International Journal of Network Dynamics and Intelligence Scilight

Article

Underwater Detection: A Brief Survey and a New Multi-task Dataset

Yu Wei [1,2], Yi Wang [1,*], Baofeng Zhu [1], Chi Lin [1], Dan Wu [1], Xinwei Xue [1], and Ruili Wang [3,4]

[1] School of Software Technology, Dalian University of Technology, Dalian 116620, China
[2] Harbin Boiler Co., Ltd, Harbin, 150000, China
[3] School of Mathematical and Computational Sciences, Massey University, Auckland 0632, New Zealand
[4] School of Computer Science, University of Nottingham Ningbo China, Ningbo 315100, China
[*] Correspondence: dlutwangyi@dlut.edu.cn

Received: 27 June 2023
Accepted: 25 April 2024
Published: 25 December 2024

Abstract: Underwater detection poses significant challenges due to the unique characteristics of the underwater environment, such as light attenuation, scattering, water turbidity, and the presence of small or camouflaged objects. To gain a clearer understanding of these challenges, we first review two common detection tasks: object detection (OD) and salient object detection (SOD). Next, we examine the difficulties of adapting existing OD and SOD techniques to underwater settings. Additionally, we introduce a new Underwater Object Multitask (UOMT) dataset, complete with benchmarks. This survey, along with the proposed dataset, aims to provide valuable resources to researchers and practitioners to develop more effective techniques to address the challenges of underwater detection. The UOMT dataset and benchmarks are available at https://github.com/yiwangtz/UOMT.

Keywords: object detection; salient object detection; underwater image enhancement; underwater dataset

1. Introduction

Marine resources are extremely valuable to humans. Underwater detection has broad applications in many areas, such as oceanography, marine navigation, and fish farming [1]. However, the complexities of underwater environments and ecosystems present formidable obstacles to effective underwater detection. For example, underwater environments exhibit considerable variability in lighting conditions, depending on factors such as depth and quality [2−3]. Intricate and multifaceted shapes characterize diverse organisms. Certain organisms also evolved a camouflage mechanism. Furthermore, these organisms often face obfuscation or disruption due to the integration of sand and gravel [4]. Therefore, it calls for the integration of related technologies in computer vision and artificial intelligence to address these challenges in underwater detection.

In this context, Object Detection (OD) [5−7] and Salient Object Detection (SOD) [8−10] are crucial to understanding and analyzing underwater scenes. OD is designed to detect and classify objects accurately while providing their precise spatial locations. SOD focuses on the localization and segmentation of salient objects that conform to the human visual system (HVS) [11]. Several studies and surveys have been conducted in the fields of underwater OD and SOD, highlighting the significance of these tasks in underwater image analysis [12−19].

Our work begins with a brief overview of recent breakthroughs and advances within the OD and SOD fields. Different from existing surveys on these two topics [5−10, 12−14, 20−21], we classify OD and SOD models into two primary categories: (i) Convolutional Neural Networks (CNNs) [22] based approaches and (ii) Transformer [23] based approaches. Furthermore, we delve into the unique challenges posed by underwater detection tasks, as well as the strengths and weaknesses of these approaches.

A comprehensive dataset entitled Underwater Object Multitask (UMOT) is also proposed, which contains various underwater scenes with more than 7K instances encompassing three organism types. UOMT provides COCO dataset [24] format labels for OD and binary segmentation masks for SOD, as illustrated in Figure 1. Evaluations of

state-of-the-art OD and SOD models are conducted on the UMOT dataset.

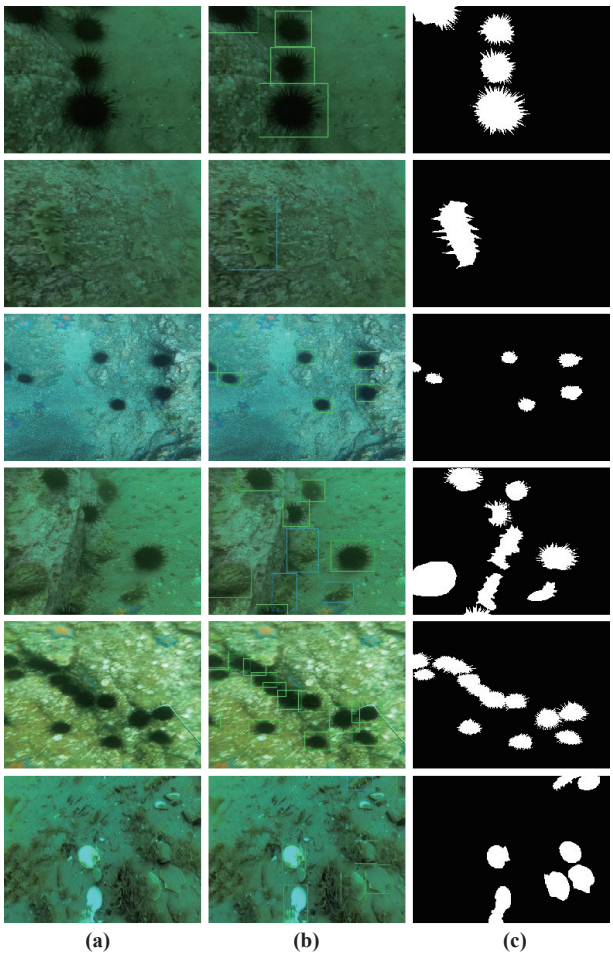

Figure 1. Examples in the proposed Underwater Object Multitask (UOMT) dataset: **(a)** Original images; **(b)** Annotations for Object Detection; **(c)** Annotations for Salient Object Detection.

The rest of this article is presented accordingly. Section II briefly reviews OD methods. Section III briefly reviews SOD methods. Section IV explains underwater OD and SOD challenges. Section V describes the evaluation metrics for OD and SOD. Section VI describes the proposed UOMT dataset. Section VII presents benchmarks for OD and SOD for the UOMT dataset through quantitative and qualitative experiments. Section VIII discusses the future development of underwater OD and SOD. Section IX summarizes the main points of this work.

2. Review of Object Detection (OD)

Object detection (OD) endeavors to accurately detect and classify objects, denoting their positions by rectangular boxes [25]. Object detection encompasses various applications, including face detection, instance segmentation, autonomous driving, surveillance systems, and sports analytics [5].

In recent years, many object detection surveys have been conducted. In 2019, Jiao et al. [5] conducted a comparative analysis of different deep learning-based OD methods. It also introduced some commonly used public datasets, analyzed their characteristics, and highlighted their strengths and weaknesses. In 2020, Wu et al. [6] concluded techniques useful in OD, such as attention-based models [26−28], end-to-end models [29−30], depth separable convolutions [31−32], etc. This survey also presented challenges and requirements for practical applications, including detecting small objects, object tracking, real-time performance, and other related topics. Li et al. [7] extensively explored various aspects of OD in images captured by optical sensors or satellites. The survey also proposed an optical remote sensing image dataset with benchmarks. Padilla et al. [12] provided a detailed and comprehensive exposition of the commonly used performance evaluation metrics in object detection methods. This survey provided details on the principles, advantages, and disadvantages of different indicators and their applicability to different application scenarios. In 2022, Cheng et al. [13] provided a comprehensive study and summary of the problem of small object detection. In 2023, Zou et al. [14] systematically reviewed the development of OD over the past 20

years, offering valuable insights into the field of SOD.

Our review of OD categorizes its methods into CNN-based and Transformer-based models. Table 1 lists the models discussed in this section.

Table 1 Summary of object detection (OD) methods

No.	Year	Method	Backbone	Stage	Anchor
1	2014	R-CNN [33]	AlexNet	Two-Stage	AB
2	2014	SPP-Net [34]	AF-5	Two-Stage	AB
3	2015	Fast R-CNN [26]	VGG-16	Two-Stage	AB
4	2015	Faster R-CNN [27]	VGG-16	Two-Stage	AB
5	2016	R-FCN [35]	ResNet-101	Two-Stage	AB
6	2017	Mask-RCNN [28]	ResNeXt-101	Two-Stage	AB
7	2015	YOLOv1 [36]	GoogleNet	One-Stage	AB
8	2016	SSD [32]	VGG-16	One-Stage	AB
9	2017	YOLOv2 [37]	DarkNet-19	One-Stage	AB
10	2018	RefineDet [38]	VGG16	One-stage	AB
11	2018	YOLOv3 [39]	DarkNet-53	One-Stage	AB
12	2019	EfficientDet [31]	Efficient-B2	One-Stage	AB
13	2020	YOLOv4 [40]	CSPDarkNet-53	One-Stage	AB
14	2021	YOLOv5 [41]	CSPDarkNet-53	One-Stage	AB
15	2021	YOLOF [42]	ResNet-101	One-Stage	AB
16	2021	PP-YOLOv2 [43]	ResNet-101	One-Stage	AB
17	2021	Deformable-DETR [44]	Transformer	One-stage	AB
18	2022	YOLOv7 [45]	ELAN	One-Stage	AB
19	2018	CornerNet [46]	Hourglass-104	One-Stage	AF
20	2020	CircleNet [47]	ResNet-50	One-Stage	AF
21	2021	YOLOX [48]	DarkNet-53	One-Stage	AF
22	2022	YOLOv6 [49]	EfficientRep	One-Stage	AF
23	2022	PP-YOLOE [50]	CSPResNet	One-Stage	AF
24	2023	YOLOv8 [29]	CSPDarkNet	One-Stage	AF
25	2022	DAB-DETR [51]	Transformer	One-Stage	AB
26	2022	DINO [52]	SwinL/ResNet50	One-Stage	AB
27	2023	Mask DINO [53]	SwinL/ResNet50	One-Stage	AB
28	2023	Grounding DINO [54]	Transformer	One-Stage	AB
29	2020	DETR [55]	Transformer	One-Stage	AF
30	2021	YOLOS [56]	Transformer	One-Stage	AF
31	2021	YOLOR [57]	CSPDarknet53	One-stage	AF
32	2022	Detic [58]	Swin Transformer	One-stage	AF
33	2022	DN-DETR [59]	Transformer	One-stage	AF

Note: 'AB' means Anchor-Based, and 'AF' means Anchor-Free.

2.1. CNN-based Object Detection Models

Convolutional Neural Networks (CNNs) [22] exhibit profound influences on object detection [14] by using mechanisms such as ReLU activation, Dropout, Anchor, and GPU acceleration. Among them, the anchor mechanism [27] is a key technique in object detection. We elaborate on OD models based on CNNs by classifying them into two groups: anchor-based detectors and anchor-free detectors, in the following.

1) Anchor-Based Approaches: Anchor [5] mechanism was proposed in 2014. It first generates a series of predefined bounding boxes (anchors) at different locations, and then each anchor is matched to a real object during detection. In the following years, many OD models were proposed using anchors. These methods can be further categorized into two-stage and one-stage detection approaches.

Two-Stage Detectors: Girshick et al. [33] proposed R-CNN in 2014, which revolutionized OD by introducing anchors. R-CNN converts OD into a two-step process: candidate region extraction and classification. R-CNN exhibits remarkable progress compared to conventional OD methods. He et al. [34] introduced SPP-Net, a significant OD breakthrough. SPP-Net facilitates the conversion of an input image with varying dimensions into a feature map with a consistent size, enabling holistic image comprehension. Nonetheless, it faces limitations as it cannot be trained end-to-end. In 2015, Girshick et al. introduced Fast R-CNN [26] and Faster R-CNN [27]. Fast R-CNN uses the entire image as input, avoiding redundant feature computations, and employs a multitask loss function to integrate classification and regression tasks, improving OD efficiency and accuracy. However, it still relies on a region proposal algorithm to generate potential regions, which has speed limitations and compromises the accuracy of the selected regions. The Faster R-CNN, therefore, generates candidate regions using a Region Proposal Network (RPN) instead of a selective search algorithm. Dai et al. [35] proposed the R-FCN, which features a fully convolutional network instead of the region of interest (ROI) pooling operation, avoiding the traditional ROI operation, and can be trained end-to-end. In 2017, Mask-RCNN [28] improved the previous versions with increased detection speed and accuracy.

One-Stage Detectors: In 2015, Redmon et al. [36] proposed YOLOv1, a one-stage detector with faster speed and better real-time performance than two-stage detectors. This method divides an image into several grids, each predicting a bounding box and a category probability. However, it has some disadvantages, such as low detection accuracy, poor detection of small objects, and easy occlusion of dense objects. In 2016, Liu et al. [32] proposed the SSD, which features only one forward propagation to complete the object detection process. In 2017, YOLOv2 [37] was introduced, which improved YOLOv1 with increased detection speed and accuracy. Various excellent methods emerged after 2017. RefineDet [38] and YOLOv3 [39] were proposed in 2018. RefineDet is the first real-time method to achieve detection accuracy greater than 80% *mAP* on PASCAL VOC 2007 [60]. In YOLOv3, performance was further enhanced by a more efficient backbone network, multiple anchors, and spatial pyramid pooling. In 2019, Tan et al. [31] introduced EfficientDet, incorporating a revolutionary integrated scaling technique to improve detection accuracy and computational efficiency.

In 2020, Alexey et al. [40] proposed YOLOv4, and Cai et al. [41] proposed YOLOv5. In YOLOv4, mosaic data augmentation, improved anchor-free detection head, and a new loss function were introduced. YOLOv5 features hyperparameter optimization, integrated experiment tracking, and automatic export to popular export formats.

In 2021, YOLOF [42] was proposed, which used a Dilated Encoder and Uniform Matching that improved detection speed and accuracy without FPN [61]. PP-YOLOv2 [43] is an efficient real-time object detection method that uses a compact foundational architecture. It introduced an adaptive weighted loss function and adaptive label smoothing techniques to better handle objects of different sizes and difficulties. PP-YOLOv2 significantly improved detection performance while maintaining faster speed and practicality.

In 2022, YOLOv7 [45] was introduced using the Extended Efficient Layer Aggregation Network (E-ELAN). YOLOv7 also added additional tasks, such as pose estimation, and surpassed the previous state-of-the-art for real-time applications. DAB-DETR [51] utilizes coil coordinates in Transformer decoders and performs soft ROI pooling.

Anchor-based object detection approaches have certain limitations that can impact training and inference. Objects with small sizes and aspect ratios may be missed or falsely detected due to their size and aspect ratio limitations. Additionally, multiple detection frames may overlap due to anchor overlap, increasing follow-up difficulties and computational effort.

2) Anchor-Free Approaches: In 2018, Zhou et al. [5] introduced the concept of Anchor-Free detection, eliminating the need for predefined bounding boxes. This makes the model simpler, faster to train and infer, and more adaptable to object shapes and sizes. Law and Deng et al. [62] proposed CornerNet in 2018, a CNN-based Anchor-Free detector, that detects the object bounding box as two key points, the upper left and lower right. CornerNet also proposed an effective corner pooling operation that captures boundary information. In 2019, Fei et al. [61] proposed NAS-FPN, which introduces a Neural Architecture Search (NAS) technique that automatically searches the structure of a neural network to obtain a better feature pyramid network. It also combines the advantages of NAS and FPN to automatically search for the optimal pyramid structure of characteristics. That year, it achieved first place in the COCO Object Detection Challenge [24]. In 2020, CircleNet [47] was developed, which treats an object as a circular region consisting of points. It detects the object by predicting the center and radius of the circle. CircleNet detects spherical biomedical objects accurately. In 2021, YOLOX [48] was proposed as an anchor-free version of YOLO, with a simpler design but better performance. YOLOv6 [49] was developed in 2022 and used in many autonomous delivery robots. YOLOv8 [29] improves YOLOv5 by separating classification and detection heads and using Distribution Focal Loss.

Anchor-free methods, however, require more training data and longer training times to achieve the same accuracy as anchor-based detectors, and cluttered images may further hinder Anchor-Free Detectors.

2.2. Transformer-based Object Detection Models

In 2020, DETR [55] (Detection Transformer) first applied the Transformer [23] to object detection. Unlike traditional object detection methods, DETR does not need prior frames or anchor points. Instead, it predicts the positions and classes of objects directly in the image. Specifically, DETR transforms the object detection problem into an ensemble matching problem. YOLOR [57] leverages Transformer to boost feature representation without increasing computation. YOLOR has a unified network that integrates explicit and implicit knowledge to learn a unified representation capable of performing multiple tasks. YOLOS [56] utilizes a Swin Transformer and Focal Loss that improve detection without convolutional layers.

In 2022, Detic [58] proposed using the image classification dataset to train the classification head of the target detector. A strong end-to-end object detector, DINO [52] was developed based on DN-DETR [59], DAB-DETR [51], and Deformable-DETR [44]. DINO improves over previous DETR-like models in performance and efficiency by using a contrastive way to denoize training, a look forward twice scheme for box prediction, and a mixed query

selection method for anchor initialization.

In 2023, a unified object detection and segmentation framework, Mask-DINO [53], was developed. The DINO Mask extends DINO by adding a mask prediction branch that supports all image segmentation tasks (instance, panorama, and semantic). In the same year, another improved DINO model, Grounding DINO [54], was proposed. Grounding DINO can detect specified targets based on text descriptions.

3. Review of Salient Object Detection (SOD)

Salient object detection (SOD) [87] is a computer vision task that aims to identify visually distinctive objects or regions within an image. This task is crucial in various applications, including image editing, visual tracking, image retrieval, and scene understanding [9–10].

Several surveys have focused on detecting salient objects in the last few years. For example, Borji et al. [8] reviewed SOD methods using CNNs in 2019. In 2020, Kumar et al. [9] investigated weakly supervised/pseudo-supervised and adversarial training learning approaches. In 2021, Zhour et al. [88] reviewed RGB-D-based SOD models, as well as related benchmark datasets, using traditional and deep learning methods. In 2022, Fu et al. [20] provided a review and benchmarks for light-field SOD. In 2022, Zhou et al. [21] summarized the latest SOD models and pointed out that different implementation details may affect performance.

In this section, our review focuses on SOD using CNNs and Transformers. Table 2 shows the models presented in this section.

Table 2 Summary of salient object detection (SOD) methods

No.	Year	Method	Backbone	Network Architecture	Features
1	2016	ICANet [63]	-	-	Using the locate-by-exemplar strategy.
2	2016	ELD [64]	VGG-16	Encoder-decoder	Combination of high- and low-level features.
3	2017	FIN [65]	VGG-16	-	The first weakly-supervised learning SOD model.
4	2018	PAGR [66]	VGG-19	-	Adding attention mechanisms to the network.
5	2019	CPD [67]	VGG-16	Encoder-decoder	A new cascaded partial decoder is proposed.
6	2019	PoolNet [68]	VGG-16/ ResNet-50	Encoder-decoder	Pooling-based techniques to supplement advanced semantic information.
7	2019	BASNet [69]	ResNet-34	Encoder-decoder	Prediction-refinement architecture and new hybrid loss function.
8	2019	EGNet [70]	VGG-16/ ResNet-50	Encoder-decoder	Making full use of edge information.
9	2020	LDF [71]	ResNet-50	Encoder-decoder	A label decoupling framework is proposed.
10	2020	MINet [72]	VGG-16/ ResNet-50	Encoder-decoder	A aggregate interaction modules are proposed.
11	2020	UCNet [73]	VGG-16	Encoder-decoder	Learning from the data annotation process to use uncertainty for detection.
12	2020	SAC [74]	ResNet-101	-	Integration of local and global image contexts within, around, and outside of salient objects.
13	2021	PA-KRN [75]	ResNet-50	Encoder-decoder	A progressive strategy to simulate the mechanisms that restore the human visual system.
14	2021	SGL-KRN [75]	ResNet-50	Encoder-decoder	Efficient and lightweight PA-KRN.
15	2021	HQSOD [76]	ResNet-50	Encoder-decoder	SOD in high resolution scenes.
16	2021	DCN [77]	ResNet-50	Encoder-decoder	A multitasking network.
17	2022	EDN [78]	ResNet-50	Encoder-decoder	Extreme downsampling to locate and segment objects.
18	2022	TNet [79]	VGG16/ResNet50	Encoder-decoder	SOD by using thermal infrared images.
19	2022	TRACER[80]	ResNet50 +EfficientNet	Encoder-decoder	Attention-guided tracing module.
20	2023	MENet[81]	ResNet-50	Encoder-decoder	Multi-enhancement and iteratively refinement.
21	2021	SwinNet [82]	Swin	Encoder-decoder	A new cross-modal fusion model is proposed.
22	2021	EBMG[83]	VIT/U-Net	Encoder-decoder	Energy-based latent variable a priori models and generative vision transformer networks.
23	2022	SelfReformer [84]	PVT	Encoder-decoder	Combined with vision transformers and self-refining mechanisms.
24	2022	PGNet [85]	ResNet+Swin	Encoder-decoder	Combined with CNNs and Transformer backbone and graft the features from transformer branch to CNN branch.
25	2023	ICON-P[86]	PVT	Encoder-decoder	Integrity cognitive network proposed based on integrity learning.
26	2023	ICON-S[86]	Swin	Encoder-decoder	Integrity cognitive network proposed based on integrity learning.

3.1. CNN-based SOD Approaches

2014 was the year of pioneering deep learning applied in SOD [89]. In many SOD models, CNNs are used to extract high-level features from images and combine local and global contextual information. These methods capture

semantic information and the salient features of images in complex scenes more effectively than traditional methods. In 2015, FCN [90] networks revolutionized the field using a semantic segmentation architecture at the pixel level. From then on, many SOD models have been developed based on FCNs. Liu et al. [91] used multiscale deep features to express saliency comparisons and prior knowledge. He et al. [63] developed an exemplar-driven top-down saliency detection model that employed a deep association network to learn similarities between exemplars and images. Lee et al. [64] used encoding distance maps, high-level features, and an encoder-decoder structure to generate saliency maps. Wang et al. [65] proposed a weakly supervised SOD model, which reduces the costs of manual annotation. Zhang et al. [66] proposed a progressive attention-guided mechanism to improve the quality of the prediction.

Many excellent SOD algorithms have been proposed since 2019. Most of them employ VGG [92] or ResNet [93] as the backbone network, such as CPD [67], PoolNet [68], BASNet [69], EGNet [70], etc. CPD [67] employs a bidirectional feature pyramid network and a feedback optimization module to generate high-quality saliency maps. PoolNet [68] proposed a Global Guidance Module (GGM) and a Feature Aggregation Module (FAM) based on pooling techniques that combine multiple layers of features at different scales. Contextual information about the image and edge information are also considered. BASNet [69] is a boundary-aware method that uses a prediction-optimization architecture and a hybrid loss function to improve the precision of boundary delineation on saliency maps. EGNet [70] extracted the features of the salient objects through progressive fusion, integrating the local edge information with the global location information and utilizing complementary features to detect objects at various resolutions.

The representative models proposed in 2020 were LDF [71], MINet [72], and UCNet [73]. LDF [71] separated a GT (Ground-Truth) map into a body map and a detailed map. It then performs feature learning on the inner and boundary regions in a two-branch decoder, respectively. UC-Net [73] improved the generalization and robustness of the model by introducing hidden variables to represent the uncertainty of the input data. SAC [74] adaptively propagates, and aggregates image context features with different attenuation factors through a spatially attenuated context (SAC) module and attention mechanisms.

The representative models proposed in 2021 were KRN [75], HQSOD [76], and DCN [77]. KRN [75] introduced a progressive approach to replicate the human visual system restoration process, with the coarse localization module and the fine segmentation module developed. HQSOD [76] extended the SOD for high-resolution images by a low-resolution saliency classification network (LRSCN) to capture semantics at low resolution and a high-resolution refinement network (HRRN) to refine the salient values of pixels in uncertain regions. DCN [77] proposed a multitasking network to predict salient maps, edges, and skeleton maps simultaneously at first. Then it designed cross-task aggregation and cross-layer aggregation modules to integrate multi-level and multi-tasking features in the final results.

In 2022, EDN [78] effectively exploited multiscale learning while improving collaboration between high- and low-level features. TNet [79] used thermal infrared images to get effective object localization and integrity information for RGB decoded features by controlling the interaction between RGB images and thermal images. Using attention-guided tracing modules, Lee et al. [80] proposed a model called TRACER, which eliminates multi-decoder structures and minimizes the use of learning parameters.

In 2023, Zhuge et al. [86] proposed ICON-R, which facilitates the extraction of multiscale feature maps through the integrity cognitive module, edges, and the Integrity Optimization module. Wang et al. [81] proposed MENet, which gradually enhances the cognition of complex targets repeatedly from the perspective of pixels, regions, and objects in images.

3.2. Transformer-based SOD Approaches

Since the Transformer model was applied to vision tasks, the accuracy of SOD has improved greatly [10]. The first proposal that used the Transformer in computer vision was the Vision Transformer (VIT) [94], proposed by the Google Brain team in 2020. It is an image classification model based entirely on self-attentive mechanisms.

In 2021, Liu et al. [95] proposed a pure Transformer-based model for SOD in RGB and RGB-D. The model receives image segments as input and employs the Transformer architecture to disseminate global contextual information among the image segments. The model does not require convolution operations and captures long-range dependencies to improve saliency detection performance. The paper also presents an RGB-D saliency detection dataset for evaluating the model's generalization capability. Zhang et al. [83] proposed EBMG, an energy-based latent variable prior model that defines the distribution of latent variables by an energy function, then samples the latent variables by Markov chain Monte Carlo methods and uses them to optimize the output of the Vision Transformer network.

The Swin Transformer mentioned in the previous section also applies to SOD. In 2022, Liu et al. [82] proposed

SwinNet, which achieved accurate detection using multimodal information RGB-D and RGB-T and local/global interaction mechanisms. SwinNet also provides a pre-trained edge sensing module to better use edge information, achieving better edge retention and refinement capabilities. Yun et al. proposed SelfReformer [84], a transformer-based self-refined network that utilizes global and local contextual information to improve the completeness of saliency maps. Xie et al. proposed PGNet [85], which uses Transformer and CNNs to extract features from images of different resolutions. Cross-Model Grafting Modules (CMGM) are proposed for CNN branches to combine broken detailed knowledge holistically by guiding decoding by different source features. Attention Guided Loss (AGL) is designed to actively supervise the attention matrix generated by CMGM to help the network interact better with attention generated by other models. In 2023, Liu et al. proposed ICON [86], which has three key components to achieve integral SOD, namely the aggregation of various features, the enhancement of the integrity channel, and the verification of the whole. ICON-P employs PVT [96] as the backbone, and ICON-S uses Swin [97] as the backbone.

Given the success of Transformer-based models in salient object detection, it is likely that more models based on this architecture will be developed in the future.

4. Underwater Detection

Accurately detecting and segmenting objects in underwater scenes is more challenging than on land [16−17]. We elaborate on underwater (salient) object detection research from the aspects of image enhancement-based OD/SOD, small object detection, and underwater datasets in the following.

4.1. Underwater Image Enhancement

For underwater OD and SOD, image enhancement is essential to improve object visibility in the water. Due to factors such as light attenuation, scattering, and water turbidity, underwater images often suffer from poor visibility and degraded image quality. To address these issues, various image enhancement techniques are applied, including contrast enhancement, color correction, and noise reduction, to improve the visual quality of underwater images [15−16]. By improving image visibility, the OD and SOD algorithms can more accurately detect and locate underwater objects.

Early underwater image enhancement methods relied on traditional land image enhancement techniques. For example, an adaptive threshold Sobel operator [98] can enhance underwater images by extracting boundaries. In [99], a method to analyze aquatic imagery was proposed by combining maximum RGB with grayscale. First, the maximum RGB value of each underwater image pixel is used as a white reference. Normalization removes the color bias from underwater images. In [100], an adaptive global histogram stretching algorithm was proposed to eliminate low contrast and color loss in underwater scenes.

Currently, deep-learning techniques are used to enhance underwater images for underwater detection tasks. These methods effectively improve the quality of underwater images by addressing challenges such as poor visibility, color distortion, and image degradation. For example, in [101], the authors use the Water-Net [102] to address low contrast, color distortion, and blurring issues in underwater images. In work [103], the authors propose an improved spatial transformation network that adaptively enriches image features based on perspective transformation. This alleviates the limitations of underwater object images taken from different angles. In work [104], the authors develop a perceptual underwater image enhancement model based on two physical priors. Detection perception first provides feedback to an enhancement model, which guides the enhancement model to generate visually satisfactory or detection-friendly images. In work [105], the authors proposed a method for detecting underwater species using a channel sharpening attention mechanism to improve the image channels relevant to the target species and suppress irrelevant channels. In work [106], the authors proposed a color conversion method that transforms underwater images into a more natural color space that improves object detection accuracy. In general, deep learning-based underwater image enhancement techniques demonstrate their value in optimizing underwater detection tasks.

4.2. Small Object Detection

The reason why small objects [107] are more challenging to detect for both OD and SOD is due to two main factors.

Resolution and scale: When the resolution of an image is limited, either due to low-quality sensors or distance from the object, the details and fine-grained features of small objects may not be adequately captured. As a consequence, information is lost, making it more difficult for the OD and SOD methods to detect and localize these objects accurately.

Context and occlusion: Small objects are more susceptible to occlusion by larger objects or environmental factors. This occlusion hinders their visibility and makes it difficult for OD and SOD algorithms to differentiate them from the background or larger objects.

To overcome these challenges, researchers in the field of OD and SOD have explored various techniques and approaches. For example, in work [108], a technique was presented to detect edges in multiple directions, providing a comprehensive representation of the object's boundaries. In work [107], the authors leveraged the YOLOv4 architecture and incorporated multiscale feature aggregation to improve detection accuracy. Furthermore, this work used the fusion of MobileNet-V2 [109] and the depthwise separable convolution [110] to reduce network parameters and size significantly. In work [111], the authors added depthwise separable convolution to the YOLOv4 backbone network with a (152×152) feature map to improve small object detection. They also incorporated a spatial pyramid pooling module to increase model complexity and improve detection accuracy. Transformer-YOLOv5 [112] replaced the prediction head of YOLOv5 with the Transformer module, increasing its detection capability on different scales and in dense environments.

4.3. Underwater Detection Datasets

Given the complexity and dynamics of aquatic ecosystems, datasets for detecting objects in underwater scenes are often scarce and limited. This poses difficulties and challenges for algorithm design and performance evaluation of underwater OD and SOD. There have been efforts to develop such datasets. Table 3 lists some underwater detection datasets. We discuss a few of them in the following.

Table 3 Summary of Underwater Detection Datasets

Year	Dataset	Image Number	Category Number	Annotation	Task	Content
2012	Fish4Knowledge [113]	27,370	23	Bounding Box	UOD	fishes
2019	Brackish [120]	14,518	6	Bounding Box	UOD	big fish, small fish, jellyfish, crabs, etc.
2019	Marine Litter [121]	5,720	3	Bounding Box	UOD	plastic waste, man-made targets, organisms
2019	MUED [122]	8,600	430	Bounding Box	UOD	seafloor objects
2020	URPC2020-DL [114]	8,975	4	Bounding Box	UOD	sea cucumbers, sea urchins, scallops, starfish
2020	RUIE-UHTS [118]	300	3	Bounding Box	UOD	sea cucumbers, sea urchins, scallops
2020	UWD [115]	10,000	4	Bounding Box	UOD	sea cucumbers, sea urchins, scallops, starfish
2020	TrashCan [123]	7,212	22	Bounding Box	UOD	seabed garbage, flora and fauna, etc.
2020	SUIM [124]	1,635	8	Bounding Box	UOD	fish, coral, plants, people, debris, etc.
2020	UIEB [102]	950	8	Bounding Box	UOD	all kinds of corals and marine life, etc.
2021	URPC2021 [114]	10,000	4	Bounding Box	UOD	sea cucumbers, sea urchins, scallops, starfish
2021	DUO [117]	7,782	4	Bounding Box	UOD	sea cucumbers, sea urchins, scallops, starfish
2021	UODD [105]	3,194	3	Bounding Box	UOD	sea cucumbers, sea urchins, scallops
2022	UDD [116]	2,227	3	Bounding Box	UOD	sea cucumbers, sea urchins, scallops
2017	OUC-Vision [125]	4,400	-	Bounding Box	USOD	220 individual objects with four pose variations
2019	MUED [122]	8,600	-	Bounding Box	USOD	430 underwater objects
2020	UFO-120 [126]	1,620	-	Pixel-Wise labels	USOD	multiple locations having different water types
2020	USOD [127]	300	-	Pixel-Wise labels	USOD	various underwater objects
2022	USOD10K [119]	10,225	70	Pixel-Wise labels	USOD	various underwater objects

Fish4Knowledge (F4K) dataset [113] is an extensive collection of fish species, as shown in Figure 2. It contains more than 27,370 images of 23 species of fish collected from various locations and depths. Each image in the dataset is carefully annotated, providing accurate and detailed labels for training and evaluation. However, this dataset has unbalanced numbers of different fish and variable image quality.

Figure 2. Images from the Fish4Knowledge dataset [113].

URPC dataset [114] is the dataset from the Underwater Robot Picking Contest, which has been held every year since 2017. The UPRC2020 dataset consists of 6,575 training images and 2,400 testing images. These images have a high resolution of (3,840×2,160). For UPRC2021, 7,600 images are used for training, and 2,400 images are used for testing. Examples of UPRC2021 are shown in Figure 3.

Figure 3. Images from the URPC2021 dataset [114].

DWD [115], **UDD** [116], **DUO** [117] **datasets** are based on URPC datasets. DWD has 10,000 images of four species: sea cucumbers, sea urchins, scallops, and starfish. DWD has no specific division of training and testing sets. UDD is an underwater marine pasture object detection dataset consisting of 2,227 images in three species categories: sea cucumbers, sea urchins, and scallops. Some examples are shown in Figure 4. DUO has 7,782 images with more accurate annotations and diversity for sea cucumbers, sea urchins, scallops, and starfish.

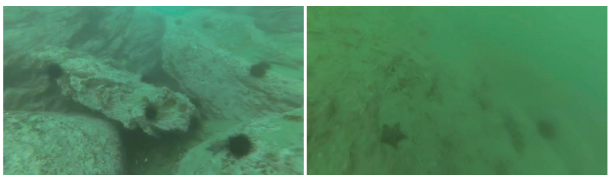

Figure 4. Images from the UDD dataset [116].

UODD dataset [105] has 3,194 images in this dataset from the RUIE [118] dataset in MS COCO format [24]. Diverse underwater scenes are presented in the UODD dataset, e.g., low contrast, multiple objects, large objects, and small objects. Examples are shown in Figure 5.

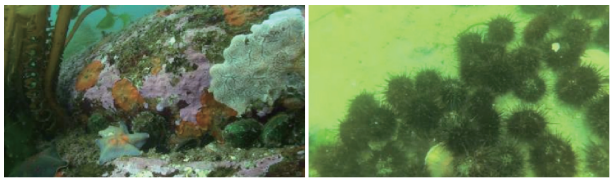

Figure 5. Images from the UODD dataset [105].

USOD10K dataset [119] is the first large-scale underwater SOD dataset. This dataset is significant for its diversity, complexity, and scalability. It contains 10,255 underwater images of seventy classes in various underwater scenes, as shown in Figure 6. The depth and boundary GT maps are also included in this dataset.

Figure 6. Images from the USOD10K [119] dataset.

5. Evaluation Metrics

It is essential to choose evaluation metrics that align with the research objectives to obtain meaningful insights [128]. In the following, we detail the metrics used in the OD and SOD tasks.

5.1. OD Metrics

Following are some of the most common evaluation metrics for object detection.

Precision & Recall [129]: *Precision* is calculated by dividing the true positives by anything that was predicted as a positive:

$$Precision = \frac{TP}{TP+FP}. \tag{1}$$

Recall (or True Positive Rate) is calculated by dividing the true positives by anything that should have been predicted as positive:

$$Recall = \frac{TP}{TP+FN}, \tag{2}$$

where *TP* denotes True Positives, *FP* denotes False Positives, and *FN* denotes False Negatives.

***IoU* (Intersection over Union)** [130]: This metric measures the overlap between the model prediction box and the annotation box. *IoU* can be defined as follows:

$$IoU = \frac{B \cap G}{B \cup G}, \tag{3}$$

where B is the prediction box of an object, and G is the annotation box of an object.

***AP* (Average Precision)** [33]: AP is a commonly used evaluation metric for OD. Using different thresholds, one constructs PR curves to obtain multiple Precision-Recall values. The AP value is derived by integrating the area enclosed by a PR curve and its coordinate axis. AP is defined by:

$$AP = \int_0^1 p(r)dr, \tag{4}$$

where p denotes Precision; r denotes the *Recall*, and $p(r)$ denotes the function with r as the parameter. The PR curve is usually smoothed, i.e., each point on the PR curve comes to the highest *precision* value to the right. It is expressed as follows:

$$P_{smooth}(r) = max_{r' \geqslant r} P(r'). \tag{5}$$

The commonly used AP value is the Interpolated AP, which takes the *Precision* values of $[0, 0.1, ..., 1.0]$ on the horizontal axis and calculates the average value [33]. The AP is obtained by calculating the average of the *Precision* values of 10 equal points of $[0, 0.1, ..., 1.0]$, which is expressed as:

$$AP = \frac{1}{11} \sum_{i=0}^{1.0} P_{smooth}(i). \tag{6}$$

***mAP* (Mean Average Precision)** [33]: The commonly used COCO metrics [24] for the accuracy evaluation are as follows.

- AP is the mAP at $IoU = 0.50 : 0.05 : 0.95$;
- AP_{50} is the mAP at $IoU = 0.50$;
- AP_{75} is the mAP at $IoU = 0.75$;
- AR_{100} represents the AR when considering up to 100 detection points in each image;
- AR_S is calculated for objects of compact dimensions, with an area of less than 32×32;
- AR_M is calculated for objects with intermediate dimensions, with an area between 32×32 and 96×96;
- AR_L is calculated for objects of large dimensions, with an area bigger than 96×96.

5.2. SOD Metrics

The following metrics are typically used to evaluate salient object detection models. [8].

Mean Absolute Error (MAE) [131]: This metric assesses the average pixel-level discrepancy at which the model-generated prediction map differs from the GT map, and is defined by:

$$MAE = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{i=1}^{H} |G(i,j) - Pred(i,j)|, \tag{7}$$

where G represents a binary ground-truth (GT) map; $Pred$ is the predicted map after normalization; W and H are the input image dimensions.

Structure-measure (S_m) [132]: This metric measures how similar the predicted map is to the ground truth and is defined by

$$S_m = \alpha \times S_o + (1 - \alpha) \times S_r, \tag{8}$$

where S_o and S_r represent region- and object-oriented level structural similarity, respectively, α is usually set to 0.5.

Enhanced Alignment Measure (E_m) [133]: This metric identifies local-pixel matching and image-level statistics from binary mapping using an enhanced alignment matrix Q_s, which is defined by

$$E_m = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{i=1}^{H} Q_s(i,j), \tag{9}$$

In this work, we use Adapted E-measure (AE_m) and Mean E-measure (ME_m) via the PySODEvalToolkit [134] in our experiments. The difference is that when calculating Q_s, a threshold is applied to filter the matching pixels in AE_m and ME_m.

AE_m adopts an adaptive threshold (denoted by τ_{ad}) which is the minimum of two times the mean value of the predicted map (denoted by $Pred_m$) and 1, as follows:

$$\tau_{ad} = \min\{2 \times Pred_m, 1\}. \tag{10}$$

Let AT denote the pixels in the prediction map whose grayscale values are greater than or equal to the adaptive threshold τ_{ad}, then we have

$$Q_s = f(\xi_{AT}), \tag{11}$$

where ξ_{AT} denotes the alignment matrix using AT, and $f(x)$ is a convex function selected (e.g., $\frac{1}{4}(1+x)^2$) for calculating the enhanced alignment matrix Q_s [133].

For ME_m, each grayscale value in the histogram of *Pred* is set to be a threshold η_i $(i = 1, ..., N_H)$, where N_H represents the total number of grayscale values in the histogram. As a result, using the set of pixels filtered by each threshold, there is a set of E_m values. By averaging these E_m values, ME_m is determined. For more details, please refer to PySODEvalToolkit [134].

F-measure (F_β) [135]: This metric represents the weighted average of *Precision* and *Recall*, and can be mathematically represented by the following formula:

$$F_\beta = \frac{(1+\beta^2)Precision \times Recall}{\beta^2 Precision + Recall}, \tag{12}$$

where β^2 is usually set to 0.3, to increase the weight of *Precision* and weaken the proportion of *Recall*. Here we also use Adapted F-measure (AF_m) and Mean F-measure (MF_m) [136] in our experiments.

Similar to AE_m and ME_m, AF_m and MF_m employ adaptive thresholds τ_{ad} as defined in formula (10), as well as histogram thresholds η_i (where i ranges from 1 to N_H) for pixel filtering, respectively. Precision and Recall are computed using these two types of thresholds for AF_m and MF_m according to formula (12), respectively. For more details, refer to the PySODEvalToolkit [134].

Weighted F-measure (wF_m) [137]: As an improvement of $F-measure$, wF_m solves the Dependency flaw and Equal-importance flaw well, and the formula can be expressed as:

$$F_\beta^\omega = \frac{(1+\beta^2)Precision^\omega \times Recall^\omega}{\beta^2 Precision^\omega + Recall^\omega}, \tag{13}$$

where $Precision^\omega$ is weighted *Precision* and $Recall^\omega$ is weighted *Recall*. The weighting terms ω for the four fundamental measures (*TP*, *TN*, *FP*, and *FN*) are calculated by the spatial relationship between foreground pixel positions and background pixel positions concerning the foreground.

6. Underwater Object Multitask Dataset

In this section, we present a new underwater object multitask dataset (UOMT), to facilitate underwater research.

6.1. Data Deduplicating

RUIE [118] dataset and UODD [105] dataset are the sources of the proposed UOMT dataset. RUIE dataset contains more than 4K images of underwater objects and environments with different astigmatisms. Over 3K images of underwater cultured products can be found in the UODD dataset, which uses the MS COCO dataset [24] format labels for object detection. Both datasets originate from video streams, so many images are duplicated. Additionally, these two datasets contain numerous tiny objects and extremely blurry environments, which make them unsuitable for the SOD segmentation task.

To support a comprehensive evaluation of multiple underwater tasks, we collected $1,766$ images (over 7K instances for three subjects, including $1,400$ training sets and 366 test sets) from the RUIE dataset and the UODD dataset to make the UOMT dataset. The UOMT dataset goes beyond a simple combination of existing datasets; it represents a series of enhancements and optimizations built on these datasets. Specifically, we ensure that the selected images encompass a broad range of underwater scenes, incorporating various lighting and scattering conditions, as depicted in Figure 1. Meanwhile, a selection of representative images is meticulously curated to ensure uniqueness and variety. Furthermore, to avoid data bias, images of various scenes and objects are collected as much as possible, and duplicate images are avoided to be selected. In addition, we ensure that each target object is presented in at least two distinct images. These strategies enhance the dataset's comprehensiveness and robustness and guarantee broader coverage of objects' appearances.

6.2. Data Annotation

Due to the difficulty in observing underwater objects, we invited 20 observers to detect objects in each image. We use a computer vision annotation tool (CVAT) [138] to label the targets in each image to obtain pixel-level ground truth masks. CVAT supports many annotation types, such as rectangles, polygons, dots, labels, text, etc. We

use CVAT Segmentation Mask 1.1 format for our pixel-level segmentation labels.

To ensure the quality and diversity of the UOMT dataset, we follow several principles in the annotation process.

1) For overlapping objects, they are labeled according to whether they belong to the same category. Objects of the same category are labeled as a whole; objects of different categories are labeled as different categories, and overlapping parts are treated as edges.

2) For objects with very complex boundaries, such as sea urchins with many sharp spines, we label the outline of each spine as meticulously as possible rather than simply drawing an approximate shape.

3) For some fuzzy or difficult-to-distinguish objects, we correct the annotation results through mutual review among 20 annotators to ensure accuracy and consistency.

The UOMT dataset spans a wide spectrum of underwater objects and scenarios and features an array of demanding situations, including category imbalance and small objects. This diverse composition enables the assessment of model generalization and adaptability in real-world underwater environments.

6.3. Data Statistics

The UOMT dataset records different environments, illuminations, categories, dimensions, locations, and quantities of objects, along with backgrounds, etc. Specifically, the following key factors are considered when constructing the UOMT dataset.

Object Size: All objects in each image are manually filtered, including multiple objects, small objects, and camouflaged objects, as shown in Figure 1. This aligns with both the OD and the SOD tasks.

Illumination Conditions: RUIE includes images under various underwater lighting conditions (such as blue and green environments), as well as high-definition and low-illumination environments. The low and varying illumination in different environments poses challenges for OD and SOD tasks. We keep many sample images of different lighting conditions.

Numbers of Object: The UOMT dataset accommodates images that contain multiple objects. The distribution of images that contain multiple objects is shown in Figure 7, and the distribution of images with varying subject types is presented in Figure 8. This inclusion addresses the challenges associated with multiobject OD and SOD tasks.

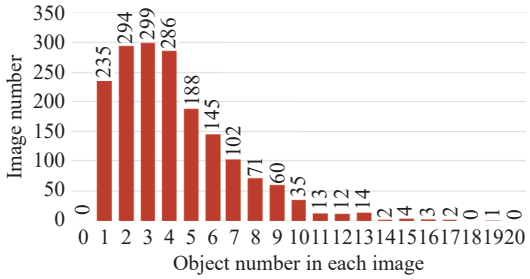

Figure 7. Number of images with multiple objects in the UOMT dataset.

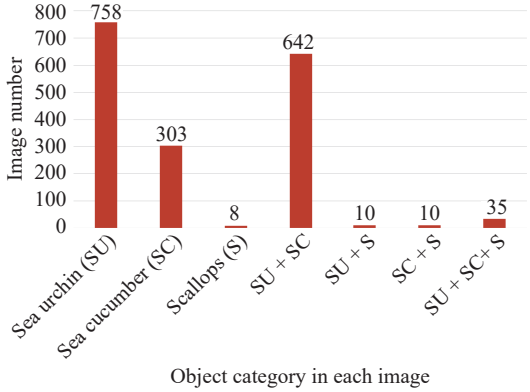

Figure 8. Number of images with different subject types in the UOMT dataset.

Background Diversity: The UOMT dataset intentionally encompasses a range of challenging backgrounds. This includes scenes with obscured views due to aquatic plants, underwater rocks, and corals, as well as cluttered backgrounds such as turbid underwater scenes. In addition, scenes are designed to feature extraneous objects, such as rocks and debris. This diversity of backgrounds has been carefully included to raise detection and segmentation diffi-

culty.

7. Benchmarks

In this section, we evaluate the state-of-the-art methods of object detection and salient object detection on the UOMT, respectively. We give a unified division of training and test sets by randomly selecting 1,400 images as the training set and the rest 366 images as the test set. In addition, to ensure a fair comparison, official codes are used to generate the results of other methods.

7.1. Object Detection Experiments

1) Experimental Settings: Our experiments are based on an open-source toolbox MMDetection (V3.0.0) [143]. During the experiments, we set up the following configurations:

• ImageNet parameters are used to initialize all backbone models. During training, each image is flipped horizontally with a probability of 0.5. SGD [144] method is adopted to optimize all models, and WarmUp [145] is used in each method. All experiments were performed on GTX 1080TI-11G and Tesla K80-11G.

• For Faster R-CNN [27], Cascade R-CNN [139], Mask R-CNN [28], Grid R-CNN [140], ATSS [142], FCOS [141], YOLOF [42] and DINO [52], we resize each image into (320×320) pixels both in training and testing. With batch size 2, the learning rate is initially set at 0.001, decreasing by 0.1 at the 16th and 22nd epochs.

• For CornerNet511 [46], each image is resized to (511×511) pixels for training and testing. The initial learning rate is $1.0/3$ with batch size 9 and decreases by 0.1 in the 180th epoch.

• For YOLOv3 [39], we resize each image into (320×320) pixels in both training and testing. The nine clusters are as follows: (10×13), (16×30), (33×23), (30×61), (62×45), (59×119), (116×90), (156×198), (373×326), and are uniformly distributed across three distinct scales, deviating from the distribution observed in the MS COCO dataset. Initially, the learning rate is set to 0.1 with a batch size of 5 and is decreased by 0.1 at the 218th epoch and the 246th epoch.

• For YOLOX [48], we resize each image in (320×320) pixels both in training and testing. The learning rate was originally set at 0.000625 with a batch size of 2 and was adaptively adjusted during the training process.

2) Quantitative comparison: Table 4 reports quantitative experimental results. One-stage detectors are generally low in accuracy and high in efficiency, while multistage detectors are generally accurate and high in efficiency. In terms of precision, there is no obvious difference between multistage methods (e.g., Cascade R-CNN) and the one-stage methods for *AP* (e.g., FCOS [141]), and the average value of AP_S is always lower than that of AP_M and AP_L. For *AR*, one-stage methods outperform multistage methods. However, multistage methods outperform one-stage methods in AP_L and AR_L. Furthermore, considerable potential remains for enhancing both *AP* and *AR*.

Table 4 OD benchmarks on the UOMT dataset. The symbol ↑ indicates the higher the evaluation metric, the better the model is. The top results are highlighted in bold

Method	Year	Backbone	$AP\uparrow$	$AP_{50}\uparrow$	$AP_{75}\uparrow$	$AP_S\uparrow$	$AP_M\uparrow$	$AP_L\uparrow$	$AR_{100}\uparrow$	$AR_S\uparrow$	$AR_M\uparrow$	$AR_L\uparrow$
Two-stage detectors												
Faster R-CNN w FPN [27]	2015	ResNet-50	39.70	72.00	39.80	31.90	41.50	46.30	50.10	47.20	50.30	50.00
Faster R-CNN w FPN [27]	2015	ResNet-101	41.80	76.00	42.00	34.50	42.90	51.90	51.40	46.90	51.90	55.60
Mask R-CNN w FPN [28]	2017	ResNeXt-101-64x4d	41.80	75.10	42.10	37.30	42.50	49.60	49.00	46.80	49.30	53.30
Cascade R-CNN [40]	2018	ResNet-101	43.20	77.70	39.20	36.10	43.00	51.80	52.10	48.50	50.50	56.20
Grid R-CNN w FPN [140]	2019	ResNeXt-101	42.30	79.30	44.20	36.60	41.70	50.60	52.10	52.70	49.40	55.70
One-stage detectors												
CornerNet511 [46]	2018	Hourglass-104	31.90	59.20	30.00	24.20	35.60	36.80	51.70	46.50	53.80	49.80
YOLOv3 [39]	2018	DarkNet-53	38.80	75.10	36.90	30.10	41.40	49.00	47.90	41.20	50.30	53.60
FCOS [141]	2019	ResNeXt-101-64x4d-FPN	43.50	76.70	46.40	**38.80**	45.50	50.10	57.80	57.80	57.40	54.50
ATSS [142]	2020	ResNet-101	42.60	76.40	41.60	35.40	43.60	49.00	57.10	52.70	58.80	52.80
YOLOv5 [41]	2021	CSPDarkNet-53	**44.40**	**84.00**	**84.00**	28.90	45.10	**52.20**	-	-	-	-
YOLOF [42]	2021	ResNet-50	36.60	73.10	30.30	30.10	39.50	42.00	54.40	50.60	55.40	49.60
YOLOX [48]	2021	YOLOX-I	44.10	78.50	43.00	32.40	**47.00**	51.10	56.70	53.90	57.10	57.60
DINO [52]	2022	ResNet-50	42.10	75.90	43.30	35.00	44.40	43.40	**60.30**	**58.30**	**61.50**	**60.70**

3) Qualitative comparison: We select a few challenging scenes that include large objects, small objects, complex multiple objects, and complex and low-contrast backgrounds for comparison, as shown in Figure 9 and Figure 10. Except for CornerNet [46], other methods detect multi-objects well. FCOS [141] has the highest detection performance and detects large and small objects.

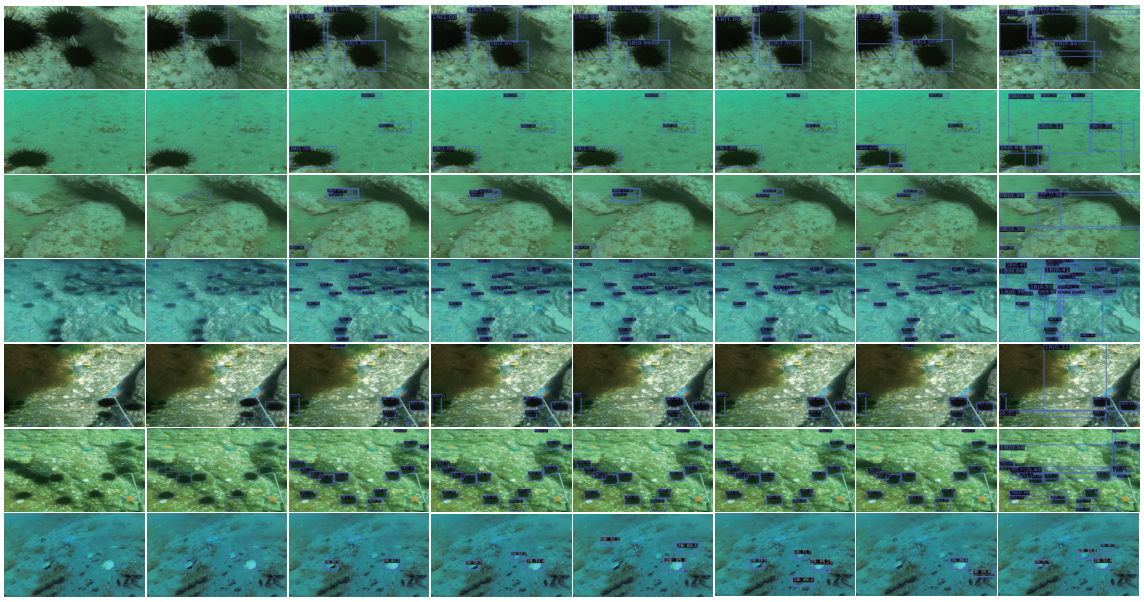

Figure 9. Visual comparison of OD methods on the UOMT dataset. The symbols 0B, 1B, and 2B denote three categories of objects, respectively.

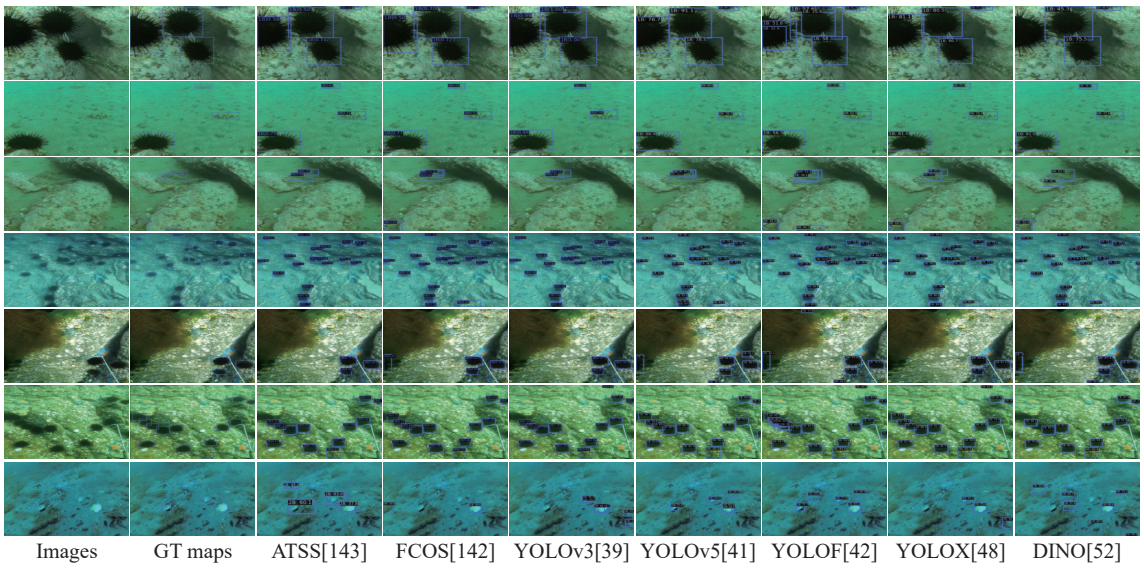

Figure 10. Visual comparison of OD methods on the UOMT dataset. The symbols 0B, 1B, and 2B denote three categories of objects, respectively.

7.2. Salient Object Detection Experiments

1) Experimental Settings: In this section, we train eleven SOD models to demonstrate their performance on the UOMT dataset, including SCRN [146], PoolNet [68], EGNet [70], CPD [67], BASNet [69], LDF [71], Joint-SOD-COD [147], TRACER [80], PGNet [85], EDN [78] and MENet [81]. The relevant configuration of the experiments is as follows.

• The parameters (e.g., learning rate, weight decay, momentum, etc.) of each model in experiments are initialized according to the settings described in their papers.

• To ensure a fair comparison, we use the PySODEvalToolkit [134] as the evaluation tool.

• We perform joint training on GTX 1080TI-11G and Tesla K80-11G.

2) Quantitative Comparisons: Table 5 shows the benchmark for the different SOD methods on the UOMT dataset. The results indicate that reducing the multilevel features (e.g., SCRN [146]), refining the high-level semantic features (e.g., PoolNet [68]), and incorporating deeper layer features (e.g., CPD [67]) produces relatively precise

structural similar salient maps. Integrating fine-grained edge details and comprehensive spatial contexts (i.e., CPD [67] and EGNet [70]) to obtain salient edge features can lead to an increase ME_m value. For MF_m, wF_m, AF_m, and AE_m, a densely guided encoder-decoder architecture, and a residual fine-tuning module, edge information as auxiliary supervision in feature interaction (i.e., LDF [71] and BASNet [69]) has more advantages. Due to the use of both ResNet and Swin-Transformer as the backbone and the reasonable fusion of their features, PGNet [85] has better extraction results and performs the best among all models.

Table 5 SOD benchmarks on the UOMT dataset. The symbol ↑ indicates the higher the evaluation metric, the better the model is, and the symbol ↓ indicates the lower the evaluation metric, the better the model is. The top results are highlighted in bold

Method	Year	Backbone	***MAE*** ↓	S_m ↑	ME_m ↑	MF_m ↑	wF_m ↑	AE_m ↑	AF_m ↑
SCRN [146]	2019	ResNet50	0.019	0.826	0.916	0.766	0.725	0.928	0.711
PoolNet [68]	2019	ResNet50	0.019	0.826	0.917	0.760	0.721	0.924	0.702
PoolNet [68]	2019	VGG16	0.053	0.670	0.710	0.485	0.342	0.657	0.367
EGNet [70]	2019	ResNet50	0.019	0.815	0.927	0.755	0.726	0.932	0.724
EGNet [70]	2019	VGG16	0.024	0.794	0.867	0.708	0.647	0.834	0.586
CPD-ResNet50 [67]	2019	ResNet50	0.019	0.827	0.923	0.765	0.728	0.926	0.712
CPD-VGG16 [67]	2019	VGG16	0.018	0.825	0.929	0.776	0.744	0.940	0.752
BASNet [69]	2019	ResNet50	0.018	0.808	0.920	0.779	0.738	0.945	0.744
LDF [71]	2020	ResNet50	0.018	0.808	0.920	0.779	0.738	0.945	0.744
Joint-SOD [147]	2021	ResNet50	0.017	0.822	0.934	0.788	0.759	0.944	0.764
TRACER [80]	2021	ResNet50	0.020	0.785	0.904	0.737	0.697	0.930	0.710
EDN [78]	2022	VGG16	0.029	0.770	0.917	0.745	0.571	0.929	0.709
EDN [78]	2022	MobileNetV2	0.066	0.676	0.849	0.631	0.344	0.874	0.602
EDN [78]	2022	ResNet50	0.042	0.734	0.900	0.726	0.456	0.911	0.683
PGNet [85]	2022	ResNet-Swin	**0.016**	**0.831**	**0.939**	**0.795**	**0.769**	**0.947**	**0.776**
MENet [81]	2023	ResNet50	0.023	0.721	0.856	0.743	0.613	0.856	0.701

3) Qualitative comparisons: As we can see from Figure 11 and Figure 12, PGNet [85] detects edges better for large objects. Joint-SOD [147] and LDF [71] are sensitive to blurred edges. The remaining methods are not sensitive to edges. With multiple small objects, CPD [67] and EDN [78] perform better than the other methods.

In summary, we can conclude that densely supervised networks introduce edge information as auxiliary supervision, the network refines multilevel features, and integrating local and global location information can improve the accuracy of underwater salient object detection. Transformers increase detection accuracy and efficiency. However, existing SOD methods are still far behind GT in underwater scenes. More research is needed to explore more accurate and efficient underwater salient object detection.

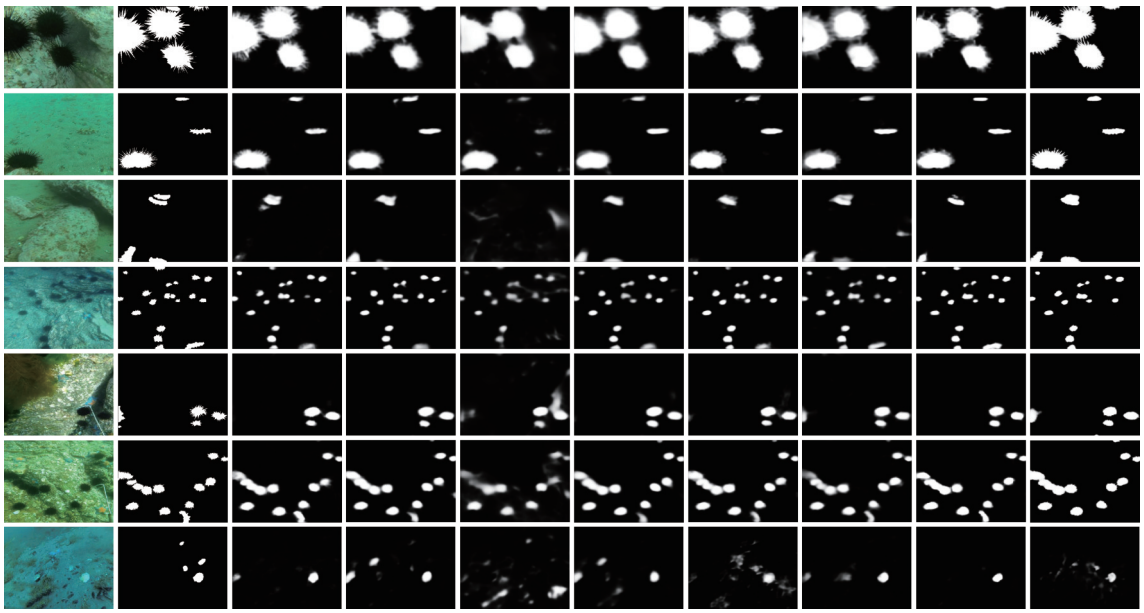

Figure 11. Visual comparison of some SOD methods on the UOMT dataset.

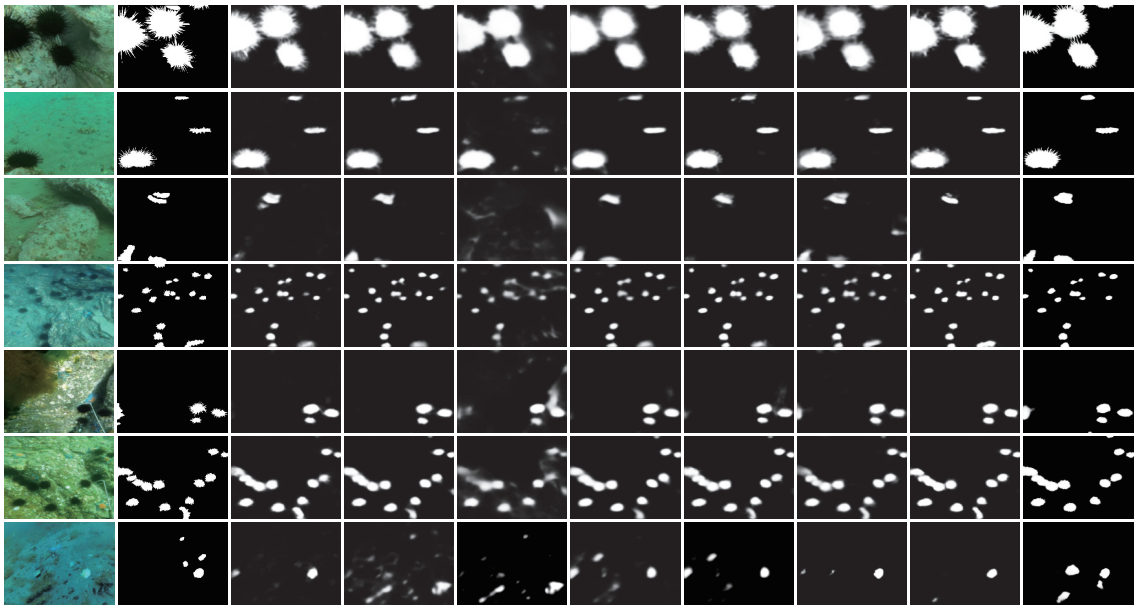

Figure 12. Visual comparison of some SOD methods on the UOMT dataset.

8. Future Development Discussion

Given that we are still in the nascent stages of underwater object detection (OD) and salient object detection (SOD), substantial scope remains for scholarly exploration. We propose the following considerations to provide a guiding framework for future development.

Enhanced Data Diversity: A concerted effort should be directed towards expanding the diversity of underwater datasets, encompassing an even broader range of environmental conditions, object categories, and challenges. This will encourage the development of more robust and adaptable algorithms.

Integration of Multi-Modal Data: Underwater environments are intricate, encompassing many data types that can be gathered, such as optical, acoustic, and magnetic data. Integrating these different data modalities can help enhance underwater object detection accuracy and robustness.

Active Learning and Semi-Supervised Learning: Collecting labeled data for underwater object detection and salient object detection can be challenging and time-consuming. Active learning and semi-supervised learning techniques reduce labeled data requirements and improve learning efficiency.

Incorporation of Domain-Specific Knowledge: The underwater environment has specific properties that differ from the land environments. Incorporating domain-specific knowledge and expertise about these properties into object detection and salient object detection algorithms could help to improve their performance.

Real-Time Processing: Real-time processing is essential for many underwater tasks, including underwater robotics and monitoring. Developing algorithms that can detect objects and detect salient objects in real time could greatly improve the usefulness of these techniques in underwater environments.

Pre-trained Model Migration: In underwater small object detection, models pre-trained on other domains or larger datasets can be used, and migration learning can be performed. By fine-tuning the pre-trained model on underwater data, it is possible to speed up model convergence and improve small object detection performance.

Combination of OD and SOD: The combination of OD and SOD is an exciting direction for underwater detection research, which enables a more comprehensive understanding of scenes. There are several advantages to this. Integrating SOD with OD can help refine object localization. SOD can provide additional information about the most visually prominent parts of an object, aiding in accurate localization. This saves computational resources and increases processing speed, especially in large-scale datasets or real-time applications. However, the output of OD can be used to guide the SOD process, ensuring that the salient regions are associated with the correct objects. This can be particularly useful in crowded scenes where instance separation is challenging. In addition, combining OD- and SOD-generated masks can create diverse training data for both tasks, improving the models' generalization capabilities. However, when designing a model that involves the fusion or serialization of SOD and OD, potential conflicts or inconsistencies between them must be addressed. It is also important to carefully design the integration to avoid redundancy and ensure meaningful enhancements.

9. Conclusion

In this paper, we present a comprehensive overview of object detection (OD) and salient object detection (SOD) techniques specifically tailored to challenging underwater environments. Although there has been notable progress in these domains, it is crucial to acknowledge that research on OD and SOD for underwater scenarios is still relatively young. Significant challenges remain, including the lack of reliable and diverse underwater datasets.

We perform a thorough analysis to address these ongoing challenges and provide insightful recommendations. In addition, we contribute to the research community by introducing a novel underwater object multitask dataset (UOMT). The UOMT dataset is carefully curated and includes various underwater scenes. It offers two essential types of annotation: object detection annotations in the COCO format and salient object detection masks with pixel-level labels.

In addition to providing the dataset, we establish a comprehensive benchmark for OD and SOD tasks. This benchmark encompasses a diverse set of accuracy metrics, making it a valuable resource for academic research and practical industrial implementations. These evaluations enable researchers and practitioners to assess the performance of OD and SOD algorithms under challenging underwater conditions, further advancing the field.

Author Contributions: **Yu Wei:** Conceptualization, data management, evaluation of salient object detection models, and draft writing. **Yi Wang:** Conceptualization, funding acquisition, and manuscript revision. **Baofeng Zhu:** Data management, validation, evaluation of object detection models, and draft writing. **Chi Lin:** Data management, validation, and manuscript revision. **Dan Wu:** Data management, validation, and evaluation of object detection models. **Xinwei Xue:** Evaluation of object detection models, formal analysis, and data management. **Ruili Wang:** Overall supervision and manuscript revision.

Funding: This work is supported in part by the National Natural Science Foundation of China under contract Nos. 62476037, 62172069, and 61976037. This work is also partially supported by 2020 Catalyst: Strategic NZ-Singapore Data Science Research Programme Fund, MBIE, New Zealand.

Data Availability Statement: The data is available at: https://github.com/yiwangtz/UOMT.

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

References

1. B. J. Boom, J. He, S. Palazzo, P. X. Huang, C. Beyan, H.-M. Chou, F.-P. Lin, C. Spampinato, and R. B. Fisher. A research tool for long-term and continuous analysis of fish assemblage in coral-reefs using underwater camera footage. Ecological Informatics, **2014**, *23*: 83−97. doi: 10.1016/j.ecoinf.2013.10.006
2. O. A. Aguirre-Castro, E. Inzunza-González, E. E. García-Guerrero, E. Tlelo-Cuautle, O. R. López-Bonilla, J. E. Olguín-Tiznado, and J. R. Cárdenas-Valdez. Design and construction of an rov for underwater exploration. Sensors, **2019**, *19*(24): 5387. doi: 10.3390/s19245387
3. Z. Chen, R. Wang, W. Ji, M. Zong, T. Fan, and H. Wang. A novel monocular calibration method for underwater vision measurement. Multimedia Tools and Applications, **2019**, *78*: 19437−19455. doi: 10.1007/s11042-018-7105-z
4. S. Fayaz, S. A. Parah, and G. Qureshi. Underwater object detection: architectures and algorithms–a comprehensive review. Multimedia Tools and Applications, **2022**, *81*(15): 20871−20916. doi: 10.1007/s11042-022-12502-1
5. L. Jiao, F. Zhang, F. Liu, S. Yang, L. Li, Z. Feng, and R. Qu. A survey of deep learning-based object detection. IEEE access, **2019**, *7*: 128837−128868. doi: 10.1109/ACCESS.2019.2939201
6. X. Wu, D. Sahoo, and S. C. Hoi. Recent advances in deep learning for object detection. Neurocomputing, **2020**, *396*: 39−64. doi: 10.1016/j.neucom.2020.01.085
7. K. Li, G. Wan, G. Cheng, L. Meng, and J. Han. Object detection in optical remote sensing images: A survey and a new benchmark. ISPRS journal of photogrammetry and remote sensing, **2020**, *159*: 296−307. doi: 10.1016/j.isprsjprs.2019.11.023
8. A. Borji, M.-M. Cheng, Q. Hou, H. Jiang, and J. Li. Salient object detection: A survey. Computational Visual Media, **2019**, *5*(2): 117−150. doi: 10.1007/s41095-019-0149-9
9. A. K. Gupta, A. Seal, M. Prasad, and P. Khanna. Salient object detection techniques in computer vision—a survey. Entropy, **2020**, *22*(1174): 1−49. doi: 10.3390/e22101174
10. W. Wang, Q. Lai, H. Fu, J. Shen, H. Ling, and R. Yang. Salient object detection in the deep learning era: An in-depth survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, **2021**, *44*(6): 3239−325. doi: 10.1109/TPAMI.2021.3051099
11. M. J. Nadenau, S. Winkler, D. Alleysson, and M. Kunt. Human vision models for perceptually optimized image processing—a review. Proc. IEEE, **2000**, *32*: 1−16
12. R. Padilla, S. L. Netto, and E. A. Da Silva, "A survey on performance metrics for object-detection algorithms," in *2020 international conference on systems, signals and image processing (IWSSIP)*, pp. 237–242, IEEE, 2020.
13. G. Cheng, X. Yuan, X. Yao, K. Yan, Q. Zeng, X. Xie, and J. Han. Towards large-scale small object detection: Survey and benchmarks. IEEE Transactions on Pattern Analysis and Machine Intelligence, **2023**, *45*(11): 13467−13488
14. Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye. Object detection in 20 years: A survey. Proceedings of the IEEE, **2023**, *111*(3): 257−276. doi: 10.1109/JPROC.2023.3238524

15. M. Jian, X. Liu, H. Luo, X. Lu, H. Yu, and J. Dong. Underwater image processing and analysis: A review. Signal Processing: Image Communication, **2021**, *91*: 116088. doi: 10.1016/j.image.2020.116088
16. T. Xu, W. Zhao, L. Cai, H. Chai, and J. Zhou, "An underwater saliency detection method based on grayscale image information fusion," in *2022 International Conference on Advanced Robotics and Mechatronics (ICARM)*, pp. 255–260, IEEE, 2022.
17. M. Reggiannini and D. Moroni. The use of saliency in underwater computer vision: A review. Remote Sensing, **2021**, *13*(1): 22
18. M. Zong, R. Wang, X. Chen, Z. Chen, and Y. Gong. Motion saliency based multi-stream multiplier resnets for action recognition. Image and Vision Computing, **2021**, *107*: 104108. doi: 10.1016/j.imavis.2021.104108
19. C. Jing, J. Potgieter, F. Noble, and R. Wang, "A comparison and analysis of rgb-d cameras' depth performance for robotics application," in *2017 24th International Conference on Mechatronics and Machine Vision in Practice (M2VIP)*, pp. 1–6, IEEE, 2017.
20. K. Fu, Y. Jiang, G.-P. Ji, T. Zhou, Q. Zhao, and D.-P. Fan. Light field salient object detection: A review and benchmark. Computational Visual Media, **2022**, *8*(4): 509−534. doi: 10.1007/s41095-021-0256-2
21. H. Zhou, Y. Lin, L. Yang, J. Lai, and X. Xie, "Benchmarking deep models for salient object detection," *arXiv preprint arXiv: 2202.02925*, 2022.
22. Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. Neural computation, **1989**, *1*(4): 541−551. doi: 10.1162/neco.1989.1.4.541
23. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. Advances in neural information processing systems, **2017**, *30*:
24. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, European Conference on Computer Vision (ECCV), September 2014.
25. S. S. A. Zaidi, M. S. Ansari, A. Aslam, N. Kanwal, M. Asghar, and B. Lee. A survey of modern deep learning-based object detection models. Digital Signal Processing, **2022**103514
26. R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015.
27. S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, **2017**, *39*(6): 1137−1149. doi: 10.1109/TPAMI.2016.2577031
28. K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, 2017.
29. ultralytics, "Yolov8," 2023. https://github.com/ultralytics/ultralytics, Last accessed on 2023-06-24.
30. J. Wang, L. Song, Z. Li, H. Sun, J. Sun, and N. Zheng, "End-to-end object detection with fully convolutional network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15849–15858, 2021.
31. M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10781–10790, 2020.
32. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pp. 21–37, Springer, 2016.
33. R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587, 2014.
34. K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, **2015**, *37*(9): 1904−1916. doi: 10.1109/TPAMI.2015.2389824
35. J. Dai, Y. Li, K. He, and J. Sun. R-fcn: Object detection via region-based fully convolutional networks. Advances in neural information processing systems, **2016**, *29*:
36. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (Los Alamitos, CA, USA), pp. 779–788, IEEE Computer Society, Jun 2016.
37. J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7263–7271, 2017.
38. S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Single-shot refinement neural network for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4203–4212, 2018.
39. J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *ArXiv*, vol. abs/1804.02767, 2018.
40. A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv: 2004.10934*, 2020.
41. G. Jocher, A. Stoken, J. Borovec, A. Chaurasia, L. Changyu, A. Hogan, J. Hajek, L. Diaconu, Y. Kwon, Y. Defretin, *et al*., "ultralytics/yolov5: v5. 0-yolov5-p6 1280 models, aws, supervise. ly and youtube integrations," *Zenodo*, 2021.
42. Q. Chen, Y. Wang, T. Yang, X. Zhang, J. Cheng, and J. Sun, "You only look one-level feature," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13039–13048, 2021.
43. X. Huang, X. Wang, W. Lv, X. Bai, X. Long, K. Deng, Q. Dang, S. Han, Q. Liu, X. Hu, *et al*., "Pp-yolov2: A practical object detector," *arXiv preprint arXiv: 2104.10419*, 2021.
44. X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv: 2010.04159*, 2020.
45. C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7464–7475, 2023.
46. H. Law and J. Deng. Cornernet: Detecting objects as paired keypoints. International Journal of Computer Vision, **2020**, *128*(3, SI): 642−656. doi: 10.1007/s11263-019-01204-1
47. E. H. Nguyen, H. Yang, R. Deng, Y. Lu, Z. Zhu, J. T. Roland, L. Lu, B. A. Landman, A. B. Fogo, and Y. Huo. Circle representation for medical object detection. IEEE Transactions on Medical Imaging, **2022**, *41*(3): 746−754. doi: 10.1109/TMI.2021.3122835
48. Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "Yolox: Exceeding yolo series in 2021," *ArXiv*, vol. abs/2107.08430, 2021.
49. C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, Z. Ke, Q. Li, M. Cheng, W. Nie, *et al*., "Yolov6: A single-stage object detection framework for industrial applications," *arXiv preprint arXiv: 2209.02976*, 2022.
50. S. Xu, X. Wang, W. Lv, Q. Chang, C. Cui, K. Deng, G. Wang, Q. Dang, S. Wei, Y. Du, and B. Lai, "Pp-yoloe: An evolved version of yolo," 2022.

51. S. Liu, F. Li, H. Zhang, X. Yang, X. Qi, H. Su, J. Zhu, and L. Zhang, "DAB-DETR: dynamic anchor boxes are better queries for DETR," *CoRR*, vol. abs/2201.12329, 2022.
52. H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H.-Y. Shum, "Dino: Detr with improved denoising anchor boxes for end-to-end object detection," *arXiv preprint arXiv: 2203.03605*, 2022.
53. F. Li, H. Zhang, H. Xu, S. Liu, L. Zhang, L. M. Ni, and H.-Y. Shum, "Mask dino: Towards a unified transformer-based framework for object detection and segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3041–3050, June 2023.
54. S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu, and L. Zhang, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," 2024.
55. N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision (ECCV)*, pp. 213–229, Springer, 2020.
56. Y. Fang, B. Liao, X. Wang, J. Fang, J. Qi, R. Wu, J. Niu, and W. Liu. You only look at one sequence: Rethinking transformer in vision through object detection. Advances in Neural Information Processing Systems, **2021**, *34*: 26183−26197
57. C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao, "You only learn one representation: Unified network for multiple tasks," *arXiv preprint arXiv: 2105.04206*, 2021.
58. X. Zhou, R. Girdhar, A. Joulin, P. Krähenbühl, and I. Misra, "Detecting twenty-thousand classes using image-level supervision," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, pp. 350–368, Springer, 2022.
59. F. Li, H. Zhang, S. Liu, J. Guo, L. M. Ni, and L. Zhang, "Dndetr: Accelerate detr training by introducing query denoising," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13619–13627, 2022.
60. M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results." http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.
61. G. Ghiasi, T.-Y. Lin, and Q. V. Le, "Nas-fpn: Learning scalable feature pyramid architecture for object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7036–7045, 2019.
62. K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "Centernet: Keypoint triplets for object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6569–6578, 2019.
63. S. He, R. W. Lau, and Q. Yang, "Exemplar-driven top-down saliency detection via deep association," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5723–5732, 2016.
64. G. Lee, Y.-W. Tai, and J. Kim, "Deep saliency with encoded low level distance map and high level features," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 660–668, 2016.
65. L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan, "Learning to detect salient objects with image-level supervision," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3796–3805, 2017.
66. X. Zhang, T. Wang, J. Qi, H. Lu, and G. Wang, "Progressive attention guided recurrent network for salient object detection," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 714–722, 2018.
67. Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3902–3911, 2019.
68. J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3917–3926, 2019.
69. X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "Basnet: Boundary-aware salient object detection," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7471–7481, 2019.
70. J. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, and M.-M. Cheng, "Egnet: Edge guidance network for salient object detection," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8778–8787, 2019.
71. J. Wei, S. Wang, Z. Wu, C. Su, Q. Huang, and Q. Tian, "Label decoupling framework for salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13025–13034, 2020.
72. Y. Pang, X. Zhao, L. Zhang, and H. Lu, "Multi-scale interactive network for salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9413–9422, 2020.
73. J. Zhang, D.-P. Fan, Y. Dai, S. Anwar, F. S. Saleh, T. Zhang, and N. Barnes, "Uc-net: Uncertainty inspired rgb-d saliency detection via conditional variational autoencoders," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8582–8591, 2020.
74. X. Hu, C.-W. Fu, L. Zhu, T. Wang, and P.-A. Heng. Sac-net: Spatial attenuation context for salient object detection. IEEE Transactions on Circuits and Systems for Video Technology, **2020**, *31*(3): 1079−1090
75. B. Xu, H. Liang, R. Liang, and P. Chen, "Locate globally, segment locally: A progressive architecture with knowledge review network for salient object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 3004–3012, 2021.
76. L. Tang, B. Li, Y. Zhong, S. Ding, and M. Song, "Disentangled high quality salient object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3580–3590, 2021.
77. Z. Wu, L. Su, and Q. Huang. Decomposition and completion network for salient object detection. IEEE Transactions on Image Processing, **2021**, *30*: 6226−6239. doi: 10.1109/TIP.2021.3093380
78. Y.-H. Wu, Y. Liu, L. Zhang, M.-M. Cheng, and B. Ren. Edn: Salient object detection via extremely-downsampled network. IEEE Transactions on Image Processing, **2022**, *31*: 3125−3136. doi: 10.1109/TIP.2022.3164550
79. R. Cong, K. Zhang, C. Zhang, F. Zheng, Y. Zhao, Q. Huang, and S. Kwong. Does thermal really always matter for rgb-t salient object detection?. IEEE Transactions on Multimedia, **2022**, *25*: 6971−6982
80. M. S. Lee, W. Shin, and S. W. Han. Tracer: Extreme attention guided salient object tracing network (student abstract). Proceedings of the AAAI Conference on Artificial Intelligence, **2022**, *36*: 12993−12994. doi: 10.1609/aaai.v36i11.21633
81. Y. Wang, R. Wang, X. Fan, T. Wang, and X. He, "Pixels, regions, and objects: Multiple enhancement for salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10031–10040, 2023.
82. Z. Liu, Y. Tan, Q. He, and Y. Xiao. Swinnet: Swin transformer drives edge-aware rgb-d and rgb-t salient object detection. IEEE Transactions on Circuits and Systems for Video Technology, **2022**, *32*(7): 4486−4497. doi: 10.1109/TCSVT.2021.3127149
83. J. Zhang, J. Xie, N. Barnes, and P. Li. Learning generative vision transformer with energy-based latent space for saliency prediction. Advances in Neural Information Processing Systems, **2021**, *34*: 15448−15463
84. Y. K. Yun and W. Lin, "Selfreformer: Self-refined network with transformer for salient object detection," *arXiv preprint arXiv:*

2205.11283, 2022.
85. C. Xie, C. Xia, M. Ma, Z. Zhao, X. Chen, and J. Li, "Pyramid grafting network for one-stage high resolution saliency detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11717–11726, 2022.
86. M. Zhuge, D.-P. Fan, N. Liu, D. Zhang, D. Xu, and L. Shao. Salient object detection via integrity learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, **2023**, *45*(3): 3738−3772
87. L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, **1998**, *20*(11): 1254−1259. doi: 10.1109/34.730558
88. T. Zhou, D.-P. Fan, M.-M. Cheng, J. Shen, and L. Shao. Rgb-d salient object detection: A survey. Comput. Vis. Media, **2021**, *7*(1): 37−69. doi: 10.1007/s41095-020-0199-z
89. R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1265–1274, 2015.
90. J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.
91. N. Liu and J. Han, "Dhsnet: Deep hierarchical saliency network for salient object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 678–686, 2016.
92. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," pp. 1–14, Computational and Biological Learning Society, 2015.
93. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
94. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al*., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv: 2010.11929*, 2020.
95. N. Liu, N. Zhang, K. Wan, L. Shao, and J. Han, "Visual saliency transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4722–4732, October 2021.
96. W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao. Pvt v2: Improved baselines with pyramid vision transformer. Computational Visual Media, **2022**, *8*(3): 415−424. doi: 10.1007/s41095-022-0274-8
97. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.
98. A. Saini and M. Biswas, "Object detection in underwater image by detecting edges using adaptive thresholding," in *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, pp. 628–632, IEEE, 2019.
99. F. Han, J. Yao, H. Zhu, and C. Wang. Underwater image processing and object detection based on deep cnn method. Journal of Sensors, **2020**, *2020*(1): 6707328
100. Z. Liu, Y. Zhuang, P. Jia, C. Wu, H. Xu, and Z. Liu. A novel underwater image enhancement algorithm and an improved underwater biological detection pipeline. Journal of Marine Science and Engineering, **2022**, *10*(9): 1204. doi: 10.3390/jmse10091204
101. P. Athira., T. Mithun Haridas, and M. Supriya, "Underwater object detection model based on yolov3 architecture using deep neural networks," in *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, vol. 1, pp. 40–45, 2021.
102. C. Li, C. Guo, W. Ren, R. Cong, J. Hou, S. Kwong, and D. Tao. An underwater image enhancement benchmark dataset and beyond. IEEE Transactions on Image Processing, **2020**, *29*: 4376−4389. doi: 10.1109/TIP.2019.2955241
103. X. Li, F. Li, J. Yu, and G. An, "A high-precision underwater object detection based on joint self-supervised deblurring and improved spatial transformer network," *arXiv preprint arXiv: 2203.04822*, 2022.
104. L. Chen, Z. Jiang, L. Tong, Z. Liu, A. Zhao, Q. Zhang, J. Dong, and H. Zhou. Perceptual underwater image enhancement with deep learning and physical priors. IEEE Transactions on Circuits and Systems for Video Technology, **2020**, *31*(8): 3078−3092
105. L. Jiang, Y. Wang, Q. Jia, S. Xu, Y. Liu, X. Fan, H. Li, R. Liu, X. Xue, and R. Wang. Underwater species detection using channel sharpening attention. Proceedings of the 29th ACM International Conference on Multimedia, **2021**4259−4267
106. C. Yeh, C. Lin, L. Kang, C. Huang, M. Lin, C. Chang, and C. Wang. Lightweight deep neural network for joint learning of underwater object detection and color conversion. IEEE Transactions on Neural Networks and Learning Systems, **2021**, *33*: 6129−6143
107. T.-S. Pan, H.-C. Huang, J.-C. Lee, and C.-H. Chen. Multi-scale resnet for real-time underwater object detection. Signal, Image and Video Processing, **2021**, *15*: 941−949. doi: 10.1007/s11760-020-01818-w
108. K. Hu, F. Lu, M. Lu, Z. Deng, and Y. Liu. A marine object detection algorithm based on ssd and feature enhancement. Complexity, **2020**, *2020*: 1−14
109. M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.
110. A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," 2017.
111. W. Hao and N. Xiao, "Research on underwater object detection based on improved yolov4," in *2021 8th International Conference on Information, Cybernetics, and Computational Social Systems (ICCSS)*, pp. 166–171, IEEE, 2021.
112. Y. Yu, J. Zhao, Q. Gong, C. Huang, G. Zheng, and J. Ma. Real-time underwater maritime object detection in side-scan sonar images based on transformer-yolov5. Remote Sensing, **2021**, *13*(18): 3555. doi: 10.3390/rs13183555
113. R. B. Fisher, Y.-H. Chen-Burger, D. Giordano, L. Hardman, F.-P. Lin, *et al*., *Fish4Knowledge: collecting and analyzing massive coral reef fish video data*, vol. 104. Springer, 2016.
114. L. Chen, Z. Liu, L. Tong, Z. Jiang, S. Wang, J. Dong, and H. Zhou, "Underwater object detection using invert multi-class adaboost with deep learning," in *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, IEEE, 2020.
115. B. Fan, W. Chen, Y. Cong, and J. Tian, "Dual refinement underwater object detection network," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pp. 275–291, Springer, 2020.
116. C. Liu, Z. Wang, S. Wang, T. Tang, Y. Tao, C. Yang, H. Li, X. Liu, and X. Fan. A new dataset, poisson gan and aquanet for underwater object grabbing. IEEE Transactions on Circuits and Systems for Video Technology, **2021**, *32*(5): 2831−2844
117. C. Liu, H. Li, S. Wang, M. Zhu, D. Wang, X. Fan, and Z. Wang, "A dataset and benchmark of underwater object detection for robot picking," in *2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pp. 1–6, IEEE, 2021.
118. R. Liu, X. Fan, M. Zhu, M. Hou, and Z. Luo. Real-world underwater enhancement: Challenges, benchmarks, and solutions under

natural light. IEEE Transactions on Circuits and Systems for Video Technology, **2020**, *30*(12): 4861−4875. doi: 10.1109/TCSVT.2019.2963772

119. L. Hong, X. Wang, G. Zhang, and M. Zhao. Usod10k: a new benchmark dataset for underwater salient object detection. IEEE transactions on image processing, **2023**, 1−1.
120. M. Pedersen, J. Bruslund Haurum, R. Gade, and T. B. Moeslund, "Detection of marine animals in a new underwater dataset with varying visibility," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 18–26, 2019.
121. M. Fulton, J. Hong, M. J. Islam, and J. Sattar, "Robotic detection of marine litter using deep visual detection models," in *2019 international conference on robotics and automation (ICRA)*, pp. 5752–5758, IEEE, 2019.
122. M. Jian, Q. Qi, H. Yu, J. Dong, C. Cui, X. Nie, H. Zhang, Y. Yin, and K.-M. Lam. The extended marine underwater environment database and baseline evaluations. Applied Soft Computing, **2019**, *80*: 425−437. doi: 10.1016/j.asoc.2019.04.025
123. J. Hong, M. Fulton, and J. Sattar, "Trashcan: A semantically-segmented dataset towards visual detection of marine debris," *arXiv preprint arXiv: 2007.08097*, 2020.
124. M. J. Islam, C. Edge, Y. Xiao, P. Luo, M. Mehtaz, C. Morse, S. S. Enan, and J. Sattar, "Semantic segmentation of underwater imagery: Dataset and benchmark," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1769–1776, IEEE, 2020.
125. M. Jian, Q. Qi, J. Dong, Y. Yin, W. Zhang, and K.-M. Lam, "The ouc-vision large-scale underwater image database," in *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1297–1302, IEEE, 2017.
126. M. Islam, P. Luo, and J. Sattar, "Simultaneous enhancement and super-resolution of underwater imagery for improved visual perception," in *Robotics* (M. Toussaint, A. Bicchi, and T. Hermans, eds.), Robotics: Science and Systems, MIT Press Journals, 2020.
127. M. J. Islam, R. Wang, and J. Sattar, "Svam: Saliency-guided visual attention modeling by autonomous underwater robots," in *Robotics: Science and Systems (RSS)*, (NY, USA), 2022.
128. D. M. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," *arXiv preprint arXiv: 2010.16061*, 2020.
129. D. L. Olson and D. Delen, *Advanced data mining techniques*. Springer Science & Business Media, 2008.
130. M. A. Rahman and Y. Wang, "Optimizing intersection-over-union in deep neural networks for image segmentation," in *Proc. Int. Symp. Vis. Comput.*, pp. 234–244, Springer, 2016.
131. F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *2012 IEEE conference on computer vision and pattern recognition*, pp. 733–740, IEEE, 2012.
132. D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *Proceedings of the IEEE international conference on computer vision*, pp. 4548–4557, 2017.
133. D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pp. 698–704, International Joint Conferences on Artificial Intelligence Organization, 7 2018.
134. lartpang Pang, "Pysodevaltoolkit." https://github.com/lartpang/PySODEvalToolkit, 2022.
135. P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. IEEE transactions on pattern analysis and machine intelligence, **2010**, *33*(5): 898−916
136. R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 1597–1604, IEEE, 2009.
137. R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps?," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 248–255, 2014.
138. B. Sekachev, A. Zhavoronkov, and N. Manovich, "Computer vision annotation tool." Website, 2019. https://github.com/opencv/cvat.
139. Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6154–6162, 2018.
140. X. Lu, B. Li, Y. Yue, Q. Li, and J. Yan, "Grid r-cnn," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7355–7364, 2019.
141. Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," in *IEEE/CVF International Conference on Computer Vision (ICCV 2019)*, pp. 9626–9635, IEEE; IEEE Comp Soc; CVF, 2019.
142. S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9756–9765, 2020.
143. K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, *et al*., "Mmdetection: Open mmlab detection toolbox and benchmark," *arXiv preprint arXiv: 1906.07155*, 2019.
144. S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv e-prints*, p. arXiv: 1609.04747, Sept. 2016.
145. H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1487–1495, 2019.
146. Z. Wu, L. Su, and Q. Huang, "Stacked cross refinement network for edge-aware salient object detection," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7263–7272, 2019.
147. A. Li, J. Zhang, Y. Lyu, B. Liu, T. Zhang, and Y. Dai, "Uncertainty-aware joint salient object and camouflaged object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10071–10081, 2021.

Citation: Wei, Y.; Wang, Y.; Zhu, B.; *et al*. Underwater Detection: A Brief Survey and a New Multitask Dataset. *International Journal of Network Dynamics and Intelligence*. 2024, 3(4), 100025. doi: 10.53941/ijndi.2024.100025