ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)

https://doi.org/10.17559/TV-20241219002204
Original scientific paper

Research on Human Body Detection Algorithm in Car Cabin Based on RGB-IR Dual Light Fusion

Siyu CHEN, Juhua HUANG, Yinyin LIU*, Fengping XU

Abstract: A strategy of RGB-IR dual optical fusion is proposed, that is, each component of multiple color Spaces is detected individually, several channels with the best human detection performance are counted, and the prospects of these channels are fused to obtain the final human detection result. In addition, to address the challenge of insufficient contour prediction accuracy in the semantic segmentation task of RGB-IR dual-modal images, we propose an innovative multi-scale contour enhancement dual-modal semantic segmentation method, and introduce a novel location and channel attention mechanism module, which can effectively promote cross-scale feature fusion. Thus, the contour prediction ability of various scales can be accurately improved. Through the detection of human body in car cabin based on infrared and visible light mode fusion, in order to verify the effectiveness of the proposed modal fusion detection algorithm for the actual traffic scene, the practical application of the model is realized through the fusion detection algorithm. The experimental results show that the detection results of this algorithm are better than those of human body based on single color space in complex scenes, and the algorithm can effectively deal with dynamic background problems.

Keywords: dual light fusion; human body detection in the car compartment; mode fusion; semantic segmentation; RGB-IR

1 INTRODUCTION

As a hot topic in computer vision research, human detection in vehicle cabin has received more and more attention, and has been widely used in many fields such as human-computer interaction, video surveillance, motion analysis and so on. For example, gestures, human posture and other information can usually be used for harmonious interaction with computers [1, 2], and such systems often include human body detection, human body tracking and other modules, among which human body detection is the basic component of the system, and its performance directly or indirectly affects the performance of human body tracking and human behavior recognition. Therefore, human detection plays a key role in systems based on computer vision technology. The complexity of the practical application environment of moving human body detection determines the high reliability and high practicability of human body motion detection.

Visible images are the main focus of attention because they conform to the visual perception characteristics of the human eye, and can capture high-resolution images with clear details and textures, thus providing powerful scene description capabilities. However, such images have strong photosensitive characteristics, and once the external conditions change (such as overexposure, insufficient light, fog, rain and snow, etc.), the image quality will be greatly reduced, so that accurate information of the target cannot be obtained, and further affect the execution of subsequent tasks [3]. In response to this shortcoming, targets with higher surface temperatures (such as the human body and the interior of a car) are more prominent in the image. This feature allows infrared images to capture clear silhouettes of targets in harsh environments or at night. However, due to the restriction of imaging mechanism, infrared images are fuzzy as a whole, and it is difficult to fully express a large amount of detailed information, while there is a large amount of interference noise, resulting in reduced recognition accuracy [4].

This research focuses on infrared and visible image fusion technology in traffic environment to optimize the human detection capability inside the vehicle cabin. The goal is to achieve a more comprehensive and accurate perception of the car interior and human body in traffic scenes by integrating information from these two image sources, and thus improve the safety performance and reliability of the automatic driving system. To solve the two major problems of camouflage phenomenon and dynamic background interference, this study innovatively proposes a multi-color space fusion strategy, which is composed of two core components: one is dedicated to solving the camouflage problem, and the other is dedicated to addressing the challenge of dynamic background. In the case of camouflage problems, human detection results based on a single color space are not very ideal. The general step of the algorithm in this paper is to find several channels with the best effect of detecting moving human body in multiple color Spaces, and fuse the prospects of these channels to obtain complete human body detection results. The first part is the introduction, the second part is related work, The third part is research framework of human body detection algorithm in car cabin based on RGB-IR dual optical fusion, the fourth part is detection algorithm of human body in car cabin based on attention mode fusion, the fifth part is simulation verification, and the sixth part is conclusion.

2 RELATED WORK

At present, the target tracking algorithm mainly adopts three main steps: selecting the target feature and representation method, tracking algorithm structure and target position prediction in subsequent image frames. The feature selection of the target is an intuitive information description of the target, which can select the contour, texture and color of the target. The most similar regions are matched from subsequent image sequences. Target prediction is a reasonable prediction of the target position through some filters, which can track the target more accurately. At present, for different application scenarios, tracking algorithms are mainly summarized as follows:

(1) Human body target tracking based on deep learning generates a mapping function to reflect the relationship between input and output to detect and track the human

body. The MOSSE tracking algorithm [5] and KCF tracking algorithm [6] are proposed, and the features of images extracted by intensive sampling are combined with Fourier transform (FFT) for classification training. In order to improve the effect of feature extraction, this study proposes a sufficiency semi-supervised feature extraction method, and specially designs a feature extraction algorithm for infrared images [7]. Ensure that the feature information has a higher degree of differentiation, and further improve the accuracy of target tracking. In recent years, machine learning-based methods have become an important trend in target tracking algorithms [8]. Using diverse network structure for training and classification, the classifier can better adapt to complex background changes. Among target tracking algorithms, online Boosting algorithm [9] and online AdaBoost algorithm [10] use rectangular features to describe the local shape of the human body. These algorithms show good results for human body tracking at different distances in intelligent monitoring systems. In addition, there is a method to track video data by using neural network offline training [11], which allows the target model to be updated online in real time. Once the detection tracking error occurs, the new online target model can be immediately switched to match, thus improving the flexibility and accuracy of target tracking.

(2) Based on the tracking method of 3D model, HOG features are proposed to complete the detection of human body structure [12] and are widely used in the representation of behavioral characteristics, but they cannot describe the diversity of human body forms. An adaptive target intensity hypothesis density filtering algorithm [13] is proposed, and sample points generated by the traceless transform are used to improve the filtering strategy, effectively solving the problem of normalization imbalance in the tracking algorithm. Adaptive linear prediction method is proposed to carry out non-recursive background subtraction operation [14], to conduct background modeling of image information, and to predict the position of target objects through the background model.

(3) Based on the tracking method of feature point matching, the similarity of the feature point fields of two adjacent frames of images is evaluated to determine the corresponding points to complete the tracking. Commonly used feature point algorithms include SIFT [15], KLT [16], Harris [17] and Moravec feature [18]. Harris is an Angle detection algorithm that calculates the first derivative of X, Y (vertical and horizontal) to represent the gray difference in each direction. The tracking algorithm using SIFT (scale invariant Feature Transform) features [19], the core of which is to achieve dynamic update of the model by accumulating stable features. The algorithm divides the tracking task into two main stages: feature matching and model updating. The matching algorithm combining KTL and SIFT features [20, 21] can improve the matching accuracy in the tracking process. However, the change between the matching feature point and the relative position has a great influence on the tracking result. When the human body changes shape or is blocked in the image, the matching tracking with feature points has a good adaptability compared with other features.

(4) The association algorithm is used to obtain the association region consistent with the target [22, 23], which has low complexity and can efficiently detect the target in the scene with obvious differences in the background. Interframe difference method and level set are used to detect moving targets [24] to realize adaptive target detection and tracking in dynamic environment and to overcome the target stretching and void defects caused by target extraction by difference method.

In recent years, some researchers have begun to introduce infrared images to compensate for the shortcomings of visible light images alone. Visible light images are limited to wavelengths between 0.4 and 0.76 microns, while infrared images cover a wide wavelength range of 0.1 to 100 microns, effectively complementing a large amount of information beyond visible light. Infrared images can provide relatively complete information and are not affected by various lighting changes. Using infrared images as a complement to visible images can enhance information integrity in a variety of complex lighting conditions. Therefore, the use of visible and infrared dual-band images for semantic segmentation is expected to significantly improve the stability and adaptability of autonomous driving systems. However, more modes mean more information [25], what information to fuse, when to fuse, and how to fuse are the challenges facing the current RGB-IR dual-band semantic segmentation problem. According to the time of fusion, the current working network structure can be divided into three categories: decoding side fusion, coding side fusion, and coding-decoder side fusion. The work of decoding side fusion includes: MFNet [26], the network fused the features of the two modes with a jump connection during the process of down sampling, and used the mini-inception module with void convolution to build an independent encoder to process visible and infrared images, and then carried out feature fusion in the decoder part. Since no pre-trained model was used, although the speed was superior, but the accuracy is low. FuNNet fuses the information of the two bands in the decoding process and uses block convolution [27] to reduce the number of parameters in the model. The proposed PSTNet introduces global semantic information to enhance segmentation effect [28]. Coding side fusion work includes: the proposed RTFNet uses pre-trained ResNet [29, 30] as the encoder. AFNet [31] realizes the fusion of two band feature maps at the base layer of the encoder. The work of coding-decoder fusion includes: FuseSeg is proposed to add the feature maps of two bands in the encoder.

3 RESEARCH FRAMEWORK OF HUMAN BODY DETECTION ALGORITHM IN CAR CABIN BASED ON RGB-IR DUAL OPTICAL FUSION

3.1 Human Body Detection Algorithm Framework Based on RGB-IR Fusion

By using cross-modal feature correction module and explicit attention enhancement module before the features enter the fusion network, the model training process is guided to pay attention to the fusion quality of the features, so that the network can better pay attention to the features related to the human detection task in the car cabin, thereby improving the model's ability to locate the target, and enhancing the detection accuracy and stability of the algorithm. The abbreviations is added in Tab. 1.

Table 1 Abbreviations

Abbreviations	Full name
FFT	Fourier transform
RGB-IR	RGB-Infrared
RGD	Random Gradient Descent

In view of the lack of feature extraction capability of the traditional two-branch trunk, how to achieve full modal interaction while maintaining the recognition of each mode, we propose an innovative human body detection model in the automotive cabin, which uses the self-attention mechanism of Transformer to achieve modal fusion. The design was inspired by the remarkable ability of self-attention mechanisms to capture global information. A self-attention-guided modal interaction module is designed to fully mine and aggregate the context information within and between a single mode. The traditional isolated two-branch trunk feature extraction structure is broken, and a two-branch cross-modal interaction trunk combined with CNN and Transformer is constructed to learn the correlation between modes at different levels and guide the original trunk to constantly adjust the information components. In the feature extraction stage, the mode information is fully interactive while maintaining the unique identification of each mode, ensuring the quality of the fusion features, and thus improving the detection ability of the algorithm in the road traffic scene. The architecture diagram is shown in Fig. 1.

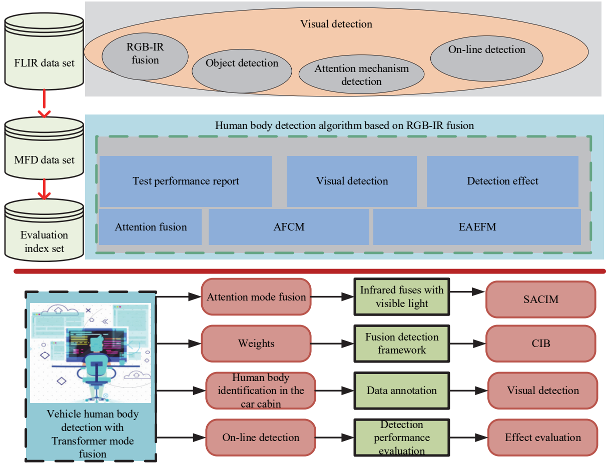


Figure 1 Human body detection algorithm block diagram in the car cabin

As shown in Fig. 1, the algorithm first predicts target contours of different scales through fusion features of various scales between encoders, and obtains more valuable pixels and channels and enhances the features through location attention and channel attention.

3.2 Two-Band Semantic Segmentation Algorithm Based on Contour Enhancement

When different objects have a similar color or appearance, it is usually not good to separate them. After dual-band fusion, the feature map of a channel is obtained by using 1 × 1 convolution. The probability of whether the feature is a contour is calculated by Sigmoid activation function. The obtained semantic contour information is multiplied by pixels with the fusion feature map of the input to enhance the contour of the feature map. Finally, the enhanced feature map is added to the feature map of the input to form a residual connection to avoid information loss. The symbolic variables are shown in Tab. 2.

Table 2 Symbolic variables

Variables	Meaning
y_i	true contour
$\hat{y}_i$	predicted profile
r	weight decay rate
C	the number of prediction categories
P_{ij}	number of pixels belonging to class i that are predicted to be class j.
X_i and Y_i	both represent two inputs for connecting channels and adding corresponding feature graphs
F	indicates global average pooling
Z_{vi}	the feature matrix of visible image

The predicted contour is supervised by the real contour label, and the contour label can be obtained by semantic label. The semantic contour supervision loss function uses binary cross entropy, and the formula is shown in Eq. (1):

$$\varsigma_{\text{egde}} = \frac{1}{N}\sum_{i=1}^{N} y_i \log \hat{y}_i + (1 - y_i)\log(1 - \hat{y}_i) \tag{1}$$

where, N represents the number of pixels; y_i represents the true contour; $\hat{y}_i$ represents the predicted profile.

Using this design has two benefits: (1) supervision through cross entropy loss, explicit constraint of contour information, and optimization of encoder features using gradient backpropagation. (2) The reconstructed features contain enhanced semantic contour information.

In order to improve the accuracy of multi-scale fusion feature maps, inspired by SENet [21], this paper weighted feature maps from two aspects of location and channel, and proposed a new location and channel attention module SAC to enhance multi-scale fusion feature maps, as shown in Fig. 2.

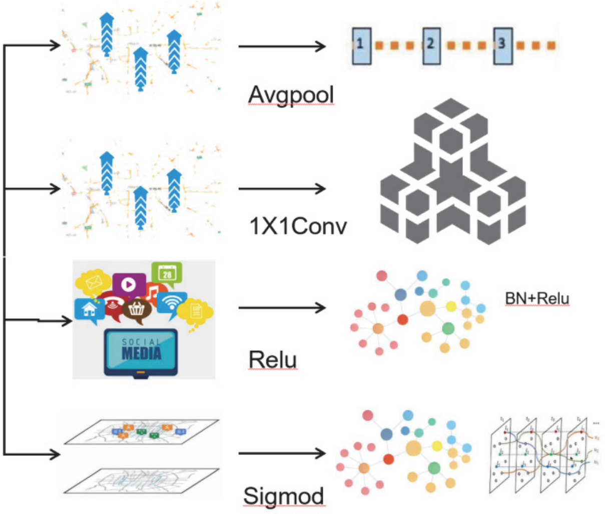


Figure 2 Location and channel attention module

The feature map with enhanced uncertainty only updates the pixels in the uncertain regions, and the attention mechanism still needs to update the weights in these regions to enhance the expression of the feature map. In the training phase, we use an RGD (Random Gradient Descent) optimizer with momentum, which provides a stronger ability to jump out of local optimal or gradient vanishing saddle points. In order to ensure the consistency and fairness of the experiment, the same training hyper parameter configuration was adopted for all the models involved in this paper (including the comparison algorithm): the training batch size was fixed at 8, and the weight decay rate was set to 0.0005. The initial learning rate r is set at 0.01.

$$r_i = r_0(1-\frac{t}{t_{\max}})^{0.8} \tag{2}$$

Generally, the main loss function of a segmented network is the cross-entropy loss function, and the formula is as follows:

$$\varsigma_{\text{egde}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y^{c}_{i}\, \hat{y_i} \log(1-\hat{y_i}) \tag{3}$$

where, N represents the number of pixels, C represents the number of prediction categories, y_i represents the prediction probability of pixel i to category c, and y is the real label. In this paper, Lovasz-loss [12] is combined with cross-entropy loss λ as the total loss function, and the formula is as follows:

$$L_{\text{total}} = L_{ce} + \lambda L_{\text{lovas}} \tag{4}$$

The model proposed in this paper uses the average intersection ratio as the main index to evaluate the performance of semantic segmentation. The calculation formula is as follows:

$$mIoU = \frac{1}{N}\sum_{i=1}^{N}\frac{P_{ii}}{\sum_{j=1}^{N}(P_{ij}+P_{ji})} \tag{5}$$

where N is the number of classes and P_{ij} is the number of pixels belonging to class i that are predicted to be class j. For the MFNetDataset, unlabeled pixels are also taken into account in the calculated metrics. The higher the score of the above evaluation indexes in the segmentation results, the better the segmentation accuracy of the algorithm.

3.3 Human Body Detection Algorithm Based on RGB-IR Space Fusion

The whole process covers three aspects: feature extraction, classification and regression, and its network architecture is shown in Fig. 3. In the feature extraction stage, the input image will go through a multi-feature fusion processing process, aiming to extract more representative features that are beneficial to human detection. These features will then be used in human classification and regression tasks.

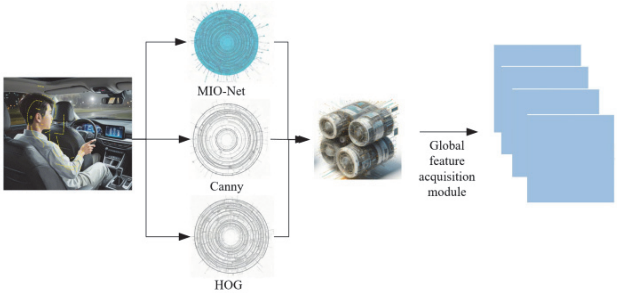


Figure 3 Schematic diagram of human detection network based on multi-feature fusion

The number of channels in the corresponding feature map remains unchanged, and the complete fusion process is as follows:

$$Z_{con} = \sum_{i=1}^{c} X_i * K_i + \sum_{i=1}^{c} Y_i * K_{i+c} \tag{6}$$

where X_i and Y_i both represent two inputs for connecting channels and adding corresponding feature graphs, assuming the number of channels is c.

The fused features contain not only deep semantic information, but also shallow appearance information such as edge and gradient, which is conducive to realizing more accurate human detection.

$$F_f = \sum_{i=1}^{d} D_i * K_i + \sum_{i=1}^{e} E_i * K_{i+d} \tag{7}$$

where, F_j represents the features fused by the channel joining operation, and the number of channels matches the number of output channels of the convolution kernel.

The specific implementation process of the global feature acquisition module is as follows:

1) The original input is a feature graph with dimension (C, H, V) obtained after the fusion of traditional features and depth features. The weight matrix of output dimension (1, H, V) is changed into a matrix of (HIV, 1, 1) through re shape operation, and then the softrnax function is used to obtain the global feature weights.

2) Then, the original input is transformed by 3 × 3 convolution, and the global feature weight obtained in step 1 is multiplied, that is, the global feature weight is applied to the corresponding position of each feature graph of all channels, and the feature value of each output channel is the global context feature shared by the feature graph on this channel.

3) By adding the original input features with the global context features obtained in Step 2, the global context features are integrated into the features of each position of the original input feature image.

The implementation principle of the global feature acquisition module is to obtain the information of the global context by calculating the weighted sum of all position features in the input feature map.

$$z_i = x_i + \sum_{j=1}^{N} \frac{e^{W_k x_j}}{\sum_{m=1}^{N} e^{W_k x_m}} (W_v x_j) \tag{8}$$

On the basis of improving the original algorithm, Kalman model [18] is added to predict the prediction of the target state region, predict and update the position of the target region obtained by each frame image sequence, effectively narrow the detection range of the detector, improve the processing speed of the detector, and track the result of the fast moving target more accurately.

$$X^{'}(k) = X^{'}(k-1) + K_k \times H_k(k-1) \tag{9}$$

The state prediction equation of Kalman predictor is:

$$X^{'}(k+1) = AX^{'}(k-1) + H_k X^{'}(k \mid k-1) \tag{10}$$

Set a rectangular area 4 times larger than the target bounding box output in the previous frame, and determine that this area is the area where the target appears. In this paper, the steps of the color feature histogram probability distribution tracking algorithm improved by RGB-IR space fusion are shown in Tab. 3:

Table 3 Improved distribution tracking algorithm for RGB-IR space fusion

1: Capture the RGB basic image in the human body foreground identification area.
2: Convert RGB image to HSV and YCbCr color space image.
3. Using iterative technology, core points were extracted from HSV and YCbCr respectively for preliminary treatment.
4: Through iterative calculation, determine the offset of the core point and its value range.
5: Set the recognition boundary of the next frame image.
6: Establishes the threshold selection criteria.
7: Update the core point and identify the boundary.
8: Perform core point detection and restore its original position.
9: Determine the target position of the K frame and apply Kalman filter to predict its state.
10: Using Markov model to predict the possible direction of movement of the target.
11: Combine the above information to calculate the target position of the K + 1 frame.

In order to deal with the problem of light intensity change, target occlusion and interference of objects similar to the target in the background, the central point of the detection body is constantly updated to achieve adaptive target detection effect. According to the prediction of the target state region, the detection range is narrowed to improve the processing speed and enhance the robustness of the tracking algorithm.

4 DETECTION ALGORITHM OF HUMAN BODY IN CAR CABIN BASED ON ATTENTION MODE FUSION

4.1 Human Body Detection Algorithm in Car Cabin

In this paper, a new attention-based cross-modal feature correction module is proposed, which is used for the feature correction of two modes before feature fusion. In order to deal with the noise and uncertainty of different modes, input features are processed in both channel and spatial dimensions to achieve the calibration of the overall features, so as to extract and interact the features of infrared and visible light modes more effectively.

Channel feature correction encodes the spatial information of the input features of the two modes into the channel dimension to obtain four attention vectors.

$$\begin{aligned} R &= F_{MLP}(F_{GAP}(F_{RGB})) \\ T &= F_{MLP}(F_{GAP}(F_T)) \end{aligned} \tag{10}$$

where: R and T are weights extracted from visible and infrared features respectively; F indicates global average pooling. In the previous research work, the fusion of features is usually carried out as follows:

$$F_{\text{fusion}} = \varsigma(R) \times F_{RGB} + \varsigma(T) \times F_T \tag{11}$$

The attention weights associated with the two modes are obtained, and then applied to the original visible and infrared features respectively by multiplying channel by element.

In the overall structure of the network, three visible light and infrared fusion modules based on the feature pyramid structure are designed to fuse the two modal features of different depth levels. Assuming that the feature matrix of visible image is Z_{vi}, and the feature matrix C, D of infrared image is Z_{inf}, the process of calculating the feature entropy of $H_i(Z_{vis})$ and $H_i(Z_{\text{inf}})$ of the two images is as follows:

$$\begin{aligned} H_i(Z_{\text{inf}}) &= C_i(Z_{vis}, Z_{\text{inf}}) - D_i(Z_{vis} \parallel Z_{\text{inf}}) \\ H_i(Z_{vis}) &= C_i(Z_{\text{inf}}, Z_{vis}) - D_i(Z_{\text{inf}} \parallel Z_{vis}) \end{aligned} \tag{12}$$

In order to intuitively evaluate the model detection effect, the proposed model was qualitatively compared with the single-modal YOLOvS model and the multi-modal fusion model of SOTA for human object detection tasks in the car cabin. A typical night scene is selected and the detection effect is shown in Fig. 4.

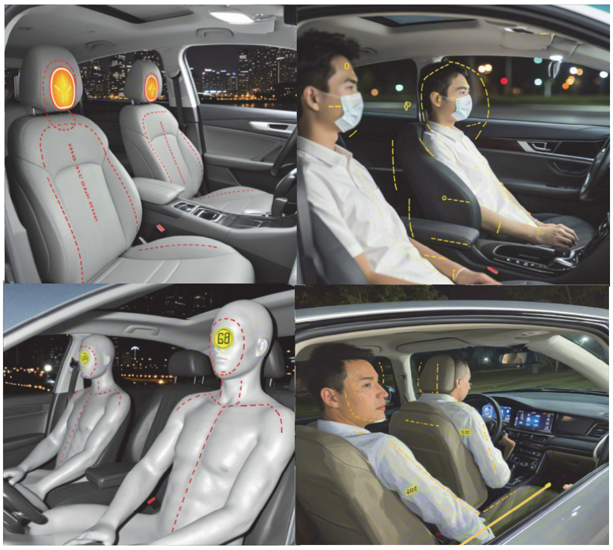


Figure 4 Night scene detection results of the proposed algorithm and some comparison algorithms

Fig. 4 shows the visualized results of the night street scene. As you can see, this scenario is particularly complex. In visible light images, due to various lighting interference, the oncoming car only presents a halo, which is difficult to distinguish. At the same time, the dense human body at the zebra crossing is submerged in the complex background, which is difficult to identify with the naked eye. In the detection results of the single mode visible light model, all human bodies are missed, which is consistent with the naked eye recognition, and there is also a false detection, which mistakenly identifies the background as the human body. In contrast, the single-infrared model detected most human bodies at zebra crossings. The human body features are more prominent than the background in infrared images, thus improving the recognition rate of human body. However, the model still missed a white car in the far right lane and was difficult to distinguish. However, for small and dense human targets, individual detection will still be missed. The algorithm in this chapter accurately detects all the vehicle cabin and human body targets in the scene, showing good detection accuracy.

In summary, the single visible light or infrared mode model has limitations in the recognition effect of human body in the car cabin, but the recognition effect of the combination of the two is better. Especially in the case of poor lighting conditions, the information from a single mode is often insufficient to adequately characterize the target characteristics. It can fully detect the human body target in the cabin of the car, which is of great significance to strengthen the scene understanding and reduce the accident rate.

4.2 Human Body Detection Algorithm in Car Cabin Based on Transformer Mode Fusion

At present, in the field of modal fusion detection, two-branch feature extraction backbone is usually constructed to extract the feature representation of different levels of the two modes respectively, and then simple methods such as element-by-element addition and element-by-element average are used to combine the features. In this study, two identical CSPDarknet53 networks were used to extract the features of the input sources respectively, and then the fused features were obtained by directly adding the two modal features of 8 times and 32 times

$$m = f(r, f(r), f(t)) \tag{13}$$

where: f represents the feature extraction function of the trunk; r and t represent visible and infrared features, respectively.

The proposed algorithm captures the correlation between the two modes at different feature levels and feeds back to each branch to guide the original branch to constantly correct the output and adjust the information component. This correction and adjustment process is iterated throughout the network, gradually optimizing the information components of specific modes on each branch, and finally superimposed the corrected and adjusted features to obtain fusion features with stronger complementarity.

Subsequently, multiple sensing layers were deployed to fine-tune the features generated by the W-MSA (Window Multi-head Self-Attention Mechanism) by introducing additional nonlinear transformations, further improving the expressibility of the model. At the same time, Layer Normalization (LN) is always performed after W-MSA and MLP (Multi-Layer Perceptron) to ensure model stability. In addition, residual connections are constructed to alleviate the problem of disappearing gradients in deep networks and accelerate training.

A picture of F data set is randomly selected for actual detection. In Fig. 5, the characteristic heat map comparison of YOLOvS models based on two different backbone networks in the actual detection process is shown.

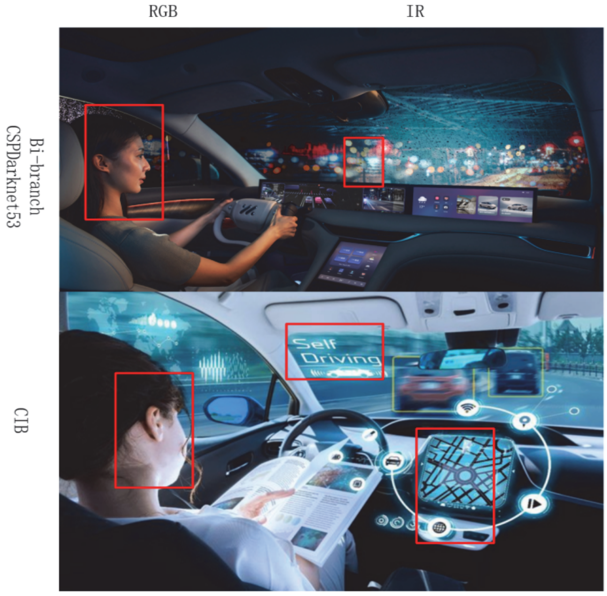


Figure 5 Feature visualization results

As shown in Fig. 5, one model uses the basic two-branch CSPDarknet53 as its core network architecture, while the other model innovatively introduces the algorithm of SACIM as the backbone network. It can be clearly observed that there are significant differences in feature extraction and activation between them. In particular, the model based on the proposed algorithm shows stronger feature concentration in the process of feature extraction. In contrast, the distribution of activation features of the model using basic two-branch CSPDarknet53 as the main trunk network shows a more obvious dispersion, which means that the degree of differentiation between features is not clear. In addition, there is a miscalculation in the model, that is, the background features are incorrectly identified as the activation features of the target class, which are marked with arrows in the graph. This wrong judgment results in the false detection or omission of the model in the process of detection. The visualization results once again demonstrate the effectiveness of the proposed method in extracting complementarity between modes.

5 SIMULATION VERIFICATION

The training and validation of all models in this study were performed on a computer equipped with an Intel Core i7-10700F CPU and a NIVIDA RTX 3080 12G GPU. The following experimental parameters were determined: the size of the two modal images for training input was fixed to 640 × 640, 16 images were randomly selected for each iteration, and data enhancement methods were used to enrich input diversity and accelerate model convergence.

SGD was used to optimize the entire network for enough rounds, the initial learning rate was set to 0.01, and the weight decay gradient and momentum Settings were set to 0.0005 and 0.937, respectively.

In order to verify the effectiveness of the improved method, six groups of controlled experiments were performed in Tab. 3. In experiment 1, two-branch CSPDarkNet53 was used as the backbone feature extraction network of the algorithm, and features of different modes with different depths were extracted respectively. Then, the feature maps corresponding to 8×, 16× and 32× downsampling of the two modes are directly added and fused, and the experimental results are input into the neck network as the benchmark results. Experiment 2, Experiment 3 and experiment 4 introduce and propose three improvements on the benchmark model respectively. According to the experimental results, compared with the single point improvement of the benchmark model, the improvement range of mAP 50 and mAP is 0.9%-3.7% and 0.5%-3.7%, respectively. This shows that each of the proposed improvements is effective for improving the accuracy of the algorithm when it is acted on alone.

Table 2 The ablation experiment of the algorithm is presented

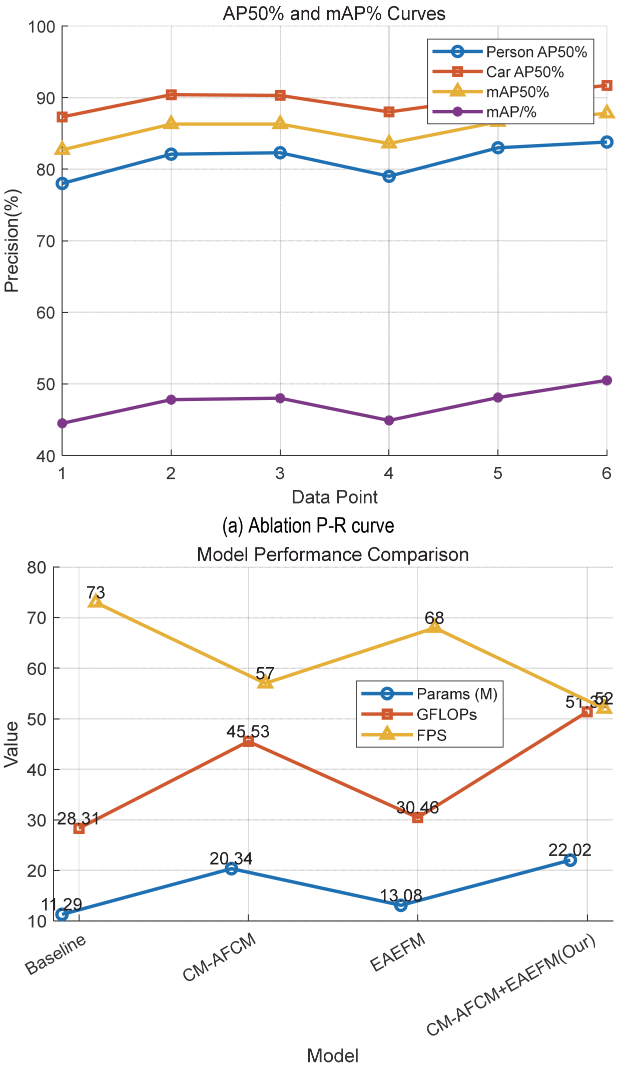
ID	AP50 / %		mAP50 / %	mAP / %
	Person	Car		
1	78	87.3	82.7	44.5
2	82.1	90.4	86.3	47.8
3	82.3	90.3	86.3	48
4	79	88	83.6	44.9
5	83	90.1	86.6	48.1
6	83.8	91.7	87.8	50.5

Experiments 5 and 6 are to verify the effectiveness of the improved method combination. Among them, complete improvement measures were introduced in experiment 6 to form the proposed *t* algorithm, which significantly improved the detection accuracy, and the two accuracy indexes increased by 5.1% and 6.1% compared with the benchmark respectively, and finally achieved the detection accuracy of 91.7% in the car cabin and 83.8% in the human body. In particular, for the category of human body with greater difficulty in identification, the detection accuracy has been significantly improved, with AP50 increased by 5.8%. The improvement has improved the detection accuracy, but also inevitably brought about the increase in the number of model parameters, computational complexity and detection speed, and the FPS has decreased by 21 frames //s, as shown in Tab. 4 for specific changes. In addition, the P-R curve and the training loss curve before and after the improvement were further drawn to compare the ablation experiment results more intuitively, as shown in Fig. 6.

Table 3 Changes in the number of model parameters, computational complexity and detection speed

Model	Input pixel	Params	GFLOPs	FPS
B aseline	640 × 640	11.29	28.31	73
+CM-AFCM	640 × 640	20.34	45.53	57
+EAEFM	640 × 640	13.08	30.46	68
+CM-AFCM+EAEFM(Our)	640 × 640	22.02	51.38	52

The change curves of mAP50 and mAP indexes in the training process of relevant models are drawn in Fig. 7. It can be clearly observed that both single-mode and multi-mode models achieve full convergence under our experimental parameter Settings, which verifies the rationality of the experimental parameter Settings. However, the single mode visible light model shows great instability at the initial stage of training, the accuracy fluctuates significantly, and the inflection point of the model convergence appears late. Although the infrared mode situation has improved, the final two accuracy indicators are improved by about 10%. This confirms the advantages of infrared detection of the human body in the car cabin under the influence of light compared to visible light detection. The stability and accuracy of the model are improved compared with that of a single mode, but the improvement is not significant. A lot of noise information is introduced, and the discriminant feature fusion is not sufficient. In contrast, the AMF-Net training process is more stable, and the training results are excellent, and the best detection accuracy is achieved.



(a) Ablation P-R curve

(b) Improved pre and post training loss curve

Figure 6 P-R curve of ablation experiment and training loss curve before and after improvement

In order to further demonstrate that the model proposed in this paper can effectively integrate the information of RGB image and infrared image, the ablation experiment is carried out on the data set, and RGB and RGB-IR are respectively used as inputs of MFFNet network. When RGB is used as network input, the MFFNet network is fine-tuned and IR detail branches are removed. The overall segmentation performance comparison is shown in Fig. 8, and the category segmentation performance comparison is shown in Fig. 9. The unfilled bar chart is the experimental result of RGB image as network input, while the filled bar chart is the experimental result of RGB-IR image as network input.

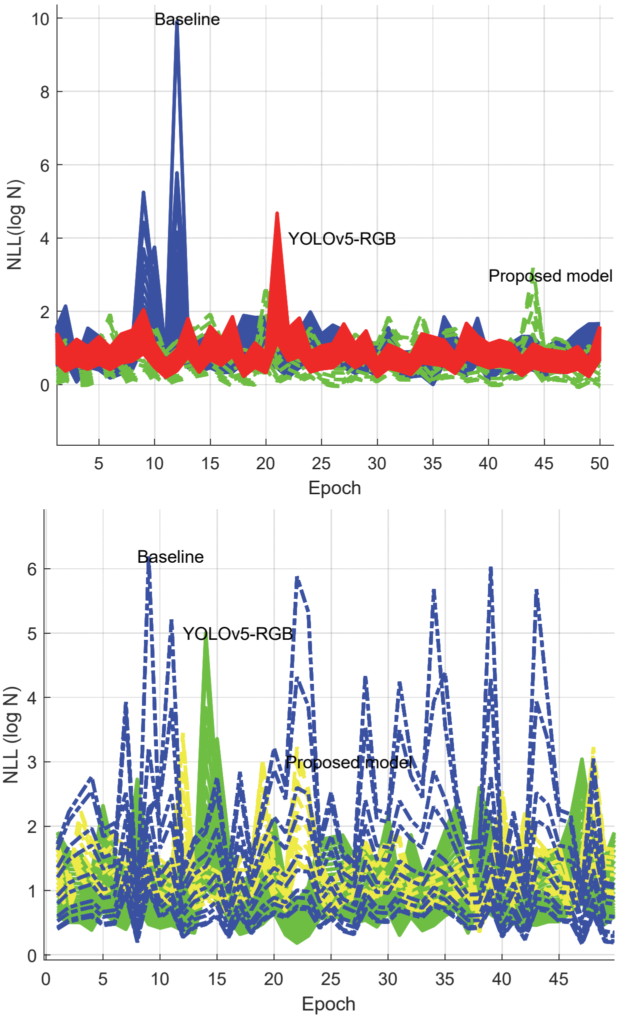


Figure 7 Model training comparison

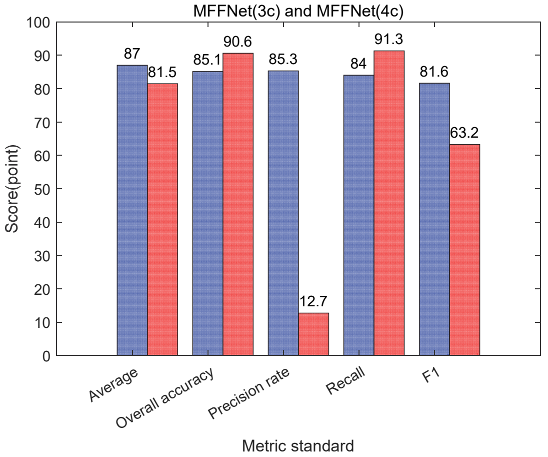


Figure 8 RGB and RGB-IR are respectively input to the MFFNet network global segmentation performance

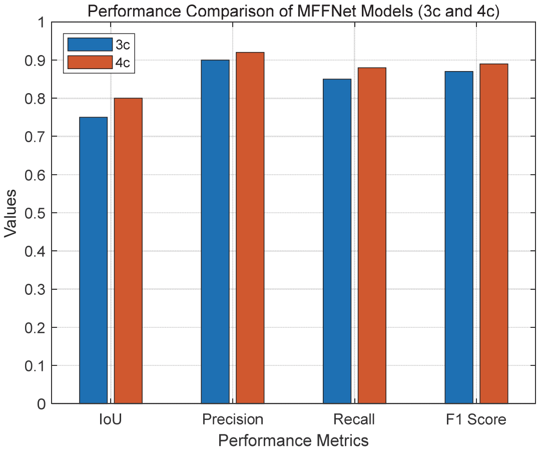


Figure 9 RGB and RGB-IR are respectively input to the MFFNet network car class segmentation performance

The performance improvement of AP^{50} indicates that there are indeed many multi-scale features that need to be learned during the fusion process. This also indicates that the proposed GrFANet performs better under low requirements than the accurate human key point detection. In order to more intuitively reflect the advantages of the balance between the number of parameters and the accuracy, as shown in Fig. 10, it can be seen that the number of parameters has a great advantage compared with other networks. Compared with the proposed network, the number of parameters increases slightly due to the introduction of a new attention mechanism, but the accuracy is further close to nearly twice the number of parameters, and the advanced performance of network detection is realized.

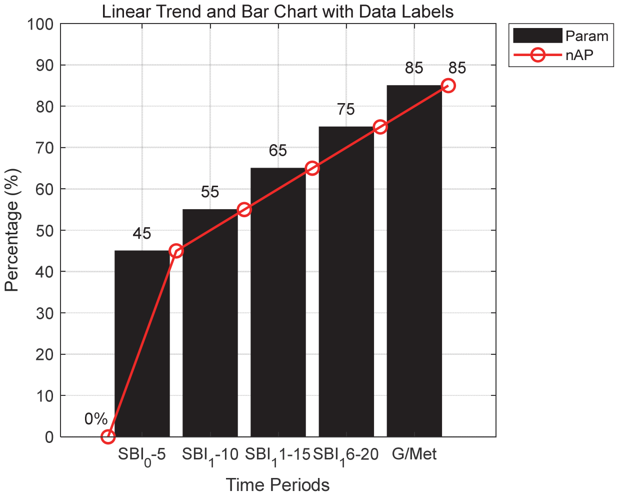


Figure 10 Comparison of network parameters and accuracy

The attention mechanism of RGB-IR dual optical fusion is added in the fusion process to extract and learn human features of different scales at different resolutions, which can improve the single-scale problem caused by the constant use of RGB-IR dual optical fusion in the fusion process, and improve the extraction effect of human features of different scales, as shown in Fig. 11.

Based on the idea of self-attention mechanism, a new human body detection framework TMF-Net based on Transformer is proposed. Through the design of self-attention guided modal interaction module as the core module. At the same time, the isolated two-branch backbone structure is broken to build a cross-mode interaction backbone combined with CNN and Transformer, which strengthens the interaction and guidance between modes at different depth layers by cascading multiple core modules. A series of ablation experiments prove that the proposed core module and improved method are effective for improving the accuracy of modal fusion detection model, and are also suitable for POLO series algorithms.

Figure 11 (a) shows the original image and (b) shows the cropped image

6 CONCLUSION

In this paper, a human body detection algorithm based on RGB-IR dual optical fusion is proposed, which mainly focuses on the feature extraction backbone, optimizes the interaction process between modes, and enhances the ability of modal information exchange. Considering the important influence of feature extraction capability of the feature extraction backbone on downstream tasks, a two-branch cross-mode interaction backbone combined with CNN and Transformer is constructed. In feature extraction, the correlation between visible and infrared features is accurately captured, and these correlation information is effectively supplemented into another mode. The predicted contour information is used to weight the feature map, and the contour of the two-band fusion feature is enhanced. Finally, the multi-scale fusion features are weighted with location information and channel information. The experimental results show that the model significantly improves the detection accuracy of human body in the car cabin under the road scene, and shows good robustness. The main research content of this paper is limited to complex scene problems, but in real life in addition to the above complex scene problems, there are also problems such as occlusion, light change, fog, rain, etc., which also need to be solved. Therefore, it is the future development direction to develop a human body detection and motion analysis algorithm that can deal with a variety of complex scenes.

7 REFERENCES

[1] Bae, S., Shin,H., & Kim, H. (2023). Deep Learning-Based Human Detection Using RGB and IR Images from Drones. *International Journal of Aeronautical and Space Sciences*, *25*(1), 164-175. https://doi.org/10.1007/s42405-023-00632-1

[2] Khanfar, N. O. (2024). Leveraging Multimodal Large Language Models (MLLMs) for Enhanced Object Detection and Scene Understanding in Thermal Images for Autonomous Driving Systems. *Automation*, *5*, 29-45. https://doi.org/10.3390/automation5040029

[3] Tran, H. T., Pham, T. H., & Mun, Y. S. (2024). Drone Detection Using Dynamic-DBSCAN and Deep Learning in an Indoor Environment. *Journal of Electromagnetic Engineering & Science*, *24*(5), 253-267. https://doi.org/10.26866/jees.2024.5.r.253

[4] Li, N., Feng, G., & Zhao, Y. (2024). A Deep-Learning-Based Algorithm for Landslide Detection over Wide Areas Using InSAR Images Considering Topographic Features. *Sensors (14248220)*, *24*(14), 583-602. https://doi.org/10.3390/s24144583

[5] Tang, Y. W., Ji, J., & Lin, J. W. (2023). Automatic Detection of Peripheral Retinal Lesions From Ultrawide-Field Fundus Images Using Deep Learning. *Asia-Pacific Journal of Ophthalmology*, *12*(3), 284-292. https://doi.org/10.1097/APO.0000000000000599

[6] Mittal, P., Sharma, A., & Singh, R. (2022). A Simulated Dataset in Aerial Images using Simulink for Object Detection and Recognition. *International Journal of Cognitive Computing in Engineering*, *3*, 144-151. https://doi.org/10.1016/j.ijcce.2022.07.001

[7] Muradás Odriozola, G. & Pauly, K. (2024). Automating Ground Control Point Detection in Drone Imagery: From Computer Vision to Deep Learning. *Remote Sensing*, *16*(5), 794-807. https://doi.org/10.3390/rs16050794

[8] Goswami, P. & Hossain, A. B. A. (2023). Street Object Detection from Synthesized and Processed Semantic Image: A Deep Learning Based Study. *Human-Centric Intelligent Systems*, *3*(4), 43-54. https://doi.org/10.1007/s44230-023-00043-1

[9] Talala, S., Shvimmer, S., & Simhon, R. (2024). Emotion Classification Based on Pulsatile Images Extracted from Short Facial Videos via Deep Learning. *Sensors (14248220)*, *24*(8), 620-643. https://doi.org/10.3390/s24082620

[10] Rida, M., Khovalyg, D., & Abdelfattah, M. A. (2023). Toward contactless human thermal monitoring: A framework for Machine Learning-based human thermo-physiology modeling augmented with computer vision. *Building and environment*, *245*(Nov.), 1.1-1.15. https://doi.org/10.1016/j.buildenv.2023.110850

[11] Goto, T. & Ishigami, G. (2021). CNN-Based Terrain Classification with Moisture Content Using RGB-IR Images. *Journal of Robotics & Mechatronics*, 33(6), 1294-1306. https://doi.org/10.20965/jrm.2021.p1294

[12] Dai, X., Yuan, X., & Wei, X. (2022). Data augmentation for thermal infrared object detection with cascade pyramid generative adversarial network. *Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies*, *52*(1), 967-981. https://doi.org/10.1109/ICCV.2022.194

[13] Proietti, C., Beni, E. D., & Cantarero, M. (2023). Rapid provision of maps and volcanological parameters: quantification of the 2021 Etna volcano lava flows through the integration of multiple remote sensing techniques. *Bulletin of Volcanology*, *85*(10), 1673-1687. https://doi.org/10.1007/s00445-023-01673-w

[14] Fan, C., Liao, L., & He, W. (2023). Real-time machine learning-based recognition of human thermal comfort-related activities using inertial measurement unit data. *Energy and buildings*, *2023*(Sep.), 294-312. https://doi.org/10.1016/j.enbuild.2023.113216

[15] Yang. B., Li. X., & Liu. Y. (2022). Comparison of models for predicting winter individual thermal comfort based on machine learning algorithms. *Building and Environment*, *215*, 108970-108987. https://doi.org/10.1016/j.buildenv.2022.108970

[16] El Mahdi, B. M. (2024). A Novel Multispectral Vessel Recognition Based on RGB-to-Thermal Image Translation. *Unmanned Systems*, *12*(5), 110-132. https://doi.org/10.1142/S2301385024500110

[17] Chen, B., Chen, L., & Khalid, U. (2024). IFSrNet: Multi-Scale IFS Feature-Guided Registration Network Using Multispectral Image-to-Image Translation. *Electronics (2079-9292)*, *13*(12), 2240-2256. https://doi.org/10.3390/electronics13122240

[18] Dippold, E. J. & Tsai, F. (2024). Enhancing Building Point Cloud Reconstruction from RGB UAV Data with Machine-Learning-Based Image Translation. *Sensors (14248220)*, *24*(7), 72358-72376. https://doi.org/10.3390/s24072358

[19] He, M., Wu, Q., & Ngan,K. N. (2023). Misaligned RGB-Infrared Object Detection via Adaptive Dual-Discrepancy Calibration. *Remote Sensing*, *15*(19), 887-895. https://doi.org/10.3390/rs15194887

[20] Tang, Y., He, H., & Wang, Y. Z. (2023). Multi-modality 3D object detection in autonomous driving: A review. *Neurocomputing*, *553*(Oct.7), 1.1-1.18. https://doi.org/10.1016/j.neucom.2023.126587

[21] Zhao, Z., Yang, Q., & Yang, S. (2021). Depth Guided Cross-modal Residual Adaptive Network for RGB-D Salient Object Detection. *Journal of Physics: Conference Series*, *1873*(1), 12024-12031. https://doi.org/10.1088/1742-6596/1873/1/012024

[22] Zhao, F., Lou, W., & Feng, H. (2024). MFMG-Net: Multispectral Feature Mutual Guidance Network for Visible-Infrared Object Detection. *Drones (2504-446X)*, *2024*, 8(3), 112-128. https://doi.org/10.3390/drones8030112

[23] Zheng, N. W. (2022). CNN Deep Learning with Wavelet Image Fusion of CCD RGB-IR and Depth-Grayscale Sensor Data for Hand Gesture Intention Recognition. *Sensors*, *22*, 803-824. https://doi.org/10.3390/s22030803

[24] Rajoria, M., Purohit, A., & Hota, A. (2022). RAD@home Inter-University Collaboratory for citizen science in galaxy evolution with multi wavelength RGB images. *Proceedings of the International Astronomical Union*, *17*(S375), 686-698. https://doi.org/10.1017/S1743921323000686

[25] Gao, G., Shao, H., & Wu, F. (2022). Leaning compact and representative features for cross-modality person re-identification. *World Wide Web*, *25*(4), 1649-1666. https://doi.org/10.1007/s11280-022-01014-5

[26] Chung, Y. & Lee, H. (2023). Joint triplet loss with semi-hard constraint for data augmentation and disease prediction using gene expression data. *Scientific Reports*, *13*(1), 467-475. https://doi.org/10.1038/s41598-023-45467-8

[27] Zhou, C., Chen, H., & Hu, J. Z. (2024). Quintuple-based Representation Learning for Bipartite Heterogeneous Networks. *ACM transactions on intelligent systems and technology*, *15*(3), 61.1-61.19. https://doi.org/10.1186/1471-2407-11-13

[28] Alawadhi, A., Eliopoulos, C., & Bezombes, F. (2024). Temporal Monitoring of Simulated Burials in an Arid Environment Using RGB/Multispectral Sensor Unmanned Aerial Vehicles. *Drones (2504-446X)*, *8*(9), 444-456. https://doi.org/10.3390/drones8090444

[29] Amelin, V. G., Emel'Yanov, O. E., & Shogah, Z. A. C. (2024). Detection and Identification of Starch and Flour Adulteration by Digital Colorimetry and Fourier-Transform Near-IR Spectroscopy. *Journal of Analytical Chemistry*, *79*(11), 1515-1523. https://doi.org/10.1134/S1061934824700916

[30] Kim, D. H. & Ruy, W. S. (2022). CNN-based fire detection method on autonomous ships using composite channels composed of RGB and IR data. *International Journal of Naval Architecture and Ocean Engineering*, *14*, 489-497. https://doi.org/10.1016/j.ijnaoe.2022.100489

[31] Ligocki, A., Jelinek, A., & Zalud, L. (2021). Fully Automated DCNN-Based Thermal Images Annotation Using Neural Network Pretrained on RGB Data. *Sensors*, *21*(4), 1552. https://doi.org/10.3390/s21041552

Contact information:

Siyu CHEN
School of Advanced Manufacturing, Nanchang University,
Nanchang 330031, China

Juhua HUANG
School of Advanced Manufacturing, Nanchang University,
Nanchang 330031, China

Yinyin LIU
(Corresponding author)
School of Economics and Management, Nanchang University,
Nanchang, 330031, China
E-mail: yinyin15618428680@126.com

Fengping XU
Yicheng Automotive Technology (Shanghai) Co.,
Shanghai, 201800, China