

Article

Robust Tilapia Disease Diagnosis Based on Prompt-Enhanced Segment Anything Model and Neuro-Fuzzy Inference

Yicheng Gao and Guofu Feng *

College of Information Technology, Shanghai Ocean University, Shanghai 201200, China;
m230951651@st.shou.edu.cn

* Correspondence: gffeng@shou.edu.cn

Abstract

Diagnosing tilapia diseases in complex aquaculture environments is severely hindered by noisy backgrounds and limited high-quality pathological data. To overcome these bottlenecks, this study presents a two-stage diagnostic framework integrating an enhanced Segment Anything Model (SAM) with an Adaptive Neuro-Fuzzy Inference System (ANFIS). In the first stage, SAM is augmented with a Convolutional Block Attention Module (CBAM) feature adapter and a Region Proposal Network (RPN)-based prompt encoder. This design enables the automated and precise extraction of irregular disease lesions by self-generating spatial prompts, thereby isolating water background noise. In the second stage, clinical color features extracted from the lesion masks are classified using ANFIS. To optimize performance on small-scale datasets, ANFIS parameters are trained via Particle Swarm Optimization (PSO) under a numerically stable One-vs-Rest (OvR) binary cross-entropy loss. Validated on the public dataset “Enhancing Disease Detection in Nile Tilapia”, our method delivers an average segmentation Dice coefficient of 86.2% and a classification accuracy of 93.5%. This hybrid approach demonstrates strong potential as a foundational baseline for the automated monitoring of aquaculture diseases.

Keywords: fish disease identification; image segmentation; prompt learning; Segment Anything Model; particle swarm optimization; Adaptive Neuro-Fuzzy Inference System

1. Introduction

As global fisheries expand, the aquaculture industry has emerged as a vital pillar of the agricultural economy. However, diseases in farmed fish cause substantial economic losses, posing a major threat to the sector. As a primary aquaculture species, tilapia faces threats from common diseases such as Streptococcosis (Str), Columnaris Disease (Col), and Tilapia Lake Virus (TiLV). Infections lead to reduced growth and reproduction rates, increased mortality, and declined production. Therefore, the timeliness and accuracy of diagnosing diseased fish are crucial for preventing outbreaks and administering effective treatments [1].

Driven by recent advancements in deep learning, an increasing number of computer vision models have been applied to fish disease detection. However, visual recognition of fish diseases still faces critical bottlenecks. A primary challenge is data scarcity. The lack of high-quality annotated images for aquaculture diseases often causes overfitting and poor generalization in parameter-intensive Convolutional Neural Networks (CNN). Furthermore, aquatic environments present complex backgrounds, surface reflections, and



Academic Editors: Atsushi Mase and Douglas O'Shaughnessy

Received: 12 May 2026

Revised: 18 June 2026

Accepted: 22 June 2026

Published: 25 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

water ripples. Healthy scale textures are easily confused with genuine lesions such as micro-ulcers or hemorrhages, making precise extraction difficult for traditional segmentation networks. Additionally, because the visual symptoms of fish diseases are often highly ambiguous, traditional hard classifiers frequently lack the biological interpretability and decision-making logic required to handle these uncertain, transitional features.

To address these challenges, this paper proposes a two-stage “segmentation-classification” model based on an Enhanced SAM (Segment Anything Model) and an Optimized ANFIS (Adaptive Neuro-Fuzzy Inference System). We introduce the large visual model SAM into the fish disease identification domain to enhance generalization. Leveraging its prompt-driven and class-agnostic segmentation capabilities, we improve SAM with self-generated prompt learning. The goal is to extract specific infection regions, exclude background noise and irrelevant fish body information, reduce the spatial dimension of feature maps, and make features more consistent and quantifiable. Subsequently, an ANFIS classifier optimized by Particle Swarm Optimization (PSO) performs inference and decision-making based on this imprecise lesion information, achieving high-precision diagnosis of tilapia diseases.

The main contributions of this paper are as follows:

- (1) We propose a two-stage deep learning model for diagnosing tilapia diseases, achieving high accuracy in multi-class disease classification on a benchmark dataset by combining the state-of-the-art large visual model SAM with the neuro-fuzzy network ANFIS.
- (2) Targeting the SAM backbone, we designed a lightweight Feature Adapter and a Region of Interest (ROI) Prompt Encoder. The feature adapter, based on attention mechanisms, implements the fusion of multi-scale features. The ROI Prompt Encoder generates guidance prompts for segmentation by supervised learning.
- (3) For classifier optimization, the PSO algorithm employs a numerically stable One-vs-Rest (OvR) Binary Cross-Entropy loss with a constrained Sigmoid activation function. It optimizes the premise and consequent parameters of the ANFIS network, maximizing classification performance on small-scale datasets.
- (4) We conduct comprehensive performance comparisons to highlight the superiority of the proposed model over existing State-Of-The-Art (SOTA) methods. Furthermore, extensive ablation studies verify the effectiveness and necessity of each designed module.

2. Related Work

This section reviews the application of computer vision in aquaculture to provide context for the proposed framework.

2.1. Deep Learning in Fish Disease Identification

The application of deep learning in automatic fish disease identification has evolved from basic classification to complex multi-stage frameworks. Early studies primarily utilized the powerful feature extraction capabilities of Convolutional Neural Networks. For example, Waleed et al. [2] improved the recognition accuracy for Epizootic Ulcerative Syndrome, Ichthyophthiriasis, and Columnaris using an Alex-Net algorithm based on CNNs, providing a viable solution for real-time fish disease detection. Cho et al. [3] investigated the performance enhancement of deep CNN models in aquaculture applications through image preprocessing, normalization, and data augmentation techniques; they applied the classic deep CNN VGGNet-19 architecture to multi-class fish disease classification, achieving an overall accuracy of 90.96% in their experiments. Furthermore, Gupta et al. [4] proposed a CNN model for fish louse and wound detection, whose architecture comprises

15 convolutional layers and 5 fully connected layers. This model demonstrated superior performance compared to VGG-19 and VGG-16 models, reaching a detection accuracy of 96.7%. Later, Khabusi et al. [5] proposed an enhanced U-Net and PSO-optimized ANFIS model for automatic fish disease identification from underwater images. The U-Net was enhanced with multi-head channel and spatial attention to segment infected regions and extract color pixel intensity features. ANFIS was optimized by PSO, confirming the method's effectiveness.

To address the problem of lesion localization in complex environments, researchers have integrated object detection models. Yasruddin et al. [6] used the Faster R-CNN architecture with Inception ResNet-V2 as the feature extraction network to identify diseased fish suffering from ulcers and white spot disease, achieving a detection precision of 51.8%. Ouyang et al. [7] proposed a high-precision fish disease detection model with multi-module synergy based on a deeply modified YOLO architecture to identify six types of fish diseases, including ulcers and parasites. This model achieved a mean Average Precision (mAP) of 97.2% on an established dataset, demonstrating outstanding performance in handling minute lesion features under challenging conditions.

Recently, the integration of attention mechanisms and Transformer-based architectures has further improved recognition precision. Huang et al. [8] proposed a novel hybrid method based on an attention-based multi-layer fusion network and an Online Sequential Extreme Learning Machine (OSELM) to identify six types of fish diseases in aquaculture. The proposed method achieved an accuracy of 94.28% on a custom dataset with background elimination, proving its applicability for fish disease identification in complex underwater environments. Wang et al. [9] proposed a detection model named DyFish-DETR, which utilizes a hybrid encoder and a CNN network for extracting texture features, combined with the multi-head self-attention mechanism of Transformers. It achieved a mAP of 96.6% on a freshwater fish dataset.

2.2. Image Segmentation Based on Deep Learning

Previous research indicates that precise lesion segmentation is a prerequisite for accurate fish disease diagnosis. To address the challenge of increasingly complex image segmentation scenarios in production and daily life, a series of deep learning-based image segmentation networks and methods have been proposed. Long et al. [10] proposed the FCN (Fully Convolutional Network), which features convolutional layers instead of fully connected ones, realizing end-to-end, pixel-to-pixel semantic segmentation training and laying the foundation for semantic segmentation using deep neural networks. Ronneberger et al. [11] proposed U-Net for segmenting biological microscope images; its structure features a down-sampling part similar to the FCN architecture, while the up-sampling part uses up-convolutions to reduce the number of feature maps while increasing their dimensionality. Finally, the feature maps are processed through convolutions to classify each pixel of the input image and generate a segmentation map. Currently, there are various improved models of U-Net; for instance, Zhou et al. [12] developed a nested U-Net architecture, and Xiao et al. [13] used Res2Net as the backbone of U-Net to segment Computed Tomography images. In the application of object detection, R-CNN and their extended models have achieved success. Among them, He et al. [14] proposed Mask R-CNN for object instance segmentation, which jointly trains losses for bounding box coordinates, predicted classes, and segmentation masks. It ultimately achieved instance segmentation on the basis of semantic segmentation, resulting in not only higher segmentation accuracy but also a more flexible model. The DeepLab series was initially introduced by Chen et al. [15] with the V1 model, which used dilated convolutions to address the issue of resolution reduction during sampling and employed fully connected Conditional

Random Fields to optimize the post-processing of segmented images to obtain more feature and context information. Currently, DeepLab-v3+ [16] adopts an encoder–decoder structure, using dilated depth-wise separable convolutions to replace max pooling for refining segmentation boundaries, ultimately achieving end-to-end semantic segmentation and fine-grained image segmentation. The large vision model SAM [17] was developed by the Meta AI research team. As an automated image segmentation model, it has received widespread attention for its powerful ability to generate accurate object masks through fully automatic or interactive modes. The introduction of the large SAM model marks the beginning of the prompt-driven paradigm in the field of image segmentation, opening up new, unexplored areas for artificial intelligence and image segmentation.

2.3. Prompt Learning

In the fully supervised learning paradigm, machine learning models are trained on labeled data of target tasks to accomplish specific objectives [18]. Subsequently, fine-tuning pre-trained models for downstream tasks has become a common training pattern in the field. However, with the continuous growth in model size and the number of tasks, fine-tuning, storing, and deploying models for specific tasks result in significant resource wastage. Consequently, a novel learning paradigm has emerged with research progression, known as prompt learning [19]. In this approach, downstream tasks are reformulated by manually designing prompts that resemble training tasks, thereby aligning the downstream task objectives with pre-training objectives. This method is equally applicable to downstream visual tasks [20]. Among current large models, CLIP [21] and CoOp [22] utilize prompt learning to align the feature spaces of text and images, thereby enhancing semantic understanding capabilities, narrowing the gap between pre-training and fine-tuning, and preventing model overfitting. RSPrompter [23] leverages prompt learning to facilitate SAM's instance segmentation capability in remote sensing imagery. Currently, methods for constructing prompts mainly include manual templates, generation, and word embedding fine-tuning [24].

2.4. Neuro-Fuzzy Inference System

Fuzzy logic is a mathematical logic that deals with uncertainty and imprecision in real-world decision-making processes [25]; it represents vague and uncertain information by processing degrees of truth. Fuzzy inference systems [26] are based on fuzzy logic and simulate human-like reasoning and decision-making processes under uncertain or imprecise conditions, serving as tools for decision-making and control systems. Unlike traditional control systems that use precise mathematical models and algorithms to describe system behavior, fuzzy logic control employs linguistic variables and rules to handle uncertainty and imprecision. Neuro-fuzzy inference systems have evolved from fuzzy inference systems by integrating fuzzy logic with neural networks. In this integration, fuzzy logic provides a framework for handling data uncertainty and fuzziness, while neural networks offer powerful learning algorithms for data-driven modeling and prediction. This enables the construction of input–output mappings based on human knowledge, effectively modeling nonlinear relationships to classify different values and make decisions using imprecise information. Consequently, neuro-fuzzy inference systems demonstrate robust classification capabilities in pattern recognition [27].

3. Methods

While recent works demonstrate the growing efficacy of deep learning in aquaculture, several gaps remain: a heavy reliance on large, high-quality annotated datasets, susceptibility to underwater background noise, and a lack of diagnostic interpretability for

transitional lesions. This section details our proposed solution. Specifically, we introduce a novel two-stage architecture that leverages the enhanced SAM and PSO-Optimized ANFIS. Compared to recent instance segmentation models and SAM variants that require manual intervention or rely on generic semantic features, our Enhanced SAM introduces a self-generating ROI prompt encoder and a CBAM-based feature adapter. This design enables the foundation model to autonomously target irregular lesions, isolating background noise without requiring dense pixel-level convolutions. Our optimized ANFIS excels in handling overlapping clinical symptoms by mathematically modeling symptom ambiguity via fuzzy rules, acting as a highly interpretable classifier that prevents overfitting.

3.1. Network Architecture

This study proposes a two-stage tilapia disease identification model that identifies disease types through lesion region segmentation and neuro-fuzzy inference classification. The overall architecture is shown in Figure 1, the network is designed to address a 3-class classification problem: identifying Streptococcosis, Columnaris Disease, and Tilapia Lake Virus.

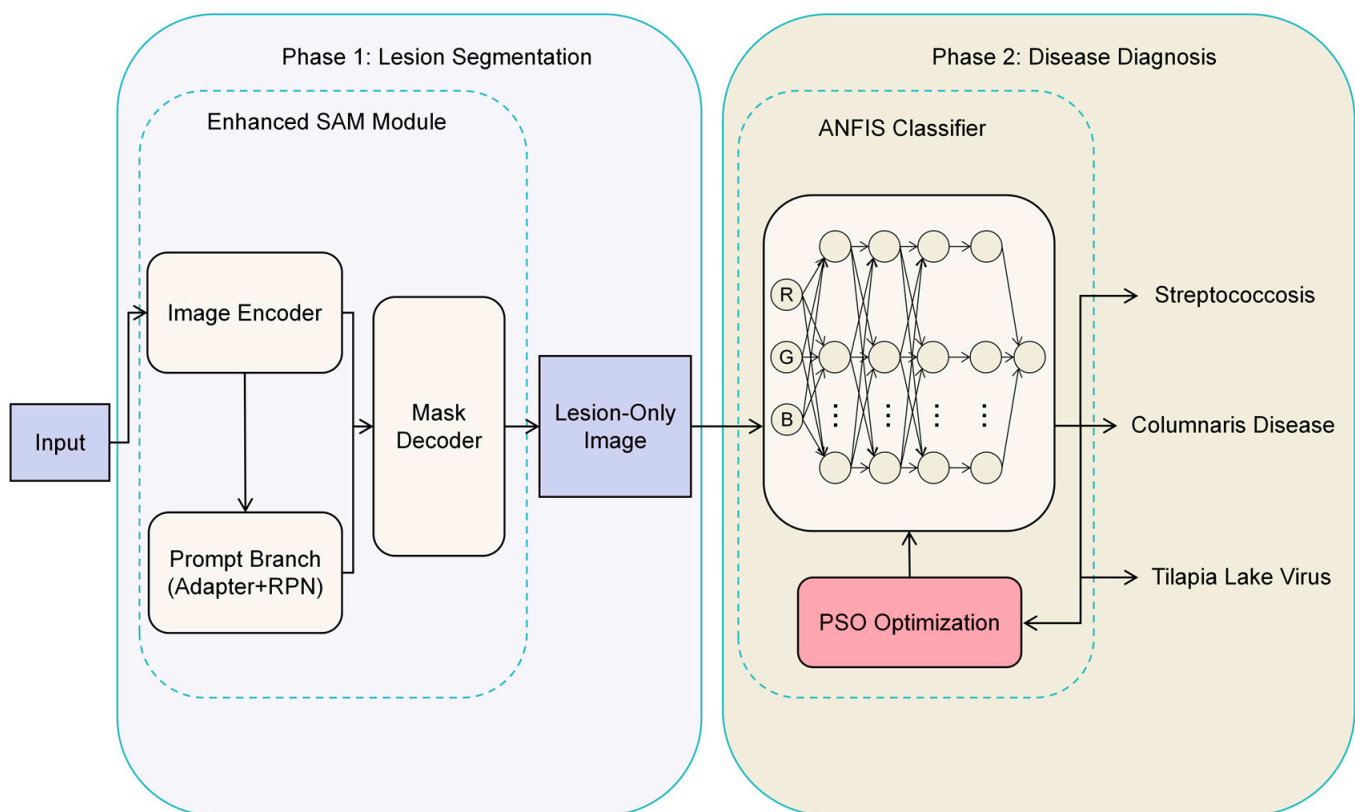


Figure 1. Architecture of the proposed two-stage fish disease identification method.

3.2. Segmentation Module

SAM [17] is a class-agnostic interactive instance segmentation method that relies on provided prior manual prompts, such as points, boxes, and masks, to guide segmentation. The structure of the original SAM is primarily divided into three components: image encoder, prompt encoder, and mask decoder. SAM processes the input image into intermediate feature maps via a pre-trained MAE (Masked Autoencoder) based on the ViT (Vision Transformer). The prompt encoder encodes prompts like points, boxes, and masks into embedding tokens, and the cross-attention mechanism in the mask decoder combines image feature maps, prompt embeddings, and output tokens to generate segmentation

masks. In the context of tilapia disease segmentation, SAM lacks the ability to automatically generate accurate segmentation masks for specific fish disease features. Therefore, in our method, we optimize SAM's prompt encoder to utilize generated prompts to guide the segmentation task, thereby enhancing SAM's automatic segmentation performance for tilapia diseases.

We utilize the original SAM network as the basic model for the segmentation module. As shown in Figure 2, its complete module architecture includes the original SAM's image encoder and mask decoder, as well as our optimized ROI Prompt Encoder designed based on the RPN [28] network. The original SAM image encoder and prompt encoder will be frozen, whereas the SAM mask decoder and our proposed prompt generation branch (comprising the Feature Adapter and the RPN) will participate in the training process.

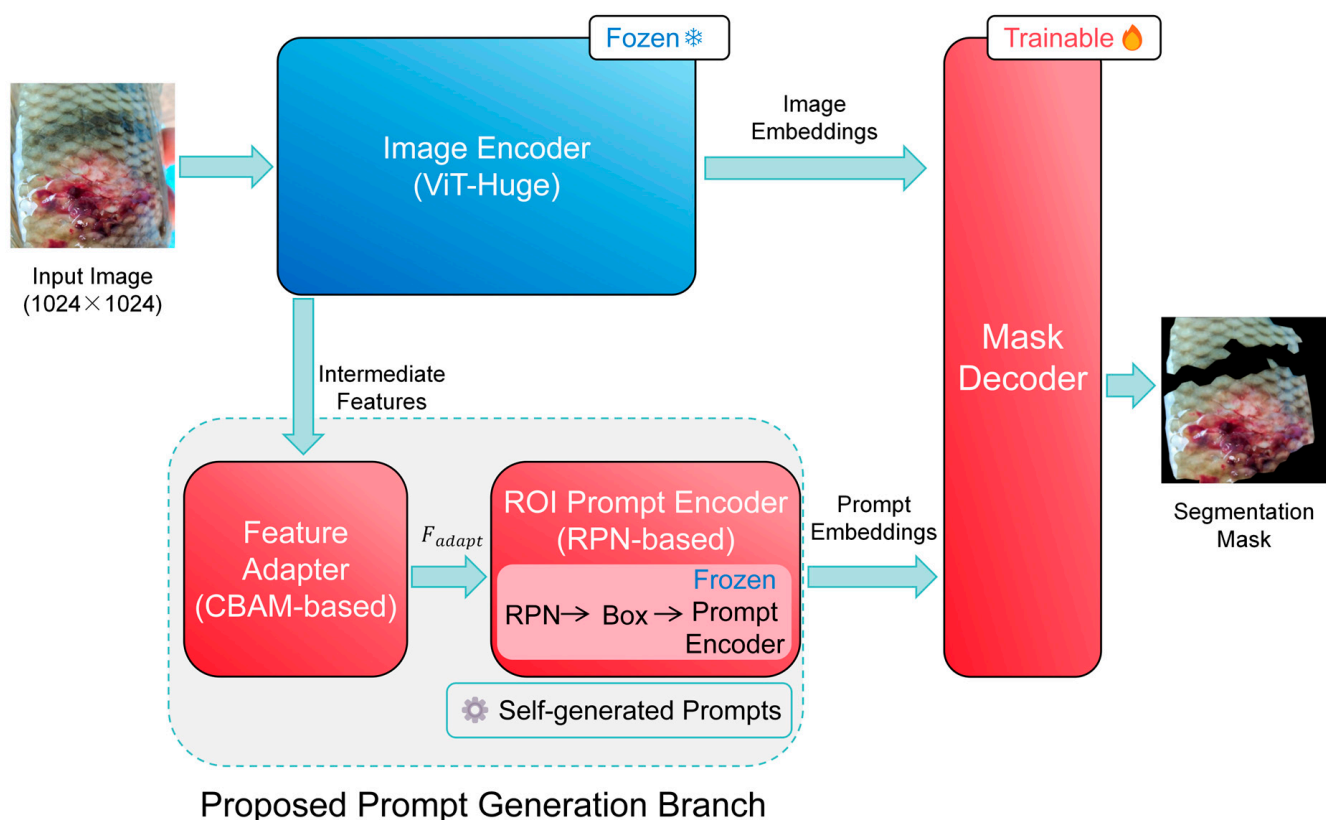


Figure 2. Framework diagram of the segmentation module.

3.2.1. Feature Adapter

Intermediate features of large vision foundation models primarily focus on object edges and geometric structures, lacking sensitivity to minor local lesions in tilapia (early-stage congestion spots or slight scale loss). Directly feeding unprocessed intermediate feature maps into the Region Proposal Network would result in a substantial number of subtle lesion features being neglected in deep networks. Therefore, we propose and design a lightweight feature adapter based on the Convolutional Block Attention Module [29] at the front end of the prompt encoder. This adapter adaptively learns weight distributions for multi-scale features without disrupting the original pre-trained feature space structure, thereby suppressing water background noise and amplifying the response intensity of lesion areas. As illustrated in Figure 3, the CBAM is composed of a channel attention module and a spatial attention module arranged in series.

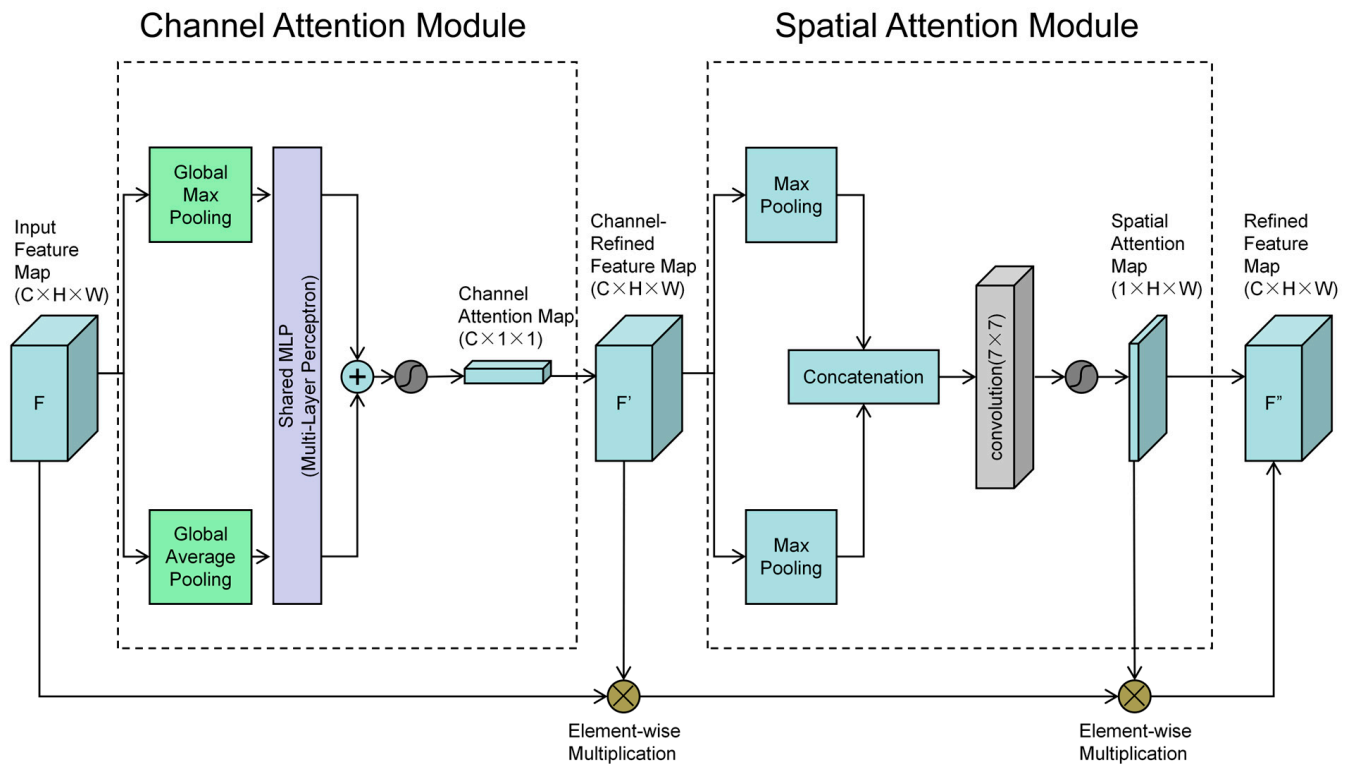


Figure 3. Flowchart of the CBAM.

The core computational pipeline of the channel attention module can be divided into three stages: spatial information compression, channel dependency modeling, and feature weight recalibration.

(1) Spatial information compression

Assume the feature map tensor extracted from the intermediate layer of ViT-Huge and fed into the feature adapter is $F \in \mathbb{R}^{C \times H \times W}$. Given that the input image undergoes 1024×1024 pixel standardization and network downsampling, the specific dimensions are $C = 256$, $H = 64$, and $W = 64$. To eliminate computational redundancy in the spatial dimension, the feature map must be squeezed spatially. This module employs parallel Global Average Pooling (GAP) and Global Max Pooling (GMP). GAP extracts the overall perceptual distribution of the tilapia and its aquatic environment by calculating the arithmetic mean of all 64×64 spatial pixels within each channel. GMP captures discriminative local pathological information by extracting the maximum response value within each channel. These operations compress the original feature map F into two 1D spatial context descriptor vectors, denoted as $F_{avg}^c \in \mathbb{R}^{C \times 1 \times 1}$ and $F_{max}^c \in \mathbb{R}^{C \times 1 \times 1}$. The mathematical expressions are formulated as follows:

$$F_{avg}^c = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W F_c(i, j), \quad (1)$$

$$F_{max}^c = \max_{i,j} F_c(i, j). \quad (2)$$

(2) Channel dependency modeling

To capture the complex non-linear interdependencies among the 256 channels, the two context vectors are fed into a shared-weight Multi-Layer Perceptron (MLP). To enhance the model's non-linear fitting capability while controlling the parameter size, this MLP adopts a typical bottleneck architecture. This architecture comprises two fully connected layers

and a reduction ratio r , which is set to 16 based on experimental evaluations balancing computational cost and performance. In the dimensionality reduction layer, the channel dimension is compressed from C to C/r (256 to 16) with weight matrix $W_0 \in \mathbb{R}^{\frac{C}{r} \times C}$, followed by a Rectified Linear Unit (ReLU) activation function δ to enhance non-linear expression and filter redundant negative response signals. In the dimensionality expansion layer, the channel dimension is mapped back to the original C (16 to 256) with weight matrix $W_1 \in \mathbb{R}^{C \times \frac{C}{r}}$. The shared-weight mechanism ensures that F_{avg} and F_{max} are processed using the identical weight matrices W_0 and W_1 , guaranteeing a unified evaluation standard for both average and maximum features.

(3) Feature weight generation and recalibration

The two feature vectors mapped by the MLP are aggregated via element-wise addition to fuse global and local information. Subsequently, a Sigmoid activation function σ smoothly maps the fused values to a continuous probability interval of $(0, 1)$, generating the final channel attention weight vector $M_c(F) \in \mathbb{R}^{C \times 1 \times 1}$. This process is defined as:

$$M_c(F) = \sigma(W_1(\delta(W_0 F_{avg})) + W_1(\delta(W_0 F_{max}))). \quad (3)$$

After obtaining the weight vector $M_c(F)$, the final feature recalibration is executed by broadcasting $M_c(F)$ along the spatial dimension and performing element-wise multiplication with the original input feature map F :

$$F' = M_c(F) \otimes F. \quad (4)$$

Through this recalibration process, the output feature map F' retains the same dimensions as the input F , but its internal information distribution has been restructured. For channels that distinctly represent tilapia diseases, the Sigmoid output of $M_c(F)$ approaches 1, preserving or amplifying the features. Conversely, for channels capturing only irrelevant lighting or water backgrounds, the weight decays toward 0, suppressing those features. This dynamic channel recalibration mechanism effectively eliminates high-dimensional channel noise for subsequent spatial localization.

The feature map F' recalibrated by the channel attention mechanism contains more explicit channel semantic information. However, to further guide the model in spatial localization of lesion coordinates on a 2D plane, a spatial attention module is cascaded following the channel attention module. The spatial attention mechanism utilizes inter-channel information complementarity to construct a 2D global spatial response mask. The implementation pipeline includes channel information compression, cross-dimensional feature aggregation, large-receptive-field convolution for context modeling, and final spatial weight assignment.

(4) Channel information compression

In contrast to pooling along the spatial dimension in the channel attention module, the spatial attention module compresses information along the channel axis. Assuming the input feature map is $F' \in \mathbb{R}^{C \times H \times W}$ ($C = 256$, $H = 64$, $W = 64$), the algorithm computes the maximum and average values across all 256 channels at each spatial pixel (i, j) . Cross-channel max pooling captures prominent priors with strong activation responses, effectively localizing sharp-edged, color-mutated tilapia ulcers or hemorrhagic patches. Cross-channel average pooling synthesizes responses from all channels to provide a smooth global context background, aiding the network in understanding the transitional relationship between

lesions and surrounding healthy tissues. Consequently, the 3D feature map F' is compressed into two 2D spatial descriptor matrices, $F'_{max}(i, j) \in \mathbb{R}^{1 \times H \times W}$ and $F'_{avg}(i, j) \in \mathbb{R}^{1 \times H \times W}$:

$$F'_{max}(i, j) = \max_{c \in \{1, 2, \dots, C\}} F'_c(i, j), \quad (5)$$

$$F'_{avg}(i, j) = \frac{1}{C} \sum_{c=1}^C F'_c(i, j). \quad (6)$$

(5) Cross-dimensional feature concatenation and context modeling

To fuse these two complementary spatial priors, the algorithm concatenates F'_{max} and F'_{avg} along the channel dimension to generate an aggregated feature map $F'_{cat} \in \mathbb{R}^{2 \times H \times W}$. A standard 2D convolution layer is introduced to extract continuous spatial dependencies from this dual-channel mapping and generate the final attention mask. The kernel size k determines the receptive field of the spatial attention; a small kernel captures only local neighborhood variations and fails to establish macroscopic spatial correlations between lesions and the overall fish structure, leading to over-focus on minor noise like water ripples. Based on experimental evaluations, k is set to 7×7 , employing zero-padding ($P = 3$) and a stride ($S = 1$) to maintain spatial resolution. This 7×7 convolution reduces the dual-channel F'_{cat} back to a single-channel 2D activation map. Finally, a Sigmoid function σ maps each element to a continuous probability space of $(0, 1)$, forming the spatial attention mask $M_s(F') \in \mathbb{R}^{1 \times H \times W}$:

$$M_s(F') = \sigma\left(f^{7 \times 7}\left(\left[F'_{avg}; F'_{max}\right]\right)\right). \quad (7)$$

(6) Feature weight generation and recalibration

Upon acquiring the 2D spatial attention mask $M_s(F')$, the final spatial feature recalibration is executed. Since $M_s(F')$ has 1 channel and the input F' has 256 channels, $M_s(F')$ is broadcasted along the channel dimension to align with F' . An element-wise spatial multiplication is then performed to output the final enhanced multi-scale feature map:

$$F'' = M_s(F') \otimes F'. \quad (8)$$

3.2.2. ROI Prompt Encoder

Under the prompt learning paradigm, the construction of high-quality prompts directly dictates the fine-tuning upper bound of pre-trained foundation models in downstream tasks. Addressing the irregularity in morphology and randomness in the distribution of tilapia lesions, and to achieve automated detection and extraction of these regions, this study innovatively introduces an ROI Prompt Encoder based on a Region Proposal Network. This module aims to mine potential lesion regions from multi-scale feature maps and transform them into high-dimensional prompt embeddings that the SAM mask decoder can comprehend. Unlike conventional semantic segmentation networks, this module focuses solely on generating class-agnostic lesion bounding box prompts, leaving specific disease classification to the subsequent classification network. As shown in Figure 4, the specific workflow mainly consists of three parts:

(1) Candidate region generation and feature extraction

Initially, the encoder utilizes the RPN to perform dense sampling on the enhanced multi-scale feature map F_{adapt} outputted by the feature adapter. We design various initial anchors with different scales (e.g., 32×32 , 64×64 , 128×128 pixels) and aspect ratios (e.g., 1:1, 1:2, 2:1) to encompass a wide range of lesion morphologies, from minute ulcer spots to large subcutaneous hemorrhages. Through a sliding window over the feature

map, the system generates predefined initial anchors at each spatial location to formulate a preliminary set of candidate boxes. To eliminate highly overlapping redundant proposals, a Non-Maximum Suppression (NMS) algorithm with an Intersection over Union (IoU) threshold of 0.7 is applied, filtering out the most representative high-quality foreground proposal set $\{P_i\}$.

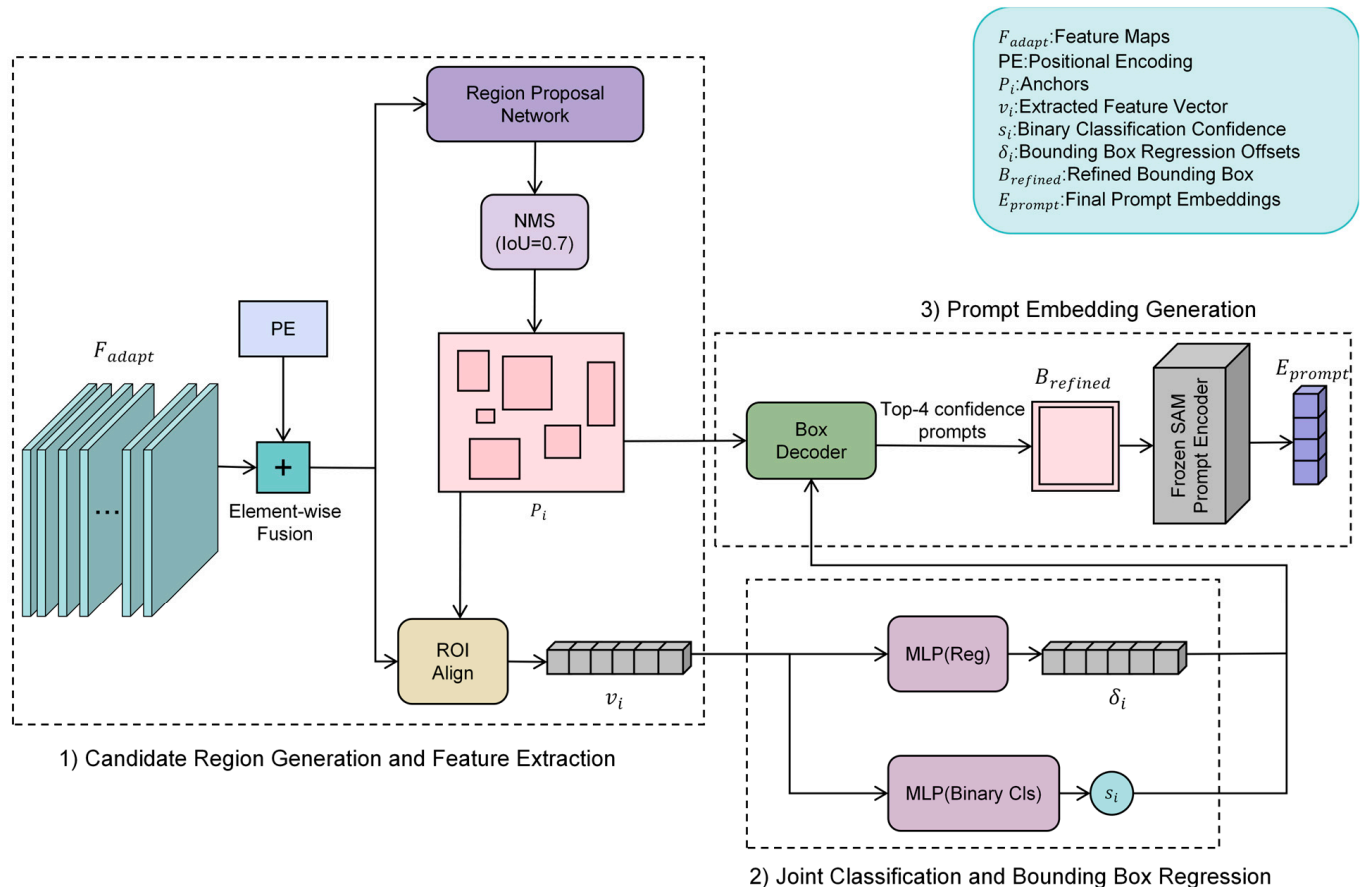


Figure 4. Architecture diagram of the ROI Prompt Encoder.

Subsequently, the ROI Align method is employed to extract a fixed-dimensional local feature vector v_i for each valid proposal box P_i . Since the subsequent SAM mask decoder is highly sensitive to spatial positions, to prevent the loss of relative positional information during pooling, absolute spatial Positional Encoding (PE) is added to the pixel-level features before extraction:

$$v_i = \text{RoIAlign}(F_{\text{adapt}} + \text{PE}, P_i). \quad (9)$$

This guarantees that each local feature vector not only encompasses the color and texture representations of the lesion but also implicitly encodes its absolute geometric coordinates on the entire tilapia body.

(2) Joint classification and bounding box regression

To accomplish the subtasks of determining whether a region is a lesion and pinpointing its exact location, the extracted ROI feature vectors v_i are respectively fed into two structurally parallel but parameter-independent lightweight MLP branches: the Binary Classification Head and the Regression Head.

The primary function of the binary classification head is to perform class-agnostic foreground/background discrimination; it determines whether the current candidate region contains any type of tilapia lesion tissue without specifying the disease cate-

gory. This branch consists of two fully connected layers, utilizing an intermediate ReLU activation function for non-linear feature extraction and a terminal Sigmoid function to map the output to a probability space. The forward inference is mathematically expressed as:

$$s_i = \sigma \left(W_{cls}^2 \cdot \max \left(0, W_{cls}^1 \cdot v_i + b_{cls}^1 \right) + b_{cls}^2 \right), \quad (10)$$

where W_{cls}^1 and W_{cls}^2 are the weight matrices of the first and second fully connected layers, respectively, and b_{cls1} and b_{cls2} are their corresponding bias vectors. The final scalar output $s_i \in (0, 1)$ represents the confidence score that the current proposal belongs to the lesion foreground.

During training, the confidence threshold τ is set to 0.5, retaining only candidate boxes with $s_i > \tau$, which efficiently filters out most invalid proposals from water backgrounds or healthy fish. Furthermore, an IoU matching mechanism is introduced to define ground-truth positive and negative samples. A candidate anchor is labeled as a positive sample if its IoU with any ground-truth lesion bounding box exceeds 0.7 or if it is the anchor with the highest IoU for a specific ground-truth box. Conversely, it is labeled as a negative sample (background noise) if its IoU with all ground-truth boxes is less than 0.3. Anchors with IoUs between 0.3 and 0.7 are considered transitional samples and are ignored during the binary classification loss calculation to prevent gradient oscillations caused by ambiguous optimization targets.

Because the initial sliding-window anchors are predefined with fixed scales and aspect ratios, they can only roughly cover the lesion areas and fail to tightly fit the irregular edges of pathological tissues. The regression head is tasked with predicting parameters to compensate for these geometric errors. Its network topology mirrors the classification head, but the final layer omits any activation function, allowing the network to output continuous values across the $(-\infty, +\infty)$ real number domain. Its forward inference formula is:

$$\delta_i = W_{reg}^2 \cdot \max \left(0, W_{reg}^1 \cdot v_i + b_{reg}^1 \right) + b_{reg}^2. \quad (11)$$

The regression head outputs a 4D coordinate offset vector $\delta_i = (\Delta x, \Delta y, \Delta w, \Delta h)$. These variables are not absolute pixel shifts, but a set of scale-invariant relative transformation coefficients. Using relative proportions effectively mitigates the impact of lesion scale variances, enabling the network to apply consistent optimization logic across lesions of differing sizes.

(3) Prompt embedding generation

After completing joint classification and bounding box regression, the network obtains a 4D relative offset vector $\delta_i = (\Delta x, \Delta y, \Delta w, \Delta h)$ to correct the initial anchors. To convert these low-dimensional geometric values into high-dimensional sparse prompt embeddings E_{prompt} capable of directly driving the SAM mask decoder, the system executes the following mapping stages.

First, the network must revert the relative offsets back into absolute pixel coordinates on the original tilapia image. Using the center point and dimensional parameters $(x_{refined}, y_{refined}, w_{refined}, h_{refined})$ of the refined box, the system calculates the top-left (x_{min}, y_{min}) and bottom-right (x_{max}, y_{max}) coordinates. In early regression stages, gradient perturbations might cause decoded coordinates to exceed the image's physical boundaries. To prevent introducing anomalous spatial indices, the system rigidly clips the decoded coordinates. For an input image of size $W \times H$, the clipping operation is defined as:

$$x'_{min} = \max(0, \min(x_{min}, W - 1)), \quad (12)$$

$$y'_{min} = \max(0, \min(y_{min}, H - 1)). \quad (13)$$

The four extreme coordinates, processed identically, form the final geometric bounding box $B_{refined} = (x'_{min}, y'_{min}, x'_{max}, y'_{max})$, which optimally encloses the irregular tilapia lesions in the 2D spatial plane.

Directly feeding raw pixel coordinates (ranging from 0 to 1024) into a Transformer-based vision model is inefficient, as neural networks struggle to capture minute spatial discrepancies from low-frequency linear coordinate signals. Therefore, this study invokes the frozen prompt encoder from the SAM pre-trained weights, which utilizes random Fourier feature positional encoding. It first normalizes the absolute coordinates into a continuous interval of $(0, 1)$. Let an arbitrary normalized coordinate be $p = (x, y) \in (0, 1)^2$. The system uses a random Gaussian weight matrix $B \in \mathbb{R}^{2 \times 128}$, fixed during pre-training, to project the 2D coordinates into a 256-dimensional high-frequency sine-cosine space:

$$PE(p) = [\sin(2\pi B^T p), \cos(2\pi B^T p)] \in \mathbb{R}^{256}. \quad (14)$$

By introducing the Gaussian matrix B to amplify the high-frequency components, even two minutely separated lesion points produce discernible shifts in the mapped 256-dimensional feature space, enhancing the mask decoder's boundary delineation accuracy.

For the top-left corner p_{tl} and bottom-right corner p_{br} of the refined bounding box $B_{refined}$, the system adds their respective positional encodings to two learned vectors, $e_{tl_type} \in \mathbb{R}^{256}$ and $e_{br_type} \in \mathbb{R}^{256}$:

$$E_{tl} = PE(p_{tl}) + e_{tl_type}, \quad (15)$$

$$E_{br} = PE(p_{br}) + e_{br_type}. \quad (16)$$

These two vectors attach semantic labels to the spatial coordinates, informing the network of the prompt box's start and end boundaries. Finally, E_{tl} and E_{br} are concatenated along the sequence dimension to form a high-dimensional sparse prompt embedding matrix $E_{prompt} \in \mathbb{R}^{2 \times 256}$, representing the candidate lesion box.

In practice, this automatic prompt generation process fundamentally shifts the segmentation paradigm. Unlike traditional convolutional segmentation models (such as U-Net or FCN), which rely solely on fixed, learned convolutional weights to map pixels to classes and often struggle with ambiguous lesion boundaries, our proposed RPN-based encoder provides explicit spatial guidance. The process operates dynamically: the RPN autonomously scans the purified feature maps, filters out background regions via binary classification, and generates tight, refined bounding boxes $B_{refined}$ precisely enclosing the irregular lesions. These deterministic geometric coordinates are then mapped into sparse, high-frequency prompt embeddings E_{prompt} using the frozen SAM encoder. By feeding these explicit spatial priors into the SAM mask decoder, the model is compelled to concentrate its powerful attention mechanisms exclusively within the proposed regions. This self-generated prompt strategy effectively bridges the gap between SAM's requirement for manual prompts and the need for full automation, contributing to improved boundary precision and a reduction in false positives when compared to traditional, unguided segmentation pipelines.

3.2.3. Comprehensive Loss Function

To synergistically optimize the Region Proposal Network, prompt generation accuracy, and final segmentation quality, this paper designs a multi-task joint loss function. This

function is composed of three parts: the RPN proposal loss, the prompt encoder regression loss, and the SAM decoder segmentation loss. The total loss is defined as follows:

$$L_{total} = \lambda_1 (L_{rpncls} + L_{rpnreg}) + \lambda_2 L_{promptreg} + \lambda_3 (L_{focal} + L_{dice}). \quad (17)$$

λ_1 , λ_2 , and λ_3 are hyperparameters that balance the weights of each task. L_{rpncls} represents the RPN binary classification loss, utilizing Binary Cross-Entropy to determine whether the generated anchors are background or foreground (whether they contain diseased regions); this ensures that the RPN can effectively distinguish tilapia lesions from the background. L_{rpnreg} represents the RPN localization loss, employing Smooth L1 Loss to regress the coordinate offsets of the anchors. This corrects the positions of the initially generated proposals, ensuring that ROI Align can extract effective local features. $L_{promptreg}$ is the critical loss connecting the RPN and SAM. Since this paper utilizes a frozen original SAM prompt encoder, the precision of its input geometric coordinates directly determines the quality of the generated prompt embeddings. Therefore, we utilize Smooth L1 Loss to supervise the offsets δ_i output by the localization regression head. This loss further finely tunes the proposal boxes, ensuring that the decoded coordinates $B_{refined}$ can provide precise spatial guidance for SAM, thereby reducing the interference of background noise on segmentation. L_{focal} and L_{dice} jointly represent the losses derived from the final image segmentation. L_{focal} is used to address the imbalance between lesion pixels and background pixels; by reducing the weight of easy-to-classify samples, it forces the model to focus more attention on the fine segmentation of lesion edges during training. L_{dice} directly optimizes the Dice coefficient; as Dice loss is insensitive to target scale variations and focuses more on mining foreground regions, it is suitable for handling the segmentation tasks of tilapia lesions, which are characterized by irregular edges and variable shapes.

3.3. Classification Module

In the field of fish disease identification, due to the varying manifestations and severity of disease symptoms, fish disease identification possesses subjectivity and relativity. In fuzzy neuro-inference systems, fuzzy logic can effectively handle the uncertainty and imprecision in classification problems. Therefore, we utilize ANFIS [25] as the classification model and optimize it with the PSO [30] algorithm. By taking the disease region segmentation mask generated by the segmentation module as input, we diagnose and identify the categories of tilapia diseases.

3.3.1. Fuzzy C-Means Clustering

First, we extract the color pixel intensity features exclusively from the disease infection regions segmented by the Enhanced SAM. Specifically, the processing involves calculating the global mean pixel values of the Red (R_{avg}), Green (G_{avg}), and Blue (B_{avg}) channels strictly within the generated spatial mask. The selection of pure color features, as opposed to morphological or complex texture features, is fundamentally driven by the clinical pathology of tilapia diseases. In aquaculture, the shape, scale, and specific texture of lesions vary unpredictably depending on the infection stage and fish size, making geometric features unreliable. Conversely, color alterations serve as highly consistent and definitive clinical biomarkers. For instance, Streptococcosis typically presents with severe hemorrhaging (erythema) on the body surface or exophthalmia, which mathematically manifests as a prominently high R value coupled with suppressed G and B values. Columnaris disease often forms pale-yellow or whitish plaque-like “saddleback” lesions, resulting in high intensities across all three channels. In contrast, Tilapia Lake Virus frequently causes deep, necrotic skin erosions and darkening, yielding universally low values across the RGB spectrum. By processing these features as spatial averages exclusively within the

precise SAM boundaries, we completely eliminate background color interference (such as water turbidity). This quantifies the subjective pathological color shifts into a deterministic, noise-free 3-dimensional feature vector $[R_{avg}, G_{avg}, B_{avg}]$, which serves as the ideal input for the ANFIS classification model.

Subsequently, FCM (Fuzzy C-Means) clustering is employed to cluster the input data, calculating cluster centers and the data distribution within clusters. Since the number of FCM clusters directly impacts the structure of the fuzzy system, we utilize Bayesian optimization to determine that the optimal number of clusters is 5. These clustering results are used to determine the key parameters of the ANFIS membership functions, ensuring that the membership functions accurately fit the data distribution. Regarding the construction of fuzzy rules, we decoupled the number of clusters from the number of fuzzy rules, employing Bayesian optimization to independently search for the optimal values for each. This is because while conventional cluster evaluation methods can identify the geometrically optimal distribution of data, geometric optimality does not necessarily equate to prediction optimality. To achieve high-precision classification, the model requires more complex logical reasoning; therefore, the number of rules exceeds the number of clusters. Based on the clustering partition, each cluster corresponds to the premise part of a fuzzy rule. After initializing five basic rules, the remaining fuzzy rules are generated by splitting the cluster center with the largest error. We defined five distinct fuzzy semantic levels (VL, L, M, H, VH) and calculated the membership degrees of samples belonging to each fuzzy set using Gaussian membership functions. Three of the fuzzy rules are listed here to demonstrate the final generation:

Rule1 : IF R_{avg} is VH AND G_{avg} is L AND B_{avg} is L, THEN Output is Streptococcus

Rule2 : IF R_{avg} is H AND G_{avg} is H AND B_{avg} is H, THEN Output is Columnaris Disease

Rule3 : IF R_{avg} is VL AND G_{avg} is VL AND B_{avg} is VL, THEN Output is Tilapia Lake Virus

3.3.2. ANFIS Network

The ANFIS network framework is illustrated in Figure 5. We designed it based on the Takagi–Sugeno fuzzy inference system [31], and its structure can be divided into five layers. The choice of the specific ANFIS network topology was driven by a combination of clinical relevance and empirical optimization. The 3 input nodes correspond to the clinical color biomarkers (R_{avg} , G_{avg} , B_{avg}) extracted by the segmentation module. The 5-layer architecture was selected because its linear consequent parameters are computationally efficient and suitable for subsequent global optimization via PSO. This specific configuration balances the network's capacity to map complex, transitional lesion features against the structural simplicity required for generalization.

(1) First, the fuzzy rules in the neuro-fuzzy inference system partition the input data variables into multiple fuzzy sets. Each node in this layer utilizes a Gaussian membership function to calculate the membership degree values of the input image's n -dimensional feature vector under different fuzzy sets. The formula can be expressed as:

$$f(x; s, c) = e^{-\frac{(x-c)^2}{2s^2}}, \quad (18)$$

where s is the standard deviation controlling the width of the curve, and c represents the center of the curve. These parameters are the premise parameters of the ANFIS network, initialized by the FCM clustering results. The fuzzification layer ultimately transforms the input data into a membership degree matrix to represent the membership degree of each input variable x corresponding to each fuzzy set.

(2) In the second layer, each node calculates the firing strength of each input sample corresponding to different fuzzy rules via the product T-norm.

(3) In this layer, the single firing strength ratio is obtained by calculating the ratio of the firing strength of each fuzzy rule to the sum of the firing strengths of all rules. This normalized result represents the weight of a single rule in the system's inference. The calculation formula for each node can be expressed as:

$$\bar{w}_i = \frac{w_i}{\sum_{j=1}^N w_j}. \quad (19)$$

(4) Each node in the fourth layer calculates the weighted output of the input data x corresponding to a single fuzzy rule for a single class using the consequent parameters and the firing strength ratio. The calculation process can be represented by Equation (20), where f_i represents the consequent function of the i -th rule, usually expressed as a linear combination of input variables, and $\{p_i, q_i, r_i\}$ are referred to as consequent parameters, which are optimized by the PSO algorithm during training.

$$\bar{w}_i f_i = \bar{w}_i (p_i x + q_i y + r_i). \quad (20)$$

(5) To achieve defuzzification, the nodes in this layer sum the weighted outputs of all fuzzy rules corresponding to a single class for the input data x . Through this process, ANFIS is capable of constructing an extremely complex nonlinear mapping surface. Its calculation process can be expressed by the formula:

$$Total = \sum_{i=1}^N \bar{w}_i f_i. \quad (21)$$

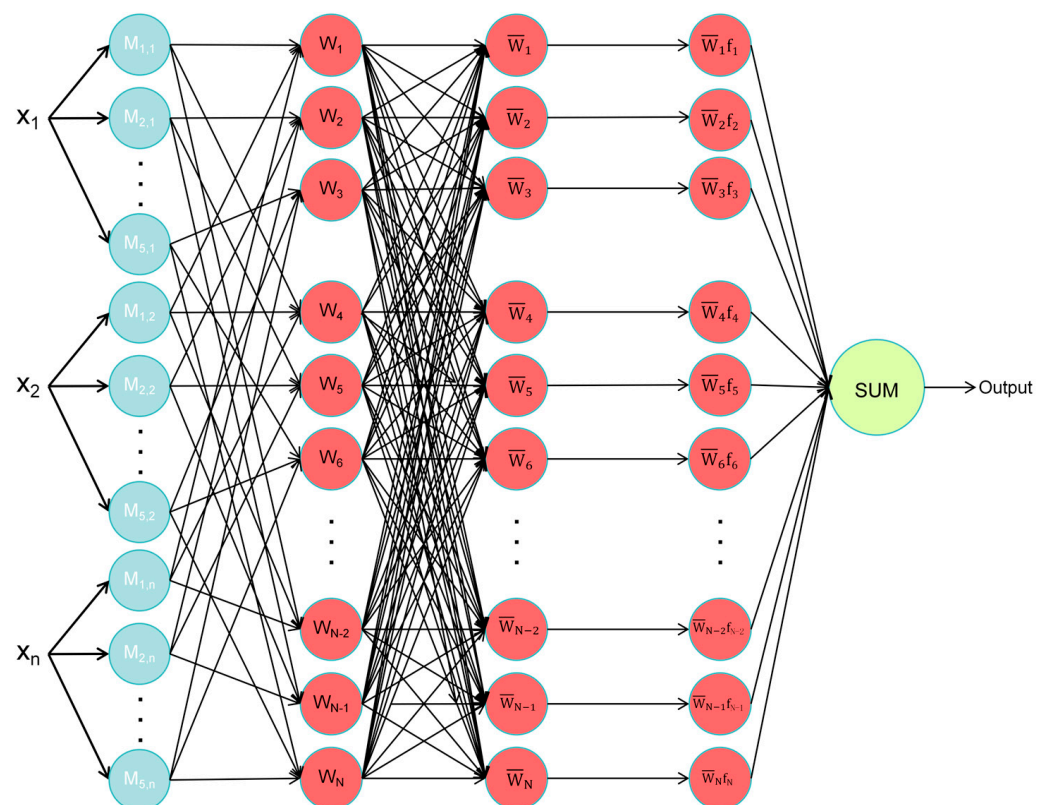


Figure 5. Schematic diagram of the 5-layer structure of the ANFIS network.

Crucially, the implementation of this defuzzification process equips the model to effectively handle ambiguous or transitional clinical cases. In real-world aquaculture, disease symptoms are rarely absolute. For instance, a tilapia exhibits early-stage overlapping symptoms: mild subcutaneous hemorrhaging mixed with early localized plaque formation. Mathematically, the extracted features R_{avg} , G_{avg} , and B_{avg} would fall into the transitional boundaries between the 'Middle' (M) and 'High' (H) fuzzy sets.

Instead of forcing a binary decision, the ANFIS handles this ambiguity by simultaneously activating multiple fuzzy rules. In this scenario, both Rule 1 (leaning towards Streptococcosis due to elevated R) and Rule 2 (leaning towards Columnaris due to elevated G and B) would fire with intermediate strengths. The final defuzzified output aggregates these weighted probabilities. This mechanism acts as a mathematical proxy for diagnostic uncertainty, enabling the system to output a soft probability distribution that faithfully reflects the ambiguity of the transitional symptoms.

3.3.3. Particle Swarm Optimization

In the ANFIS inference process, the PSO algorithm is employed to optimize the premise and consequent parameters of the first and fourth layers. Traditional gradient-based learning methods, such as backpropagation, are susceptible to becoming trapped in local minima, particularly when navigating the complex, non-convex loss landscapes associated with limited aquaculture datasets. As a derivative-free global heuristic search algorithm, PSO effectively overcomes this limitation. By leveraging collective spatial exploration, PSO ensures robust optimization, reduces the risk of overfitting, and enhances classification stability. In the proposed method for ANFIS parameter optimization, we select the minimization of the fitness function to determine the global optimal position. After initializing the PSO parameters, the algorithm enters a loop to traverse each particle and calculate the updated fitness. Among the personal best positions of all particles, the one with the minimum fitness is selected to update the global optimum.

The pseudo-code for the PSO algorithm is shown in Algorithm 1; we encode all tunable parameters of the neuro-fuzzy system into a high-dimensional particle vector X_i . This includes the premise parameters (c and σ of the Gaussian membership functions in Layer 1) and the consequent parameters (p , q , r of the linear functions in Layer 4). For a system with N_{rules} fuzzy rules and n input features, each particle represents a candidate solution in the parameter space. During each iteration, the particles explore the search space by updating their velocities based on personal and global best positions to minimize the fitness function, which is defined as the BCE (Stable Binary Cross-Entropy) loss on the training set.

3.3.4. Binary Cross-Entropy Loss Based on OvR Strategy

While techniques such as few-shot learning or Generative Adversarial Networks are prevalent for small-sample problems, they often introduce immense computational overhead and carry a high risk of severe overfitting on specialized pathological datasets. We address the small-sample constraint through a dual approach: geometric data augmentation at the image level, and an optimized loss function at the classifier level. Specifically, we implement a numerically stable One-vs-Rest Binary Cross-Entropy strategy. Standard Cross-Entropy (CE) paired with Softmax forces output probabilities to sum to one, creating a highly coupled and complex non-convex search space. For a gradient-free algorithm like PSO operating on limited data, this complexity often leads to premature convergence or severe oscillations. By decoupling the multi-class task into independent binary sub-tasks using the OvR strategy, we reduce the dimensionality of the loss landscape for each specific disease. This decoupling allows PSO to optimize the ANFIS parameters independently

and efficiently, acting as an implicit regularization technique that prevents the model from overfitting to small-scale training samples.

Algorithm 1: PSO-Optimized ANFIS with Stable BCE Loss

Input: Training data D , Population size N , Max Iterations T_{max} , Coefficients ω, c_1, c_2

Output: Optimal ANFIS parameters θ^*

```

1: Initialize particle positions  $X$  and velocities  $V$  randomly.
2: Set personal best  $P_i \leftarrow X_i$  and global best  $G \leftarrow \operatorname{argmin} F(X_i)$ .
3: for  $t = 1$  to  $T_{max}$ , do
4:   for  $i = 1$  to  $N$ , do
5:     Decode  $X_i$  to ANFIS parameters (Premise & Consequent).
6:     Compute ANFIS output  $z$  and apply constraint:  $\hat{y} \leftarrow \sigma(\operatorname{clamp}(z))$ .
7:     Evaluate Fitness  $F_i$  using Stable Binary Cross-Entropy (OvR).
8:     if  $F_i < F(P_i)$  then  $P_i \leftarrow X_i$ .
9:     if  $F(P_i) < F(G)$  then  $G \leftarrow P_i$ .
10:   end for
11: Update Particles:
12:   for  $i = 1$  to  $N$  do
13:      $V_i \leftarrow \omega V_i + c_1 r_1 (P_i - X_i) + c_2 r_2 (G - X_i)$ 
14:      $X_i \leftarrow X_i + V_i$ 
15:   end for
16: end for
17: Return  $\theta^* \leftarrow G$ 

```

In terms of loss function design, we decompose the multi-class classification task into multiple independent binary classification sub-tasks based on the OvR strategy. This is because the output form of ANFIS allows each output to correspond to a binary sub-task without significant modifications to the model structure. Furthermore, the PSO algorithm demonstrates higher optimization efficiency for continuous, low-complexity loss functions. Therefore, for these independent binary sub-tasks, we utilize the Sigmoid function to map the raw outputs of ANFIS to probabilities. However, a standard Binary Cross-Entropy implementation is susceptible to numerical instability in this specific context. In conventional deep learning, gradient-descent optimizers update weights in small, controlled steps. In contrast, during the exploration phase of the PSO algorithm, particles pseudo-randomly search the parameter space. This stochastic behavior frequently causes the ANFIS network to generate extreme, unbounded raw output logits z . If these extreme values are fed directly into a standard Sigmoid and subsequent logarithmic BCE function, it triggers floating-point overflow ($\exp(-z) \rightarrow \infty$) or logarithmic singularities ($\log(0) \rightarrow -\infty$). Such numerical collapse produces NaN fitness values, immediately terminating the entire swarm iteration. Therefore, an explicit numerical stabilization mechanism is necessary. Our numerically stable BCE differs from the standard implementation by introducing a rigorous pre-activation clipping boundary. As defined in Equation (22), we strictly clamp the raw logits z to an empirically safe range of $[-20, 20]$. This preemptive constraint ensures that the subsequent Sigmoid probability outputs safely fall within the evaluable precision bounds. This stabilization enables PSO algorithm to accurately evaluate particle fitness even at extreme spatial coordinates. The formula can be expressed as:

$$z_{\text{clamped}} = \max(-20, \min(z, 20)), \quad (22)$$

$$s(z) = \frac{1}{1 + \exp(-z)}, \quad (23)$$

$$\text{BCE} = -\frac{1}{N} \sum_{i=1}^N [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]. \quad (24)$$

4. Experiments

To validate the theoretical framework of the Enhanced SAM and the PSO-Optimized ANFIS, this section outlines the experimental setup, detailing the public dataset, evaluation metrics, and the quantitative results obtained from comparative baselines and ablation studies.

4.1. Dataset and Pre-Processing

To verify the effectiveness of the proposed method, we utilized the Enhancing Disease Detection in Nile Tilapia dataset, which is a publicly available dataset of diseased tilapia images. It was acquired in collaboration with a local aquaculture farm near the Sirinthorn Dam. These datasets encompass a wide array of images capturing Nile Tilapia in various states of health, including both normal conditions and those affected by specific diseases. Our experiments were conducted using three categories of diseased fish image samples included in the NTD-1 subset of this dataset. Specifically, it comprises 327 images of Streptococcosis, 362 images of Columnaris Disease, and 354 images of Tilapia Lake Virus, resulting in a total of 1043 tilapia images. To establish the ground truth for segmentation and localization tasks, a hierarchical annotation protocol was implemented. All bounding boxes and pixel-level lesion masks were manually and independently generated using LabelMe by the first author, a researcher specializing in computer vision for aquatic diagnostics. To ensure objectivity and eliminate chronological bias, a random 20% subset of the dataset was blindly re-annotated by the same researcher after a two-week interval, yielding the mean Intersection over Union (mIoU) of 90.7%, which demonstrates temporal stability. The annotation protocol strictly followed these criteria: only explicitly visible pathological features were bounded to avoid subjective over-segmentation; for disconnected multiple lesions on a single fish body, independent instance masks were delineated; and ambiguous transitional boundaries between diseased and healthy tissues were conservatively excluded to minimize false positives. Finally, the review of all annotations was performed by the corresponding author, a senior expert in intelligent aquaculture and computer vision applications in our laboratory. Any inaccurate delineations were returned for mandatory revision to guarantee biological accuracy.

To ensure a rigorous evaluation, we implemented a Subject-Level Stratified 5-Fold Cross-Validation strategy on the 1043 images. We partitioned the dataset based on independent subject IDs (individual fish) and acquisition sequences. This process ensures that multiple images originating from the same fish are grouped into the same fold. In each of the five iterations, three folds (60%) were allocated for training, one fold (20%) was used for validation, and the remaining fold (20%) was reserved for independent testing.

Following this data split, we applied the Multi-Scale Retinex with Color Restoration (MSRCR) pipeline. Images captured in real-world aquaculture environments often suffer from issues such as turbidity, color distortion, and low contrast. To address these challenges and provide high-quality inputs for the subsequent SAM module, we incorporated a specialized image pre-processing step MSRCR. The core principle of Retinex theory is to decompose an image into illumination and reflectance components, aiming to estimate the true reflectance of objects by minimizing illumination interference.

For a given image, the Multi-Scale Retinex (MSR) output for the i -th color channel ($i \in \{R, G, B\}$) is computed as the weighted sum of Single-Scale Retinex (SSR) out-

puts across N different scales (typically $N = 3$, representing small, medium, and large Gaussian surrounds):

$$R_{MSR_i}(x, y) = \sum_{n=1}^N w_n [\log I_i(x, y) - \log (F_n(x, y) * I_i(x, y))], \quad (25)$$

where $I_i(x, y)$ is the intensity of the input image, w_n is the weight of the n -th scale, $*$ denotes the convolution operation, and $F_n(x, y)$ represents the Gaussian surround function at scale n .

However, standard MSR tends to cause color distortion and unnatural color shifts (often manifesting as a “graying out” effect), which are particularly detrimental to identifying minute disease lesions. To address this, MSRCR introduces a Color Restoration Function (CRF), denoted as $C_i(x, y)$, to adjust the proportional relationship among the color channels dynamically:

$$C_i(x, y) = \beta \log \left(\alpha \frac{I_i(x, y)}{\sum_{j \in \{R, G, B\}} I_j(x, y)} \right), \quad (26)$$

where α is a nonlinear adjustment control parameter and β is the color restoration gain constant. The final MSRCR output is formulated by multiplying the MSR result by the color restoration factor:

$$R_{MSRCR_i}(x, y) = G \cdot [C_i(x, y) \cdot R_{MSR_i}(x, y) + b], \quad (27)$$

where G and b denote the global gain and offset parameters, respectively, mapping the processed data back to the standard displayable dynamic range. By executing this MSRCR pipeline, we effectively compensated for non-uniform illumination and water-induced color attenuation, thereby significantly improving local contrast and restoring the true color representation of the minute lesions.

Finally, we applied “padding-to-square” operation and proportional resizing to 1024×1024 pixels to adapt the data for the SAM input. Furthermore, to enhance model robustness, geometric data augmentation (including horizontal/vertical flips, random rotations, and random shearing) was applied exclusively to the training folds. The validation and test sets remained completely independent and contained only original unaugmented images. All final performance metrics reported in this study represent the average across all five independent test folds, ensuring the true assessment of the model’s performance on unseen data.

4.2. Evaluation Metrics

(1) Dice Similarity Coefficient (DSC)

Since tilapia infection regions may be small in area or possess irregular edges, the Dice coefficient can more accurately reflect the alignment accuracy of the segmentation. Where A represents the ground truth and B represents the prediction results, the formula is defined as:

$$Dice = \frac{2|A \cap B|}{|A| + |B|}. \quad (28)$$

(2) Precision (P)

Precision represents the ratio of correctly predicted positive instances to all instances predicted as positive.

$$Precision = \frac{TP}{TP + FP}. \quad (29)$$

(3) Recall (R)

Recall represents the ratio of correctly predicted positive instances to all actual positive instances.

$$Recall = \frac{TP}{TP + FN}. \quad (30)$$

(4) F1-Score (F)

F1-Score is the harmonic mean of Precision and Recall; it balances the risk of false detections, ensuring both the comprehensiveness and precision of the classification.

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}. \quad (31)$$

(5) Accuracy (A)

Accuracy directly measures the proportion of correct classifications among all predictions. It effectively reflects the overall classification performance of the ANFIS model and serves as the most intuitive core metric for fish disease identification.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \quad (32)$$

(6) Brier Score (BS)

Brier Score measures the mean squared difference between the predicted probability and the actual one-hot encoded label on samples. A lower BS indicates that the model's probabilistic predictions are highly calibrated.

$$BS = \frac{1}{N} \sum_{i=1}^N (y_i - p_i)^2. \quad (33)$$

(7) Computational Complexity Metrics

Parameters (Params) measure the total number of trainable variables, reflecting the model's storage requirements and memory footprint. Floating Point Operations (GFLOPs) quantify the computational cost of a single forward pass, indicating the theoretical execution complexity. Inference Time (T_{inf}) measures the average latency for processing a single image.

4.3. Experimental Details

In our experiments, the hardware and software specifications for the training environment included Intel Core i5-12600KF 3.7 GHz CPU, 16 GB RAM, GPU (NVIDIA GeForce RTX 4060 Ti (NVIDIA, Santa Clara, CA, USA)), Python 3.9, and PyTorch 1.10.0. All random seeds across PyTorch frameworks were globally fixed to 42. The parameters were set as follows: For SAM, we selected the ViT-Huge as the backbone network and utilized its pre-trained weights. For the prompter component, the RPN hyperparameter definitions followed Faster R-CNN, and the number of generated box prompt embedding vectors was set to 4. During training, the original SAM image encoder and prompt encoder were frozen, while the SAM mask decoder and our prompt generation branch (comprising the Feature Adapter and the RPN) participated in the training process. The input image size was 1024×1024 . We adopted the AdamW optimizer with an initial learning rate of 1×10^{-4} and a batch size of 4, and trained for 200 epochs using the comprehensive loss function. For

the ANFIS part, hyperparameters were determined by defining a search space and utilizing Bayesian optimization to sample different combinations. We selected the Tree-structured Parzen Estimator as the surrogate model and Expected Improvement as the acquisition function. The optimal hyperparameter combination was evaluated and selected after training validation, with final values shown in Table 1; in PSO implementation, the population size N_{pop} was set to 44, and the maximum number of iterations T_{max} was 200, as determined by Bayesian optimization. The optimization process was governed by specific convergence criteria: the algorithm terminates either upon reaching the maximum iterations T_{max} or if the absolute change in the global best fitness value remains below a threshold of 1×10^{-6} for 10 consecutive iterations. The cognitive c_1 and social c_2 coefficients were respectively set to 1.483 and 1.880. The training process was based on PSO using binary cross-entropy loss for 200 epochs.

Table 1. Bayesian optimization results.

Parameter Category	Symbol	Parameter	Search Space	Optimal Value
PSO Parameters	N_{pop}	Swarm size	[20, 100]	44
	w	Inertia weight	[0.4, 0.9]	0.719
	c_1	Cognitive coefficient	[1.0, 2.5]	1.483
	c_2	Social coefficient	[1.0, 2.5]	1.880
ANFIS Parameters	$N_{cluster}$	FCM cluster number	[3, 8]	5
	m	Fuzzy exponent	[1.1, 2.5]	1.7
	N_{rules}	Fuzzy rule number	[10, 30]	18

4.4. Segmentation Result

We selected several state-of-the-art instance segmentation methods for testing and comparison with the proposed Enhanced SAM. These include specialized models for biomedical image segmentation such as U-Net [11], DeepLabv3+ [16], and the related improved model TransU-Net [32]; classic instance segmentation models Mask R-CNN [14] and Mask2Former [33]; and SAM-related variants Original SAM [17] and HQ-SAM [34]. All baseline models (both segmentation and classification) were trained and evaluated using the same Stratified 5-Fold Cross-Validation data splits, identical MSRCR preprocessing pipeline, and uniform geometric data augmentation policies; the hyperparameters were optimized based on the validation folds.

Table 2 presents the comparative results of different methods on the dataset, demonstrating the superior performance of the proposed framework across evaluation metrics. Specifically, the model recorded segmentation mask Dice coefficients of 86.0%, 87.7%, and 84.9% for Streptococcosis, Columnaris, and Tilapia Lake Virus, respectively (averaging 86.2%). To further confirm its stability across the 5 independent test folds, as shown in Table 3, the low standard deviation ($\pm 0.9\%$) across the 5 independent test folds underscores the model's stability. This validates the necessity of introducing large vision models into biological pathology domains and highlights the efficacy of our architecture in mitigating complex background interference.

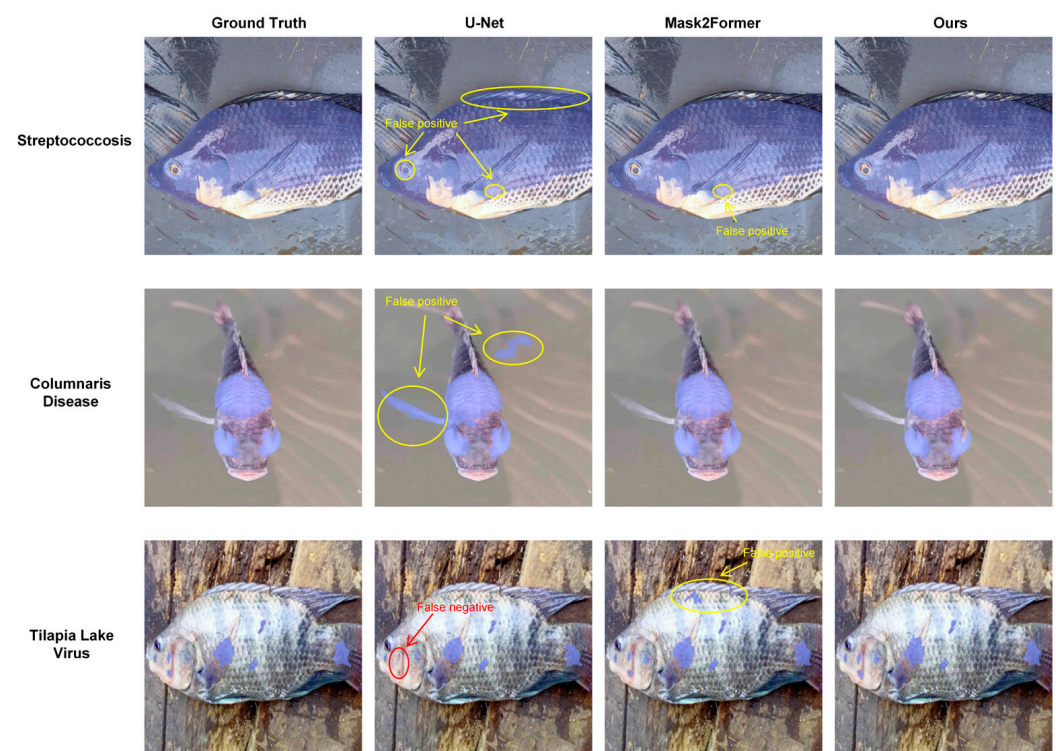
Additionally, Figure 6 displays partial segmentation instances of samples under different methods, showing that the segmentation results of the Enhanced SAM are more accurate and closer to the ground truth masks.

Table 2. Segmentation results: comparison of proposed method with SOTA model.

Method	Dice (%)			
	Str	Col	TiLV	Avg
U-Net	69.3	73.4	71.2	71.3
TransU-Net	80.7	82.8	79.3	80.9
DeepLabV3+	78.5	77.6	75.5	77.2
Mask R-CNN	80.3	76.5	74.1	77.0
Mask2Former	82.8	83.3	84.3	83.5
SAM	72.3	78.4	70.9	73.9
HQ-SAM	85.1	81.4	83.9	83.5
Proposed	86.0	87.7	84.9	86.2

Table 3. Per-fold segmentation results of proposed method.

Fold	Dice (%)			
	Str	Col	TiLV	Avg
Fold 1	84.8	86.8	83.5	85.0
Fold 2	85.4	87.3	84.2	85.6
Fold 3	85.9	87.7	84.8	86.1
Fold 4	86.7	88.1	85.6	86.8
Fold 5	87.2	88.6	86.4	87.4
Mean	86.0	87.7	84.9	86.2
Std ($\pm$)	0.9	0.7	1.1	0.9

**Figure 6.** Comparison of visualized segmentation results.

4.5. Classification Result

We compared the PSO-optimized ANFIS with various classification models, including deep learning-based models such as ResNet50, Swin Transformer, and MLP, as well as traditional machine learning classifiers like SVM, Random Forest, and XGBoost. To ensure a methodologically fair evaluation, the traditional machine learning classifiers (SVM, Random Forest, XGBoost, and MLP) and our proposed ANFIS were

trained on the low-dimensional mean RGB features extracted from the lesion masks. The deep learning architectures (ResNet-50 and Swin Transformer) were not restricted to low-dimensional features; they were trained on the 2D lesion masks outputted by the segmentation module.

The experimental results, as shown in Table 4, indicate that the PSO-optimized AN-FIS achieved a high level of performance in Precision, Recall, F1-score, and Accuracy. Specifically, the proposed method achieved an overall Precision of 92.8%, Recall of 92.5%, F1-score of 92.6%, Accuracy of 93.5%, and Brier Score of 0.065. In Table 5, the standard deviations were $\pm 0.8\%$ for Accuracy, $\pm 1.0\%$ for Precision, $\pm 1.2\%$ for Recall, and $\pm 1.1\%$ for F1-Score.

Table 4. Classification results: comparison of proposed method with SOTA model.

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Brier Score
SVM	89.2	88.7	88.5	88.6	0.095
Random Forest	88.5	88.1	87.8	87.9	0.102
XGBoost	91.8	91.5	91.2	91.3	0.078
MLP	87.4	86.8	87.1	86.9	0.115
ResNet-50	90.3	89.9	90.1	90.0	0.088
Swin Transformer	91.5	91.1	91.4	91.2	0.082
Proposed	93.5	92.8	92.5	92.6	0.065

Table 5. Per-fold classification results of proposed method.

Fold	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Fold 1	92.3	91.5	90.9	91.2
Fold 2	92.8	92.1	91.8	91.9
Fold 3	93.8	92.8	92.6	92.7
Fold 4	94.2	93.5	93.3	93.4
Fold 5	94.2	94.1	93.9	94.0
Mean	93.5	92.8	92.5	92.6
Std ($\pm$)	0.8	1.0	1.2	1.1

Furthermore, Figure 7 illustrates the per-class confusion matrix derived from the independent test folds. The matrix reveals that while the model achieves a high overall accuracy of 93.5%, misclassifications primarily occur between specific disease pairs with transitional or overlapping clinical manifestations. For instance, the model occasionally misclassifies early-stage Streptococcosis (presenting with mild erythema) as Columnaris. Similarly, advanced Tilapia Lake Virus (characterized by dark necrotic skin) and shadows in Columnaris lesions can cause minor diagnostic confusion.

4.6. Computational Complexity Analysis

In this section, we provide a quantitative analysis of the computational cost to assess the trade-off between identification accuracy and efficiency. Table 6 compares the proposed method framework with several baseline models.

The Stage 1 segmentation module, driven by the ViT-Huge backbone, accounts for virtually all the computational footprint (~638.2 M total parameters and 2614.7 GFLOPs). However, by adopting a fine-tuning strategy—freezing the heavy image encoder and updating only the custom Feature Adapter, ROI Prompt Encoder, and mask decoder—the actual trainable parameters are aggressively reduced to approximately 5.8 M.

Conversely, the Stage 2 classification module proves to be lightweight. Unlike deep neural network classifiers that require millions of dense weight matrices, the neuro-fuzzy system operates on a concise set of logical rules and membership functions. It contains

fewer than 10,000 parameters (<0.01 M) and contributes a negligible 0.2 ms to the total inference latency.

Table 6. Comparison of computational complexity with SOTA model.

Method	Total Params (M)	GFOPs	T_{inf} (ms)	Dice/Acc (%)
TransU-Net	105.3	340.8	18.7	80.9/-
Mask2Former	107.0	265.4	25.4	83.5/-
HQ-SAM	634.5	2621.3	49.5	83.5/-
Proposed (Segmentation)	638.2	2614.7	48.1	86.2/-
SVM	<0.01	<0.01	0.5	-/89.2
XGBoost	<0.01	<0.01	0.8	-/91.8
Proposed (Classification)	<0.01	<0.01	0.2	-/93.5
Proposed (Total)	638.2	2614.7	48.3	86.2/93.5

		Predicted Class		
		Streptococcosis	Columnaris Disease	Tilapia Lake Virus
True Class	Streptococcosis	305	18	4
	Columnaris Disease	12	338	12
	Tilapia Lake Virus	6	16	332

Figure 7. Confusion matrix on 5-Fold test sets.

Despite the higher total latency compared to SOTA models, the pipeline maintains an overall inference speed of approximately 20 FPS (48.3 ms per image) on an NVIDIA RTX 4060 Ti GPU. It demonstrates that the framework successfully offers a balance between accuracy and efficiency.

4.7. Ablation Experiments

4.7.1. Ablation Study on Segmentation

In the segmentation module, we utilized the original SAM (employing the ViT-Huge backbone with pre-trained weights) as the baseline model. We individually integrated the CBAM-based Feature Adapter (FA) and the RPN-based ROI Prompt Encoder (RPE) to perform a comparative analysis. Table 7 presents the ablation results for the segmentation module. “Baseline (SAM)” denotes the exclusive use of the original SAM; due to the absence of automated prompts targeting specific disease regions, we employed a

conventional grid point prompt strategy to guide the segmentation process. “SAM + FA” refers to the incorporation of the CBAM-based Feature Adapter into the baseline model. This module enhances the model’s focus on the minute features of tilapia lesions and suppresses background noise through channel and spatial attention mechanisms. “SAM + RPE” indicates the introduction of the ROI Prompt Encoder to the baseline. This module utilizes the RPN to automatically generate high-quality candidate box prompts, replacing manual or grid prompts. “Proposed” represents the complete segmentation model proposed in this paper, which simultaneously integrates feature adapter and automated prompt generation.

Table 7. Ablation study results of the segmentation module.

Method	Feature Adapter (FA)	ROI Prompt Encoder (RPE)	Dice (%)
Baseline (SAM)	×	×	73.9
SAM + FA	✓	×	78.5
SAM + RPE	×	✓	80.3
Proposed	✓	✓	86.2

The analysis results indicate a significant improvement in the Dice coefficient after the introduction of the FA. This demonstrates that the CBAM effectively refined segmentation boundaries through multi-scale feature fusion. Specifically, for lesions with colors similar to the background of the fish body (such as the erythema of early Streptococcosis), the spatial attention mechanism played a crucial focusing role. Furthermore, the increase in the Dice coefficient after introducing the RPE indicates that the RPN-based automatic prompt generation mechanism can precisely localize lesion regions. When both FA and RPE were combined, the model achieved optimal segmentation quality. This suggests that the two modules are functionally complementary, jointly realizing high-precision segmentation of tilapia disease regions.

4.7.2. Ablation Study on Classification

In the classification module, we validated the superiority of the PSO algorithm compared to the traditional ANFIS training method (the hybrid learning algorithm, combining Least Squares Estimation with Gradient Descent), as well as the impact of the improved loss function. The input data consisted of color pixel intensity features extracted following segmentation by the Enhanced SAM. The experiments adopted Precision, Recall, F1-score, and Accuracy as evaluation metrics. Table 8 presents the ablation study results for the classification module. Specifically, “Standard ANFIS” utilizes traditional backpropagation gradient descent to adjust premise parameters and the Least Squares Estimation method to adjust consequent parameters. “ANFIS + PSO” introduces the Particle Swarm Optimization algorithm for parameter tuning but employs standard Cross-Entropy loss. “Proposed” represents the complete classification scheme proposed in this article.

Table 8. Ablation study results of the classification module.

Method	Optimizer	Loss Function	A (%)	P (%)	R (%)	F1 (%)
Standard ANFIS	Gradient Descent	Standard CE	85.9	84.5	83.1	83.8
ANFIS + PSO	PSO	Standard CE	91.1	91.2	91.8	91.5
Proposed	PSO	Stable BCE (OvR)	93.5	92.8	92.5	92.6

The analysis results indicate a significant improvement in accuracy after introducing the PSO algorithm compared to the standard ANFIS. This is attributed to the relatively limited sample size of the dataset in this study, where traditional gradient descent methods are prone to getting trapped in local optima within the non-convex fuzzy inference parameter space, in contrast, PSO can more effectively seek global optimal parameter solutions through the cooperative evolution of the particle swarm. Additionally, the employment of an improved numerically stable loss function further fine-tuned the model's performance. Utilizing the Standard CE loss with PSO yielded an accuracy of 91.1%; in contrast, replacing it with the proposed Stable BCE (OvR) loss boosted the accuracy to 93.5%, alongside consistent improvements in Precision, Recall, and F1-score. This empirical performance margin demonstrates that the OvR-based loss function is more suitable for the PSO-ANFIS architecture. By mitigating cross-class competition during optimization, the OvR strategy provides a smoother optimization gradient and establishes superior classification boundaries for small sample sets.

4.7.3. Impact of Segmentation on Classification Performance

To further verify the necessity of the segmentation module within the overall identification network and to quantify the specific contribution of prior segmentation to subsequent feature extraction and classification performance, we designed this experiment. It aims to investigate the difference in disease identification accuracy between removing the segmentation prior (using whole-image features) and retaining the segmentation prior (using ROI features). In the experiment, the classifier was maintained as the PSO-Optimized ANFIS model proposed in this paper, and only the source of input feature extraction was altered. The experimental results are shown in Table 9. “Global Features (*w/o* SAM)” does not utilize SAM for segmentation; instead, the mean pixel values of the R, G, B channels are calculated directly from the entire original image (including water background, fish body, and lesions) as feature inputs. “ROI Features (Proposed)” utilizes the mask generated by the Enhanced SAM to calculate the mean pixel values of the R, G, B channels specifically within the lesion region as feature inputs.

Table 9. Comparison of classification performance with and without segmentation prior.

Feature Source	Segmentation Prior	A (%)	P (%)	R (%)	F1 (%)
Global Features (<i>w/o</i> SAM)	×	70.2	70.0	68.4	69.2
ROI Features (Proposed)	✓	93.5	92.8	92.5	92.6

From the data in the table, it can be observed that in the absence of a segmentation prior, the model's classification performance declined significantly, with an accuracy of only 70.2%, which is far lower than the 93.5% achieved by the method proposed in this paper. This performance disparity is primarily attributed to the feature dilution effect and interference from background noise. Tilapia aquaculture environments typically contain complex backgrounds with varying water bodies and lighting conditions. When extracting whole-image features, these high-proportion background pixel values dominate the calculation of color means, causing the weak color signals of the lesion regions to be drowned out. After introducing the Enhanced SAM segmentation, the model is able to precisely strip away the background and focus solely on the diseased tissue. It reduces the intra-class variance of the feature regions and enhances feature separability, thereby enabling the ANFIS classifier to achieve high-precision disease diagnosis.

In summary, the ablation experiments demonstrate that each improvement module proposed in this paper has made a positive contribution to the final tilapia disease

identification performance, validating the scientific validity and necessity of the joint architecture comprising the Enhanced SAM front-end segmentation and the optimized ANFIS back-end classification.

5. Discussion

The experimental results demonstrate the efficacy of the proposed two-stage framework. To fully contextualize these findings, this section provides a deeper analysis of the model's performance relative to existing methods, examines the impact of complex environments, and critically assesses the inherent limitations of the current approach.

5.1. Performance Comparison and Contextualization

Our proposed method, integrating an Enhanced SAM with a PSO-optimized ANFIS, demonstrates the superiority over existing methodologies in fish disease identification. Traditional CNN models such as U-Net and DeepLabV3+ are limited by local receptive fields, making it difficult to handle scenarios with low contrast between lesions and the background. Although TransU-Net incorporates Transformer models, its performance remains insufficient when dealing with minute or blurred lesions due to the lack of targeted "prompt guidance". Mask R-CNN is a typical two-stage "detection-then-segmentation" method; however, its mask branch defaults to outputting low-resolution soft masks which are then upsampled back to ROI dimensions. This process tends to lose high-frequency details along fine edges during downsampling, resulting in an average Dice coefficient of only 77.0%. Although Mask2Former also demonstrates excellent performance, it still trails our method by 2.7% when handling minute lesions or those with blurred edges. This indicates that while general foundation models possess powerful feature extraction capabilities, efficient fine-tuning for specific biomedical tasks is an effective approach to achieving optimal precision.

The average Dice of the original SAM was only 73.9%; it is primarily attributed to the original SAM's lack of understanding of the specific semantic concept of "disease" in the absence of precise prompts, leading it to easily misidentify water ripples, fish fins, or light spots as foreground. In contrast, our method achieved a massive performance leap of 12.3% by introducing the RPN prompt encoder. This provides evidence that the prompt boxes generated by the RPN offer critical localization priors for SAM, solving the "cold-start" problem where general foundation models fail to locate targets in specific domains. Although HQ-SAM is an advanced variant optimized for SAM segmentation quality, our method still outperformed it by 2.7%. This advantage is attributed to the CBAM feature adapter designed in this study. While HQ-SAM improves mask edge quality, it still relies on generic ViT features. Conversely, our feature adapter specifically learns the unique color and texture features of tilapia lesions through channel and spatial attention mechanisms, thereby performing better in terms of internal lesion consistency and edge alignment.

To further justify the selection of the PSO-optimized ANFIS over modern deep neural network (DNN) classifiers, several critical factors inherent to the tilapia disease identification task must be considered.

First, the risk of overfitting is paramount. While DNNs excel at learning high-dimensional representations from massive datasets, they are prone to overfitting when applied to niche pathological datasets with limited samples. In our experiments, Swin Transformer and ResNet-50 achieved accuracies of 91.5% and 90.3% respectively, notably lower than the 93.5% achieved by our method. Rather than extracting clinical biomarkers,

these heavily parameterized DNNs tend to over-optimize on the training set. Consequently, their generalization ability on unseen data is compromised.

Second, the neuro-fuzzy paradigm provides superior interpretability. Modern deep classifiers function as black boxes, offering little insight into the decision-making logic. MLP (87.4%), when processing low-dimensional features (only R/G/B means), is prone to falling into local optima and lacks a clear logical reasoning rule structure like ANFIS; in contrast, ANFIS maps low-dimensional, clinically relevant color features into a structured set of “If-Then” fuzzy rules. This mirrors the diagnostic process of human pathologists, who identify diseases based on the intensity and distribution of symptoms like erythema or necrotic erosion.

Finally, the imprecision of biological symptoms necessitates a “soft computing” approach. Fish diseases do not manifest as discrete, hard-coded states; they exist on a continuum of symptom severity. DNNs typically utilize a Softmax layer to force a hard probability distribution, which may struggle with the overlapping feature spaces of early-stage infections. Similarly, XGBoost (91.8%) and SVM (89.2%) are both based on hard decision boundaries for classification. Beyond hard-classification metrics, the Brier Score analysis reveals the distinct advantage of the proposed framework in handling diagnostic uncertainty. Swin Transformer (BS = 0.082) and ResNet-50 (BS = 0.088) exhibit significant calibration errors, meaning they tend to generate overconfident but incorrect predictions when processing ambiguous transitional lesions. Conversely, the proposed PSO-optimized ANFIS achieved lower Brier Score of 0.065. It quantitatively proves that the fuzzy inference mechanism maintains well-calibrated confidence, avoiding serious misclassifications in complex aquatic environments.

5.2. Limitations and Future Perspectives

Although the proposed network achieves a relatively high diagnostic accuracy in experimental settings, it retains certain limitations regarding its transition toward industrial application and deployment in complex environments. The following directions warrant further investigation:

(1) Lightweight model architecture and edge deployment

To maximize the precision of lesion mask edge delineation, the current segmentation network retains ViT-Huge, a parameter-intensive architecture, as the image encoder backbone. While this guarantees depth in feature representation, it inevitably increases computational overhead and memory consumption, rendering direct deployment on computationally constrained edge Internet of Things (IoT) devices in aquaculture farms highly challenging. Future research could incorporate lightweight techniques, such as knowledge distillation, or substitute the backbone network with lightweight visual model variants explicitly designed for mobile platforms. Trading a negligible loss in accuracy for accelerated inference speeds and reduced computational costs remains critical for enabling real-time, on-site early warning systems in aquaculture.

(2) Generalization validation and few-shot learning

A critical limitation of the current study is the reliance on a single dataset. We acknowledge that the high classification accuracy achieved by the ANFIS module heavily reflects the dataset-specific color separability of the selected diseases. Because the model predominantly leverages these mean RGB values, its robust performance cannot guarantee equivalent diagnostic generalization across independent cohorts. Variations in water turbidity or different geographical farming conditions could alter the baseline color distribution, potentially leading to domain shift failures. Future work could integrate more comprehensive features (such as microscopic textural patterns and morphological geometry) to

reduce the model's over-reliance on color distributions, and validate the architecture on independent multi-center datasets. Furthermore, deeply integrating domain adaptation or few-shot learning paradigms into the existing framework will endow the model with strong transferability, enabling it to rapidly fine-tune and adapt to novel farm environments utilizing only a minimal number of newly acquired samples. This Few-Shot Learning could be implemented in two stages. In the vision stage, few-shot visual prompting could be employed to fine-tune the ROI Prompt Encoder, enabling the network to establish accurate spatial priors for novel lesion morphologies utilizing limited support images. Concurrently, in the diagnostic stage, the rules of the ANFIS classifier would be dynamically adapted. Rather than retraining the entire PSO swarm from scratch, the system would utilize the feature distributions of the few-shot support set to dynamically shift the Gaussian membership function centers and recalibrate the consequent parameters, thereby achieving rapid diagnostic transferability.

(3) Multi-modal data fusion for comprehensive health status perception

Currently, the disease diagnostic decision-making in this study relies exclusively on the visual color and texture features derived from static two-dimensional images. However, in actual biological manifestation, early indicator of infection is frequently not confined to static structural lesions on the body surface; it is often preceded by dynamic behavioral alterations. Future intelligent fishery monitoring systems should transcend the limitations of a single visual modality by incorporating underwater video stream sensor data. Research could explore the application of three-dimensional convolutional neural networks (3D-CNNs) or spatial-temporal graph convolutional networks (ST-GCNs) to extract the continuous motion trajectories and behavioral time-series features of fish schools. Fusing these dynamic temporal features with the static localized lesion features established in this study into a spatiotemporally coordinated, multidimensional perception network will significantly enhance the system's proactive warning capabilities and overall fault tolerance. To address the practical challenge of tracking individual fish across consecutive video frames, Multi-Object Tracking algorithms could be implemented prior to behavioral feature extraction to assign and maintain consistent temporal IDs. Regarding the architectural integration, we plan to adopt a decision-level fusion strategy. Specifically, the static image pipeline and the temporal video pipeline will evaluate the targets in parallel. Their independent diagnostic probabilities will then be fused using an adaptive weighting module. This parallel processing strategy is viable, as it ensures that the rule-based interpretability of our current static system is fully preserved while leveraging dynamic behavioral data for final diagnostic calibration.

(4) Hybrid optimization algorithms and improved convergence efficiency

Within the classification network, although the Particle Swarm Optimization algorithm successfully enabled the ANFIS network to circumvent local optima in the parameter space and achieve high classification accuracy, the standard PSO algorithm still encounters bottlenecks—such as constrained convergence speeds and a loss of population diversity during late-stage iterations—when navigating substantially larger parameter spaces. Subsequent research could consider introducing adaptive inertia weight mechanisms into the PSO framework or hybridizing it with other optimization techniques, such as Genetic Algorithms (GA) or Simulated Annealing (SA). Striking a more optimal balance between broad global exploration and deep local exploitation will further minimize the time costs associated with model training on servers.

6. Conclusions

This study introduces an innovative two-stage “segmentation–classification” framework that addresses the persistent challenges of complex backgrounds, minute lesions, and ambiguous symptoms in tilapia disease identification. In the segmentation phase, by introducing a CBAM-based feature adapter and an RPN-based ROI Prompt Encoder, we successfully transform the general-purpose SAM into a domain-specific model equipped with automated prompt generation capabilities, achieving precise extraction of disease regions. In the classification phase, utilizing an ANFIS network optimized by the PSO algorithm, combined with a numerically stable OvR binary cross-entropy loss function, we realized high-precision disease diagnosis based on lesion color features within the evaluated dataset bounds.

Experimental results confirm that the proposed Enhanced SAM outperforms SOTA models like Mask2Former with a Dice coefficient of 86.2%. Concurrently, the PSO-ANFIS classifier achieved a superior accuracy of 93.5%, outperforming prominent deep learning and machine learning baselines. Ablation studies validate the critical contributions of the self-generated prompts and the neuro-fuzzy logic. Ultimately, this hybrid framework provides a robust, interpretable, and computationally viable foundational model for automated aquaculture monitoring systems.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, methodology, validation, investigation, data curation, writing—original draft preparation, writing—review and editing: Y.G.; supervision, project administration: G.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data available in a publicly accessible repository. The original data presented in the study are openly available in kaggle at <https://www.kaggle.com/datasets/engineeringubu/enhancing-disease-detection-in-nile-tilapia> (accessed on 23 March 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Dongyu, Q. 2024 the state of world fisheries and aquaculture-blue transformation in action. In *The State of World Fisheries and Aquaculture*; Food and Agriculture Organization of the United Nations: Rome, Italy, 2024; pp. 1–232.
2. Waleed, A.; Medhat, H.; Esmail, M.; Osama, K.; Samy, R.; Ghanim, T.M. Automatic recognition of fish diseases in fish farms. In Proceedings of the 2019 14th International Conference on Computer Engineering and Systems (ICCES), Cairo, Egypt, 17 December 2019; pp. 201–206.
3. Cho, K.-W.; Kim, T.; Na, I.S. A Deep Convolutional Neural Network for Multi-class Fish Disease Classification using VGGNet-19 Architecture. *Indian J. Anim. Res.* **2025**, *60*, 981–988. [\[CrossRef\]](#)
4. Gupta, A.; Bringsdal, E.; Knausgård, K.M.; Goodwin, M. Accurate wound and lice detection in Atlantic salmon fish using a convolutional neural network. *Fishes* **2022**, *7*, 345. [\[CrossRef\]](#)
5. Khabusi, S.P.; Huang, Y.-P.; Lee, M.-F.; Tsai, M.-C. Enhanced U-Net and PSO-Optimized ANFIS for Classifying Fish Diseases in Underwater Images. *Int. J. Fuzzy Syst.* **2024**, *26*, 2518–2535. [\[CrossRef\]](#)
6. Yasruddin, M.L.; Ismail, M.A.H.; Husin, Z.; Tan, W.K. Feasibility study of fish disease detection using computer vision and deep convolutional neural network (dcnn) algorithm. In Proceedings of the 2022 IEEE 18th International Colloquium on Signal Processing & Applications (CSPA), Selangor, Malaysia, 12 May 2022; pp. 272–276.
7. Ouyang, C.; Peng, H.; Tan, M.; Yang, L.; Deng, J.; Jiang, P.; Hu, W.; Wang, Y. YOLO-TPS: A multi-module synergistic high-precision fish-disease detection model for complex aquaculture environments. *Animals* **2025**, *15*, 2356. [\[CrossRef\]](#) [\[PubMed\]](#)
8. Huang, Y.-P.; Khabusi, S.P. A CNN-OSELM multi-layer fusion network with attention mechanism for fish disease recognition in aquaculture. *IEEE Access* **2023**, *11*, 58729–58744. [\[CrossRef\]](#)

9. Wang, Z.; Ruan, Z.; Chen, C. DyFish-DETR: Underwater fish image recognition based on detection transformer. *J. Mar. Sci. Eng.* **2024**, *12*, 864. [\[CrossRef\]](#)
10. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
11. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
12. Zhou, Z.; Rahman Siddiquee, M.M.; Tajbakhsh, N.; Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In Proceedings of the International Workshop on Deep Learning in Medical Image Analysis, Granada, Spain, 20 September 2018; pp. 3–11.
13. Xiao, Z.; Liu, B.; Geng, L.; Zhang, F.; Liu, Y. Segmentation of lung nodules using improved 3D-UNet neural network. *Symmetry* **2020**, *12*, 1787. [\[CrossRef\]](#)
14. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
15. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [\[CrossRef\]](#) [\[PubMed\]](#)
16. Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
17. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.-Y. Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–3 October 2023; pp. 4015–4026.
18. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *60*, 84–90. [\[CrossRef\]](#)
19. Lester, B.; Al-Rfou, R.; Constant, N. The power of scale for parameter-efficient prompt tuning. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, 7–11 November 2021; pp. 3045–3059.
20. Jia, M.; Tang, L.; Chen, B.-C.; Cardie, C.; Belongie, S.; Hariharan, B.; Lim, S.-N. Visual prompt tuning. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 709–727.
21. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 8748–8763.
22. Zhou, K.; Yang, J.; Loy, C.C.; Liu, Z. Learning to prompt for vision-language models. *Int. J. Comput. Vis.* **2022**, *130*, 2337–2348. [\[CrossRef\]](#)
23. Chen, K.; Liu, C.; Chen, H.; Zhang, H.; Li, W.; Zou, Z.; Shi, Z. RSPrompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4701117. [\[CrossRef\]](#)
24. Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; Neubig, G. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.* **2023**, *55*, 1–35. [\[CrossRef\]](#)
25. Jang, J.-S. ANFIS: Adaptive-network-based fuzzy inference system. *IEEE Trans. Syst. Man Cybern.* **1993**, *23*, 665–685. [\[CrossRef\]](#)
26. Huang, Y.-P.; Wang, T.-H.; Basanta, H. Using fuzzy mask R-CNN model to automatically identify tomato ripeness. *IEEE Access* **2020**, *8*, 207672–207682. [\[CrossRef\]](#)
27. Yang, J.; Chen, T.; Chen, L.; Zhao, J. Towards A Clustering Guided Rule Interpolation for ANFIS Construction. In Proceedings of the 2024 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Yokohama, Japan, 30 June 2024–5 July 2024; pp. 1–6.
28. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *39*, 1137–1149. [\[CrossRef\]](#) [\[PubMed\]](#)
29. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
30. Xia, X.; Gui, L.; Yu, F.; Wu, H.; Wei, B.; Zhang, Y.-L.; Zhan, Z.-H. Triple archives particle swarm optimization. *IEEE Trans. Cybern.* **2019**, *50*, 4862–4875. [\[CrossRef\]](#)
31. Cheung, N.J.; Ding, X.-M.; Shen, H.-B. OptiFel: A convergent heterogeneous particle swarm optimization algorithm for Takagi–Sugeno fuzzy modeling. *IEEE Trans. Fuzzy Syst.* **2013**, *22*, 919–933.
32. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306.

33. Cheng, B.; Misra, I.; Schwing, A.G.; Kirillov, A.; Girdhar, R. Masked-attention mask transformer for universal image segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 1290–1299.
34. Ke, L.; Ye, M.; Danelljan, M.; Tai, Y.-W.; Tang, C.-K.; Yu, F. Segment anything in high quality. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 29914–29934. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.