frontiers | Frontiers in Neurorobotics

TYPE Original Research
PUBLISHED 14 December 2023
DOI 10.3389/fnbot.2023.1315251

OPEN ACCESS

EDITED BY
Ming-Feng Ge,
China University of Geosciences Wuhan, China

REVIEWED BY
Lixiong Gong,
Hubei University of Technology, China
Hua Yin,
Jiangxi Agricultural University, China
Junsuo Qu,
Xi'an University of Post and Telecommunications, China

*CORRESPONDENCE
Mei Yu
✉ yumei@ctgu.edu.cn

RECEIVED 10 October 2023
ACCEPTED 23 November 2023
PUBLISHED 14 December 2023

CITATION
Yu J, Zheng H, Xie L, Zhang L, Yu M and Han J (2023) Enhanced YOLOv7 integrated with small target enhancement for rapid detection of objects on water surfaces.
Front. Neurorobot. 17:1315251.
doi: 10.3389/fnbot.2023.1315251

Enhanced YOLOv7 integrated with small target enhancement for rapid detection of objects on water surfaces

Jie Yu[1,2,3], Hao Zheng[1,2], Li Xie[3], Lei Zhang[1,2], Mei Yu[1,2]* and Jin Han[3]

[1]Hubei Key Laboratory of Intelligent Vision Based Monitoring for Hydroelectric Engineering, School of Computer and Information, China Three Gorges University, Yichang, China, [2]School of Computer and Information, China Three Gorges University, Yichang, China, [3]State Grid Yichang Electric Power Supply Company, Yichang, China

Unmanned surface vessel (USV) target detection algorithms often face challenges such as misdetection and omission of small targets due to significant variations in target scales and susceptibility to interference from complex environments. To address these issues, we propose a small target enhanced YOLOv7 (STE-YOLO) approach. Firstly, we introduce a specialized detection branch designed to identify tiny targets. This enhancement aims to improve the multi-scale target detection capabilities and address difficulties in recognizing targets of different sizes. Secondly, we present the lite visual center (LVC) module, which effectively fuses data from different levels to give more attention to small targets. Additionally, we integrate the lite efficient layer aggregation networks (L-ELAN) into the backbone network to reduce redundant computations and enhance computational efficiency. Lastly, we use Wise-IOU to optimize the loss function definition, thereby improving the model robustness by dynamically optimizing gradient contributions from samples of varying quality. We conducted experiments on the WSODD dataset and the FIOW-Img dataset. The results on the comprehensive WSODD dataset demonstrate that STE-YOLO, when compared to YOLOv7, reduces network parameters by 14% while improving AP50 and APs scores by 2.1% and 1.6%, respectively. Furthermore, when compared to five other leading target detection algorithms, STE-YOLO demonstrates superior accuracy and efficiency.

KEYWORDS

object detection, water target detection, YOLOv7, unmanned surface vessel, small object detection

1 Introduction

Unmanned surface vessel (USV) are now widely used in the fields of harbor surveillance, fisheries monitoring, maritime management, and military intelligence analysis, such as target detection and environmental monitoring (Zhang et al., 2022, 2023; Zhou et al., 2022). Equipped with cameras, lasers, and an array of sensors, USVs enable autonomous detection and recognition of their surroundings. This capability ensures the safe and efficient navigation of unmanned vessels, while also enhancing the effectiveness of monitoring tasks. Target detection technology plays a pivotal role in evaluating the performance of unmanned vessels for object detection and recognition tasks. Deep learning-based target detection algorithms have garnered significant attention due to their precision in identifying objects. However, the high-speed mobility of unmanned vessels introduces considerable scale

variations in the objects detected. This demands robust multi-scale detection capabilities from the algorithm.

At present, several researchers are actively delving into the realm of waterborne target detection. Moosbauer et al. (2019) publicly released the Singapore Maritime Dataset (SMD), which has provided invaluable resources to drive progress in the field of sea surface target detection from a horizontal perspective. Shin et al. (2020) pioneered the utilization of instance segmentation techniques to extract ship targets from the SMD dataset. These targets were subsequently merged with ocean backgrounds, resulting in a synthetic dataset that significantly enhances the precision of sea surface target detection. Chen et al. (2018) curated a dataset comprising 1,500 images of sea surface targets, drawing from three distinct sources: MS COCO (Lin et al., 2014), Pascal VOC (Everingham, 2010), and SMD. This dataset played a pivotal role in validating their proposed hierarchical, multi-scale deep convolutional neural network-based sea surface target detection algorithm. Furthermore, Zhou et al. (2021) have contributed to the field by curating the WSODD dataset, which stands as a notable advancement in the domain of water target detection, Cheng et al. (2021) collected a variety of local floating garbage to form the FloW dataset—the world's first unmanned ship-view of floating garbage detection dataset, promoting the rapid development of floating garbage detection technology.

To address the limitations of existing water target detection algorithms, particularly in the realms of multi-scale and small target detection, we present STE-YOLO, an enhanced model built upon the foundation of YOLOv7. The overview of the detection pipeline utilizing STE-YOLO is depicted in Figure 1. This framework achieves a reduction in network size by meticulous redesign and optimization of the network structure. Additionally, it integrates multi-scale and multi-level information, augmenting the network capacity to characterize objects effectively. Moreover, a residual module is incorporated to heighten the model sensitivity to small targets. The contributions of this paper are delineated as follows:

1. We establish a specialized detection head, focused on precisely detecting small targets. This strategic design empowers the network to adeptly leverage shallow-level information, thereby enhancing its efficacy in identifying small targets. Consequently, this improvement extends to the network's overall capability to detect targets across diverse scales.
2. The introduction of the lite visual center (LVC) module seamlessly merges coordinate convolution with target-relative positional information within the network architecture. This fusion facilitates refined feature extraction from the target region, thereby intensifying attention and precision in detecting small targets. This integration culminates in an elevated overall performance in target detection.
3. The integration of the lite efficient layer aggregation networks (L-ELAN) module into the network results in reduced parameters and operations, all while upholding accuracy.
4. Optimization of the loss function computation is achieved through the utilization of Wise-IOU (Tong et al., 2023), thereby boosting the model confidence level and enhancing its robustness.

2 Related works

2.1 Object detection based on deep learning

Modern deep learning-based methods for target detection can be categorized into two primary types: two-stage target detectors and single-stage target detectors. Two-stage target detectors, represented by architectures like Mask-RCNN (He et al., 2017), VFNet (Zhang et al., 2021), and CenterNet (Zhou et al., 2019), generate a set of region proposals using a region selection network. These proposals then undergo feature extraction through a learning module, followed by classification and regression processes. However, this method of extracting features for each proposal can lead to significant computational costs and might not capture a comprehensive global feature representation effectively. On the other hand, prevalent single-stage detectors, exemplified by networks like YOLOX (Ge et al., 2021), FCOS (Tian et al., 2022), Scaled-YOLOv4 (Wang et al., 2021), and EfficientDet (Tan et al., 2020), rely on a backbone network to extract feature maps for the entire input image. These feature maps are subsequently used to predict bounding boxes, enabling concurrent prediction and classification through box generation.

Furthermore, in terms of the network's architectural components, it can be divided into two primary elements: the Backbone, responsible for extracting image features, and the Head, utilized for predicting object categories and bounding box coordinates. Additionally, some researchers have introduced an intermediary module known as the Neck between the Backbone and the Detection Head to optimize detection performance.

Backbone. Commonly utilized backbone networks include ResNet (Wightman et al., 2021), CSPDarknet53 (Bochkovskiy et al., 2020), Swin Transformer (Liu et al., 2021), and FasterNet (Chen J. et al., 2023). These backbones exhibit robust feature extraction capabilities, particularly for classification tasks. Typically, researchers only need to fine-tune the backbone to optimize its performance for specific tasks.

Neck. The Neck network is designed to enhance the utilization of features derived from the Backbone network. It accomplishes this by reprocessing the feature maps extracted by the Backbone at various stages. The Neck network typically consists of multiple bottom-up and top-down pathways. What sets this network apart is its direct multi-stage feature mapping approach, which omits feature layer aggregation operations and aligns directly with the Head. Prominent Neck network architectures include PANet (Wang et al., 2019), NAS-FPN (Ghiasi et al., 2019), and SFAM (Zhao et al., 2019). These architectures often utilize combinations of up-sampling and down-sampling iterations, concatenation, element-wise summation, and dot products to establish effective aggregation strategies. Additionally, complementary modules like MSFFM (Wan et al., 2023), ASPP (Weber et al., 2021), SPPCSPC (Wang et al., 2023), and GFFAP (Sun et al., 2021) are incorporated into Neck networks to enhance feature fusion and improve detection accuracy. Adding an attention module is also a great way to do this, such as MFINEA (Sun et al., 2023) and Box-Attention (Nguyen et al., 2022).

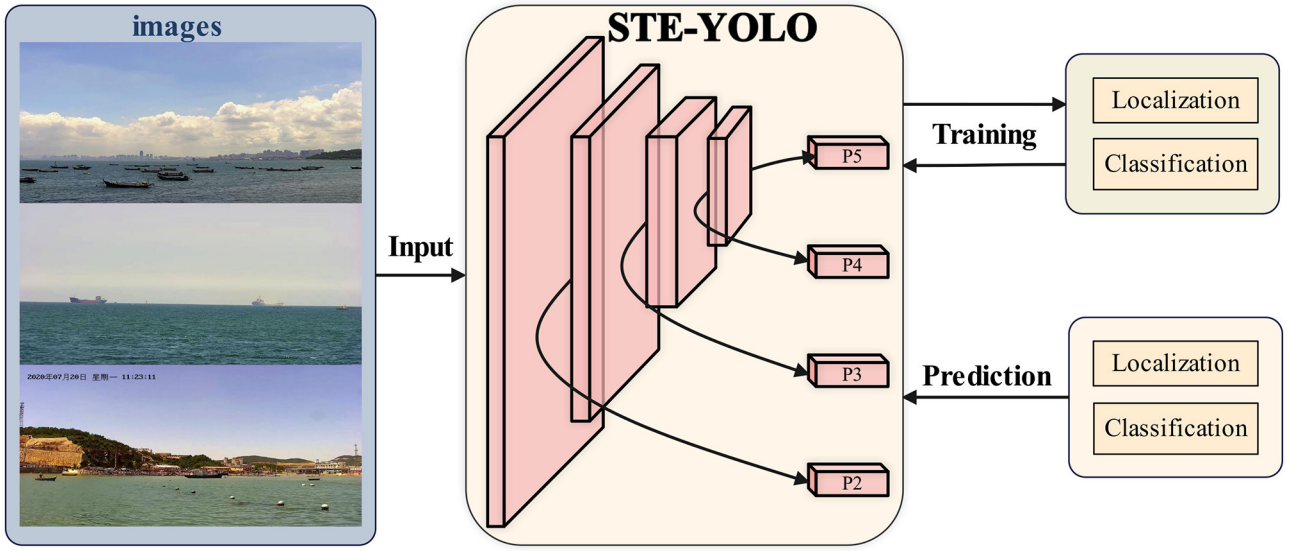

FIGURE 1
The overview of working pipeline using STE-YOLO. Compared to original version, we mainly improve the method by adding one more head (P2) to better detect different scale objects. In addition, we also employ LVC and L-ELAN to make STE-YOLO stronger.

Head. While the backbone network primarily functions as a classification network, it is insufficient to perform the localization task independently. Hence, the head network plays a crucial role in achieving both target localization and categorization, utilizing the feature maps extracted by the backbone network. Head networks are generally divided into two main categories: single-stage detectors and two-stage detectors (He et al., 2017; Ren et al., 2017). stands out as a notable representative of two-stage detectors. On the other hand, one-stage detectors offer faster prediction by simultaneously estimating bounding boxes and target categories, but they may sacrifice accuracy. Due to the real-time constraints in water target detection, a majority of algorithms used in this domain opt for one-stage detectors, such as YOLO (Wang et al., 2023) and SSD (Zalesskaya et al., 2022) .

2.2 Small target detection

Small target detection presents significant challenges within the realm of target detection, necessitating algorithms with robust fine feature extraction capabilities (Shamsolmoali et al., 2022; Gong, 2023). Typically, two primary criteria are employed to define small targets: absolute size and relative size. In terms of absolute size, targets with dimensions smaller than 32×32 pixels are categorized as small. In the case of relative size, targets with an aspect ratio <0.1 times the original image size are considered small (Lin et al., 2014). Currently, small target detection algorithms fall into three main categories: those utilizing data augmentation, those emphasizing multi-scale learning, and those leveraging contextual information.

Data augmentation. Kisantal et al. (2019) elevated the percentage of small targets within the dataset through replication, thereby enhancing their significance in the network. This augmentation aimed to bolster the model's proficiency in detecting small targets. Yu et al. (2020) introduced the scale-matching strategy, aligning pre-trained network features with those obtained by the detector. This strategy ensures the comprehensive utilization of pre-trained network capabilities, thereby enhancing overall performance.

Multiscale learning. The deep network has large receptive field and strong representation ability of semantic information, but weak representation ability of geometric information. The lower layer network has relatively small receptive field and strong representation ability of geometric details, but weak representation ability of semantic information. Thus, Multiscale learning often enhances network representation ability by fusing shallow detail information with deep semantic information, thus improving small target detection. However, multiscale learning can increase parameters and slow down inference speed. The Feature Pyramid Network (FPN), proposed by Lin et al. (2017), is a classic multiscale learning network structure. In FPN, the image undergoes bottom-up feature extraction, followed by top-down feature fusion, before being fed into the detection head for regression prediction. Deng et al. (2022) extended this approach with the Enhanced Feature Pyramid Network (EFPN), which incorporates a feature texture migration module for ultra-high-resolution feature extraction, further enhancing small target detection.

Utilization of contextual information. Zhu et al. (2021) introduced TPH-YOLOv5, a novel strategy that integrates the Transformer (Vaswani et al., 2017) into the prediction head of YOLOv5 (Jocher et al., 2022). This integration enhances predictive regression capabilities while employing an attention mechanism to focus intensively on small targets. In a different vein, QueryDet (Yang et al., 2022) utilizes a querying mechanism

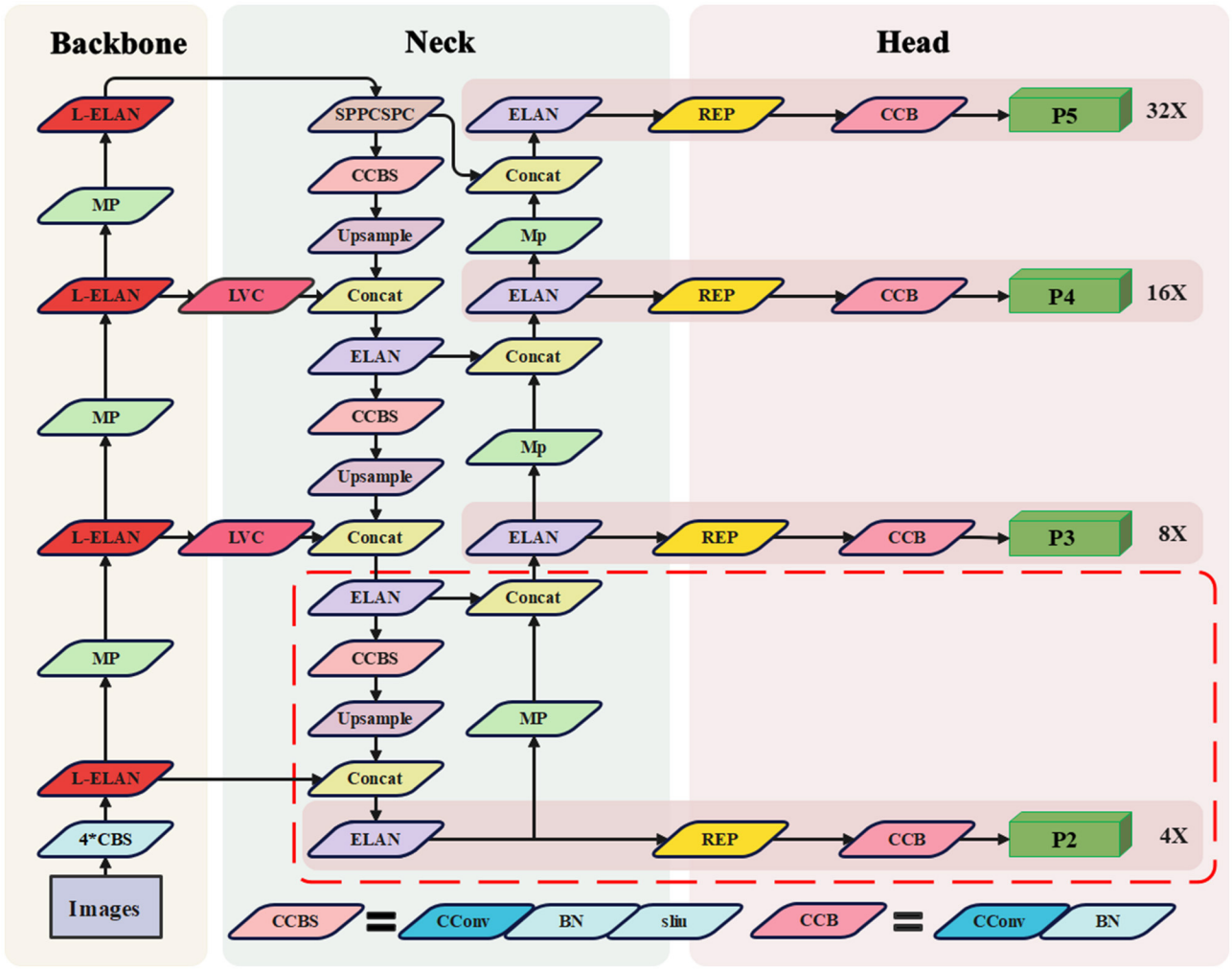

FIGURE 2
Architecture of STE-YOLO. CSPDarknet53 backbone with four L-ELAN blocks. The Neck use the structure like PANet with LVC. Four prediction heads use the different size of feature maps Neck. The CBS module is a basic module, which contains three operations of convolution, normalization and activation function, the CConv means CoordConv.

to expedite target detector inference. It leverages low-resolution features for preliminary localization predictions, which then guide higher-resolution features, contributing to increased accuracy in predictive regression.

3 STE-YOLO

YOLOv7 (Wang et al., 2023) stands out as a highly proficient one-stage detector, renowned for its exceptional overall performance. The model structure comprises three central components: Backbone, Neck, and Head. The YOLOv7 framework incorporates a wide range of advanced techniques, substantially enhancing its detection capabilities. Operating as a one-stage detector, YOLOv7 boasts impressive real-time processing capabilities, rendering it particularly well-suited for meeting the demands of real-time water-based target detection.

However, due to the susceptibility of water target detection to adverse weather conditions like storms, coupled with the high-speed nature of USVs leading to significant variations in target scale, YOLOv7 encounters certain challenges. These include limitations in recognizing targets across diverse scales, which in turn reduces accuracy in rapidly detecting small targets in water scenarios. With a focused approach, our aim is to enhance YOLOv7's proficiency in detecting small targets while optimizing computational efficiency. Our proposed solution is a waterborne rapid target recognition algorithm named STE-YOLO. The architectural diagram, illustrated in Figure 2, introduces a novel detection branch known as P2 head, represented by dashed lines. The P2 head primarily enhances the detection performance of small targets while simultaneously accommodating multi-scale target detection. The LVC module employs coordinate convolution to amplify focus on small targets, replacing the original fusion operation in the FPN structure and amalgamating multi-level information to enhance detection accuracy. Additionally, the L-ELAN module, a lightweight feature extraction addition, aims to reduce network computation while bolstering network robustness and computational efficiency. Furthermore, we have fine-tuned

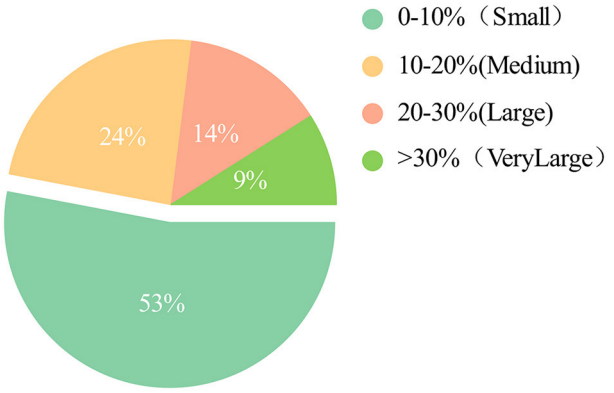

FIGURE 3
Percentage of targets of different sizes in the WSODD dataset.

the loss function and integrated Wise-IOU to achieve improved detection performance.

3.1 Small target head P2

In the context of actual waterborne target detection tasks, encountering a substantial volume of small target samples (Zhou et al., 2021) is common. Through in-depth data analysis of the prevalent water target dataset WSOOD, a notable insight has emerged. This dataset comprises an impressive 53% of small targets, as depicted in Figure 3. Consequently, the task of waterborne target detection demands that the model possesses a strong capability to extract localized information, persisting across the intricate layers of the network. This is crucial to ensure the comprehensive retention of information relevant to small targets. YOLOv7 adheres to the downsampling mechanism typical of conventional convolutional networks, where increasing depths lead to augmented downsampling factors. This mechanism fosters a wider spatial perception, which is advantageous for recognizing overarching target contours. However, this approach also carries the risk of losing detailed information, which could significantly hinder small target recognition.

With this challenge in mind, we have developed a specialized prediction head network called P2 head that places a primary focus on recognizing small targets. As an innovative extension, a shallow feature obtained through a 4× downsampling rate is introduced as one of the inputs to the neck network. Subsequently, the neck network orchestrates the fusion of four distinct sets of features, each corresponding to varying scales downsampling rates of 4×, 8×, 16×, and 32×. These amalgamated features are then directed to the head network, resulting in the creation of four distinct detector head structures. Each of these structures is optimized for detecting objects of different sizes. The architectural arrangement is vividly depicted in Figure 4.

The newly introduced P2 detection branch is meticulously crafted for the explicit purpose of detecting exceedingly small targets. Given the inherent possibility of significant information loss in deeper feature maps due to consecutive convolutional pooling, and the risk of larger target features overpowering those of smaller targets, a challenge of misdetection and omission emerges. As a response, there is a proactive approach in place for the input to the P2 detection branch structure, predominantly originating from the shallow convolutional layer. This particular layer encapsulates a wealth of localized information, spanning attributes such as shape, position, and size. Consequently, it greatly assists in precisely localizing small targets. This strategic enhancement effectively bolsters the efficiency of small target detection, all the while catering to the broader capabilities of multi-scale target detection. Furthermore, this architectural extension comprises four distinct detector head structures, each serving as a mitigation strategy against the adverse consequences of significant variations in target scale. This configuration, in turn, ultimately contributes to the elevation of the comprehensive detection performance.

Furthermore, YOLOv7 exhibits a high degree of adaptability to the configuration of the anchor frame dimensions. Therefore, when utilizing the P2 detection branch for predictive regression, meticulous assessment of the anchor frame dimensions becomes imperative. This evaluation involves a tailored K-means cluster analysis that aligns with the dataset's characteristics. Its goal is to determine the most suitable anchor frame size. Once this determination is made, the anchor frame settings for each branch are established, as outlined in Table 1. As depicted in Table 1, the P2 detection branch holds the potential to address scenarios where the target object might evade detection due to its diminutive size combined with an excessively large anchor frame. The strategic alignment effectively alleviates challenges of misdetection and omission that may arise from suboptimal anchor frame configurations P3 head. This approach ensures a robust and reliable detection process by fine-tuning the anchor frame settings to harmonize with the unique characteristics of the detection task.

3.2 Lite visual center–LVC

Given the intricate and ever-changing nature of the aquatic environment, a multitude of dynamic factors come into play. These factors encompass light reflections, precipitation, interference induced by fog, and the constant fluctuations in wind and wave patterns. These phenomena often converge to create challenging conditions. These conditions, frequently presenting as shadows, reflections, or image blurring, severely compromise the visibility of targets. Consequently, the intricacies associated with target identification become more pronounced. Therefore, there is a compelling imperative to devise innovative and efficient strategies for feature extraction. These strategies are specifically designed to enhance the capacity for acquiring both broad context and localized information. This strategic endeavor assumes paramount importance in the pursuit of heightening the overall effectiveness of detection.

While YOLOv7 employs a singular convolutional fusion to link the backbone network with the neck network, effectively facilitating the extraction and fusion of features across diverse levels and thereby enhancing the capability for multi-level feature extraction to some extent, its adequacy in recognizing small targets falls short. This insufficiency gives rise to the introduction of the lite visual

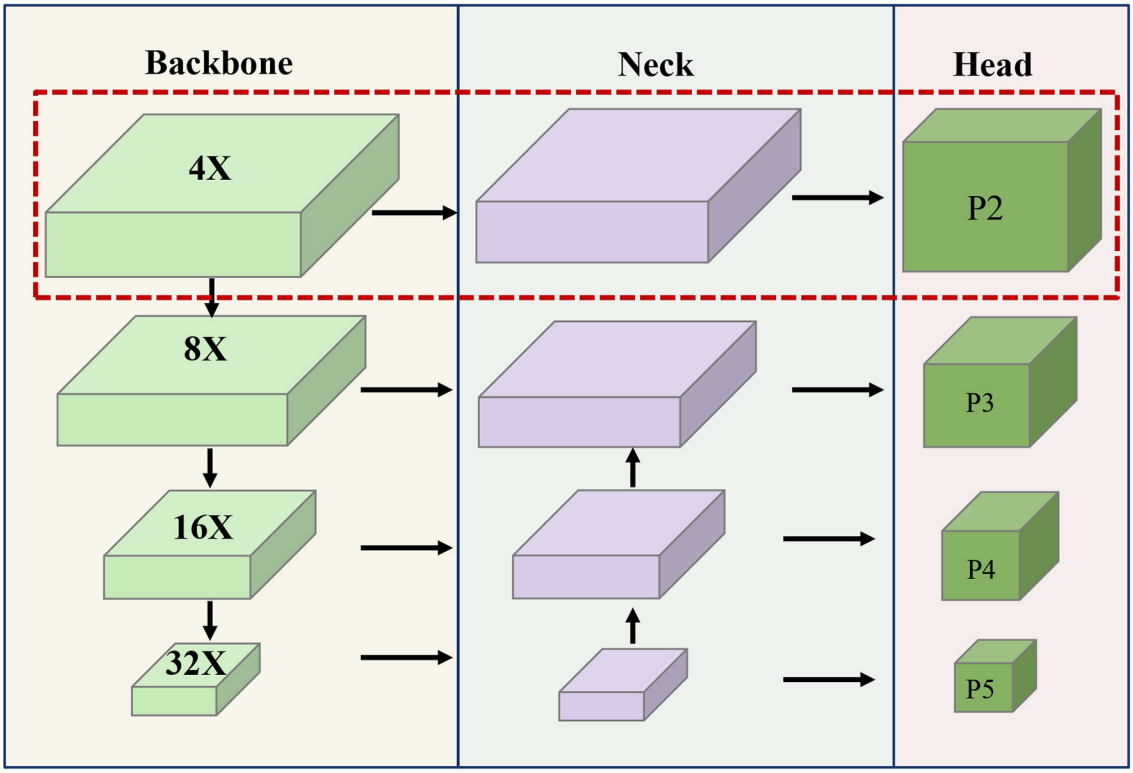

FIGURE 4
The overview of Head network optimization results, the portion of the red box is the newly added P2 detector head.

TABLE 1 The head setting of STE-YOLO.

Prediction head	Anchor box setting
P2	[8,10, 9,22, 17,16]
P3	[10,13, 16,30, 33,23]
P4	[30,61, 62,45, 59,119]
P5	[116,90, 156,198, 373,326]

center (LVC) module. This module is inspired by the structural blueprint of EVC (Quan et al., 2023), which is present within the framework of CFP. The architectural details of this novel module are visually represented in Figure 5.

Within the LVC module, two critical components are seamlessly integrated the multilayer perceptron [MLP (Tolstikhin et al., 2021)] and CCBS [CoordConv (Liu et al., 2018) with bilateral strategy]. The incorporation of the MLP component serves a pivotal role in capturing extensive global long-term dependencies within deep features, which results in the encapsulation of global information and ultimately enhancing the precision of holistic recognition. Concurrently, the introduction of CCBS involves the application of CoordConv, a convolutional mechanism enriched with relative position coordinate information during the convolution process. This integration empowers the network to discern and ascertain the relative positioning of targets more effectively through the amalgamation of localized area features. Ultimately, the feature maps originating from the MLP and CCBS modules are amalgamated across the channel dimension, thus constituting the output of the LVC module. This amalgamated feature suite seamlessly integrates the strengths of both modules, preserving significant information concerning elements within the block or pool, and capable of supplying more comprehensive and enriched multi-level fusion features. Consequently, the detection model is able to acquire a holistic range of feature representations, thereby amplifying the recognition capacity for small targets.

Simultaneously, this paper undertakes the replacement of selected CBS modules in the original model with CCBS modules, allowing the network to harness the amalgamated positional coordinate data and global information, ultimately enhancing the overall system performance. These innovative addition serves as a means to further enrich the model feature extraction capabilities, especially in scenarios where small targets play a pivotal role.

3.3 L-ELAN

In the realm of water target detection tasks, the inherent limitations of equipment often require the downsizing of the model for enhanced practical applicability. With this consideration, a L-ELAN lightweight module is put forth, meticulously striking a balance between model accuracy and computational efficiency. The schematic representation of this module can be observed in Figure 6. The water target detection task, due to the limitation

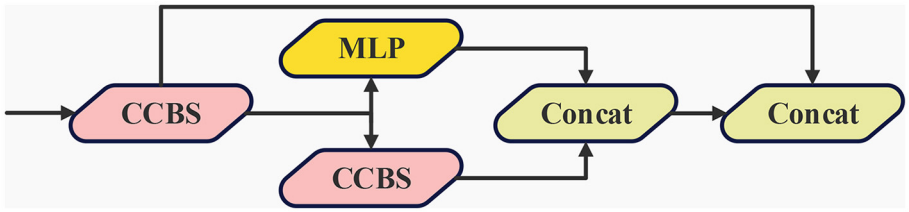

FIGURE 5
The architecture of LVC, which contains two main blocks, a Multilayer Perceptron (MLP) and CCBS, residual paths are also used.

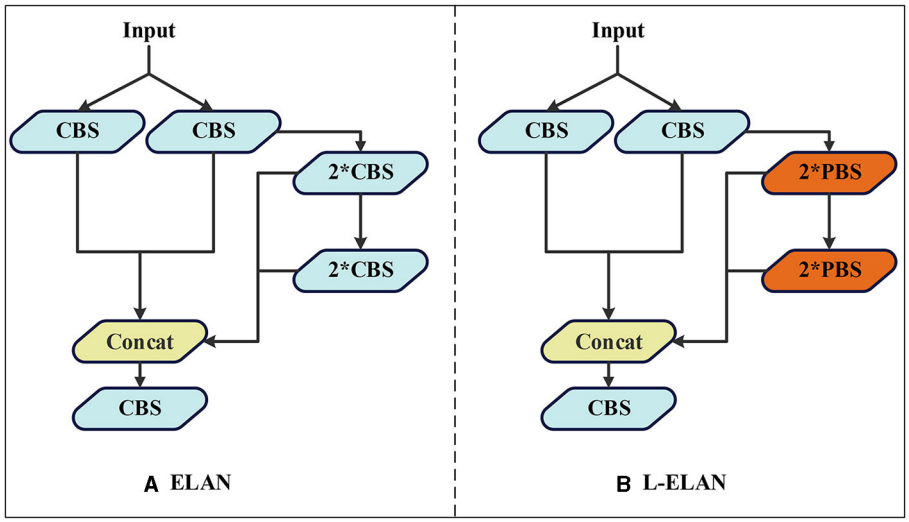

FIGURE 6
The structure of ELAN **(A)** and L-ELAN **(B)**, where PBS modules are combinations of Pconv, BN, and silu activation functions, and all PBS modules in the ELAN structure are CBS.

of equipment, it is often necessary to reduce the model size as much as possible for ease of use. For this reason, a lightweight module L-ELAN is proposed by considering model accuracy and computational efficiency, as shown in Figure 6.

In the original YOLOv7 framework, the efficient layer aggregation networks (ELAN) module predominantly employs conventional convolutions for feature extraction. While this methodology ensures a commendable level of detection accuracy, it doesn't inherently excel in terms of computational efficiency. In response to this challenge, the lite efficient layer aggregation networks (L-ELAN) module is introduced. It integrates positional convolution [PConv (Chen J. et al., 2023)] to replace specific segments of the traditional convolution processes The structure of PCONV is shown in Figure 7. This strategic substitution leverages the inherent traits of PConv minimal computational requirements and heightened efficiency. The integration of PConv within L-ELAN serves a dual purpose: enhancing the network computational efficiency and reducing the overall count of network parameters. This innovative enhancement aligns with the overarching goal of optimizing the network performance not only in terms of detection accuracy but also in terms of computational resource utilization. By strategically selecting and incorporating PConv within the L-ELAN module, the model achieves a balance between accuracy and efficiency a crucial consideration in real-time waterborne target detection scenarios.

When handling input $\mathbf{I} \in \mathbb{R}^{c\times h\times w}$, conventional convolution utilizes c filters $\mathbf{W} \in \mathbb{R}^{k\times k}$ to perform computations and generate an output $\mathbf{O} \in \mathbb{R}^{c\times h\times w}$. In contrast, PConv employs standard convolution for spatial feature extraction exclusively on select input channels, while leaving the remaining channels unaffected. This methodology ensures an alignment between the number of channels in the input and the resulting output feature maps. The computational cost, quantified in terms of FLOPS (Floating Point Operations Per Second), associated with the standard convolution operation is as follows:

$$FLOPs = h \times w \times k^2 \times c \tag{1}$$

where h, w is the size of the feature map, k is the size of the filters, c is the number of filters used. The FLOPS calculation formula for PConv is as follows:

$$FLOPs = h \times w \times k^2 \times c_p^2. \tag{2}$$

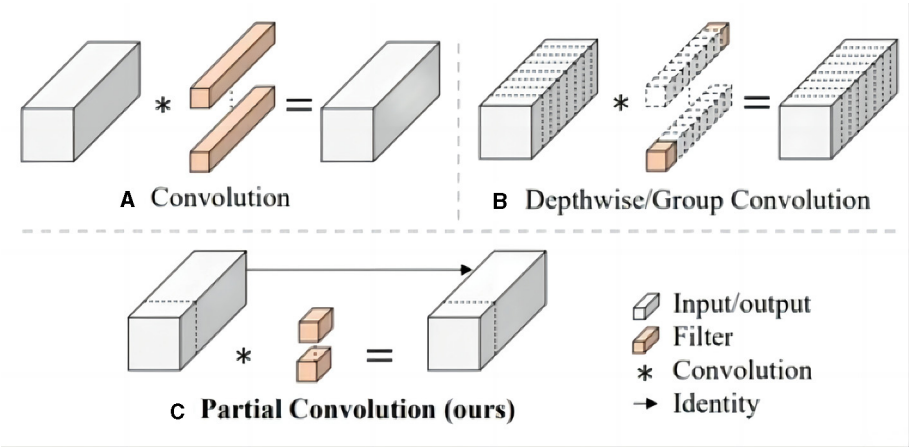

FIGURE 7
The structure of Conventional convolution **(A)**, Dconv **(B)**, and Pconv **(C)**, the Pconv is only applied to a part of the input channel for spatial feature extraction without a affecting the rest of the channels.

where c_p is the number of partial channels. When the convolution channel percentage $r = \frac{c_p}{c} = \frac{1}{4}$, the FLOPS of PConv is just $\frac{1}{16}$ of the traditionally convolution.

Furthermore, PConv exhibits reduced memory accesses (M), which can be computed as demonstrated in Eq. 3, while the conventional convolution is represented by Eq. 4.

$$M = h \times w \times 2c_p + k^2 \times c_p^2 \approx h \times w \times 2c_p \tag{3}$$

$$M = h \times w \times 2^c + k^2 \times c^2 \approx h \times w \times 2c \tag{4}$$

The computational effort is merely a quarter of that required by the normal convolution when r equals $\frac{1}{4}$.

The analysis provided above leads to a clear realization that the incorporated L-ELAN module within this study achieves two pivotal objectives. Primarily, it significantly reduces the computational load while simultaneously effectively addressing memory access demands. This harmonious accomplishment notably streamlines and expedites the practical feasibility of real-world deployment.

3.4 Wise-IOU

In the realm of water target recognition tasks, it's commonplace to encounter a notable prevalence of suboptimal samples, particularly within datasets that display limitations and imbalances (Chen X. et al., 2023). After thorough examination, this study establishes that the CIOU loss function, when integrated into the original YOLOv7 model, exerts a substantial negative impact on the cumulative regression loss. This issue predominantly arises from samples characterized by subpar regression quality. Concurrently, it introduces difficulties in effectively fine-tuning samples that possess relatively higher regression quality, thereby compromising the overall effectiveness of recognition. To navigate these challenges, this paper deliberately chooses to adopt the Wise-IOU loss function (WIOUv3 version), as outlined in Eq. 5.

$$\mathcal{L}_{WIOU_{v3}} = rR_{WIOU}\mathcal{L}_{IOU} \tag{5}$$

where $\mathcal{L}_{IOU} \in (0, 1)$ is the ratio of the intersection of the prediction frame and the true frame, $R_{WIOU} \in [1, e)$ means distance attention, r is non-monotonic focusing factor, which can adaptively adjust the gradient gain assignment strategy according to the degree of outliers of the anchor frame.

The Wise-IOU loss function incorporates a dynamic non-monotonic focusing mechanism, leveraging "outliers" in place of IOUs for the assessment of anchor frame quality. This approach also encompasses an intelligent strategy for assigning gradient gains. The expressions for R_{WIOU} and r are outlined below:

$$R_{WIOU} = \exp\left(\frac{(x - x_{gt})^2 + (y - y_{gt})^2}{(W_g{}^2 + H_g{}^2)^*}\right) \tag{6}$$

$$r = \frac{\beta}{\delta\alpha^{\beta-\delta}} \tag{7}$$

where x and y denote the anchor frame dimensions in terms of length and width, while W_g and H_g represent the dimensions of the minimum enclosing frame. The symbol (*) serves to mark the detachment of W_g and H_g from the computational graph, effectively excluding them from participating in backpropagation.

This measure is undertaken to avoid the emergence of gradients that might hinder convergence, particularly concerning the R_{WIOU} computation. β assumes significance as the outlier value, functioning as a descriptor for the anchor frame quality. δ and α take on the role of hyperparameters This description takes the following form.

$$\beta = \frac{\mathcal{L}^*_{IOU}}{\mathcal{L}_{IOU}} \in [0, +\infty) \tag{8}$$

Due to the small data set, we hope that the model can reach the high gradient gain earlier in the training process, so we choose to increase and decrease to improve the speed of reaching the peak, so that the model anchor frame can obtain the highest gradient gain earlier. The final experimental results also prove this.

In our established framework, a smaller departure from the norm signifies an anchor frame boasting heightened quality. Consequently, these high-quality anchor frames receive minimal gradient gains, directing the emphasis of bounding box regression toward anchor frames of average quality. Conversely, anchor frames exhibiting larger deviations are coupled with relatively subdued gradient gains, effectively countering the emergence of unfavorable gradients originating from subpar samples. This strategy not only tempers the influence wielded by high-quality anchor frames but also mitigates the adverse gradients arising from low-quality instances. This carefully calculated approach serves to harmonize the weighting assigned to a diverse spectrum of samples, steering attention toward anchor frames that epitomize an intermediate standard. This intentional focal point on anchor frames with moderate performance significantly fortifies the model resilience, ultimately yielding a conspicuous enhancement in the overall performance of the detection system.

4 Results

4.1 Implementation details

In the scope of this research, experimentation unfolds within the Ubuntu environment, specifically the 18.04.6 version. Both training and testing occur on a sole NVIDIA RTX3090 GPU. The chosen deep learning framework is PyTorch 1.10.1, harnessed alongside CUDA version 11.7. The model foundational weights stem from prior training on the COCO dataset. The encompassing training regimen spans 200 epochs, commencing with an initial learning rate of 0.01, which subsequently tapers to 0.001 in the concluding cycle set banchsize to 1 when calculating FPS. The training pipeline leverages the Stochastic Gradient Descent (SGD) optimization technique, with the incorporation of momentum, anchored at 0.937. Simultaneously, the weight decay coefficient stands at 0.0005. Importantly, the model operates on a consistent scale of input images, each adopting dimensions of 640 × 640 pixels, while each training batch encompasses 12 samples.

Within this study, we use the WSODD dataset Zhou et al. (2021), which is commonly used in water target detection, and consists of images from oceans, lakes, and rivers with different climatic conditions and shooting times, with a total of 7,467 images, and the resolution of each image is 1,920*1,080. In addition, the dataset consists of a total of 14 common object classes and 21,911 instances . In this paper, it is divided into training set, testing set, and validation set, and the performance of the model is evaluated with reference to the criteria of COCO dataset (Lin et al., 2014).

We also used the FIOW-Img subdataset Cheng et al. (2021), which is the world's first floating garbage detection dataset in a real inland river scene from the perspective of an unmanned ship. The FloW-Img subdataset contains 2,000 images and 5,271 marked targets. One thousand two hundred images are randomly selected as the training set and the rest as the verification set and test set. Small targets (size in 32 × 32) in this dataset account for a large proportion (about 60%), which is beneficial for testing the performance of relevant algorithms on small targets.

TABLE 2 Ablation study of proposed method on WSODD test dataset.

Methods	AP50 (%)↑	FLOPs (G)↓	Para (M)↓
YOLOv7	81	103.4	34.56
YOLOv7 + P2	81.9 (↑0.9)	107.5 (↑4.2)	36.07 (↑1.21)
YOLOv7 + LVC	81.7 (↑0.7)	105.6 (↑2.3)	35.89 (↑1.03)
YOLOv7 + L-ELAN	81.3 (↑0.3)	83.1 (↓20.3)	30.56 (↓4.3)
YOLOv7 + Wise-IOU	81.6 (↑0.6)	103.4 (−)	34.86 (−)
STE-YOLO	83.1 (↑2.1)	89.6 (↓13.8)	32.8 (↓2.06)

↑ Means the larger the better, ↓ means the smaller the better.

4.2 Ablation studies

To assess the efficacy of the proposed P2 head, LVC, L-ELAN, Wise-IOU presented in this paper, we conducted a series of comparative experiments. The outcomes of these experiments are detailed in Table 2.

4.2.1 P2 head

The integration of the P2 Head to enhance small object detection slightly increases the network parameters. However, this adjustment corresponds to a notable 0.8% enhancement in the AP value. This outcome underscores the beneficial influence of the P2 Head in bolstering the detection of small targets.

4.2.2 LVC

Upon incorporation of the LVC module, the computational complexity of the model in this study registers a 2% growth in GFLOPs and a 1% increase in parameters. Remarkably, the AP50, a key performance metric, experiences a substantial 0.9% improvement. This observation substantiates the effectiveness of the LVC module in significantly enhancing target recognition rates.

4.2.3 L-ELAN

Upon activation of the L-ELAN module, the GFLOPs of the model in this paper reduce from 103.4 to 83.1, accompanied by a parameter reduction from 34.86 to 30.56. This translates to a reduction of 20% and 13% respectively. Notably, this reduction in computational demands is accompanied by a 0.3% increase in AP. The implications of these findings are twofold: the L-ELAN module not only achieves an effective reduction in network size, but also plays a pivotal role in elevating the accuracy of target detection.

4.2.4 Further validation of L-ELAN

To further substantiate the performance of the L-ELAN module, a comparative evaluation involving distinct convolutional modules (Yang et al., 2019);

(Zhu et al., 2019) is carried out. The ensuing results are detailed in Table 3. Notably, the use of the PConv module leads to an improvement in target detection performance, coupled with lower GFLOPs and a reduced parameter count.

TABLE 3 Comparison of the performance in L-ELAN module.

Methods	AP50 (%)↑	FLOPs (G)↓	Para (M)↓
YOLOv7	81	103.4	34.56
Conv	81	103.4	34.56
DCNv2	80.8 (↓ 0.2)	89.1 (↓ 14.3)	36.07 (↑ 1.40)
DCNv3	81.1 (↑ 0.1)	81.1 (↓ 17.9)	35.89 (↓ 1.97)
PConv	81.3 (↑ 0.3)	83.1 (↓ 20.3)	30.56 (↓ 4.30)

↑ Means the larger the better, ↓ means the smaller the better.

4.2.5 Wise-IOU

As evidenced by the data in Table 2, the incorporation of the novel loss function Wise-IOU results in unchanged model parameters and computational operations. However, owing to the influence of the dynamic non-monotonic focusing mechanism, the model described in this paper exhibits a 0.6% enhancement in AP50. This improvement underscores the efficacy of the introduced loss function and its capacity to effectively refine model performance.

4.3 Comparisons with the state-of-the-art

To comprehensively validate the overall detection prowess of the model presented in this paper, a thorough comparison is conducted against the benchmark model YOLOv7, along with

TABLE 4 The comparison of the performance in WSODD.

Methods	AP50 (%)↑	AP95 (%)↑	APs (%)↑	APm (%)↑	APl (%)↑	Para (M)↓	FLOPS (G)↓
DETR	79.9	4.60	14.7	34.4	61.8	39.37	86
D-DETR	80.1	38.9	21.1	35.7	56.1	40.01	173
YOLOv5s	79.6	46.3	17.6	37.2	58.3	6.72	15.9
YOLOv5m	80.9	47.3	17.4	38.2	62.0	19.94	48
DAMO-YOLOt	81.3	48.2	20.8	40.3	60.3	8.5	18.2
DAMO-YOLOs	81.6	48.7	19.0	**40.8**	57.5	16.3	37.8
DAMO-YOLOm	81.8	**50.1**	18.7	40.3	61.8	28.2	61.8
YOLOv7t	77.7	46.0	15.5	37.1	60.3	5.78	13.1
YOLOv7m	81.0	49.3	20.3	39.7	62.7	34.86	103.4
YOLOv7x	83.0	49.2	21.6	39.3	62.6	67.62	188.3
STE-YOLO	**83.1**	49.4	**21.9**	39.8	**63.2**	32.8	89.6
YOLOv8n	80.3	47.7	15.8	37.8	62.7	**3.01**	**8.1**
YOLOv8s	81.8	49.2	18.7	38.7	62.5	11.13	28.5
YOLOv8m	82.2	49.9	18.3	39.9	63.0	25.85	78.7

Bold values means the best performance under the same evaluation criteria.
↑ Means the larger the better, ↓ means the smaller the better.

TABLE 5 The comparison of the performance in FIOW-Img.

Methods	AP50 (%)↑	AP95 (%)↑	APs (%)↑	APm (%)↑	FPS↑	Para (M)↓	FLOPS (G)↓
YOLOv5s	81.1	**38.3**	23.9	52.6	141.5	6.71	15.8
YOLOv5m	81.7	31.1	24.7	53.2	87.1	19.88	47.9
YOLOv5l	82.8	36.4	25.9	55.2	51.3	46.12	107.6
YOLOv7t	80.9	35.7	23.8	53.0	**143.7**	5.76	13.0
YOLOv7	81.7	36.8	24.2	54.1	47.4	34.81	103.2
YOLOv7x	83.0	38.2	25.6	52.4	30.1	67.53	188.0
STE-YOLO	**83.2**	37.3	**26.1**	55.9	56.8	32.74	89.2
YOLOv8n	80.7	30.0	23.5	52.2	125.3	**3.0**	**8.1**
YOLOv8s	81.3	33.4	24.2	55.8	118.7	11.1	28.4
YOLOv8m	81.8	33.5	25.1	**56.3**	92.6	25.77	78.7

Bold values means the best performance under the same evaluation criteria.
↑ Means the larger the better, ↓ means the smaller the better.

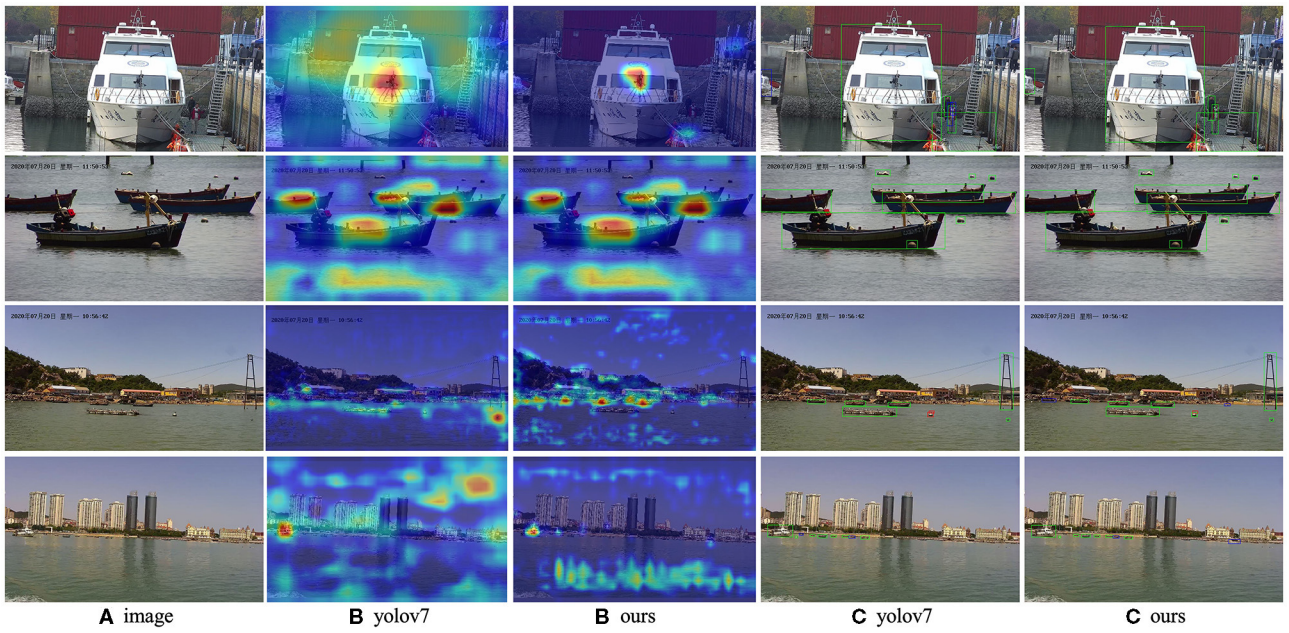

FIGURE 8
Some visualization results from our STE-YOLO on WSODD. Where **(A)** represents the original picture, **(B)** represents the thermal map, and **(C)** represents the detection map. We are using heat maps and calibration boxes to show the the effect. Use different colors for different test results (green means correct detection, blue means wrong detection, red means missed detection). In our model, the heat map is more centralized, and there are fewer misdetections and omissions compared to the original modele.

various cutting-edge target detectors (Carion et al., 2020; Zhu et al., 2020; Jocher et al., 2022; Xu et al., 2022). The ensuing comparative experiments yield results showcased in Tables 4, 5. Notably, the findings in Table indicate that the model proposed in this paper attains the highest score on AP50, along with leading APs and APl scores. This resounding achievement stands as compelling evidence affirming the efficacy of the strategies put forth in this paper.

When juxtaposed with the benchmark model YOLOv7 on WSODD, our STE-YOLO demonstrates a notable 2.1% enhancement in AP50, as well as a 1.6% improvement in APs, reaching 21.9%. Additionally, the parameter count in this paper model is 14% lower than that of YOLOv7, with a 6% reduction in GFLOPs. Noteworthy is the comparison with DAMO-YOLO and DETR, models with similar GFLOPs. STE-YOLO outperforms both in terms of detection performance. Furthermore, in comparison to the larger YOLOV7x model, the model detailed in this paper reduces parameter count by 53% and GFLOPs by 51%, while simultaneously surpassing AP50 and Aps by 0.1% and 0.3%, respectively.

When the model was tested on the FIOW-Img dataset, our STE-yolo achieved the best results on AP50 and APs, with 1.5% and 1.9% improvements over the original model, respectively, and APs in particular was significantly ahead of the larger version model. In addition, the FPS of this model has also improved, having the highest FPS among models of similar size, reaching 56.8. This also proves the fast detection and optimization effect of the proposed strategy for small targets.

In summary, the STE-YOLO model proposed in this paper demonstrates substantial enhancements across diverse metrics when compared to the standard YOLOv7 model, and it also showcases clear advantages over alternative algorithms in some extent.

4.4 Visualization of results

To provide a visual representation of the experimental outcomes, a selection of images was curated for display purposes. As depicted in Figure 8, a compilation of heat maps and detection results is presented, offering insights into the model's performance across images featuring both large and small objects. This visualization serves to further illustrate the efficacy and versatility of the proposed approach in capturing objects of varying scales within the detection framework. As depicted in Figure 9, We also compared the detection results of different models on small targets and marked the detection situation. After testing the whole verification set, we made statistics on the detect results, as shown in Table 6. Compared to YOLOV7, STE-YOLO's error detection rate has been greatly reduced this shows that STE-YOLO has better small target detection performance.

5 Discussion

Addressing the intricate task of accurately identifying and discerning small targets within the realm of waterborne unmanned crafts' target detection, this study introduces the STE-YOLO algorithm. This algorithm represents a swift and precise technique for waterborne target detection, bolstered by enhanced small-target detection capabilities. Its primary objective is to elevate

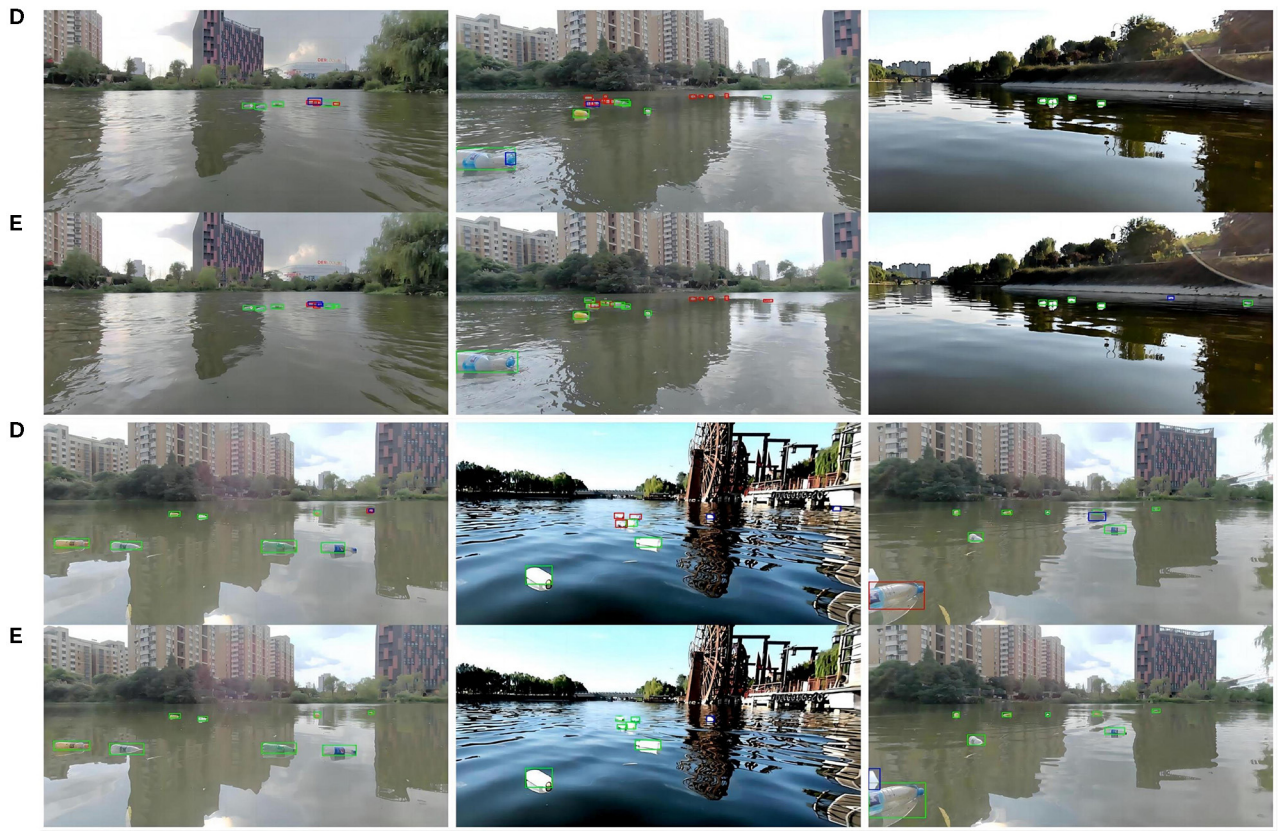

FIGURE 9
Some visualization results from our STE-YOLO on FIOW-Img. Where **(D)** represents the detection effect of the original model and **(E)** means the detection effect of STE-YOLO, it is obvious that our model has better detection effect of small targets.

the precision of detecting small targets within this context. Achieving this aim, the algorithm integrates a pioneering P2 detection head branch that adeptly tackles the challenge of detecting targets spanning a wide range of scales. To further enhance its ability to recognize small targets, a lightweight vision center (LVC) module is seamlessly integrated to effectively synchronize cross-layer information. Simultaneously, to optimize the algorithm's computational efficiency for efficient real-world deployment, an aggregation network named L-ELAN is seamlessly woven into the backbone network architecture, thereby enhancing computational efficiency. In a strategic move to fortify the algorithm's robustness across varying scales, the Wise-IOU loss function is introduced. This dynamic loss function optimizes gradient influence arising from samples of varied quality, further enhancing model performance. Significantly, empirical evidence garnered from experiments conducted on the WSODD dataset affirms the supremacy of STE-YOLO in terms of detection accuracy, particularly its proficiency in detecting small targets, and its efficient utilization of model parameters. This advancement carries theoretical importance for the pragmatic deployment of water surface target detection applications.

As future research avenues unfold, a key focus will be refining the network structure. Promising methodologies such as model pruning or knowledge distillation will be employed to comprehensively reduce the number of network parameters. This strategic effort aims to enhance the model's applicability and performance during deployment, especially in scenarios marked by limited computational resources.

TABLE 6 **Statistics of detection results on FIOW-Img.**

Methods	Right	Miss	Error
YOLOv7	1,501	368	455
STE-YOLO	1,616	288	351

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

JY: Writing—original draft. HZ: Writing—original draft. LX: Conceptualization, Writing—original draft. LZ: Writing—review

& editing. MY: Writing—review & editing. JH: Writing—review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported in part by National Natural Science Foundation of China under Grant Number 62271286, in part by Natural Science Foundation of Hubei Province under Grant Number 2022CFB752, in part by Yichang Natural Science Research Project under Grant Number A22-3-004, and in part by State Grid Yichang Company Management Technology Project under Grant Number B715H023XT08.

Conflict of interest

JY, LX, and JH were employed by State Grid Yichang Electric Power Supply Company.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Bochkovskiy, A., Wang, C.-Y., and Liao, H.-Y. M. (2020). YOLOv4: optimal speed and accuracy of object detection. *arXiv* [preprint]. doi: 10.48550/arXiv.2004.10934

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., et al. (2020). "End-to-end object detection with transformers," in *Computer Vision-ECCV 2020*, eds A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm (Cham: Springer International Publishing), 213–229. doi: 10.1007/978-3-030-58452-8_13

Chen, J. Kao, S.-h., He, H., Zhuo, W., Wen, S., Lee, C.-H., Chan, S.-H. G. (2023). "Run, don't walk: chasing higher flops for faster neural networks," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Vancouver, BC), 12021–12031. doi: 10.1109/CVPR52729.2023.01157 Available online at: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/10806/1080616/A-maritime-targets-detection-method-based-on-hierarchical-and-multi/10.1117/12.2503030.short?SSO=1

Chen, W., Li, J., Xing, J., Yang, Q., and Zhou, Q. (2018). "A maritime targets detection method based on hierarchical and multi-scale deep convolutional neural network," in *Tenth International Conference on Digital Image Processing (ICDIP 2018)*, Volume 10806, eds X. Jiang, and J.-N. Hwang (International Society for Optics and Photonics, SPIE), 1080616. doi: 10.1117/12.2503030

Chen, X., Yuan, M., Yang, Q., Yao, H., and Wang, H. (2023). Underwater-ycc: underwater target detection optimization algorithm based on YOLOv7. *J. Mar. Sci. Eng.* 11, 995. doi: 10.3390/jmse11050995

Cheng, Y., Zhu, J., Jiang, M., Fu, J., Pang, C., Wang, P., et al. (2021). "Flow: a dataset and benchmark for floating waste detection in inland waters," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (Montreal, QC: IEE), 10953–10962. doi: 10.1109/ICCV48922.2021.01077

Deng, C., Wang, M., Liu, L., Liu, Y., and Jiang, Y. (2022). Extended feature pyramid network for small object detection. *IEEE Trans. Multimedia* 24, 1968–1979. doi: 10.1109/TMM.2021.3074273

Everingham, M., Van Gool, L., Williams, C. K. I., Win, J., Zisserman, A., et al. (2010). The pascal visual object classes (VOC) challenge. *Int. J. Comput. Vis.* 88, 303–338. doi: 10.1007/s11263-009-0275-4

Ge, Z., Liu, S., Wang, F., Li, Z., and Sun, J. (2021). YOLOX: exceeding yolo series in 2021. *arXiv* [preprint]. doi: 10.48550/arXiv.2107.08430

Ghiasi, G., Lin, T.-Y., and Le, Q. V. (2019). "NAS-FPN: learning scalable feature pyramid architecture for object detection," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Long Beach, CA: IEEE), 7029–7038. doi: 10.1109/CVPR.2019.00720

Gong, L., Huang, X., Chao, Y., Chen, J., and Li, B. (2023). An enhanced ssd with feature cross-reinforcement for small-object detection. *Appl. Intell.* 53, 19449–19465. doi: 10.1007/s10489-023-04544-1

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). "Mask R-CNN," in *2017 IEEE International Conference on Computer Vision (ICCV)* (Venice: IEEE), 2980–2988. doi: 10.1109/ICCV.2017.322

Jocher, G., Stoken, A., Borovec, J., Chaurasia, A., TaoXie, Changyu, L., et al. (2022). *ultralytics/yolov5: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation*. Zenodo. Available online at: https://ui.adsabs.harvard.edu/abs/2022zndo...7347926J

Kisantal, M., Wojna, Z., Murawski, J., Naruniec, J., and Cho, K. (2019). Augmentation for small object detection. *arXiv* [preprint]. doi: 10.48550/arXiv.1902.07296

Lin, T.-Y., Dollar, P., Girshick, R., He, K., Hariharan, B., Belongie, S., et al. (2017). "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Honolulu, HI: IEEE). doi: 10.1109/CVPR.2017.106

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., et al. (2014). "Microsoft COCO: Common Objects in Context," in *Computer Vision - ECCV 2014. ECCV 2014. Lecture Notes in Computer Science*, Vol. 8693, eds D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars (Cham: Springer).

Liu, R., Lehman, J., Molino, P., Petroski Such, F., Frank, E., Sergeev, A., et al. (2018). "An intriguing failing of convolutional neural networks and the coordconv solution," in *Advances in Neural Information Processing Systems*, Vol. 31, eds S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Red Hook, NY: Curran Associates, Inc) 9605–9616.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (2021). "Swin transformer: hierarchical vision transformer using shifted windows," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (Montreal, QC: IEEE), 9992–10002. doi: 10.1109/ICCV48922.2021.00986

Moosbauer, S., König, D., Jäkel, J., and Teutsch, M. (2019). "A benchmark for deep learning based object detection in maritime environments," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (Long Beach, CA), 916–925. doi: 10.1109/CVPRW.2019.00121

Nguyen, D.-K., Ju, J., Booij, O., Oswald, M. R., and Snoek, C. G. M. (2022). "BoxeR: box-attention for 2D and 3D transformers," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (New Orleans, LA: IEEE), 4763–4772. doi: 10.1109/CVPR52688.2022.00473

Quan, Y., Zhang, D., Zhang, L., and Tang, J. (2023). Centralized feature pyramid for object detection. *IEEE Trans. Image Process.* 32, 4341–4354. doi: 10.1109/TIP.2023.3297408

Ren, S., He, K., Girshick, R., and Sun, J. (2017). Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 1137–1149. doi: 10.1109/TPAMI.2016.2577031

Shamsolmoali, P., Zareapoor, M., Yang, J., Granger, E., and Chanussot, J. (2022). "Enhanced single-shot detector for small object detection in remote sensing images," in *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium* (Kuala Lumpur: IEEE), 1716–1719. doi: 10.1109/IGARSS46834.2022.9884546

Shin, H.-C., Lee, K.-I., and Lee, C.-E. (2020). "Data augmentation method of object detection for deep learning in maritime image," in *2020 IEEE International Conference on Big Data and Smart Computing (BigComp)* (Busan: IEEE), 463–466. doi: 10.1109/BigComp48618.2020.00-25

Sun, H., Li, B., Dan, Z., Hu, W., Du, B., Yang, W., et al. (2023). Multi-level feature interaction and efficient non-local information enhanced channel attention for image dehazing. *Neural Netw.* 163, 10–27. doi: 10.1016/j.neunet.2023.03.017

Sun, H., Zhang, Y., Chen, P., Dan, Z., Sun, S., Wan, J., et al. (2021). Scale-free heterogeneous cyclegan for defogging from a single image for autonomous driving in fog. *Neural Comput. Appl.* 35, 3737–3751. doi: 10.1007/s00521-021-06296-w

Tan, M., Pang, R., and Le, Q. V. (2020). "Efficientdet: scalable and efficient object detection," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Seattle, WA: IEEE), 10778–10787. doi: 10.1109/CVPR42600.2020.01079

Tian, Z., Shen, C., Chen, H., and He, T. (2022). FCOS: a simple and strong anchor-free object detector. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 1922–1933. doi: 10.1109/TPAMI.2020.3032166

Tolstikhin, I. O., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., et al. (2021). "Mlp-mixer: an all-mlp architecture for vision," in *Advances in Neural Information Processing Systems*, Vol. 34, eds M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan (Red Hook, NY: Curran Associates, Inc), 24261–24272.

Tong, Z., Chen, Y., Xu, Z., and Yu, R. (2023). Wise-IoU: bounding box regression loss with dynamic focusing mechanism. *arXiv* [preprint]. doi: 10.48550/arXiv.2301.10051

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," inAdvances *in Neural Information Processing Systems*, Vol. 30, eds I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Red Hook, NY: Curran Associates, Inc).

Wan, J., Liu, J., Zhou, J., Lai, Z., Shen, L., Sun, H., et al. (2023). Precise facial landmark detection by reference heatmap transformer. *IEEE Trans. Image Process.* 32, 1966–1977. doi: 10.1109/TIP.2023.3261749

Wang, C.-Y., Bochkovskiy, A., and Liao, H.-Y. M. (2021). "Scaled-YOLOv4: scaling cross stage partial network," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Nashville, TN: IEEE), 13024–13033. doi: 10.1109/CVPR46437.2021.01283

Wang, C.-Y., Bochkovskiy, A., and Liao, H.-Y. M. (2023). "YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Vancouver, BC: IEEE), 7464–7475. doi: 10.1109/CVPR52729.2023.00721

Wang, K., Liew, J. H., Zou, Y., Zhou, D., and Feng, J. (2019). "Panet: few-shot image semantic segmentation with prototype alignment," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (Seoul: IEEE), 9196–9205. doi: 10.1109/ICCV.2019.00929

Weber, M., Wang, H., Qiao, S., Xie, J., Collins, M. D., Zhu, Y., et al. (2021). DeepLab2: a tensorflow library for deep labeling. *arXiv* [preprint]. doi: 10.48550/arXiv.2106.09748

Wightman, R., Touvron, H., and Jégou, H. (2021). ResNet strikes back: an improved training procedure in timm. *arXiv* [preprint]. doi: 10.48550/arXiv.2110.00476

Xu, X., Jiang, Y., Chen, W., Huang, Y., Zhang, Y., Sun, X., et al. (2022). DAMO-YOLO: a report on real-time object detection design. *arXiv* [preprint]. doi: 10.48550/arXiv.2211.15444

Yang, C., Huang, Z., and Wang, N. (2022). "QueryDet: cascaded sparse query for accelerating high-resolution small object detection," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (New Orleans, LA: IEEE), 13658–13667. doi: 10.1109/CVPR52688.2022.01330

Yang, Z., Liu, S., Hu, H., Wang, L., and Lin, S. (2019). RepPoints: point set representation for object detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (Seoul: IEEE), 9656–9665. doi: 10.1109/ICCV.2019.00975

Yu, X., Gong, Y., Jiang, N., Ye, Q., and Han, Z. (2020). "Scale match for tiny person detection," in *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1246–1254. doi: 10.1109/WACV45572.2020.9093394

Zalesskaya, G., Bylicka, B., and Liu, E. (2022). How to train an accurate and efficient object detection model on any dataset. *arXiv* [preprint]. doi: 10.48550/arXiv.2211.17170

Zhang, H., Wang, Y., Dayoub, F., and Sanderhauf, N. (2021). "Varifocalnet: an iou-aware dense object detector" in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Snowmass, CO: IEEE), 8510–8519. doi: 10.1109/CVPR46437.2021.00841

Zhang, L., Liu, T., and Ding, X. (2022). Large-scale WSNS resource scheduling algorithm in smart transportation monitoring based on differential ion coevolution and multi-objective decomposition. *IEEE Trans. Intell. Transp. Syst.* 1–10. doi: 10.1109/TITS.2022.3208699 Available online at: https://ieeexplore.ieee.org/abstract/document/9904963

Zhang, R., Zhang, L., Su, Y., Yu, Q., and Bai, G. (2023). Automatic vessel plate number recognition for surface unmanned vehicles with marine applications. *Front. Neurorobot.* 17, 1131392. doi: 10.3389/fnbot.2023.1131392

Zhao, Q., Sheng, T., Wang, Y., Tang, Z., Chen, Y., Cai, L., et al. (2019). M2det: a single-shot object detector based on multi-level feature pyramid network. *Proc. AAAI Conf. Artif. Intell.* 33, 9259–9266. doi: 10.1609/aaai.v33i01.33019259

Zhou, X., Wang, D., and Krähenbühl, P. (2019). Objects as points. *arXiv* [preprint]. doi: 10.48550/arXiv.1904.07850

Zhou, Z., Hu, X., Li, Z., Jing, Z., and Qu, C. (2022). A fusion algorithm of object detection and tracking for unmanned surface vehicles. *Front. Neurorobot.* 16, 808147. doi: 10.3389/fnbot.2022.808147

Zhou, Z., Sun, J., Yu, J., Liu, K., Duan, J., Chen, L., et al. (2021). An image-based benchmark dataset and a novel object detector for water surface object detection. *Front. Neurorobot.* 15:723336. doi: 10.3389/fnbot.2021.723336

Zhu, X., Hu, H., Lin, S., and Dai, J. (2019). "Deformable convnets v2: more deformable, better results," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Long Beach, CA: IEEE), 9300–9308. doi: 10.1109/CVPR.2019.00953

Zhu, X., Lyu, S., Wang, X., and Zhao, Q. (2021). "TPH-YOLOv5: improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios," in *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)* (Montreal, BC: IEEE), 2778–2788. doi: 10.1109/ICCVW54120.2021.00312

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J., et al. (2020). Deformable DETR: deformable transformers for end-to-end object detection. *arXiv* [preprint]. doi: 10.48550/arXiv.2010.04159