MCIR-YOLO: White Medication Pill Classification Using Multi-Band Infrared Images

Mohan Wang, Yang Jiang, Baohui Xu, Mengqiang Huang, Xu Xue, Xu Wu, Wenjian Kuang, Xiang Liu, and Harm Tolner

***Abstract*—The identification and categorization of pills constitute critical tasks within a contemporary hospital, particularly for avoiding medication errors. Conventional approaches to visual recognition and classification predominantly rely on visible light imagery, proving inadequate for discerning white pills with similar visual characteristics. However, white pills exhibit distinctive infrared properties across various spectral bands. Building upon these observations, this paper introduces the MCIR-YOLO algorithm, a multi-band infrared image object detection system, which enhances the YOLOv5s model through multimodal fusion techniques. This study presents a novel dataset comprising IR images of white round pills captured across six channels, with peak wavelengths ranging from approximately 1400 nm to 1650 nm. Furthermore, a multimodal fusion strategy is proposed, facilitating multi-level feature integration across the six IR channels. This fusion technique exploits the scale features inherent to each IR modality, thereby enabling comprehensive information fusion across multiple modalities. Additionally, the model incorporates an auxiliary detection branch, independent of the backbone, which utilizes fused feature information to calculate a distinct loss, effectively mitigating overall loss. Attention mechanism modules are integrated after two distinct fusion points to enhance feature precision. Leveraging mean and scaling of IR features, these attention mechanisms significantly boost detection accuracy. Experimental results demonstrate that the improved model outperforms the baseline YOLOv5s model, particularly evident in a self-constructed dataset of white round pill IR images, where mAP0.5 increased by 5.47% and 7.96% for single-channel (peak at 1650 nm) and six-channel configurations, respectively. Notably, the utilization of the MCIR-YOLO model for six-channel recognition yields a substantial advantage of 12.05% over the best-performing single-channel IR image recognition.**

***Index Terms*—Near infrared images, object detection, multi-band IR, pill categorization, feature fusion.**

Manuscript received 21 May 2024; revised 30 June 2024; accepted 8 July 2024. Date of publication 11 July 2024; date of current version 19 July 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 61905116 and in part by the Postgraduate Research & Practice Innovation Program of Jiangsu Province under Grant KYCX24_1498 and Grant SJCX24_0472. *(Corresponding author: Wenjian Kuang.)*

Mohan Wang, Yang Jiang, Baohui Xu, Mengqiang Huang, Xu Xue, Xu Wu, and Xiang Liu are with the School of Electronic & Information Engineering, Nanjing University of Information Science & Technology, Nanjing 210044, China (e-mail: wmh0829@126.com).

Wenjian Kuang is with the School of Electronic & Information Engineering, Nanjing University of Information Science & Technology, Nanjing 210044, China, and also with the Suzhou Texpe technology Ltd., Suzhou 210059, China (e-mail: kuang@nuist.edu.cn).

Harm Tolner is with the Tolner Technology, 5615 EG Eindhoven, The Netherlands (e-mail: harm.tolner@gmail.com).

Digital Object Identifier 10.1109/JPHOT.2024.3426929

I. Introduction

With the continuous advancement of deep learning-based object detection techniques in the field of computer vision, object detection has found widespread application and profound impact in various domains, including medical-assisted diagnosis [1], remote sensing target recognition [2], autonomous driving [3], facial recognition [4], military target recognition [5], etc.

Correctly identifying prescription medication stands is a critically important issue in medication reconciliation [6], while the recognition and classification of white round pills containing different ingredients presents a significant challenge. Traditional computer vision methods for pill identification primarily rely on benchmark algorithms such as the R-CNN series [7], [8], [9] and YOLO series [10], [11], [12], [13]. These algorithmic models are typically constructed based on large-scale pill datasets and utilize two-stage detectors for efficient object detection [14]. However, the medical pills are generally manufactured by active pharmaceutical ingredients (APIs) mixed with powdered excipients. Due to the minimal variations in surface characteristics among the white round pills with similar excipients, mostly lactose or starch, there is a significant margin of error in the detection results. In this case, traditional object recognition approaches have encountered a bottleneck in prescription pill identification.

Currently, the majority of object recognition algorithms primarily rely on datasets composed of visible light data [15], leading to suboptimal performance in identifying objects that share similar colors, shapes, and sizes. As IR spectroscopy is well-known for the identification of active substances and adulterants [16], IR images would be an effective approach to identify pills of the same color with different ingredients. Additionally, IR images that are generated through the sensing and processing of IR radiation, possess advantages in coating penetration that are not present in visible light [17], [18]. Despite progress in utilizing IR images for object recognition, single-wavelength near-infrared image data alone is insufficient to provide all the necessary feature information required for pill identification.

To address the aforementioned issue, this paper proposes an improved algorithm, Multi Channels for IR images YOLO (MCIR-YOLO), designed for multi-band near-IR datasets, based on enhancements to YOLOv5s [19], [20]. YOLOv5s has exhibited exceptional object detection performance, coupled with the advantage of being a lightweight model, making it suitable for deployment in embedded platforms and therefore

serves as the baseline model for our study. Considering the limited distinguishing features of white pills in visible light images, we collected near-IR images of white pills across six bands (1400-1650 nm) by using a short-wave IR camera, to constitute the dataset. Prior to training, the dataset underwent preprocessing techniques such as contrast enhancement to augment the IR features of pill images, thereby facilitating the acquisition of more informative data. We improved upon conventional multimodal fusion methods [21], [22], [23] and devised a multi-level feature fusion module tailored specifically for six channels. Furthermore, drawing inspiration from the auxiliary detection branch in YOLOv7 [24], [25], we conducted re-fusion of feature information with varying levels of importance across three different positions, enabling the calculation of a novel branch loss to optimize the final loss output. Finally, we applied a dual attention module at the location where the detection head features are transmitted, focusing on the most critical information of the pills to be detected while attenuating the influence of irrelevant background information. Experimental results demonstrate that the incorporation of these methods enhances the classification of white round pills, thereby improving detection precision and yielding promising outcomes. This approach holds potential as a promising avenue for future research in the field of white pill recognition.

The main contributions of this study are summarized as follows:

1) This paper proposes an IR six-band image object detection method for classifying white round pills. We construct a dataset of IR images of pills with six IR bands (range from 1400 nm to 1650 nm, with 50 nm interval). Multiband IR pill images provide more feature information, significantly improving the accuracy of identifying white pills.
2) In response to the features of IR images of white circular pills, we propose the Multi Channels for IR images (MCIR) Fusion model to perform multi-level feature fusion by gradually increasing fusion levels at four different positions. This approach effectively preserves the characteristics of the IR images and significantly improves detection accuracy.
3) To further optimize detection efficiency, an IR-Image Auxiliary Detection Branch (IADB) was introduced into the model. This branch independently calculates losses to optimize the overall model loss output. Additionally, a dual IR Attention (IRA) mechanism was designed at the detection head position. IRA scales and optimizes features based on the required information for IR image detection. These two improvements effectively enhanced the accuracy of the model's detection.

The MCIR-YOLO model, designed specifically based on the characteristics of white pills, incorporates three innovative model optimizations, effectively enhancing target detection performance for this type of pill. Compared to the currently popular YOLO series object detection models, our model exhibits a notable leading advantage, particularly in terms of recognition accuracy.

II. Related Work

This section commences with an overview of the current research status regarding traditional pill classification, followed by a discussion on the pertinent applications of IR datasets in object detection. Subsequently, various representative optimization models or preprocessing approaches are examined. Finally, an analysis is conducted to compare the advantages and disadvantages of three fusion schemes in multimodal fusion: pixel fusion, decision fusion, and feature fusion. Additionally, a distinction is made between single-level feature fusion and multi-level feature fusion within the context of feature fusion.

A. Traditional Pill Classification Research

The traditional method for pill classification detection primarily relies on distinguishing pills based on their color, shape, and imprinted features, often coupled with CNN models for classification and recognition [26]. Due to substantial objective differences in such characteristics, certain distinctions can also be made by visual observation. In real-world scenarios of pill classification, the quantity of pills to be classified is typically substantial. Addressing the inevitable issue of pills adhering together, Kwon et al. [27] proposed an improved model integrating the Mask R-CNN algorithm, which effectively segments multiple adhered pills and enhances the efficiency of classification and recognition. Thi et al. [28] proposed a pill defect recognition model using Gaussian filtering and smoothing techniques as preprocessing methods, coupled with the YOLOv3 model. Their study demonstrated that the YOLO model, when appropriately preprocessed, could effectively identify pill datasets.

Due to the uniform color and shape, traditional CNN models and visual observation struggle to effectively identify white round pills. In light of this, we propose the MCIR-YOLO model, which is tailored for more accurate and efficient classification and recognition of various types of the pills.

B. Infrared Datasets for Object Detection

The advantage of IR images lies in their ability to provide richer information. In the field of object recognition and detection, integrating them with multimodal fusion schemes can yield improved detection performance, especially in challenging scenarios where object recognition is difficult. IR datasets are widely utilized in the domain of object detection. Popular public datasets include the FLIR dataset [29], the KAIST Multimodal dataset [30], the NIR dataset [31], etc.

Despite the aforementioned prominent advantages, IR datasets exhibit certain limitations and issues in object detection due to their inherent characteristics. To address the commonly encountered issue of low resolution in IR datasets, Zhou et al. [32] proposed an IR image enhancement algorithm for detecting low-resolution IR images. This approach optimizes the focus on detected targets and enhances the efficiency of feature fusion. In response to the limitations posed by conducting single-modal target detection independently on IR datasets, Yu et al. [33] proposed a multi-band IR image synchronous fusion method. This approach achieves synchronous fusion and noise reduction

optimization at the detail level, effectively mitigating the impact of noise on the fusion results. To solve the issue of poor target recognition caused by insufficient IR image contrast in IR datasets, Zha et al. [34] proposed a model that combines the YOLO model with Canny edge detection and Gabor filtering. This model is supplemented by preprocessing techniques such as local contrast multi-scale enhancement, non-local methods, and contrast-limited adaptive histogram equalization. These approaches effectively enhance the accuracy and robustness of detection.

Drawing on the prominent advantages and objective limitations of IR datasets in the field of object detection, this paper proposes a feature fusion algorithm for IR multi-band data (six bands). Through preprocessing methods such as contrast enhancement, and further incorporating tailored loss optimization and attention mechanisms designed for IR datasets into the model, it achieves improved performance in pill detection.

C. Multimodal Fusion

Currently, multimodal fusion algorithms have been widely and profoundly applied in various practical scenarios of object detection. For instance, in medical imaging for lesion detection and diagnosis [35], small target identification in remote sensing [36], road condition recognition for autonomous driving vehicles [37], and pedestrian detection in street scenes [38], etc.

Research in multimodal fusion, particularly in the domain of multispectral information fusion, primarily focuses on fusion methodologies. These fusion schemes can be systematically categorized into three main types: pixel-level fusion[39], feature-level fusion [40], and decision-level fusion [41]. Pixel-level fusion emphasizes the direct fusion of data from different sensors or modalities at the pixel level. After extracting feature information from multispectral images, the fusion typically occurs at the head of the model's image information input. Pixel values from different sources are merged using pixel-level weighting or fusion rules to reconstruct the fused image, which is then used as input for the model [42]; Decision-level fusion integrates data from different modalities or sensors at the decision level, which involves combining the decision results from different modalities or sensors to derive the final outcome [43]; Feature-level fusion refers to the fusion of data at the level of feature representation. Furthermore, feature fusion can be further subdivided into single-level fusion and multi-level fusion. Single-level fusion primarily involves the fusion of data from different modalities at a single position during the overall fusion stage [44]. Multi-level fusion involves the gradual fusion of multimodal data information at multiple levels or stages. Overall, each fusion step builds upon the previous one by integrating data at a higher level of abstraction [45].

Through the comparison of various fusion methods, multi-level feature-level fusion aligns more closely with the requirement to retain additional information features from multispectral IR images of pills. Therefore, this paper proposes the MCIR-Fusion method based on multi-level feature-level fusion. By incrementally fusing features over four fusion levels, MCIR-Fusion attains a greater degree of detail in capturing feature information from different spectral bands of IR images, thereby enhancing the model's detection accuracy.

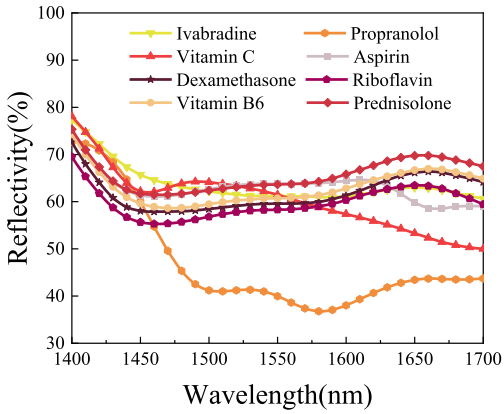

Fig. 1. The reflectivity of eight types of pills in the 1400-1700 nm infrared wavelength range.

III. Proposed Method

To verify whether there are indeed differences in the IR images of eight different pills at various IR wavelengths used in the experiment, we tested the reflectance curves of these eight pills under IR light illumination. The penetration capabilities of IR mean that the reflected light can contain deeper information about the ingredients from inside pills than that of the visible light. In our approach, we have selected the 1400-1700 nm range, as the reflectance differentiation between 800 and 1400 nm are relatively smaller. As illustrated in Fig. 1, the reflectance of different types of pills varies significantly across different wavelength bands. This variation in reflectance across bands indicates that the characteristic information of tablets presented in infrared images differs depending on the wavelength. This finding further substantiates the feasibility of using infrared images from different wavelength bands to create a multispectral dataset, which effectively meets the need for distinguishing the characteristics of white pills.

Referring to the model architecture depicted in Fig. 2, we propose the following new contributions based on the original YOLO algorithm framework. Firstly, based on the varying reflection effects of the target pharmaceuticals under different spectral bands, we assembled a dataset of six-channel IR images. Leveraging the advantages of multimodal fusion in target detection, we evaluated the strengths and weaknesses of various fusion schemes and selected the multi-level feature fusion approach with the most optimal fusion effect to merge the multi-channel IR images. Furthermore, we introduce an image enhancement detection branch based on the detection characteristics of the IR dataset. This detection branch extracts three sets of features with different scales from three different locations of the detection backbone. After fusion and computation, independent losses are obtained for each branch. After fusion and computation, independent losses are obtained for each branch. These independent losses, along with the backbone loss, are then proportionally combined to optimize the model loss output.

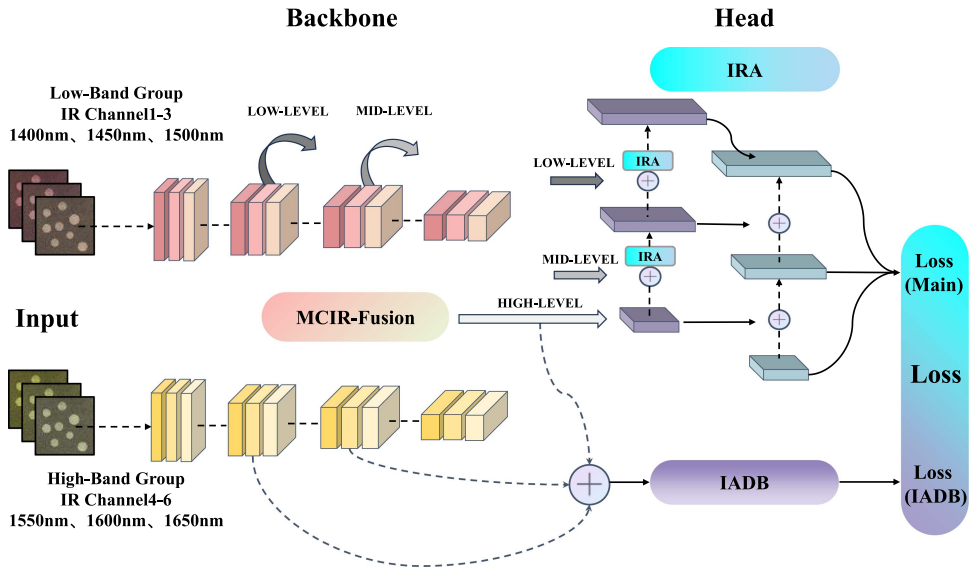

Fig. 2. Overall structure of MCIR-YOLO object detection network.

Finally, we incorporate two attention mechanisms at the head position of the original architecture. These attention mechanisms enhance the detection efficiency and accuracy by amplifying relevant IR feature information through mean calculation and tensor scaling.

A. *Multimodal Fusion of Multi-band Infrared Images*

The detection of white round pills has long been a challenging task in the field of pill detection due to their highly similar features such as color, shape, and size. It is widely recognized that the more information the detection model collects about the target objects, the higher the discriminability and detection accuracy will be. Therefore, providing the detection model with more information about the pills is crucial for addressing this detection challenge. According to the IR spectral reflection detection results of the tested pills under different spectral bands of IR light as shown in Fig. 1, it can be observed that there are significant differences in the pills' absorption capabilities of IR light within the wavelength range of 1400 nm to 1650 nm. The IR imaging information collected from the pills under different spectral bands of IR light sources varies considerably. Multimodal fusion aims to gather more information by integrating image data from different modalities. Traditional multimodal fusion often involves combining visible light images with IR images or visible light images with thermal imaging, among others. Our proposed MCIR-Fusion module integrates IR images from different IR bands, offering greater diversity in imaging information across the scale of IR images.

Currently, three widely used fusion mechanisms are pixel-level fusion, feature-level fusion, and decision-level fusion. Each of these fusion schemes has its advantages and disadvantages. In brief, pixel-level fusion mainly focuses on pixel-level operations, this fusion scheme primarily emphasizes preserving the original image features and may not meet the requirement for extracting deeper IR features in pill detection. On the other hand, decision-level fusion has lower computational efficiency, poorer generalization, and feature expression capabilities, which do not align well with the requirements. Based on the above considerations, the model opts for feature-level fusion for multimodal fusion. This choice aims to maximize computational efficiency and achieve optimal recognition results while ensuring the most comprehensive retention of IR image features. Taking into account the differences in information provided by each IR band image, our MCIR-Fusion module employs a feature concatenation fusion method to integrate the information from all bands, enabling the model to comprehensively learn from the characteristics of each band. As illustrated in Fig. 3(a), the six IR images are initially grouped based on the characteristic differences in imaging at different IR bands. These are categorized into a set of relatively proximate near-IR bands, including 1400 nm, 1450 nm, and 1500 nm, and a set of relatively distant near-IR bands, comprising 1550 nm, 1600 nm, and 1650 nm. Then, a multi-level feature fusion operation is conducted on the two bands of the spectrum. Following the input of the images, they undergo the first round of feature fusion after passing through two CBS modules. After passing through the MCIR-Fusion module, the IR image information from all bands is concatenated. Following this, the fused information undergoes a segmentation process before being returned to the original two branches.

Subsequently, after passing through each CBS module, the MCIR-Fusion module fusion operation is performed once. This process is repeated three times, and at the same time, feature information is output to the neck after each iteration.

The detailed structure of the MCIR-Fusion module is illustrated in Fig. 3(b). After the input of the short and long wavelength-band groups of IR images, they are divided into main branches and fusion branches. The main branch undergoes one CBS convolution, reducing the number of channels to half of the initial input, which is then directly fed into the later

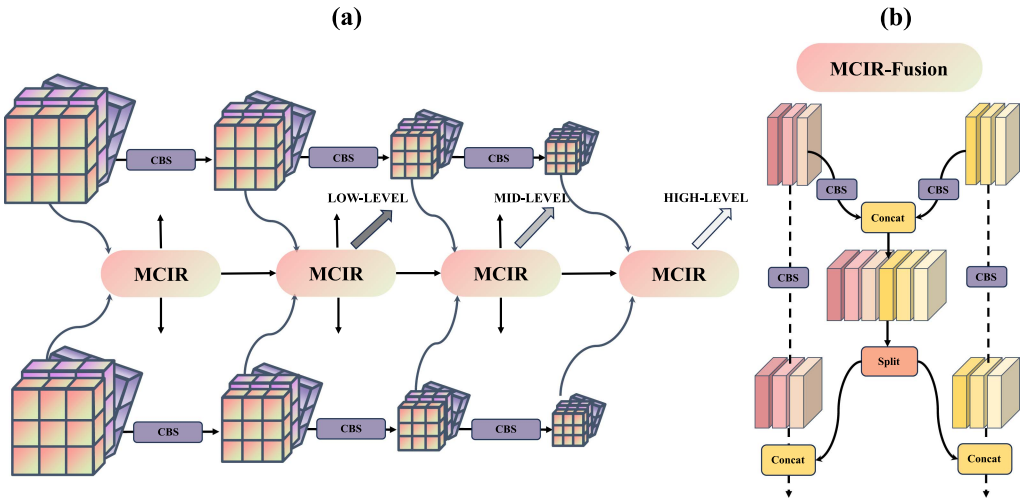

Fig. 3. (a) The structure of the MCIR-YOLO six-channel multi-level fusion. (b) Feature fusion of the MCIR-Fusion.

part awaiting fusion. Meanwhile, the fusion branches of the low-band and high-band groups undergo a CBS convolution simultaneously, reducing the number of channels by half for both groups. Subsequently, a feature fusion operation is performed for the low-band and high-band groups.

To ensure that the subsequent feature maps concatenated in the main branch preserve both the original features and the features obtained from the fusion of the two band groups, we subject the newly fused image obtained from the fusion branch to segmentation processing once again. After undergoing the effect of the separation module consisting of consecutive residual units, the fused image is again split into two parts. The formula for separating the feature information fused by the MCIR Fusion module using the separation module is as follows:

$$S_L, S_H = r^n(M_L^{in}, M_H^{in}) \tag{1}$$

which are then separately fed into the main branches of the low-band group and high-band group. Subsequently, they are concatenated with the feature maps in the main branch that retain the original information. The concatenated new feature information undergoes a final step of transmission through the last CBS module, completing the final stage of the process. With this, the MCIR-Fusion module accomplishes a complete fusion operation. The expression of the features output after fusion is as follows:

$$M_L^{out} = (M_L^{in}, S_L), M_H^{out} = (M_H^{in}, S_H) \tag{2}$$

B. Infrared-Image Auxiliary Detection Branch

In the realm of deep learning for object detection in IR image datasets, the loss coefficients of the model outputs largely determine the efficacy of model learning and the quality of detection accuracy. To ultimately optimize the learning effectiveness of the model, we drew inspiration from the auxiliary detection branch module of YOLOv7 and introduced the IR-Image Auxiliary Detection Branch (IADB) for IR image enhancement. This branch performs a new fusion and calculation of the independently branched feature information outputs, derived from three different positions of the backbone, each with varying levels of feature characteristics. Subsequently, the independent branch loss is computed, combined proportionally with the backbone loss, and optimized to enhance the overall model loss, thereby improving detection accuracy.

Starting from the second MCIR-Fusion module in the input layer, for each fusion operation, feature information is outputted to the neck layer. Specifically, three sets of feature information are sequentially outputted after three fusion operations: High-Level, Mid-level, and Low-Level. As each feature fusion operation occurs, the number of channels is halved, leading to a corresponding reduction in retained feature information. Consequently, as the crucial contrast and resolution information in the IR image features diminish, the accuracy of detection naturally becomes harder to improve. To address this issue, the introduction of the detection branch (IADB) aims to establish an independent loss optimization scheme that runs parallel to the main detection head. This branch is designed to effectively enhance detection accuracy. From Fig. 3(a), it can be observed that after the second feature fusion, we extract IR image feature information once as the low-level feature input for IADB. Secondly, following the third fusion, it is further extracted as the mid-level feature input. Finally, after the last feature fusion (fourth time), it is extracted once more as the high-level feature input.

The loss calculation of the IADB channels is illustrated in Fig. 4, where the first cascade is performed between the inputs at the high-level and mid-level (Fusion Output 1), followed by the second cascade between the input at the low level (Fusion Output 2). Fusion Output 1 and Fusion Output 2 are further merged to produce another fusion output (Fusion Output 3), and subsequently, Fusion Output 3 is further merged with the input at the high position to generate another fusion output (Fusion Output 4). The outputs of the conventional three-stage loss computation are referred to as Fusion Output 2 (low-level output), Fusion Output 3 (mid-level output), and Fusion Output 4 (high-level output). The loss of the backbone is finally denoted as loss(loc), loss(obj), and loss(cls) respectively, the three also

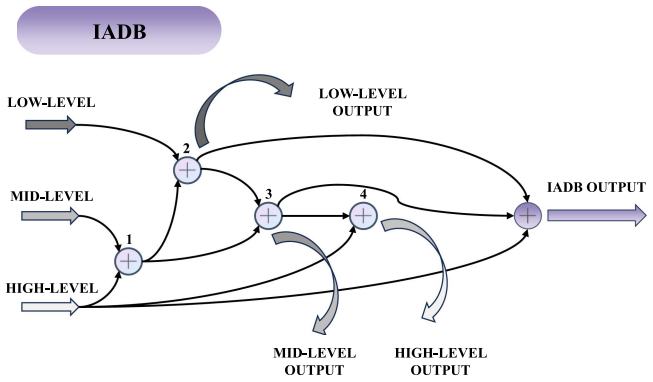

Fig. 4. The branch structure of IADB.

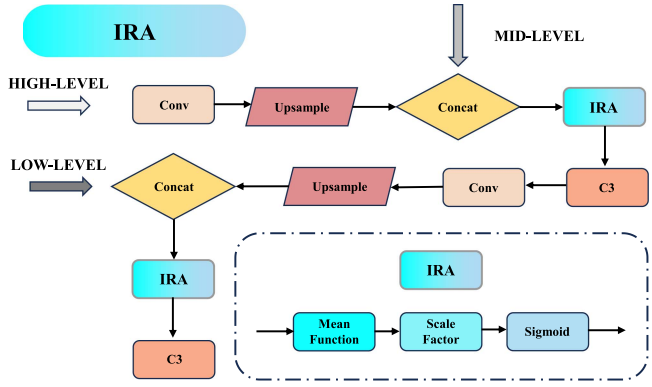

Fig. 5. The structure of dual IRA attention mechanism at the head position.

referred to as loss (Main) [13], [46], is formulated as follows:

$$Loss_{Main} = \lambda_{loc} \sum_{i=0}^{2} a_i Loss_{loc} + \lambda_{obj} \sum_{i=0}^{2} b_i Loss_{obj} + \lambda_{cls} \sum_{i=0}^{2} c_i Loss_{cls} \quad (3)$$

The loss output of the IADB branch is obtained by further integrating and calculating the losses of output 2, output 3, and the output at the high position through a tertiary fusion process based on this. This output is denoted as loss (IADB), which also contains three parts: loss(loc2), loss(obj2), and loss(cls2), and its expression is formulated as follows:

$$Loss_{IADB} = \lambda_{loc2} \sum_{i=0}^{2} d_i Loss_{loc2} + \lambda_{obj2} \sum_{i=0}^{2} e_i Loss_{obj2} + \lambda_{cls2} \sum_{i=0}^{2} f_i Loss_{cls2} \quad (4)$$

After obtaining the main branch loss and the IADB branch loss, they are combined and calculated according to a certain proportion. λ(Main) and λ(IADB) are balancing coefficients obtained during model training for the backbone and branch. The formula for the overall loss calculation is as follows:

$$Loss_{MCIR} = \lambda_{Main} Loss_{Main} + \lambda_{IADB} Loss_{IADB} \quad (5)$$

C. Dual IRA Attention Mechanism

The IR image dataset and the visible light dataset inherently exhibit significant differences in feature information. To achieve better fusion effects at the concatenation positions of the head layer, we designed two attention mechanism modules at different hierarchical feature concatenation positions and named them IR Attention (IRA). The structure of this attention mechanism is illustrated in Fig. 5. After the concatenated feature information enters the module, it first undergoes a mean calculation, which computes the mean along the spatial dimensions of the tensor. Following this, based on the distinctive characteristics of the IR images, a scaling factor is set to perform similarity scaling. The scaled tensor then enters the final activation module (Sigmoid),

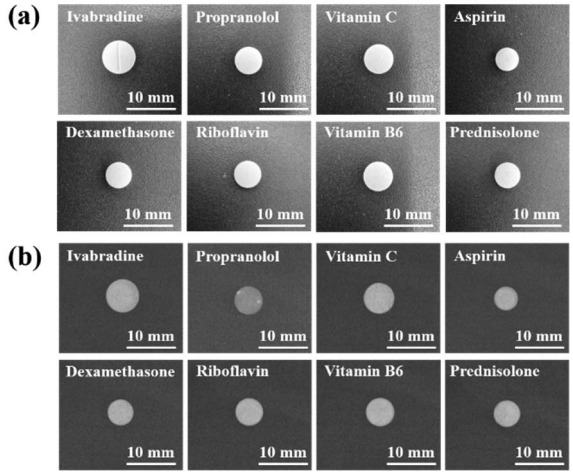

Fig. 6. (a) Visible Light Images of Eight Pills. (b) IR Images of Eight Pills.

where it is multiplied by the attention weights, yielding the final output value. The attention weight e_λ is set to 1×10^{-4}. Regarding the placement of attention mechanisms, the first IRA is positioned after the output of Low-level features following the SPPF module, where it merges with Mid-level features. The second IRA is placed after the first fusion of features, where the information is fused with High-level features. This setup effectively supervises a dual attention mechanism, thereby enhancing both detection efficiency and accuracy.

IV. Experiments and Results

A. Dataset Collection and Infrared Image Annotation

This study selected eight representative white round pills, namely “Ivabradine”, “Propranolol”, “Vitamin C”, “ Aspirin”, “Dexamethasone”, “Riboflavin”, “ Vitamin B6”, and “ Prednisolone”. The images of the pills under visible light and IR light at 1650 nm is depicted in Fig. 6. All eight types of pills share the same color, which is white; they have similar shapes, all being round; and their sizes are comparable, with diameters ranging from 6mm to 12mm.

All the samples of white round medication used in the self-built dataset were provided by Suzhou Science and Technology City Hospital. The IR imaging system utilized the SW1412

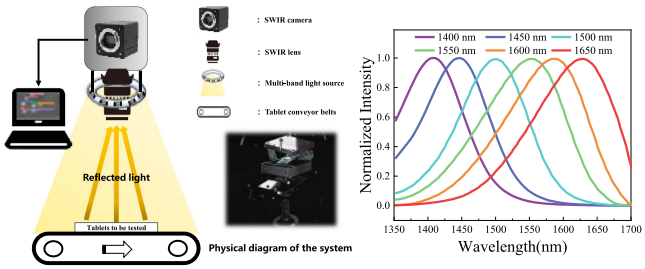

Fig. 7. Physical diagram of the system and the normalized emission spectra of the six wavelength bands.

short-wave IR camera lens from Suzhou Zhisheng Technology Co., Ltd. The lens has a focal length of 12mm, a minimum object distance of 15 cm, and a resolution of 5MP. It operates in the wavelength range of 800 nm to 1700 nm. The IR sensor employed in this study is the 640 short-wave focal plane detectors from IR Sensor Technology (Beijing) Co., Ltd. It supports three sampling specifications: 640×512, 640×480, and 512×512. In this experiment, the sampling specification of the short-wave focal plane detector is set to 640×512. The light source for the IR spectrum consists of an IR LED ring light source. The LED light source provides a total of six bands with peak wavelengths at 1400 nm, 1450 nm, 1500 nm, 1550 nm, 1600 nm, and 1650 nm, with intervals of 50 nm between adjacent bands. Fig. 7 illustrates the entire process of acquiring multispectral infrared images of white pills and presents the normalized spectra for six bands using an infrared LED light source. Each band is equipped with six LED light sources, totaling thirty-six sources. With this IR image collection system, a total of 211 sets of IR images of pills were acquired, with six bands per set, resulting in 1266 images in total.

Fig. 8 displays the IR images of the pills obtained under the IR LED light sources at six different IR bands. The grayscale histograms of the IR images at each band exhibit slight variations, indicating the objective influence of IR light irradiation at different bands on the imaging performance of the IR images.

B. Preprocessing of Infrared Images

Considering the performance limitations of the shortwave IR detector used in collecting the IR image dataset, the original IR images of white round pills are deficient in their ability to present the pill's characteristic information. To address this issue, we employed the method of IR image contrast stretching [47] to enhance the contrast of the IR images, thereby highlighting the characteristic information of the pills in the images. This enhancement makes it easier for algorithmic models to learn and understand the images, ultimately improving the overall accuracy of the algorithms. By comparing the zoomed-in images of the same position of the pills in the two images before and after preprocessing in Fig. 9, it is evident that the IR images of the white round pills, after contrast enhancement preprocessing, have been significantly improved in terms of both local details and feature presentation.

C. Implementation Details

The model framework used in our experiments is based on PyTorch version 2.0.1, with Python version 3.10.10, CUDA version 11.8, using an 11th Gen Intel(R) Core (TM)i9-11900K 3.50 GHz CPU, and an NVIDIA GeForce RTX 3090 GPU. The input image has a pixel size of $640 * 512$. The IR image dataset is formatted according to the standard YOLOv5 dataset format. The dataset consists of six bands, with 211 images per band. Following a tenfold cross-validation scheme, 189 images are used for training, while 22 images are allocated for validation and testing purposes. During training, the model was trained with a Batch Size of 2, an initial learning rate of 0.01, initial momentum of 0.937, for a total of 300 epochs. Additionally, data augmentation techniques such as multi-scale training, translation, horizontal flipping, mosaic, and padding were applied to augment the dataset. However, these augmentation strategies were disabled during the testing phase.

D. Accuracy Metrics

Precision is one of the key metrics used to evaluate the performance of classification models. It measures the proportion of true positive samples (TP) among all samples predicted as positive by the model, compared to the number of false positives (FP). The formula for precision(P) is as follows:

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

Recall, another crucial metric for evaluating classification model performance, measures the model's ability to correctly identify all true positive samples, compared with the number of false positives. The formula for recall (r) is as follows:

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

The mean average precision mAP is a comprehensive metric to compare the effectiveness of the different methods. It calculates the area enclosed by the precision-recall curve and the coordinate axes using integral methods for all categories. The formula for mAP is as follows:

$$mAP = \frac{1}{n} AP = \frac{1}{n} \int_0^1 p(r) dr \tag{8}$$

The accuracy metric mentioned below, mAP0.5, refers to the performance of the object detection model in multi-class detection using an IoU threshold parameter of 0.5; mAP0.5: 0.95 represents the average mAP value obtained over the IoU threshold parameter ranging from 0.5 to 0.95.

E. Ablation Study

The ablation study firstly compares the detection accuracy of single-band and multi-band approaches, as shown in Table I, demonstrating the rationale for selecting multi-band fusion. Additionally, the detection results of six individual bands are compared, revealing that the 1650 nm single band achieves the highest classification accuracy. Consequently, all subsequent tables compare the detection results of the 1650 nm single band

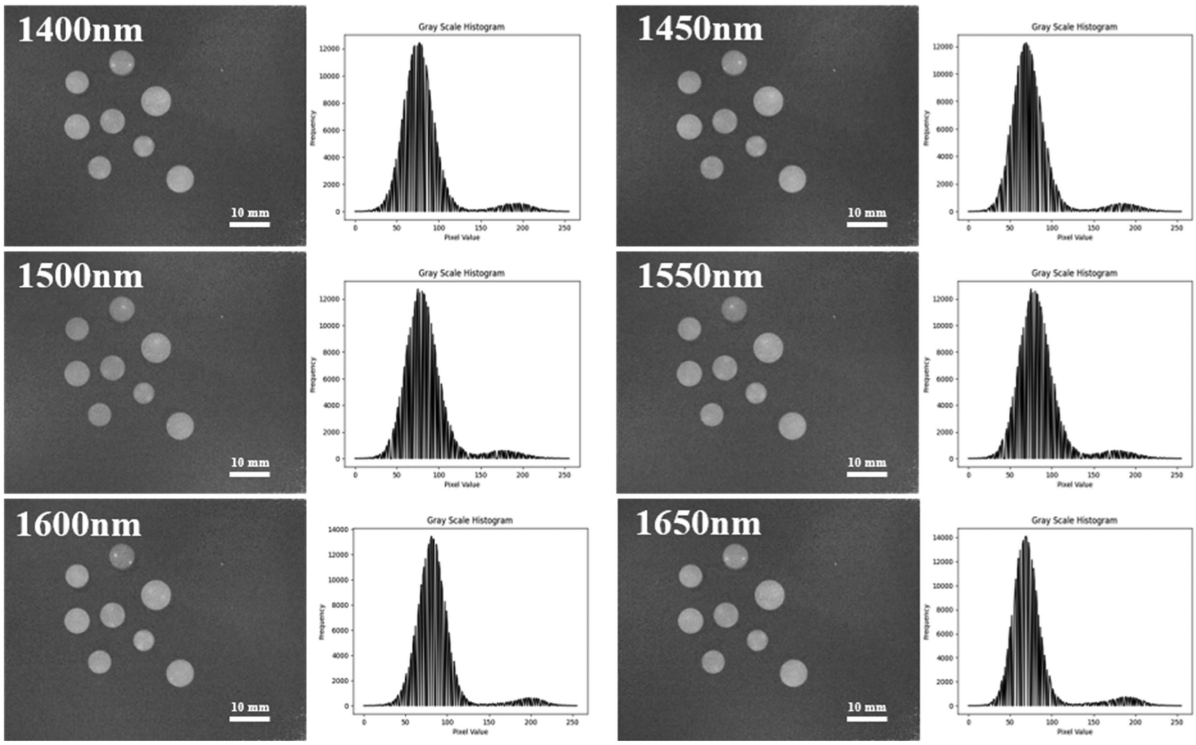

Fig. 8. The IR images of the pills at six different bands along with their corresponding grayscale histograms.

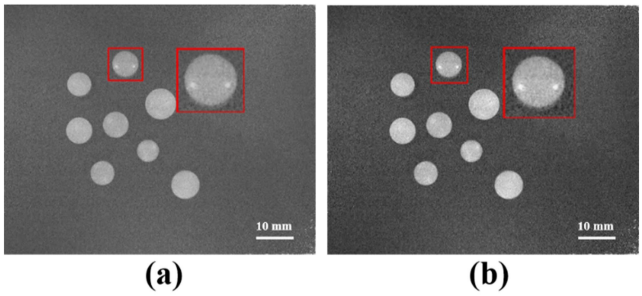

Fig. 9. (a) The original IR images of pills and locally magnified views. (b) The IR images of pills after contrast enhancement and locally magnified views.

with other methods, verifying that the addition of new innovations effectively enhances accuracy. The experiment compared various fusion schemes within the multimodal fusion framework, as shown in Table II, ultimately determining that the MCIR-Fusion scheme yielded the best results. Table III demonstrates that the inclusion of the IADB detection branch positively enhances detection accuracy. Table IV highlights that the IRA attention mechanism effectively improves detection accuracy. To ensure the rigor and accuracy of the experiments, the results in Tables III and IV are based on the previous innovation results and integrated with the new innovative algorithms, thus proving that the accumulation of innovations significantly enhances result accuracy.

1) Comparison of IR Datasets Across Different Spectral Bands: In order to evaluate the distinctions in detection outcomes across different spectral bands of IR datasets and simultaneously compare the variances in detection outcomes between

TABLE I
COMPARISON OF RECOGNITION RESULTS ACROSS DIFFERENT IR CHANNELS

Modality	Precision(%)	Recall(%)	map0.5(%)	map0.5:0.95(%)
1400nm	61.33	79.55	73.81	66.93
1450nm	57.42	84.90	74.65	67.40
1500nm	66.33	81.32	74.79	67.17
1550nm	62.26	83.41	76.29	67.81
1600nm	69.11	82.59	76.99	67.08
1650nm	63.08	85.37	77.25	67.99
3 Channels (1400nm-1500nm)	67.51	**86.66**	78.27	71.22
3 Channels (1550nm-1650nm)	61.99	83.50	79.04	71.99
6 Channels (1400nm-1650nm)	**73.62**	82.78	**81.34**	**73.72**

TABLE II
COMPARISON OF DIFFERENT FUSION METHODS

Method	Modality	Precision(%)	Recall(%)	mAP0.5(%)	mAP0.5:0.95(%)
Concatenation	6channels	73.62	82.78	81.34	73.72
Pixel-Level Fusion	6channels	66.61	86.48	76.67	68.78
Feature-Level Fusion 1	6channels	61.32	**90.42**	78.91	71.82
Feature-Level Fusion 2	6channels	62.83	85.42	79.03	71.54
Feature-Level Fusion 3	6channels	68.59	82.32	76.59	68.79
MCIR-Fusion (Ours)	6channels	**72.07**	85.66	**83.53**	**73.93**

TABLE III
IMPACT OF IADB BRANCH

Method	Modality	Precision(%)	Recall(%)	mAP0.5(%)	mAP0.5:0.95(%)
YOLOv5s	1channel (1650nm)	63.08	85.37	77.25	67.99
YOLOv5s+ IADB	1channel (1650nm)	**68.35**	**91.02**	**79.18**	**72.92**
YOLOv5s	6channels	73.62	82.78	81.34	73.72
YOLOv5s+IADB	6channels	**75.42**	83.64	83.88	76.53
YOLOv5s+IADB +MCIR-Fusion	6channels	69.68	**91.62**	**86.68**	**76.99**

TABLE IV
THE ENHANCEMENT OF IRA ATTENTION MECHANISM

Method	Modality	Precision(%)	Recall(%)	mAP0.5(%)	mAP0.5:0.95(%)
YOLOv5s	1channel (1650nm)	63.08	**85.37**	77.25	67.99
YOLOv5s+ IRA	1channel (1650nm)	**74.16**	74.31	**78.86**	**71.09**
YOLOv5s	6channels	73.62	82.78	81.34	73.72
YOLOv5s+ IRA	6channels	71.70	86.17	82.72	73.32
YOLOv5s+ IRA+MCIR-Fusion	6channels	77.83	89.12	85.85	76.51
YOLOv5s+ IADB+IRA	6channels	71.05	**89.92**	85.60	78.59
YOLOv5s+ IRA+IADB+ MCIR-Fusion	6channels	**78.05**	85.59	**89.30**	**80.31**

single-band single-channel and six-band multi-channel datasets, we employed our benchmark model YOLOv5s as the testing model for comparative experiments. As shown in Table I, it is evident from the experimental results that the most critical metric, mAP0.5, exhibits a continuous enhancement from the 1400 nm band to the 1650 nm band. This observation broadly aligns with the findings obtained from the spectral analysis conducted earlier, indicating the maximal discrepancy in IR light absorption by the medication at the 1650 nm band. Furthermore, the mAP0.5 value obtained from the testing of the three-channel IR image dataset of the short wavelength-band group (1400 nm, 1450 nm, and 1500 nm) is 78.27%, surpassing that of any single band within it. Similarly, the mAP0.5 value for the high band group (1550 nm, 1600 nm, and 1650 nm) is 79.04%, corroborating this outcome. Finally, the experimental evaluation of the model with simultaneous input of all six channels yielded an mAP0.5 value of 81.34%, significantly higher in accuracy compared to all other single-channel and three-channel test results. Therefore, through experimental comparisons, we have confirmed the rationale behind employing multi-channel IR bands for detection. Moreover, among the multiple bands, the performance of utilizing all six bands surpasses that of using only three bands, making it the optimal choice at present.

2) Comparison of Different Fusion Methods: Currently, mainstream fusion methods include pixel-level fusion mainly conducted at the backbone level and feature-level fusion. Decision fusion performed at the head level is also prevalent. In our approach, fusion occurs at the input layer, thus excluding decision fusion. To select the most suitable fusion scheme, comparative experiments were conducted, including pixel-level fusion at the input position and three different single-level feature-level fusion methods at various positions. These experiments were performed using a six-channel IR dataset and compared against our proposed MCIR-Fusion, the multi-level feature fusion approach.

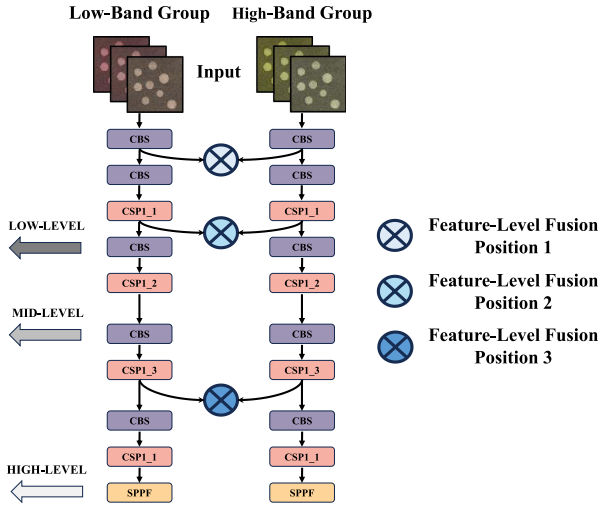

Fig. 10. The ablative experiments of feature fusion methods, showcasing three instances of unipolar feature fusion at different positions.

The comparative results, as shown in Table II, indicate that the MCIR-Fusion method achieved the best fusion performance. The Precision value reached 72.07%, while the mAP0.5 and mAP0.5:0.95 values reached 83.53% and 76.53%, respectively, surpassing other fusion approaches by a significant margin.

In the other comparative schemes, we conducted pixel-level fusion immediately after the first CBS convolution at the position of the six-channel input. The IR features of the six channels were concatenated and then inputted into the subsequent layers, yielding an mAP0.5 value of 76.67%. Regarding feature fusion, as illustrated in Fig. 10, three different positions were selected for monopolar fusion comparison. The first position, chosen identical to the pixel fusion position, achieved the highest recall value of 90.42% among all experiments; The second position, just before the third CBS convolution, was chosen for feature fusion, yielding the best mAP0.5 value among the three monopolar fusions conducted; The final monopolar fusion position was selected just before the Low-Level output, and its fusion effect was the worst among the three instances. In comparison, the overall performance of feature-level fusion is superior to pixel-level fusion, but the differences in results manifested at different fusion positions are objectively present. After comprehensive comparison of the strengths and weaknesses of various schemes, we have chosen MCIR-Fusion, which demonstrates the highest accuracy and is most suitable for drug detection, as the fusion method for our model.

3) Impact of the IADB Branch: To evaluate whether the introduced auxiliary detection branch can effectively enhance medication detection, experiments were conducted by integrating the IADB branch into both single-channel and multi-channel detection frameworks. As shown in Table III, the single-channel baseline selected the 1650 nm band, which exhibited the highest accuracy among the six channels. Upon integration of the

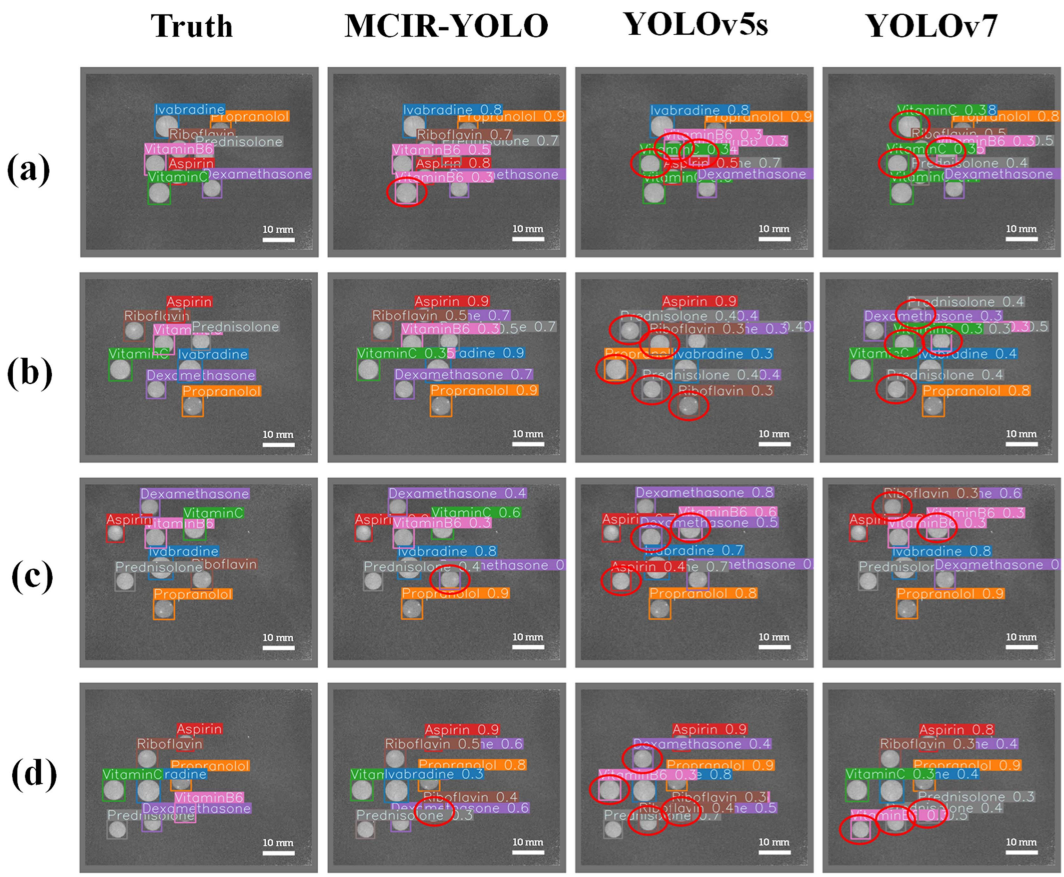

Fig. 11. Comparative analysis of the practical detection results for pill classification using different models.

branch, notable improvements were observed across all four evaluation metrics used to assess detection performance. Specifically, the mAP0.5 increased from 77.25% to 79.18% (+1.93%). In the case of the six-channel setup, employing the most basic cascaded fusion approach yields a result of 81.34%. However, by solely incorporating the IADB branch on top of the cascaded fusion, the test results improved by 2.54%. Finally, if we replace the cascaded fusion with MCIR-Fusion and further incorporate the IADB branch for auxiliary detection, the detection accuracy significantly improves to 86.68%, resulting in an increase of 5.34%. In summary, whether it is a single-channel or six-channel detection scheme, the IADB branch has objectively shown improvements. Additionally, when combined with the MCIR-Fusion method, superior detection results can be achieved.

4) The Enhancement of Model Accuracy By IRA: Similar to Table III, in Table IV, we also compare the detection performance of the attention module in both single-channel and six-channel settings. In the single-channel setting of 1650 nm IR images, there is a slight improvement in detection accuracy when the attention mechanism is added to the detection head position. After fusing the detection across all six channels, employing the YOLOv5s+IRA method with cascade fusion and the addition of the attention mechanism yielded an mAP0.5 result of 82.72%; then, by altering the fusion approach to incorporate MCIR-Fusion followed by the addition of the attention mechanism, we achieved a result of 85.85%, indicating a significant improvement in accuracy; furthermore, retaining the cascaded fusion method but incorporating the IADB detection branch, we achieved an output of 85.60%; Finally, by integrating all the modules proposed in this paper, namely MCIR-Fusion method along with IADB detection branch and the attention mechanism at the head position (YOLOv5s+MCIR-Fusion+IADB+IRA), we obtained the best results among all experiments, with a Precision value of 78.05, Recall value of 85.59%, mAP0.5 value of 89.30%, and mAP0.5:0.95% value of 80.31%.

F. Qualitative Analysis

Fig. 11 illustrates the visual detection results of our MCIR-YOLO model compared to three other models, YOLOv5s, YOLOv7, at four different positions of medication. The red ellipses indicate instances where detection errors or confusion occurred. From the figure, it can be observed that MCIR-YOLO exhibits significantly fewer detection errors or confusion on white round pills compared to the other two models. Therefore, making targeted improvements based on the IR image characteristics of white round pills on the traditional YOLO model holds certain value and significance

The detection results of eight medications using the MCIR-YOLO model proposed in this paper are compared with the baseline model YOLOv5s in both single-channel and six-channel testing environments. As shown in Fig. 12, it is evident that the MCIR-YOLO model achieved significant improvements in detection accuracy.

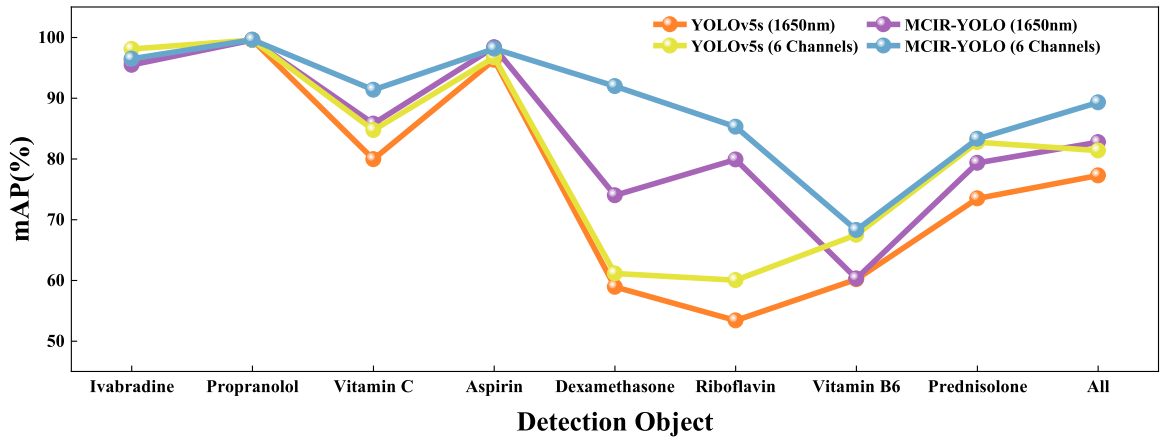

Fig. 12. Comparison of the detection accuracy of the MCIR-YOLO model and the YOLOv5s model using single-channel (1650 nm) and six-channel IR images.

TABLE V
COMPARISON BETWEEN TEN MODELS IN SINGLE-CHANNEL AND SIX-CHANNEL SCENARIOS

Method	Modality	Ivabradine	Propranolol	Vitamin C	Aspirin	Dexamethasone	Riboflavin	Vitamin B6	Prednisolone	Precision(%)	mAP0.5 (%)	mAP0.5:0.95 (%)
YOLOv3	1channel	95.43	99.57	50.15	97.45	46.59	85.16	36.87	85.61	61.71	74.60	67.88
YOLOv4	1channel	92.30	99.56	62.94	96.94	66.24	59.71	43.46	91.49	69.04	76.58	69.70
YOLOv5m	1channel	90.61	99.56	81.34	95.89	81.52	64.52	44.92	72.53	70.16	78.86	71.09
YOLOv5x	1channel	99.55	99.56	51.97	97.11	74.29	**91.89**	54.19	64.89	68.35	79.18	72.95
YOLOv5l	1channel	88.75	99.56	67.57	97.60	65.00	72.16	49.92	76.76	70.10	77.17	70.36
YOLOv5s	1channel	96.18	**99.58**	79.94	96.30	58.92	53.41	60.14	73.50	63.08	77.25	67.99
YOLOv7	1channel	95.48	**99.58**	83.55	96.98	76.47	48.85	40.55	87.83	64.96	80.13	73.07
YOLO-FIRI	1channel	**99.57**	**99.58**	48.15	97.48	**87.01**	83.04	28.25	**89.14**	65.37	79.58	71.89
YOLO-CIR	1channel	96.67	99.41	84.12	97.36	76.59	47.28	49.25	86.56	66.43	79.91	72.21
MCIR-YOLO (Ours)	1channel	95.44	**99.58**	**85.79**	**98.36**	73.97	79.89	**60.32**	79.34	**71.70**	**82.72**	**73.32**
YOLOv3	6channels	90.61	99.56	81.34	95.89	81.52	64.52	44.92	72.53	74.16	78.86	71.09
YOLOv4	6channels	96.91	99.56	63.44	97.27	70.22	75.93	48.04	90.30	69.69	80.46	71.59
YOLOv5m	6channels	95.92	99.56	67.01	96.64	71.45	**90.78**	58.42	70.52	67.61	81.29	73.26
YOLOv5x	6channels	97.37	99.57	54.53	97.82	91.33	86.90	39.08	93.01	71.37	82.46	73.62
YOLOv5l	6channels	93.96	99.56	73.25	97.27	56.00	80.86	67.57	85.16	75.56	81.70	73.33
YOLOv5s	6channels	98.10	99.56	84.73	96.69	61.14	60.03	67.47	82.76	73.62	81.34	73.72
YOLOv7	6channels	97.95	99.57	82.53	97.82	88.81	62.25	55.58	**93.17**	73.80	84.72	76.34
YOLO-FIRI	6channels	**99.39**	99.57	85.71	96.33	77.24	85.80	44.57	79.65	72.07	83.53	73.93
YOLO-CIR	6channels	96.07	99.56	77.38	96.60	87.45	75.58	51.78	91.60	75.97	84.50	75.38
MCIR-YOLO (Ours)	6channels	96.49	**99.59**	**91.35**	**98.13**	**91.97**	85.29	**68.29**	83.29	**77.83**	**89.30**	**80.31**

Table V summarizes and synthesizes the recognition performance of ten different YOLO models (YOLOv3, YOLOv4, YOLOv5s, YOLOv5m, YOLOv5l, YOLOv5x, YOLOv7, YOLO-FIRI [48], YOLO-CIR [32]) and the proposed model (MCIR-YOLO) on our self-collected IR pill image dataset, both for single-channel (1650 nm) and six-channel scenarios. The table displays the mAP0.5 detection results for eight types of pills: "Ivabradine", "Propranolol", "Vitamin C", "Aspirin", "Dexamethasone", "Riboflavin", "Vitamin B6", and "Prednisolone". Similarly, the results also present the metrics of Precision, mAP0.5, and mAP0.5:0.95, which collectively evaluate the overall detection performance.

From the comparison of single-channel testing, it is evident from the experiments that our new model achieved an mAP0.5 value improvement of 2.6% in overall performance. For individual types of medication, the MCIR-YOLO model achieved varying degrees of enhancement in accuracy for "Prednisolone", "Vitamin C", "Aspirin", "Vitamin B6". In the comparison of

multi-channel testing, due to the incorporation of six different IR bands and the MCIR-Fusion method, our model demonstrated better performance in multimodal fusion experiments compared to single-modal detection. Overall, the MCIR-YOLO model achieved first place in Precision, mAP0.5, and mAP0.5:0.95 among the eight comparative models, with respective leads of 2.3%, 4.6%, and 4.0%. Breaking down by individual medications, while maintaining the lead in "Prednisolone", "Vitamin C", "Aspirin", "Vitamin B6", the model also made advancements in "Dexamethasone".

From the detection results presented by the eight different medications, "Ivabradine", "Prednisolone", and "Aspirin" consistently achieved outstanding mAP0.5 scores exceeding 90% across all ten models. The detection results of the other five traditional models fluctuated considerably, but they achieved significant improvements to varying degrees after incorporating the MCIR-Fusion method, IADB branch, and dual IRA attention mechanism at the head position. However, for instance, the detection mAP0.5 accuracy of "Vitamin B6" still has not surpassed the threshold of 70%, indicating that further improvement is needed in the detection accuracy of challenging-to-identify medications.

V. Conclusion

This paper presents an object detection algorithm with multi-band image fusion, named MCIR-YOLO, for classifying and detecting white pills by using six-band IR images. First of all, the refined model incorporates a multi-level feature fusion technique. This method effectively amalgamated scale-specific feature information from the six IR modalities. Secondly, we introduced an image auxiliary detection branch, which optimized the main loss results through the calculation of independent branch losses. Finally, we designed two attention mechanism modules, significantly enhancing the IR feature information at the detection head position, thereby greatly improving the detection accuracy of white pills.

The MCIR-YOLO model utilized in the study achieved an mAP0.5 value of 89.30% on the self-built dataset using six-channel IR images. This represents an improvement of 7.96% over the baseline model YOLOv5s in six-channel detection, and a notable enhancement of 12.05% over the best-performing single-channel detection. The prediction accuracy should be further increased by additional IR bands. The impressive performance of the MCIR-YOLO model paves the way for valuable directions in the classification, recognition, and detection of white pills in the medical field. Simultaneously, it offers a novel approach for the multi-channel fusion of different IR spectral images.

References

[1] E. H. Nguyen et al., "Circle representation for medical object detection," *IEEE Trans. Med. Imag.*, vol. 41, no. 3, pp. 746–754, Mar. 2022, doi: 10.1109/TMI.2021.3122835.

[2] X. Li, J. Deng, and Y. Fang, "Few-shot object detection on remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Feb. 2021, Art. no. 5601614, doi: 10.1109/TGRS.2021.3051383.

[3] B. Mahaur and K. K. Mishra, "Small-object detection based on YOLOv5 in autonomous driving systems," *Pattern Recognit. Lett.*, vol. 168, pp. 115–122, Apr. 2023, doi: 10.1016/j.patrec.2023.03.009.

[4] F. Ma, B. Sun, and S. Li, "Facial expression recognition with visual transformers and attentional selective fusion," *IEEE Trans. Affect. Comput.*, vol. 14, no. 2, pp. 1236–1248, Apr.–Jun. 2023, doi: 10.1109/TAFFC.2021.3122146.

[5] J. Dai, X. Zhao, L. P. Li, and X. F. Ma, "GCD-YOLOv5: An armored target recognition algorithm in complex environments based on array LiDAR," *IEEE Photon. J.*, vol. 14, no. 4, Aug. 2022, Art. no. 3937711, doi: 10.1109/JPHOT.2022.3185304.

[6] N. Usuyama, N. L. Delgado, A. K. Hall, and J. Lundin, "ePillID dataset: A low-shot fine-grained benchmark for pill identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2020, pp. 3971–3977, doi: 10.1109/CVPRW50498.2020.00463.

[7] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587, doi: 10.1109/CVPR.2014.81.

[8] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1440–1448, doi: 10.1109/ICCV.2015.169.

[9] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.

[10] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 779–788, doi: 10.1109/CVPR.2016.91.

[11] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[12] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.

[13] Ultralytics, "YOLOv5: End-to-end object detection with YOLO' 2020. [Online]. Available: https:// github.com/ultralytics/yolov5

[14] Y. Wang, J. Ribera, C. Liu, S. Yarlagadda, and F. Zhu, "Pill recognition using minimal labeled data," in *Proc. IEEE 3rd Int. Conf. Multimedia Big Data*, Apr. 2017, pp. 346–353, doi: 10.1109/BigMM.2017.61.

[15] R. Huang, J. Pedoeem, and C. Chen, "YOLO-LITE: A real-time object detection algorithm optimized for non-GPU computers," in *Proc. IEEE Int. Conf. Big Data*, Dec. 2018, pp. 2503–2510, doi: 10.1109/BigData.2018.8621865.

[16] S. F. Williams, R. Stokes, P. L. Tang, and A. M. Blanco-Rodriguez, "Detection & identification of hazardous narcotics and new psychoactive substances using Fourier transform infrared spectroscopy (FTIR)," *Anal. Methods*, vol. 15, no. 26, pp. 3225–3232, 2023, doi: 10.1039/D3AY00766A.

[17] A. Sun, H. Guo, Q. Gan, L. Yang, Q. Liu, and L. Xi, "Evaluation of visible NIR-I and NIR-II light penetration for photoacoustic imaging in rat organs," *Opt. Exp.*, vol. 28, no. 6, Mar. 2020, Art. no. 9002, doi: 10.1364/OE.389714.

[18] E. Wan, Q. Zhang, L. Li, Q. Xie, X. Li, and Y. Liu, "The online in situ detection of indoor air pollution via laser induced breakdown spectroscopy and single particle aerosol mass spectrometer technology," *Opt. Lasers Eng.*, vol. 174, Mar. 2024, Art. no. 107974, doi: 10.1016/j.optlaseng.2023.107974.

[19] Y. Zheng et al., "A lightweight ship target detection model based on improved YOLOv5s algorithm," *PLoS One*, vol. 18, no. 4, Apr. 2023, Art. no. e0283932, doi: 10.1371/journal.pone.0283932.

[20] R. Yang, W. Li, X. Shang, D. Zhu, and X. Man, "KPE-YOLOv5: An improved small target detection algorithm based on YOLOv5," *Electronics*, vol. 12, no. 4, Feb. 2023, Art. no. 817, doi: 10.3390/electronics12040817.

[21] F. Qingyun, H. Dapeng, and W. Zhaokui, "Cross-modality fusion transformer for multispectral object detection," 2020, *arXiv:2111.00273*.

[22] Y. Xie, L. Zhang, X. Yu, and W. Xie, "YOLO-MS: Multispectral object detection via feature interaction and self-attention guided fusion," *IEEE Trans. Cogn. Dev. Syst.*, vol. 15, no. 4, pp. 2132–2143, Dec. 2023, doi: 10.1109/TCDS.2023.3238181.

[23] X. Tu et al., "An improved YOLOv5 for object detection in visible and thermal infrared images based on contrastive learning," *Front. Phys.*, vol. 11, May 2023, Art. no. 1193245, doi: 10.3389/fphy.2023.1193245.

[24] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new State-of-the-art for real-time object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 7464–7475, doi: 10.1109/CVPR52729.2023.00721.

[25] H. Yang et al., "Improved Apple fruit target recognition method based on YOLOv7 model," *Agriculture*, vol. 13, no. 7, Jun. 2023, Art. no. 1278, doi: 10.3390/agriculture13071278.
[26] K. Al-Hussaeni, I. Karamitsos, E. Adewumi, and R. M. Amawi, "CNN-based pill image recognition for retrieval systems," *Appl. Sci.*, vol. 13, no. 8, Apr. 2023, Art. no. 5050, doi: 10.3390/app13085050.
[27] H.-J. Kwon, H.-G. Kim, and S.-H. Lee, "Pill detection model for medicine inspection based on deep learning," *Chemosensors*, vol. 10, no. 1, Dec. 2021, Art. no. 4, doi: 10.3390/chemosensors10010004.
[28] T. T. Mac, "Application of improved Yolov3 for pill manufacturing system," *IFAC-PapersOnLine*, vol. 54, no. 15, pp. 544–549, 2021, doi: 10.1016/j.ifacol.2021.10.313.
[29] J. Liu, X. Fan, J. Jiang, R. Liu, and Z. Luo, "Learning a deep multi-scale feature ensemble and an edge-attention guidance for image fusion," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 1, pp. 105–119, Jan. 2022, doi: 10.1109/TCSVT.2021.3056725.
[30] Y. Cao, X. Luo, J. Yang, Y. Cao, and M. Y. Yang, "Locality guided cross-modal feature aggregation and pixel-level fusion for multispectral pedestrian detection," *Inf. Fusion*, vol. 88, pp. 1–11, Dec. 2022, doi: 10.1016/j.inffus.2022.06.008.
[31] X. Deng, E. Liu, S. Li, Y. Duan, and M. Xu, "Interpretable multi-modal image registration network based on disentangled convolutional sparse coding," *IEEE Trans. Image Process.*, vol. 32, pp. 1078–1091, Jan. 2023, doi: 10.1109/TIP.2023.3240024.
[32] J. Zhou et al., "YOLO-CIR: The network based on YOLO and ConvNeXt for infrared object detection," *Infrared Phys. Technol.*, vol. 131, Jun. 2023, Art. no. 104703, doi: 10.1016/j.infrared.2023.104703.
[33] D. Yu, S. Lin, X. Lu, B. Wang, D. Li, and Y. Wang, "A multi-band image synchronous fusion method based on saliency," *Infrared Phys. Technol.*, vol. 127, Dec. 2022, Art. no. 104466, doi: 10.1016/j.infrared.2022.104466.
[34] C. Zha, S. Luo, and X. Xu, "Infrared multi-target detection and tracking in dense urban traffic scenes," *IET Image Process.*, Feb. 2024, Art. no. ipr2.13053, doi: 10.1049/ipr2.13053.
[35] M. A. Azam et al., "A review on multimodal medical image fusion: Compendious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics," *Comput. Biol. Med.*, vol. 144, May 2022, Art. no. 105253, doi: 10.1016/j.compbiomed.2022.105253.
[36] J. Zhang, J. Lei, W. Xie, Z. Fang, Y. Li, and Q. Du, "SuperYOLO: Super resolution assisted object detection in multimodal remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Mar. 2023, Art. no. 5605415, doi: 10.1109/TGRS.2023.3258666.
[37] Y. Xiao, F. Codevilla, A. Gurram, O. Urfalioglu, and A. M. Lopez, "Multimodal end-to-end autonomous driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 1, pp. 537–547, Jan. 2022, doi: 10.1109/TITS.2020.3013234.
[38] Q. Li, C. Zhang, Q. Hu, H. Fu, and P. Zhu, "Confidence-aware fusion using dempster-shafer theory for multispectral pedestrian detection," *IEEE Trans. Multimedia*, vol. 25, pp. 3420–3431, Mar. 2022, doi: 10.1109/TMM.2022.3160589.
[39] S. Li, X. Kang, L. Fang, J. Hu, and H. Yin, "Pixel-level image fusion: A survey of the state of the art," *Inf. Fusion*, vol. 33, pp. 100–112, Jan. 2017, doi: 10.1016/j.inffus.2016.05.004.
[40] H.-J. Li, Z. Wang, J. Pei, J. Cao, and Y. Shi, "Optimal estimation of low-rank factors via feature level data fusion of multiplex signal systems," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 6, pp. 2860–2871, Jun. 2022, doi: 10.1109/TKDE.2020.3015914.
[41] W. Li, C. Chen, H. Su, and Q. Du, "Local binary patterns and extreme learning machine for hyperspectral imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 7, pp. 3681–3693, Jul. 2015, doi: 10.1109/TGRS.2014.2381602.
[42] H. Zhou, F. Luo, H. Zhuang, Z. Weng, X. Gong, and Z. Lin, "Attention Multihop graph and multiscale convolutional fusion network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Apr. 2023, Art. no. 5508614, doi: 10.1109/TGRS.2023.3265879.
[43] Y. Zhang and W. Chen, "Decision-level information fusion powered human pose estimation," *Appl. Intell.*, vol. 53, no. 2, pp. 2161–2172, Jan. 2023, doi: 10.1007/s10489-022-03623-z.
[44] H. Yang, T. Wang, and L. Yin, "Adaptive multimodal fusion for facial action units recognition," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2982–2990, doi: 10.1145/3394171.3413538.
[45] J. Li et al., "Deep learning in multimodal remote sensing data fusion: A comprehensive review," *Int J. Appl. Earth Observ. Geoinformation*, vol. 112, Aug. 2022, Art. no. 102926, doi: 10.1016/j.jag.2022.102926
[46] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Trans. Comput. Imag.*, vol. 3, no. 1, pp. 47–57, Mar. 2017, doi: 10.1109/TCI.2016.2644865.
[47] L. Li, L. Jiang, J. Zhang, S. Wang, and F. Chen, "A complete YOLO-based ship detection method for thermal infrared remote sensing images under complex backgrounds," *Remote Sens.*, vol. 14, no. 7, Mar. 2022, Art. no. 1534, doi: 10.3390/rs14071534.
[48] S. Li, Y. Li, Y. Li, M. Li, and X. Xu, "YOLO-FIRI: Improved YOLOv5 for infrared image object detection," *IEEE Access*, vol. 9, pp. 141861–141875, Oct. 2021, doi: 10.1109/ACCESS.2021.3120870.