IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2022.Doi Number

Improvement of the Image Analysis of Medicinal Herbs with YOLOv7-based Object Detection

KAIHUA CHE[1], YUHENG LIANG[1], PENGJIE WANG[1], WEI LV[1], TONGFEI LI[2] and XIAOLIN ZHU[1]

[1] School of Aliyun Big Data Application, Zhuhai College of Science and Technology, Zhuhai 519041 China

[2] School of Computer Science and Engineering Faculty of Innovation Engineering, Macau University of Science and Technology Macao SAR 999078 China

Corresponding author: XIAOLIN ZHU (paullooy@163.com)

This work is supported in part by the Guangdong University Research Platform and Project Support (2023ZDZX1049), and in part by the Innovation and entrepreneurship project for college students of Zhuhai University of Science and Technology(S202313684041), and in part by Zhuhai College of Science and Technology Doctoral Enhancement Program(ZCST2021-133)

ABSTRACT The burgeoning significance of the medical sector aligns with ongoing improvements in human quality of life. Within this context, herbal medicines serve as key modalities for treating both human and animal pathologies. Yet, the precise harvesting, identification, and taxonomic classification of medicinal herbs present substantial challenges. This is compounded by the extensive diversity and phenotypic similarities among different herb varieties, hindering accurate identification by non-specialists. Moreover, extant commercial datasets of Chinese medicinal herbs are plagued by significant redundancy and noise. To address these issues, we assembled ten labeled datasets of morphologically similar Chinese herbs. Leveraging advancements in deep learning, particularly the YOLOv7 algorithm—an industry standard for image classification—we introduce an augmented YOLOV7 framework for herb identification. Our modifications include the incorporation of self-attention mechanisms and novel convolutional networks in both the head and backbone sections of the original YOLOv7 architecture, along with a redefined loss function. Experimental evaluations reveal significant performance gains: a 10.55% improvement in MAP_0.5, a 13.23 % increase in accuracy, a 13.83% enhancement in MAP_0.5:0.95, and a 4% rise in recall compared to the baseline YOLOv7 model. These results substantiate that our modified YOLOv7 algorithm outperforms its predecessor for this specific dataset.

INDEX TERMS YOLOv7, Chinese medicinal herbs, self-attention, novel convolutional networks

I. INTRODUCTION

The escalation of economic prosperity has precipitated a parallel rise in societal expectations for healthcare quality. Within the framework of healthcare, the therapeutic efficacy of Chinese herbal medicines serves as an integral component [1], [2]. Consequently, public awareness and demand for Chinese herbal medical treatments and health supplements have surged. Data from the Third Census of Chinese Herbal Medicine Resources in China enumerates 12,807 distinct Chinese herbal medicines, including 28 classifieds as toxic. Categorically, these include 11,146 types of medicinal plants, 1,581 types of medicinal animals, and 80 types of medicinal minerals [3]. Nevertheless, the physical characteristics—such as shape, texture, and color—of many of these herbal medicines are frequently indistinguishable. This morphological resemblance presents a significant challenge, not only to the general populace but also to medical professionals specializing in this field. Accurate identification is paramount, as errors can jeopardize treatment efficacy and may induce severe adverse reactions, including patient mortality [4]. Thus, the precise identification of herbal constituents remains a pressing issue that warrants immediate attention.

Existing methodologies for herbal medicine identification present significant constraints. These methodologies are principally categorized into macro-identification and micro-identification techniques. While

micro-identification yields accurate results, it necessitates laboratory conditions, rendering it impractical for general consumer application [5]. Macro-identification, often employed in everyday settings, relies on observable characteristics such as shape, size, color, and olfactory properties. However, this approach demands specialized expertise and a wealth of experiential knowledge for accurate identification. Moreover, such identification is subject to environmental variables and remains inherently subjective and prone to uncertainty. To illustrate, consider the Chinese herbal medicines "JiXueteng" and "Mistletoe." JiXueteng is characterized by resinous secretions in its bast and wood formations consisting of either concentric elliptic or eccentric semicircular rings, with a pitch that is often skewed to one side. In contrast, Mistletoe exhibits a surface that ranges in color from yellowish-green to golden-yellow or yellowish-brown, and similarly features a pith that is typically inclined to one side. The prevalence of such morphologically similar herbs, both domestically and globally, compounds the complexity of accurate identification.

Given the imperative for reliable, objective herbal identification methods, advancements in the domain of computer vision offer a promising avenue. Selection of a robust image classification model has been an area of sustained focus within this field. The YOLO series of models has garnered considerable attention for their efficacy in real-time image classification and object localization. These models execute target detection and localization in a single forward pass, thereby facilitating rapid inference rates essential for real-time applications. Specifically, Joseph Redmon et al. [6] presented the YOLOV3 model, characterized by high detection accuracy on the COCO dataset, albeit at the expense of computational speed. Conversely, Jocher et al. [7] implemented data augmentation techniques in the YOLOv5 model, achieving a satisfactory trade-off between model speed and performance metrics.

In the present study, we propose advancements to the YOLOV7 model [8] for the specific application of herbal classification. The YOLOV7 model builds upon its predecessors by incorporating Efficient Layer Aggregation Network (E-ELAN) and employing model scaling strategies based on concatenation. Additionally, it employs coarse labeling strategies for auxiliary heads and fine labeling for dominant heads. We have further enhanced the model by refining its attention mechanism, specifically by replacing operations such as upsampling. These modifications have yielded significant improvements in model performance. Consequently, we posit that the application of the optimized YOLOV7-CHM model for herbal image classification constitutes both a scientifically rigorous and operationally effective approach.

The primary objective of this research is to facilitate the identification and classification of plants pertinent to the field of medicine. Such categorization is instrumental in assessing the efficacy of herbal treatments and mitigating the risks associated with erroneous herb identification. The salient contributions of this work are enumerated below:

- Data Collection: We have amassed a substantial image dataset that incorporates various backgrounds and includes images of multiple small herbs. This dataset provides a more realistic comparison base relative to existing publicly available datasets, which typically feature clean, isolated herb images.
- Loss Function Modification: To address the limitations of the CIOU loss function, specifically its inability to manage overlaps between actual and predicted frames, we employed the SIOU loss function. This alteration enhances both the model's convergence rate and its recognition accuracy.
- Attention Mechanism Enhancement: We integrated GAM Attention into the backbone layer of the model to enable heightened focus on salient micro-features, thereby reducing background noise. Additionally, CoordConv was implemented in the head layer to enrich the feature map with spatial coordinate information, thus elevating the accuracy of small target detection.
- Operator Replacement: The lightweight CARAFE operator was substituted for the upsampling component in the YOLOV7 model. This adjustment ameliorates issues related to image resolution, enhances the detection accuracy for smaller targets, and offers computational efficiency gains.
- Structural Augmentation: We incorporated a residual network architecture into the SPPCSPC module. This inclusion enhances the model's capability to process intricate targets and variable scenes, mitigates overfitting risks, simplifies the training process, and fortifies the model's robustness.

II. RELATED WORK

Advancements in deep learning techniques have steadily matured, leading to increased integration into medical applications. Such integration not only enhances diagnostic and treatment accuracy but also elevates operational efficiency. Wang et al. [9] developed a method for automated detection of abnormal respiratory activity via infrared video, thereby minimizing choking incidents. Zhang et al. [10] employed computer vision systems to identify absent tubing in infusion bag assemblies, ensuring proper fluid delivery during clinical interventions. Bouhissi et al. [11] utilized machine learning algorithms for the prediction of gestational diabetes, while Venkat et al. [12] employed a random forest algorithm to ascertain the correlation between cardiovascular diseases and demographic variables. In parallel, the burgeoning

advancements in deep learning have precipitated significant growth in the field of computer vision, with YOLO models standing as a cornerstone. Won et al. [13] reduced YOLOv3's feature extraction and box detection layers to increase detection speed and improved accuracy by

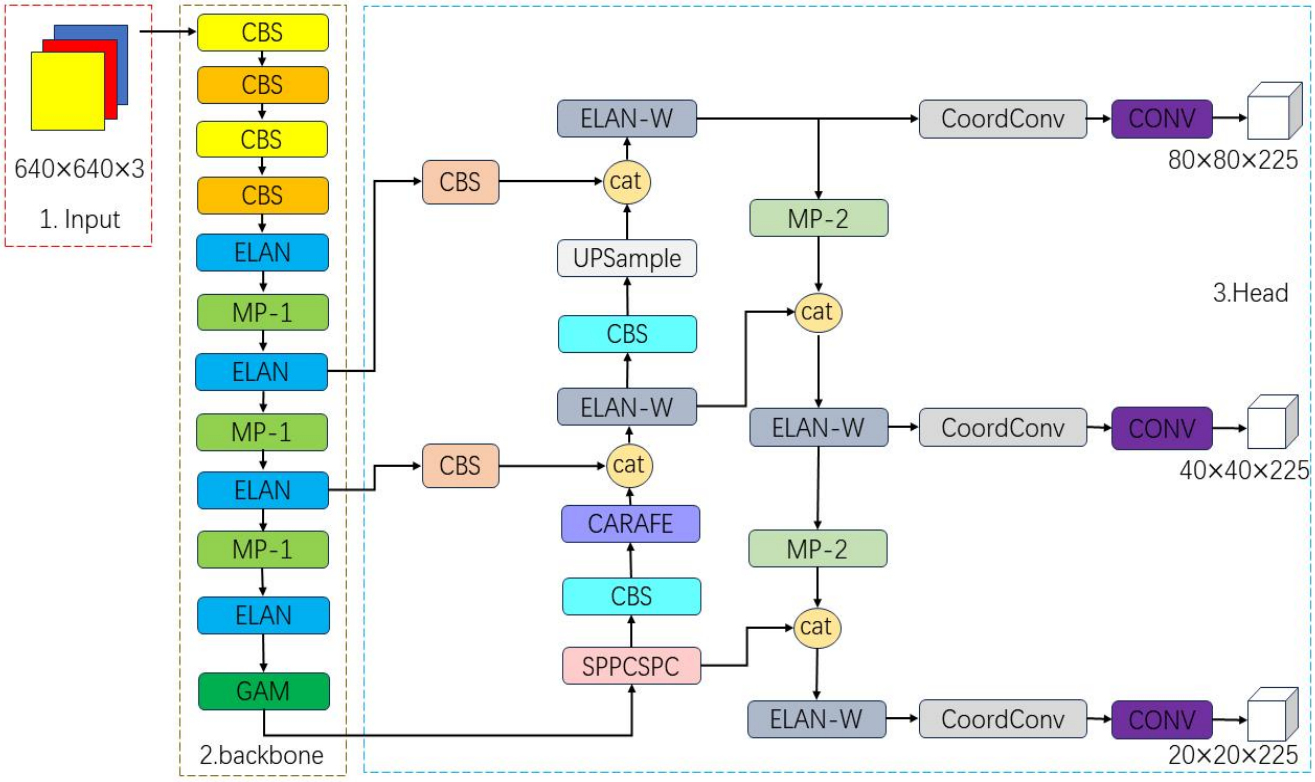

Figure 1 THE STRUCTURE OF YOLOV7-CHM

using de-identification technology. Zhao et al. [14] refined the bounding box width and height prediction methods in YOLOV3 via improved clustering techniques. Bochkovskiy et al. [15] incorporated self-adversarial training and hyper-parameter optimization through genetic algorithms in YOLOV4. Tanbunan et al. [16] compared the performance metrics of YOLOV4 and YOLOV4-tiny in robots utilizing omnidirectional cameras. Finally, Zheng et al. [17] introduced a low-visibility target detection method based on YOLOV5, and Li et al. [18] proposed an enhanced YOLOV6 architecture, dubbed EfficientRep, which is built on RepVGG. Continued advancements in the application of YOLO models to various sectors have been noteworthy. Kaur et al. [19] fused the YOLOV6 model with an augmented logistic regression algorithm to optimize traffic sign recognition. Concurrently, Wang et al. [20] implemented a novel scaling strategy within the YOLOV7 framework to preserve the model's optimal architecture. This particular YOLOV7 model has garnered considerable attention within the medical community. For example, Ahmad et al. [21] developed an attention-based YOLO7 variant aimed at automatic detection of gastric lesions in endoscopic imaging, effectively enhancing the capability to identify minute gastric anomalies. Technol et al. [22] introduced an enhanced YOLOv7-E6E model for real-time detection of gestational sacs in ultrasound images, thereby informing farm production plans. Wu et al. [23] implemented a coordinate attention mechanism within a YOLOV7 framework, successfully enhancing the accurate detection of nasopharyngeal carcinoma lesions in magnetic resonance imaging.

In the realm of herbal medicine, the utility of deep learning for plant classification is increasingly acknowledged. Previously, identification primarily relied on macroscopic characteristics such as color and shape. Recent developments, however, focus on leveraging computational models for more nuanced analysis. Sun et al. [24] utilized a convolutional neural network along with activation functions for herbal medicine identification, thereby mitigating the influence of varying backgrounds. Liu et al. [25] incorporated the Inception module of the Google Net framework to both enhance performance and minimize computational load. Hao et al. [26] devised an 'MTAL' model that amalgamates mutual learning and triplet attention techniques to elevate classification performance through supervised learning and imitation loss. Luo et al. [27] examined the efficacy of Principal Component Analysis (PCA) against Support Vector Machines (SVM) in herbal medicine classification, finding SVM to yield superior accuracy. Huang et al. [28] enhanced the feature

extraction module within the AlexNet model, there by improving its effectiveness in classifying Chinese herbal medicine images. Xing et al. [29] developed a model based on DenseNet and applied transfer learning techniques, resulting in a significant boost in the accuracy of traditional Chinese medicine recognition. Yue et al. [30] introduced a VCSEL-based time-delay reservoir optical feedback computing system. Through the integration of parallel processing and orthogonal optical feedback, they achieved a minimum recognition error rate of 1.7% at high processing speeds. Consequently, the integration of deep learning techniques in herbal medicine classification emerges as a scientifically robust approach.

III. Improved Work

The proposed YOLOV7-CHM model is bifurcated into two primary components: the Backbone and the Head. The Backbone serves as the feature extraction network, comprising convolutional, pooling, and other layers designated for the extraction of high-level features from the input imagery. Conversely, the Head, or the detection head network, conducts target detection procedures based on the feature maps produced by the Backbone. The comprehensive architecture is depicted in Figure 1.

A. GLOBAL ATTENTION MECHANISM

The Global Attention Module (GAM) [31] functions as a self-attention mechanism amalgamating both channel and spatial attentiveness to augment image feature performance. Specifically, in scenarios requiring the detection of diminutive or partially concealed objects, GAM excels in extracting pivotal features from relevant regions to facilitate object recognition. Derived from the CBAM network architecture, GAM incorporates modifications to its internal sub-modules. The mathematical formulations governing GAM are presented in equations (1) and (2):

$$F_2 = M_C(F_1) \otimes F_1 \#(1)$$

$$F_3 = M_S(F_2) \otimes F_2 \#(2)$$

In Equations (1) and (2), F_1 denotes the input feature, F_2 signifies the intermediate state, and F_3 represents the output feature. Additionally, M_C and M_S stand for the channel attention map and spatial attention map, respectively. The symbol $\otimes$ designates element-wise multiplication. The architecture of the GAM is illustrated in Figure 2.

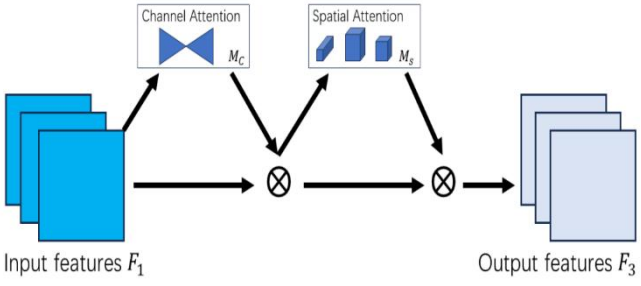

Figure 2 THE STRUCTURE OF GAM

As discernible from Figure 2, the GAM is comprised of two principal components: Channel Attention and Spatial Attention. The Channel Attention module employs 3D convolutional operations to preserve 3D spatial information, which is subsequently fed into a two-layer Multilayer Perceptron (MLP). This operation amplifies the correlation between channel and spatial dimensions while minimizing information dispersion, thereby enhancing overall model performance.

The Spatial Attention module commences by processing the input features through a 7x7 convolutional network. To mitigate computational complexity and memory overhead, the resolution of the feature map is reduced via a downsampling rate r. Subsequent to another set of 7×7 convolutional operations and channel number adjustments, the final output undergoes a sigmoid activation function. Spatial Attention aims to isolate critical feature maps and mitigate noise interference. Through learning the interrelation between channel and spatial dimensions, the GAM self-attention mechanism is proficient in either enhancing or attenuating features, consequently elevating both performance and feature representation of the model.

B. THE SHORTCUT-SPPCSPC

The SPPCSPC module in YOLOv7 serves as an advanced spatial pyramid pooling mechanism designed for multi-scale feature extraction. The module operates by performing pooling on input feature maps across varying scales, followed by channel-wise fusion via 1x1 convolution operations. This facilitates more comprehensive feature information extraction and thus enhances YOLOv7's adaptability to target detection across different scales, consequently improving detection accuracy

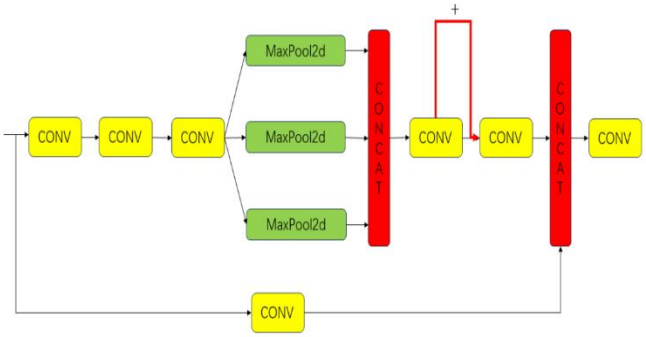

Figure 3 THE STRUCTURE OF SPPCSPC

and recall rates.

In the present study, we integrate residual connections into the SPPCSPC architecture. Residual connectivity serves as a pivotal mechanism in deep neural networks, mitigating issues associated with gradient vanishing and explosion. The underlying principle involves the introduction of skip connections, enabling direct information flow across network layers, thereby facilitating more efficient training. Traditional backpropagation methods require gradient descent to traverse sequentially

through each layer, which can contribute to gradient vanishing, adversely affecting model convergence rates. The introduction of residual connections allows for element-wise addition of input and output features from adjacent layers, simplifying gradient updates without necessitating a full mapping operation. The synergy between residual connections and SPPCSPC efficiently curtails computational complexity. Specifically, residual connectivity minimizes the depth of gradient

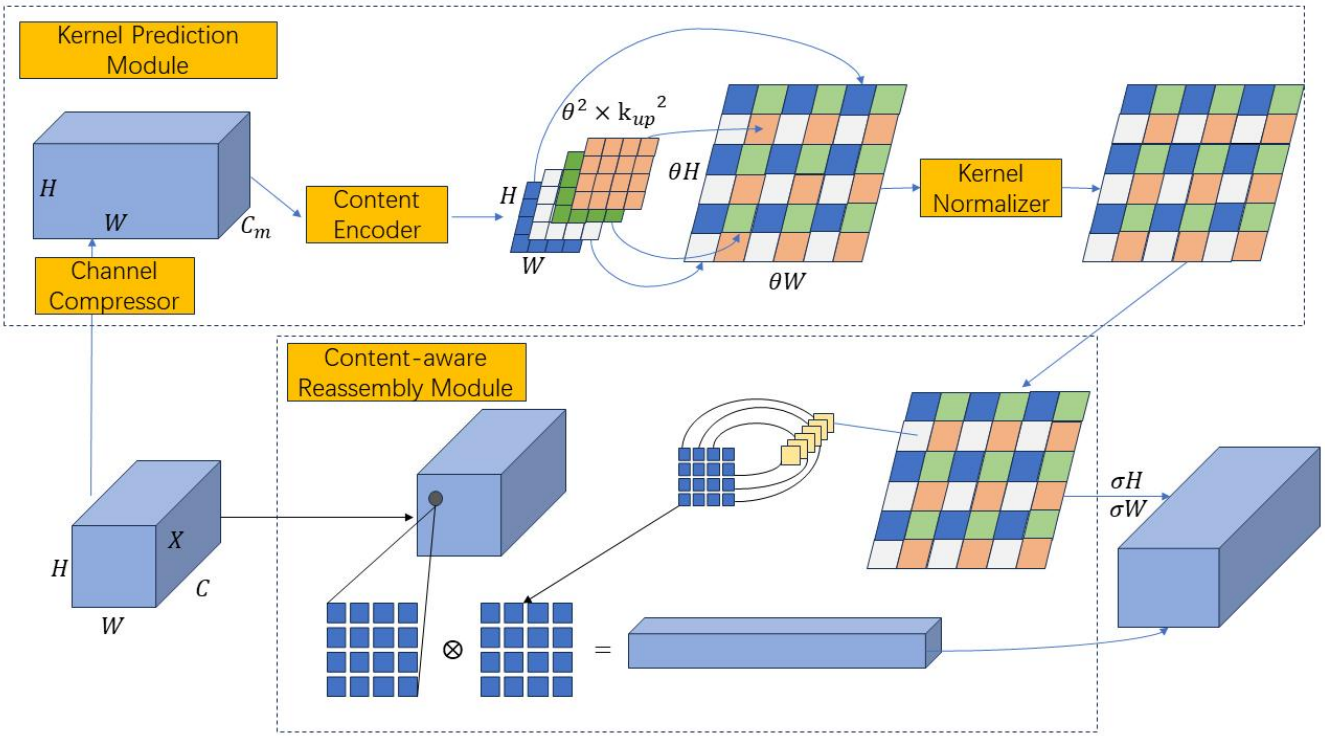

Figure 4 THE STRUCTURE OF CARAFE

propagation via skip connections, while the SPPCSPC mechanism reduces the computational cost associated with pooling operations. The architecture featuring the integration of residual connections within the SPPCSPC module is illustrated in Figure 3, with the implemented shortcut denoted by a thick red line.

C. THE CARAFE STRUCTURE

Traditional up-sampling techniques generally employ bilinear interpolation methods characterized by expanding low-resolution feature maps to a higher resolution, dictated by the task requirements. These methods calculate new pixel values based on adjacent pixels to populate the upscaled feature map. However, conventional interpolative up-sampling approaches may inadequately capture complex, non-linear pixel-to-pixel relationships, leading to detail-deficient and semantically poor feature maps. Such inadequacies are particularly detrimental in high-precision segmentation or object detection tasks. Additionally, the interpolation process can introduce information loss and image blurring. To address these limitations, we propose the incorporation of the CARAFE [32] module as a replacement for the traditional up-sampling module.

The CARAFE module consists of two primary components: the upsampling kernel prediction unit and the feature reorganization unit. In the former, a 1×1 convolutional layer is first utilized to compress the initial feature map dimensions $C \times H \times W$ into $C_m \times H \times W$, significantly reducing subsequent computational requirements. The upsampling kernel, assumed to be $k_{up} \times k_{up}$, is designed to be unique for each pixel, yielding a final kernel size of $\sigma H \times \sigma W \times k_{up}^2$. This kernel is then normalized to ensure the convolutional weights sum to unity.

In the feature reorganization unit, a 1×1 convolutional layer outputs the compressed feature map $C_m \times H \times W$, Each pixel in this map is mapped one-to-one with the input feature map to obtain a $k_{up} \times k_{up}$ region. A dot product operation is performed between the predicted upsampling kernel and its corresponding k_{up} region, with different channels at the same spatial location sharing the same upsampling kernel. This CARAFE operation enables feature reorganization within local regions, assigning higher weights to salient features and thereby directing sampling focus towards target regions while ignoring background noise. Moreover, the CARAFE module's content-aware upsampling kernel generation enhances the effective

receptive field, facilitating more informative sampling while maintaining computational efficiency and a minimal memory footprint. Figure 4 is shown as the structure of carafe.

D. THE COORDCONV MODULE

The YOLOv7 model originally employs a Repconv module, which, through its multi-projection layers, is capable of efficient multi-scale feature extraction from images. This enables superior object and structure detection across varying scales, but at the cost of increased computational resources. Particularly in our herbal medicine dataset, challenges arise due to background occlusion, varying lighting conditions, and instances of multiple small herbal constituents. Moreover, the data often exhibit low-contrast herbal targets with indistinct boundaries, and images may also have inconsistent dimensions.

To mitigate these issues, we propose the integration of the CoordConv feature extraction module. CoordConv not only excels in high-level multi-scale feature extraction but also augments this with coordinate information, thereby enhancing precise localization and classification accuracy for herbal targets. This is particularly advantageous for images featuring multiple small herbal components under diverse lighting conditions and with background occlusions. CoordConv improves the model's spatial perception by including coordinate information as part of the convolutional operation, subsequently enhancing classification metrics. Figure 5 is shown as the structure of

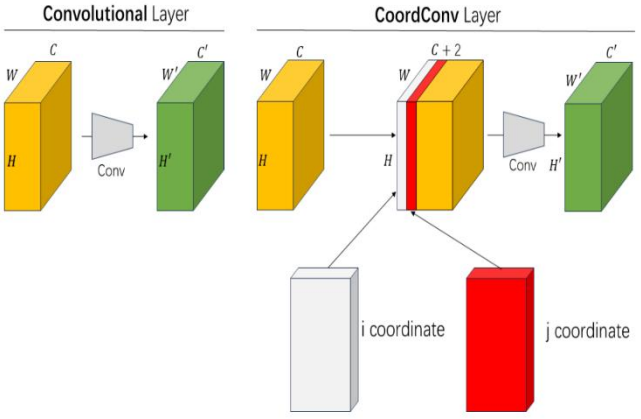

Figure 5 THE STRUCTURE OF CARAFE

CoordConv.

Distinguishing CoordConv from conventional convolutional networks is its initial addition of two coordinate channels—an i-coordinate and a j-coordinate—to the output feature map. Following this, coordinate data and original pixel values are merged within the channel dimension to produce a composite input feature map. This enriched feature map then undergoes a subsequent convolution operation to acquire more advanced features that include pixel-level coordinate information. This facilitates the capturing of essential attributes like shape, position, and object boundaries. Through the application of CoordConv, the model achieves better spatial understanding of the herbal medicines' location and morphological traits. Notably, the addition of coordinate channels within the convolution operation has a minimal impact on model complexity and computational overhead.

E. THE SIOU FUNCTION

The loss function serves as a critical component in machine learning, facilitating both the evaluation of model performance and the optimization of model parameters. Specifically, it quantifies the discrepancies between a model's predictions and actual data, thereby informing the model's accuracy. Given the task- and problem-specific needs, the selection of an appropriate loss function is vital for optimizing performance and mitigating overfitting.

In the YOLOv7 framework, the default loss function employed is the CIoU [34] loss function. This function incorporates three geometric metrics: the overlap between the predicted and actual bounding boxes, the distance between their centroids, and the consistency of their aspect ratios. Furthermore, a penalty term is introduced to account for scenarios in which the aspect ratios of the predicted and actual boxes differ, despite coinciding centroids and identical Intersection-over-Union (IoU) values. The mathematical formulations for IoU and CIoU are detailed in Equations (3) and (4), respectively:

$$IoU = \left|\frac{b \cap b^{gt}}{b \cap b^{gt}}\right| \#(3)$$

$$L_{CIoU} = 1 - IoU + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha v \#(4)$$

Where b is the prediction box, b^{gt} is the real box, v is used to calculate the consistency of the aspect ratio between the prediction box and the target box, and α is a compensation parameter, with the coefficient increasing as the prediction box and the target box IoU become larger. c is the diagonal length of the smallest bounding box contained by the two boxes (usually rectangular boxes). This diagonal length is conventionally employed to ascertain the relative position or orientation between two bounding boxes.

One limitation of the CIoU loss function is its inadequate consideration of non-overlapping scenarios between the actual and predicted boxes. To address this, we employ the SIoU [35] loss function, which incorporates directional information between the actual and predicted frames. For the first time, it utilizes prediction targets on the X/Y axis and redefines the penalty metrics by calculating the angle between the predicted and actual frames. The SIoU loss function comprises two components, as formalized in Equation (5):

$$L = W_{box}L_{box} + W_{cls}L_{cls} \#(5)$$

Where L_{cls} is the focus loss, W_{box} is the loss weight of the detection box and W_{cls} is the loss weight of the classification result. The equation (6) about the L_{box} is as follow:

$$L_{box} = 1 - IoU + \frac{\Delta + \Omega}{2} \#(6)$$

The parameters of this equation (6) are The Angle cost, The Distance cost Δ, The Shape cost Ω and The IoU cost SIoU. The Angle cost calculates the angle α between the two frames and dynamically adjusts the angle of the two frames according to this angle. If α is less than π/4, it continues to minimize α, otherwise it minimizes π/2 minus α. The Distance Cost component minimizes the centroidal distance by computing the distance between the centroids and applying a penalty function to mitigate misalignment. The

Figure 6 THE ORIGINAL YOLOV7 DETECT

Figure 7 THE YOLOV7-CHM DETECT

Shape Cost adjusts the aspect ratios of the predicted and actual boxes to improve shape similarity. The IoU calculates the overlap ratio by intersecting and concatenating the predicted and actual frames. Our enhanced YOLOv7 model leverages the SIoU loss function, achieving faster convergence and superior localization accuracy compared to traditional CIoU.

IV. EXPERIMENT WORK

A. EXPERIMENT SETTING AND DATASET

1) EXPERIMENT DETAILS

In the experimental setup, the hardware employed was an NVIDIA Tesla V100S, while the software versions for PyTorch and CUDA were 1.14 and 12.0, respectively. YOLOv7 served as the initial model for training. The training parameters included pre-training weights, an epoch setting of 300, a batch size of 16, an image size of 640, a

learning rate of 0.01, a weight decay of 0.0005, and the utilization of SGD as the optimizer.

2) DATASET INFORMATION

The quality of the dataset is often pivotal in enhancing model performance. In the domain of Chinese herbal medicine research, there exists a conspicuous absence of large, accurate, and professionally certified datasets. Consequently, we based our dataset on sales data from a local Chinese herbal medicine store, focusing on the top 10 categories of herbs. A total of 3,987 images were initially obtained for these categories through web scraping techniques. However, the dataset was compromised by a high level of duplication and noise. To rectify this, we employed a hashing algorithm to identify and remove similar images. Additionally, images where text occupied more than 70% of the frame were manually eliminated. To maintain a balanced distribution of training samples across herb categories, the final dataset consisted of 1,329 images. The dataset distribution is detailed in Table 1.

Table 1 THE DISTRIBUTION OF DATASET

Dataset	Amount
Aiye	117
Baibiandou	136
Baibu	137
Baihe	157
Gancao	121
Gouqi	166
Dangshen	153
Cangzhu	130
Jinyinhua	137
Zicao	75

B. EXPERIMENT SETTING AND DATASET

To rigorously assess the efficacy of our model, we employed four key performance indicators: accuracy, recall, mean of Average Precision (mAP), and Precision at a Confidence Threshold of 0.95 with a 0.5 IoU threshold, denoted as mAP@0.95_0.5. The relevant formulas for these metrics (7) – (10) are provided subsequently.

$$Recall = \frac{TP}{TP + FN} \#(7)$$

$$Precision = \frac{TP}{TP + FP} \#(8)$$

$$AP = \int_0^1 P(Recall) \#(9)$$

$$mAP = \frac{\Sigma_{i=0}^{n} AP_i}{n} \#(10)$$

In this nomenclature, TP stands for True Positive, indicating a correct classification of a positive example; FP signifies False Positive, representing an incorrect classification of a negative example as positive; FN denotes False Negative, marking a positive example incorrectly classified as negative; and AP refers to Average Precision. Among these metrics, Recall quantifies the model's ability to accurately identify all true positive examples, with higher recall scores indicating fewer false negatives. Accuracy gauges the overall correctness of the model's predictions, with higher values suggesting greater reliability. AP serves as a performance measure in target detection and information retrieval tasks, quantifying accuracy across various confidence thresholds and mapping the area under the precision-recall curve.

C. EXPERIMENTAL RESULTS

1) DIFFERENT MODELS EXPERIMENT

For empirical validation, we conducted experiments contrasting our model with several extant YOLO algorithms. To ensure methodological rigor, identical equipment, experimental parameters, and datasets were employed across all experimental setups. As indicated in Table 2, our model demonstrated significant superiority in accuracy, recall, mAP_0.5, and mAP@0.95_0.5 metrics. Notably, the mAP_0.5 value attained a peak of 0.991, underscoring the enhanced effectiveness of our model relative to the referenced YOLO algorithms.

2) DIFERENT EXPERIMENT LOSS

Table 3 DIFFERENT LOSS

Different Loss	YOLOV7	YOLOV7-CHM
Train-cls-loss	0.01419	0.007288
Train-box-loss	0.02241	0.01529
Train-obj-class	0.008323	0.006681
Val-cls-loss	0.001816	0.000778
Val-box-loss	0.006896	0.00636
Val-obj-loss	0.003566	0.003047

As delineated in Table 3, subsequent to training the model over 300 epochs, the derived metrics for classification loss, bounding box position loss, and object loss in both test and validation sets markedly outperform those of the original YOLOv7 model. Notably, the classification loss metric is halved, underscoring the superior performance of our optimized model in object classification and localization tasks.

3) SELF ATTENTION ALALYSIS

In an effort to validate the efficacy of the newly incorporated attention mechanism, comparative experiments were executed against established attention mechanisms, namely SimAM and SE attention. The ensuing data, illustrated in the referenced figure, demonstrates that the incorporation of the GAM significantly elevates various performance indicators relative to the baseline model. Specifically, precision improved by 11.07%, recall by 7.81%, mAP_0.5 by 10.55%, and mAP@0.5_0.95 by 11.71% when compared to SimAM integration. Furthermore, relative to the SE model, GAM integration yielded improvements of 3.09% in precision, 1.83% in mAP_0.5, and 3.12% in mAP@0.5_0.95. These empirical findings unequivocally

affirm the superiority of integrating GAM into our model.
For further clarification,

Table 2 THE DIFFERENT MODELS EXPERIMENT

Model	Precision	Recall	mAP_0.5	mAP@0.5_0.95
Yolov4-tiny	0.5876	0.8532	0.8478	0.4090
Yolov4	0.9683	0.9327	0.9683	0.5742
Yolov5n	0.9128	0.9431	0.9573	0.7231
Yolov7	0.8211	0.9074	0.8855	0.6210
Yolov7-CHM	0.9534	0.9486	0.991	0.7593

Table 4 THE DIFFERENT SELF-ATTENTION EXPERIMENT

Model	Precision	Recall	mAP_0.5	mAP@0.5_0.955
Yolov7	0.8211	0.9074	0.8855	0.621
Yolov7+SiMam	0.8427	0.8755	0.9288	0.6422
Yolov7+SE	0.9225	0.9633	0.9727	0.7281
Yolov7_CHM	0.9534	0.9486	0.991	0.7593

Table 5 ABLATION EXPERIMENT

Model	Precision	Recall	mAP_0.5	mAP@0.5_0.95
Yolov7	0.8211	0.9074	0.8855	0.6210
Yolov7+Siou	0.8818	0.8771	0.9525	0.6944
Yolov7+Siou+GAM	0.9177	0.9612	0.9733	0.7286
Yolov7+Siou+GAM+CoordConv	0.9622	0.9079	0.9822	0.7286
Yolov7+Siou+GAM+CoordConv+Carafe	0.9285	0.9226	0.9602	0.7237
Yolov7+Siou+GAM+CoordConv+Carafe+Shortcut	0.9534	0.9486	0.991	0.7539

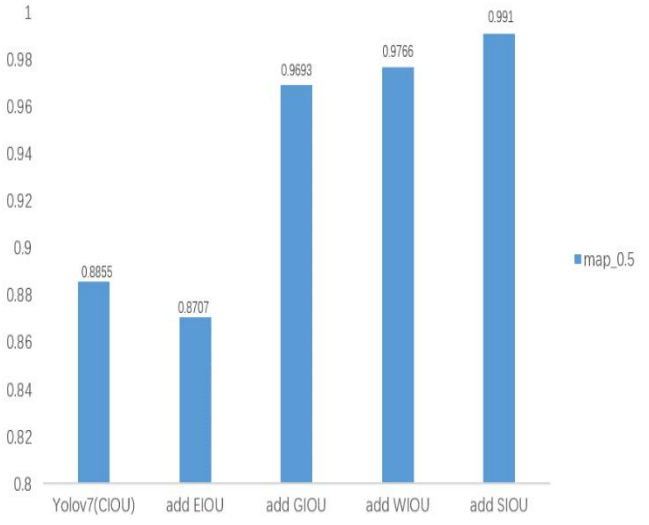

Figure 8 DIFFERENT LOSS FUNCTION EXPERIMENT

Table 4 depicts the comparative analysis of different self-attention mechanisms employed in the experiment.

4) IMPROVED LOSS FUNCTION ALALYSIS

This segment of the study aims to scrutinize the performance metrics of the model under different loss functions. As illustrated in the aforementioned figure, our baseline model employs the CIoU loss function, while subsequent iterations utilize EIou, GIou, WIou, and SIou loss functions, respectively. Experimental data reveal a 10.55% improvement in mAP_0.5 for the baseline model

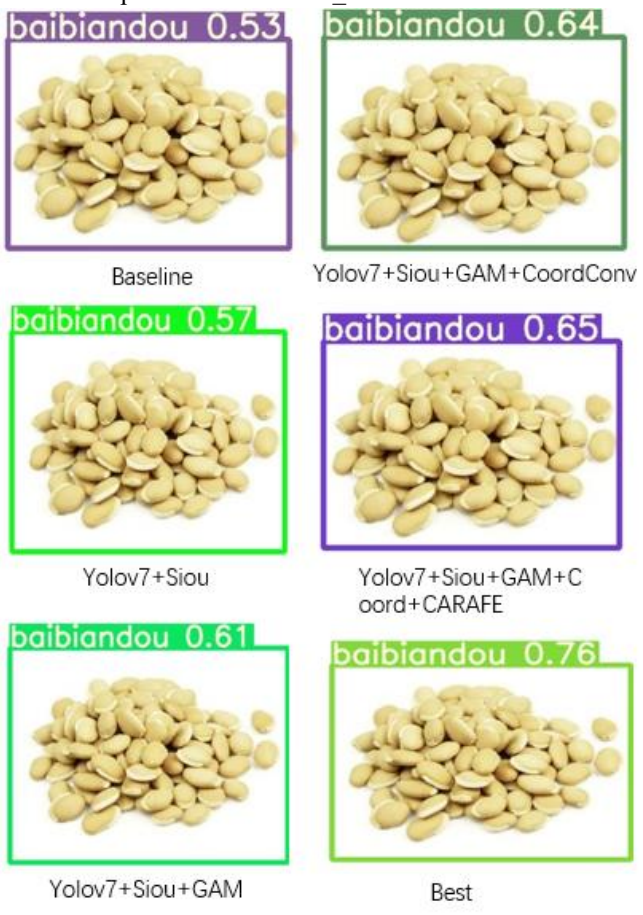

Figure 9 ABLATION ANALYSIS

when employing the SIoU loss function. Varied degrees of improvement were also observed for alternative loss functions, thereby substantiating the relative efficacy of the SIoU loss function in our experimental context. Figure 8 elucidates the comparative performance of different loss functions employed in the study.

D. Ablation Analysis

To elucidate the influence of incremental modifications on the model's performance, ablation experiments were conducted, with results tabulated in Table 5. Initially, a baseline was established using the original model featuring the CIoU loss function. Subsequently, the CIoU was substituted with the SIoU loss function. Although this change induced a minor decrement in recall, it manifested significant gains of 6.07% in precision and 7.31% in mAP_0.5, implying enhanced capabilities in accurate bounding box localization.

Further incorporation of the GAM yielded an 8.41% increment in recall, which suggests that the GAM effectively mitigates recognition errors. Replacing RepConv with CoordConv led to a 4.45% amelioration in model accuracy, evidencing that the integration of coordinate data into feature layers contributes positively to object recognition efficacy.

In an effort to optimize computational efficiency, the upsampling module was supplanted by CAREFE, resulting in a performance improvement of 1.52% relative to the original model. This modification achieves a judicious equilibrium between parameter reduction and performance enhancement.

Lastly, the introduction of residual structures within the SPPCSPC module culminated in an mAP_0.5 score of 0.991 and an mAP@0.5_0.95 score of 0.7593, marking the pinnacle of performance metrics within these ablation studies. In summary, each implemented modification engendered variable but consistent improvements in the model's performance metrics, corroborating the model's scientific robustness and operational utility. Figure 8 delineates the differential impact of these modifications, further validating the effectiveness of the implemented changes.

Figure 9 presents the results of the ablation experiments conducted on a singular image. The figure elucidates the differential impact of each implemented modification, thereby substantiating the efficacy of our model improvements.

V. CONCLUSION

In the present study, we introduce an innovative approach for the classification and recognition of herbal images. Initially, datasets comprising diverse backgrounds and accurately annotated labels were collated. The GAM was then integrated into the backbone architecture, enhancing the model's capacity for feature extraction. Subsequent modifications were implemented based on the residual network structure, resulting in a revamped SPPCSPC network architecture that markedly elevated the model's accuracy.

Concomitantly, in an effort to attenuate computational demands, the CARAEF lightweight operator was incorporated and the original RepConv module was substituted with CoordConv, enabling a more nuanced exploration of feature map coordinate relationships. Additionally, the adoption of the SIoU loss function accelerated the model's convergence process by computing the angular deviation between actual and predicted images.

The aggregate impact of these modifications yielded substantial improvements in key performance metrics, including accuracy and recall. The importance of this enhanced model is underscored by its applicability to the critical task of herbal medicine classification and recognition. The accurate categorization and verification of herbal varieties bear direct implications for consumer safety and the overall efficacy of herbal preparations. Furthermore, this study serves as a valuable tool for both regulatory agencies and consumers in ensuring that marketed herbal medicines meet requisite quality standards, thereby contributing to the mitigation of risks associated with adulteration or contamination.

REFERENCES

[1] G. Zheng, Y. -J. Zhang, H. Zhang, M.-F. Li, X.-W. He, Y.-T. Qi, J.-P. Zhan, H.-T. Guo, "Biological functional analysis of Chinese herbal medicines against wind-cold-dampness syndrome," in 2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD). IEEE,2016, pp. 1364-1368.

[2] L. Yan, B.-Y. Xu, P. Liu, A.-H. Ou, Y. Xu, B.-Y. Ou, R.-Y. Xu, Z.-Z. Wei, L. Zeng, R. Mai, J.-B. Chen, W. Zhou, H. M. Zeng, J. L. Zeng. "A randomized controlled study of emotion and quality of life on mild to moderate insomnia intervened by Chinese Medicine in integrated program," 2012 IEEE International Conference on Bioinformatics and Biomedicine Workshops. IEEE, 2012, pp. 428-434.

[3] J. Ming, L. Chen, Y. Cao, C. Yu, B.-S. Huang, K.-L. Chen, "Rapid identification of nine easily confused mineral traditional Chinese medicines using Raman spectroscopy based on support vector machine," Journal of Spectroscopy, vol. 2019, pp. 6967984, Jan.2019. doi:10.1155/2019/6967984

[4] E. Ernst. "Adulteration of Chinese herbal medicines with synthetic drugs: a systematic review," in Journal of Internal Medicine, vol. 252, pp: 107-113, Aug.2002. 10.1046/j.1365-2796.2002.00999.x

[5] J. Zhou, L. Zhou, Y.-S. Zheng, Z. Zhu, X.-M. Xu, G.-H. Jiang, "Terahertz Spectroscopic Identification of Different Species of Herbal Medicine Fritillaria", in 2019 44th International Conference on Infrared, Millimeter, and Terahertz Waves (IRMMW-THz), IEEE, 2019, pp. 1-2

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

[7] J. Glenn, C. Ayush, S. Alex, et al. "ultralytics/yolov5: v7. 0-YOLOv5 SOTA Realtime instance segmentation." Zenodo, Nov, 2022. doi:10.5281/zenodo.7347926.

[8] C. Y. Wang, A. Bochkovskiy, H.-Y. Mark Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," 2022, arXiv:2207.02696.

[9] C.-W. Wang, A. Ahmed and A. Hunter, "Vision Analysis in Detecting Abnormal Breathing Activity in application to Diagnosis of Obstructive Sleep Apnoea," 2006 International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, pp. 4469-4473, doi:10.1109/IEMBS.2006.260648.

[10] Q. Zhang, K. Liu, B. Huang, “Research on Defect Detection of The Liquid Bag of Bag Infusion Sets Based on Machine Vision,” Academic Journal of Science and Technology, 2023, vol. 5, pp. 186-197.
[11] H. El Bouhissi, R. E. Al-Qutaish, A. Ziane, K. Amroun, N. Yaya and M. Lachi, "Towards Diabetes Mellitus Prediction Based on Machine-Learning," in 2023 International Conference on Smart Computing and Application (ICSCA), pp. 1-6.Feb. 2023.
[12] V. Venkat, H. Adbelhalim, W. DeGroat, S. Zeeshan, Z. Ahmed, “Investigating genes associated with heart failure, atrial fibrillation, and other cardiovascular diseases, and predicting disease using machine learning techniques for translational research and precision medicine,” Genomics, vol. 115, pp. 0888-7543
[13] J. -H. Won, D. -H. Lee, K. -M. Lee and C. -H. Lin, "An Improved YOLOv3-based Neural Network for De-identification Technology," 2019 34th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC), JeJu, Korea (South), 2019, pp. 1-2.
[14] L.-Q. Zhao, S.-Y. Li, “Object Detection Algorithm Based on Improved YOLOv3”, Electronics, Mar. 2020, vol. 9, pp. 537.
[15] A. Bochkovskiy, C.-Y. Wang, H.-Y. Mark Liao, “YOLOv4: Optimal Speed and Accuracy of Object Detect,” 2020, arXiv:2004.10934
[16] I. H. Tambunan, D. Silaen, F. Michael, B. A. Sihotang and K. G. Sitanggang, "Performance Comparison of YOLOv4 and YOLOv4-Tiny Algorithm for Object Detection on Wheeled Soccer Robot," 2022 IEEE International Conference of Computer Science and Information Technology (ICOSNIKOM), IEEE, 2022, pp. 1-6.
[17] Y. Zheng, Y. Zhan, X. Huang and G. Ji, "YOLOv5s FMG: An Improved Small Target Detection Algorithm Based on YOLOv5 in Low Visibility," IEEE Access, 2023,
[18] C.-Y, Li, L.-L. Li, H.-J. Jiang, K.-H. Weng, et al, “YOLOv6: A Single-Stage Object Detection Framework for Industrial Applications,” 2022, arXiv:2209.02976
[19] R. Kaur and J. Singh, "Local Regression Based Real-Time Traffic Sign Detection using YOLOv6," 2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), IEEE, 2022, pp. 522-526.
[20] C.-Y. Wang, A.Bochkovskiy, H.-Y. Mark Liao, “YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors,” IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7464-7475
[21] S. Ahmad, J. -S. Kim, D. K. Park and T. Whangbo, "Automated Detection of Gastric Lesions in Endoscopic Images by Leveraging Attention-Based YOLOv7," in IEEE Access, 2023, vol. 11, pp. 87166-87177.
[22] T.-K. Kim, J.-S. Kim, H.-C. Cho, “Deep-learning-based gestational sac detection in ultrasound images using modified YOLOv7-E6E model,” Journal of Animal Science and Technology, 2023, pp. 627-637. doi:10.5187/jast.2023.e43
[23] H.-X. Wu, X. Zhao, G.-H. Han, H.-J. Li, Y.-H. Kong, J.-H, Li, “MWSR-YLCA: Improved YOLOv7 Embedded with Attention Mechanism for Nasopharyngeal Carcinoma Detection from MR Images,” Electronics 2023, vol. 12, pp.1352
[24] X. Sun, H.-N, Qian, “Chinese Herbal Medicine Image Recognition and Retrieval by Convolutional Neural Network,” PLOS ONE, June. 2016, doi:10.1371/journal.pone.0156327
[25] S. Liu, W. Chen and X. Dong, "Automatic Classification of Chinese Herbal Based on Deep Learning Method," 2018 14th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), IEEE, 2018, pp. 235-238
[26] W. Hao, M. Han, S. Li, and F.-Z. Li, “A Novel Chinese Herbal Medicine Classification Approach with Mutual Triplet Attention Learning,” Hindawi, vol 2022, pp. 8034435. doi:10.1155/2022/8034435
[27] L. Dehan, W. Jia, C. Yimin and G. Hamid, "Classification of Chinese Herbal medicines based on SVM," 2014 International Conference on Information Science, Electronics and Electrical Engineering, IEEE, 2014, pp. 453-456.
[28] F. Huang, L. Yu, T. Shen and L. Jin, "Chinese Herbal Medicine Leaves Classification Based on Improved AlexNet Convolutional Neural Network," 2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), IEEE, 2019, pp. 1006-1011
[29] C. Xing, Y. Huo, X. Huang, C. Lu, Y. Liang and A. Wang, "Research on Image Recognition Technology of Traditional Chinese Medicine Based on Deep Transfer Learning," 2020 International Conference on Artificial Intelligence and Electromechanical Automation (AIEA), IEEE, 2020, pp. 140-146,
[30] D. Yue, Y. Hou, C. Hu, C. Zang and Y. Kou, "Chinese Herbal Medicine Recognition Using a VCSEL-Based Time-Delay Reservoir Computing System," in IEEE Photonics Journal, vol. 15, no. 3, pp. 1-8
[31] Y.-C. Lin, Z.-G. Shao, N. Hoffmann, “Global Attention Mechanism: Retain Information to Enhance Channel-Spatial Interactions,” 2021, arXiv:2112.05561
[32] J.-Q. Wang, K. Chen, R. Xu, Z.-W. Liu, C.C. Loy, D.-H. Lin, “CARAFE: Content-Aware ReAssembly of Features,” IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3007-3016
[33] R. Liu, J. Lehman, P. Molino, F. P. Such, E. Frank, A. Sergeev, J. Yosinski, “An intriguing failing of convolutional neural networks and the CoordConv solution,” Advances in neural information processing systems, vol. 31, 2018, pp: 31
[34] Z. Zheng, P. Wang, W. Liu, J. Li. R. Ye, and D. Ren, “Distance-loU loss: Faster and better learning for bounding box regression,’’ in Proc. AAAI Conf. Artif. Intell., New York, NY, USA, vol. 34. Feb. 2020pp.12993-13000.
[35] Z. Gevorgyan, “SIoU loss: More powerful learning for bounding box regression,” 2022, arXiv:2205.12740

KAIHUA CHE was born in 2002. He is currently studying in Zhuhai College of Science and Technology , majoring in Data Science and Big Data Technology. He was an intern in the product development department of Guangzhou HEXIN Technologies CO. His research interests include computer vision and big data.

was born in 2003. He is Zhuhai College of Technology , majoring Big Data Technology. include machine

YUHNEG LIANG currently studying in Science and in Data Science and His research interests learning and big data.

PENGJIE WANG was born in 2001. He is currently studying in Zhuhai College of Science and Technology , majoring in Data Science and Big Data Technology. His research interests include machine learning and computer vision.

TONGFEI LI received the MS. degree from the Institute of Data Science, City University of Macau, Macau, China, in 2021. He is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering, Faculty of Innovation Engineering, Macau University of Science and Technology, Macau. He has lectured with the Huike Group and has additional teaching experience in many universities, mainly on subjects.such as hadoop, machine learning, deep learningand computer vision.

WEI LV received the B.S., M.S., and Ph.D. degrees from the Mathematics Department, Software Research Institute, Sun Yat-sen University.in 2009. He has been a Visiting Scholar with Princeton University, Nanyang Technological University, The City University of New York, and RWTH Aachen University. He is currently the Dean of the School of Aliyun Big Data Applications, Zhuhai College of Science and Technology, and a Visiting Professor with the Institute of Data Science, City University of Macau. His research interests include big data and cloud computing.

XIAOLIN ZHU received the PH.D. degree from the Institute of Data Science, City University of Macau, Macau, China. He has recently been in charge of the Young Innovative Talents of Guangdong Ordinary Colleges and Universities (Natural Sciences) project.