ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 3 - 2026

Intelligent Recognition and classification of UAV floating objects in Rivers and Lakes based on deep Learning

Yin Zhang[1], *Weijia Han*[2], *Liqun He*[2] *and Fei Chen*[2,*]

[1] Shunjiangyuan Provincial Nature Reserve Management Center (Tangpu Reservoir Management Center), Shaoxing, 312000, Zhejiang, China
[2] Shaoxing Water Conservancy and Hydropower Survey and Design Institute Co., Ltd, Shaoxing, 312000, Zhejiang, China

SUMMARY: *The detection of floating objects in rivers and lakes based on unmanned aerial vehicle (UAV) is the basis for automatic water surface monitoring and fine-grained environmental monitoring. Manual inspection and traditional image processing methods are limited by complex ripples, shore shadows, small target scales, and unstable imaging angles. This paper proposes a multi-source deep learning framework for intelligent recognition and classification of floating objects in UAV river patrol images. The proposed framework combines visible light images, multispectral cues, and spatial state encoding to enhance the boundary representation of floating objects against a reflected water background. The lightweight detection branch locates the suspected floating area, and the classification branch is used to distinguish the categories of floating objects such as plastic bottles, foam boards, branches and leaves, bags and algae. A dataset consisting of 18,420 UAV images and 73,600 annotated objects is constructed from different river, lake, and wetland scenes. Experimental results show that the model achieves 91.8% mAP, 94.2% classification accuracy, 92.7% F1-score, 42.6 FPS, and FLOPs of 15.6B, which supports stable river patrol deployment and online analysis tasks.*

KEYWORDS: *Deep learning; Uav river patrol; Floating object recognition; Image classification*

1 Introduction

The intelligent recognition and classification of floating objects in rivers and lakes by unmanned aerial vehicle (UAV) is a key technology in water intelligent perception. It can automatically locate plastic bottles, foam boards, branches, bag-like garbage and algal masses by using low-altitude river patrol images, and complete category discrimination. This technology is oriented to scenes with dense rivers, complex coastlines and high inspection frequency, and can convert visible light images, multispectral information and flight pose data into visual features, which provides a data basis for abnormal detection and spot review. There are many factors in the river patrol image of UAV, such as water reflection, ripple disturbance, shoreline shadow and scale change. Floating objects often appear in the form of small size, low contrast and irregular boundaries, and it is difficult to maintain stable results by threshold or texture matching.

The existing research provides a methodological basis for the detection of floating objects

*chenfei8478@163.com
https://doi.org/10.65102/is20261055
Published: 10 June, 2026

on the water surface. Maharjan et al. used UAV sensing data and deep learning to detect river plastics, and verified the feasibility of low-altitude images in river garbage identification [1]. Goncalves and Andriolo studied the operational usage of UAV multi-spectral images in macro garbage mapping and classification, indicating that spectral information can make up for the deficiency of visible light in water reflection conditions [2]. Andriolo et al. compared the operation differences of Uavs in the investigation of beach garbage and floating garbage, and proposed that the flight height and imaging scale would affect the target visibility [3]. Goddijn-Murphy et al. used UAV thermal infrared cameras to monitor floating Marine plastics, providing a complementary path for low-light water perception [4]. Armitage et al. proposed a floating plastic detection and classification method based on on-board video and deep learning, and proved that continuous video frames can be used for target state recognition [5]. Sannigrahi et al. developed an automated Marine floating plastic detection system combining Sentinel-2 images and machine learning, indicating that remote sensing information can provide regional priors [6]. Gnann et al. reviewed the short-range remote sensing and artificial intelligent-assisted macroplastic recognition methods and pointed out that robust feature representation was still needed under complex water background [7]. Goncalves et al. discussed the standardization process of UAV garbage investigation and emphasized that image resolution, route and labeling caliber would affect model reuse effect [8]. Corrigan et al. proposed a real-time instance segmentation method for underwater garbage to provide reference for boundary extraction [9]. Alboody et al. embedded the hyperspectral imaging system into the unmanned water surface platform and used machine learning to identify floating plastics, showing the advantages of multi-source sensing fusion [10]. Andriolo et al. analyzed the suitable flight altitude and image resolution in coastal and river garbage monitoring, providing parameter basis for river patrol data collection [11].

Focusing on the problems that small and medium-sized targets are not easy to separate, the category boundaries are close, and the features of multi-source images are difficult to express jointly, this paper constructs a deep learning recognition and classification method based on multi-source UAV vision. Method In the feature extraction stage, multi-spectral information is introduced as auxiliary features to enhance the boundary of floating objects, the difference of water texture, and the semantic response of categories. The detection branch is responsible for generating the suspected floating object area, and the classification branch further discriminates the target types such as plastic, foam, branches and leaves, pouches and algae. This method connects low-altitude inspection image acquisition, visual feature modeling and category output into a unified calculation process, which provides a technical path for automatic identification of floating objects in complex water scenes.

2 Related work

2.1 Traditional inspection and image recognition methods of floating objects in rivers and lakes

Artificial river patrol, fixed point video viewing and image recognition based on artificial features are the main methods in the early inspection of floating objects in rivers and lakes. Relying on on-site observation and photo recording, artificial river patrol can determine the location, type and accumulation range of floating objects. However, in scenes with a large number of river channels, tortuous shoreline, and obvious occlusion of wetlands and Bridges and culverts, the inspection coverage is easily affected by route, weather and personnel experience. A fixed camera can obtain continuous images, but the Angle of view is fixed,

which is difficult to cover the curved river and the blind corner of the bank. The image quality is also disturbed by reflection, fog, backlight and water ripple.

Traditional image recognition methods usually use color threshold, edge detection, texture statistics, morphological segmentation and SVM classifier to extract the floating object area, which has certain recognition ability for plastics, foam or branches with obvious color difference. However, the stability of feature expression is insufficient when the target is of low contrast background, small scale and similar class appearance. Zaaboub et al. used UAV and machine learning technology to evaluate sebaceous residue and plastic waste, indicating that low-altitude images can make up for the spatial blind area of manual inspection, but shallow features still rely on manual selection, and the generalization ability under complex water surface is limited [12]. Saeed et al. studied airborne small target detection of Uavs and pointed out that detection tasks on edge devices are constrained by target size, reasoning speed and computing resources [13].

Dewangan et al. applied image processing technology to UAV target detection and analyzed the recognition effect combined with distance change, showing that traditional image enhancement and geometric features are sensitive to the change of line of sight [14]. Terven and other systems comb the evolution of YOLO architecture in computer vision, showing that the end-to-end detection framework has gradually replaced the process that relies on manual features [15]. Diwan et al. analyzed the structure, dataset and application challenges of YOLO target detection, and pointed out that complex background, small target and real-time deployment are still computational problems that need to be solved for water patrol tasks [16].

In the dense area of rivers and lakes, the algorithm also needs to take into account the connection between the batch acquisition of routes, the quality of video return, the consistency of image archiving and the requirements of subsequent review tasks. Therefore, traditional inspection and traditional image recognition methods can be used as basic data collection and initial screening methods, but they are difficult to support fine positioning, category discrimination and online processing under multi-source UAV vision conditions. Subsequent methods need to introduce deep feature extraction, scale adaptation and lightweight reasoning structures.

2.2 UAV water target detection and classification method based on deep learning

Deep learning object detection provides a stable feature expression for the recognition of floating objects in UAV water images. Different from traditional threshold segmentation and texture description, convolutional network can extract edge, shape, color and context information in multi-layer features, and the single-stage detection framework can also complete candidate region regression and category prediction, which is suitable for rapid analysis after the return of river tour video. Zaidi et al. systematically combed the modern deep learning object detection model and pointed out that the two-stage method was stable in regional positioning, but the calculation link was long, and the single-stage method was more suitable for edge deployment in real-time detection [17]. In UAV river and lake scenes, floating objects are often affected by water highlights, shoreline shadows and shooting height. The model should not only maintain the recall rate of small targets, but also control the inference delay.

As shown in Table 1, different deep learning methods form different applicable characteristics, and model selection needs to take into account detection accuracy, classification ability and deployment overhead.

Table 1: Comparison of UAV water target detection methods

Method type	Main computational features	Applicable scenario	Limitation
Faster R-CNN	Candidate region generation and feature pooling	High-precision offline review	Slow inference speed
Improved YOLOv4 model	Multi-scale detection and anchor-box regression	Low-altitude UAV target recognition	Small floating objects are easily missed
Improved YOLO algorithm	Lightweight convolution and fast classification	Real-time river patrol video processing	Strongly affected by reflection interference
Remote sensing segmentation model	Pixel-level annotation and semantic segmentation	Large-scale litter extraction	Depends on high-quality annotations
Proposed method	Multi-source visual fusion and dual detection-classification branches	Recognition and classification of floating objects in rivers and lakes	Requires multi-source image registration

Dadrass Javan et al. proposed an improved YOLOv4 network for vision-based UAV identification, indicating that the YOLO structure can be adapted to low-altitude shooting tasks through feature layer reorganization and detection head adjustment [18]. Jawaharlalnehru et al. proposed a UAV image target detection method based on improved YOLO algorithm, which improved the detection efficiency under complex background and provided a reference for real-time positioning in river patrol images [19]. Kikaki et al. constructed the MARIDA Marine debris remote sensing detection benchmark, indicating that the recognition of water surface debris needs unified category labels, pixel boundaries and background samples to reduce the deviation between different data sources [20]. Nunkhaw and Miyamoto studied the deep learning image analysis method of river floating waste, and proved that floating objects in river images can be automatically identified through deep features, but water changes and occlusion still affect the classification results [21].

Combined with the characteristics of image acquisition, video transmission and multi-source perception in UAV river patrol, the existing deep learning methods still need to adapt to the structure of river and lake scenes. The single visible light model is prone to target boundary drift under the conditions of reflection, low light and turty water. A single classification network lacks positioning output, which is difficult to support spot review. The detection and classification framework based on multi-source UAV vision can combine the visible light texture, multi-spectral response and spatial position coding into the same feature space, extract the water surface background difference in the backbone network, locate the floating object in the detection branch, and output the categories of plastic, foam, branches and leaves, bags and algae in the classification branch, serving the online recognition of river and lake inspection.

3 Intelligent recognition and classification method of floating objects in rivers and lakes based on multi-source UAV vision

3.1 Principle and calculation process of floating object recognition based on UAV river patrol image

The identification of floating objects based on UAV river patrol image is a calculation process that converts the river and lake surface images collected at low altitude into target regions, category labels and confidence results. The process takes the visible image as the main input, and receives the multispectral channel, flight attitude, relative altitude and timestamp information at the same time. After image denoising, scale normalization and spatial registration, the unified visual tensor is formed. Floating objects in rivers and lakes often show the characteristics of small scale, weak boundary and low contrast. The surface reflection, ripple disturbance, shoreline shadow and bridge and culverts can weaken the target texture. Therefore, this section organizes the multi-source image input, deep feature extraction, background suppression, candidate box generation and category discrimination as a continuous computational flow.

To illustrate the processing sequence of river patrol images after entering the model, Fig. 1 shows the process of floating object recognition and classification. The process starts from the UAV acquisition end, and enters the detection and classification network after image cleaning, multi-source alignment and depth coding. Finally, the reviewable position and category results of the floating object are output.

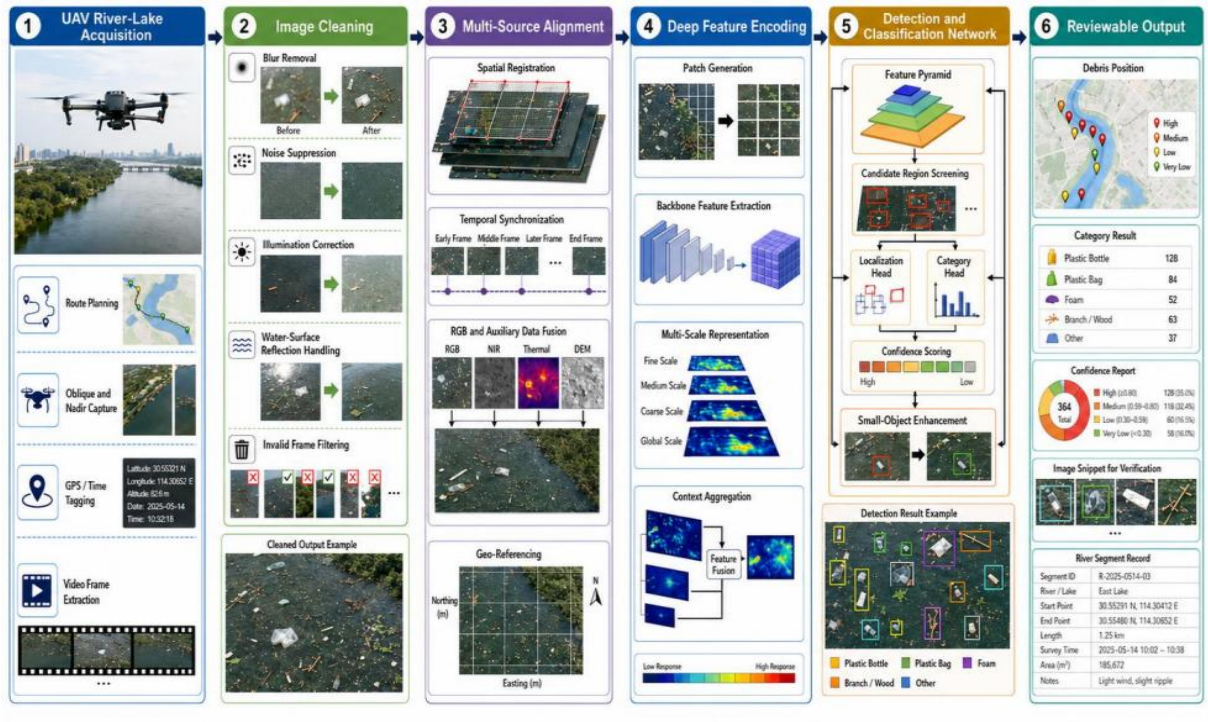

Figure 1: UAV river and lake floating object recognition and classification calculation process

In order to ensure that different sensing channels participate in the calculation in the same coordinate system, the input needs to encode the optical image, multispectral image and flight

state into a unified tensor, and the construction process is shown in the following equation:

$$X_t = \Phi_v(I_t^v) \oplus \Phi_s(T_{s\to v}(I_t^s)) \oplus \Phi_g(a_t, r_t, \tau_t) \quad (1)$$

Here, I_t^v represents the visible image at time t,I_t^s represents the multispectral image, $T_{s\to v}(\cdot)$ represents the spatial registration map, $\Phi_v(\cdot)$, $\Phi_s(\cdot)$, and $\Phi_g(\cdot)$ represent the three types of encoding functions, respectively, a_t, r_t, τ_t represent the attitude Angle, relative height, and timestamp, and Xt represents the unified input feature. This formula is used to eliminate scale, Angle and acquisition time differences, so that multi-source information enters the same feature space.

After input alignment, the model needs to weaken background responses such as ripples, reflections and shore shadows, so that the edges of floating objects and material textures are clear in the deep features. The feature correction process is shown in the following equation:

$$F_t = \Gamma(X_t) \odot [1 - \Omega(B_t)] + \Lambda(X_t) \odot \Omega(U_t) \quad (2)$$

Here, F_t represents the target perception feature,$\Gamma(X_t)$ represents the global semantic feature, $\Lambda(X_t)$ represents the local boundary feature, B_t represents the water surface background response map, U_t represents the suspected floating object response map, $\Omega(\cdot)$ represents the normalized weight map, and $\odot$ represents the element-wise weighting. This formula models the background and the target separately, which makes the floating object form a more stable visual response in the complex water surface.

After background suppression, the detection branch and the classification branch need to output the candidate box, category and confidence synchronously, and use the multi-source consistency to correct the judgment bias in the low-quality image. The joint output process is shown in the following equation:

$$O_t = \{(b_n, c_n, \alpha_n) \mid \alpha_n = \sigma(S_n + \lambda R_n), n = 1, \dots, M\} \quad (3)$$

where O_t represents the set of final recognition results, b_n represents the n candidate box, c_n represents the category of floating objects, α_n represents the confidence, S_n represents the basic detection score, R_n represents the multi-source consistency score, $\lambda\lambda$ represents the correction coefficient, and M represents the number of candidate targets. This formulation puts localization, classification and credibility calibration into the same output structure, so that the results can correspond to specific surface targets.

The calculation process focuses on the imaging characteristics of the river patrol image, which not only retains the edge and texture information in the visible light, but also uses the multispectral response to supplement the material difference. At the same time, the process retains the flight position and collection time information, which facilitates the alignment of multiple inspection results in the same river reach, and facilitates the tracking of target position, moving direction and category changes in subsequent manual review, so that the recognition results have better scene continuity, engineering usability and high stability.

3.2 Visual feature adaptation and recognition mechanism for floating objects in complex river patrol scenes

Compared with general UAV target detection, floating object recognition in rivers and lakes depends more on scene adaptation. The floating objects are often in the reflection of the water surface, the shadow of the shoreline, the occlusion of the bridge and the disturbance of the ripple. The plastic bottles, foam boards, branches, bags and algae have small differences in

appearance, and it is difficult to maintain stable discrimination by only relying on the visible light texture. The proposed method does not regard the river patrol image as a common target detection image, but takes the water background state, the change of flight Angle and the material characteristics of floating objects into the visual representation, so that the model can complete the target localization and category discrimination in complex river patrol scenes.

To illustrate the role of the visual feature adaptation mechanism in the model, Fig. 2 shows the feature modification path in a complex river patrol scene. The mechanism starts with scene perturbation estimation, assigns weights to multi-scale features, and then unifies spectral differences, boundary strengths, and classification confidence into the floating object recognition output.

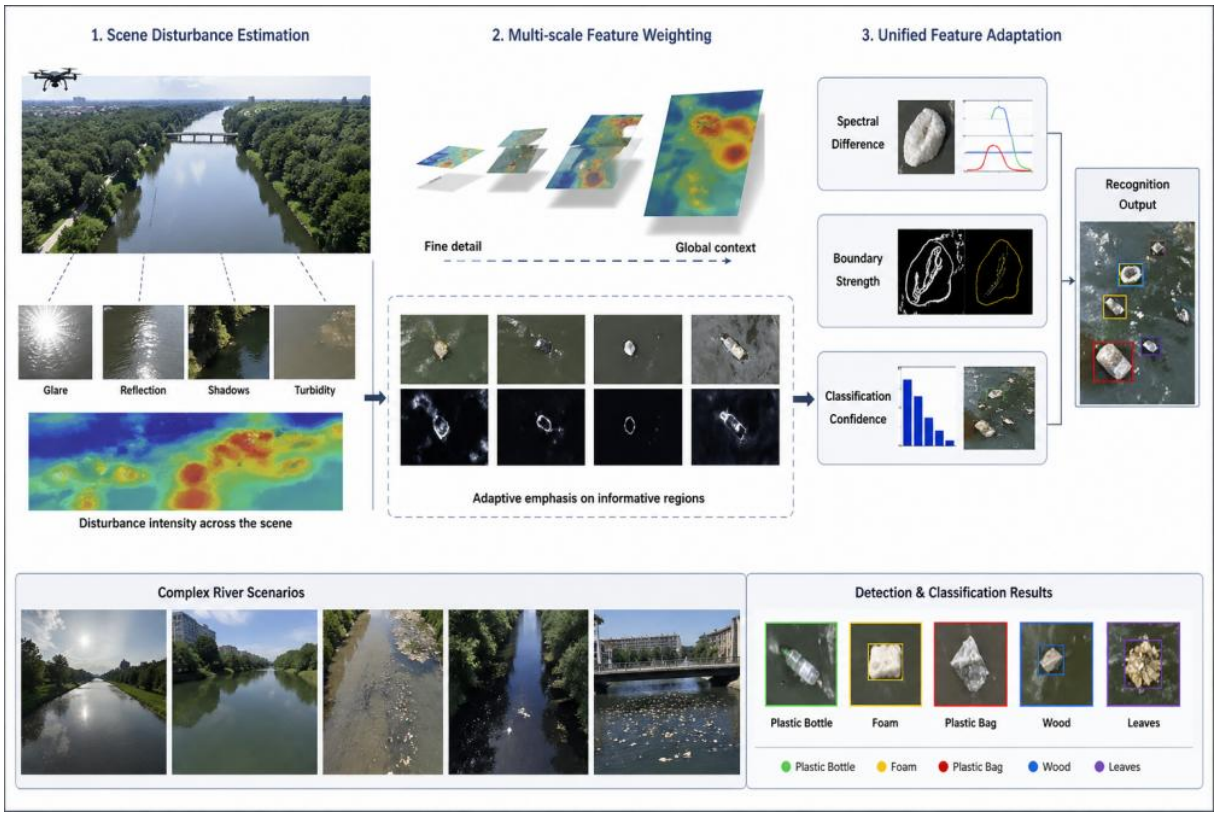

Figure 2: Visual feature adaptation mechanism for floating objects in complex river patrol scenes

In order to describe the influence of reflection, shadow and ripple on the visual expression of floating objects in the river patrol image, it is necessary to construct the scene disturbance vector and participate in the feature correction calculation. The expression process is as follows:

$$D_t = \mathrm{Norm}(\mu(W_t), \sigma(W_t), \nabla L_t, \kappa(S_t)) \tag{4}$$

Here, D_t represents the scene disturbance vector at time t, W_t represents the local surface texture patch, $\mu(W_t)$ and $\sigma(W_t)$ represent the texture mean and dispersion degree, ∇L_t represents the illumination gradient, $\kappa(S_t)$ represents the surface ripple curvature statistics, and $\mathrm{Norm}(\cdot)$ represents the normalized mapping. This formula encodes the background disturbance explicitly, so that the model can obtain interpretable background correction basis in strong reflection and dark shadow regions.

In order to maintain the separable expression of floating objects of different scales in the network hierarchy, it is necessary to generate hierarchical weights according to the target candidate scale and complete multi-layer aggregation calculation. The scale adaptation process is shown in the following equation:

$$M_t = \sum_{l=1}^{L} \frac{\exp(\beta_l \cdot u_t - \gamma d_l)}{\sum_{r=1}^{L} \exp(\beta_r \cdot u_t - \gamma d_r)} F_t^{(l)} \tag{5}$$

Here, M_t represents the fusion feature after scale adaptation, $F_t^{(l)}$ represents the feature map of the l layer, u_t represents the candidate target scale vector, d_l represents the deviation between the receptive field of the layer and the target scale, β_l and γ represent the learnable parameters, and L represents the total number of feature layers. This formula avoids small floatings being compressed by high-level semantics, and avoids large algal masses being misjudged by low-level textures.

In order to ensure that the material category is consistent with the spatial position output, spectral difference, boundary strength and classification confidence should be integrated into the discriminant calculation constraint term. The category discrimination process is shown in the following equation:

$$C_t = \arg\max_c [P_c(M_t) + \lambda_1 E_c + \lambda_2 Q_c - \lambda_3 R_c] \tag{6}$$

where C_t represents the final class output, $P_c(M_t)$ represents the depth classification probability of class c, E_c represents the multispectral response difference, Q_c represents the object boundary integrity, R_c represents the background confusion penalty term, and λ_1 to λ_3 represent the weight coefficients. This formula takes material properties, shape boundaries and background interference into the category judgment synchronously, so that similar objects such as plastic, foam, branches and algae can maintain clearer discrimination boundaries.

Different from the traditional fixed feature method, the core of the adaptation mechanism is to dynamically adjust the visual response according to the state of the river patrol screen. Under the condition of multi-source image acquisition, visible light assumes the responsibility of texture localization, and multi-spectral assumes the responsibility of material discrimination. Both of them jointly constrain the spatial consistency and category stability of candidate targets, and reduce the fluctuation of false detection and missed detection. When the reflection of the water surface is strong, the model weakens the influence of bright speckle on the candidate box. When the shoreline shadow is obvious, the model increases the weight of boundary continuity and spectral difference. When floating objects gather, the model uses multi-scale aggregation to separate adjacent objects. After scene disturbance modeling, scale adaptation and category constraints, the identified links can better correspond to different inspection images such as rivers, lakes and wetlands, which provides a stable feature basis for subsequent model design.

3.3 Model Design

3.3.1 Multi-source UAV vision network structure

The network structure of the proposed method is constructed around the multi-source input and rapid recognition task of UAV river and lake inspection images, which is composed of visible light branch, multi-spectral branch, spatial state coding layer, cross-modal fusion layer, floating object detection head and category discrimination head. The visible light image is

used to provide the edge, color and texture information of the floating object, the multispectral image is used to supplement the material response difference, and the flight altitude, camera pose and time stamp are used as the spatial state information to participate in the encoding. The network structure does not directly concatenate multi-source data, but retains the independent feature channels of visible light and multi-spectral in the backbone extraction stage, and then completes the semantic alignment through the fusion layer, so that the model can adapt to different river cruise images such as rivers, lakes, wetlands and coastlines.

In order to maintain the correspondence between visible and multispectral features in different network depths, it is necessary to construct a hierarchical mapping relationship of two-stream backbones and control the information transfer between channels. The calculation process is as follows:

$$H_l = \phi_l(W_l^v * V_l + W_l^m * M_l + E_l(G_l)) \tag{7}$$

Here, H_l represents the l layer output feature, V_l represents the visible light feature, M_l represents the multispectral feature,G_l represents the flight state encoding, W_l^v and W_l^m represent the branch convolution parameters, $E_l(\cdot)$ represents the state embedding function, and $\phi_l(\cdot)$ represents the activation and normalization combination. This formula is used to describe how the two-stream backbone receives image information and spatial state information in the same level, so that feature extraction is not limited to a single visual channel.

To clearly illustrate the internal connection relationship of the multi-source UAV vision network, Fig. 3 shows the overall structure from image input to recognition output. In the figure, the visible image and the multispectral image enter the dual-stream backbone respectively, and the spatial state encoding enters the fusion layer. The detection head is responsible for locating the floating object, the classification head is responsible for output the target category, and the confidence calibration unit is used to correct the recognition bias in the reflection, occlusion and low-definition images.

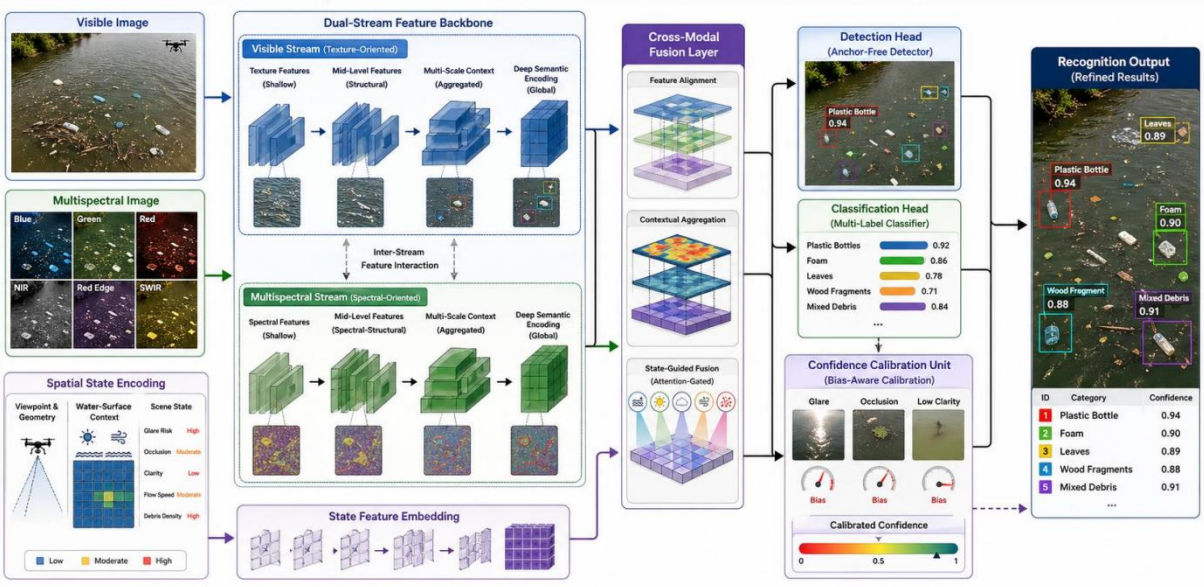

Figure 3: Multi-source UAV vision network structure diagram

In order to ensure that the network maintains local boundary information in small target

floating object recognition, multi-layer features should be reorganized across scales, and the association between shallow texture and deep semantics should be preserved. The reorganization process is shown in the following equation:

$$R_l = \alpha_l \mathrm{Up}(H_{l+1}) + (1 - \alpha_l) H_l + \beta_l \mathrm{Down}(H_{l-1}) \tag{8}$$

Here, R_l represents the restructured features at layer l, $\mathrm{Up}(\cdot)$ and $\mathrm{Down}(\cdot)$ represent the upsampling and downsampling operations, respectively, and α_l and β_l represent the learnable scale weights. This formulation enables the network to use both shallow edges and deep category semantics to reduce the false detection caused by water surface texture when dealing with different scale objects such as foam fragments, plastic bottles, and branches.

In this network structure, the image acquisition conditions of the river patrol map, the visual morphology of the floating object and the multi-source input features were put into a unified computing link. The visible branch is responsible for the detailed texture, the multispectral branch is responsible for the material difference, and the state encoding is responsible for the viewpoint and scale compensation. The detection and classification head shares the fusion features, but the output tasks are independent of each other, so that the localization results and category results can be generated at the same time, and provide a stable image basis for subsequent review. This structure takes into account both the reasoning efficiency of the edge end and the stability of the fine-grained expression ability of the water target.

3.3.2 Image input module and model parameter configuration

The image input module is responsible for organizing the heterogeneous images obtained by UAV river patrol into trainable tensors. The raw data contains visible light images, multispectral images, flight altitude, camera pose, acquisition time, and flight route number, and different sources differ in spatial resolution, imaging Angle, and lighting conditions. The input module performs image denoising, distortion correction and scale unification, and then completes visible and multispectral registration according to geographical location and timestamp. Subsequently, the module writes image channels, flight states and scene labels into batch samples to ensure that the network receives stable inputs in the training and inference phases.

Table 2 lists the image input module with the model parameter configuration. The input size is used to preserve the boundary of the floating object, the batch size is used to balance the memory occupation and gradient stability, and the threshold parameter is used to filter the weak response candidate box.

Table 2: Image input module and model parameter configuration

Parameter item	Configuration value	Computational function
Input size	1280×720	Preserves small-object textures
Visible-light channels	3	Extracts color and boundary features
Multispectral channels	5	Represents material differences
Batch size	16	Stabilizes gradient updates
Number of fusion layers	4	Covers multi-scale targets
Initial learning rate	0.001	Controls the convergence range
Confidence threshold	0.35	Filters background responses

In order to weaken the influence of water reflections and local shadows on the input distribution, the image channels need to be robust normalized instead of simple mean scaling,

which is calculated as follows:

$$\tilde{I}_c = \frac{I_c - Q_{0.5}(I_c)}{Q_{0.75}(I_c) - Q_{0.25}(I_c) + \epsilon} \tag{9}$$

Here, $\tilde{I}_c$ represents the normalized image of the c channel, I_c represents the original channel, $Q_{0.5}$, $Q_{0.75}$, and $Q_{0.25}$ represent the median, upper quartile, and lower quartile, respectively, and ϵ represents the stabilization term that prevents the denominator from being zero. The formula uses quantiles to constrain abnormal bright spots, so that the input image still maintains a stable numerical range in the strong reflection region.

In order to ensure the reasonable contribution of the samples in the training, the sample weights need to be adjusted according to the image clarity, target density and registration error. The weight calculation process is as follows:

$$\omega_i = \frac{\log(1 + n_i) \cdot q_i}{1 + \exp(e_i - \bar{e})} \tag{10}$$

Here, ω_i represents the training weight of the i image, n_i represents the number of labeled targets, q_i represents the image sharpness score, e_i represents the multi-source registration error, and $\bar{e}$ represents the batch average error. In this formula, the target density, image quality and registration state are incorporated into the input weight, so that the low-quality but effective tour images are not completely ignored, and the samples with poor registration quality are avoided to cause too strong interference to the parameter update.

In the actual river patrol scene, the same river segment may be collected by a fixed airport, a mobile airport, or a temporary airline, and the image scale and top view Angle are not consistent. The input module retains the route number and collection time, which can push multiple images of the same water into the same index space. The subsequent model can continuously compare the position of the floating object, and also reduce the sample bias caused by repeated images. This processing makes the parameter configuration consistent with the UAV river patrol acquisition characteristics and also facilitates subsequent batch reasoning.

3.3.3 Fusion mechanism of visible light and multispectral features

The optical and multispectral feature fusion mechanism is used to put texture boundary and material response into the same feature space. Visible images can present the contour, color, shadow boundary and local shape of floating objects, which is suitable for candidate region localization. Multispectral images can show the reflection differences of different materials in the blue, green, red, red edge and near infrared channels, which is suitable for distinguishing plastic, foam, branches and algae and other targets. The two types of images have differences in spatial resolution and acquisition time. The fusion mechanism needs to complete feature alignment, assign weights according to the water surface state, and finally generate fusion features for detection and classification.

In order to reduce the influence of multi-source image registration error on fusion results, it is necessary to jointly determine the cross-modal matching relationship by spatial proximity and semantic similarity. The matching weight calculation process is shown as follows:

$$A_{pq} = \frac{\exp\left((f_p^v)^\top f_q^m / \sqrt{d} - \lambda \|p - q\|_2\right)}{\sum_{q' \in \mathcal{N}(p)} \exp\left((f_p^v)^\top f_{q'}^m / \sqrt{d} - \lambda \|p - q'\|_2\right)} \tag{11}$$

Here, A_{pq} represents the matching weight between visible light position p and multispectral position q, f_p^v represents the visible light feature, f_q^m represents the multispectral feature, d represents the feature dimension, λ represents the spatial distance penalty coefficient, and $\mathcal{N}(p)$ represents the neighborhood set. This formulation enables the fusion layer to preferentially select spatially close and semantically similar multispectral features to avoid misaligned responses entering the floating phenological constituency.

In order to make the model adaptively select feature sources in reflective, shadow and turquoise water, it is necessary to calculate the fusion gating coefficient according to the water surface state, and the gating calculation process is shown as follows:

$$g_p = \sigma(W_g[f_p^v, \sum_q A_{pq} f_q^m, l_p, b_p] + b_g) \tag{12}$$

Here, g_p represents the fusion gating at position p, l_p represents the local brightness perturbation, b_p represents the boundary strength, and W_g and b_gg represent the learnable parameters. The formula uses visual texture, spectral response, brightness change and boundary information together for weight calculation, so that the visible light plays a dominant role when the boundary is clear, and the multispectrum bears higher weight when the material difference is obvious.

In order to generate the fusion features of the final input detection classification head, the gating results and cross-modal matching results should be combined in the feature reconstruction process, and the residual information should be retained. The fusion process is shown in the following equation:

$$F_p^{fuse} = (1 - g_p)f_p^v + g_p \sum_q A_{pq} f_q^m + \Theta([f_p^v, f_p^m]) \tag{13}$$

Here, F_p^{fuse} represents the fused multi-source visual features, $\Theta(\cdot)$ represents the residual mapping function, and $[\cdot]$ represents the feature concatenation. This formulation enables the fusion result to contain both the spatial details of the visible light and the multispectral material response, and the local information in the original image is preserved by residual compensation.

The mechanism is also suitable to deal with the perspective bias caused by route changes. The floating objects in the river patrol image may move with the current, and the same target presents different directions and scales in adjacent frames. The fusion layer retains the spatial matching weight, which can maintain the consistency of the target response in consecutive images, reduce the category jump caused by short-term reflection and ripple occlusion, and make the subsequent output have better temporal coherence.

3.3.4 Floating object localization detection and category discrimination branch

The floating object location detection and class discrimination branch is the core part of the model output layer, which is responsible for candidate box regression, target confidence assessment and class probability prediction. After receiving the fusion features in the detection branch, the candidate regions were generated on the feature maps of different scales to determine the center position, width, height and spatial extent of the floating object. The classification branch discriminates the material and shape of the candidate regions, and outputs the categories of plastic bottles, foam boards, branches and leaves, bags and algae. Since the boundary of floating objects is easily cut by water waves, and similar objects have

different appearance due to wetness, illumination Angle and drift attitude, the output branch needs to use regional features, boundary integrity and multi-source consistency scores at the same time.

In order to make the candidate box adapt to the morphological changes of floating objects, the positioning branch uses the center offset and scale extension joint regression method to generate the target boundary, and the regression process is shown in the following equation:

$$B_n = (x_n + \Delta x_n, y_n + \Delta y_n, w_n \exp(\Delta w_n), h_n \exp(\Delta h_n)) \tag{14}$$

where B_n represents the n predicted candidate box, x_n, y_n, w_n, h_n represent the center coordinates and dimensions of the base box, and $\Delta x_n, \Delta y_n, \Delta w_n, \Delta h_n$ represent the offset predicted by the fused features. This formula keeps the width and height as positive values by exponential scaling, and allows the candidate box to be corrected according to the boundary changes of the floating object, so that the positioning result is closer to the real area.

In order to reduce the interference of bright spots on the water surface, shadows on the shore and floating textures on the candidate target, the target response and background suppression terms need to be incorporated into the confidence calculation. The confidence evaluation process is shown in the following equation:

$$s_n = \sigma(w_s^\top r_n - \eta_1 \mathcal{B}_n + \eta_2 \mathcal{E}_n + \eta_3 \mathcal{M}_n) \tag{15}$$

Here, s_n represents the confidence of the n candidate target, r_n represents the candidate region feature, $\mathcal{B}_n$ represents the background similarity, $\mathcal{E}_n$ represents the boundary closure degree, $\mathcal{M}_n$ represents the multi-source consistency score, and η_1 to η_3 represent the weight coefficients. The formula reduces the confidence output in the high background similar region, and enhances the target response in the complete boundary region, so that the candidate box screening is more in line with the imaging characteristics of floating objects on the water surface.

In order to make the category discrimination branch distinguish similar material targets, texture, spectral and morphological features need to form a joint category vector, and the calculation process of the category probability is as follows:

$$p_n(c) = \frac{\exp(W_c^\top [r_n^t, r_n^m, r_n^b] + b_c)}{\sum_{c'=1}^{C} \exp(W_{c'}^\top [r_n^t, r_n^m, r_n^b] + b_{c'})} \tag{16}$$

Here, $p_n(c)$ represents the probability that the n candidate object belongs to class c, r_n^t represents the texture feature, r_n^m represents the multispectral material feature, r_n^b represents the boundary shape feature, and C represents the total number of classes. The formula uses material, shape and local texture together for class discrimination, avoiding confusion between plastic, foam, pouches and algae caused by color judgment alone.

The location detection and category discrimination branches adopt the structure of sharing fusion features and separating task outputs. Sharing features can reduce repeated computations, and separating outputs can avoid interference between localization tasks and classification tasks. The detection branch focuses on the spatial accuracy of the candidate region, and the classification branch focuses on the target material and shape differences. The confidence calculation provides the basis for screening between the two branches. After candidate box regression, background suppression and joint category discrimination, the model can transform the floating objects in the river patrol image into structured results that can be located, classified and checked, and provide stable output for online patrol recognition.

3.3.5 Model Training Strategy and parameter update method

The model training strategy focuses on multi-source image alignment, joint learning of detection and classification, and stable parameter update. The training process adopts a phased strategy, and the parameters of the multispectral branch are fixed in the early stage, so that the visible light backbone obtains stable boundary expression first. In the middle stage, the fusion layer and the detection and classification head were opened to enable the model to learn the target response in different scenarios. In the later stage, a smaller learning rate is used to fine-tune the full network parameters to reduce the gradient fluctuation caused by high reflective images and low definition images. Data augmentation uses brightness perturbation, random cropping, scale transformation and water texture noise simulation, so that the model can maintain stable training under the conditions of river bend, wetland occlusion, dark area of Bridges and culverts, and strong reflection on the lake surface. The main parameter Settings during model training are shown in Table 3.

Table 3: Model training parameter configuration

Parameter item	Configuration value	Description
Optimizer	AdamW	Decouples weight decay
Training epochs	120	Supports convergence of multi-source features
Batch size	16	Controls GPU memory consumption
Weight decay	0.0005	Limits overfitting
Learning rate decay	Cosine	Smooths parameter updates
Augmentation method	Cropping, scaling, brightness perturbation	Simulates changes in UAV river patrol images

In order to keep the multi-task training stable in different stages, the detection loss, classification loss and multi-source consistency loss should be formed into a joint objective function, and the training objective is shown in the following equation:

$$\mathcal{L} = \lambda_{\mathrm{d}}\mathcal{L}_{\mathrm{det}} + \lambda_{\mathrm{c}}\mathcal{L}_{\mathrm{cls}} + \lambda_{\mathrm{m}}\mathcal{L}_{\mathrm{match}} + \lambda_{\mathrm{r}}\|\Theta\|_2^2 \tag{17}$$

Here, $\mathcal{L}$ represents the total loss, $\mathcal{L}_{\mathrm{det}}$ represents the localization detection loss, $\mathcal{L}_{\mathrm{cls}}$ represents the class discrimination loss, $\mathcal{L}_{\mathrm{match}}$ represents the consistency loss between visible light and multispectral, $\|\Theta\|_2^2$ represents the parameter regularization term, and λ_{d},λ_{c},λ_{m}, and λ_{r} represent the weight coefficients. This formulation incorporates localization, classification, and multi-source alignment into training at the same time, so that the model is not biased towards a single task.

In order to control the gradient fluctuation in the training process, the learning rate update amplitude should be dynamically adjusted according to the change of validation loss and the quality of batch samples. The learning rate adjustment process is shown in the following equation:

$$\eta_{\mathrm{t}} = \eta_0 \cdot \frac{1 + \cos(\pi \mathrm{t}/\mathrm{T})}{2} \cdot (1 + \rho \mathrm{Q}_{\mathrm{t}})^{-1} \tag{18}$$

Here, η_{t} represents the learning rate at the t round, η_0 represents the initial learning rate, T represents the total training rounds, Q_{t} represents the current batch image quality fluctuation coefficient, and ρ represents the adjustment parameter. The formula smoothly reduces the learning rate in the later stage of training, and automatically compreses the update

amplitude when the proportion of low-quality images is high to reduce parameter oscillation.

The parameter update method is matched with the image characteristics of the river patrol map. The visible branch learns the boundary and texture through the enhanced image, the multispectral branch learns the material response through the consistency constraint, and the detection branch and the classification branch converge synchronously under the joint loss. After training, the model can output candidate boxes, categories and confidence with limited computing resources, and maintain a relatively stable inference speed.

4 Experimental Evaluation

4.1 Experimental Setup

The experimental setup was carried out around the task of recognition and classification of floating objects in rivers and lakes. The data consists of low-altitude river patrol images, visible light zoom images, multispectral channels and flight state records, covering scenes such as rivers, lakes, wetlands, Bridges, culverts, shore-lines and drainage outlets. The sample contains 18420 images and 73600 labeled targets, and the categories are set as plastic bottles, foam boards, branches and leaves, bags and algae floating objects. The images were processed by distortion correction, denoising, scale unification and multi-source registration, and the input size was fixed to 1280×720. The data were divided into training set, validation set and test set by 70%, 15% and 15%, and the random seed was set to 42 to keep the contrast conditions consistent.

In order to ensure the reproducibility of the experimental process, and clearly present the data scale, input channels, division methods and computing environment, Table 4 lists the data composition and operating environment Settings of this experiment.

Table 4: Experimental data and running environment Settings

Item	Setting	Description
Image scale	18,420 images	Covers multiple UAV river patrol scenarios
Annotated objects	73,600 objects	Supports detection and classification training
Input channels	3+5	Joint input of visible-light and multispectral images
Data split ratio	70:15:15	Keeps training, validation, and testing consistent
Hardware environment	RTX 3090	Supports image inference and model training
Software framework	PyTorch 2.0, CUDA 11.8	Supports deep learning computation

The comparison methods are HOG+SVM, FasterR-CNN, YOLOv5s, YOLOv8n and the model of this paper. All methods use the same data partition, augmentation strategy and evaluation metric. In the training phase, random cropping, brightness perturbation, water texture noise and scale scaling are used to reduce the influence of reflection, shadow and shooting height differences on model learning. The evaluation metrics include mAP, Accuracy, F1-score, FPS, and FLOPs to measure positioning accuracy, category judgment, real-time inference, and computational overhead. The experimental process fixed the initial value of learning rate 0.001, batch size 16, training 120 rounds, the validation set was used for parameter selection, and the test set was only used for the final performance evaluation. This setting corresponds to the online identification task of UAV river patrol, and maintains the balance of sample distribution and the reproducibility of results.

4.2 Experimental results and analysis

4.2.1 Recognition effects under different river patrol scenes

The imaging differences of different river patrol scenes directly affect the recognition results of floating objects. In the river scene, the shoreline is narrow and the moving direction of the target is stable, in the lake scene, the water surface is open and the reflective range is large, in the wetland scene, the vegetation is occluded obviously, in the bridge and culverts scene, there are dark areas and shadow edges, and in the shoreline scene, there are floating objects and shore debris. The model uses the same test set partition and the same confidence threshold in all five types of scenes, and outputs mAP, Accuracy, F1-score and FPS. To visualize the scene differences, Fig. 4 shows the recognition performance heatmaps under different river patrol scenes, where higher values indicate more stable detection and classification results.

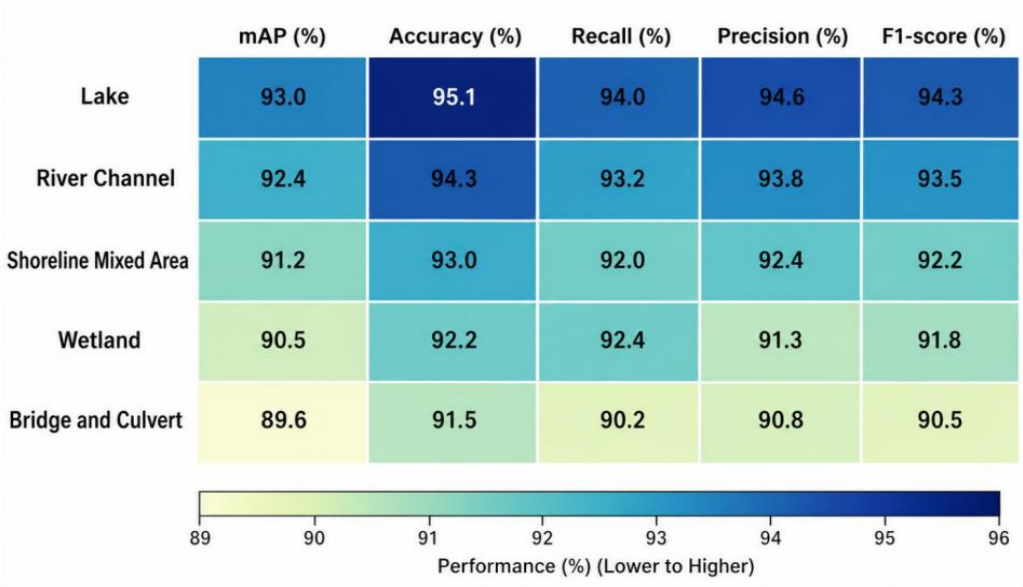

Figure 4: Heatmaps of recognition performance for different river patrol scenarios

Fig. 4 shows that the mAP of the lake scene reaches 93.0% and the Accuracy reaches 95.1%, the main reason is that the floating object contour is relatively separated from the water surface background, and the multispectral branch can supplement the material response. The mAP of the bridge and culverts scene is 89.6%, which is lower than the other scenes, and the error mainly comes from the shadow edge and the low-light area under the bridge. The F1-score of wetland scene is 91.8%, and vegetation texture can interfere with the judgment of floating objects in branches and leaves, but the model still maintains a high recall level. To further observe the error sources, Fig. 5 shows the distribution of false detections and missed detections in five types of scenes.

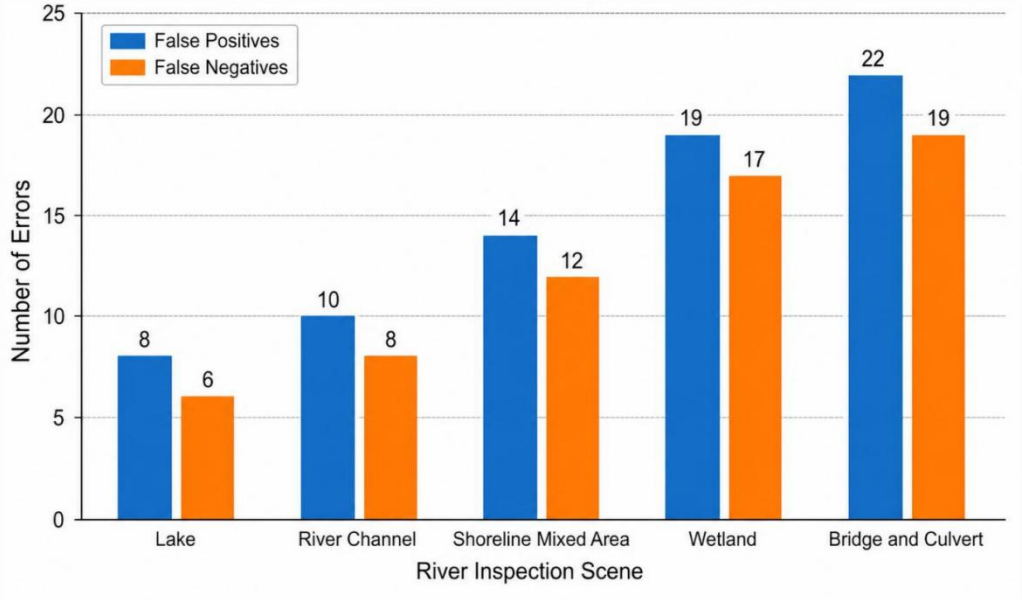

Figure 5: False detection and missed detection distribution of river patrol scene

Fig. 5 shows that Bridges and culverts and wetlands are the scenes with more concentrated errors. The dark area in the bridge and culvert image will weaken the edge of floating objects, and the wetland vegetation will form similar textures with branches and leaves. The low error of the river and lake scenes indicates that the multi-source visual structure can maintain stable recognition in the conventional river patrol images. On the whole, the model does not obviously rely on a single scene under complex water background, and can adapt to scale changes, illumination changes and shoreline interference in UAV river patrol.

The recognition results of different river patrol scenes show that the proposed model maintains a relatively stable detection and classification performance in the mixed area of open lake, regular river and shoreline, and does not show significant performance degradation in complex scenes such as Bridges and culverts. The errors are mainly concentrated in areas with low illumination, strong reflection and vegetation occlusion, which is consistent with the real acquisition conditions of UAV water images. Multi-source visual features play a constraint role in complex background, so that the model can maintain good recognition continuity under the coexistence of water surface texture variation, target scale fluctuation and shoreline interference, which indicates that the method is suitable for online detection tasks in river patrol scenes.

4.2.2 Classification results of different types of floating objects

Different types of floating objects have obvious differences in appearance, material and water surface attitude. Plastic bottles and foam boards often exhibit highly reflective edges, branches and leaves have similar natural textures to algae, and the bag morphology is unstable and susceptible to water flow stretching and folding. The model classification results were evaluated using AP, Precision, Recall and F1-score. In order to present the discrimination differences between each class, Fig. 6 shows the classification result matrix of different floating object categories, where the main diagonal line in the matrix represents the proportion of correct classification, and the off-diagonal line represents the proportion of category confusion.

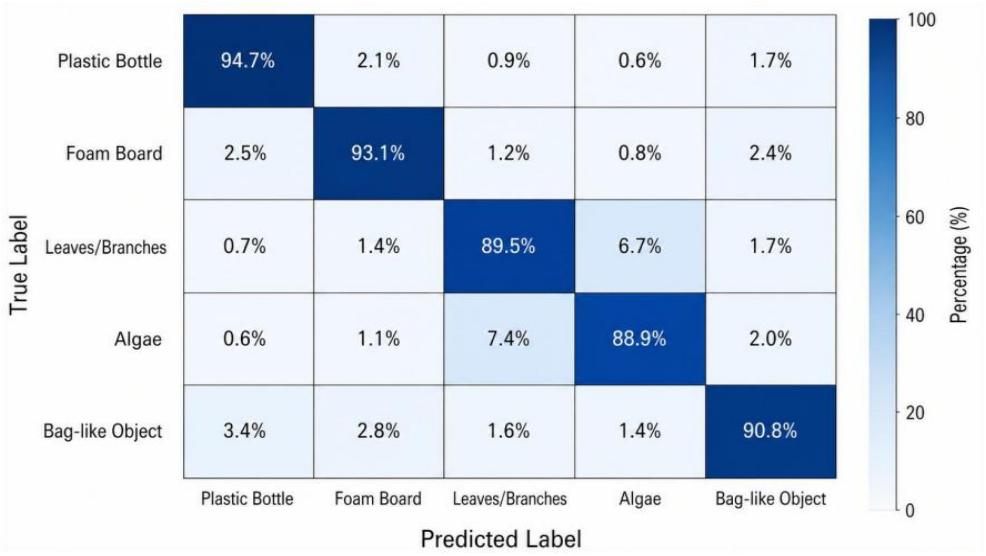

Figure 6: Confusion matrix for classification of floating object categories

Fig. 6 shows that the recognition ratio of plastic bottle and foam board reaches 94.7% and 93.1%, respectively, and the two types of targets have clear edges and obvious differences in material reflection. The confusion ratio between branches and algae was high, with 6.7% of branches and leaves judged as algae and 7.4% of algae judged as branches and leaves, indicating that similar textures still exist between natural flotters. Bags were correctly classified 90.8% of the time, with the majority of errors coming from similar boundaries to foam boards and plastic bottles. To observe the distribution of category features, Fig. 7 shows a schematic diagram of the inter-class distances of various types of targets in the fused feature space.

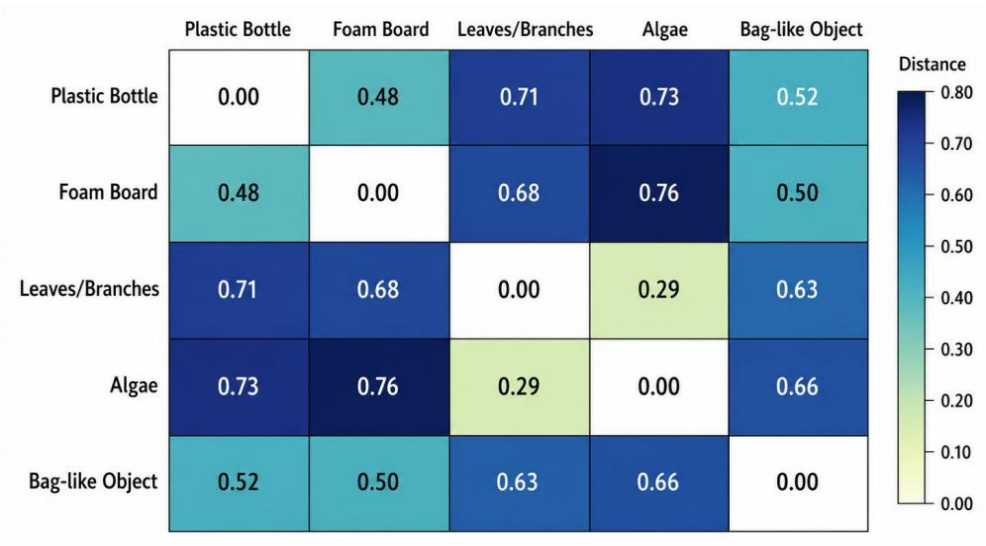

Figure 7: Inter-class distance plot of fusing features of floaters

Fig. 7 shows that the characteristic distance between artificial and natural floaters is large, the distance between plastic bottle and branches reaches 0.71, and the distance between foam board and algae reaches 0.76, indicating a clear classification boundary. The distance between branches and algae is only 0.29, which is the closest group of categories. Combined with Fig. 6, it can be seen that the multi-source visual branch can distinguish the artificial floaters well,

but the natural floaters still need to rely on the joint judgment of multispectral response and boundary morphology. The overall classification accuracy is 94.2%, and the F1-score is 92.7%.

From the classification results of different types of floating objects, the recognition boundary of artificial floating objects is clear, and the plastic bottle, foam board and bag have high separability in the fusion feature space. The differences between natural floaters are relatively subtle, and branches and algae are still the main sources of confusion. The multispectral information complements the expression of material differences, and the boundary features support the distinction between bags and foam sheets. The overall results show that the proposed model can not only complete the target localization of floating objects, but also form a stable category discrimination for different types of floating objects, which provides an experimental basis for subsequent multi-source visual branch analysis.

4.2.3 Influence of multi-source visual branches on model performance

The role of the multisource visual branch is reflected in the collaborative representation of visible light texture, multispectral material response, and spatial state encoding. To avoid duplication with the scene recognition and category classification graphs of the previous two sections, branch combination experiments and module ablation experiments are used to analyze structural contributions in this section. The branch combination experiment focuses on the influence of different input combinations on the overall detection and classification performance, and the ablation experiment focuses on the performance change after the internal modules of the model are removed. Both types of experiments use the same test set and the same threshold, and the indicators include mAP, Accuracy, F1-score, FPS and FLOPs.

Table 5 lists the results under different combinations of visual branches. The mAP of single branch in the visible light is 86.9%, which performs well in the images with clear boundaries, but there are more false detections under the reflective water surface. The mAP of multi-spectral single branch is 84.7%, which has strong material discrimination ability but insufficient spatial details. In the late stage, the splicing fusion reached 89.5%. In this paper, the multi-source fusion branch reaches 91.8%, the Accuracy is 94.2%, and the F1-score is 92.7%.

Table 5: Results of multi-source visual branch combination experiments

Method	mAP/%	Accuracy/%	F1-score/%	FPS	FLOPs/B
Visible-light single branch	86.9	89.8	88.4	47.8	10.2
Multispectral single branch	84.7	88.5	86.9	45.6	11.4
Late concatenation fusion	89.5	92.1	90.8	43.5	13.8
Proposed multi-source fusion	91.8	94.2	92.7	42.6	15.6

Table 6 further presents the results of ablation experiments. After removing the multispectral fusion, the mAP decreased to 88.6%, and the confusion between branches, algae and foam board increased. After removing the background suppression, the mAP was 87.9%, and the false detection of the dark area of the bridge and culvert and the reflective area of the lake increased. After removing the spatial state encoding, the F1-score decreased to 90.3%. After removing the confidence calibration, the mAP is 90.1%, and the screening of candidate boxes in low-quality images is not stable enough.

Table 6: Results of ablation experiments for multi-source vision models

Model setting	mAP/%	Accuracy/%	F1-score/%	FPS	FLOPs/B
Without multispectral fusion	88.6	91.3	89.7	44.1	13.2
Without background suppression	87.9	90.8	89.0	44.6	13.6
Without spatial state encoding	89.2	92.0	90.3	43.8	14.1
Without confidence calibration	90.1	92.7	91.2	43.1	14.8
Complete model	91.8	94.2	92.7	42.6	15.6

The two sets of results show that the multi-source vision branch does not simply increase the input channels, but makes the water surface texture, material response and flight state into a unified feature space. Compared with the single-branch structure, the full model increases the mAP by at least 4.9 percentage points, and the FPS still maintains 42.6, which can support the online reasoning of river patrol images.

4.3 Discussion

The proposed model has strong scene adaptability in the task of floating object recognition in rivers and lakes. Experiments on different river cruise scenes show that the mAP of the mixed areas of lakes, rivers and shorelines is maintained above 91%. Although the scenes of Bridges and culverts are affected by low illumination, vegetation occlusion and water surface texture, the F1-score is still maintained above 90%, which indicates that multi-source visual structure can alleviate the false detection of a single visible image under complex backgrounds. Compared with the traditional single-branch detection method, the proposed model unifies the visible texture, multi-spectral material response, and spatial state coding into the fusion feature, so that artificial floating objects such as plastic bottles, foam boards, and bags have clearer category boundaries. Ablation results also show that multispectral fusion, background suppression and confidence calibration have a direct impact on performance stability. The mAP, Accuracy and F1-score of the full model reach 91.8%, 94.2% and 92.7%, respectively, and the online processing ability is still maintained at 42.6FPS. The results show that the model design for river patrol scene cannot only pursue high accuracy, but also need to take into account the computational disturbance caused by flight altitude variation, image return quality, and complex water surface texture. The model still has some shortcomings, the feature distance between branches and algae is close, the missed detection rate of the shadow area of Bridges and culverts is high, and the candidate box screening in low-definition images also has fluctuations. In the future, temporal frame association, lightweight attention and uncertainty estimation can be further introduced to form a more robust and continuous joint judgment of target trajectory, class confidence and abnormal samples in continuous river patrol images.

5 Conclusion

Focusing on the intelligent recognition and classification task of UAV floating objects in rivers and lakes, this paper constructs a deep learning method based on multi-source UAV vision. The model takes the visible light river survey image as the main input, introduces multi-spectral features and spatial state coding, and forms a continuous computing link from image input, feature fusion, floating object localization to category discrimination. The experimental results show that the complete model achieves 91.8% mAP, 94.2% classification accuracy, 92.7% F1-score and 42.6FPS on 18420 river patrol images and 73600 labeled

targets, and maintains a relatively stable recognition effect in the mixed areas of rivers, lakes, wetlands, Bridges and culverts, and coastlines. Compared with the single visible light branch and late splicing fusion methods, the proposed model performs more stable in complex background suppression, material differential expression and candidate box confidence calibration, indicating that multi-source visual fusion can adapt to small target detection and fine-grained classification tasks in UAV river patrol images. There are still some limitations in this paper. The spectral response and texture morphology of natural floating objects are relatively close, and the branches and leaves are still easy to be confused with algae. The shadow of Bridges and culverts, strong reflective water surface and low definition return image will affect the quality of candidate frames. Although the multi-source input improves the recognition accuracy, it also increases the image registration and model inference overhead. The follow-up research can introduce the temporal correlation of consecutive video frames to enhance the consistency of the target trajectory. A lighter fusion module and quantitative reasoning strategy were designed to adapt to the edge-end deployment. Samples in different seasons, at night and in flood season were extended to construct a floating object identification dataset with wider coverage, so as to better adapt the model to the online river patrol review scene.

About the Author

Fei Chen was born in Rongcheng, Shandong, China, in 1985. He obtained a Master of Engineering degree from Three Gorges University in China. Currently, he works at Shaoxing Water Conservancy and Hydropower Survey and Design Institute Co., Ltd. His main research directions are the planning and design of water conservancy projects and related research on informatization.chenfei8478@163.com

Yin Zhang was born in Shaoxing, Zhejiang,China，in 1986 . He obtained a Master's degree in Safety Engineering from Jiangsu University in China. He is employed at the Water Conservancy Bureau of Shaoxing, His main research interests focus on the planning, protection, management and utilization of rivers and lakes.zhang1002yin@126.com

Weijia Han was born in Shaoxing, Zhejiang Province, P.R. China, in 1989. She graduated from the Yuanpei College of Shaoxing University, majoring in Civil Engineering, with a bachelor's degree. She is currently working at Shaoxing Water Conservancy and Hydropower Survey and Design Institute Co., Ltd. Her main research direction is water conservancy planning and design as well as cost estimation.770908923@qq.com

Liqun He was born in Shaoxing, Zhejiang, China in 1991. He obtained a Master of Engineering degree from Hohai University, in China. He is currently working at Shaoxing Water Conservancy and Hydropower Survey and Design Institute Co., Ltd. His research direction is the planning and consulting of water conservancy projects.wwwhlq1@qq.com

References

[1] Maharjan N, Miyazaki H, Pati B M, et al. Detection of river plastic using UAV sensor data and deep learning[J]. Remote Sensing, 2022, 14(13): 3049.

[2] Gonçalves G, Andriolo U. Operational use of multispectral images for macro-litter mapping and categorization by Unmanned Aerial Vehicle[J]. Marine Pollution Bulletin, 2022, 176: 113431.

[3] Andriolo U, Garcia-Garin O, Vighi M, et al. Beached and floating litter surveys by unmanned aerial vehicles: operational analogies and differences[J]. Remote sensing, 2022, 14(6): 1336.

[4] Goddijn-Murphy L, Williamson B J, McIlvenny J, et al. Using a UAV thermal infrared camera for monitoring floating marine plastic litter[J]. Remote Sensing, 2022, 14(13): 3179.

[5] Armitage S, Awty-Carroll K, Clewley D, et al. Detection and classification of floating plastic litter using a vessel-mounted video camera and deep learning[J]. Remote Sensing, 2022, 14(14): 3425.

[6] Sannigrahi S, Basu B, Basu A S, et al. Development of automated marine floating plastic detection system using Sentinel-2 imagery and machine learning models[J]. Marine Pollution Bulletin, 2022, 178: 113527.

[7] Gnann N, Baschek B, Ternes T A. Close-range remote sensing-based detection and identification of macroplastics on water assisted by artificial intelligence: a review[J]. Water Research, 2022, 222: 118902.

[8] Gonçalves G, Andriolo U, Gonçalves L M S, et al. Beach litter survey by drones: Mini-review and discussion of a potential standardization[J]. Environmental Pollution, 2022, 315: 120370.

[9] Corrigan B C, Tay Z Y, Konovessis D. Real-time instance segmentation for detection of underwater litter as a plastic source[J]. Journal of Marine Science and Engineering, 2023, 11(8): 1532.

[10] Alboody A, Vandenbroucke N, Porebski A, et al. A new remote hyperspectral imaging system embedded on an unmanned aquatic drone for the detection and identification of floating plastic litter using machine learning[J]. Remote Sensing, 2023, 15(14): 3455.

[11] Andriolo U, Topouzelis K, van Emmerik T H M, et al. Drones for litter monitoring on coasts and rivers: suitable flight altitude and image resolution[J]. Marine pollution bulletin, 2023, 195: 115521.

[12] Zaaboub N, Guebsi R, Chaouachi R S, et al. Using unmanned aerial vehicles (UAVs) and machine learning techniques for the assessment of Posidonia debris and marine (plastic) litter on coastal ecosystems[J]. Regional Studies in Marine Science, 2023, 67: 103185.

[13] Saeed Z, Yousaf M H, Ahmed R, et al. On-board small-scale object detection for unmanned aerial vehicles (UAVs)[J]. Drones, 2023, 7(5): 310.

[14] Dewangan V, Saxena A, Thakur R, et al. Application of image processing techniques for uav detection using deep learning and distance-wise analysis[J]. Drones, 2023, 7(3): 174.

[15] Terven J, Córdova-Esparza D M, Romero-González J A. A comprehensive review of yolo architectures in computer vision: From yolov1 to yolov8 and yolo-nas[J]. Machine learning and knowledge extraction, 2023, 5(4): 1680-1716.

[16] Diwan T, Anirudh G, Tembhurne J V. Object detection using YOLO: challenges, architectural successors, datasets and applications[J]. multimedia Tools and Applications, 2023, 82(6): 9243-9275.

[17] Zaidi S S A, Ansari M S, Aslam A, et al. A survey of modern deep learning based object detection models[J]. Digital Signal Processing, 2022, 126: 103514.

[18] Dadrass Javan F, Samadzadegan F, Gholamshahi M, et al. A modified YOLOv4 Deep Learning Network for vision-based UAV recognition[J]. Drones, 2022, 6(7): 160.

[19] Jawaharlalnehru A, Sambandham T, Sekar V, et al. Target object detection from Unmanned Aerial Vehicle (UAV) images based on improved YOLO algorithm[J]. Electronics, 2022, 11(15): 2343.

[20] Kikaki K, Kakogeorgiou I, Mikeli P, et al. MARIDA: A benchmark for Marine Debris detection from Sentinel-2 remote sensing data[J]. PloS one, 2022, 17(1): e0262247.

[21] Nunkhaw M, Miyamoto H. An image analysis of river-floating waste materials by using deep learning techniques[J]. Water, 2024, 16(10): 1373.