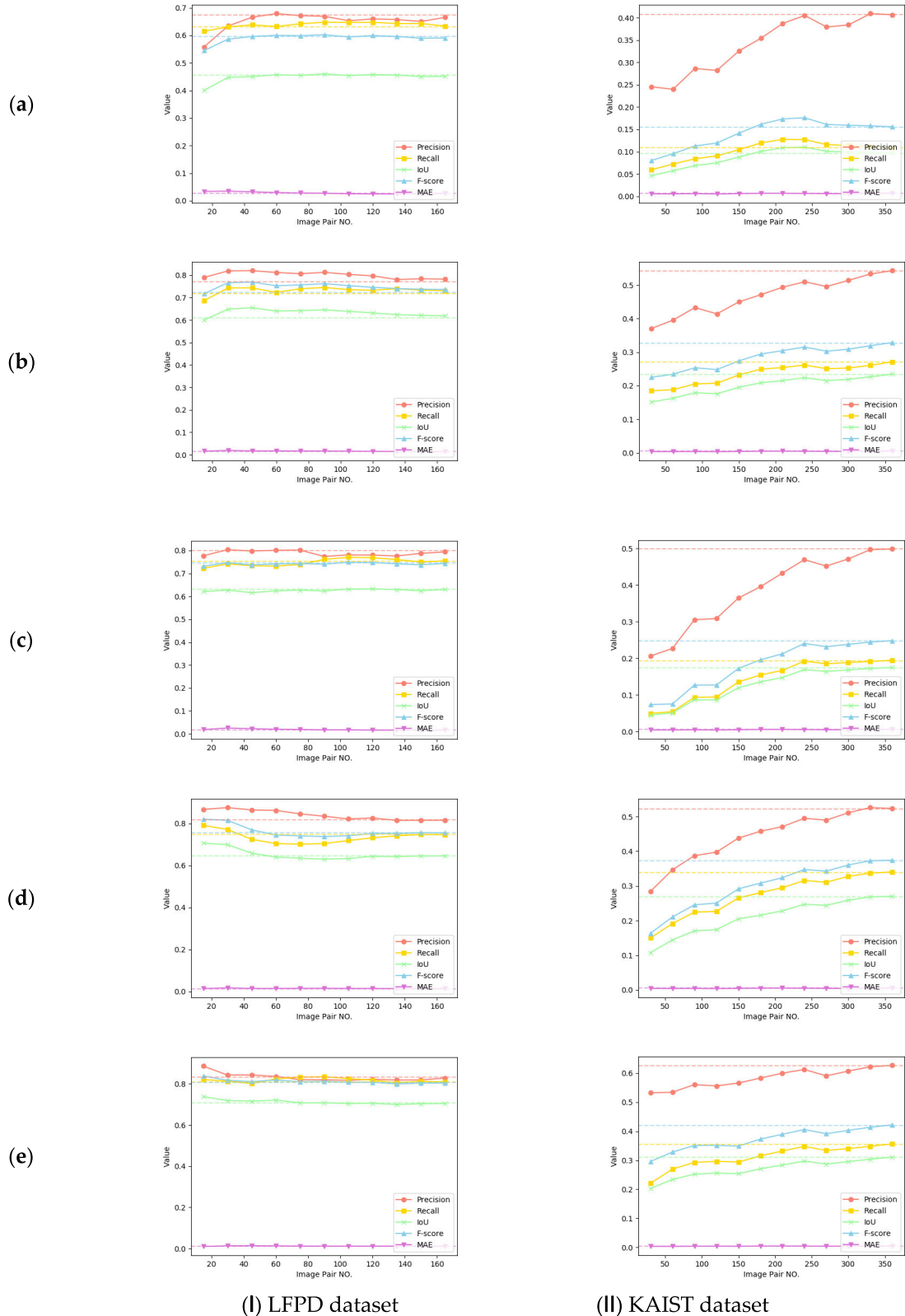


FIGURE 11. Test results of different algorithms in two datasets. (a) is the UNet, (b) is the FCN8, (c) is the PSPNet, (d) is the SegNet, (e) is FSGNet.

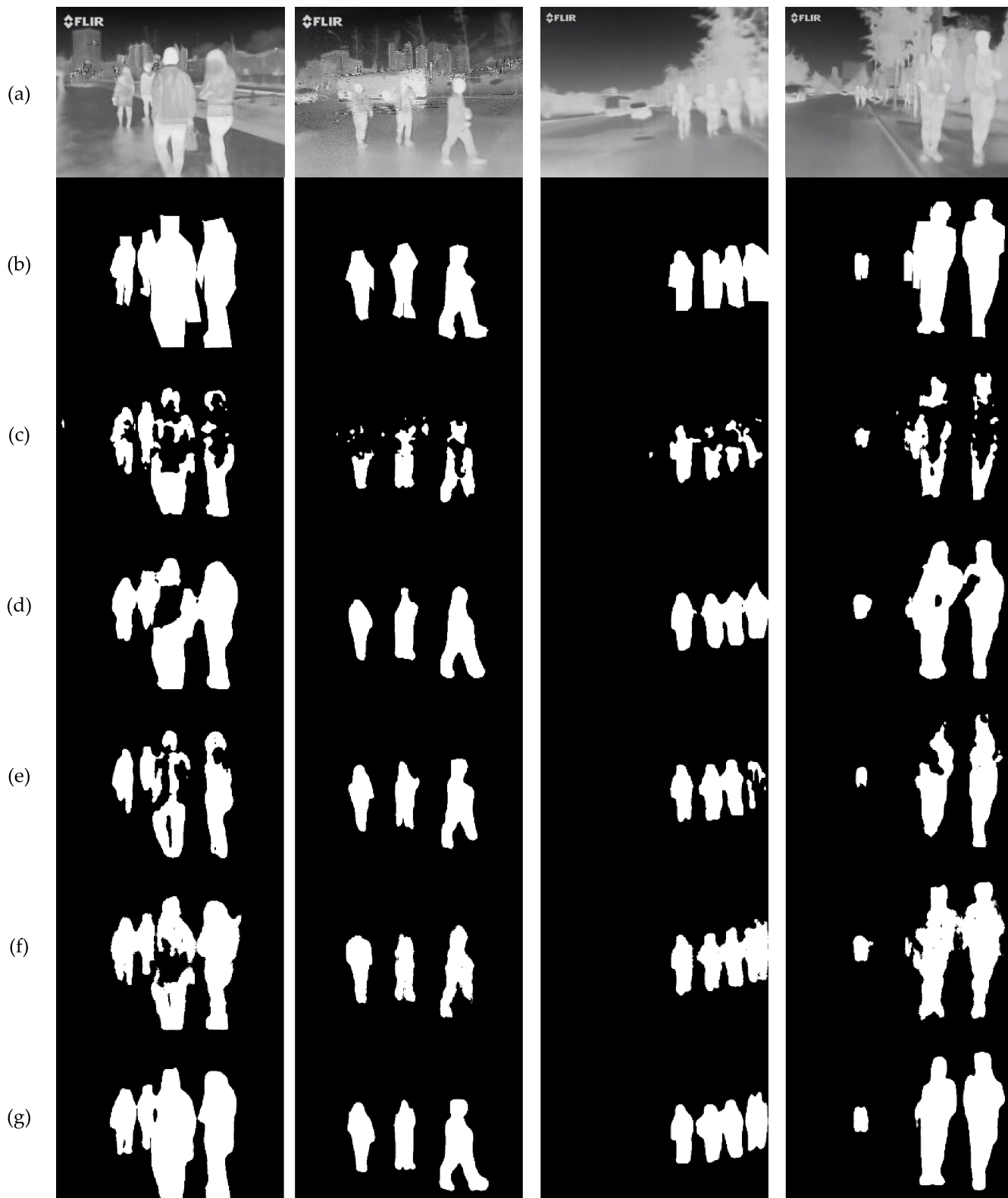


FIGURE 12. Test image results display of LFPD dataset. (a) is the input image, (b) is the label image, (c) is the UNet, (d) is the FCN8, (e) is the PSPNet, (f) is the SegNet, (g) is FSGNet.

uses multiple receptive fields and feature enhancement modules, so that features with more expression capabilities are selected. At the same time, the semantic guidance

structure is used to transmit the obtained feature map with strong expressive power to the decoding step, which makes the network more capable of segmenting far infrared images.

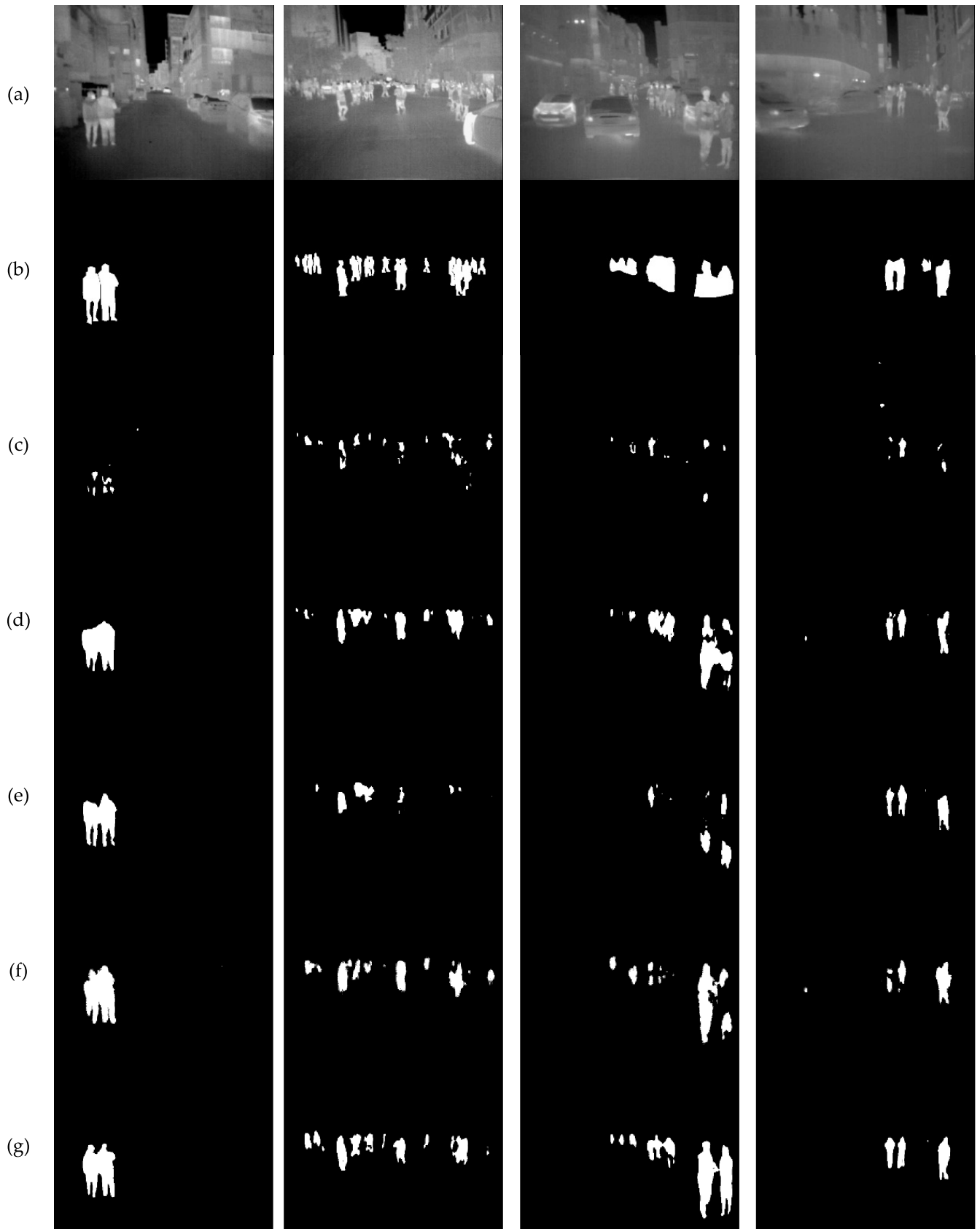


FIGURE 13. Test image results display of KAIST dataset. (a) is the input image, (b) is the label image, (c) is the UNet, (d) is the FCN8, (e) is the PSPNet, (f) is the SegNet, (g) is FSGNet.

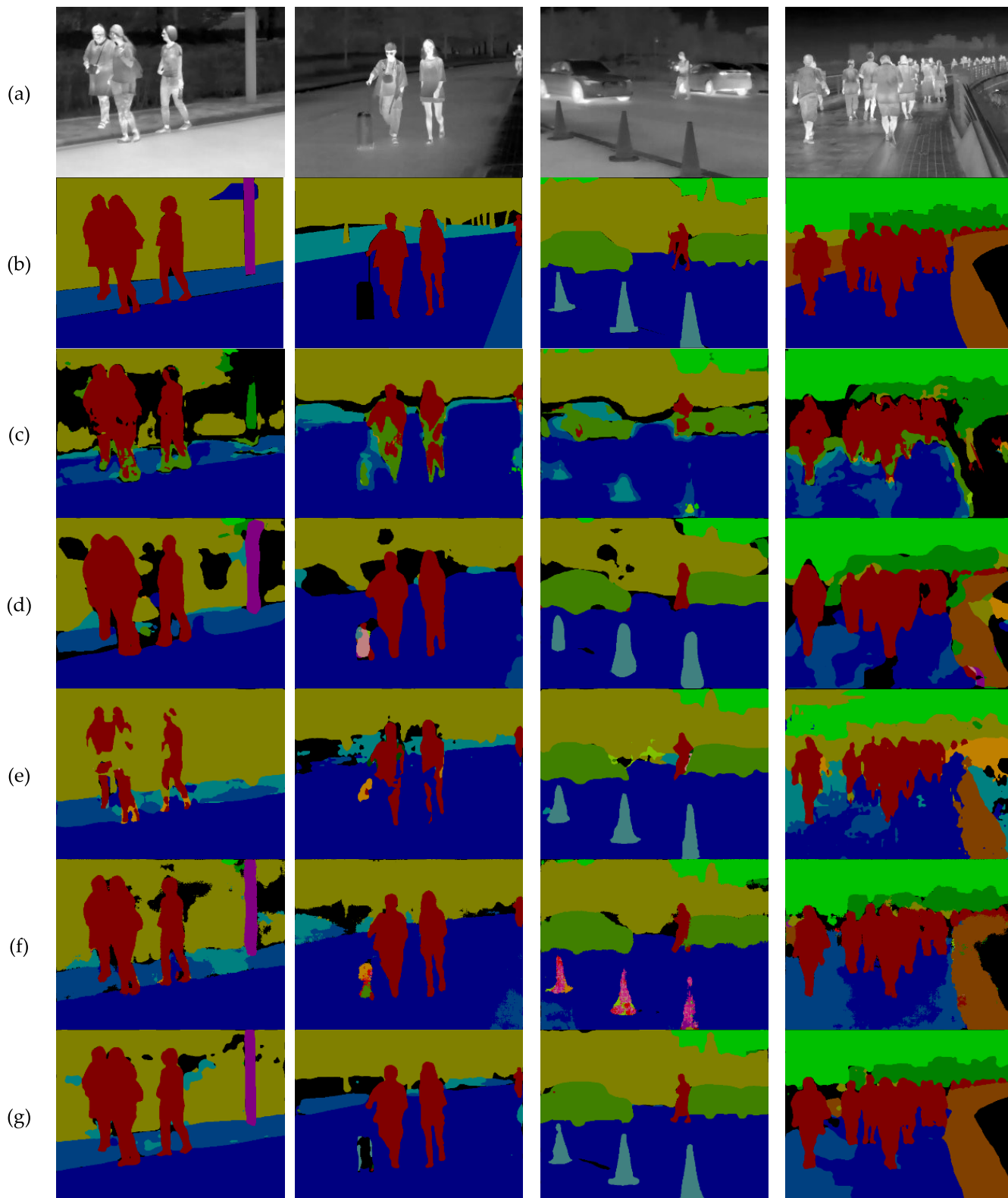


FIGURE 14. Test image results display of SODA dataset. (a) is the input image, (b) is the label image, (c) is the UNet, (d) is the FCN8, (e) is the PSPNet, (f) is the SegNet, (g) is FSGNet.

Therefore, our method has a strong semantic extraction capability for object segmentation in far infrared images. And for some special cases, our method not only depends on the

brightness information of the object, but also considers its outline and appearance information, so as to obtain a stronger segmentation ability.

TABLE 5. FSGNet ablation experiments on three datasets (Use mIoU).

Dataset	Semantic guidance structure	Multi-receptive field	Feature enhancement module	Mean IoU(%)
Baseline: ResNet 50				
LFPD				65.47
	✓			66.52
	✓	✓		69.81
	✓		✓	68.23
	✓	✓	✓	70.59
KAIST	✓			28.01
	✓	✓		28.65
	✓		✓	30.79
	✓	✓	✓	30.26
	✓	✓	✓	30.98
SODA	✓			54.39
	✓	✓		55.17
	✓		✓	58.87
	✓	✓	✓	56.52
	✓	✓	✓	60.67

E. ABLATION STUDY

The last experiment in this article aims to verify the functionality of the module we designed. In Table 5, the basic experimental framework uses ResNet50 to realize an encoding-decoding network. On this basis, we first add a semantic guidance structure to the feature map of the last layer so that the network has a total of three branches. Adding this structure improves the very small segmentation performance of the network. Our multi-receptive field module and feature enhancement module are designed based on the semantic guidance structure, so two different experiments are designed to verify the effectiveness of the module. From the test results of the three datasets, we can see that the multi-receptive field module has a greater impact on the network segmentation performance than the feature enhancement module. However, all of them have improved the segmentation performance of the IoU. Finally, we add all modules to the basic network and obtain the best results. The best mIoU results of the LFPD, KAIST and SODA datasets were 70.59%, 30.98% and 60.67%, respectively. All the above experiments prove the effectiveness of the modules proposed in this article, and they have a very important influence on the segmentation of far-infrared objects.

According to the experimental results, the values of Mean IoU have increased by 3.29% for LFPD datasets, 2.14% for KAIST datasets, and 3.70% for SODA datasets after adding multiscale, respectively. As shown from the experimental data, the performance has been improved obviously by adding multiscale.

V. CONCLUSION

In this study, far-infrared images and deep learning technology are used to segment the objects in the driving scene. Considering the diversity of the dataset, this

article releases a low-resolution far-infrared pedestrian dataset. Next, a convolutional neural network algorithm for far-infrared image segmentation is designed. Based on the general encoding-decoding structure, a multi-receptive field module is designed to expand the visual range and reduce the impact of missing texture features on segmentation. The feature enhancement module is used to select representative feature channels, including the contours. It is used to solve the problem of object temperature reversal and background similarity. Finally, the semantic guidance structure is used to help obtain more detailed information between features of different scales. Our method is tested on three diverse datasets. The experimental results show that the FSGNet has stronger segmentation capabilities for far-infrared images than many popular segmentation networks. It has achieved the best results in all indicators. In the future, we will compare more segmentation methods to make the network structure of this article more convincing. At the same time, we will also work to improve the real-time performance of the network.

ACKNOWLEDGMENT

(Ying Zang and Bo Yu are co-first authors.)

REFERENCES

- [1] W. Wang, J. Zhang, and C. Shen, "Improved human detection and classification in thermal images," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2010, pp. 2313–2316. [Online]. Available: <https://ieeexplore.ieee.org/document/5649946>
- [2] B. Besbes, A. Rogozan, A.-M. Rus, A. Bensrhair, and A. Broggi, "Pedestrian detection in far-infrared daytime images using a hierarchical codebook of SURF," *Sensors*, vol. 15, no. 4, pp. 8570–8594, Apr. 2015.
- [3] X. Zhao, Z. He, S. Zhang, and D. Liang, "Robust pedestrian detection in thermal infrared imagery using a shape distribution histogram feature and modified sparse representation classification," *Pattern Recognit.*, vol. 48, no. 6, pp. 1947–1960, Jun. 2015.
- [4] Y. Chen and H. Shin, "Pedestrian detection at night in infrared images using an attention-guided encoder-decoder convolutional neural network," *Appl. Sci.*, vol. 10, no. 3, p. 809, Jan. 2020.
- [5] J. Park, J. Chen, Y. K. Cho, D. Y. Kang, and B. J. Son, "CNN-based person detection using infrared images for night-time intrusion warning systems," *Sensors*, vol. 20, no. 1, p. 34, Dec. 2019.
- [6] N. Deckers, M. Yildirim, and R. Reulke, "Sensor fusion-based learning for the improvement of person segmentation by means of a low-resolution thermal infrared array sensor," in *Proc. Int. Conf. Comput. Graph. Digit. Image Process. CGDIP*, 2017, pp. 1–6, doi: [10.1145/3110224.3110237](https://doi.org/10.1145/3110224.3110237).
- [7] R. Brehar, F. Vancea, T. Marita, and S. Nedeveschi, "A deep learning approach for pedestrian segmentation in infrared images," in *Proc. IEEE 14th Int. Conf. Intell. Comput. Commun. Process. (ICCP)*, Sep. 2018, pp. 253–258. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/8516630>
- [8] Q. Liu and J. Zhuang, "A generalized thresholding algorithm of pedestrian segmentation for far-infrared images," in *Proc. IEEE Int. Conf. Imag. Syst. Techn.*, Jul. 2012, pp. 338–343. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/6295515>
- [9] G. Wang, Q. Liu, and Z. Wang, "Segmentation of far-infrared pedestrians for advanced driver-assistance systems," in *Proc. IEEE Int. Conf. Imag. Syst. Techn. (IST)*, Sep. 2015, pp. 1–6. [Online]. Available: <https://ieeexplore.ieee.org/document/7294521>
- [10] D. Zhou, S. Qiu, Y. Song, and K. Xia, "A pedestrian extraction algorithm based on single infrared image," *Infr. Phys. Technol.*, vol. 105, Mar. 2020, Art. no. 103236.
- [11] Y. Qiao, Z. Wei, and Y. Zhao, "Thermal infrared pedestrian image segmentation using level set method," *Sensors*, vol. 17, no. 8, p. 1811, Aug. 2017.

[12] D. Olmeda, A. D. L. Escalera, and J. M. Armingol, "Phase spread segmentation of pedestrians in far infrared images," in *Advanced Microsystems for Automotive Applications*. Berlin, Germany: Springer, 2011, doi: [10.1007/978-3-642-21381-6_13](https://doi.org/10.1007/978-3-642-21381-6_13).

[13] X. Bai, Y. Wang, H. Liu, and S. Guo, "Symmetry information based fuzzy clustering for infrared pedestrian segmentation," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 4, pp. 1946–1959, Aug. 2018.

[14] K. Piniarski and P. Pawlowski, "Efficient pedestrian detection with enhanced object segmentation in far IR night vision," in *Proc. Signal Process., Algorithms, Archit., Arrangements, Appl. (SPA)*, Sep. 2017, pp. 160–165. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/8166857>

[15] D. Jin, X. Bai, and Y. Wang, "Integrating structural symmetry and local homoplasmy information in intuitionistic fuzzy clustering for infrared pedestrian segmentation," *IEEE Trans. Syst., Man, Cybern. Syst.*, early access, Aug. 28, 2019, doi: [10.1109/TSMC.2019.2931699](https://doi.org/10.1109/TSMC.2019.2931699).

[16] F. He, Y. Guo, and C. Gao, "An improved pulse coupled neural network with spectral residual for infrared pedestrian segmentation," *Infr. Phys. Technol.*, vol. 87, pp. 22–30, Dec. 2017.

[17] Y. Sun, W. Zuo, and M. Liu, "RTFNet: RGB-thermal fusion network for semantic segmentation of urban scenes," *IEEE Robot. Autom. Lett.*, vol. 4, no. 3, pp. 2576–2583, Jul. 2019.

[18] S. S. Shivakumar, N. Rodrigues, A. Zhou, I. D. Miller, V. Kumar, and C. J. Taylor, "PST900: RGB-thermal calibration, dataset and segmentation network," 2019, *arXiv:1909.10980*. [Online]. Available: <http://arxiv.org/abs/1909.10980>

[19] J. Zhao and S.-C.-S. Cheung, "Human segmentation by geometrically fusing visible-light and thermal imageries," *Multimedia Tools Appl.*, vol. 73, no. 1, pp. 61–89, Nov. 2014.

[20] N. Salamati, D. Larlus, G. Csurka, and S. Süstrunk, "Semantic image segmentation using visible and near-infrared channels," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 461–471, doi: [10.1007/978-3-642-33868-7_46](https://doi.org/10.1007/978-3-642-33868-7_46).

[21] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440. [Online]. Available: <https://arxiv.org/abs/1411.4038>

[22] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[23] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2881–2890, [Online]. Available: <https://arxiv.org/abs/1612.01105>

[24] Z. He, S. Tang, J. Yang, Y. Cao, M. Ying Yang, and Y. Cao, "Cascaded deep networks with multiple receptive fields for infrared image super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 8, pp. 2310–2322, Aug. 2019.

[25] Z. He, Y. Cao, L. Du, B. Xu, J. Yang, Y. Cao, S. Tang, and Y. Zhuang, "MRFN: Multi-Receptive-Field network for fast and accurate single image super-resolution," *IEEE Trans. Multimedia*, vol. 22, no. 4, pp. 1042–1054, Apr. 2020.

[26] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241, doi: [10.1007/978-3-319-24574-4_28](https://doi.org/10.1007/978-3-319-24574-4_28).

[27] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141, [Online]. Available: <https://arxiv.org/abs/1709.01507>

[28] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[29] S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1037–1045. [Online]. Available: <https://www.computer.org/csdl/proceedings-article/cvpr/2015/07298706/12OmNwDAnC>

[30] Q. Hou, M. M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr, "Deeply supervised salient object detection with short connections," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3203–3212, [Online]. Available: <https://arxiv.org/abs/1611.04849>

[31] D. Ghose, S. M. Desai, and S. Bhattacharya, "Pedestrian detection in thermal images using saliency maps," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2019, pp. 988–997. [Online]. Available: https://openaccess.thecvf.com/content_CVPRW_2019/html/PBVS/Ghose_Pedestrian_Detection_in_Thermal_Images_Using_Saliency_Maps_CVPRW_2019_paper.html



YING ZANG was born in December 1981. She is currently pursuing the Ph.D. degree with the University of Chinese Academy of Sciences. She is an AI Engineer with the School of Information Engineering, Huzhou University. She is working on the research of object detection, semantic segmentation, and image processing.



BO YU was born in July 1994. He is currently pursuing the Ph.D. degree in artificial intelligence with Jilin University. He is working on the research of computer vision, digital image processing, and infrared technology.



LONGJIAO YU was born in September 1995. She is currently a Graduate Student with the College of Information Science and Technology, Dalian Maritime University. She is working on the research of deep learning, multiobject detection, and pedestrian re-identification.



DONGSHENG YANG was born in 1965. He is currently a Researcher with the Shenyang Institute of Computing Technology, Chinese Academy of Sciences. He is working on the research of real-time control systems, numerical control technique, and computer vision.



QINGSHAN LIU was born in December 1979. He is currently an Associate Professor with the School of Life Sciences, Huzhou University. He is working on the research of ionic liquids, deep learning, and image processing.

...