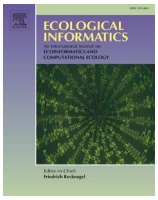
Ecological Informatics 83 (2024) 102802

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



Camouflage detection: Optimization-based computer vision for *Alligator sinensis* with low detectability in complex wild environments

Yantong Liu [a], Sai Che [c], Liwei Ai [d], Chuanxiang Song [a], Zheyu Zhang [a], Yongkang Zhou [e], Xiao Yang [f], Chen Xian [b,*]

[a] Department of Computer Information Engineering, Kunsan National University, Gunsan 54150, Republic of Korea
[b] Department of Mechanical Engineering, Kunsan National University, Gunsan 54150, Republic of Korea
[c] Department of Mechanical Engineering, Jiangsu University of Science and Technology, Zhenjiang 212000, China
[d] State Key Laboratory of Hydraulics and Mountain River Engineering, College of Water Resource and Hydropower, Sichuan University, Chengdu 610065, China
[e] Alligator sinensis Research Center of Anhui Province, Xuanzhou 242000, China
[f] Department of Poultry Science, University of Georgia, Athens, GA 30605, USA

ARTICLE INFO

Keywords:
Alligator sinensis
YOLO v8 optimization
Wildlife conservation technology
Camouflage pattern recognition
Camouflaged object detection
Ecological monitoring algorithms

ABSTRACT

Alligator sinensis is an extremely rare species that possesses excellent camouflage, allowing it to fit perfectly into its natural environment. The use of camouflage makes detection difficult for both humans and automated systems, highlighting the importance of modern technologies for animal monitoring. To address this issue, we present YOLO v8-SIM, an innovative detection technique specifically developed to significantly enhance the identification precision. YOLO v8-SIM utilizes a sophisticated dual-layer attention mechanism, an optimized loss function called inner intersection-over-union (IoU), and a technique called slim-neck cross-layer hopping. The results of our study demonstrate that the model achieves an accuracy rate of 91 %, a recall rate of 89.9 %, and a mean average precision (mAP) of 92.3 % and an IoU threshold of 0.5. In addition, the model operates at a frame rate of 72.21 frames per second (FPS) and excels at accurately recognizing objects that are partially visible or smaller in size. To further improve our initiatives, we suggest creating an open-source collection of data that showcases *A. sinensis* in its native environment while using camouflage techniques. These developments collectively enhance the ability to detect disguised animals, thereby promoting the monitoring and protection of biodiversity, and supporting ecosystem sustainability.

1. Introduction

The *A. sinensis*, also known as Chinese *A. sinensis*, is currently in a critically endangered state (Pan et al., 2019; Platt et al., 2022; Wan et al., 2013). This is primarily attributed to the loss of natural habitats and encroachment of human activities (He et al., 2017). Presently, it is confronted with a severe peril to its existence, as the number of individuals in the wild is dwindling to less than 1400, with fewer than 1000 mature adults as of 2023. The preservation of this ancient organism is essential for both the equilibrium of the ecosystem and the preservation of biodiversity (Marshall, 2017). Conventional animal identification methods include remote sensing, hand tracking, and tags (Hahn et al., 2022; Technological advances in biodiversity monitoring: applicability, opportunities and challenges - ScienceDirect, 2024). Nevertheless, remote sensing frequently suffers from limited precision in wildlife surveys because of intricate backdrops and disruptions (Fang et al., 2016; Han et al., 2021). Although remote sensing encounters difficulties in terms of the overall accuracy, satellite monitoring presents a distinct set of hurdles. In addition, satellite monitoring is limited to larger animals because of the need for high-resolution images (CSIRO PUBLISHING | Marine and Freshwater Research, 2024; Whales from space: Four mysticete species described using new VHR satellite imagery - Cubaynes - 2019 - Marine Mammal Science - Wiley Online Library, 2024). *A. sinensis* effectively uses camouflage in muddy waters or dense vegetation in its natural habitats, which poses challenges for monitoring operations because of its unpredictable behavior. Manual tracking, although feasible, is arduous and time-consuming, and sometimes leads to incomplete or erroneous data. Recent research on the detection of

* Corresponding author.
E-mail addresses: lyt1994@kunsan.ac.kr (Y. Liu), Che_sai@stu.just.edu.cn (S. Che), ailiwei@stu.scu.edu.cn (L. Ai), zheyu@kunsan.ac.kr (Z. Zhang), xy50573@uga.edu (X. Yang), xc2021@kunsan.ac.kr (C. Xian).

https://doi.org/10.1016/j.ecoinf.2024.102802
Received 10 January 2024; Received in revised form 26 August 2024; Accepted 26 August 2024
Available online 28 August 2024

camouflaged objects has shown promise in enhancing monitoring capabilities and offering new ways to address these difficulties. These developments utilize sophisticated features that are sensitive to textures and use strategies that involve learning from interconnected graphs to improve object detection in challenging contexts. This has the potential to significantly boost animal monitoring efforts (Fan et al., 2020; Ren et al., 2021; Zhai et al., 2021). Tag tracking in *A. sinensis* may be subject to various challenges, including illness, physical harm, chip displacement, technical malfunctions, and ethical implications. Although traditional monitoring approaches provide valuable insights, they present difficulties such as labor-intensive processes and intrusive tactics. Consequently, there is a growing trend toward the adoption of sustainable and ecologically sensitive strategies (Hahn et al., 2022; Technological advances in biodiversity monitoring: applicability, opportunities and challenges - ScienceDirect, 2024). Given the pressing requirements for successful preservation efforts, the field of wildlife monitoring technology is rapidly progressing toward the development of non-invasive, all-encompassing, and less disruptive approaches to save endangered species (Adams, 2019; Farrell et al., 2022; Linden et al., 2017).

The deployment of drones and automated camera traps has markedly transformed wildlife monitoring, enabling the acquisition of copious visual data (Gonzalez et al., 2016; Nazir and Kaleem, 2021; Pimm et al., 2015). Rapid advancements in remote sensing and artificial intelligence, particularly in the domain of deep-learning (DL) methodologies, have advanced areas such as image classification, computer vision, and object detection. Deep learning, distinguished by its use of multilayer neural networks, has demonstrated a remarkable capacity for pattern recognition. Deep learning models autonomously assess the precision of their forecasts through the operation of their neural networks, without the need for human input. Islam and Valles (Islam and Valles, 2020) proposed the use of convolutional neural networks (CNN) for automatic detection and categorization of small creatures, including toads, frogs, snakes, and lizards, in camera trap photographs. Roy et al. (A Computer Vision-Based Object Localization Model for Endangered Wildlife Detection by Arunabha Mohan Roy, Jayabrata Bhaduri, Teerath Kumar, Kislay Raj: SSRN, 2024) introduced WilDect-YOLO, a deep learning model designed to detect endangered species in real-time. The model achieved a mean average precision of 96.89 %, F1-score of 97.87 %, and precision of 97.18 % at a frame rate of 59.20 frames per second (FPS). Carl et al.. conducted an experiment using the FasterRCNN+InceptionResNetV2 model to detect and classify wild European mammals in camera-trap photos. The researchers achieved a detection accuracy of 94 % and a classification accuracy of up to 93 % when using an intersection-over-union (IoU) threshold of 0.5. Kellenberger et al. (Kellenberger et al., 2017) developed a CNN approach to enhance the accuracy of detecting large animals in unmanned aerial vehicle (UAV) photos. The system demonstrates real-time processing at a frequency of 72 Hz. Advancements in this field have significantly enhanced the capacity to detect animals, offering a vast repository of data that enables automated identification of diverse species. In their investigation, Y. X et al. (Yang et al., 2022) proposed a deep learning model, designated YOLOv5x-hens, based on the YOLOv5 architectural framework. The objective of this model was to monitor the behavior of hens on free-range farms. The accuracy of real-time detection exceeded 95 %.

The complexity and similarity of field settings are vitally important in the process of environmentally discerning and safeguarding *A. sinensis*. Moreover, industry currently lacks precise and up-to-date techniques for identifying indistinguishable objects in natural surroundings. To address the limitations of the current method, we have enhanced the YOLO v8 algorithm with the specific aim of detecting items that are difficult to differentiate. Consequently, we devised the YOLO v8-SIM model, which is capable of real-time detection of camouflaged objects in their natural surroundings.

The enhanced structure of the algorithm can be articulated as follows:

1. The YOLO v8-SIM model leverages the ResNet−18 architecture as its foundational backbone network, integrating it with a dual-layer attention mechanism comprising former and reverse attention (RA). This combination facilitated the precise detection of targets within camouflaged environments by effectively managing the focus of the network.
2. The inner IoU loss function was used to augment the model's capacity for generalizing target detection across various scales. The inner IoU introduces a scale factor that adjusts the size of the auxiliary bounding box used in the loss computation, thus fine-tuning the bounding box regression process for both high- and low-IoU samples.
3. Utilization of the slim-neck cross-layer hopping mechanism enables pruning of the network model. This approach strategically reduces the number of parameters in the model, thereby achieving model compression without compromising the accuracy of target detection.

This study presents a dataset designed to improve the identification of *A. sinensis* in field environments characterized by low recognition rates.

2. Detection algorithm

2.1. The YOLO algorithm development

The YOLO (You Only Look Once) series, operating in a single stage, is a real-time object identification model built upon a CNN. The success of YOLO stems from its ability to effectively combine features and produce extremely accurate detection results, while maintaining a lightweight network design. YOLO v8, the most recent iteration of the YOLO detection model, incorporates novel characteristics and enhancements compared with prior versions of YOLO. These updates aim to increase the ability of the model to detect objects and provide greater flexibility in the performance. The YOLO v8 model is anchor-free, resulting in a reduction in the number of box predictions, acceleration of non-maximum suppression, and enhancement of detection efficiency. The YOLO v8 network consists of three primary modules: the Backbone, Neck, and Head. These modules were responsible for extracting features, fusing multiple features, and producing prediction outputs. The network configuration is shown in Fig. 1..

The YOLO v8 model incorporates the ELAN (Efficient Layer Aggregation Network) concept (Wang et al., 2022) into its basic structure, replacing the C3 architecture of YOLOv5 with a novel C2f design. The C2f framework is a notable improvement that combines two parallel gradient flow branches, while eliminating one convolutional layer from the original C3 configuration. This modification enables enhanced and stronger extraction of gradient flow information while preserving the efficiency and lightweight nature of the model. In addition, the inclusion of the Spatial Pyramid Pooling Fusion (SPPF) module, which is inherited from YOLOv5 (Zhu et al., 2021), effectively addresses visual distortion issues and improves the capacity of the model to adapt to different settings.

The neck module in YOLO v8 utilizes the PAN (Path Aggregation Network) paradigm for feature fusion, which is influenced by the FPN (Feature Pyramid Network) concept. The technique used is a bidirectional fusion method that combines high- and low-level characteristics. This method significantly improves the ability to identify objects of various sizes by enhancing the low-level characteristics using smaller receptive fields. YOLO v8 has made more strategic modifications than YOLOv5. The convolutional layers were removed in the upsampling step of the FPN, and the standard C3 module was replaced with the C2f module to enhance the performance optimization.

The head module of YOLO v8 utilizes a Decoupled Head mechanism that is specifically designed to minimize interdependencies among the many tasks involved in target detection (Arora et al., 2004). The model effectively separates the categorization and detection components,

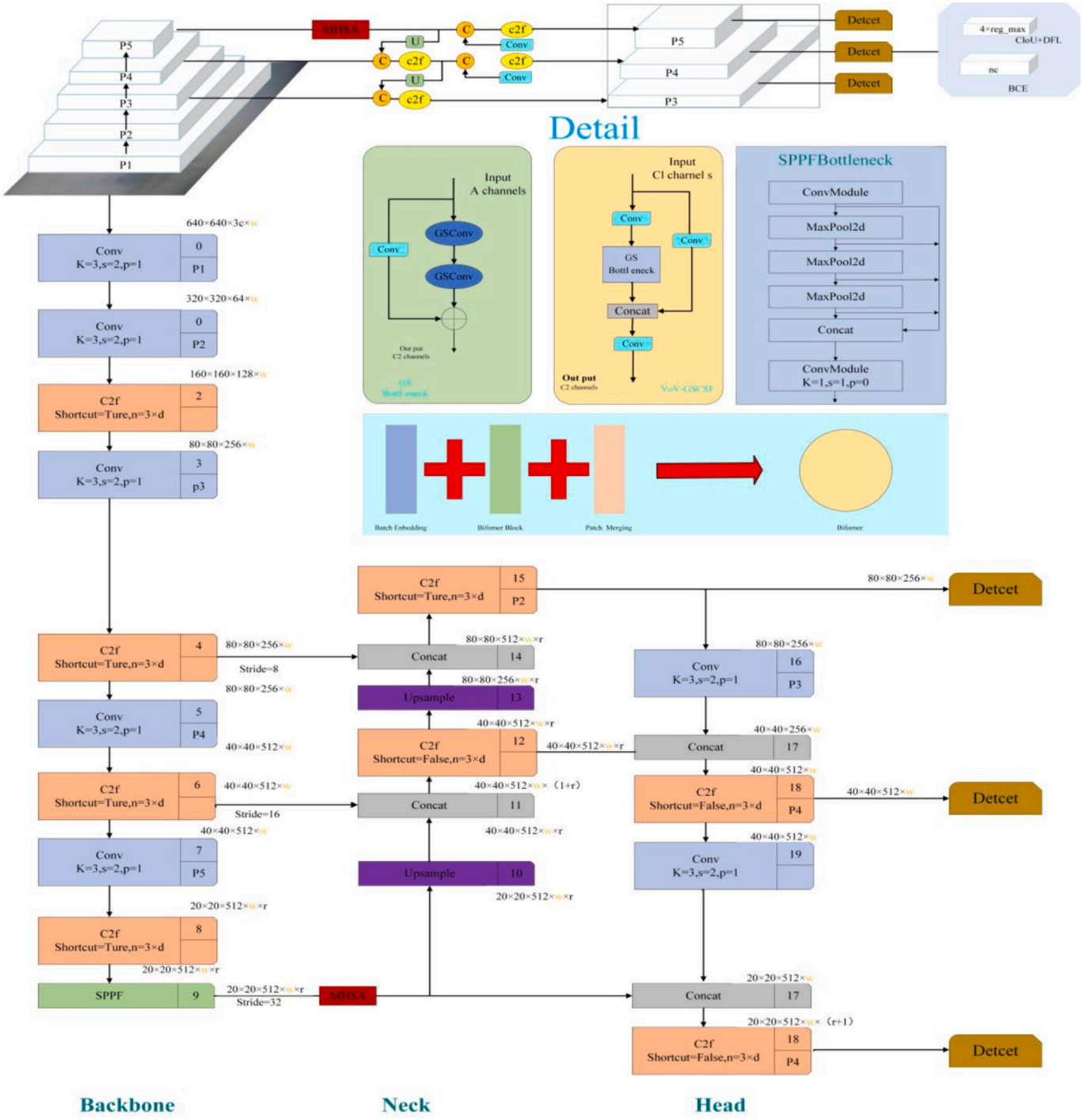


Fig. 1. YOLO v8 network structure.

allowing for efficient extraction and processing of positional and category information. This segregation enables more efficient integration of data from various network branches, thereby decreasing the extra latency commonly linked to convolutions in the decoupling head. The outcome is an augmented capacity for generalization and greater resilience of the model.

2.2. Related work

In the field of object detection, particularly with regard to wildlife conservation, our research addresses the distinctive challenges of detecting cryptic species such as *A. sinensis* in their natural habitats.

In Section 3.2.1, we introduce a dual-layer attention mechanism comprising bias and reverse attention (RA) in the YOLO v8-SIM model architecture. This innovation markedly enhances the model's ability to focus on the target species by amplifying relevant features and suppressing complex backgrounds, thereby improving detection accuracy. Furthermore, in Section 3.2.2, we propose the integration of an inner IoU-optimized loss function to address the varying scales of *A. sinensis* in disparate environmental contexts. This approach addresses the shortcomings of conventional IoU loss functions, which often exhibit slow convergence and poor generalization, particularly when dealing with partially visible or juvenile *A. sinensis*. In Section 3.2.3, we introduce a slim-neck cross-layer hopping mechanism for network pruning and

compression. The objective was to decrease model complexity while preserving detection accuracy. This technique is of great importance for optimizing the computational efficiency, particularly in environments with limited resources or real-time applications. Moreover, in Section 3.2.4, we address the challenge of preventing attention mechanisms from overfitting highly textured background features, which could potentially result in neglecting the target species. By meticulously calibrating the former attention mechanism and inner IoU scale, we enhanced the responsiveness of the model to both the target and background, thereby ensuring its resilience in diverse environmental settings. Finally, in Section 3.2.5, the slim-neck mechanism was implemented with the objective of reducing information loss during network pruning, thereby enhancing the model's ability to detect *A. sinensis* across different scales and environmental contexts. Collectively, these innovations signify significant advancements in the field of object detection, particularly in wildlife conservation, by enhancing the detection accuracy of cryptic species in their natural habitats.

3. Materials and methods

3.1. Dataset

A total of two datasets are used in this study, The first is the open source dataset COD10K for camouflaged object detection, which was proposed by Dengping Fan et al. in 2020 (Fan et al., 2021). which is used for the model to undergo the migration learning of camouflage learning first, and then the migration learning preparation.

The second dataset was collected from the *A. sinensis* National Nature Reserve in Xuancheng City, Anhui Province and contained 1027 images. These images were captured from different angles using an iPhone 13 Pro mobile phone camera (California, USA) and a Xiaomi 12 5G mobile phone camera (Beijing, China). The shooting heights were varied between 1 and 1. 5 m. The collected images were saved in JPG format. The dataset was expanded to 2464 images through processing techniques, including rotation, horizontal flipping, and random cropping, and was subsequently made open-source.

Because of the extreme rarity of *A. sinensis*, locating them in the wild is nearly impossible, and even within the dedicated *A. sinensis* National Park, their numbers are scarce. Consequently, the dataset size was constrained, and there are no available open-source datasets specifically tailored for camouflaged *A. sinensis*. Initially, we selected 5066 images (primarily occlusion and environment-mimicking images) from the widely used open-source camouflage recognition dataset COD10K to train all models, following an 8:1:1 ratio for the training, validation, and test sets. This compensates for the absence of an *A. sinensis* dataset, enhancing the camouflage recognition capabilities across all models. Notably, all the models were standardized after COD10K training, ensuring equitable and unbiased outcomes upon subsequent training with our proprietary dataset.

Performing a statistical analysis of the dataset, considering its composition and characteristics, is crucial to understanding the data distribution and potential biases. In a given dataset, each image presents unique challenges and variations that must be considered in the analysis. First, we consider the distribution of image types within the dataset.

Normal Images: These are images in which the target object, in this case, *A. sinensis*, is clearly visible without any occlusion or camouflage. Typically, normal images serve as a baseline for comparison with more challenging images.

Overlapping Objects: These are images in which multiple objects, including *aA. sinensis* and other elements, overlap or intersect, making it difficult for the model to accurately detect or identify the target.

Ambient Color Artifacts: These images in which environmental factors such as lighting conditions, shadows, or reflections introduce color artifacts that may obscure or distort the appearance of the target object.

Environmental Proximity: These images where *A. sinensis* is in close proximity to its natural surroundings, such as foliage, water bodies, or terrain, making it challenging to distinguish the target from the background owing to camouflage or blending effects.

Given the nature of the task and the focus of the dataset on detecting camouflaged *A. sinensis* in their natural habitats, it is reasonable to expect that a significant portion of the dataset will comprise images with environmental proximity. This category likely constitutes the largest proportion of the dataset, reflecting real-world challenges in detecting cryptic species in their natural environments.

Following this reasoning, we can estimate the distribution of images in the dataset:

Normal Images: Relatively few images account for approximately 2 % of the total dataset, because these images represent ideal detection scenarios where the target is clearly visible.

Overlapping Objects: This category may account for approximately 20 % of the dataset, because overlapping objects introduce additional complexity into the detection algorithm.

Ambient Color Artifacts: This category is likely to make up approximately 13 % of the dataset because color artifacts can occur in a variety of environmental conditions but may not be as prevalent as other challenges.

Environmental proximity: This category is expected to make up the majority of the dataset, perhaps 65 % or more, reflecting the primary focus of the task of detecting *A. sinensis* in natural habitats.

Analyzing the dataset further to identify specific environmental artifacts, such as puddles, duckweed, shores, and ledges, provides deeper insights into the challenges faced in detecting camouflaged *A. sinensis* in different contexts. These artifacts play a significant role in obscuring target objects and complicating the detection process.

Puddle: Images featuring puddles can introduce challenges related to reflections, distortions, and varying water depths, affecting the visibility of *A. sinensis*. Given the diverse range of environmental conditions under which puddles can occur, the number of images featuring puddle artifacts may vary widely. In a dataset of 2464 images, it is reasonable to estimate that approximately 10–20 % of the images may feature puddle artifacts.

Duckweed: Duckweed, a common aquatic plant, often forms dense mats on the surface of water bodies, concealing the objects beneath and making detection challenging. The prevalence of duckweed artifacts may depend on factors, such as the type of water bodies represented in the dataset and the seasonality of duckweed growth. In a dataset focused on detecting *A. sinensis* in aquatic habitats, images featuring duckweed artifacts constituted a substantial portion, potentially ranging from 20 % to 30 % of the total dataset.

Shore: Where land meets water, unique challenges for detection are presented owing to the complex interplay between terrestrial and aquatic environments. Shore artifacts may include features such as rocks, vegetation, mud banks, and varying terrain textures, which can obscure the presence of *A. sinensis*. Depending on the emphasis of the dataset on shoreline habitats, images featuring shore artifacts could range from 15 % to 25 % of the total dataset.

Ledge: Ledges, elevated or protruding features along the water edge or within aquatic environments, can create shadowed areas and visual obstructions. The detection of *A. sinensis* near ledges may be challenging because of partial concealment, shadows, and perspective distortions. In a dataset capturing diverse aquatic and terrestrial habitats, images featuring ledge artifacts may account for approximately 10–15 % of the total dataset.

These values summarize the potential distribution of the artifact categories in the dataset and highlight the complexity of the detection task in a natural environment. Detailed labeling of images to categorize them according to specific artifact types and locations would yield more precise proportions for each category, thereby facilitating targeted analyses and model development.

Low recognition rates for *A. sinensis* at different locations are shown in Fig. 2.

To avoid model overfitting, data enhancement methods, such as

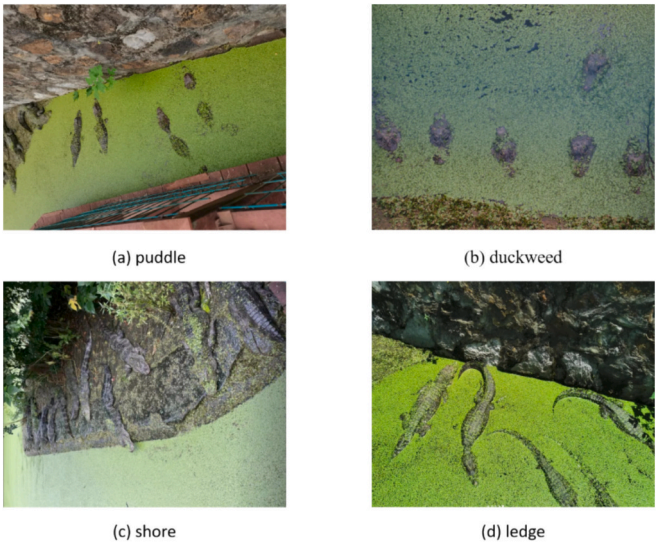


Fig. 2. *A. sinensis* in different locations.

cropping, flipping, and panning, were applied to various cropped images, expanding the dataset to three times its original size. The entire image dataset was divided into three categories: training, validation, and testing, with split ratios of 80 %, 10 %, and 10 %, respectively, with 2464 images in the training set, 308 images in the validation set, and 308 images in the testing set.

The rectangular regions of the cropped images were manually annotated using the commercial annotation tool ROBOFLOW (https://roboflow.com) and the results were saved as bounding box coordinates in an XML file. The XML file was processed using the Python image library and converted into TXT format.

3.2. Model

YOLO v8 has five distinct scale models (n, s, m, l, and x) that cater to the requirements of various research projects. These models were designed using scale factors similar to those found in YOLOv5, resulting in sequential increases in both the depth and width.

Images of *A. sinensis* in their natural habitat exhibit low detectability owing to features such as coloration that closely resemble the surrounding environment and the relatively small target size. A comprehensive parameter breakdown for each layer of the enhanced YOLO v8-SIM algorithm is presented in Table 1.

This paper presents a comprehensive overview of the strategic enhancements implemented in the YOLO v8-SIM model, which was specifically designed to optimize the detection of *A. sinensis* in its natural habitats, particularly in instances where the species is camouflaged. This section establishes a foundation for comprehending the sophisticated methodologies integrated into the model, each designed to address distinct challenges in ecological monitoring. We begin by introducing a two-layer attention mechanism that markedly enhances a model's capacity to discern subtle nuances in intricate visual scenes. Subsequently, we describe the adaptation of an inner IoU-optimized loss function that facilitates precise target localization, which is essential for improving the model accuracy under varying environmental conditions. Further enhancements include the incorporation of a slim-neck cross-layer hopping technique that strikes a balance between computational efficiency and performance integrity. In addition, we provide a comprehensive account of the modifications made to the model architecture to facilitate robust and efficient feature extraction, which is crucial for handling the intricacies of camouflaged object detection. Collectively, these modifications not only enhance the technical robustness of the model but also ensure its applicability in critical conservation efforts, offering a significant advancement over traditional detection systems.

Table 1
Overall parameters of YOLO v8-SIM.

Layers	from	parameters	module	arguments
0	−1	11,176,512	Resnet18	[False]
1	−1	394,240	SPPF	[512, 256, 5]
2	−1	265,728	Biformer	[256, 8, 7]
3	−1	0	Upsample	[None, 2, 'nearest']
4	[−1,3]	0	Concat	[1]
5	−1	145,984	VoVGSCSP	[512, 128, 1]
6	−1	0	Upsample	[None, 2, 'nearest']
7	[−1,2]	0	Concat	[1]
8	−1	37,152	VoVGSCSP	[256, 64, 1]
9	−1	19,360	GSconv	[64, 64, 3, 2]
10	[−1,9]	0	Concat	[1]
11	−1	105,024	VoVGSCSP	[192, 128, 1]
12	−1	75,584	GSconv	[128, 128, 3, 2]
13	[−1,6]	0	Concat	[1]
14	-1	414,848	VoVGSCSP	[384, 256, 1]
15	[12, 15, 18]	897,664	Detect	[80, [64, 128, 256]]

"–1" indicates downward passage through one layer, and "3″ indicates upward passage through three layers, reflecting the data connectivity between various layers of the network.
"Parameters" refers to the number of parameters, serving as an evaluative metric for a model. The greater the number of parameters, the more intricate is the model, representing a dimensionless metric.

3.2.1. Introducing the ResNet-18 network

The ResNet-18 [35]architecture represents a form of deep CNN that is composed of 18 weighted layers as shown in Fig. 3. and is predominantly utilized for image recognition and classification. At its foundation, it incorporates a residual learning framework designed to address the degeneracy problems often encountered during the training of deep networks. The architecture can be described as follows:

The initial layer of the network begins with a convolutional layer of 7×7 dimensions with a stride of two, which extracts preliminary features and reduces the image size. This was immediately achieved using a maximum pooling layer with 3×3 dimensions and a stride of two, which further decreased the spatial dimensionality of the resulting feature maps.

The core of the network was formed by four sequential sets of residual blocks, each comprising two convolutional layers of 3×3 dimensions with an equal count of output feature maps. The first block utilizes 64 channels, with each subsequent block doubling the channel count to 128, 256, and 512 channels. These blocks incorporate jump connections that facilitate the direct addition of input to the output of the convolutional layers, thus allowing the inputs to bypass certain layers and ensure uninterrupted signal transmission through the network, even in the absence of learning.

Down-sampling within the network was accomplished by implementing a convolution with a stride of two in the inaugural convolutional layer of each residual block. Concurrently, as the number of channels increases, dimensional compatibility is maintained at the junctions through 1×1 convolutions.

Posterior to every convolutional layer is a Batch Normalization layer, followed by a ReLU activation function. This sequence normalizes the output and introduces non-linearity into the system.

The network concludes with a global mean-pooling layer that succeeds in all the residual blocks. This layer effectively condenses the spatial dimensions of each feature map to one, thereby reducing the model parameters and mitigating the risk of overfitting. The final stage of the network consists of a fully connected layer that transforms the average pooled features into definitive category predictions.

This intricate structure endows ResNet-18 with the capacity to train deep networks proficiently while circumventing the performance degradation that typically accompanies deeper network configurations. The incorporation of jump connections is pivotal, as it allows for a direct

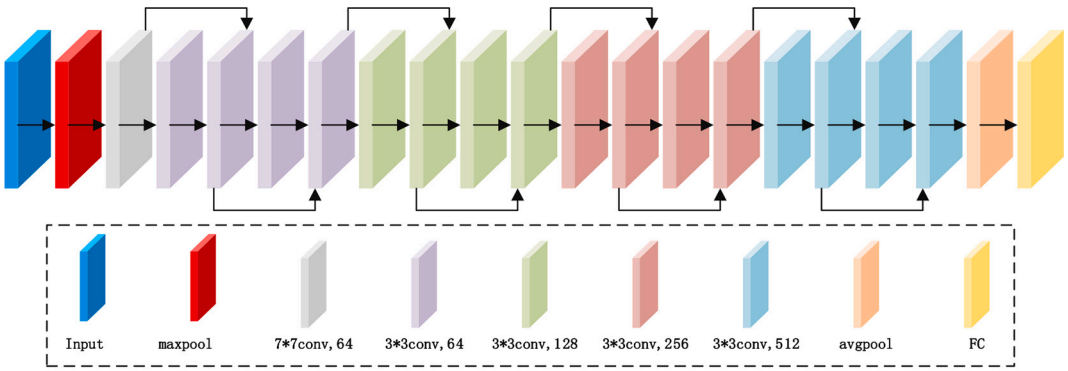


Fig. 3. ResNet-18 network structure.

pathway for gradient flow, thus resolving the vanishing gradient problem in deep networks and facilitating effective training and convergence, even for substantially deep networks. The practicality of this architecture has been substantiated through its utility in complex tasks, such as camouflage recognition, where it demonstrates exceptional performance.

3.2.2. RA module

In the domain of deep learning, particularly in CNNs tasked with object recognition, attention deviation during training is a significant challenge. This deviation can result in an inadvertent shift in focus from the primary object of interest to surrounding contextual information. In situations where the objective is to isolate a specific object from an image, the network must navigate the intricacies of the image's texture and color features as well as potential interference from background elements that may share similarities with the target.

To address this issue, a reverse attention (RA) (Chen et al., 2018) mechanism is used to redirect the focus of the network onto background elements that are distinct from the target object. In this instance, the target species was *A. sinensis*. This redirection is essential for enhancing the capacity of the network to differentiate between the background and objects with analogous features. The RA mechanism accentuates the peripheral features of *A. sinensis* boundary contours, thereby improving the precision and distinctness of the contours. The RA mechanism in a neural network is shown in Fig. 4. This mechanism involves upsampling high-level features, applying a sigmoid function, subtracting the result from unity, and multiplying it element-wise with low-level characteristics to improve object identification.

The 'low -level' and 'high -level' in the section are located in the ninth and tenth layers of the YOLOv8 model, respectively. This mechanism effectively enhances feature fusion by leveraging both low- and high-level information, contributing to improved object identification performance.

The architecture of RA integrates a series of operations: high-level feature maps, which encapsulate more profound semantic content and are critical for identifying the target object region, undergo upsampling through bilinear interpolation to align with the dimensions of the antecedent feature map layers. The normalization of these feature values to the [0,1] range is achieved through a sigmoid activation function, preparing the data for subsequent differential computations with a matrix composed entirely of ones.

The computation yields an output feature map derived from element-wise multiplication of the normalized high-level feature map with the lower-level feature map. After calculating the inverse augmentation weight matrix, the final output feature map is produced by element-wise multiplication with the lower-level feature map. Encapsulation was performed using the following equation:

$$\text{out} = \text{low} \otimes (1 - \sigma(\text{Up}(\text{high}))) \quad (1)$$

Here, $\otimes$ represents element-wise multiplication, low and high denote the lower and higher level feature maps, σ signifies the Sigmoid activation function, and Up stands for the bilinear interpolation upsampling operation. This methodology effectively enables the network to accentuate the target object against complex backgrounds, thereby optimizing the network training phase for more accurate object recognition.

3.2.3. BIformer

The integration of the Biformer (Zhu et al., 2023) attention mechanism into the slim-neck network, which represents the superior stratification of the neck network, significantly enhances the network's attentional capabilities, engendering improved harmony and performance between the layers. The former mechanism is characterized by a four-tiered pyramidal structure.

Commencing with the inaugural stage, the mechanism uses a patch-embedding module that serves the dual purpose of reducing the input-space resolution while concurrently amplifying the channel count. The subsequent stages, specifically the second through the fourth, utilize Patch Merging to further reduce the spatial resolution and increase the

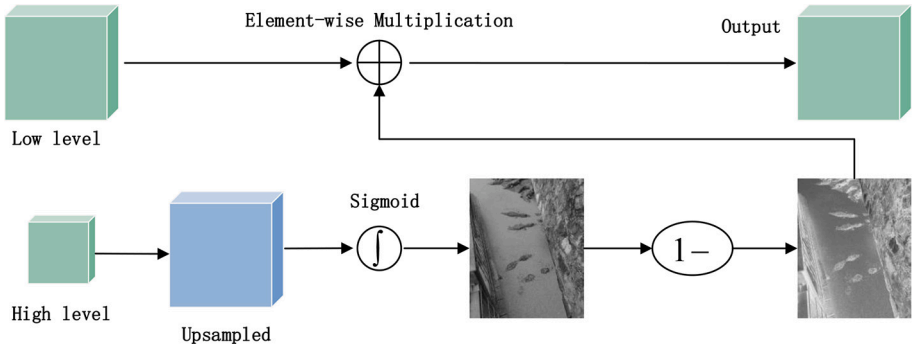


Fig. 4. RA module diagram.

number of channels. This reduction and increase are integral to the preparatory phase of the feature transformation carried out by successive biformer blocks.

Each module commences with a deep convolution to implicitly encode the Location Number (LN), thus facilitating the intrinsic understanding of the spatial hierarchy of the network. The former attention mechanism then uses the BRA module, followed by a multilayer perceptron (MLP) (Cuevas-Vargas et al., 2022). These modules are sequentially applied to model cross-positional relationships and embed features on a per-position basis.

The algorithmic structure of the attention mechanism is delineated in the associated Fig. 5**.** Initially, an input feature map undergoes linear mapping to derive the QKV (Query, Key, Value) matrices. Subsequently, a directed graph was formulated using the adjacency matrix to discern the participation relationships among the varying key-value pairs. This process aims to identify the pertinent regions for each region. The culmination of this process involves the establishment of a region-to-region indexing matrix, which is then applied to token-to-token attention.

The former method economizes on the parameter count and computational demand by aggregating key-value pairs from the foremost k relevant windows and using a sparsity operation that omits the computation of nonpertinent regions. Because of its dynamically sparse nature, the former attention mechanism facilitates a more adaptable allocation of computations and heightened content awareness, which translates into notable performance and computational efficiency. This innovation underscores the potential of the former attention mechanism to augment the functionality of CNNs in tasks requiring a refined attentional focus and efficiency. To summarize. An image is segmented into patches, and the key features (Q, K, and V) are extracted and processed using the attention module (A). The module captures intricate spatial relationships by leveraging a sparse attention pattern, which is then condensed through matrix multiplication to yield an enhanced feature representation (θ).

The former neural network architecture, which is an advanced attention mechanism, is shown in Fig. 6. This schematic represents the transformation of an input image through a series of stages: patch embedding, former blocks, and patch merging, culminating in the former output. Each stage strategically reduces the spatial resolution while amplifying the channel depth, and optimizes the network for efficient feature extraction and transformation.

The abbreviation “LN” in the figures stands for layer norm, which is a normalization technique applied to stabilize the inputs of each layer. By normalizing the inputs, the Layer Norm helps maintain consistent scaling and variance, thereby accelerating training and improving model performance.

3.2.4. Slim-neck by GSConv and VoV-GSCSP

In the advanced slim-neck architecture of the YOLO v8 model, the typical C2f module was replaced by the VoV-GSCSP module, (Li et al., 2022) and the standard convolution (SC) was succeeded by the GSConv module. This strategic enhancement is represented in the schematic designated as Fig. 7**.** The GSConv module is designed to minimize the loss of semantic information usually associated with the processes of spatial compression and channel expansion in CNNs.

The computational workflow of GSConv is outlined in Fig. 8**.** commences with a downsampling operation using a SC. This is followed by a deep convolution (DWConv), which amalgamates the results through concatenation, along with the initial convolution. This sequence culminates in a Shuffle Convolution step that adjusts the channel count by concatenating the channels from the two preceding convolutions.

The GSConv mechanism is characterized by dense convolution operations, which bolster the preservation of latent interchannel connections. However, its application across all model stages resulted in a denser network layer structure, substantially increasing the inference time. As a solution, GSConv is selectively used within the neck network, where it operates on feature maps that have reached their peak channel dimensions and minimizes the width and height, negating the need for further transformations.

The utilization of GSConv in the attention mechanism at this juncture ensures minimal information redundancy, highlighting the need for compression and enhancing its efficacy of the attention mechanism. GSConv distinguishes itself as a lightweight convolutional approach, incurring 60–70 % of the computational cost associated with SC.

Conversely, the VoV-GSCSP module aims to reduce both the computational complexity and network structural complexity. This is achieved by implementing a cross-level network structure that maximizes the retention of semantic information while preserving sufficient accuracy. This module represents a pivotal design decision in the slim-neck network, reflecting a concerted effort to balance the computational efficiency while maintaining the model accuracy.

In the VoV-GSCSP architecture, the terms “C1” and “C2” denote the number of channels at different stages of the convolutional layers. Initially, the input feature map contained C1 channels, which were reduced to C1/2 channels when entering the GS bottleneck. This reduction minimizes computational complexity and enhances efficiency. Within the GS bottleneck, grouped convolutions (GSConv) process the features, and the channel count is restored to the C1 channels to maintain consistency and prevent information loss. The processed feature map was finally output with C2 channels, completing the transformation through the VoV-GSCSP architecture.

3.2.5. Inner IoU

To address the limited generalization capabilities and protracted

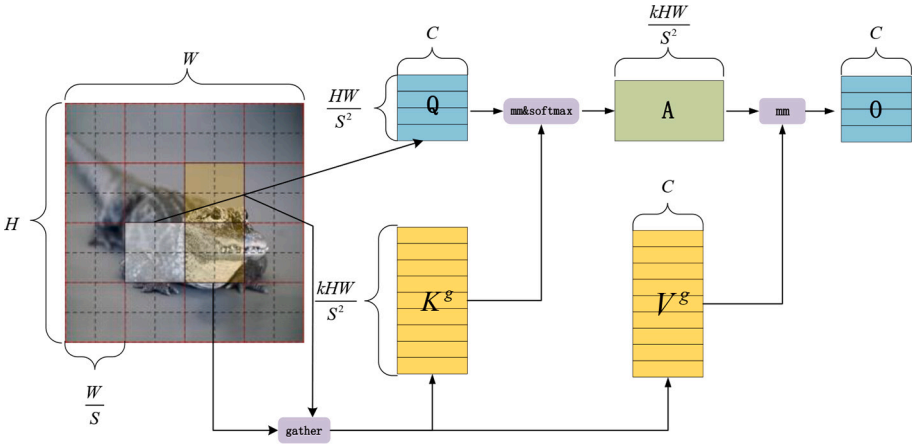


Fig. 5. Feature extraction eIformer.

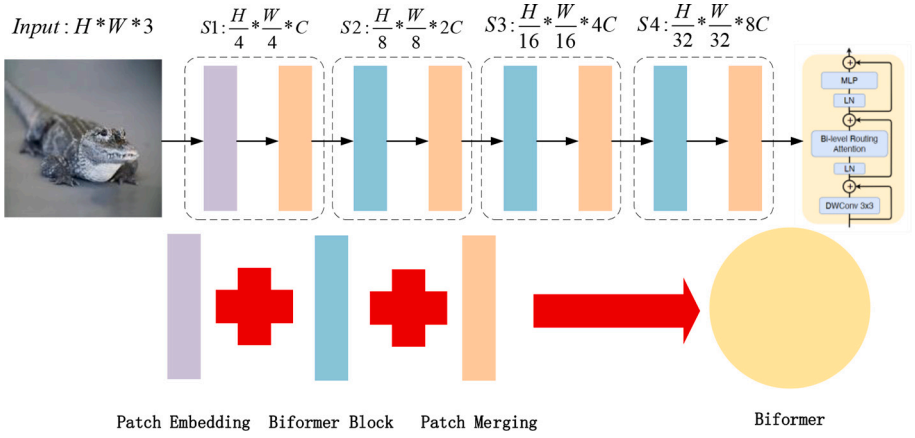


Fig. 6. BIformer flowchart.

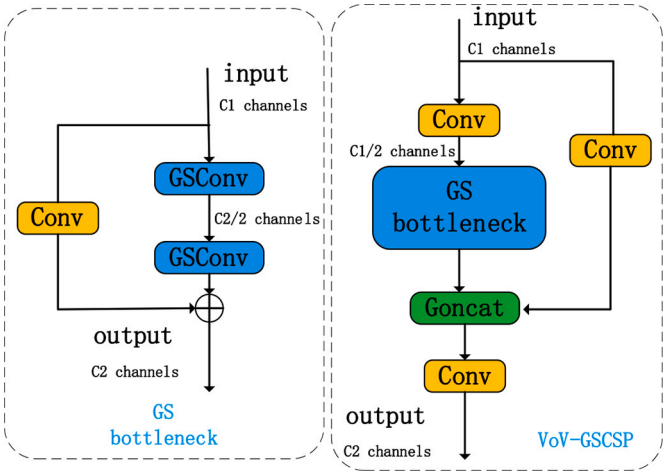


Fig. 7. Schematic Representation of VoV-GSCSP Architecture.

convergence rates associated with current IoU losses in various detection tasks, the introduction of an auxiliary bounding box to compute the losses and expedite the bounding box regression process has been proposed (Zhang et al., 2023). This approach, termed inner IoU, incorporates a scale factor that modulates the dimensions of the auxiliary bounding box. The generalization shortcomings of existing IoU methodologies can be effectively circumvented by using auxiliary bounding boxes of varying scales tailored to specific datasets and detectors.

As depicted in Fig. 9., the ground truth (GT) and anchor boxes are designated as b^{gt} and b, respectively. The centroids of the GT box and the internal GT box are represented by the coordinates (x_c^{gt}, y_c^{gt}), whereas (x, y) denote the centroids of the preselector box and the internal preselector box. The dimensions of the GT box are expressed as w^{gt} for width and h^{gt} for height, and similarly, w and h denote the width and height of the preselector box.

The ratio variable acts as a scale proportion factor, and is confined to the interval *[0.5, 1. 5]*. When this scale factor is less than unity, the auxiliary bounding box is smaller than the actual bounding box, thereby narrowing the effective regression range relative to the IoU loss. However, this reduction in range was counterbalanced by an increase in the absolute value of the gradient, which in turn catalyzed the convergence of samples with high IoU values. Conversely, a ratio exceeding unity signifies that the larger-scale auxiliary bounding box broadens the effective range of the regression, thus amplifying the potency of the regression for samples with a low IoU.

The integration of the inner IoU loss into the prevailing IoU-based bounding box regression loss functions enables a more nuanced calculation of the IoU, as expressed in the corresponding equation. This innovation facilitates a more refined and swift convergence of bounding box regression, thereby augmenting the performance of the detection models across diverse datasets.

The formula for the calculation process of inner IoU is expressed as follows:

$$b_l^{gt} = x_c^{gt} - \frac{w^{gt} * ratio}{2}, b_r^{gt} = x_c^{gt} + \frac{w^{gt} * ratio}{2} \tag{2}$$

$$b_t^{gt} = y_c^{gt} - \frac{h^{gt} * ratio}{2}, b_b^{gt} = y_c^{gt} + \frac{h^{gt} * ratio}{2} \tag{3}$$

$$b_l = x_c - \frac{w^{*} ratio}{2}, b_r = x_c + \frac{w^{*} ratio}{2} \tag{4}$$

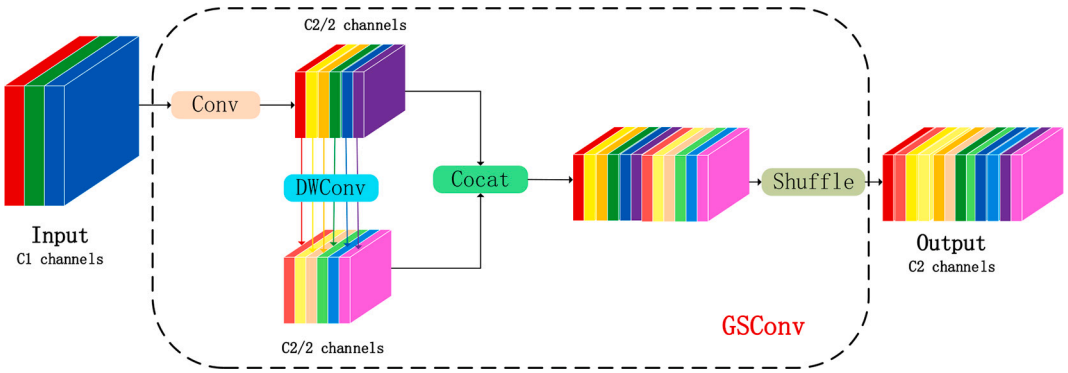


Fig. 8. GSConv module workflow diagram.

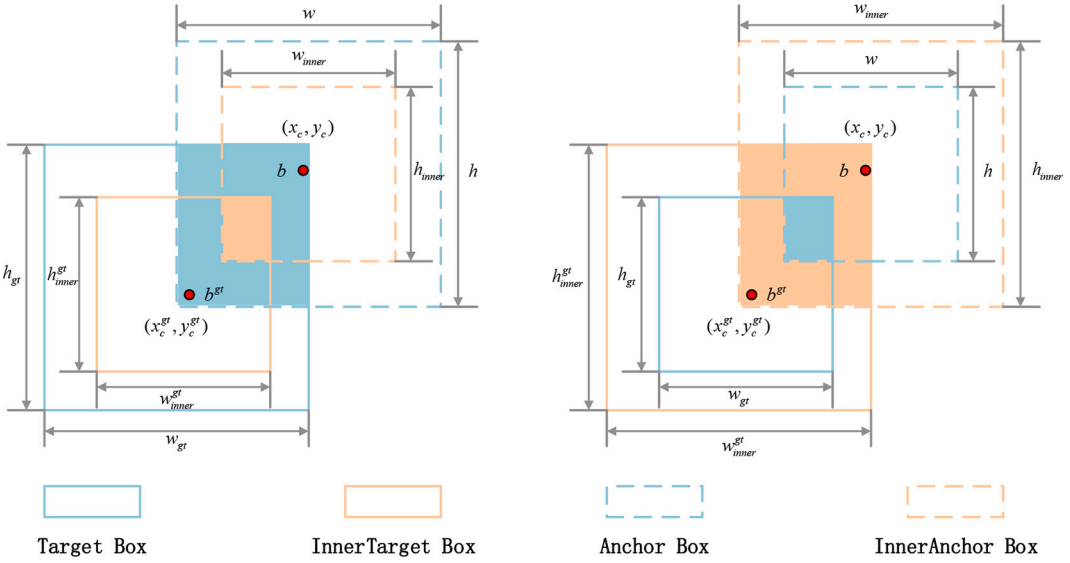


Fig. 9. Inner IoU bounding box analysis.

$$b_t = y_c - \frac{h^* ratio}{2}, b_b = y_c + \frac{h^* ratio}{2} \quad (5)$$

$$union = \left(w^{gt} * h^{gt}\right) * (ratio)^2 + (w * h) * (ratio)^2 - inter \quad (6)$$

$$inter = \left(min\left(b_r^{gt}, b_r\right) - max\left(b_l^{gt}, b_l\right)\right) * \left(min\left(b_b^{gt}, b_b\right) - max\left(b_t^{gt}, b_t\right)\right) \quad (7)$$

$$IoU^{inner} = \frac{inter}{union} \quad (8)$$

$$L_{Inner-IoU} = 1 - IoU^{inner} \quad (9)$$

4. Experiments

4.1. Experimental configuration

The hardware setup used for image processing and model training in this work comprises:

- CPU: Intel(R) Xeon(R) E5 2689 @2.60 GHz;
- Memory: 16 GB DDR4 RAM;
- Graphics processor: NVIDIA GeForce RTX 3070TI.

The deep learning model was implemented using the following software configuration:

- OS: Microsoft Windows 10 64-bi;
- CUDA: CUDA version: 11. 8.0;
- Python: 3.9.17;
- PyTorch: 1. 9.0;
- numpy 1. 26.0.

4.2. Parameter setting

The calculated values of the parameters following the adjustment of different hyperparameters to attain the best possible performance of the model are listed in Table 2. To improve the speed at which convergence occurs, the learning rate was adjusted to 0.01. The optimization technique relies on two crucial components: momentum and weight-decay. Increasing the value of momentum improves the stability of the optimization path, indicating a higher level of stability. On the other hand, weight-decay controls the extent of L2 regularization, where a smaller value indicates a lesser impact of regularization on the loss of function of the model. The parameters were configured as follows:

Table 2
Hyperparameterisation.

Parameter name	Parameter value
Momentum	0.937
Weight_decay	0.0005
Batch_size	32
Learning_rate	0.01
Epochs	300
Workers	8
Patience	100

The momentum was set to 0.937, the weight decay was set to 0.0005, and a maximum of 300 epochs were imposed to prevent potential overfitting of the model. Table concisely summarizes the hyperparameters used in the study.

4.3. Evaluation criteria

The YOLO v8 model predominantly uses a set of core evaluation metrics, including precision (P), recall (R), AP, and mAP. These metrics are represented mathematically as follows:

$$P = \frac{TP}{TP + FP} \quad (10)$$

$$R = \frac{TP}{TP + FN} \quad (11)$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (12)$$

$$AP = \int_0^1 Precision_{(Recall)} dR \quad (13)$$

$$AP = \frac{TP + TN}{TP + FN + FP} \quad (14)$$

$$mAP = \frac{1}{Q} \sum_{q=Q} AP(q) \quad (15)$$

True Positive (TP) represents the number of correctly identified positive samples, False Positive (FP) represents the number of

mistakenly identified negative samples, False Negative (FN) represents the number of falsely predicted negative samples, and True Negative (TN) represents the number of correctly identified negative samples. The AP metric calculates the average accuracy by considering all possible recall rates. To enhance the accuracy of the assessment, the average AP value for each category was calculated using an Intersection Over Union (IOU) threshold of 0.5.

In situations where the result is favorable, R denotes the complete count of precisely remembered events. Relying exclusively on precision (P) and recall (R) as the sole criteria for evaluating model performance may yield inconsistent outcomes owing to the inherent limits of these measures. Therefore, it is advisable to include Average Precision (AP) as it offers a thorough perspective on the precision–recall curve along both axes. As the Intersection over Union (IOU) threshold approaches 0.5, the model performance is projected to increase, resulting in a maximum Average Precision (mAP) score at an IOU threshold of 0.5.

A comprehensive breakdown of all the test results is provided in Table 3. Categories include the true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). This table serves as a confusion matrix, providing graphical and numerical depictions of the performance of the model. Using this complete method, we can guarantee a more precise and sophisticated assessment of the model's ability to make correct predictions.

This confusion matrix displays the classification accuracy of the proposed model, categorizing the predictions into true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). This representation is crucial for evaluating the performance of a model in terms of precision and recall.

4.4. Experimental results

The objective of this project was to improve YOLO v8 by making targeted improvements to address the challenge of detecting items that are inherently difficult to identify, such as those with hidden or disguised characteristics. During the training phase, each model was trained using consistent parameter settings and dataset images with an input image resolution of 640 × 640 pixels. The models were initially trained using the COD10K dataset for camouflage detection. Subsequently, the acquired models are subjected to migration learning using the collected datasets.

To confirm the superiority of the YOLO v8-SIM model, which uniquely combines several scenarios (Chen et al., 2018; Cuevas-Vargas et al., 2022; He et al., 2016; Li et al., 2022; Zhang et al., 2023; Zhu et al., 2023), we compared it with the original YOLO v8 model network. In addition, we trained and tested several well-known object recognition models, including SSD, Faster-R CNN, YOLOv5, and WilDect-YOLO, using identical datasets and model parameters.

The fluctuations in the performance metrics for each model throughout the training process as observed from the saved training logs are shown in Fig. 10, and the fluctuations in the loss function during training are shown in Fig. 11.

The plot illustrates the mAP over the training epochs for various object detection models. The initial training phase exhibited periodic fluctuations, which are characteristic of the intrinsic properties of the dataset. Notably, the model YOLO v8-SIM achieves a superior mAP early in training, approximately at the 50-epoch mark, and thereafter sustains a plateau that outperforms competing models. This indicates that YOLO v8-SIM not only converges swiftly to a high mAP, but also maintains its leading performance, underscoring its robustness and effectiveness in the given detection tasks.

Table 3
Confusion matrix for model evaluation.

Confusion matrix		Actual value	
		Actual positive	Actual negative
Predicted value	Predicted positive	True positive	False positive
	Predicted negative	False negative	True negative

The loss function trajectories for the same set of object recognition models across training epochs are shown in Fig. 11. Models such as SSD, Faster RCNN, and WilDect-YOLO demonstrated erratic patterns in loss reduction, suggesting instability in the learning process. In contrast, YOLO v8, YOLO v5, and YOLO v8-SIM exhibit a more stable decrease in loss values, reflecting the sophistication and reliability of their underlying architectures. Among these, YOLO v8-SIM stands out with the most pronounced and consistent decline, reinforcing its high-performance credentials in effectively optimizing the loss function over the training period.

The comprehensive detection outcomes of the various models on the test dataset are presented in Fig. 12 and Table 4. The YOLO v8-SIM model achieves higher precision, F1-score, and mAP@0.5 values than the original YOLO v8 model. Specifically, the precision, recall, and mAP@0.5 of the YOLO v8-SIM model were 91. 0 %, 89.9 %, and 92.3 %, Moreover, the other metrics of precision (P) demonstrate an improvement of 4.2 %, recall (R) by 3.3 % and average accuracy (mAP) by 4.5 %.

The YOLO v8 method, although faster than YOLO v8-SIM, exhibits a 4.29 % decrease in precision. Compared with the SSD, Faster-RCNN, YOLOv5, YOLO v8, and WilDect-YOLO models, the mean average precision (mAP) increased by 22.58 %, 7.45 %, 3.35 %, 4.53 %, and 1. 54 % respectively.

4.5. Comparison of results

4.5.1. Heat map

We generated heat maps using the gradient-weighted class activation mapping (Grad-CAM) technique, which provides insight into the parts of the input image that are most influential in the model's classification decision. This method captures the gradients that are fed into the final convolutional layer of the YOLO v8-SIM model to visualize the important regions of the image for predicting the presence of camouflaged *A. sinensis*.

The heat maps of YOLO v8-SIM algorithm with the YOLO v8 algorithm are compared in Fig. 13. These heat maps represent areas of interest or activation regions identified as significant by the algorithms.

The left side shows the original image and the middle image corresponds to the heatmap generated using the YOLO v8-SIM algorithm. The image on the side shows the original YOLO v8 algorithm-generated heat map.

4.5.2. Test results

The delineate comparative analyses between the original scene (leftmost image), YOLO v8-SIM optimized algorithm detections (middle image), and YOLO v8 algorithm outcomes (rightmost image) (Figs. 14–16).

The YOLO v8-SIM algorithm performance, where multiple bounding boxes with varying confidence levels surround *A. sinensis*, are shown in Fig. 14. The algorithm assigned high confidence scores to several instances of *A. sinensis*, such as 0.91, 0.93, and 0.88, indicating a robust certainty in the detections. Lower confidence scores, such as 0.32 and 0.67, are attributed to more distant *A. sinensiss*, likely due to factors like occlusion or the angle of capture. Conversely, the original YOLO v8 results on the right side suggest a lower degree of confidence in its detection, with scores of 0.67, 0.71, and 0.70, and a notable detection at a higher confidence of 0.83. The original YOLO v8 failed to detect more remote *A. sinensis* and erroneously identified rock formations as targets. Compared with YOLO v8-SIM, there are fewer detections and less variation in confidence levels.

A comparative analysis of algorithmic detection in the low recognition rate state of *A. sinensis* on shore is presented in Fig. 15. The YOLO v8-SIM algorithm distinctly delineates several *A. sinensis* organisms with bounding boxes and associated confidence scores ranging ranging

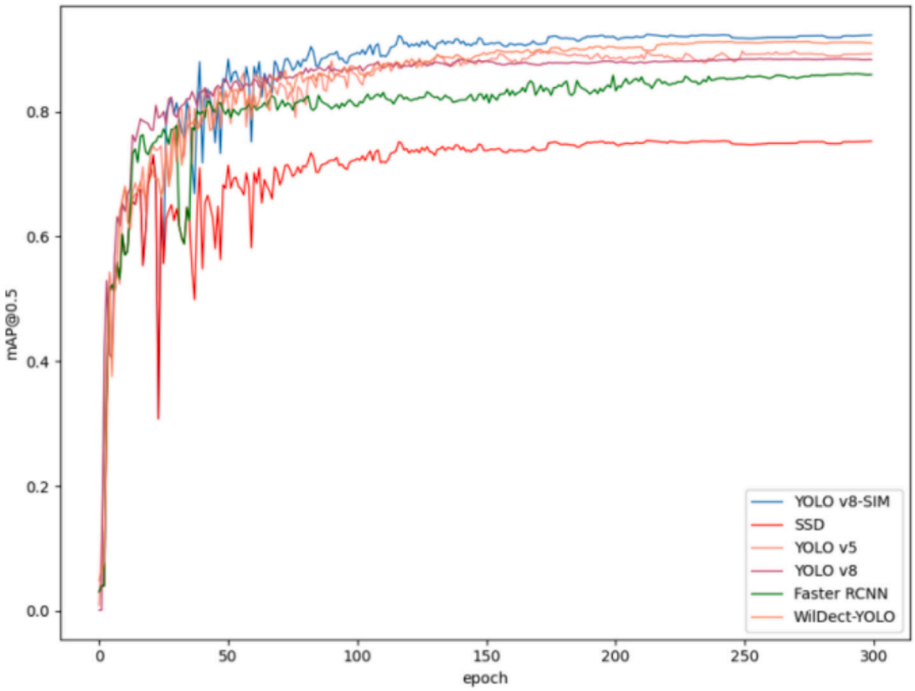


Fig. 10. Plot of changes in mAP@0.5.

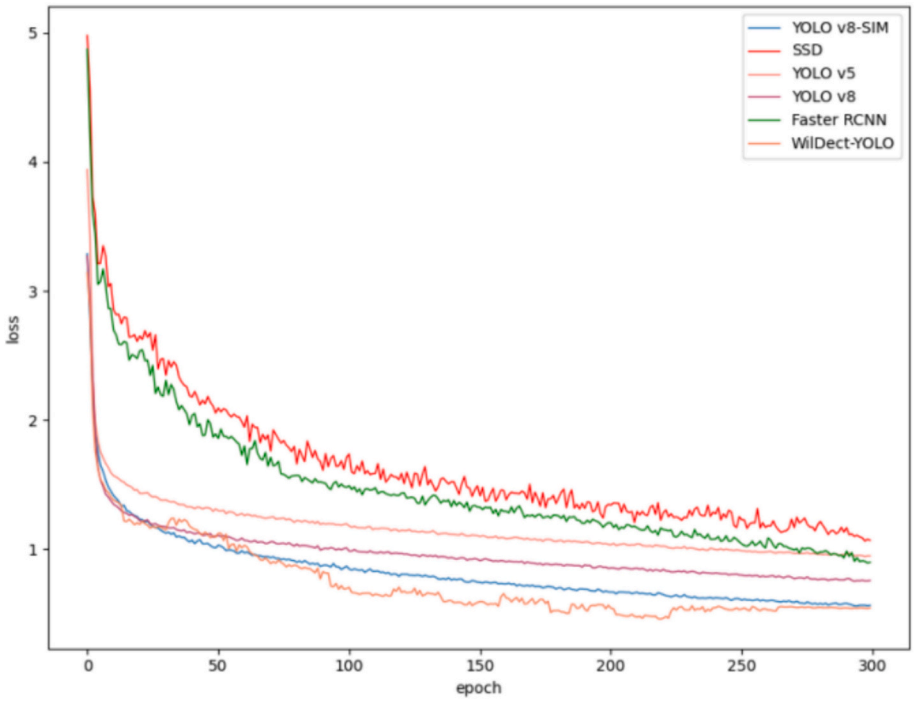


Fig. 11. Plot of change in loss function.

0.54–0.83. This indicates a comprehensive detection across the scene, encompassing the identification of the more distant *A. sinensis*. Conversely, the original YOLO v8, illustrated on the right, also acknowledged the presence of *A. sinensis* but with generally lower confidence scores of 0.57 and 0.39. Crucially, it failed to detect *A. sinensis* that YOLO v8-SIM identified with a confidence score of 0.83, and the number of *A. sinensis* detected in comparison to YOLO v8-SIM was significantly reduced.

A comparison of the YOLO v8-SIM and YOLO v8 algorithms for detecting *A. sinensis* in duckweed-filled water at low recognition rates are shown in Fig. 16. The YOLO v8-SIM algorithm identified *A. sinensis* with a range of confidence levels, notably 0.70, 0.48, 0.59, and 0.27, using appropriately sized and positioned red bounding boxes. An *A. sinensis* with a 0.48 confidence score is notably enclosed by two bounding boxes. The origin YOLO v8 algorithm, on the other hand, discerns two *A. sinensis* with lower confidence scores of 0.29 and 0.28, and bounding boxes similar in placement and size to those in the YOLO v8-SIM image, yet with diminished confidence. Moreover, the detection

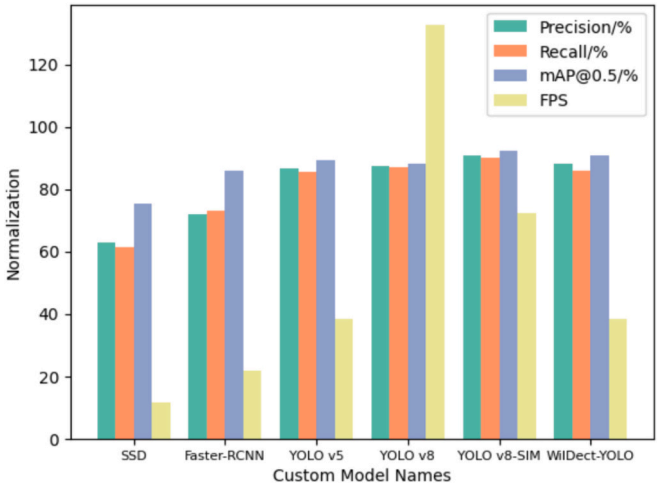


Fig. 12. YOLO v8-SIM comparison bar chart.

Table 4
Comparative analysis of experimental results across different models.

Model	Precision/%	Recall /%	mAP@0.5/%	FPS/(f·s-1)
SSD	63.1	61. 3	75.3	11. 89
Faster-R CNN	72.1	73.0	85.9	21. 92
YOLO v5	86.6	85.6	89.3	38.64
YOLO v8	87.3	87.0	88.3	132.54
YOLO v8-SIM	91. 0	89.9	92.3	72.21
WilDect-YOLO	88.1	86.1	90.9	38.50

rate is only 50 % of YOLO v8-SIM.

4.5.3. Ablation experiment

We conducted a series of ablation experiments to assess the contribution of each module to the YOLO v8-SIM model. Each key module was systematically removed or replaced, and the performance changes were recorded. The modules evaluated in our study include the ResNet-18 backbone, biformer, RA, inner IoU loss function, slim neck cross-layer hopping, and VoV-GSCSP modules. Observe the performance changes through removal or replacement. The key modules evaluated in our ablation study were as follows:

ResNet-18 Backbone: We replaced the ResNet-18 backbone with a simpler CNN to assess its impact on feature extraction.

Dual-Layer Attention Mechanism: We disabled the former and reverse attention mechanisms to observe their effects on target focus and background suppression.

Inner IoU Loss Function: We revert to a standard IoU loss function to evaluate the performance impact of the inner IoU optimized loss.

Slim Neck Cross-Layer Hopping: We replaced the Slim Neck mechanism with a standard neck structure to measure its effect on model compression and parameter reduction.

VoV-GSCSP Module: We replaced the VoV-GSCSP module with a standard bottleneck module to understand its role in enhancing the feature fusion.

The same dataset and training parameters were used for all the ablation experiments to ensure consistency. The dataset comprises 2464 enhanced images, as described in Section 3.1. Performance metrics including Precision, Recall, mAP@0.5, and FPS were recorded for each configuration. The results of the ablation experiments are summarized in Table 5.

The ablation experiments demonstrated that substituting the ResNet-18 backbone in the YOLO v8-SIM model markedly influenced feature extraction, resulting in a pronounced decline in performance metrics. Moreover, the integration of the dual-layer attention mechanism, comprising former and reverse attention, proved indispensable for

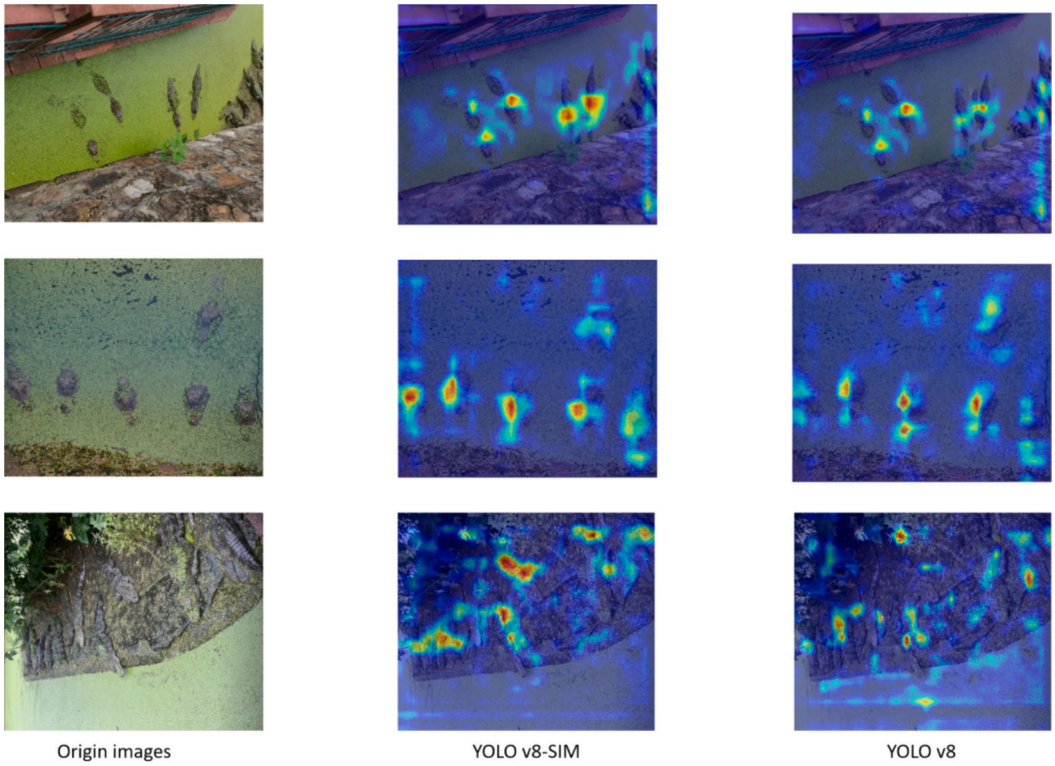


Fig. 13. Detection heat map.

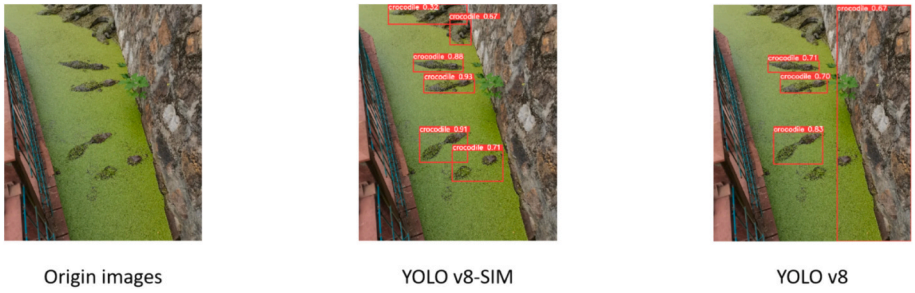


Fig. 14. *A. sinensis* camouflaged beside a ledge.

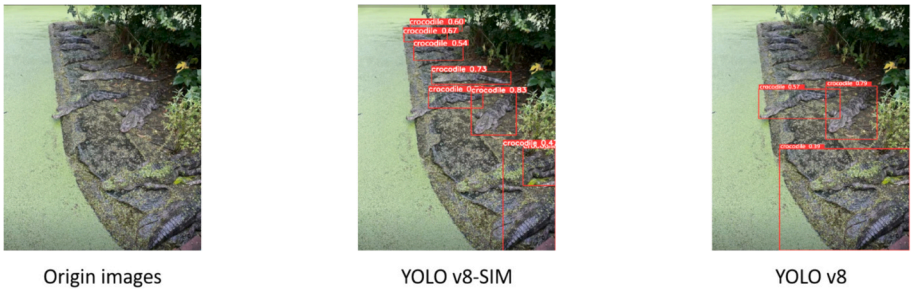


Fig. 15. *A. sinensis* camouflaged on a shore.

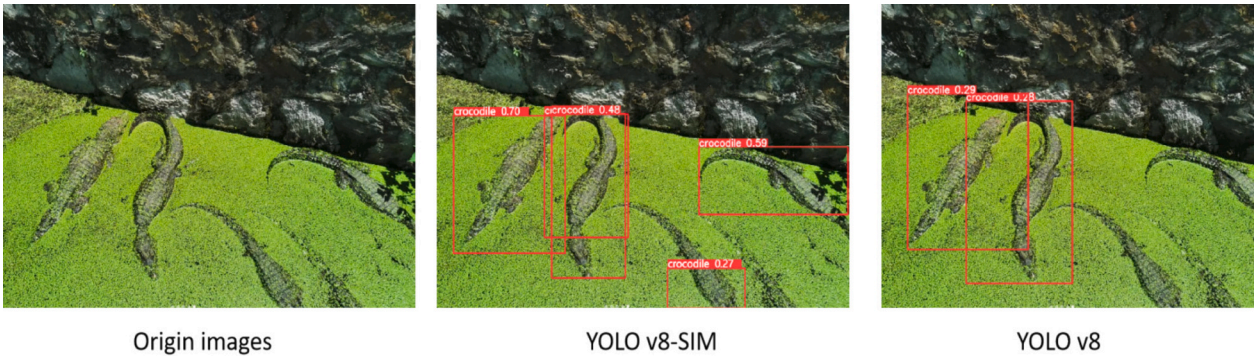


Fig. 16. *A. sinensis* camouflaged in a puddle.

Table 5
Comparison of the ablation experimental results.

Model	Precision/%	Recall /%	mAP@0.5/%	FPS/(f·s-1)
YOLO v8	87.3	87.0	88.3	132.54
YOLO v8-SIM	91	89.9	92.3	72.21
YOLO v8 + ResNet-18	88.3	87.0	89.1	122.45
YOLO v8 + RA	88.0	86.5	88.6	120.12
YOLO v8 + inner IoU	87.7	87.4	88.4	116.55
YOLO v8 + slim neck	87.9	87.6	88.7	109.90
YOLO v8 + biformer	88.2	86.9	89.0	81. 50
Without ResNet-18	87.5	87.0	88.9	75.80
Without RA	90.5	88.5	91. 8	73.50
Without inner IoU	89.0	86.9	90.1	72.90
Without slim neck	90.2	88.2	91. 2	75.30
Without biformer	87.9	85.7	88.7	72.00

enhancing the target focus and efficiently suppressing complex backgrounds. Each key module was thoroughly examined, and the results revealed its respective contributions. Notably, the absence of RA resulted in a discernible reduction in the mAP at an IoU threshold of 0.5, underscoring its pivotal role in enhancing the overall accuracy of the model. Furthermore, the inner IoU loss function was instrumental in improving generalization across diverse scales, with its absence resulting in diminished recall and mAP@0.5. In addition, the slim-neck cross-layer hopping mechanism facilitates model compression and parameter reduction while maintaining accuracy; however, its removal leads to a decline in precision and recall. Finally, the biformer module illustrates its significance in enhancing feature fusion and capturing spatial relationships. In the absence of this module, there was a discernible decline in the precision and recall metrics, substantiating its pivotal role in the efficacy of the model.

The results demonstrate that the YOLO v8-SIM model outperforms the basic YOLO v8 model and other configurations in terms of overall composite metrics. Despite having the lowest FPS in the ablation experiments, the difference was negligible, and an FPS of 72.21 is sufficient for real-time identification, thereby validating the effectiveness of the proposed enhancement.

5. Discussion

Autonomous wildlife monitoring presents a distinctive set of challenges, particularly in the detection of cryptic species, such as *A. sinensis*, within their natural habitats. This is primarily attributable to the camouflaging capabilities of *A. sinensis* and the dynamic complexity of

its natural environment. This study introduces YOLO v8-SIM, an optimized version of YOLO v8, designed to enhance the detection and identification of *A. sinensis* in foliage and water bodies.

YOLO v8-SIM incorporates a ResNet-18 backbone that facilitates the robust extraction of features while maintaining a lightweight architectural configuration. YOLO v8-SIM incorporates a biformer and RA double-layer mechanism to address the issue of attentional dilution in situations where the background and object of interest are of similar color. This combination sharpens the focus of the model and enhances its ability to distinguish *A. sinensis* from its environment, thereby enhancing target features and suppressing complex backgrounds.

During the optimization process, it became evident that a loss function capable of adapting to the varying scales of *A. sinensis* in different environmental contexts was necessary. The standard IoU loss functions were found to be inadequate, owing to their slow convergence and poor generalization capabilities. The introduction of an auxiliary bounding box through the inner IoU loss function enables the scaling of the bounding box to align with the characteristics of the dataset, thereby overcoming the limitations of existing methods in terms of generalization. This innovation markedly enhanced the detection of partially visible or juvenile *A. sinensis*, which are typically challenging to discern because of their small size and effective camouflage.

To enhance the model efficiency, we used a slim-neck cross-layer hopping mechanism for network pruning and compression. This technique significantly reduces the complexity of the model while maintaining its accuracy. Meticulous calibration ensured that this pruning did not result in loss of critical information.

5.1. Implications for ecological monitoring and conservation

The development of YOLO v8-SIM has significant implications for ecological monitoring and conservation efforts, particularly for cryptic species, such as *A. sinensis*. The enhanced detection accuracy of our model in complex natural environments addresses a critical gap in wildlife-monitoring technology.

First, the enhanced ability to detect camouflaged *A. sinensis* in various habitats, including water bodies with dense vegetation, enables more comprehensive population surveys. This improvement can facilitate more precise estimations of population size and distribution, which are essential for the formulation of effective conservation strategies and optimal allocation of resources.

Second, the capacity of the model to detect partially visible or juvenile *A. sinensis* contributes to a more comprehensive understanding of population dynamics and breeding success. This information is crucial for assessing the health and viability of *A. sinensis* populations, particularly in the context of reintroduction programs and protection.

Furthermore, the real-time detection capabilities of YOLO v8-SIM, operating at 72.21 FPS, facilitate novel possibilities for continuous monitoring and rapid response to potential threats or changes in *A. sinensis* behavior. This could be particularly valuable in managing human-wildlife conflicts or studying species responses to environmental changes.

5.2. Comparison to current studies and practical applications

The YOLO v8-SIM model that we developed draws on and expands the latest breakthroughs in technology used for detecting wildlife. Our findings can be directly compared with those of other recent studies, revealing similarities in methodology and distinctive additions.

Recent advancements in ecological informatics have focused on the creation of YOLO-based models for environmental and ecological monitoring applications (). For example, de Melo Lima et al. (Lima et al., 2024) created a streamlined model using YOLOv8, which was specifically designed to identify neotropical brown stink bugs in soybean fields. This study demonstrated the effectiveness of the model in detecting agricultural pests. Our study enhances object detection frameworks for camouflaged wildlife, aligning with the advancements made by Roy et al. (Roy et al., 2023) in their presentation of WilDect-YOLO for the detection of endangered animals. Both studies emphasize the significance of customizing YOLO structures to address the unique requirements of ecological monitoring. Bakana et al. (Bakana et al., 2024) introduced WildARe-YOLO, a model that shares similarities with ours in terms of lightweight and efficiency. This characteristic renders it suitable for real-time applications in diverse and expansive situations. The inclination toward enhancing YOLO models for ecological purposes was also apparent in Tang et al. (Tang et al., 2023), who enhanced pest detection capabilities by including an effective channel attention mechanism with a transformer encoder.

Wei et al. (Wei and Zhan, 2024) utilized a multicat module and a C2flite module to introduce the YOLO_MRC model in their research. This model effectively addresses the problem of real-time detection and counting of Tephritidae pests. Zhou et al. (Zhou et al., 2024) created the UODN model by combining cross-stage multi-branch (CSMB) and large kernel spatial pyramid (LKSP) modules, resulting in improved underwater object recognition skills. Feng et al. (Feng and Jin, 2024) improved pest identification using their CEH-YOLO model by incorporating a high-order deformable attention (HDA) mechanism and an enhanced spatial pyramid pooling fast (ESPPF) module. Our YOLO v8-SIM model stands out from previously described methodologies because of its unique capability to tackle specific difficulties in detecting camouflaged animals in intricate natural surroundings, including *A. sinensis*. The model combines a complex dual-layer attention mechanism (biformer and RA), an optimized inner IoU loss function, and a slim-neck cross-layer hopping technique. By utilizing this combination, YOLO v8-SIM achieved a notable level of precision, with an accuracy of 91 %, recall of 89.9 %, and mAP of 92.3 % at an IoU threshold of 0.5. Furthermore, it retains a real-time performance of 72.21 FPS. Unlike previous models, YOLO v8-SIM exhibits exceptional performance in identifying partially visible or smaller objects, making it suitable for monitoring biodiversity and conserving camouflaged species in natural environments.

These comparisons demonstrate that although recent research has made notable progress in wildlife identification systems, our YOLO v8-SIM model specifically tackles the difficulties of recognizing camouflaged species in intricate natural surroundings.

5.3. Potential applications

The development of YOLO v8-SIM provides a foundation for future research and applications in wildlife conservation and ecological monitoring.

The ability to detect multiple species simultaneously represents a significant advancement in wildlife conservation and ecological monitoring. Although our model has been optimized for *A. sinensis*, future studies should investigate its potential for adaptation to other camouflaged or cryptic species. This could lead to the development of a more versatile tool for monitoring biodiversity in various ecosystems.

Integration with Other Technologies: The combination of YOLO v8-SIM with other emerging technologies, such as environmental DNA analysis or acoustic monitoring, could provide a more comprehensive approach to wildlife monitoring. Such integration can facilitate the acquisition of multimodal data, thereby enhancing the robustness of ecological assessments.

Long-term Ecological Studies: The capacity of the model to consistently detect *A. sinensis* in diverse environmental contexts renders it well-suited for long-term ecological investigations. Further research should concentrate on utilizing this technology to examine behavioral patterns, habitat utilization, and responses to environmental alterations over an extended timeframe.

The potential of YOLO v8-SIM to inform conservation policies is significant. The precise data yielded by YOLO v8-SIM could serve as a basis for the formulation of conservation policies and development of

management strategies. Further research should investigate the potential integration of this technology into the decision-making processes for protected area management and species recovery plans.

Applications of Citizen Science: The adaptation of YOLO v8-SIM for use in smartphone applications could facilitate the involvement of citizen scientists in monitoring the efforts related to *A. sinensis*. This could markedly enhance the scope of data collection, while fostering public engagement in conservation initiatives.

In conclusion, our research involved overcoming several challenges, including preventing attention mechanisms from overfitting the highly textured features of the *A. sinensis* environment, which could lead to the species being missed. By modifying the biformer attention mechanism and the inner IoU scale, we enhanced the model's capacity to discern both the target and background. Furthermore, a slim neck mechanism was uses to reduce the number of parameters, thus ensuring that the model retained the capacity to detect *A. sinensis* in diverse environmental contexts and at varying scales.

The optimized model exhibited superior performance compared to the baseline YOLO v8, demonstrating a notable enhancement in the detection of camouflaged *A. sinensis* in intricate settings. This study underscores the pivotal role of targeted optimization in enhancing the efficacy of CNN for specific and challenging computer vision tasks in wildlife conservation and ecological monitoring.

6. Conclusions

To accurately detect camouflaged *A. sinensis* in their natural habitats, we propose the use of the YOLO v8-SIM model, which was designed to address the issues of low recognition rates and identification difficulties. In this study, we used a dual-layer attention mechanism with bias and reverse attention in conjunction with the inner IoU optimized loss function and a slim-neck cross-layer hopping mechanism to achieve notable improvements in detection accuracy. In accordance with the principle of openness (Huettmann and Arhonditsis, 2023), all data and program codes are available for review at https://github.com/Ap1rate/yolov8-SIM/.

The incorporation of the former and RA mechanisms facilitated the refinement of the model's focus on the target *A. sinensis*, enabling effective distinction from complex backgrounds. The inner IoU loss function influenced the model's capacity to generalize across varying scales, thereby enhancing the detection of partially visible or smaller *A. sinensis*. Furthermore, the slim-neck mechanism optimizes the network architecture, resulting in a reduction in computational requirements while maintaining the integrity of the detection capabilities.

The proposed model exhibited a precision of 91 %, recall of 89.9 %, and mAP at 0.5 of 92.3 % at a frame rate of 72.21 frames per second. These metrics demonstrate that the proposed model outperforms the other tested models, including SSD, Faster-RCNN, YOLO v5, YOLO v8, and WilDect-YOLO.

Furthermore, we propose a dataset for the identification of low-recognition *A. sinensis* in the wild, which will be made available as an open source in the near future.

Our work contributes to wildlife conservation by facilitating non-intrusive monitoring and studying low-detection-rate species such as *A. sinensis* in their natural habitats. The objective of our model was to address the low detectability of objects in the field. By improving the detection accuracy, our model is capable of collecting vital data on species behavior and population dynamics, thereby supporting conservation efforts. In animal sanctuaries, the proposed model facilitates enhanced monitoring by identifying and tracking individual animals with minimal human intervention. In a natural setting, our model markedly enhanced the probability of detecting *A. sinensis* in situations where background similarity was high, thereby facilitating a more comprehensive understanding of population metrics and habitat utilization.

This advancement also facilitates targeted conservation interventions, such as the strategic design of protected areas. Moreover, this near-background detection technology shows promise for future applications in precise monitoring of other cryptic species. However, this model requires further refinement to better handle scenarios with extremely well camouflaged or partially obscured target animals. Future studies should aim to increase the sensitivity of the model and reduce the number of false positives under difficult conditions. The aim is to continually improve the robustness of the model to provide conservationists and sanctuary staff with more reliable tools to effectively protect and study wildlife.

Funding

This research was supported by the 14th Five-Year Plan Fund for Educational Science in Shaanxi Province (No. SGKCSZ2020871), with the support and assistance of Jiangsu University of Science and Supported by Kunsan National University's Industry-Academia Cooperation Group (IACG) (grant number 2023H052).

Institutional review board statement

Not applicable.

CRediT authorship contribution statement

Yantong Liu: Writing – original draft. **Sai Che:** Data curation. **Liwei Ai:** Investigation. **Chuanxiang Song:** Formal analysis, Conceptualization. **Zheyu Zhang:** Methodology. **Yongkang Zhou:** Resources, Investigation. **Xiao Yang:** Writing – review & editing. **Chen Xian:** Funding acquisition.

Declaration of competing interest

The authors declare no conflicts of interest.

Data availability

All data and program codes are available for review at https://github.com/Ap1rate/yolov8-SIM/.

Acknowledgments

The author would like to thank Kunsan University and Jiangsu University of Science and Technology for assistance, Anhui *Alligator sinensis Alligator sinensis* Nature Reserve for providing the dataset, and the editors and reviewers for their guidance and comments.

References

A Computer Vision-Based Object Localization Model for Endangered Wildlife Detection by Arunabha Mohan Roy, Jayabrata Bhaduri, Teerath Kumar, Kislay Raj: SSRN, 2024. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4315295 (accessed December 24, 2023).

Adams, W.M., 2019. Geographies of conservation II: technology, surveillance and conservation by algorithm. Prog. Hum. Geogr. 43, 337–350. https://doi.org/10.1177/0309132517740220.

Arora, A., Dutta, P., Bapat, S., Kulathumani, V., Zhang, H., Naik, V., Mittal, V., Cao, H., Demirbas, M., Gouda, M., Choi, Y., Herman, T., Kulkarni, S., Arumugam, U., Nesterenko, M., Vora, A., Miyashita, M., 2004. A line in the sand: a wireless sensor network for target detection, classification, and tracking. Comput. Netw. 46, 605–634. https://doi.org/10.1016/j. comnet. 2004.06.007.

Bakana, Sibusiso Reuben, Zhang, Yongfei, Twala, Bhekisipho, 2024. WildARe-YOLO: a lightweight and efficient wild animal recognition model. Eco. Inform. 80, 102541.

Chen, S., Tan, X., Wang, B., Hu, X., 2018. Reverse Attention for Salient Object Detection, pp. 234–250. https://openaccess.thecvf.com/content_ECCV_2018/html/Shuhan_Chen_Reverse_Attention_for_ECCV_2018_paper.html (accessed January 4, 2024).

CSIRO PUBLISHING | Marine and Freshwater Research, 2024. https://www.publish.csiro.au/MF/MF08153 (accessed January 4, 2024).

Cuevas-Vargas, H., Camarena, J.L., Velázquez-Espinoza, N., 2022. Sustainability performance as a result of frugal innovation. The moderating effect of firm size. Proc. Comput. Sci. 214, 141–148. https://doi.org/10.1016/j.procs.2022.11.159.

Fan, Deng-Ping, et al., 2020. Camouflaged object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/CVPR42600.2020.00285.
Fan, Deng-Ping, et al., 2021. Concealed object detection. IEEE Trans. Pattern Anal. Mach. Intell. 44 (10), 6024–6042.
Fang, Y., Du, S., Abdoola, R., Djouani, K., Richards, C., 2016. Motion based animal detection in aerial videos. Proc. Comput. Sci. 92, 13–17. https://doi.org/10.1016/j.procs. 2016.07.316.
Farrell, J.A., Whitmore, L., Mashkour, N., Rollinson Ramia, D.R., Thomas, R.S., Eastman, C.B., Burkhalter, B., Yetsko, K., Mott, C., Wood, L., Zirkelbach, B., Meers, L., Kleinsasser, P., Stock, S., Libert, E., Herren, R., Eastman, S., Crowder, W., Bovery, C., Anderson, D., Godfrey, D., Condron, N., Duffy, D.J., 2022. Detection and population genomics of sea turtle species via noninvasive environmental DNA analysis of nesting beach sand tracks and oceanic water. Mol. Ecol. Resour. 22, 2471–2493. https://doi.org/10.1111/1755-0998.13617.
Feng, Jiangfan, Jin, Tao, 2024. CEH-YOLO: a composite enhanced YOLO-based model for underwater object detection. Eco. Inform. 102758.
Gonzalez, L.F., Montes, G.A., Puig, E., Johnson, S., Mengersen, K., Gaston, K.J., 2016. Unmanned aerial vehicles (UAVs) and artificial intelligence revolutionizing wildlife monitoring and conservation. Sensors 16, 97. https://doi.org/10.3390/s16010097.
Hahn, N.R., Bombaci, S.P., Wittemyer, G., 2022. Identifying conservation technology needs, barriers, and opportunities. Sci. Rep. 12, 4802. https://doi.org/10.1038/s41598-022-08330-w.
Han, W., Fan, R., Wang, L., Feng, R., Li, F., Deng, Z., Chen, X., 2021. Improving training instance quality in aerial image object detection with a sampling-balance-based multistage network. IEEE Trans. Geosci. Remote Sens. 59, 10575–10589. https://doi.org/10.1109/TGRS. 2020.3038803.
He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep Residual Learning for Image Recognition, pp. 770–778. https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html (accessed January 4, 2024).
He, F., Zarfl, C., Bremerich, V., Henshaw, A., Darwall, W., Tockner, K., Jähnig, S.C., 2017. Disappearing giants: a review of threats to freshwater megafauna. WIREs Water 4, e1208. https://doi.org/10.1002/wat2.1208.
Huettmann, Falk, Arhonditsis, George, 2023. Towards an ecological informatics scholarship that is reflective, repeatable, transparent, and sharable! Eco. Inform. 76, 102132.
Islam, S.B., Valles, D., 2020. Identification of wild species in Texas from camera-trap images using deep neural network for conservation monitoring. In: 2020 10th Annual Computing and Communication Workshop and Conference (CCWC), pp. 0537–0542. https://doi.org/10.1109/CCWC47524.2020.9031190.
Kellenberger, B., Volpi, M., Tuia, D., 2017. Fast animal detection in UAV images using convolutional neural networks. In: 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 866–869. https://doi.org/10.1109/IGARSS.2017.8127090.
Li, H., Li, J., Wei, H., Liu, Z., Zhan, Z., Ren, Q., 2022. Slim-neck by GSConv: A better design paradigm of detector architectures for autonomous vehicles. https://doi.org/10.48550/arXiv. 2206.02424.
Lima, De Melo, Pinheiro, Bruno, et al., 2024. A lightweight and enhanced model for detecting the Neotropical brown stink bug, Euschistus heros (Hemiptera: Pentatomidae) based on YOLOv8 for soybean fields. Eco. Inform. 80, 102543.
Linden, D.W., Fuller, A.K., Royle, J.A., Hare, M.P., 2017. Examining the occupancy–density relationship for a low-density carnivore. J. Appl. Ecol. 54, 2043–2052. https://doi.org/10.1111/1365-2664.12883.
Marshall, C.R., 2017. Five palaeobiological laws needed to understand the evolution of the living biota. Nat. Ecol. Evol. 1, 1–6. https://doi.org/10.1038/s41559-017-0165.
Nazir, S., Kaleem, M., 2021. Advances in image acquisition and processing technologies transforming animal ecological studies. Eco. Inform. 61, 101212 https://doi.org/10.1016/j. ecoinf. 2021. 101212.
Pan, T., Wang, H., Duan, S., Ali, I., Yan, P., Cai, R., Wang, M., Zhang, J., Zhang, H., Zhang, B., Wu, X., 2019. Historical population decline and habitat loss in a critically endangered species, the Chinese alligator (Alligator sinensis). Glob. Ecol. Conserv. 20, e00692 https://doi.org/10.1016/j. gecco. 2019.e00692.
Pimm, S.L., Alibhai, S., Bergl, R., Dehgan, A., Giri, C., Jewell, Z., Joppa, L., Kays, R., Loarie, S., 2015. Emerging technologies to conserve biodiversity. Trends Ecol. Evol. 30, 685–696. https://doi.org/10.1016/j. tree. 2015.08.008.
Platt, S.G., Li, F., He, Q., 2022. A Population and Nesting Survey of Reintroduced Chinese Alligators at Dongtan Wetland Park, Shanghai, China. https://programs.wcs.org/library/doi/ctl/view/mid/33065/pubid/DMX4303400000.aspx.
Ren, Jingjing, et al., 2021. Deep texture-aware features for camouflaged object detection. IEEE Trans. Circ. Syst. Video Technol. 33 (3), 1157–1167. https://doi.org/10.48550/arXiv.2102.02996.
Roy, Arunabha M., et al., 2023. WilDect-YOLO: an efficient and robust computer vision-based accurate object localization model for automated endangered wildlife detection. Eco. Inform. 75, 101919.
Tang, Zhe, et al., 2023. Improved Pest-YOLO: real-time pest detection based on efficient channel attention mechanism and transformer encoder. Eco. Inform. 78, 102340.
Technological advances in biodiversity monitoring: applicability, opportunities and challenges - ScienceDirect, 2024. https://www.sciencedirect.com/science/article/abs/pii/S1877343520300592 (accessed January 4, 2024).
Wan, Q.-H., Pan, S.-K., Hu, L., Zhu, Y., Xu, P.-W., Xia, J.-Q., Chen, H., He, G.-Y., He, J., Ni, X.-W., Hou, H.-L., Liao, S.-G., Yang, H.-Q., Chen, Y., Gao, S.-K., Ge, Y.-F., Cao, C.-C., Li, P.-F., Fang, L.-M., Liao, L., Zhang, S., Wang, M.-Z., Dong, W., Fang, S.-G., 2013. Genome analysis and signature discovery for diving and sensory properties of the endangered Chinese alligator. Cell Res. 23, 1091–1105. https://doi.org/10.1038/cr. 2013.104.
Wang, C.-Y., Liao, H.-Y.M., Yeh, I.-H., 2022. Designing network design strategies through gradient path. Analysis. https://doi.org/10.48550/arXiv. 2211. 04800.
Wei, Min, Zhan, Wei, 2024. YOLO_MRC: a fast and lightweight model for real-time detection and individual counting of Tephritidae pests. Eco. Inform. 79, 102445.
Whales from space: Four mysticete species described using new VHR satellite imagery - Cubaynes - 2019 - Marine Mammal Science - Wiley Online Library, 2024. https://onlinelibrary.wiley.com/doi/full/10.1111/mms.12544 (accessed January 4, 2024).
Yang, X., Chai, L., Bist, R.B., Subedi, S., Wu, Z., 2022. A deep learning model for detecting cage-free hens on the litter floor. Animals 12 (15), 1983. https://doi.org/10.3390/ani12151983.
Zhai, Qiang, et al., 2021. Mutual graph learning for camouflaged object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/TIP.2022.3223216.
Zhang, Hao, Cong, Xu, Zhang, Shuaijie, 2023. Inner-iou: more effective intersection over union loss with auxiliary bounding box. arXiv preprint arXiv:2311, 02877. https://doi.org/10.48550/arXiv.2311. 02877.
Zhou, Hui, et al., 2024. Real-time underwater object detection technology for complex underwater environments based on deep learning. Eco. Inform., 102680
Zhu, X., Lyu, S., Wang, X., Zhao, Q., 2021. TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-Captured Scenarios, pp. 2778–2788. https://openaccess.thecvf.com/content/ICCV2021W/VisDrone/html/Zhu_TPH-YOLOv5_Improved_YOLOv5_Based_on_Transformer_Prediction_Head_for_Object_ICCVW_2021_paper.html (accessed January 4, 2024).
Zhu, L., Wang, X., Ke, Z., Zhang, W., Lau, R.W.H., 2023. BiFormer: Vision Transformer With Bi-Level Routing Attention, pp. 10323–10333. https://openaccess.thecvf.com/content/CVPR2023/html/Zhu_BiFormer_Vision_Transformer_With_Bi-Level_Routing_Attention_CVPR_2023_paper.html (accessed January 4, 2024).