Artificial Intelligence Review (2024) 57:165
https://doi.org/10.1007/s10462-024-10788-1

Dynamic YOLO for small underwater object detection

Jie Chen[1] · **Meng Joo Er**[1]

Accepted: 5 May 2024 / Published online: 6 June 2024

Abstract
The practical application of object detection inevitably encounters challenges posed by small objects. In underwater object detection, a crucial method for marine exploration, the presence of small objects in underwater environments significantly hampers the performance of detection. In this paper, a dynamic YOLO detector is proposed as a solution to alleviate this problem. Specifically, a light-weight backbone network is first constructed based on deformable convolution v3, with some specialized designs for small object detection. Secondly, a unified feature fusion framework based on channel-wise, scale-wise, and spatial-aware attention is proposed to fuse feature maps from different scales. This is particularly critical for detecting small objects since it allows us to fully exploit the enhanced capabilities offered by our proposed backbone network. Finally, a simple but effective detection head is designed to handle the conflict between classification and localization by disentangling and aligning the two tasks. Extensive experiments are conducted on benchmark datasets to demonstrate the effectiveness of the proposed model. Without bells and whistles, dynamic YOLO outperforms the recent state-of-the-art methods by a large margin of + 0.8 AP and + 1.8 AP_S on the DUO dataset. Experimental results on Pascal VOC and MS COCO datasets also demonstrate the superiority of the proposed method. At last, ablation studies are conducted on DUO dataset to validate the effectiveness and efficiency of each design in dynamic YOLO. Source code will be available at https://github.com/chenjie04/Dynamic-YOLO.

Keywords Small underwater object detection · Deformable convolution · Dynamic feature fusion · Decoupled head

✉ Meng Joo Er
mjer@dlmu.edu.cn

Jie Chen
chenjie04@dlmu.edu.cn

[1] Institute of Artificial Intelligence and Marine Robotics, College of Marine Electrical Engineering, Dalian Maritime University, 1 Linghai Road, Dalian 116026, Liaoning, China

1 Introduction

Marine exploration has always held great significance for humanity, whether in the exploitation of marine resources or the preservation of ecosystems. With the rapid advancement of marine robotics, vision-based underwater object detection emerges as a cost-effective yet promising approach for marine exploration, garnering considerable attention from the marine research and engineering community (Fayaz et al. 2022; Xu et al. 2023). However, this field faces unique challenges in underwater object detection due to factors like small-sized objects, which hinder the practical implementation of AI-powered techniques (Er et al. 2022).

Most objects of interest in underwater object detection, such as marine organisms (holothurian, echinus, scallops, starfish, etc.), are typically small and tend to aggregate densely (Fig. 1a). The visualization of the Statistics of Detecting Underwater Objects (DUO) dataset (Liu et al. 2021a) depicted in Fig. 1b reveals that the majority of objects exhibit small or medium sizes.[1][2] Specifically, approximately 43.9% constitute small objects while around 53.7% represent medium-sized ones; the number of large objects is almost negligible. This ubiquitous fact poses an inevitable challenge for detecting small underwater objects (Er et al. 2023).

Insufficient visual information hampers the extraction of discriminative features for classification and localization when detecting small objects (Sun et al. 2021). The limited spatial coverage of objects restricts conventional convolutional neural networks with fixed geometric structures in their kernels from effectively extracting features (Dai et al. 2017). Unnecessary contextual information (e.g., sea-grasses in the environment around the objects) can impede representation learning. Deformable Convolution Network (DCN) overcomes this limitation by dynamically aligning sampling locations using predicted offsets, enabling more precise feature extraction (Wang et al. 2022). The incorporation of DCN facilitates adaptive interaction with short- or long-range features, which is particularly advantageous for detecting irregularly shaped small objects in underwater environments. In this paper, we propose a backbone network based on deformable convolution with specialized designs tailored for small object detection.

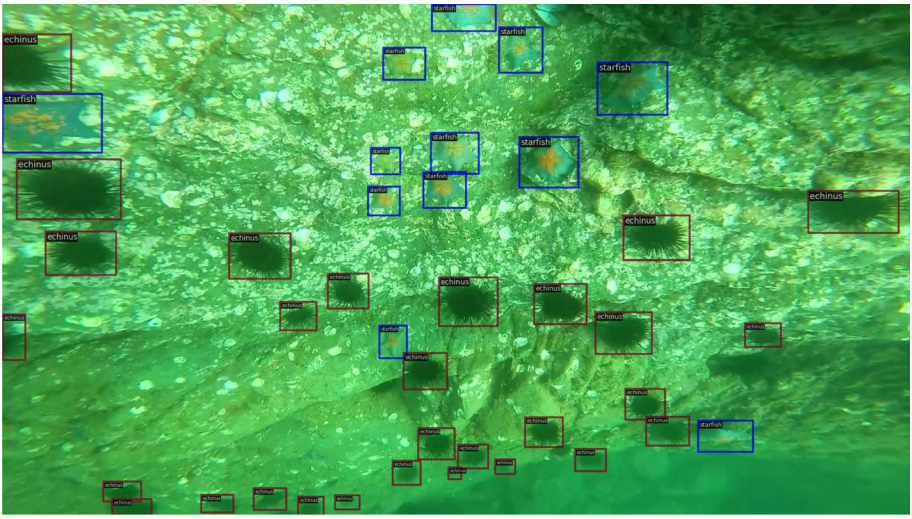
On the contrary, as convolutional neural networks delve deeper into layers, the intricate details of small objects gradually diminish in the feature hierarchy, posing a greater challenge for detection. To address this issue, multi-scale feature fusion strategies have been proposed to aggregate more comprehensive semantic information and localization signals by fusing feature maps from different stages of backbone networks (Chen et al. 2020a). Various feature fusion networks have been extensively explored in previous studies (Lin et al. 2017a; Liu et al. 2018; Tan et al. 2020), yet a unified framework for multi-scale feature fusion has not been established.

We propose three fundamental principles for the design of feature fusion networks:

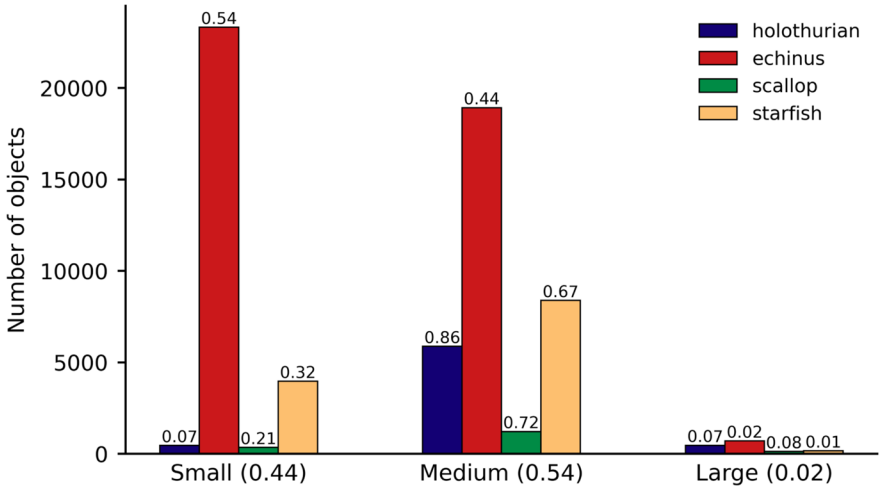
- Firstly, feature fusion should exhibit ***channel-awareness***. Our aim is to dynamically aggregate semantic information and localization signals from feature maps at different

[1] Images have been rescaled to 640×640 pixels, a commonly used scale in recent detectors (Ge et al. 2021).

[2] The definitions for scales of small objects are adopted from the MS COCO dataset (Lin et al. 2014), where an object with an area less than 32^2 pixels is considered small, between 32^2 and 96^2 pixels is considered medium, and larger than 96^2 pixels is considered large.



(a)



(b)

Fig. 1 Marine objects are usually small and tend to congregate in dense distributions. **a** Visualization of detecting small objects on DUO dataset (Liu et al. 2021a). **b** Statistics of DUO dataset on different scales. The numerical values on each bar represent the corresponding percentage within that particular category, e.g., of the sea urchins (the red bars), 54% are small, 44% are medium, and only 2% are large. Overall, a significant proportion of objects fall into the small and medium size, with 44% small, 54% medium, and 2% large

levels. Channel-aware attention enables the activation of distinct semantic information or localization signals at specific spatial locations as desired.

- Secondly, feature fusion should demonstrate ***scale-awareness***. Feature maps from various levels respond to object detection at corresponding scales. Consequently,

unequal contributions are made by feature maps with different scales from the previous module; thus, scale-aware attention facilitates the fusion of feature maps at appropriate magnitudes.
- Thirdly, feature fusion should manifest ***spatial-awareness***. Objects with diverse sizes and forms are distributed across different locations in the image space. Spatial-aware attention assists in aggregating crucial region-based information while suppressing irrelevant context.

Based on the fundamental principles, we explicitly propose a unified feature fusion framework for enhancing small object detection by sequentially applying channel-, scale-, and spatial-aware attention mechanisms to refine features.

Another challenge posed by small object detection is the need for higher localization accuracy, as even slight misalignments can result in false detections. This issue is further complicated by the inherent conflict between classification and localization tasks (Ge et al. 2021). In this study, we propose an extended decoupled head that addresses this problem through the application of a dynamic ReLU function (Chen et al. 2020b) along the channel dimension to disentangle these two tasks using dynamic activation. Subsequently, two deformable convolution layers are employed to enhance task alignment within the detection head. By disentangling and aligning classification and localization processes, our proposed approach mitigates conflicts and achieves superior localization accuracy.

In this paper, we propose a dynamic YOLO detector that effectively detects small underwater objects, leveraging the lightweight backbone network, novel feature fusion framework, and extended decoupled head. Our approach is extensively evaluated on benchmark datasets to demonstrate its effectiveness. Notably, without any additional complexities, our dynamic YOLO outperforms the recent state-of-the-art methods by a significant margin of $+0.8$ AP and $+1.8$ AP_S on the DUO dataset. Furthermore, experimental results on the Pascal VOC and MS COCO datasets validate the superiority of our proposed model consistently. Finally, ablation studies confirm the effectiveness and efficiency of each design choice.

The following are summaries of this paper's significant contributions:

1. A light-weight backbone network specially designed for underwater small target detection based on DCN v3 is proposed.
2. Three fundamental principles for multi-scale feature fusion are identified, and a unified feature fusion framework is proposed.
3. An extended decoupled head is introduced to alleviate the conflict between classification and localization tasks by disentanglement and alignment.
4. With these improvements, a dynamic YOLO detector is developed, achieving state-of-the-art performance on benchmark datasets for underwater object detection.

The remaining sections of this paper are organized as follows: Sect. 2 presents the related works. In Sect. 3, we propose a dynamic YOLO detector and provide detailed explanations on the light-weight backbone, novel feature fusion framework, and decoupled head. Experimental results and discussions on benchmark datasets are presented in Sect. 4. Finally, we summarize our conclusions in Sect. 5.

2 Related work

With the flourishing development of marine robots, vision-based underwater object detection has emerged as a prominent research area (Teng and Zhao 2020). Marine robots equipped with deep learning-powered visual perception systems demonstrate immense potential for ocean exploration. However, unlike land scenarios, underwater object detection poses greater challenges (Er et al. 2023), including image degradation, small objects, poor generalization, and real-time requirements. In this study, our focus lies on detecting small underwater objects.

2.1 Deformable convolution

In recent years, transformer-based detectors have emerged as dominant players in the common object detection community due to their robust representation capabilities and superior performance (Han et al. 2022). However, they still face challenges in detecting small objects primarily because of their limited ability to capture local information. This limitation has prompted researchers to reintroduce convolution modules into the framework (Wu et al. 2021). Among various alternatives, DCN stands out as a more competitive option owing to its adaptive feature extraction across spatial distributions.

The Deformable Convolutional Network (DCN) was initially proposed in Dai et al. (2017) to enhance the transformation modeling capability of conventional CNNs by refining the sampling locations with spatial offsets. This enables easy adaptation of feature extraction to object variations in geometry, making it desirable for visual recognition tasks requiring accurate localization. In DCN v2 (Zhu et al. 2019), a learnable modulation amplitude is introduced at each sampling location, allowing control over the relative influence of samples on recognition tasks. To further strengthen its capability, DCN v3 (Wang et al. 2022) incorporates several optimizations: Firstly, sharing projection weights among convolution neurons reduces parameters and memory complexity. Secondly, a multi-group mechanism aggregates richer information from different feature subspaces at various locations. Lastly, normalization of modulation scalars along sampling locations stabilizes the training process.

DCN, being a robust operator, has gained widespread adoption in computer vision systems for precise feature extraction. Its capability enables detectors to extract features with higher accuracy from small objects and effectively suppress interference caused by the surrounding environment. In this study, we propose a lightweight backbone network based on DCN v3 for detecting small underwater objects.

2.2 Small object detection

Small objects are widely acknowledged as a significant issue in object detection using deep learning, since the network gradually loses detailed information as it goes deeper (Liu et al. 2021b). Many multi-scale feature fusion strategies have been proposed to generate a discriminative representation for small object detection (Er et al. 2023).

In the hierarchical structure of convolutional neural networks (CNNs), there is an enhancement of semantic information, but a loss of localization signals as the network deepens (Liu et al. 2018). To address this predicament, researchers have proposed

multi-scale feature fusion to augment representations with both semantic information and localization signals from diverse scale feature maps, thereby enabling robust object detection across various sizes.

FPN represents the first endeavor towards multi-scale representation by incorporating high-level features into lower ones, facilitating the integration of high-level semantic information (Lin et al. 2017a). However, FPN is limited by a single top-down path flow, resulting in weak localization capabilities for the top feature map. To address this issue, PANet (Liu et al. 2018) introduces an additional bottom-up path to complement FPN and enhance precise localization signals throughout the entire feature hierarchy. NAS-FPN (Ghiasi et al. 2019) employs neural architecture search to obtain an optimal topology for feature fusion but results in an irregular network structure and increased computational cost. In order to establish a simple and efficient multi-scale feature fusion network, BiFPN (Tan et al. 2020) eliminates redundant nodes while extensively incorporating skip-connections to enhance output representation, achieving a better trade-off between accuracy and efficiency.

In recent years, the attention mechanism has demonstrated its superior performance in various tasks (Vaswani et al. 2017). Several attention-based feature fusion networks have been proposed to enhance this capability (Lian et al. 2021; Qin et al. 2020). In Qin et al. (2020), a novel attention module is designed by combining channel and pixel attention, which effectively treats different features and pixels unequally. Similarly, in Lian et al. (2021), an attention feature fusion block is introduced to aggregate relevant context from different network layers for improved detection of small objects in traffic scenarios.

The most relevant work to our paper is the dynamic head (Dai et al. 2021), which aims to integrate attention mechanisms into the detection head. It treats the output of a backbone network as a 3-dimensional tensor, with dimensions defined as level × space × channel. Consequently, scale-, spatial-, and channel-aware attention are sequentially applied. In this paper, we contend that there exists a significant conflict between the classification and regression tasks within the detection head, necessitating separate handling of these two tasks.

2.3 Conflict in detection head

The conflict between classification and localization tasks has long been acknowledged in the field of object detection (Feng et al. 2021; Ge et al. 2021; Song et al. 2020; Wu et al. 2020). For a given object, distinctive characteristics within specific prominent regions may offer valuable information for accurate classification. Conversely, features near the boundary can effectively aid in localizing the bounding box. This misalignment poses a challenge to aligning these two tasks during training and significantly impacts detection performance.

In Song et al. (2020), a task-aware spatial disentanglement (TSD) operator is proposed to decouple the classification and regression tasks from the spatial dimension by generating two disentangled proposals from shared proposals. This simple disentanglement leads to an improvement of approximately 3% AP on the MS COCO dataset for all backbones and models. Wu et al. (2020) revisit the fc-head and conv-head for classification and localization tasks, finding that fc-head is more suitable for classification due to its greater spatial sensitivity, while conv-head is better suited for localization. YOLOX (Ge et al. 2021) proposes a decoupled head with both classification and location branches based on convolution, achieving a better trade-off between performance and efficiency; however, spatial misalignment still exists in this approach. Feng et al. Feng et al. (2021)'s TOOD method (2021) proposes a task-aligned head that achieves a

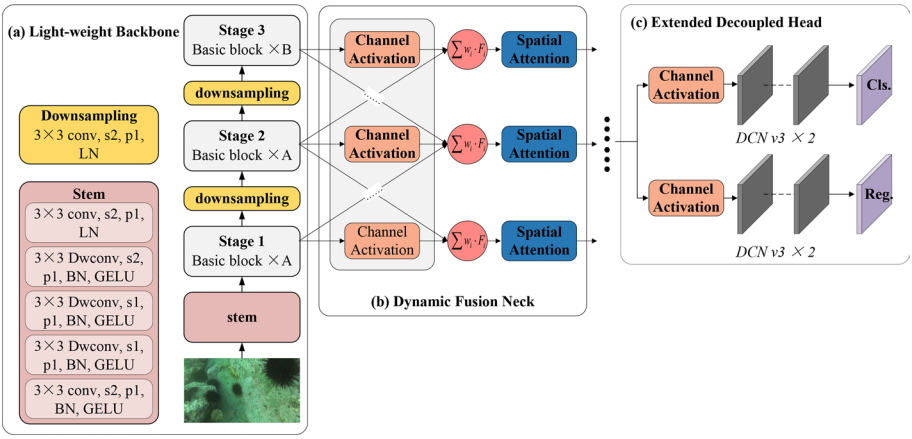


Fig. 2 The architecture of our proposed dynamic YOLO detector

better balance between learning task-interactive and task-specific features through alignment of classification and localization using a task-aligned predictor based on learned features.

In this study, we propose a task-aligned head based on the decoupled head architecture. To disentangle the two tasks, we employ channel-aware attention and introduce deformable convolution to enhance the flexibility of the head for alignment learning. The incorporation of disentanglement and alignment leads to improved detection performance.

3 Our approach

To enhance the performance of small underwater object detection, we have developed a lightweight detector called dynamic YOLO. As depicted in Fig. 2a, we formally present the design of our backbone network based on DCN v3, which is both lightweight and efficient for extracting features from small objects. Additionally, we introduce a unified framework for multi-scale feature fusion that leverages the enhanced capabilities of our backbone network. In Fig. 2b, different scale feature maps are dynamically fused using various attention modules. Furthermore, we investigate the conflict between classification and localization tasks and propose an improved decoupled head as illustrated in Fig. 2c, which proves advantageous for object detection purposes. Based on these enhancements, we propose the dynamic YOLO detector.

3.1 Light-weight backbone network based on DCN v3

The backbone network of our system is constructed based on DCN v3, which will be briefly revisited in this section before proceeding to the construction of the basic block. Subsequently, specialized designs for small object detection are proposed and integrated into the backbone.

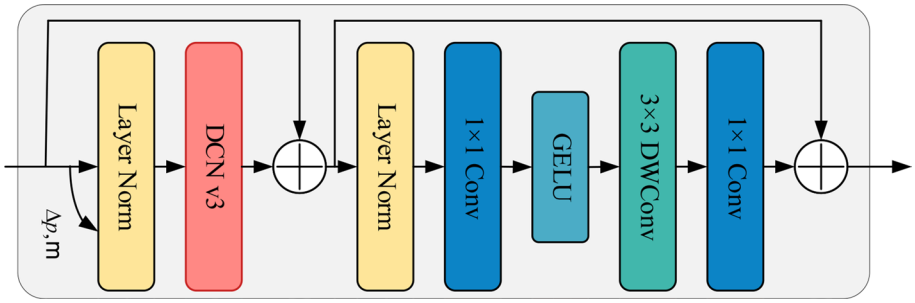


Fig. 3 The detailed structure of basic block of our backbone network

3.1.1 Revisiting DCN v3

DCN v3 is formulated as Equation (1):

$$\mathbf{y}(p_0) = \sum_{g=1}^{G} \sum_{k=1}^{K} \mathbf{w}_g \mathbf{m}_{gk} \mathbf{x}_g (p_0 + p_k + \Delta p_{gk}). \tag{1}$$

Given the input feature map $\mathbf{x} \in \mathbb{R}^{C \times H \times W}$, where C, H, and W represent the channel, height, and width of the feature map respectively. DCN v3 divides it into G groups, denoted as $\mathbf{x}_g \in \mathbb{R}^{C' \times H \times W}$ to indicate the sliced feature map with group dimension $C' = C/G$. Here, p_0 represents the current pixel while K denotes the kernel size. The pre-defined sampling locations are enumerated as p_k. The learning offset for location p_k in the g-th group is represented by Δp_{gk} which adaptively recalibrates sampling locations to achieve precise feature extraction. Modulation scalars $\mathbf{m}_{gk} \in \mathbb{R}$ control the relative influence of each sample.

Additionally, shared projection weights $\mathbf{w}_g \in \mathbb{R}^{C \times C'}$ are employed to map each sample to a hidden feature space.

3.1.2 Basic block

By introducing the multi-group mechanism and weight-sharing strategy, DCN v3 becomes a lightweight yet efficient operator for feature extraction in Wang et al. (2022). We have redesigned the basic block to achieve a higher level of weight lighting. Firstly, we have dropped the unnecessary input and output projections before and after the DCN v3 operator, as they are primarily used for creating query, key, and value vectors in transformers. Secondly, it is crucial to incorporate feed-forward networks for exchanging information between groups due to separated feature modeling in different sub-spaces. We compressed the expansion ratio of the first feed-forward layer to 1 and replaced second feed-forward layer with a 3×3 depthwise separable convolution to enhance spatial dependency incorporation. These design choices significantly enhance the capability of our basic block. Furthermore, our basic block incorporates layer normalization and employs the *GELU* activating function as shown in Fig. 3.

3.1.3 Specialized designs for small object detection

We propose specialized designs for small object detection. Typically, object detectors employ a 4-stage backbone network with an "AABA" stacking pattern, where the 1-*st* stage consists of "A" basic blocks and the 3-*rd* stage consists of "B" basic blocks, with "B" being significantly larger than "A", e.g., the backbone of InternImage-s detector is followed (4, 4, 21, 4) pattern. The stride of this 4-stage backbone is set as $stride = \{4, 8, 16, 32\}$. In our work, we merge the $stride = 4$ stage into the stem layer using depthwise separable convolution while retaining the subsequent stages with strides of $\{8, 16, 32\}$. This modification aligns with (Huang et al. 2022), aiming to enhance computational efficiency. Additionally, we adopt an "AAB" stacking pattern but assign a larger number of layers to "A", e.g., our Dynamic YOLO-s model is in (8, 8, 4) pattern. These specialized designs allow us to stack more layers in the first stage without sacrificing resolution and facilitate better extraction of semantic information crucial for small object detection.

With the utilization of specialized designs and basic blocks mentioned above, we propose the architecture of our backbone network as illustrated in Fig. 2a. To achieve weight-lighting, depthwise separable convolutions are extensively employed in the stem layer, while down-sampling is accomplished through 3×3 convolution with a stride of 2 and layer normalization.

3.2 Dynamic neck for multi-scale feature fusion

To fully exploit the potential of the backbone network, we propose a dynamic neck for multi-scale feature fusion, aiming to aggregate valuable features at different scales and enhance each feature representation with semantic information or localization signals from higher- or lower-level feature maps. Specifically, given a list of multi-scale feature maps obtained from the output of the backbone $F_{in} = \left\{F_i\right\}_{i=1}^{L}$ (where L denotes the number of feature maps), our approach seeks to improve feature representations through a transformation: $F_{out} = f(F_{in})$.

Based on the discussion of fundamental principles for multi-scale feature fusion, we propose a dynamic neck network, as illustrated in Fig. 2b, where channel-aware attention, scale-aware attention, and spatial-aware attention are sequentially applied. Specifically, distinct channel attention modules are initially employed in each connection to activate diverse semantic information or localization signals prior to feature fusion. Secondly, only adjacent feature maps are fused based on our intuition that long-range feature maps may introduce potential conflicts. By repeating the dynamic neck block $N - 1$ times, all information from the initial N feature maps can be accessed without concerns about information loss. Lastly, the spatial attention module is implemented using a basic block based on DCN v3 and applied once after feature fusion to reduce redundant information.

3.2.1 Channel-aware attention

Feature fusion aims to enhance the representation by incorporating semantic information and localization signals from higher or lower feature maps. To achieve desired feature fusion, it is essential to selectively activate different channels of feature maps. For instance, when fusing two adjacent feature maps, activating the semantic information

from the higher map while utilizing the localization signals from the lower map would be optimal. In this study, we employ dynamic ReLU (DyReLU-B) function (Chen et al. 2020b) to direct distinct feature channels towards preferred activations.

The DyReLU-B activation function is a parametric approach that dynamically adjusts the channel-wise activation using control signals, which encode the global context of the feature map through a hyperfunction. Initially, the global context is aggregated by adaptive average pooling. Subsequently, an explicit "Squeeze-and-Excitation" operation (Hu et al. 2018) is performed to model inter-dependencies between channels. Based on these inter-dependencies, control signals $\left[a_1, b_1, a_2, b_2 \mid \in \mathbb{R}\right]$ are generated to adaptively recalibrate the channel-wise feature activation. The formulation of DyReLU-B is as follows:

$$a_1, b_1, a_2, b_2 = F_{ex}\left(F_{sq}\left(\frac{1}{C \times H \times W} \sum_{C,H,W} F_i\right)\right) \tag{2}$$

$$F_i' = \max(a_1 \cdot F_i + b_1, a_2 \cdot F_i + b_2) \tag{3}$$

where F_i, F_i' denote the input and output feature maps. The squeeze function $F_{sq}(\cdot)$ and excitation function $F_{ex}(\cdot)$ are approximated by 1×1 convolution, respectively. A shifted hard-sigmoid function is employed in the excitation function $F_{ex}(\cdot)$ to normalize the output within the range of $[-1, 1]$. By incorporating the parametric DyReLU-B function, our channel-aware attention module gains the capability to selectively activate semantic information or localization signals based on specific requirements.

3.2.2 Scale-aware attention

Scale-aware attention aims to dynamically integrate features from different scales based on their semantic roles' significance. Our rationale is straightforward: feature maps at various levels exhibit varying responses to object detection at corresponding scales, thus contributing unequally to the current representation. Consequently, adaptive feature fusion becomes imperative.

In the scale-aware attention module, global context of each feature map is initially aggregated through adaptive average pooling. Subsequently, the hard-sigmoid function based on global context activates the scale-ware fusion scores. Finally, multi-scale features are weightedly summed up to achieve scale-ware feature fusion with reference to the fusion scores. The formulation for scale-aware fusion can be expressed as follows:

$$F_l = \frac{1}{L'} \sum_{i}^{L'} \sigma\left(f\left(\frac{1}{C \times H \times W} \sum_{C,H,W} F_i\right)\right) \cdot F_i, \tag{4}$$

$$\sigma(x) = \min\left(\max\left(\frac{x+3}{6}, 0\right), 1\right) \tag{5}$$

where F_i is the ith feature map from the previous fusion block. $f(\cdot)$ presents the linear mapping implemented by a 1×1 convolution, and $\sigma(x)$ denotes the hard-sigmoid function. Only feature maps at adjacent levels are fused in the scale-aware attention module. L' denotes the number of feature maps in the current fusion process, which can be varied.

3.2.3 Spatial-aware attention

Spatial-aware attention enhances representation capabilities by selectively focusing on crucial regions of the feature map and suppressing unnecessary context, enabling precise feature extraction for object detection (Guo et al. 2022). Various approaches exist to implement spatial-aware attention (Fu et al. 2019), with DCN (Dai et al. 2017) being one of the most prominent methods. By learning a 2D offset for each neuron in the convolution kernel, DCN enables interaction with specific spatial regions, thereby achieving spatial attention (Dai et al. 2017).

In this study, we have implemented a spatial-aware attention module based on the fundamental block proposed in the backbone network. By incorporating the extended DCN v3, our spatial-aware attention module demonstrates enhanced effectiveness and efficiency in strengthening representation capability. Following scale-aware fusion, spatial-aware attention is applied to the feature map to acquire a more robust representation.

3.3 Extended decouple head for task alignment

Multi-scale feature fusion has established a robust foundation for object detection. However, the conflict between classification and localization within the detection head remains a bottleneck that hampers the improvement of detection performance, particularly in small underwater object detection (Ge et al. 2021). Consequently, the adoption of decoupled heads, which disentangle classification and localization through two separate branch networks, is frequently employed in both one-stage and multi-stage detectors (Song et al. 2020; Wu et al. 2020). Nevertheless, significant spatial misalignment still persists within these decoupled heads. This misalignment poses an unfavorable circumstance for object detection. In this study, we propose an extension to the decoupled head for task alignment in a learning-based manner.

As illustrated in Fig. 2c, we introduce the DyReLU-B function as a means to disentangle features for classification and localization tasks. Consistent with the discussion presented in the section on multi-scale feature fusion, the feature maps generated by the neck module encode both semantic information and localization signals within each feature vector at a spatial point. Consequently, it becomes crucial to disentangle these features along the channel dimension.

To address the misalignment, we incorporate two deformable convolution layers into separate branches that effectively aggregate features from relevant spatial locations to cater to different tasks. For instance, while the classification branch focuses on aggregating semantic information within salient areas, the localization branch gathers localization signals primarily around object boundaries.

The aforementioned network architecture design has endowed the extended decoupled head with the capability to align classification and localization tasks. However, a learning mechanism is still required to guide the detection head towards achieving alignment. In this paper, we employ quality focal loss (QFL) (Li et al. 2020) as the classification loss function to supervise the learning process. Unlike standard focal loss, QFL incorporates softening of the usual one-hot category label by considering localization quality, which is determined by the IoU scores between predicted bounding boxes and their corresponding ground truth annotations. Specifically, $y = 0$ represents the classification label for negative samples with a quality score of 0. Meanwhile, $0 \leq y \leq 1$

denotes positive sample labels along with their corresponding IoU scores. By utilizing soft labels, QFL can be formulated as follows:

$$\boldsymbol{QFL}(p) = -|y-p|^{\beta}((1-y)log(1-p) + ylog(p)), \tag{6}$$

where p denotes the prediction output, β is the scaling factor, and $|y-p|$ measures the distance between the prediction p and its ground truth, which is used to down-weight the contribution of easy examples. By adopting QFL supervision, we can ensure that spatial points with higher classification scores also possess higher Intersection over Union (IoU) values. This property guarantees the successful alignment of classification and localization tasks in the extended decoupled head.

We employ *GIoU* (Rezatofighi et al. 2019) as our localization loss, and the total loss is set as:

$$L = \lambda_1 L_{cls} + \lambda_2 L_{reg} \tag{7}$$

where L_{cls} denotes classification loss and L_{reg} denotes localization loss. $\lambda_1 = 1.0$ and $\lambda_2 = 2.0$ are the weights of two losses by default. The alignment between the two tasks can greatly enhance detection performance.

3.4 Dynamic YOLO

This paper introduces a dynamic YOLO detector for small underwater object detection, featuring a light-weight backbone, dynamic neck, and extended decoupled head. To enhance the multi-scale representation crucial for detecting objects of different sizes, especially small ones, we incorporate multiple repetitions of the fusion block in the dynamic neck. Additionally, instead of utilizing separated detection heads on different level features as suggested by Redmon and Farhadi (2018), we choose to share the detection head along different levels to improve model efficiency.

4 Experiment

To evaluate the effectiveness of the proposed dynamic YOLO, we conducted extensive experiments on the DUO dataset (Liu et al. 2021a), which contains about 6671 images in the training set and 1111 images in the testing set, respectively. The DUO dataset was collected from Underwater Robot Professional Contest,[3] which is developed for robot picking based on underwater images. It contains four categories of underwater targets, namely holothurian, echinus, scallops, and starfish. The brief statistic of DUO is shown in Fig. 1b, there are 63,998 objects, with 44% small, 54% medium, and 2% large. We also evaluate our model on the Pascal VOC and MS COCO datasets, the most well-accepted benchmark datasets for common object detection. At last, to validate the effectiveness and efficiency of each design in the proposed model, ablation studies are performed on the DUO dataset.

[3] Underwater Robot Professional Contest: http://en.cnurpc.org.

Table 1 Comparison of dynamic YOLO with state-of-the-art methods on the number of parameters, FLOPS, and accuracy on the DUO dataset

Method	Param.	FLOPs	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Multi-stage detectors								
Faster R-CNN (Ren et al. 2015)	41.14	63.26	54.8	75.9	63.1	53.0	56.2	53.8
Cascade R-CNN (Cai and Vasconcelos 2018)	68.94	77.54	55.6	75.5	63.8	44.9	57.4	54.4
RepPoints (Yang et al. 2019)	36.60	35.60	56.0	80.2	63.1	40.8	58.5	53.7
One-stage detectors								
RetinaNet (Lin et al. 2017b)	36.17	39.68	49.3	70.3	55.4	36.5	51.9	47.6
FCOS (Tian et al. 2019)	31.84	38.84	53.0	77.1	59.9	39.7	55.6	50.5
ATSS (Zhang et al. 2020)	31.89	38.84	58.2	80.1	66.5	43.9	60.6	55.9
GFL (Li et al. 2020)	32.04	39.63	58.6	79.3	66.7	46.5	61.6	55.6
Real-time detectors								
YOLOX (Ge et al. 2021)	8.94	13.34	61.2	82.8	69.9	46.6	63.1	59.7
RTMDet (Lyu et al. 2022)	8.86	14.80	67.4	85.9	75.3	*53.3*	68.5	*67.0*
YOLOv6 (Li et al. 2022)	17.19	21.88	67.2	86.2	74.9	46.7	68.6	66.5
YOLOv7 (Wang et al. 2023)	**6.23**	**6.89**	62.3	83.5	70.5	46.6	63.7	61.7
YOLOv8 (Jocher et al. 2023)	11.14	14.27	*67.8*	*86.1*	*75.6*	48.5	*69.3*	66.8
Dynamic YOLO	*8.21*	*12.51*	**68.6**	**86.7**	**76.3**	**55.1**	**69.8**	**68.1**

The best results are in bold, and the second-best results are in italics

4.1 Implementation details

We implement our dynamic YOLO model based on the MMdetection 3.0 framework (Chen et al. 2019), with Python 3.8.18, PyTorch 2.0.0, and CUDA Toolkit 11.8. The default stacking pattern of the backbone network is set to $\{8, 8, 4\}$, and the number of groups for DCN v3 in each stage is set to $\{4, 8, 16\}$. The fusion block in the dynamic neck is repeated 4 times by default. We adopt AdamW as our optimizer with a 0.001 initial learning rate, which is scheduled by a Flat-Cosine strategy. The batch size is set to 8. The stochastic depth and layer scale techniques are also employed to increase the dynamic in training. Strong data augmentations, including cached Mosaic and MixUp (Lyu et al. 2022), are applied for a robust generalization but are switched off in the last 20 epochs to fine-tune the model on a more realistic data distribution. All our models are trained from scratch for 300 epochs on a compute node with 2 RTX A5000 GPUs, each with 24GB of memory.

4.2 Comparison with the state-of-the-arts on DUO dataset

The most straightforward method to demonstrate the effectiveness and efficiency of the proposed model is to compare it with the state-of-the-art methods on the benchmark dataset. Several representative one-stage (Li et al. 2020; Lin et al. 2017b; Tian et al. 2019; Zhang et al. 2020) and multi-stage detectors (Cai and Vasconcelos 2018; Ren et al. 2015; Yang et al. 2019) are adopted for comparison. Specifically, most experimental results are from the DUO benchmark (Liu et al. 2021a), where detectors are trained on 512×512 resolutions. The state-of-the-art real-time object detectors, including YOLOX (Ge et al.

2021), RTMDet (Lyu et al. 2022), YOLOv6 (Li et al. 2022), YOLOv7 (Wang et al. 2023), YOLOv8 (Jocher et al. 2023), and our dynamic YOLO, are trained from scratch with 640×640 resolutions.

The experimental results of the comparison on the DUO dataset are shown in Table 1. As we can see, the best method among previous state-of-the-art detectors in the benchmark dataset is GFL (Li et al. 2020), which obtains 58.6 AP and 46.5 AP_S. However, the most remarkable result on small object detection is obtained by Faster R-CNN (Ren et al. 2015), achieving 53.0 AP_S, which is a benefit of the fine-tuning process in the two-stage detectors.

The real-time detectors in Table 1 were brought from the community of common object detection. They launch a new era of light-weight detectors for underwater object detection. YOLOX employs various sophisticated detection techniques, such as a decoupled head and the leading label assignment approach SimOTA, which impressively outperforms GFL with + 2.6 AP improvement but only 27.9% parameters. YOLOv6 heavily absorbs recent ideas in network design, training strategies, testing techniques, quantization, and optimization methods, achieving significant performance improvements while also doubling the model complexity. YOLOv7 presented here is a tiny version; it outperforms YOLOX but with only 69.7% parameters, demonstrating its superiority. With continuous evolution, RTMDet pushes the boundary of performance by a large margin again, achieving cutting-edge performance with 67.4 AP, especially the 53.3 AP_S on small object detection. Without bells and whistles, RTMDet beats the previous state-of-the-art detectors in all aspects. YOLOv8 is the latest real-time detector, integrating many advanced technologies. It achieves excellent performance, 67.8 AP, but the performance of small object detection is slightly inferior to RTMDet.

Our method surpasses previous methods by a significant margin with fewer parameters. For a fair comparison, all hyperparameters of dynamic YOLO keep the same with RTMDet (Lyu et al. 2022). The proposed dynamic YOLO model delivers a new state-of-the-art performance of 68.6 AP, with an impressive + 0.8 AP improvement over YOLOv8. Meanwhile, it significantly outperforms RTMDet with 55.1 AP_S, an increase of + 1.8 AP_S, for small object detection. As shown in Table 1, the best results are bolded, and the second-best results are highlighted in italics. On the other hand, dynamic YOLO only has 73.7% parameters of YOLOv8, resulting in a much better trade-off between parameter and accuracy. The comprehensive experimental results fully confirm the effectiveness and efficiency of our dynamic YOLO model, demonstrating its superior performance in detecting small underwater objects.

We do not show the comparison between our model and transformer-based detectors (Han et al. 2022) in this paper because they cannot even converge on such small-scale datasets without being pre-trained on large-scale datasets.

4.3 Visualization of detection results on DUO dataset

For an intuitive understanding, we visualize several representative samples of underwater object detection in Fig. 4, including the common scenarios in underwater environments, such as (a) small objects, (b) low contrast, (c) occlusion, and (d) clustering.

As shown in Fig. 4a, many small objects (mainly echinus) scatter on the sea bed, making the detection extremely challenging. We can see some objects on the right-up part of the image, but YOLOX misses them all, while RTMDet and dynamic YOLO catch some (as indicated by the yellow arrows). However, in Fig. 4b, the situation

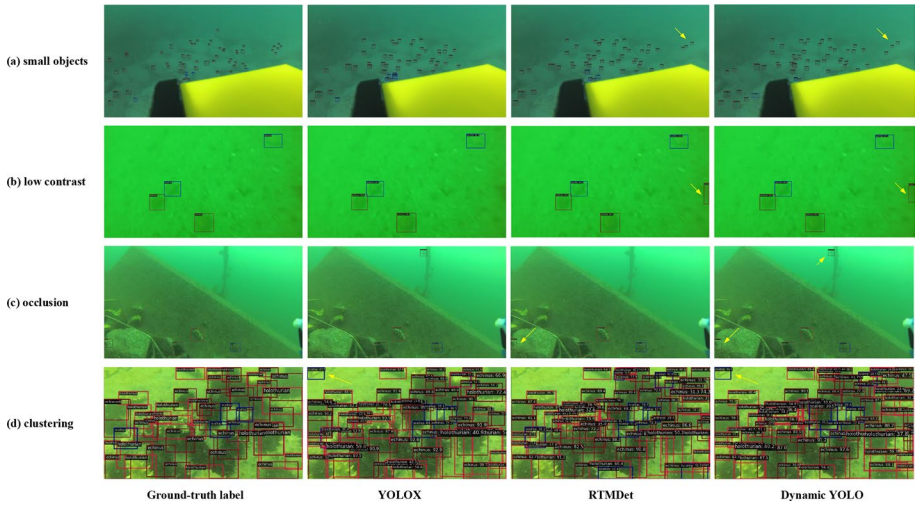


Fig. 4 Example images of underwater object detection in common scenarios of the DUO dataset: **a** small objects; **b** low contrast; **c** occlusion; and **d** clustering

Table 2 Experimental results on pascal VOC dataset

Method	Param	FLOPs	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
YOLOX	8.95	13.34	50.6	78.0	55.7	22.8	38.8	54.7
YOLOv6	17.20	21.90	53.7	75.0	58.4	14.6	34.8	61.2
YOLOv7	**6.07**	**6.63**	44.4	68.4	48.4	11.2	29.2	50.9
YOLOv8	11.14	14.29	54.3	74.3	59.2	14.8	35.5	61.9
RTMDet	8.86	14.77	*60.7*	*83.1*	*66.7*	**29.0**	*42.5*	*66.9*
Dynamic YOLO	*8.27*	*12.56*	**61.7**	**83.3**	**67.7**	*26.0*	**43.2**	**68.4**

reversed. YOLOX performs relatively well in the contrast scenario, but RTMDet and dynamic YOLO pose a false detection on the right edge of the image. This situation is due to the perplexing shadow, causing the RTMDet and dynamic YOLO to malfunction.

In Fig. 4c, there is an echinus on the left-down corner, which is impeded by rock, and only a tiny part exhibits. Both RTMDet and dynamic YOLO can detect this echinus successfully. However, on the other hand, dynamic YOLO is deceived by a knot on the discarded rope, as YOLOX was. The last scenario is (d) clustering, where piles of marine objects gather together. The behaviors of all detectors perform almost consistently; YOLOX and dynamic YOLO still raise a false detection in the left-up corner.

From the visualization of detection results, a comprehensive understanding is obtained. There are some trivial false positives in Fig. 4b and d. Actually, the MS COCO dataset also has some ambiguous objects marked by "ignore". They can be or are not the target objects. In our case, we consider these false positives to be negligible, as evidenced by the better performance on quantitative, as shown in Table 1. Dynamic YOLO is more sensitive than RTMDet.

4.4 Experimental results on pascal VOC and MS COCO datasets

We also evaluate our dynamic YOLO model on Pascal VOC (Everingham et al. 2010) and MS COCO (Lin et al. 2014) datasets. In the experiments with Pascal VOC, all detectors were trained on the 2007 and 2012 training sets and tested on the 2007 testing set. As shown in Table 2, most real-time detectors perform consistently with DUO dataset, but RTMDet outperforms YOLOX by a large margin, achieving 60.7 AP, that is, a + 6.4 AP increase over YOLOv8. Especially, the best performance of 29.0 AP_S is achieved by RTMDet for small object detection.

Dynamic YOLO model achieves a competitive result with fewer parameters, obtaining a new state-of-the-art performance of 61.7 AP. Experimental results demonstrate confidence in the superiority of our approach. However, the performance of small object detection in dynamic YOLO has dropped by about − 3.0 AP_S degradation. We conduct a statistic on the Pascal VOC training set and find that it has 7.4% small objects, 26.5% medium, and 66.1% large objects. The training process of dynamic YOLO has been dominated by large objects, leading to inadequate learning of small objects and degradation in performance. This is probably due to the insufficient learning of offset in deformable convolution. Based on the above observation, we conclude that deformable convolution is superior in detecting small objects but is sensitive to the class-imbalance in the training process.

Experimental results on the MS COCO dataset are shown in Table 3; the best results are highlighted in bold. COCO is the standard benchmark dataset for common object detection. We train our model on the COCO Train 2017 set and evaluate it on the Val 2017 set. The experimental results of other models for comparison are adopted from MMYOLO Contributors (2022). As shown in Table 3, our model achieves the best results compared with the previous real-time detectors (Glenn et al. 2022; Ge et al. 2021; Li et al. 2022; Xu et al. 2022; Lyu et al. 2022; Jocher et al. 2023) on equal conditions, achieving 45.5 AP. However, it is slightly behind YOLOv8 in small object detection, obtaining 25.4 AP_S. Dynamic YOLO achieves a better parameter-accuracy trade-off, demonstrating its superiority.

Table 3 Experimental results on MS COCO Val 2017 dataset

Method	Param.	FLOPs	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
YOLOv5 (Glenn et al. 2022)	**7.2**	**8.3**	37.7	57.1	41.0	21.7	42.5	48.8
YOLOX	9.0	13.4	40.7	59.6	44.3	23.9	45.2	53.8
YOLOv6 (Li et al. 2022)	17.2	22.1	43.7	60.8	47.0	23.6	48.7	59.8
PPYOLOE (Xu et al. 2022)	*7.9*	*8.7*	43.1	60.5	46.6	23.2	46.4	56.9
YOLOv8 (Jocher et al. 2023)	11.2	14.36	44.2	61.2	47.9	**25.6**	*49.0*	59.7
RTMDet	9.0	14.8	*44.5*	*61.9*	*48.1*	24.9	48.5	*62.5*
Dynamic YOLO	8.3	12.6	**45.5**	**62.6**	**49.5**	*25.4*	**50.2**	**64.1**

Table 4 Ablation studies on the effectiveness of each design in dynamic YOLO on the DUO dataset

Method	Param.	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
RTMDet	8.86	67.4	85.9	75.3	53.3	68.5	67.0
Backbone	11.04	67.7	86.4	75.3	**55.4**	69.2	66.8
Backbone + neck	8.47	68.3	86.6	75.5	55.0	**70.0**	67.1
Backbone + neck + head	**8.21**	**68.6**	**86.7**	**76.3**	55.1	69.8	**68.1**

By replacing the backbone, neck, and head with proposed counterparts, the performance of underwater object detection grows gradually while the model complexity decreases

Table 5 Ablation studies on the effectiveness of redesigned DCNv3 module on the DUO dataset

Method	Param	FLOPs	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Original DCNv3	21.81	29.44	**69.5**	**87.5**	**77.7**	**58.5**	**71.0**	**68.4**
Replaced FNN	11.62	15.95	68.8	86.9	76.5	56.2	70.4	67.6
Dropped projections	**8.21**	**12.51**	68.6	86.7	76.3	55.1	69.8	68.1

4.5 Ablation study

4.5.1 Ablation study on basic designs

Ablation studies are extensively conducted on the DUO dataset to validate the efficiency and effectiveness of each design in our dynamic YOLO model. We set RTMDet as a baseline and gradually replaced the backbone, neck, and detection head to evaluate their performances. As shown in Table 4, by replacing the components with our proposed counterparts, the performances of underwater object detection grow gradually while the model complexity decreases.

First, we replace the backbone network, which results in a significant improvement on most evaluation metrics. Especially in small object detection, we gain a $+2.1\,AP_S$ boost on performance. It fully proves the superiority of deformable convolution for small object detection. However, we also note the growth in model parameters. The proposed light-weight backbone network is slightly heavier than the CSPNeXt network used in RTMDet.

To fully activate the potential of the proposed backbone network, we replaced the neck of RTMDet with our dynamic neck. In the third row of Table 4, as demonstrated, the performances of the detector improve consistently, except for a slight drop in small object detection. Notably, the model parameters have decreased to $8.47\,M$, much lower than RTMDet. This experimental result demonstrates that feature fusion based on attention mechanisms is more competitive than the conventional FPN framework.

At last, the detection head is replaced by our extended decoupled head, which has the capability of task alignment. By alleviating the conflict between classification and localization, the performance of underwater object detection is continuously improved. At the same time, the model parameters decrease again, as shown in the last row of Table 4. It comprehensively outperforms the competitive RTMDet detector and achieves state-of-the-art performances. The usefulness and efficiency of each design in our dynamic YOLO model are clearly demonstrated through ablation studies.

Table 6 Ablation studies on the effectiveness of each attention mechanism in the neck block on PASCAL VOC dataset

Channel	Scale	Spatial	Param	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
✗	✗	✗	8.79	60.2	82.2	66.4	24.3	40.6	67.1
✓	✗	✗	9.38	60.4	82.5	66.4	23.2	41.1	67.1
✗	✓	✗	8.79	60.3	82.7	66.4	23.4	40.6	67.2
✗	✗	✓	7.62	61.1	83.1	67.2	22.8	42.1	67.8
✓	✓	✗	9.38	60.4	82.6	67.0	24.2	42.0	66.9
✓	✗	✓	8.21	61.2	82.4	66.7	23.8	43.0	68.1
✗	✓	✓	**7.62**	61.2	83.0	67.4	25.9	42.1	68.1
✓	✓	✓	8.21	**61.7**	**83.3**	**67.7**	**26.0**	**43.2**	**68.4**

4.5.2 Ablation study on backbone network

We rebuilt the DCNv3 module to obtain a lightweight backbone network. Table 5 illustrates that, despite achieving a greater performance of 69.5 AP with the same architecture, the model with a backbone network based on the original DCNv3 module has a computation complexity that is almost 2.5 times higher. By compressing the expansion ratio to 1 and substituting a depthwise separable convolution for the second fully connected layer, we were able to almost halve the model complexity, just sacrificing an acceptable level of performance deterioration.

We also observed that the DCNv3 module does not require the input projection layer, which is used to produce query, key, and value vectors for transformers. Furthermore, the feed-forward layer that follows allows the output project layer, which transfers information across group convolutions, to be disregarded. As shown in Table 5, by dropping the projection layer, we can further reduce the model complexity with little impact on the model performance, about − 0.2 AP degradation. Though the performance deterioration for small object detection is more severe, at roughly − 1.1 AP_S, it is still acceptable given the difficulty in small object detection. The experimental results demonstrate that our redesigned DCNv3 module achieves a superior trade-off between accuracy and efficiency.

4.5.3 Ablation study on attention mechanisms

As the key component in dynamic YOLO, the effectiveness of each attention mechanism in the neck block is validated on the PASCAL VOC dataset. A simple linear fusion was employed as the baseline, where adjacent feature maps in the pyramid were combined linearly and then processed by a convolutional module. The ablation experimental result is presented in Table 6, with "Channel", "Scale", and "Spatial" denoting channel-aware, scale-aware, and spatial-aware attention mechanisms, respectively.

The interactions between different attention mechanisms are intricate. Initially, we integrate each attention into the baseline fusion structure individually. As depicted in Table 6, it can be observed that channel-aware attention or scale-aware attention only yield slight improvements. In fact, there is even a decline in performance for small object detection. Conversely, spatial attention leads to significant enhancements across most evaluation metrics, except for small object detection, where it achieves an approximately + 0.9 AP improvement.

Table 7 Ablation studies on the effectiveness of redesigned DCNv3 module on the DUO dataset

Dy &DCN	QFL	Param	FLOPs	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
✗	✗	8.47	15.01	63.7	84.2	71.7	47.9	65.0	63.0
✗	✓	8.47	15.01	68.3	86.6	75.5	55.0	**70.0**	67.1
✓	✗	8.21	12.51	64.0	84.4	71.8	46.9	64.8	63.7
✓	✓	**8.21**	**12.51**	**68.6**	**86.7**	**76.3**	**55.1**	69.8	**68.1**

When collaborating with other attention mechanisms, scale-aware attention significantly enhances the performance of small object detection, particularly when integrated with spatial-aware attention, resulting in a notable improvement of $+1.6\,AP_S$. This validates that feature maps at different levels exhibit distinct responses to object activations across various scales, emphasizing the necessity for scale-aware feature fusion. For channel-aware attention, it is crucial to enhance the robustness of feature fusion, although neural networks can learn to fuse the localization signals and semantic information from different feature maps implicitly.

Finally, channel-aware, scale-aware, and spatial-aware attentions are applied to feature maps sequentially for feature fusion. The fully dynamic fusion module significantly improves the baseline by $+1.5$ AP and $+1.7\,AP_S$. The experimental results demonstrate that these attention mechanisms work in a coherent manner.

4.5.4 Ablation study on detection head

To improve the performance of small object detection, we proposed to disentangle and align the features for classification and localization via dynamic activation and deformable convolution (Dy &DCN). QFL is employed to guide the learning process. The results of the ablation study are shown in Table 7. As depicted, even without the supervision of QFL, the DCN-based detection head (row 3) is superior to the CNN-based one (row 1), at 64.0 AP vs. 63.7 AP. But unexpectedly, it is inferior in small object detection, suffering a $-1.0\,AP_S$ degradation. This may be due to the difficulty of learning the offsets in deformable convolutions. Under the supervision of QFL, the performances have greatly improved (row 2 and row 4). The DCN-based detection head outperforms the CNN-based detection head on almost every metric while slightly compressing the computational complexity. Experimental results demonstrated the effectiveness of our detection head.

4.6 Visualization of feature maps

To achieve a more comprehensive understanding, we illustrate the feature maps of dynamic YOLO in several common scenarios of the DUO dataset. Feature maps are extracted from the backbone network's first stage, the dynamic neck output, and the classification and regression branches in the extended decoupled head, respectively.

As shown in the first column at the top of Fig. 5, the proposed model focuses on objects well (indicated by the red region), which means more semantic information emerges via deeper convolution layers with deformable receptive fields in the first stage. This is beneficial for small object detection because it can gain more semantic information without decreasing resolution compared to other competitive models. After feature fusions, our

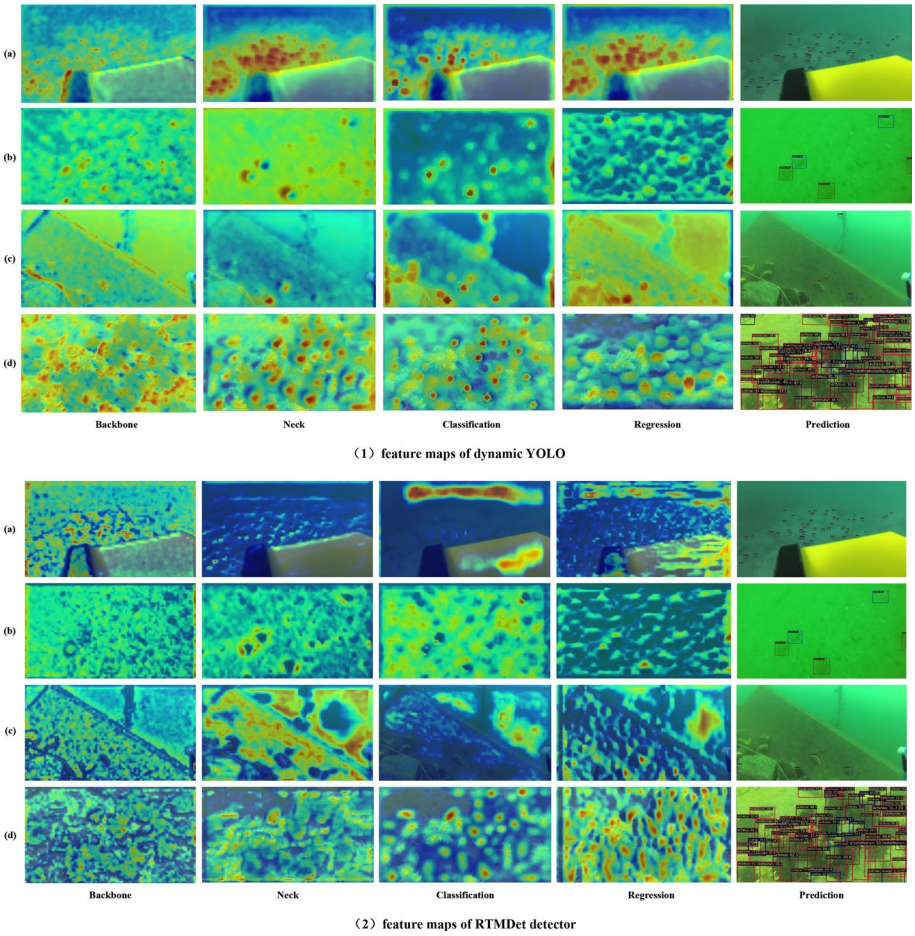


Fig. 5 Visualization of feature maps of dynamic YOLO (top) and RTMDet detector (bottom) in common scenarios of the DUO dataset: **a** small objects, **b** low contrast, **c** occlusion, and **d** clustering

dynamic YOLO model gradually focuses on the regions of objects. As shown in the second column, objects are clearly distinguished from their surrounding environments, demonstrating the effectiveness of our dynamic neck for multi-scale feature fusion. We also visualize the feature map output from the classification and regression branches of the extended decoupled head in the third and fourth columns. As expected, the feature maps of the two tasks are relatively well aligned since deformable convolutions are employed to adaptively aggregate desired features from different locations. The visualization of feature maps confirms the aforementioned discussion about the behaviors of dynamic YOLO, demonstrating the significant superiority of the proposed model.

For comparison, the visualization of RTMDet's feature maps is attached at the bottom of Fig. 5. The feature map from the first stage of CSPNeXt in RTMDet presents more lower-detail information, such as the fine-grained structures in the images. After the feature fusions, the focus of the detector is still scattered. The alignment of classification and localization tasks presented in the feature map also does not perform well enough compared

with the dynamic YOLO model. This fact serves as more evidence of the superiority of our approach.

5 Conclusion

This paper thoroughly investigates the problem of small underwater object detection. We propose a light-weight dynamic YOLO detector as a solution for this issue. Specifically, a backbone network is designed based on deformable convolution v3, which is superior for small object detection due to its capability for adaptive feature extraction. To better exploit the potential of the backbone, a dynamic feature fusion network is proposed as the neck to fuse multi-scale representation. The conflict between the classification and localization tasks in the detection head is also explored in this paper, and we propose an extended decoupled head to alleviate this problem through task alignment. With the aforementioned improvements, dynamic YOLO surpasses state-of-the-art methods by a large margin of $+0.8$ AP and $+1.8$ AP_S on performance with fewer parameters on DUO dataset. Experimental results on Pascal VOC and MS COCO datasets also demonstrate the superiority of the proposed model. At last, the effectiveness and efficiency of each design are evaluated. We anticipate that our research will shed light on small underwater object detection.

Acknowledgements The authors would like to acknowledge the support of Fundamental Research Funds for the Central Universities under Grant 3132019344 and the Leading Scholar Grant, Dalian Maritime University under Grant 00253007.

Author contributions Jie Chen wrote the main manuscript text and Meng Joo Er reviewed the manuscript.

Declarations

Competing interests The authors declare no competing interests.

References

Cai Z, Vasconcelos N (2018) Cascade R-CNN: delving into high quality object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6154–6162

Chen K, Wang J, Pang J et al (2019) MMDetection: open MMLAB detection toolbox and benchmark. arXiv preprint. arXiv:1906.07155

Chen G, Wang H, Chen K et al (2020a) A survey of the four pillars for small object detection: multiscale representation, contextual information, super-resolution, and region proposal. IEEE Trans Syst Man Cybern Syst 52(2):936–953

Chen Y, Dai X, Liu M et al (2020b) Dynamic ReLu. In: Proceedings of the European conference on computer vision. Springer, Cham, pp 351–367

Dai J, Qi H, Xiong Y et al (2017) Deformable convolutional networks. In: Proceedings of the IEEE international conference on computer vision, pp 764–773

Dai X, Chen Y, Xiao B et al (2021) Dynamic head: unifying object detection heads with attentions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 7373–7382
Er MJ, Chen J, Zhang Y (2022) Marine robotics 4.0: present and future of real-time detection techniques for underwater objects. In: Industry 4.0—perspectives and applications. IntechOpen, London. https://doi.org/10.5772/intechopen.107409
Er MJ, Chen J, Zhang Y et al (2023) Research challenges, recent advances, and popular datasets in deep learning-based underwater marine object detection: a review. Sensors 23(4):1990
Everingham M, Van Gool L, Williams CKI et al (2010) The pascal visual object classes (voc) challenge. Int J Comput Vision 88(2):303–338
Fayaz S, Parah SA, Qureshi G (2022) Underwater object detection: architectures and algorithms—a comprehensive review. Multimedia Tools Appl 81(15):20871–20916
Feng C, Zhong Y, Gao Y et al (2021) TOOD: task-aligned one-stage object detection. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV), pp 3490–3499
Fu J, Liu J, Tian H et al (2019) Dual attention network for scene segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3146–3154
Ge Z, Liu S, Wang F et al (2021) YOLOX: exceeding yolo series in 2021. arXiv preprint. arXiv:2107.08430
Ghiasi G, Lin TY, Le QV (2019) NAS-FPN: learning scalable feature pyramid architecture for object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 7036–7045
Glenn J, Ayush C, Alex S et al (2022) ultralytics/yolov5: v7.0—YOLOv5 SOTA realtime instance segmentation. https://doi.org/10.5281/zenodo.7347926
Guo MH, Xu TX, Liu JJ et al (2022) Attention mechanisms in computer vision: a survey. Comput Vis Media 8(3):331–368
Han K, Wang Y, Chen H et al (2022) A survey on vision transformer. IEEE Trans Pattern Anal Mach Intell 45(1):87–110
Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7132–7141
Huang T, Huang L, You S et al (2022) LightViT: towards light-weight convolution-free vision transformers. arXiv preprint. arXiv:2207.05557
Jocher G, Chaurasia A, Qiu J (2023) YOLO by ultralytics. https://github.com/ultralytics/ultralytics
Li C, Li L, Jiang H et al (2022) YOLOv6: a single-stage object detection framework for industrial applications. arXiv preprint. arXiv:2209.02976
Li X, Wang W, Wu L et al (2020) Generalized focal loss: learning qualified and distributed bounding boxes for dense object detection. Adv Neural Inf Process Syst 33:21002–21012
Lian J, Yin Y, Li L et al (2021) Small object detection in traffic scenes based on attention feature fusion. Sensors 21(9):3031
Lin TY, Maire M, Belongie SJ et al (2014) Microsoft COCO: common objects in context. In: Proceedings of the European conference on computer vision, pp 740–755
Lin TY, Dollár P, Girshick R et al (2017a) Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2117–2125
Lin TY, Goyal P, Girshick R et al (2017b) Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision, pp 2980–2988
Liu C, Li H, Wang S et al (2021a) A dataset and benchmark of underwater object detection for robot picking. In: Proceedings of the IEEE international conference on multimedia & expo workshops (ICMEW), pp 1–6
Liu S, Qi L, Qin H et al (2018) Path aggregation network for instance segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 8759–8768
Liu Y, Sun P, Wergeles N et al (2021) A survey and performance evaluation of deep learning methods for small object detection. Expert Syst Appl 172:114602
Lyu C, Zhang W, Huang H et al (2022) RTMDet: an empirical study of designing real-time object detectors. arXiv preprint. arXiv:2212.07784
MMYOLO Contributors (2022) MMYOLO: OpenMMLab YOLO series toolbox and benchmark. https://github.com/open-mmlab/mmyolo
Qin X, Wang Z, Bai Y et al (2020) FFA-Net: feature fusion attention network for single image dehazing. In: Proceedings of the AAAI conference on artificial intelligence, pp 11908–11915
Redmon J, Farhadi A (2018) YOLOv3: an incremental improvement. arXiv preprint. arXiv:1804.02767
Ren S, He K, Girshick RB et al (2015) Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 39:1137–1149

Rezatofighi H, Tsoi N, Gwak J et al (2019) Generalized intersection over union: a metric and a loss for bounding box regression. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 658–666
Song G, Liu Y, Wang X (2020) Revisiting the sibling head in object detector. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 11563–11572
Sun C, Ai Y, Wang S et al (2021) Mask-guided SSD for small-object detection. Appl Intell 51:3311–3322
Tan M, Pang R, Le QV (2020) EfficientDet: scalable and efficient object detection. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition, pp 10781–10790
Teng B, Zhao H (2020) Underwater target recognition methods based on the framework of deep learning: a survey. Int J Adv Rob Syst 17(6):1729881420976307
Tian Z, Shen C, Chen H et al (2019) FCOS: fully convolutional one-stage object detection. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 9627–9636
Vaswani A, Shazeer N, Parmar N et al (2017) Attention is all you need. In: Proceedings of the Advances in neural information processing systems
Wang W, Dai J, Chen Z et al (2022) InternImage: exploring large-scale vision foundation models with deformable convolutions. arXiv preprint. arXiv:2211.05778
Wang CY, Bochkovskiy A, Liao HYM (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, pp 7464–7475
Wu Y, Chen Y, Yuan L et al (2020) Rethinking classification and localization for object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10186–10195
Wu H, Xiao B, Codella N et al (2021) CvT: introducing convolutions to vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 22–31
Xu S, Wang X, Lv W et al (2022) PP-YOLOE: an evolved version of yolo. arXiv preprint. arXiv:2203.16250
Xu S, Zhang M, Song W et al (2023) A systematic review and analysis of deep learning-based underwater object detection. Neurocomputing 527:204–232
Yang Z, Liu S, Hu H et al (2019) RepPoints: point set representation for object detection. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 9657–9666
Zhang S, Chi C, Yao Y et al (2020) Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9759–9768
Zhu X, Hu H, Lin S et al (2019) Deformable ConvNets V2: more deformable, better results. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9308–9316