scientific reports

Check for updates

OPEN

Deep learning method for detecting fluorescence spots in cancer diagnostics via fluorescence in situ hybridization

Zini Jian✉, Tianxiang Song, Zhihui Zhang, Zhao Ai, Heng Zhao, Man Tang & Kan Liu✉

Fluorescence in Situ Hybridization (FISH) is a technique for macromolecule identification that utilizes the complementarity of DNA or DNA/RNA double strands. Probes, crafted from selected DNA strands tagged with fluorophore-coupled nucleotides, hybridize to complementary sequences within the cells and tissues under examination. These are subsequently visualized through fluorescence microscopy or imaging systems. However, the vast number of cells and disorganized nucleic acid sequences in FISH images present significant challenges. The manual processing and analysis of these images are not only time-consuming but also prone to human error due to visual fatigue. To overcome these challenges, we propose the integration of medical imaging with deep learning to develop an automated detection system for FISH images. This system features an algorithm capable of quickly detecting fluorescent spots and capturing their coordinates, which is crucial for evaluating cellular characteristics in cancer diagnosis. Traditional models struggle with the small size, low resolution, and noise prevalent in fluorescent points, leading to significant performance declines. This paper offers a detailed examination of these issues, providing insights into why traditional models falter. Comparative tests between the YOLO series models and our proposed method affirm the superior accuracy of our approach in identifying fluorescent dots in FISH images.

Keywords FISH images, Cancer diagnostics, Fluorescence spots, Object detection, YOLO

Fluorescence in situ hybridization (FISH) identifies and maps specific DNA sequences or RNA molecules within cells or tissues through staining techniques[1,2]. This method aids in elucidating cytogenetic variations, gene rearrangements, chromosomal abnormalities, and localization. FISH is utilized extensively in medical diagnostics, encompassing genetic disease screening, tumor identification, and embryonic genome assessment. It holds significant potential for the early detection, prognosis, and management of cancers, including leukemia, breast cancer, and gastric cancer[3–5]. By quantifying and pinpointing specific genes, FISH provides vital data for diagnosing and treating diseases. Recently, the convergence of medical imaging and computer science has deepened. Techniques for detecting medical images through deep learning have gained traction[6–11]. However, the vast and complex nature of medical imaging data, coupled with a shortage of labeled data, complicates the training of deep learning models. The sensitivity of patient information in FISH images further hinders data set collection.

Object detection in medical images involves the localization and classification of lesions and other entities. Prominent algorithms encompass R-CNN, Fast R-CNN, Faster R-CNN, PFN, PSPNet, SSD, YOLO, CenterNet, and EfficientNet[12–19]. This process unfolds in two primary phases: (1) target feature extraction and (2) classification and localization of objects. Feature extraction utilizes CNNs. Object detection frameworks are categorized into two types: two-stage and single-stage. Two-stage frameworks initially engage in preprocessing to generate a detection scheme, followed by the detection process. They first extract CNN features from image regions devoid of category information, then classify these using category-specific classifiers. In contrast, single-stage frameworks, which include SSD, YOLO, CenterNet, and EfficientNet families, utilize a priori frame techniques to generate initial prediction frames, subsequently refining these through parameter adjustments to finalize the prediction. While two-stage frameworks perform excellently in object detection tasks, particularly in complex scenarios where they can provide high accuracy, they also have some significant drawbacks, especially regarding their high computational demands. Experimental environments require an efficient method to

School of Electronic and Electrical Engineering, Wuhan Textile University, Wuhan 430200, China. ✉email: znjian@wtu.edu.cn; liukan2002@gmail.com

process images to ensure timely and accurate results in time-sensitive situations. In this context, the YOLO architecture, designed for real-time detection, is more suitable. In FISH, the target DNA or RNA sequences are marked by their small size, large quantity, and high density. These characteristics make them small targets, which are notoriously difficult to detect. While many models excel with medium to large objects, they often underperform on datasets comprised of such small targets. Challenges in detecting small objects arise primarily from two factors. Firstly, small objects often lack sufficient appearance information to differentiate them from backgrounds or similar entities. Additionally, their positional data is uncertain, necessitating higher precision for accurate localization. Furthermore, the field has accumulated limited expertise and knowledge concerning small object detection, with predominant research focusing on large or medium-sized targets. To enhance the detection of FISH fluorescent dot images, Xu et al.[9] introduced a lightweight deep learning model utilizing a rotated Gaussian kernel, achieving an analysis time of approximately 0.31 s per frame—about 800 times faster than traditional pathologist methods. T. LES et al.[20] developed a method for localizing fluorescent dots in FISH images using 3D shape analysis, with results showing less than a 3% discrepancy compared to expert evaluations across several thousand cells. Further, Chao Xu et al.[21] implemented a multiscale MobileNetYOLO-V4 network, achieving high precision detection at a fast pace. Bouilhol E et al.[22] proposed the DeepSpot model, which incorporates dilated convolutions into a module specifically designed for small object context aggregation, and uses residual convolutions to propagate this information throughout the network. This allows DeepSpot to enhance all RNA spots to the same intensity, thus eliminating the need for parameter tuning. Nevertheless, the efficacy of traditional methods significantly declines with blurred and noisy images. Consequently, we have enhanced the YOLO series by developing YOLOv8, significantly improving model accuracy for detecting FISH fluorescent spots.

The principal contributions of this study are outlined below:

(1) The integration of the space-to-depth module into YOLOv8 has mitigated the loss of fine-grained information, enhancing both the learning efficiency of feature representation and the accuracy of YOLOv8 in small object detection. To enhance low-level feature extraction, the stride of the Conv module in YOLOv8 has been reduced to 1. This adjustment aids in the precise identification of image structures and minimizes downsampling, preserving the original spatial resolution of the input data. Reflecting the characteristics of the fluorescent point dataset, the large object detection head in YOLOv8 was replaced with a small object detection head, thereby boosting small object detection performance.
(2) A novel module named CE, which merges the C2f module from YOLOv8 with the efficient channel attention (ECA) module, has been introduced, yielding significant performance enhancements with only a minimal increase in C2f parameters.
(3) The original loss function has been substituted with L_{MPDIoU}, encompassing all pertinent factors of commonly utilized loss functions. This replacement addresses issues when the bounding box shares the aspect ratio of the true value bounding box, yet the width and height differ substantially, hindering effective optimization.

Materials and methods

Sample preparation

We conducted experiments on the patient's leukocytes using the AML1/ETO fusion gene detection kit. In the FISH experiment, whole blood preprocessing is performed first by adding a lysis buffer to induce hemolysis, followed by immersion in KCl solution and fixation using methanol-acetic acid. Next, phosphate-buffered saline (PBS) is used for sample loading and washing. Then, denaturation and hybridization are conducted by adding probe buffer, AML1/ETO probe, and mineral oil, completing the hybridization at specific temperatures and pressures. After hybridization, samples are washed sequentially with sodium citrate buffer (SSC), post-hybridization wash solution, and deionized (DI) water. Finally, DAPI is used for staining to facilitate the observation of cell nuclei. The ETO probe was labeled with an orange-red fluorophore, and the AML1 probe was labeled with a green fluorophore. The probes were hybridized to the target detection sites using in situ hybridization technology. Green light (with a wavelength in the range of 500–550 nm) is used to excite the ETO probe, while blue light (with a wavelength in the range of 450–490 nm) is used to excite the AML1 probe. Since the fluorescent spots formed by the two probes have different colors but nearly identical shapes, we obtain their grayscale images for model training.

Collection of FISH images

The inverted fluorescence microscope used was an OLYMPUS IX83, equipped with a fluorescence illumination system and widefield imaging modality. The objective lens was an OLYMPUS LUCPLFLN 60X with a numerical aperture (N.A) of 0.7 and a magnification of 60. The detector used for image acquisition was a QIMAGING optiMOS camera, which is a type of sCMOS. This combination of camera and objective lens is designed to balance imaging resolution, the requirements of the detection task, and the availability of equipment under the current experimental conditions. All FISH images and biological experiments in this study were conducted and provided by professionals from our laboratory. Each pixel corresponds to an actual distance of approximately 0.1 μm, and each FISH spot is about 5 pixels in size. Figure 1 shows representative images along with their scalebars.

We annotated 199 FISH fluorescent spot images, each with a size of 1920×1080, using Labelme software. The annotation process is illustrated in Fig. 2.

To address the limited number of images, we expanded the dataset to increase the diversity of training samples, reduce overfitting, and enhance the generalization capability of the network. Various augmentation techniques were employed, such as rotation, cropping, mirror symmetry, and the addition of Gaussian noise.

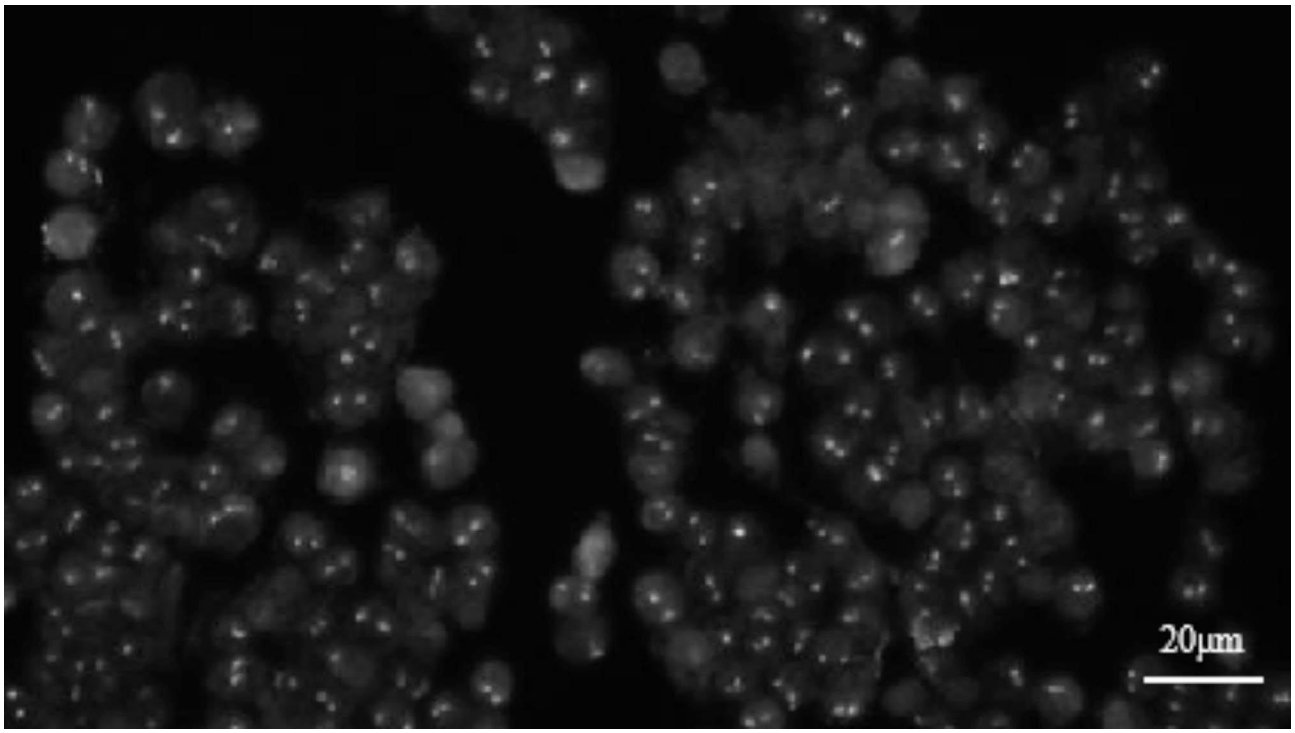

Fig. 1. Examples of data sets.

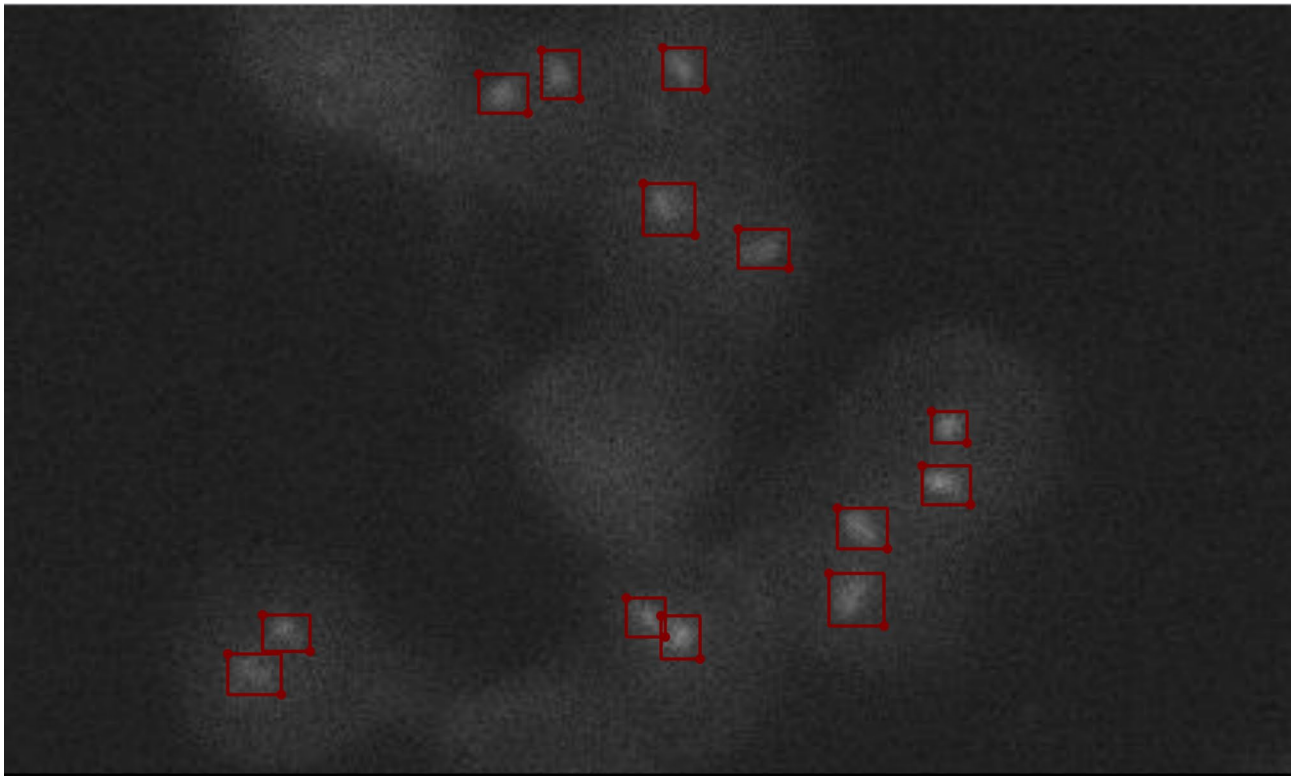

Fig. 2. Labeling process.

Consequently, we generated 995 images of fluorescent spots, each with a size of 1920 × 1080 pixels, containing 100 to 200 spots. The dataset was split into a training set and a validation set with a ratio of 9:1, comprising 895 and 100 images, respectively.

Proposed method for fluorescent spot detection

To this date, YOLOv10 represents the latest advancement in the YOLO series[23,24]. However, we have chosen to utilize YOLOv8 for our analysis. The rationale for this selection will be provided in the subsequent sections, where we will discuss its suitability for our specific application, including considerations related to processing efficiency and performance in the context of FISH data analysis. The official YOLOv8 code offers several networks; however, this article focuses primarily on YOLOv8, YOLOv8-P2, and YOLOv8-P6 for object detection tasks. YOLOv8-P2 and YOLOv8-P6 are distinguished by the inclusion of an additional small object detection head and a large object detection head, respectively, enhancing their detection capabilities. Specifically, the enhanced ability of YOLOv8-P2 to detect small objects, such as fluorescent spots, significantly improves accuracy in scenarios where detection of small objects is critical.

The YOLOv8 model comprises several modules, including C2f (Convolution to Fully Connected), Bottleneck, SPPF (Spatial Pyramid Pooling-Fast), Detect, and Conv. The C2f module utilizes cross-stage partial feature fusion to integrate low-level and high-level feature maps. This integration significantly increases the model detection precision and processing speed. The Bottleneck architecture reduces feature map channels, decreasing the computational burden. It incorporates residual connections to mitigate vanishing gradients and includes a 3 × 3 convolutional layer to expand the receptive field. SPPF, a crucial component, features Spatial Pyramid Pooling, enabling the model to process various object sizes within a single image by aggregating features from multiple receptive fields. The Detect module processes the outputs from the Neck module, which integrates inputs from the C2f, Bottleneck, SPPF, Detect, and Conv modules, as depicted in Fig. 3.

YOLO-SEM (YOLO-small object enhancement model) network framework

Inspired by the referenced models, we introduce YOLO-SEM, consisting of a backbone, neck, and head components. The backbone serves to extract relevant image information for use in subsequent network layers. This component enhances efficiency and performance, while simultaneously lowering the computational complexity involved in feature extraction. Positioned between the backbone and the head, the neck optimizes the use of extracted features and aids in feature fusion. The head utilizes these features to improve recognition capabilities.

The input image, featuring fluorescent dots, is segmented into an N × N grid structure by the network with the grid partitioning automatically performed using the k-means algorithm[25]. Although this grid-based approach may encounter some disadvantages when handling multiple adjacent spots, it also offers several key advantages. By processing the entire image in a single forward pass, this method allows the model to consider the global context, which helps reduce the effect of background noise and improves target localization. Understanding the global context is crucial for overall accurate detection. Additionally, this approach is very fast, enabling real-time or near-real-time analysis, which is particularly beneficial in high-throughput scenarios or when processing large datasets. Each cell within this grid undergoes examination to detect targets. Responsibility for detection is assigned to a cell if the target's center intersects with it. Subsequently, the network forecasts bounding boxes for each cell and allocates a confidence score to them. The computation of the confidence score is defined by the following formula (1):

$$Conf = P \times IoU_{truth}^{pred}, \quad P \in [0, 1] \tag{1}$$

The variable P is set to 1 when objects are present within the mesh; otherwise, it remains 0. The IoU quantifies the overlap between the predicted and actual bounding boxes. The confidence level measures the precision of a bounding box that contains an object and indicates the presence or absence of an object in the mesh. When multiple bounding boxes identify the same target, the YOLO network employs non-maximal suppression (NMS) to select the optimal box. This technique is essential for eliminating redundant detection frames, ensuring only the most representative frame is retained. Detection frames are organized in descending order by their confidence levels. Starting with the highest confidence frame, the NMS algorithm calculates its IoU against other frames. Frames exceeding a predefined IoU threshold with the baseline frame are discarded. This iterative selection process continues, using one highest confidence frame as the reference each time, until all frames are evaluated. NMS guarantees that only one representative frame per target appears in the final results.

In object detection tasks, detection heads are classified by the size of their corresponding feature maps. Larger detection heads are associated with feature maps of lower resolution, while smaller detection heads are linked to feature maps of higher resolution. To improve the management of small objects in object detection tasks, we have chosen to replace the large object detection head with a smaller one specifically designed for small objects, as shown in Fig. 4. This modification allows the model to concentrate more effectively on small objects, thereby improving their detection rates and contributing to targeted optimization. In the YOLO-SEM model, four feature map sizes are organized in descending order with varying resolutions, labeled as P1, P2, P3, and P4. Employing multi-scale feature maps and different sizes of detection heads enables YOLO-SEM to comprehensively detect a wide range of objects, thus enhancing overall detection performance.

Our experiments have demonstrated that the P2 and P3 detection heads significantly influence the accuracy of fluorescent point detection. To mitigate risks of model overfitting, diminished generalization capability, increased computational demands, and longer training and inference times associated with an excess of attention mechanisms, we have implemented the ECA mechanism exclusively on feature maps of corresponding sizes in P2 and P3. This strategy ensures precise control over the application levels of the attention mechanism, balancing

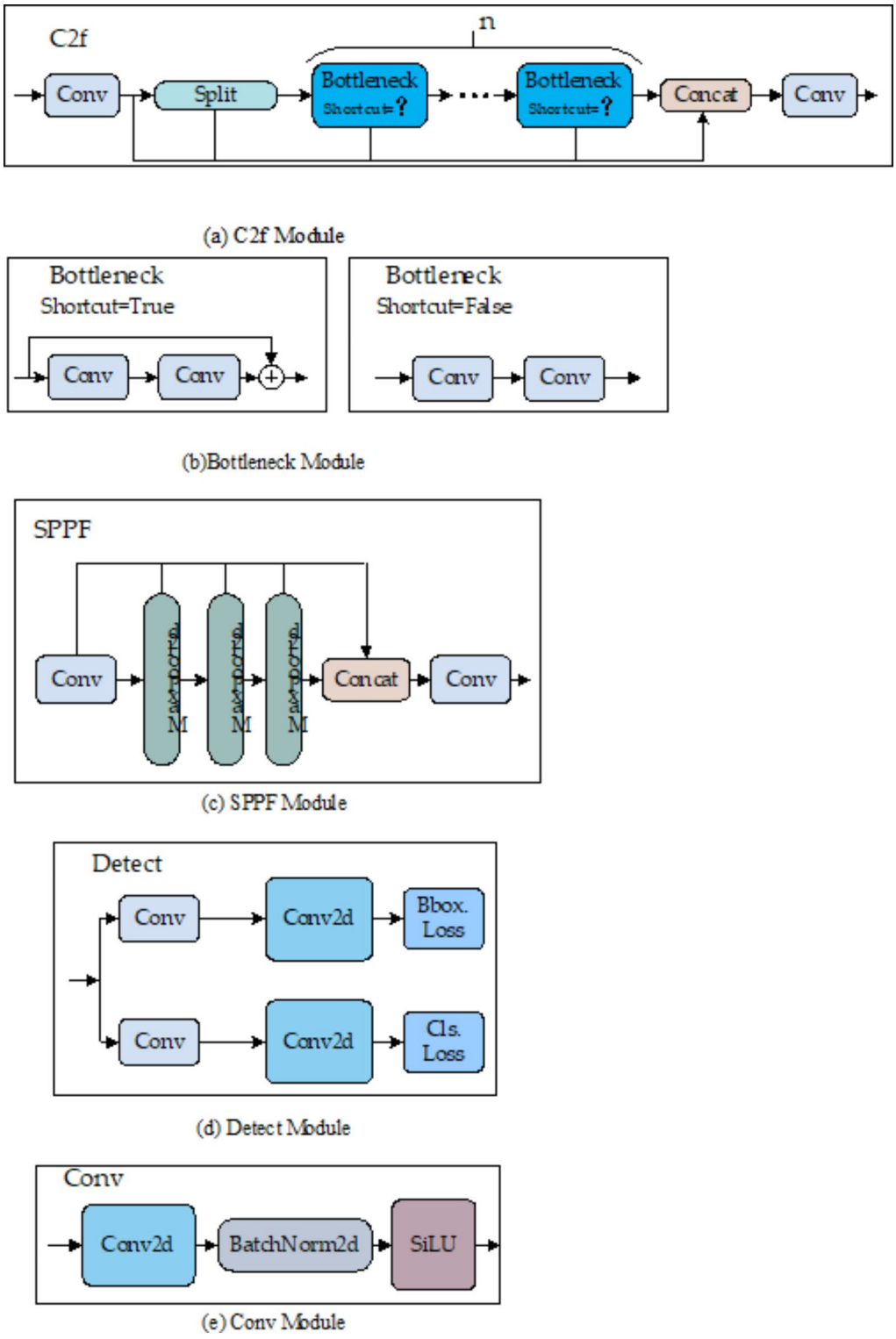

Fig. 3. YOLO-SEM network module structure diagram.

the model performance and complexity. Furthermore, this method enhances the model robustness, making it more viable for real-world applications. The model structure is depicted in Fig. 5.

SPD module

The YOLO series of architectures excels in various computer vision tasks, including object detection and image classification[26–29]. However, it exhibits a notable decline in performance when processing low-resolution images or detecting small objects[30,31]. This decrease can be attributed to the use of stride convolution and pooling layers,

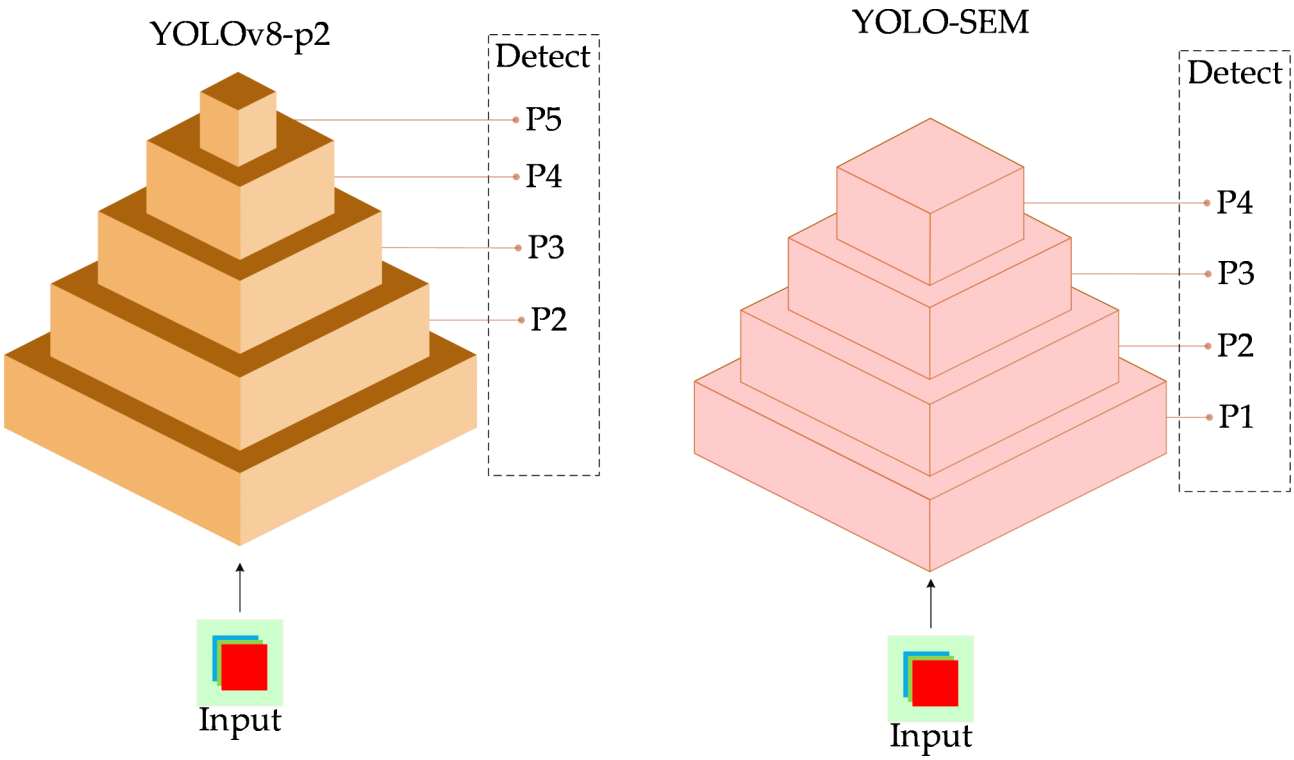

Fig. 4. Comparison of YOLO-SEM and YOLOv8-P2 models.

prevalent in YOLO architectures, which lead to reduced feature representation and a loss of precision. Typically, these adverse effects are mitigated because the images analyzed generally possess high resolution and contain large objects, allowing the model to skip redundant pixel information while still effectively learning features. However, this assumption of redundancy fails in the case of the FISH image dataset, where images are blurry, and objects are small, resulting in the loss of fine-grained information, insufficient feature learning, and a marked reduction in the model's detection capability.

To resolve this challenge, we integrated the SPD module[32] into YOLO-SEM. This module employs a spatial-depth layer to downsample feature maps, ensuring the retention of information across channel dimensions and minimizing data loss. Furthermore, the Conv module in YOLO-SEM employs a step size of 1, which eliminates additional downsampling and reduces the total downsampling instances within the model, thereby maintaining the spatial resolution of the input data. Figure 6 depicts the SPD module.

Assuming that each feature map F is of size W × H × C, the feature map is sliced into sub-feature sequences $F_{m,\,n,}$ As shown in Eq. (2):

$$F_{m,n} = \mathrm{F}[m : H : Step, n : H; Step] \tag{2}$$

Next, these sub-feature sequences are concatenated along the channel dimensions to obtain a feature mapF' of size $\frac{W}{step} \times \frac{H}{step} \times (C \times step^2)$, and this new feature map is then passed to the next layer.

CE module

The attention mechanism has gained significant attention in recent years due to its outstanding performance across various applications[33–36]. While current approaches often aim to improve overall model effectiveness by designing increasingly complex attention modules, this usually results in higher model complexity. In contrast, the ECA module[37] enhances model performance with only a slight increase in computational cost and facilitates cross-channel interactions in learning channel attention without reducing channel dimensionality, as depicted in Fig. 7.

Global average pooling (GAP) is used to obtained aggregated features map Box1, which is then extracted as a single real value to obtain feature $\chi_{avg} \in R^{(W \times H \times C)}$, as shown in Eq. (3):

$$\begin{cases} \chi_{avg} = GAP(\chi) \\ GAP(\chi) = \frac{1}{W \times H} \sum_{i=1,j=1}^{W,H} \chi_{i,j} \end{cases} \tag{3}$$

ECA generates channel weights through a rapid one-dimensional convolution of size k, which is adaptively determined by mapping the channel dimension, C. The extent of interaction, indicated by k, the size of the convolution kernel, correlates directly with the channel dimension through a mapping ϕ, as shown in Eq. (4):

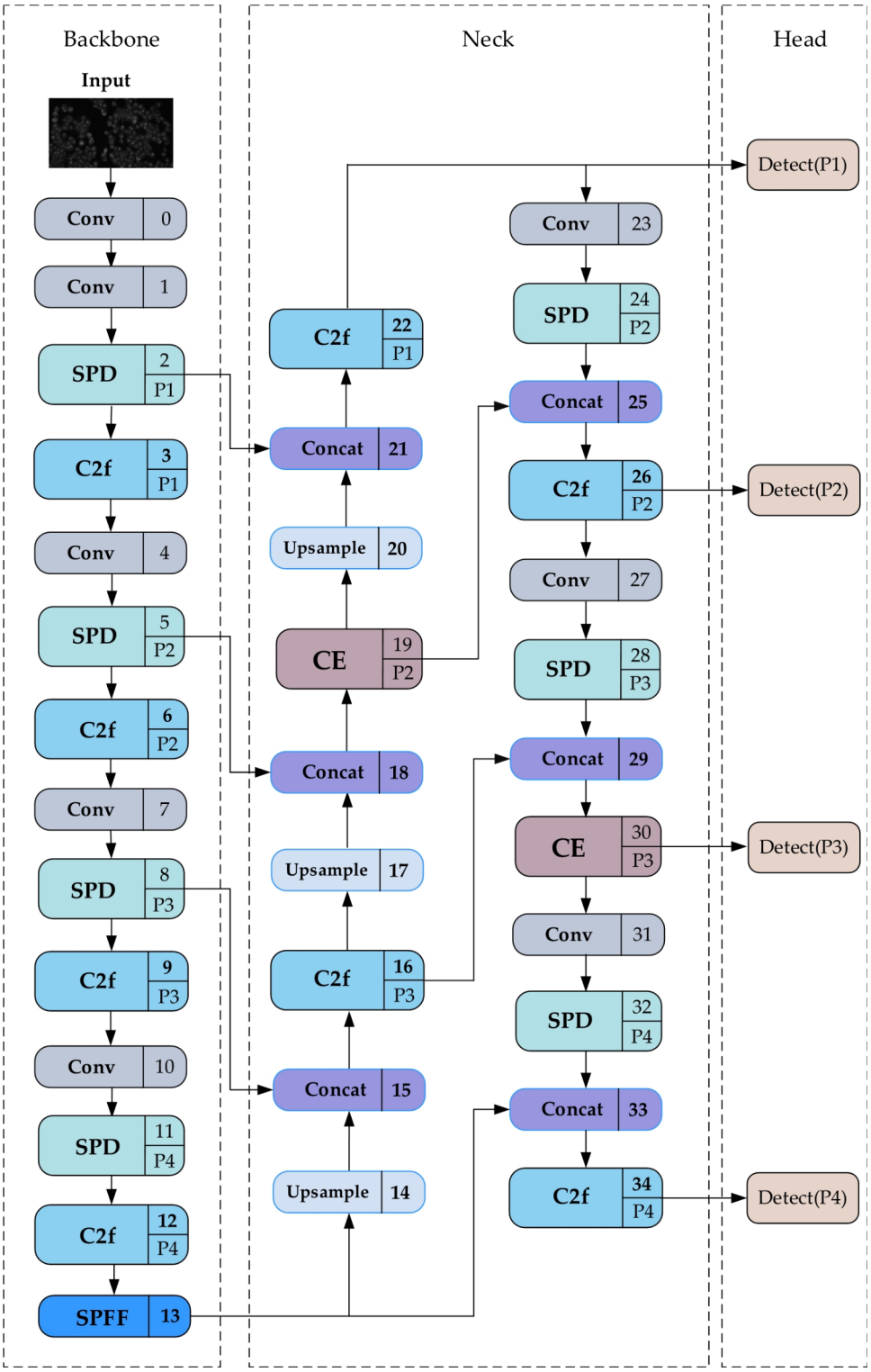

Fig. 5. YOLO-SEM model.

$$C = \phi(k) \tag{4}$$

If the mapping is represented by the linear function $\phi(k) = \lambda \times k - b$,its capacity to depict feature relationships is notably restricted. Consequently, this linear function is expanded to a nonlinear function to enhance its representational capability. Typically, the channel dimension C (number of filters) is set at 2. The relationship between these parameters is defined in Eq. (5):

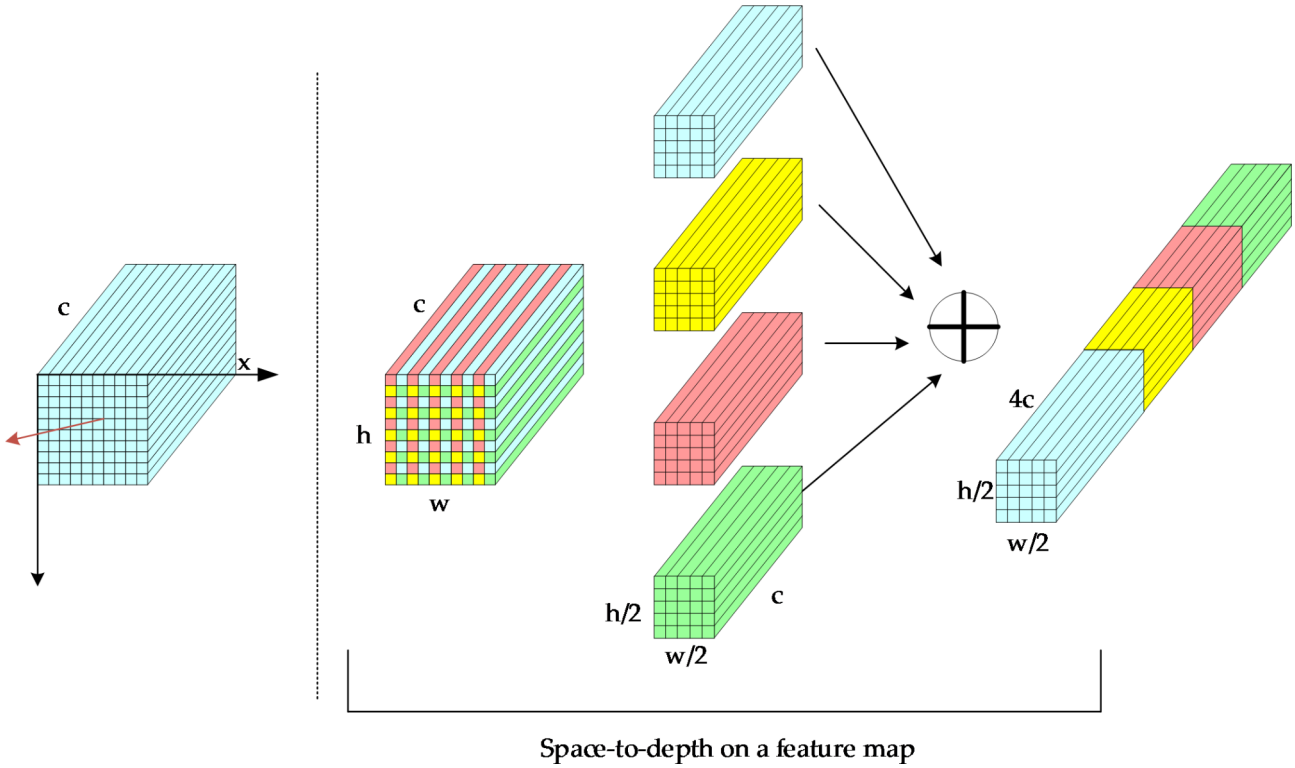

Fig. 6. Space-to-depth when step = 2.

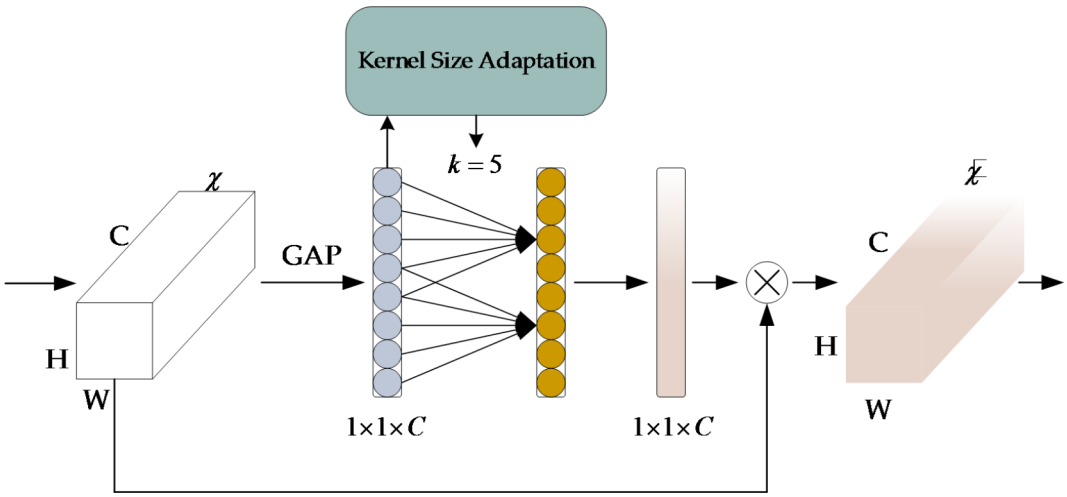

Fig. 7. ECA module.

$$C = \phi(k) = 2^{\gamma \times K - b} \tag{5}$$

The adaptive method for the kernel size K can then be determined as follows:

$$K = \psi(C) = \left| \frac{\log_2(C)}{\gamma} + \frac{b}{\gamma} \right|_{\text{odd}} \tag{6}$$

Where $|t|_{odd}$ represents the nearest odd integer to t. Activation values for one-dimensional convolutional outputs are derived using the Sigmoid function. The ECA algorithm preserves channel dimensions and assesses inter-channel relationships to address noise-induced disturbances, thereby enhancing the model's noise reduction capabilities.

To enhance the inter-channel correlation, the ECA module has been integrated into the C2f module. The C2f module, comprising multiple convolutional layers and complex feature transformations, facilitates this enhancement by leveraging the ECA module. This integration fosters richer and more discriminative feature representations, thereby improving the performance and generalization capabilities of complex networks, ultimately resulting in higher accuracy.

The Bottleneck, a critical component of C2f, primarily operates on local feature maps. Enhanced by the incorporation of ECA, the Bottleneck selectively augments task-specific channels, thereby improving the capture of essential object features. The configuration of Bottleneck, denoted by 'n', facilitates the adjustment of ECA implementations, depending on the task requirements and available computational resources. The Bottleneck structure comprises two convolutional modules tasked with feature extraction and transformation. Positioned between these convolutional layers, the ECA module optimizes channel attention weighting on the outputs from the initial convolution. This arrangement allows the weighted features to undergo further transformation in the subsequent convolutional layer, thus enabling the attention mechanism of the ECA module to influence the entire feature map comprehensively. Figure 8 illustrates the configuration of the Bottleneck with the ECA module.

L_{MPDIoU} loss function

Bounding box regression is extensively used in object detection and instance segmentation, serving as a critical step for target localization[38–40]. The original IoU is calculated as the ratio of the intersection area between the predicted bounding box and the ground truth bounding box to their combined union area, as delineated in Eq. (7):

$$IoU = \frac{B_{gt} \cap B_{prd}}{B_{gt} \cup B_{prd}} \tag{7}$$

The distance d_c between the center coordinates of the predicted box and the ground truth box can be expressed using Eq. (8).

$$d_c^2 = (x_c^{prd} - x_c^{gt})^2 \times (y_c^{prd} - y_c^{gt})^2 \tag{8}$$

Where (x_c^{prd}, y_c^{prd}) and (x_c^{gt}, y_c^{gt}) represent the center coordinates of the predicted box and the ground truth box, respectively.

Let w_c and h_c be the width and height of the minimum enclosing box, then its area c can be expressed by Eq. (9).

$$c^2 = w_c^2 + h_c^2 \tag{9}$$

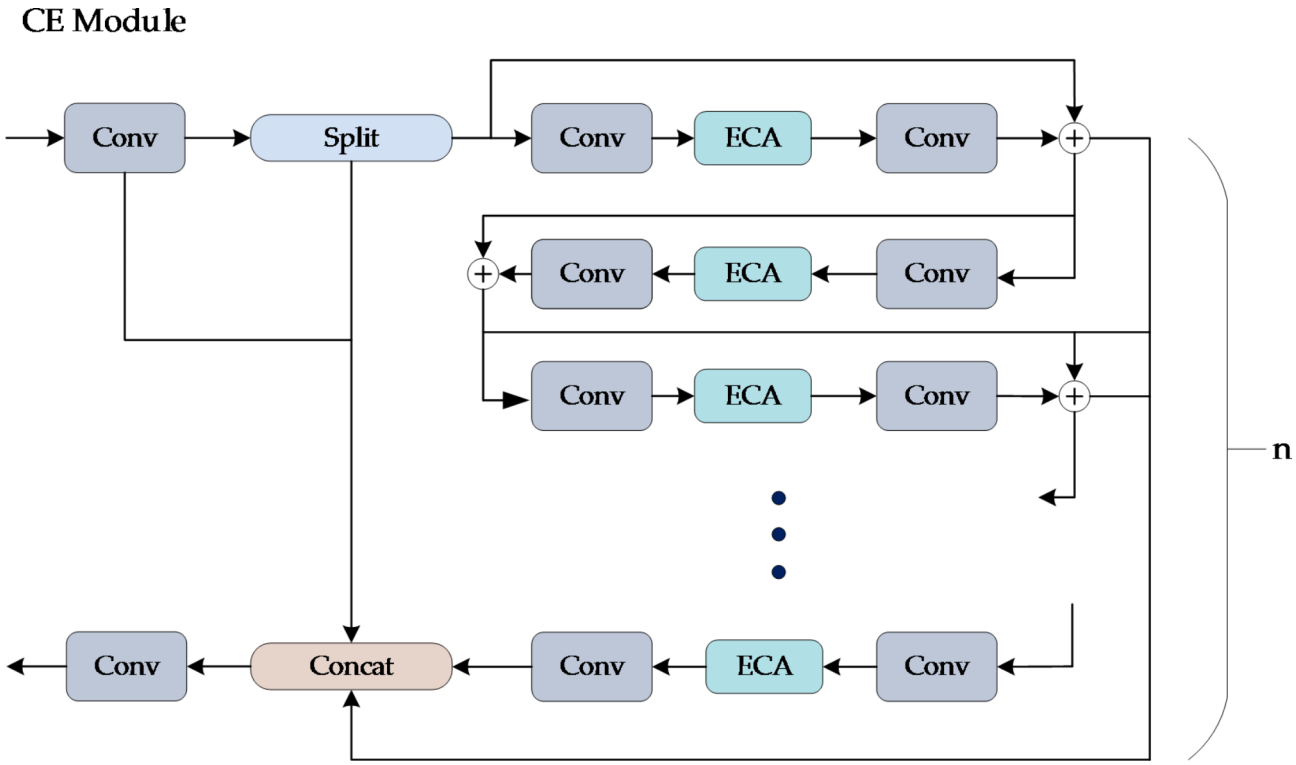

Fig. 8. CE module.

V is the aspect ratio consistency measure between the predicted box and the ground truth box, defined by Eq. (10).

$$V = \frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w^{prd}}{h^{prd}}\right)^2 \tag{10}$$

w^{gt} and h^{gt} are the width and height of the ground truth box, while w^{prd} and h^{prd} are the width and height of the predicted box. By calculating the arctan difference of the aspect ratios, *V* reflects the aspect ratio discrepancy between the predicted box and the ground truth box.

The initial bounding box regression loss function in YOLOv8-p2 is CIoU which is also used in most YOLO models. It can be described by Eq. (11).

$$CIoU = IoU - \frac{d_c^2}{c^2} - \alpha \cdot V \tag{11}$$

Where α is a weighting factor used to balance the importance between IoU and the center point distance, typically ranging between 0 and 1. However, CIoU face challenges in optimizing effectively when the predicted bounding box maintains the aspect ratio of the true bounding box but differs in size. To address this issue, a novel bounding box similarity metric, Minimum Point Distance-based Intersection over Union (MPDIoU), has been introduced. This metric leverages the geometric characteristics of horizontal rectangles and incorporates all pertinent factors considered in existing loss functions, such as overlapping or non-overlapping areas, center distance, and width and height deviations. Moreover, it simplifies the computational process. The loss function, L_{MPDIoU}, is derived from MPDIoU.

Assuming that $B_{prd} = (x_1^{prd}, y_1^{prd}, x_2^{prd}, y_2^{prd})andB_{gt} = (x_1^{gt}, y_1^{gt}, x_2^{gt}, y_2^{gt})$,where the coordinates of the top-left (x_1^{prd}, y_1^{prd}), (x_1^{gt}, y_1^{gt}) and bottom-right (x_2^{prd}, y_2^{prd}), (x_2^{gt}, y_2^{gt}) points are given, the width and height of the image are denoted as w and h respectively, as shown in Eq. (12):

$$\begin{aligned} d_1^2 &= (x_1^{prd} - x_1^{gt})^2 \times (y_1^{prd} - y_1^{gt})^2 \\ d_2^2 &= (x_2^{prd} - x_2^{gt})^2 \times (y_2^{prd} - y_2^{gt})^2 \end{aligned} \tag{12}$$

Then MPDIoU for Eq. (13):

$$MPDIoU = IoU - \frac{d_1^2}{h^2 + w^2} - \frac{d_2^2}{h^2 + w^2} \tag{13}$$

The L_{MPDIoU} for Eq. (14):

$$L_{MPDIoU} = 1 - MPDIoU \tag{14}$$

MPDIoU is suitable for multiple object detection contexts because it evaluates not only the overlap between two bounding boxes but also their spatial relationships. This enhances the accuracy in depicting the relative positions of objects within an image, especially when objects are occluded or partially visible.

Results

Evaluation indexes

Key metrics for assessing the effectiveness of the neural network model include precision, recall, mAP50, mAP50-95, and IoU. Binary classification categorizes samples into four groups: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN), based on the alignment of actual and predicted categories. Table 1 presents the confusion matrix of these classification outcomes.

Precision and Recall are defined as Eqs. (15) and (16):

$$Precision = \frac{TP}{TP + FP} \tag{15}$$

$$Recall = \frac{TP}{TP + FN} \tag{16}$$

Label	Predict	Confusion matrix
Positive	Positive	TP
Positive	Negative	FN
Negative	Positive	FP
Negative	Negative	TP

Table 1. Confusion matrix of classification results.

Name	Parameter
CPU	Intel Xeon Gold 5318Y
GPU	NVIDIA A100
Programming Language	Python 3.8.17
Deep learning framework	Pytorch 1.8.0

Table 2. Experimental configuration.

Model	mAP50 (%)	mAP50-95 (%)	Precision (%)	Recall (%)	Time (ms)
LOG	22.49	14.23	44.03	51.08	0.3
YOLOv3	37.06	13.27	40.42	34.98	1.3
YOLOv5	36.20	12.61	39.22	35.26	1.4
YOLOv8	36.80	12.95	39.88	35.68	1.7
YOLOv9	39.01	14.62	45.25	36.39	1.6
YOLOv10	39.59	15.27	47.26	35.99	0.9
Chao Xu`s	37.66	13.73	44.45	35.43	1.4
YOLOv8-p2	59.29	23.17	65.57	55.48	2.2
YOLO-SEM	70.21	29.40	74.97	61.84	1.5

Table 3. Comparison of detection result data.

For each category, average accuracy was determined by calculating the area under the Precision-Recall curve with the IoU threshold set at or above 0.5, as detailed in Eqs. (17) and (18).

$$AP_{50} = \sum_{k=1}^{n} (R_k - R_{k-1}) P_k \tag{17}$$

$$mAP_{50} = \frac{1}{N} \sum_{i=1}^{n} AP_{50i} \tag{18}$$

Where n is the total number of Recalls, R_k is the k_{th} value of the Recall at the level and P_k is the Precision value at the corresponding Recall. Averaging the AP50 across all categories gives mAP50. And mAP50-95 is the calculation of Precision and Recall at different confidence thresholds for each category, over a range of IoU thresholds from 0.5 to 0.95, as shown in Eqs. (19) and (20).

$$AP_{50-95} = \frac{1}{11} \sum_{i=0}^{n} AP_{(t+5)/10} \tag{19}$$

$$mAP_{50-95} = \frac{1}{N} \sum_{i=1}^{N} AP_{50-95i} \tag{20}$$

Where ap is the $AP_{(t+5)/10}$ at an IoU threshold of $\frac{t+5}{10}$. The mAP50-95 metric is derived by evaluating the model across a spectrum of IoU thresholds from 0.5 to 0.95, segmented into eleven equal intervals. This approach provides a thorough assessment of the model performance in object detection at various IoU thresholds, thereby furnishing a more detailed insight.

Experiments

The optimization of hyperparameters for YOLO-SEM was conducted using stochastic gradient descent (SGD) with an initial learning rate set at 0.01, momentum at 0.937, and a weight decay of 0.0005. The training was divided into two distinct phases. In the initial phase, YOLOv8-p2 underwent training over 300 epochs with a batch size of 2. Subsequently, the second phase applied the outcomes from the initial phase to facilitate transfer learning for YOLO-SEM, maintaining the same epoch count and batch size. The experimental setup is detailed in Table 2.

To assess the performance of our proposed model, we conducted a comparative analysis of YOLO-SEM against LOG (Laplacian of Gaussian filter), YOLOv8-P2, Multi-scale MobileNet-YOLO-V4 proposed by Chao Xu[21] and several conventional YOLO series models. These models were trained using the FISH fluorescent point dataset across 300 iterations. The peak results are documented in Table 3, and the training data trends are depicted in Fig. 9 (Since the LOG is not a deep learning method, there is no training process data available).

Based on the above results, the YOLO-SEM model employed in this study demonstrates the best accuracy in detecting fluorescent spots among all the compared models. First, its mAP50 score of 70.21% indicates that the

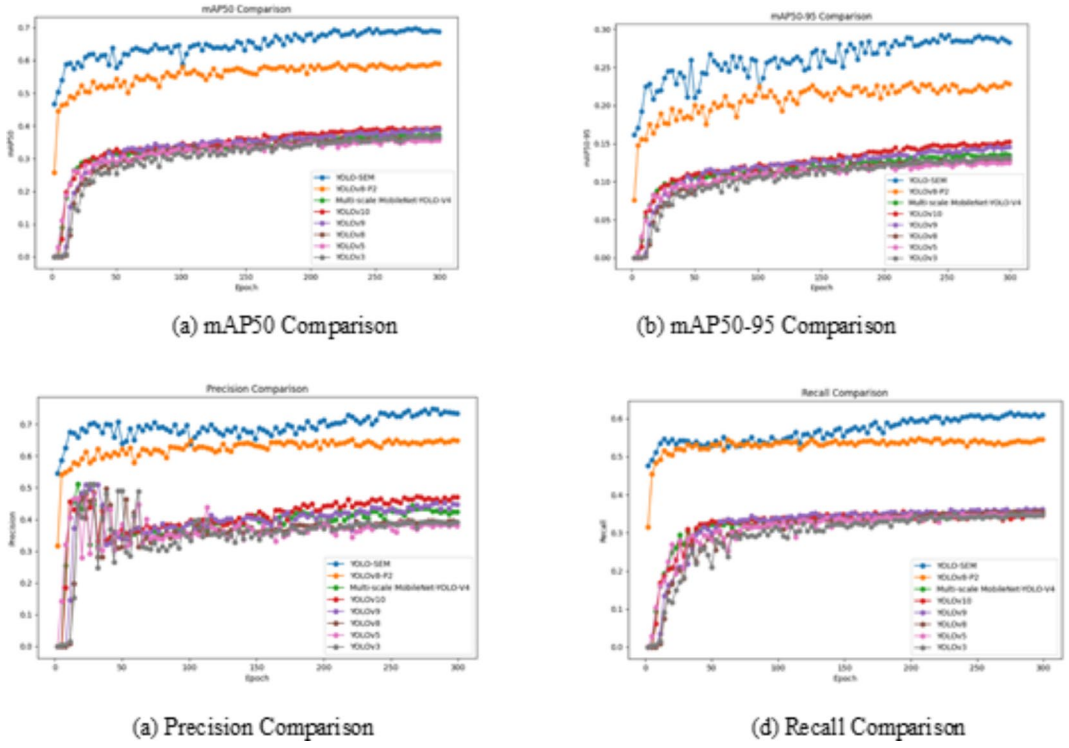

Fig. 9. Comparison charts of the models on different metrics.

model can accurately locate most targets under noisy conditions and when the targets are small, even at an IoU threshold of 0.5. Additionally, with an mAP50-95 score of 29.40%, the model demonstrates strong performance across various IoU thresholds (from 0.5 to 0.95), showcasing its robustness in handling complex scenarios. In terms of precision, YOLO-SEM achieves 74.97%, meaning that the majority of detected targets are true positives, highlighting the model's effectiveness in reducing false alarms. A recall rate of 61.84% further emphasizes YOLO-SEM's reliability in detecting most of the actual targets present. The effectiveness of fluorescent point detection is illustrated in Fig. 10.

However, due to the small size of fluorescent spots, low image resolution, and the presence of noise, YOLO models not optimized for small object detection exhibit a noticeable drop in performance when faced with these challenges. Although YOLOv8-P2 introduced a small object detection head and made some improvements in detecting smaller objects, it still suffers from a significant number of missed detections, failing to meet the high-precision demands of these tasks. While the Laplacian of Gaussian (LOG) filter is fast, its poor detection performance and need for manual adjustments limit its practicality. In contrast, the YOLO-SEM model has been specifically optimized to address these challenges, proving more effective in handling small fluorescent spots, low-resolution images, and noise interference. It not only excels in detecting tiny fluorescent spots but also maintains high accuracy and recall rates even under these difficult conditions, demonstrating its adaptability and reliability. The model's inference time is only 1.5 milliseconds, providing rapid processing speed while maintaining high detection performance. This makes it suitable for real-time or near-real-time detection tasks and capable of meeting the requirements of clinical experiments.

To ascertain the impact of individual modules on model performance, each module was integrated sequentially with YOLOv8-P2. The outcomes of these integrations are presented in Table 4. The results indicate that each module contributed to the enhancement of the network detection capabilities to varying degrees.

Discussion

The global cancer incidence has remained alarmingly high in recent years, leading to millions of fatalities annually[41]. Cancer diagnosis currently relies heavily on diagnostic imaging and pathological assessments[42,43]. Early detection is critical for improving patient survival rates, making non-invasive, efficient screening methods a focal point of research. Deep learning, a branch of machine learning, uses simulated neural networks to extract features from data. Applying deep learning to medical imaging can aid in lesion localization, facilitate diagnoses, reduce physician workload, minimize errors, and improve the accuracy and reliability of prognosis and diagnostic outcomes. Demonstrating robust capabilities, deep learning models excel in tasks such as medical image classification, segmentation, lesion detection, and alignment[44–46]. These models process various medical imaging modalities, including X-rays, MRIs, CT scans, and cytofluorograms, aiding in the diagnosis of conditions such as breast and blood cancers. This paper proposes a novel automatic disease detection approach using the discussed model.

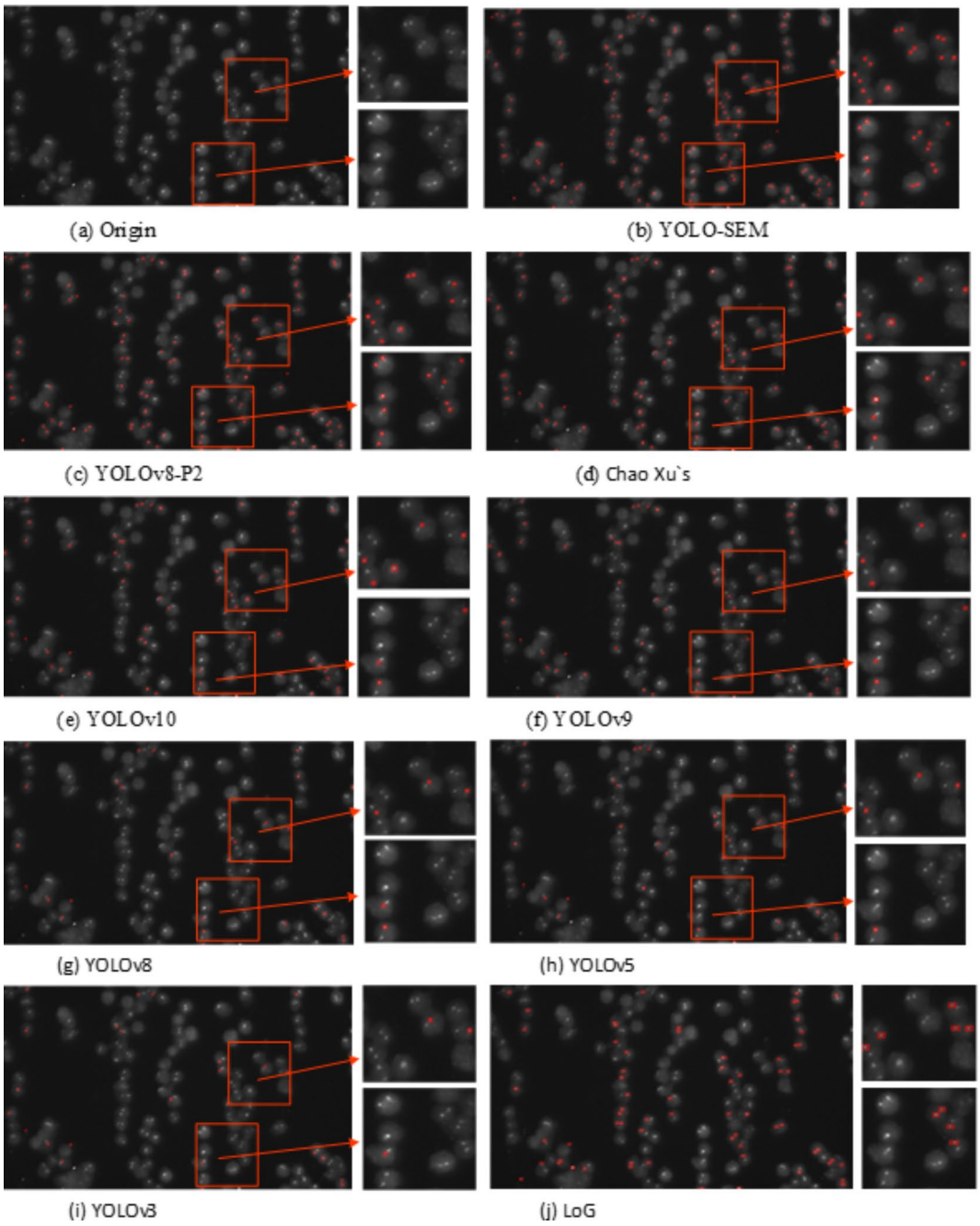

Fig. 10. Comparison of detection results.

Model	SPD	MPDIoU	CE	mAP50 (%)	mAP (%)	Precision (%)	Recall (%)
YOLOv8-P2	×	×	×	59.29	23.17	65.57	55.48
YOLO-SEM	√	×	×	62.53	25.07	68.39	56.29
YOLO-SEM	√	√	×	66.96	27.53	72.30	58.48
YOLO-SEM	√	√	√	70.21	29.40	74.97	61.84

Table 4. Ablation experiment.

FISH staining generally involves using DAPI dye to stain cell nuclei and base-pairing dyes for RNA/DNA (such as the AML1/ETO fusion gene probes used in this study's dataset). FISH result analysis involves two main steps. First, obtain images of cell outlines stained with DAPI. Then, image segmentation algorithms such as Unet[47] and Unet++[48], which have been shown to be effective[49], were utilized to segment individual cell outlines. Subsequently, locate the fluorescent probes stained with the fusion gene dye within the nucleus of each cell. The evaluation of each cell is based on the distribution of probes within it. However, in practical applications, the image localization of fluorescent probes often faces challenges such as noise and low resolution. The YOLO-SEM algorithm, introduced in this study, is utilized for detection, addressing challenges such as the small size of fluorescent spots and image resolution that typically hinder traditional models.

Conclusion

In this study, we introduce the YOLO-SEM network, which augments the widely adopted and robust YOLOv8 framework. YOLO-SEM, an object detection model, is specifically designed for detecting small objects at low resolutions and incorporates several innovative modules and techniques. The model addresses the issue of detail loss in downsampled feature maps with the introduction of the SPD module, a space-to-depth layer that preserves important information and fine structural details of small objects. Furthermore, YOLO-SEM improves sensitivity to small objects by replacing the large object detection head with a smaller, specialized detection head. The model feature representation is enhanced through the integration of C2f and the ECA attention mechanism, which processes channel information and captures essential object features more effectively. Additionally, the MPDIoU loss function refines model parameters by incorporating object location data, thereby increasing the accuracy of localization for small objects. Comparative experiments have shown that YOLO-SEM outperforms other models, including YOLOv8 and YOLOv5, particularly in detecting fluorescent points in FISH images. However, despite its superior performance, YOLO-SEM still exhibits limitations in the precise detection of fluorescent spots.

Future work will focus on broadening the application of current models in medical image processing beyond FISH images. This expansion will include assessing and improving the performance of deep learning models across various types of medical images, including cardiac ultrasound and CT scans.

Data availability

The dataset that supports the findings and conclusion of this study are available from the corresponding author on reasonable request. The data are not publicly available due to privacy.

Received: 2 April 2024; Accepted: 31 October 2024
Published online: 08 November 2024

References

1. Hwang, C. C. et al. Dual-colour chromogenic in-situ hybridization is a potential alternative to fluorescence in-situ hybridization in HER2 testing. *Histopathology* **59**(5), 984–992 (2011).
2. Jayasena Kaluarachchi, T. et al. Diagnosing human cutaneous leishmaniasis using fluorescence in situ hybridization. *Pathogens Global Health* **115**(5), 307–314 (2021).
3. Yang, R. et al. Identification of chromosomal abnormalities and genomic features in near-triploidy/tetraploidy-acute leukemia by fluorescence in situ hybridization. *Cancer Manag. Res.*, 1559–1567 (2019).
4. Shirsat, H. S. et al. HER 2 status in invasive breast cancer: immunohistochemistry, fluorescence in-situ hybridization and chromogenic in-situ hybridization. *Indian J. Pathol. Microbiol.* **55**(2), 175 (2012).
5. Kiyose, S. et al. Detection of kinase amplifications in gastric cancer archives using fluorescence in situ hybridization. *Pathol. Int.* **62**(7), 477–484 (2012).
6. Hayashida, T. et al. Establishment of a deep-learning system to diagnose BI-RADS4a or higher using breast ultrasound for clinical application. *Cancer Sci.* **113**(10), 3528 (2022).
7. Shimizu, H. & Nakayama, K. I. Artificial intelligence in oncology. *Cancer Sci.* **111**(5), 1452–1460 (2020).
8. Wei, T. et al. Survival prediction of stomach cancer using expression data and deep learning models with histopathological images. *Cancer Sci.* **114**(2), 690 (2023).
9. Xu, X. et al. A lightweight and robust framework for circulating genetically abnormal cells (CACs) identification using 4-color fluorescence in situ hybridization (FISH) image and deep refined learning. *J. Digit. Imaging*, 1–14 (2023).
10. Xue, T. et al. Deep learning to automatically evaluate HER2 gene amplification status from fluorescence in situ hybridization images. *Sci. Rep.* **13**(1), 9746 (2023).
11. Zakrzewski, F. et al. Automated detection of the HER2 gene amplification status in fluorescence in situ hybridization images for the diagnostics of cancer tissues. *Sci. Rep.* **9**(1), 8231 (2019).
12. Girshick, R. et al. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 580–587 (2014).
13. Girshick, R. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, 1440–1448 (2015).
14. Ren, S. et al. Faster r-cnn: towards real-time object detection with region proposal networks. *Adv. Neural. Inf. Process. Syst.*, 28 (2015).
15. Zhou, J. et al. Fusion PSPnet image segmentation based method for multi-focus image fusion. *IEEE Photonics J.* **11**(6), 1–12 (2019).
16. Liu, W. et al. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14* 21–37 (Springer International Publishing, 2016).
17. Jiang, P. et al. A review of Yolo algorithm developments. *Procedia Comput. Sci.* **199**, 1066–1073 (2022).
18. Duan, K. et al. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6569–6578 (2019).
19. Koonce, B. & Koonce, B. EfficientNet. Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization, 109–123 (2021).
20. Les, T. et al. Localization of spots in FISH images of breast cancer using 3-D shape analysis. *J. Microsc.* **262**(3), 252–259 (2016).
21. Xu, C. et al. An efficient fluorescence in situ hybridization (FISH)-based circulating genetically abnormal cells (CACs) identification method based on multi-scale MobileNet-YOLO-V4. *Quant. Imaging Med. Surg.* **12**(5), 2961 (2022).

22. Bouilhol, E. et al. DeepSpot: a deep neural network for RNA spot enhancement in single-molecule fluorescence in-situ hybridization microscopy images. *Biol. Imaging* **2**, e4 (2022).
23. Wang, A. et al. Yolov10: real-time end-to-end object detection. arXiv preprint arXiv:2405.14458 (2024).
24. Sapkota, R. et al. Yolov10 to its genesis: a decadal and comprehensive review of the you only look once series. arXiv preprint arXiv:2406.19407 (2024).
25. Redmon, J. & Farhadi, A. YOLO9000: better, faster, stronger. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7263–7271 (2017).
26. Huang, H. Y. et al. Classification of skin cancer using novel hyperspectral imaging engineering via YOLOv5. *J. Clin. Med.* **12**(3), 1134 (2023).
27. Du, Y. et al. Pavement distress detection and classification based on YOLO network. *Int. J. Pavement Eng.* **22**(13), 1659–1672 (2021).
28. Chen, L., Zhou, F., Wang, S. et al. SWIPENET: Object detection in noisy underwater scenes. *Pattern Recognit.* **132**, 108926 (2022).
29. Diwan, T., Anirudh, G. & Tembhurne, J. V. Object detection using YOLO: challenges, architectural successors, datasets and applications. *Multimedia Tools Appl.* **82**(6), 9243–9275 (2023).
30. Chang, Y. et al. An improved YOLO model for UAV fuzzy small target image detection. *Appl. Sci.* **13**(9), 5409 (2023).
31. Zeng, S. et al. SCA-YOLO: a new small object detection model for UAV images. *Vis. Comput.*, 1–17 (2023).
32. Sunkara, R. & Luo, T. No more strided convolutions or pooling: A new CNN building block for low-resolution images and small objects. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* 443–459 (Springer Nature Switzerland, 2022).
33. Niu, Z., Zhong, G. & Yu, H. A review on the attention mechanism of deep learning. *Neurocomputing* **452**, 48–62 (2021).
34. Liu, J. et al. Learning a deep multi-scale feature ensemble and an edge-attention guidance for image fusion. *IEEE Trans. Circ. Syst. Video Technol.* **32**(1), 105–119 (2021).
35. Wang, W., Shen, J. & Ling, H. A deep network solution for attention and aesthetics aware photo cropping. *IEEE Trans. Pattern Anal. Mach. Intell.* **41**(7), 1531–1544 (2018).
36. Hu, J., Shen, L. & Sun, G. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7132–7141 (2018).
37. Wang, Q. et al. ECA-Net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11534–11542 (2020).
38. Xu, Y. et al. Gliding vertex on the horizontal bounding box for multi-oriented object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**(4), 1452–1459 (2020).
39. Long, Y. et al. Accurate object localization in remote sensing images based on convolutional neural networks. *IEEE Trans. Geosci. Remote Sens.* **55**(5), 2486–2498 (2017).
40. Yang, X. et al. Detecting rotated objects as gaussian distributions and its 3-d generalization. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**(4), 4335–4354 (2022).
41. Chhikara, B. S. & Parang, K. Global cancer statistics 2022: the trends projection analysis. *Chem. Biol. Lett.* **10**(1), 451–451 (2023).
42. Hunter, B., Hindocha, S. & Lee, R. W. The role of artificial intelligence in early cancer diagnosis. *Cancers* **14**(6), 1524 (2022).
43. Liu, Z. et al. Instant diagnosis of gastroscopic biopsy via deep-learned single-shot femtosecond stimulated Raman histology. *Nat. Commun.* **13**(1), 4050 (2022).
44. Ker, J. et al. Deep learning applications in medical image analysis. *IEEE Access* **6**, 9375–9389 (2017).
45. Greenspan, H., Van Ginneken, B. & Summers, R. M. Guest editorial deep learning in medical imaging: overview and future promise of an exciting new technique. *IEEE Trans. Med. Imaging* **35**(5), 1153–1159 (2016).
46. Ghanem, N. M. et al. AUTO-BREAST: A Fully Automated Pipeline for Breast cancer Diagnosis Using AI technology. In *Artificial Intelligence in Cancer Diagnosis and Prognosis, Volume 2: Breast and Bladder Cancer* 6-1–6-24 (IOP Publishing, 2022).
47. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18* 234–241 (Springer International Publishing, 2015).
48. Zhou, Z. et al. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, Proceedings 4* 3–11 (Springer International Publishing, 2018).
49. Siddique, N. et al. U-net and its variants for medical image segmentation: a review of theory and applications. *IEEE Access* **9**, 82031–82057 (2021).

Author contributions

Conceptualization, Kan Liu; methodology, Zini Jian and Tianxiang Song; investigation, Zini Jian and Man Tang; software, Tianxiang Song; funding acquisition, Zini Jian, Kan Liu and Man Tang; data curation, Zini Jian and Tianxiang Song; writing—original draft preparation, Tianxiang Song; writing—review and editing, Zini Jian, Kan Liu and Man Tang; validation, Zhihui Zhang; resources, Man Tang , Zhao AI and Heng Zhao. All authors have read and agreed to the published version of the manuscript.

Funding information

This research was funded by the National Key Research and Development Program (2021YFB3801003), Innovative Research Groups of Hubei Province (2022CFA038), Hubei Provincial Natural Science Foundation of China(2023AFB284), and Wuhan Applied Foundational Frontier Project(2022013988065211).

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Z.J. or K.L.