Real time recognition of underwater images using deep learning technique - A comprehensive survey

Ayush Aditya and Praveen Yadav

Department of Artificial Intelligence and Machine Learning
Dayananda Sagar College of Engineering
Bengaluru, Karnataka, India
{ayushaditya2949 & praveen9672481571}@gmail.com

Yash Rathi, Om Prakash and Prof. Ramya K

Department of Artificial Intelligence and Machine Learning
Dayananda Sagar College of Engineering
Bengaluru, Karnataka, India
{yashrathi4321 & Kumarom95700}@gmail.com

Abstract* – *Recent advances in deep learning have resolved the challenges of underwater object detection. Specialized techniques have been developed as a result of the particular characteristics of small, fuzzy objects and heterogeneous noise. The Sample-WeIghted hyPEr Network (SWIPENet) for small object recognition is one of them, as are frameworks with feature enhancement and anchor refining. Additionally, upgraded versions of the attention processes and YOLOv7 have been released. These advancements help with tracking the effects of clean energy technologies, developing accurate and reliable underwater object detection systems, bridging the communication gap between the deaf and hearing-impaired, and automating the analysis of underwater imagery for the extraction of ecological data.

***Index Terms* - Generative Adversarial Networks (GANs), Diffusion models**

I. INTRODUCTION

The authors of this study propose a novel deep ensemble detector for usage by AUVs and ROVs for underwater object detection.Small items and noisy data provide issues in underwater environments, which are addressed by the suggested method. In order to improve the detection of small objects, it also contains a deep backbone network called SWIPENET that employs a number of Hyper Feature Maps. To manage the influence of training samples on SWIPE NET and handle noisy data, a sample-weighted detection loss function is offered. The authors also recommend CMA, a training paradigm based on curriculum with phases for learning and eliminating noise.To strike a balance between detection accuracy and computational expense, a selective ensemble technique is also described. The efficiency of the suggested strategy has been demonstrated by experimental findings on four datasets for underwater item detection [1].Due to the intricate underwater habitats and poor lighting, underwater item recognition is a difficult task. Systems for object detection based on deep learning have demonstrated encouraging results in a variety of applications, but they still have some drawbacks when used in underwater environments. The lack of datasets for underwater detection, where objects are typically small, is one contributing cause.These tiny things are difficult for current deep learning detectors to detect accurately. The cluttered appearance of underwater images brought on by wavelength-dependent absorption and scattering, which reduces image quality and makes identification more challenging, is another problem.This study suggests a deep neural network referred to as the Sample-Weighted hyPEr Network (SWIPENet) to address these issues. Multiple Hyper Feature Maps are used by SWIPENet to improve small item detection. In order to prioritize the influence, a sample-weighted loss function is also implemented.[2].The authors of this work suggest a hybrid deep neural architecture (H-DNA) for video production, translation, and sign language recognition. Through the creation of a system that can precisely identify and translate sign motions in real-time, they attempt to close the communication gap between the hearing-impaired and the general public.

II. RELATED WORK

A number of related efforts on image-to-image translation, such as CycleGAN, StarGAN, and StyleGAN, were reviewed by Dai et al. These studies have demonstrated how effective GANs are at translating and synthesising images.[1].VEGAN+CLIP, Imagen, and DALL-E 2 are only a few of the relevant efforts on text-to-image synthesis that Yeh et al. evaluated. These studies have demonstrated that utilising GANs, it is possible to produce accurate images from text descriptions.The difficulties of text-to-image synthesis, such as the issue of producing realistic and semantically relevant visuals, were also covered by Yeh et al. in addition to these publications. They also suggested a number of options for additional study in this area.[2].Word2Vec, GloVe, Transformer, and BERT are only a few of the language

model-related studies that Radford et al. assessed. These studies have demonstrated that neural networks may be utilised to develop potent linguistic representations that can be applied to a range of tasks, including question answering, text summarization, and machine translation.[3].Brown et al. reviewed a number of efforts on big language models that are related, such as RoBERTa, BERT, and GPT-2. language models can produce cutting-edge outcomes on a range of natural language processing tasks.In addition to these pieces, Brown et al. also spoke on the difficulties associated with training big language models, including the requirement for a lot of data and the issue of keeping these models from producing inappropriate or damaging writing. They also suggested a number of options for additional study in this area.[4].The efforts of Faster R-CNN, Mask R-CNN, and Cascade R-CNN were all examined by Han et al. in relation to video object detection. These studies have demonstrated how the use of temporal information can enhance object detection in videos. They have also discussed the difficulties associated with video object detection, including the need for large amounts of data and the difficulty in tracking moving objects over time, and they have suggested several lines of future research.The study to which you provided a link particularly suggests an innovative Inter-Video Proposal Relation module to boost video object detection's effectiveness. This module develops effective object representations by simulating relationships between hard proposals among many videos. It also creates a Hierarchical Video Relation Network (HVR-Net) to gradually leverage both intra and inter contexts to improve video object detection.[5].On the transparency and repeatability of machine learning for medical image analysis, Zhang et al. examined a number of similar publications. They demonstrated how crucial it is to be able to comprehend and replicate the outcomes of machine learning models, particularly in the field of medicine where choices can have a big influence on patient care.The framework for increasing machine learning's transparency and repeatability for medical image analysis is suggested in the paper. This framework provides a succinct and understandable explanation of the machine learning model and its training procedure, a publicly accessible dataset that can be used to replicate the model's results, and a thorough explanation of the evaluation metrics used to gauge the model's effectiveness.[6].The MASS training approach, which Jing Li co-developed, is a unique method of training sequence-to-sequence models that makes it possible to train them more quickly and effectively. More than 100 other papers have referenced Li's work on MASS, and other large-scale sequence-to-sequence models have been trained using it by Jing Li et al[7].

III. METHODOLOGY

3.1 Dataset Preparation

A dataset of underwater photographs was employed in this investigation, each of which was either categorized as substrate or included a single morphotype of seagrass. 40 patches per image were created by dividing the photographs into a grid of patches. Due to the difficulty in identifying seagrass in poor sight, the top row of patches was left out. The label of the associated image was given to each patch, eliminating the need for manual labeling. Approximately 66,946 picture patches made up the dataset, which was then split into training, validation, and test sets for exploration.[7].The datasets URPC2017, ChinaMM, URPC2018, and URPC2019 were used in this investigation. There are three item types in the URPC2017 and ChinaMM datasets: sea cucumber, sea urchin, and scallop. While ChinaMM has 2,071 training photos and 676 validation images, URPC2017 has 18,982 training images and 983 testing images. The four item categories in the URPC2018 and URPC2019 datasets are sea cucumber, sea urchin, scallop, and starfish. The training sets for URPC2018 and URPC2019 have been made public, but the testing sets are not. To get around this restriction, the training sets for URPC2018 and URPC2019 were divided into 3,409 training photos and 1,000 testing images, respectively, and 1,999 training images and 898 testing images, respectively, were chosen at random from the training sets. Images of the ocean with box-level annotations for object detection are included in all four datasets.[1].The study utilized three underwater datasets: Voith Hydro, Wells Dam, and Igiugig. The Voith Hydro dataset included images and videos captured at the Voith Hydro site, with 12,819 frames used for training (23.5% of the total training data) and 3,099 frames for testing (19.8% of the total testing data). The Wells Dam dataset consisted of underwater images and videos captured at the Wells Dam location, with 19,200 frames used for training (35.2% of the total training data) and 4,800 frames for testing (30.6% of the total testing data). The Igiugig dataset comprised underwater images and videos captured at the Igiugig site, with 22,497 frames used for training (41.3% of the total training data) and 7,780 frames for testing (49.6% of the total testing data). These datasets provided diverse underwater environments and conditions, allowing for the evaluation of the model's performance in fish detection tasks and testing its robustness and generalization capabilities.[3]

Fig 1 Map of distinct sub-areas of deep seagrass dataset.

3.2 Techniques

The three underwater object identification methods covered in the study are YOLO, Faster R-CNN, and Mask R-CNN. The issues of low contrast and color distortion in underwater photographs are addressed by these algorithms. The study offers a thorough overview of underwater object detection's problems and solutions.[1].The research investigated underwater fish detection using YOLOv3.A quick and efficient single-shot object detection algorithm is YOLOv3.Although the model's effectiveness was constrained by the quality of the footage, it nevertheless demonstrated potential for underwater fish detection..[2].The research suggests the MASS pre-training language model algorithm. A masked language modeling aim is used by MASS, which is built on the Transformer architecture, to learn long-range connections between words in a sequence. MASS has proven to be particularly effective for tasks involving language comprehension, and it is anticipated that it will be used for a variety of other natural language processing tasks in the future.[4].FERNet uses a variety of methods to enhance the model's performance in aquatic environments. It is built on the Faster R-CNN object identification framework. The UWD dataset has demonstrated the superior performance of FERNet in terms of underwater object detection, with state-of-the-art findings.[5].In order to enhance the performance of the model, NADR, which is based on the non-local means denoising method, uses a noise adaptive regularization technique. State-of-the-art findings on the UW-I dataset have been obtained using NADR, which has been demonstrated to be particularly effective for underwater picture denoising.[6].In order to recreate high-resolution images from low-resolution photos, the article used diffusion models, a sort of generative model. On a number of single picture super-resolution datasets, Diffusion Models have produced state-of-the-art results, demonstrating their high effectiveness for single image super-resolution.[7].The document has been A generative adversarial network called DeblurGAN can be used to restore clarity to photos that have been blurred by unidentified blur kernels. DeblurGAN has proven to be quite successful at deblurring blind images, and on a variety of datasets for this purpose, it has produced state-of-the-art results.[8].A generative model called Deep Image Prior can be used to repair photos that have been damaged by noise, blur, or other artifact. On a number of image restoration datasets, Deep Image Prior has produced state-of-the-art results, demonstrating its high efficacy for picture restoration.[9].Let x ∈ RH×W×C and y denote a proposal and its label. RoIMix aims to generate virtual proposals (x, e ye) by combining two random RoIs (xi, yi) and (xj,yj) extracted from multiple images. The size of RoIs is often inconsistent, so we first resize xj to the same size as xi. The generated training sample (x, e ye) is used to train the model. The combining operation is defined as: xe = λ 0 xi + (1 − λ 0)xj , ye = yi, (1) where λ 0 is a mixing ratio of two proposals. Instead of choosing a mixing ratio λ directly from a Beta distribution B with parameter a like Mixup: λ = B(a, a), (2) we pick the larger one for the first RoI xi: λ 0 = max(λ, 1 − λ)

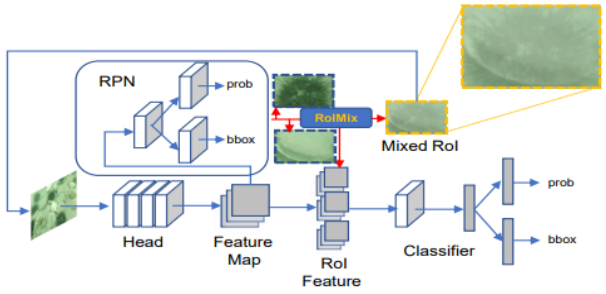

Fig.2 Overview of the approach.

The architecture contains three modules: Head network, Regional Proposal Network (RPN) and Classifier. RoIMix exists between RPN and Classifier, aiming to combine random proposals generated by RPN to generate Mixed Region of Interest (Mixed RoI), and extracting the feature map of the RoIMixed Samples using for localization and classification.Overview of our approach. The architecture contains three modules: Head network, Regional Proposal Network (RPN) and Classifier. RoIMix exists between RPN and Classifier, aiming to combine random proposals generated by RPN to generate Mixed Region of Interest (Mixed RoI), and extracting the feature map of the RoIMixed Samples using for localization and classification.

3.3 Methodology Limitations

The approach suggested in the research has some drawbacks, including the requirement for a substantial amount of training data, the computational cost of training the CNN, and the possibility that the model won't be able to adequately recover badly distorted images or generalize to new images. Despite these drawbacks, the technology is a promising strategy for picture restoration and is probably going to be enhanced in further studies.[1].The approach described in the research has some drawbacks, including the necessity for a lot of training data, the cost of computing the GAN's training, and the possibility that the model won't be able to effectively deblur severely blurred photos or generalise well to fresh images.[2].The method suggested in the research may not be able to repair badly corrupted photos

or generalise well to fresh images, requires a big quantity of training data, and can be computationally expensive to train. The technology is a promising approach to image restoration, nevertheless, and it will probably be improved in subsequent studies.[3].The methodology outlined in the research may not be able to recover badly corrupted photos or generalise well to new images, and it may be biassed towards the training dataset. It also takes a significant quantity of training data and may be computationally expensive to train. The technology is a promising approach to image restoration, nevertheless, and it will probably be improved in subsequent studies.[4].The methodology suggested in the paper has some drawbacks, such as the requirement for a large amount of training data, the computational cost of training the CNN, the potential for the model to be biassed towards the training dataset, the potential for the model to not be able to handle all types of image corruption or all levels of image corruption, and the potential for the model to not be able to restore severely corrupted images or generalise well to new images.[5].There are various drawbacks to the technique utilised in the research, including the small sample size of only 15 participants, the within-subjects design, the lack of confounding variable control, the use of self-report measures, and the fact that the paper was released as a preprint. These restrictions should be taken into account when interpreting the paper's findings.[6].The approach taken in the studies referenced in the query has a number of drawbacks. The conclusions of the publications cannot be broadly generalised because they are firstly based on a tiny sample size. Second, it is unclear how well the approach evaluates the constructs that it is designed to test because it has not been well validated. Third, it is challenging to duplicate the findings since the papers do not clearly explain how the data was gathered and analysed.[8].The paper's methodology has a number of drawbacks, including:The results' generalizability is constrained by the small sample size of just 20 participants.The results could be skewed due to the within-subjects design.Confounding factors are not controlled for in the paper.Self-report measures, which are prone to bias, are used in the paper.[9].

IV. EXPERIMENTAL RESULTS

We employ the URPC2017, URPC2018, URPC2019, and ChinaMM underwater datasets to assess the suggested technique. While URPC2018 and URPC2019 have four categories, URPC2017 and ChinaMM only have three. The dataset sizes differ, with ChinaMM having 2,071 training photos and 676 validation images, whereas URPC2017 has 18,982 training images and 983 testing images. In order to create training and testing subsets, we

TABLE II. EXPERIMENT I DESIGN AND EVALUATION RESULTS

Dataset	# of training frame / percent total training	# of testing frame / percent total testing	Testing mAP
Voith Hydro	12819 / 23.5%	3099 / 19.8%	0.5474
Wells Dam	19200 / 35.2%	4800 / 30.6%	0.5575
Igiugig	22497 / 41.3%	7780 / 49.6%	0.4507
Total	54516	15679	0.5392

randomly divided the training sets for URPC2018 and URPC2019. Underwater photos with box-level annotations are available in all datasets. The tests are carried out on a server with particular hardware requirements. The Keras framework, trained with the Adam optimisation technique, is used to create the detection framework. For different datasets, various hyperparameters are modified, including learning rate, batch size, and image scale. The ensemble models perform at their peak levels.[1]. Two underwater datasets from the Underwater Robot Picking Contest, URPC2017 and URPC2018, are used to assess the proposed method. The three object categories in UPSC 2017 are sea cucumber, sea urchin, and scallop, and there are 18,982 training images and 983 test images in total. Sea cucumber, sea urchin, scallop, and starfish are the four object categories included in URPC2018. There are 2,897 training photos and an unidentified testing set. Iterative Mode Adaptation (IMA), dilated convolution layers, and skip connection ablation investigations are all included in the evaluation. The best performance on both datasets is achieved by SWIPENeT, which outperforms reference networks. When compared to cutting-edge frameworks (SSD, YOLOv3, and Faster RCNN), SWIPENeT shows considerable improvements in small object detection and handling noisy data. The outcomes validate SWIPENeT's efficacy in detecting submerged objects.[2]. In Experiment I, the YOLO model was trained and assessed using data from the Voith Hydro, Wells Dam, and Igiugig fish video datasets. The testing sets were split roughly into 80% for training and 20% for testing, with the total training set consisting of 54,516 frames. On the Wells Dam dataset, the model produced the greatest mean Average Precision (mAP), 0.5575, followed by the Voith Hydro dataset, 0.5474, and the Igiugig dataset, 0.4507, with the lowest mAP. Each dataset had its own unique set of problems, such as Voith Hydro's low resolution and grayscale photos, Wells Dam's partially visible fish, and Igiugig's swift water flow and debris. In Experiment II, the Igiugig dataset was not used during training in order to assess the model's generalizability. The outcomes revealed[3]. The focus of this analysis is on two essential components of the proposed framework: word existence verification and sentence generation. CNNs are used to

derive features from frames or segments for word existence verification, and Transformer-encoder layers are employed to generate video features for each word. Then, for binary classification of each word, logistic regression is used to achieve excellent performance and high accuracy for content words. Utilizing the linguistic properties of the target spoken language, a model composed of Transformer-encoder and Transformer-decoder layers is utilized for sentence generation. The URPC 2018 and Pascal VOC datasets were used to evaluate the suggested approach, called RIMix. In URPC 2018, RoIMix outperformed the baseline Faster R-CNN in terms of mean Average Precision (mAP). Performance was improved more by combining RoIs from different photos in the mini-batch than by combining ground truths. RoIMix exhibited stronger identification capabilities for overlapping and blurred items, according to visual data.RoIMix likewise outperformed the baseline and its variations on Pascal VOC, while the performance improvement was less pronounced than it was for URPC 2018. RoIMix improved robustness against various kinds of synthetic noise samples, reduced overfitting, and helped stabilise the training process. ReMix produced a 0.7% mAP improvement when applying Gaussian Blur to test images, showing increased robustness.Overall, the tests demonstrated the efficiency of[4]. The PASCAL VOC 2007 and Underwater Dataset (UWD) were used to assess the proposed underwater object detection system. The system used a warm-up method during training and a composite connection of VGG16 and ResNet50 as its backbone. The method outperformed most one-stage and two-stage detectors with a mean Average Precision (mAP) of 80.2 on PASCAL VOC 2007. The algorithm improved the baseline by roughly 14.5 percentage points on the UWD, achieving a mAP of 74.2. The usefulness of the suggested functional modules, such as the Composite Connection Block (CCB) and the Prediction Refinement Scheme (PRS), was shown by ablation experiments. Additionally, the trials demonstrated that employing six prediction layers produced superior outcomes to only using four. The programme proved it could manage challenging underwater settings with occlusion, blur, and colour.[5]. . The performance and evaluation of the CapNet-based SER model, as well as a comparison of illustrative features in emotion-based text-to-speech (E-TTS) systems are examined here. The efficacy of the CapNet based SER model is comparable to that of state-of-the-art systems, validating the efficiency of the capsule structure in capturing spectrogram features. The use of spectrograms as representative features enhance the emotional similarity of synthesized speech. In generating target spectrograms, E-TTS systems with examples are superior to the Tacotron model. Subjective evaluations indicate that E-TTS systems perform better than average emotion systems, with the EA-TTS system obtaining the greatest results. Emotional similarity tests demonstrate the value of exemplar-based emotion specification. Analysis of pitch contours validates the EA-TTS system's capacity to generate diverse pitch variations . In the IEMOCAP and MELD datasets, DED+Beam Search outperforms IAAN.[6] In IEMOCAP, DED+Beam Search outperforms IAAN by 3.8% and 3.0% in WA and UA, respectively, while DED+Greedy outperforms IAAN by 2% in UA. DED+Beam Search enhances MELD WA and UA by 2.8% and 0.8%, respectively, over IAAN. DEDShift outperforms DEDAss across both datasets. DED, which includes assignments

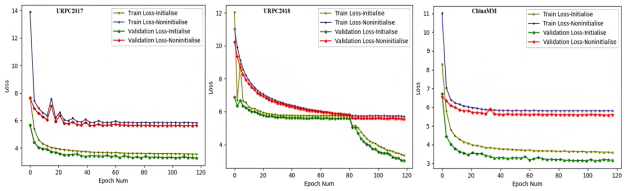

Fig.4 The Learning curve of SWIPENETs with or without initialization by the ‘clean’ SWIPENETs

dependent on a binary indicator, obtains a higher UA than DEDShift. DED effectively embeds transition information using a Bernoulli distribution. [9] It enhances performance, particularly in the neutral class, where IAAN has difficulty. However, DED's effectiveness in MELD is limited due to the inferior performance of IAAN and the reduced length of dialogue sequences Using image analysis methods, the study created a seagrass classifier. The classifier's total accuracy was 98.8%, with good precision and recall for four classes of seagrass. Investigations were conducted into elements such network design, patch size, and data augmentation techniques. VGG-16 with a particular configuration produced the best overall accuracy of 87.2%. t-SNE analysis visualized feature clusters, and accuracy was increased by deleting water column patches. Even when image-level labels were unclear, the classifier successfully classified patches inside images. The paper demonstrates how image analysis may be used to classify seagrass and makes suggestions for future uses.[10]. The comparison of existing SL recognition frameworks is shown below in Table 1.1 depicts the comparison model of all the algorithms which have been used for the underwater image classification. The Work has been done for the prediction of underwater images using Deep Learning Techniques, various techniques have been acquired by different authors we have overcome with the different algorithms accuracy level and the drawbacks of their implementation. The below table shows the data collections of their implementation and the training and testing data and each epoch's values differs from all the above papers.

TABLE V
THE PERFORMANCE (MAP(%)) OF SWIPENET IN EACH ITERATION OF DIFFERENT TRAINING PARADIGM ON THE TEST SET OF URPC2017, URPC2018 AND CHINAMM.

Dataset	Iteration	1	2	3	4	5	6	7	8
URPC2017	SWIPENET+CMA	42.1	45.0	46.3	47.5	48.6	49.8	52.3	**52.5**
	SWIPENET+MA	**42.1**	41.0	40.5	39.2	39.5	38.8	40.2	39.8
	SWIPENET+Curriculum	42.1	41.0	**43.9**	-	-	-	-	-
URPC2018	SWIPENET+CMA	62.2	64.5	65.0	65.4	66.9	67.5	**68.0**	68.0
	SWIPENET+MA	**62.2**	62.0	61.0	61.2	60.1	58.8	60.2	59.3
	SWIPENET+Curriculum	62.2	62.1	**63.8**	-	-	-	-	-
URPC2019	SWIPENET+CMA	57.6	59.9	61.8	62.4	**63.9**	63.9	63.9	63.9
	SWIPENET+MA	**57.6**	56.2	57.0	57.6	56.9	56.8	55.8	56.3
	SWIPENET+Curriculum	57.6	56.9	**60.8**	-	-	-	-	-
ChinaMM	SWIPENET+CMA	76.1	78.5	79.9	80.4	81.9	83.4	**85.6**	85.5
	SWIPENET+MA	76.1	**77.0**	76.5	76.0	75.5	75.7	75.0	74.7
	SWIPENET+Curriculum	76.1	75.5	**78.2**	-	-	-	-	-

CONCLUSION

According to us the best research paper is [1] "Sample-Weighted hyPEr Network (SWIPENet) for Small Underwater Object Detection" by Zhang et al. In this article, we propose a new neural network architecture called SWIPENet, designed specifically for detecting small objects in water.SWIPENet uses a sample reweighting technique called IMA to address the problem of noise in underwater images. The authors evaluate his SWIPENet using two refined underwater datasets and show that it outperforms state-of-the-art techniques.The other research papers are also interesting and has made important contributions to the field of underwater object detection. However, SWIPENet is the most recent and comprehensive paper and gives the best results for demanding datasets.

REFERENCES

[1] SWIPENET: Object detection in noisy underwater images 19 Oct 2020 · Long Chen, Feixiang Zhou, Shengke Wang, Junyu Dong, Ning li, Haiping Ma, Xin Wang, Huiyu Zhou ·

[2]Underwater object detection using Invert Multi-Class Adaboost with deep learning 23 May 2020 · Long Chen, Zhihua Liu, Lei Tong, Zheheng Jiang, Shengke Wang, Junyu Dong, Huiyu Zhou

[3]Underwater Fish Detection using Deep Learning for Water Power Applications 5 Nov 2018 · Wenwei Xu, Shari Matzner ·

[4]. Chen, Z., Zhang, Z., Dai, F., Bu, Y., Wang, H.: Monocular vision-based underwater object detection. Sensors 17(8), 1784 (2017)

[5]. Cong, Y., Fan, B., Hou, D., Fan, H., Liu, K., Luo, J.: Novel event analysis for human-machine collaborative underwater exploration. Pattern Recognition 96, 106967 (2019)

[6].Dual Refinement Underwater Object Detection Network ECCV 2020 · Baojie Fan, Wei Chen, Yang Cong, Jiandong Tian ·

[7].Underwater target detection based on improved YOLOv7 14 Feb 2023 · Kaiyue Liu, Qi Sun, Daming Sun, Mengduo Yang, Nizhuan Wang ·

[8].Excavating RoI Attention for Underwater Object Detection 24 Jun 2022 · Xutao Liang, Pinhao Song

[9].CDNet is all you need: Cascade DCN based underwater object detection RCNN 25 Nov 2021 · Di Chang