Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

A novel robotic visual perception framework for underwater operation*#

Yue LU[1], Xingyu CHEN[2], Zhengxing WU[1], Junzhi YU†‡[1,3], Li WEN[4]

[1]*State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China*

[2]*Ytech, Kuaishou Technology, Beijing 100085, China*

[3]*State Key Laboratory for Turbulence and Complex Systems, Department of Advanced Manufacturing and Robotics, College of Engineering, Peking University, Beijing 100871, China*

[4]*School of Mechanical Engineering and Automation, Beihang University, Beijing 100191, China*

†E-mail: junzhi.yu@ia.ac.cn

Received July 29, 2021; Revision accepted Jan. 25, 2022; Crosschecked Mar. 3, 2022; Published online May 31, 2022

Abstract: Underwater robotic operation usually requires visual perception (e.g., object detection and tracking), but underwater scenes have poor visual quality and represent a special domain which can affect the accuracy of visual perception. In addition, detection continuity and stability are important for robotic perception, but the commonly used static accuracy based evaluation (i.e., average precision) is insufficient to reflect detector performance across time. In response to these two problems, we present a design for a novel robotic visual perception framework. First, we generally investigate the relationship between a quality-diverse data domain and visual restoration in detection performance. As a result, although domain quality has an ignorable effect on within-domain detection accuracy, visual restoration is beneficial to detection in real sea scenarios by reducing the domain shift. Moreover, non-reference assessments are proposed for detection continuity and stability based on object tracklets. Further, online tracklet refinement is developed to improve the temporal performance of detectors. Finally, combined with visual restoration, an accurate and stable underwater robotic visual perception framework is established. Small-overlap suppression is proposed to extend video object detection (VID) methods to a single-object tracking task, leading to the flexibility to switch between detection and tracking. Extensive experiments were conducted on the ImageNet VID dataset and real-world robotic tasks to verify the correctness of our analysis and the superiority of our proposed approaches. The codes are available at https://github.com/yrqs/VisPerception.

Key words: Underwater operation; Robotic perception; Visual restoration; Video object detection
https://doi.org/10.1631/FITEE.2100366 **CLC number:** TP391.4

1 Introduction

Within the last few years, great efforts have been made in underwater robotics (Gong et al., 2018; Li B et al., 2018; Zhu DQ et al., 2019; Cai et al., 2020). For example, Gong et al. (2018) designed a soft robotic arm for underwater operation. Cai et al. (2020) developed a hybrid-driven underwater vehicle-manipulator system for collecting marine products. Concerning intelligent autonomous

‡ Corresponding author

* Project supported by the National Natural Science Foundation of China (Nos. 61633004, 61725305, and 62073196) and the S&T Program of Hebei Province, China (No. F2020203037)

Electronic supplementary materials: The online version of this article (https://doi.org/10.1631/FITEE.2100366) contains supplementary materials, which are available to authorized users

ORCID: Yue LU, https://orcid.org/0000-0001-7472-9935; Junzhi YU, https://orcid.org/0000-0002-6347-572X

robots, visual methods are usually adopted for underwater scene perception (Gong et al., 2018; Cai et al., 2020), and video object detection (VID) plays a decisive role. Over recent years, we have witnessed the development of temporal object detection on the ImageNet VID dataset (Russakovsky et al., 2015). However, there are two problems in underwater object detection, i.e., domain shift (discordance between the training domain and testing domain) and detection continuity and stability.

Deep learning based object detectors have achieved state-of-the-art performance (Zhang et al., 2018; Chen XY et al., 2019a). However, as data-driven detectors, they have problems in underwater object detection. Because of optical absorption and scattering, underwater visual signals usually suffer from degeneration and form low-quality images and videos (Schechner and Karpel, 2004). Therefore, visual restoration has been widely studied (Schechner and Karpel, 2004; Li CY et al., 2016; Chen XY et al., 2019b; Liu RS et al., 2020) to improve visual quality for subsequent image processing. However, visual restoration exactly changes the data domain, an important part of the data-driven learning process (Xu et al., 2014; Raj et al., 2015; Chen YH et al., 2018; Inoue et al., 2018; Khodabandeh et al., 2019). Although visual restoration has been helpful for traditional man-made features, e.g., scale-invariant feature transform (SIFT) (Lowe, 2004; Li CY et al., 2016), the relationship between image quality and convolutional representation remains unclear. In addition, with different data domains, within-domain and cross-domain detection performances have rarely been studied. That is, visual restoration can improve image quality but also creates a domain shift, and its effect on deep learning based underwater object detection remains unclear. Consequently, we have been motivated to investigate the relationship between domain shift and detection performance based on visual restoration. In our opinion, the real sea area is different from datasets, so exploring the effect of the data domain is instructive for building robust underwater detectors.

The domain problem can affect detection accuracy which is essential for robotic perception. Detection continuity and stability are also important for robotic perception. For instance, autonomous grasping needs continuous and stable detection results as feedback, but the changing target position will interfere with control of the robotic arm. As shown in Figs. 1a and 1b, in addition to the false positives/negatives captured by average precision (AP), defective temporal detection cases are subdivided into two aspects: recall continuity and localization stability. Fig. 1a shows the problem of recall continuity. Transient object recall induces short tracklet duration, which can contain only a few frames. Additionally, intermittent missing objects form tracklet fragments, which could cause identity switching. In VID, we call these phenomena "damage recall continuity." The problem of localization stability is shown in Fig. 1b; box center/size jitter frequently appears in modern object detection, and a slight pixel-level change can incur considerable location jitter. It is surmised that this phenomenon impairs localization stability in VID. Note that current VID evaluation indicators, such as AP, cannot assess these two characteristics. Although it reports object recall/missing from a spatial perspective, AP is insufficient for analysis of temporal classification. As shown in Fig. 1, based on intersection-over-union (IoU), AP can hardly reflect discontinuity (Fig. 1d vs. 1e) and instability (Fig. 1f vs. 1g), but these issues cannot be ignored because they could cause jitter and even an error in control of the robot.

In this study, we improve the performance of underwater robotic perception from the aforementioned two aspects. For detection accuracy, we jointly analyze visual restoration and underwater object detection. Typical single-stage detectors (i.e., Single Shot MultiBox Detector (SSD) (Liu W et al., 2016), RetinaNet (Lin et al., 2017), RefineDet (Zhang et al., 2018), and DRNet (Chen XY et al., 2019a)) are investigated on different quality-diverse data domains. In addition, real-world data are collected on the seabed for online object detection. As a result, although visual restoration induces adverse effects on object detection, it efficiently suppresses a domain shift between training images and practical scenes. Thus, visual restoration still plays an essential role in aquatic robotic perception. For detection continuity and stability to perform favorably in robotic scenarios, novel non-reference assessments are proposed based on multi-object tracking (MOT) rather than ground-truth labels. We modify the MOT pipeline to capture recall failures (i.e., missing objects) and design a Fourier approach for stability evaluation. Further, online tracklet refinement (OTR) is

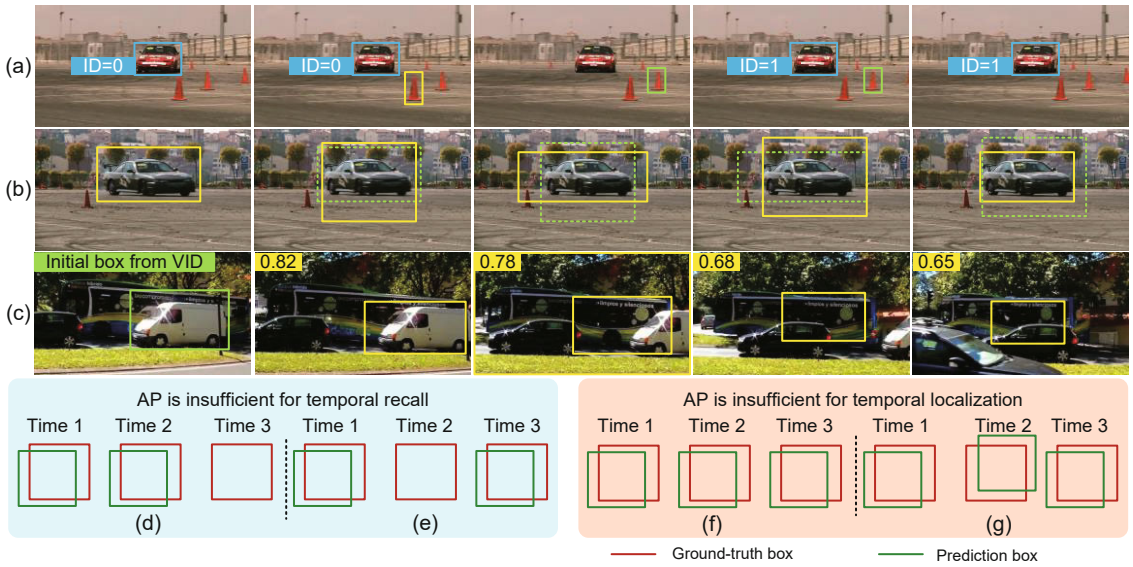

Fig. 1 Defective detection and tracking cases: (a) short tracklet duration (yellow tracklet contains only one object, whereas the green tracklet contains two) and tracklet fragment; (b) box center/size jitter (green dashed boxes denote the previous results); (c) Siamese tracker requiring an initial box from VID and being prone to drift (tracking score shown in the top-left); (d–g) AP hardly describing temporal recall/localization. AP: average precision; VID: video object detection (References to color refer to the online version of this figure)

proposed to enhance VID continuity and stability; it can be generally applied to any detector in temporal tasks. Finally, we propose a novel robotic visual perception framework for object search and grasping. Small-overlap suppression (SOS) is designed for extending VID approaches to single-object tracking (SOT) tasks; thus, SOT-by-detection is proposed, which creates faster inference, free initialization, and flexibility. In addition, visual restoration is used to suppress domain shift. Our framework has high speed and high accuracy on real-world object perception and can switch quickly between MOT and SOT for search and grasping, respectively. Our contributions are summarized as follows:

1. Based on visual restoration, we reveal domain effects on underwater detection. As a result, visual restoration is a useless operation for improving within-domain detection accuracy, leading to lower recall efficiency (Chi et al., 2019). However, it is beneficial in reducing domain shift between training data and practical aquatic scenes so that online detection performance can be boosted. Therefore, it is an essential operation in real-world object perception.

2. Two VID problems are analyzed in a novel way from the robotic perspective, i.e., continuity and stability, and then non-reference assessments are proposed for them. Our assessments can compensate for the deficiency of traditional accuracy-based evaluation. Further, OTR is proposed to generally improve detection continuity and stability, leading to more efficient and robust real-world visual perception.

3. We propose a novel robotic visual perception framework for underwater object search and grasping. SOS is designed to extend VID approaches to SOT tasks without requiring traditional SOT methods. The proposed SOT-by-detection is flexible for VID, MOT, and SOT tasks in robotic perception. In addition, visual restoration is implanted in the framework to reduce domain shift, leading to high underwater detection accuracy. With our framework, we achieve underwater autonomous object search and grasping in real sea areas.

2 Related work

2.1 Underwater visual restoration

Due to natural physical phenomena, underwater visual signals are usually degraded, leading to low-quality vision. In detail, underwater images and videos show low contrast, high color distortion, and strong haziness, creating huge difficulties in image processing. Schechner and Karpel (2004) attributed this degeneration to visual absorption and

scattering. To overcome this difficulty, some methods have been proposed (Li CY et al., 2016; Chen XY et al., 2019b; Liu RS et al., 2020). Recently, Liu RS et al. (2020) built an underwater enhancement benchmark for follow-up works, whose samples were collected on the seabed under natural light. The above-mentioned studies reveal that visual restoration is beneficial in clearing image details and producing salient low-level features. For example, canonical SIFT (Lowe, 2004) algorithms deliver a huge performance improvement based on restoration (Li CY et al., 2016). However, how visual restoration contributes to convolutional neural network (CNN) based feature representation remains unclear. Moreover, visual restoration is tightly related to the data domain, so we explore the domain effect based on restoration.

2.2 Object detection and domain adaptation

During the deep learning era, for single-stage object detection a single-shot network was used for regression and classification. As a pioneering work, Liu W et al. (2016) proposed SSD for real-time detection, and many subsequent works based on SSD further improved the single-stage detection performance (Lin et al., 2017; Zhang et al., 2018; Chen XY et al., 2019a). Although some two-stage detectors (Zhu YS et al., 2019) and anchor-free detectors (Zhou et al., 2019) could produce higher accuracy, the single-stage methods maintain a better accuracy–speed trade-off for robotic tasks.

The above detectors generally assume that training and test samples fall within an identical distribution. However, real-world data usually suffer from domain shift, which affects detection performance. Hence, cross-domain robustness of object detection was recently explored (Xu et al., 2014; Raj et al., 2015; Chen YH et al., 2018; Inoue et al., 2018; Khodabandeh et al., 2019). These works indicated how to moderate the domain shift problem, but there have been relatively few works extensively studying the domain effect on detection performance. In contrast, based on underwater scenarios, we investigate the effect of a quality-diverse data domain on object detection. Kalogeiton et al. (2016) analyzed detection performance based on different image qualities, but our work has advantages over their work: (1) We analyze deep learning based object detection, in consideration of the fact that their work was reported before the deep learning era and lacked this information; (2) Our domain change is derived from realistic visual restorations, whereas their work considered the impact of simple factors (e.g., Gaussian blur); (3) We investigate both cross-domain and within-domain performances, whereas their work analyzed only the cross-domain performance; (4) Our work contributes to aquatic robotics.

2.3 Detection and tracking metrics

AP considers static accuracy based on the detection recall rate and precision (Everingham et al., 2010), but it can hardly give a temporal evaluation for VID methods as mentioned in Section 1. Therefore, some tracking metrics are proposed. Referring to Bernardin and Stiefelhagen (2008), MOT metrics include multi-object tracking accuracy (MOTA) and multi-object tracking precision (MOTP). MOTA synthesizes false positive, false negative, and identity switching of detected objects, whereas MOTP considers static localization precision. Therefore, tracklet fragments were captured by identity switching, but some other tracklet characteristics were ignored (e.g., short tracklet duration).

The visual object tracking (VOT) benchmark uses the expected average overlap (EAO) rate for SOT evaluation (Kristan et al., 2018), including accuracy and robustness. Accuracy is determined by static IoU, and the robustness describes the tracking failure. After the tracking failure is captured by the evaluation process, the tracker would be initialized with the ground truth. EAO could describe tracking fragments, but it could not give a comprehensive evaluation for multiple tracklets.

Tracking metrics (i.e., MOTA, MOTP, and EAO) are able to describe temporal recall; however, similar to AP, they are insufficient for evaluating temporal localization. Moreover, all existing evaluations are based on the ground truth. In contrast, we propose a Fourier approach to directly describe box jitter without the need for labels.

2.4 Detection–SOT cascade

Detection and SOT are distinct in their pipelines, and researchers have tried to simultaneously use their advantages (Kim and Kim, 2016; Feichtenhofer et al., 2017; Kang et al., 2018; Luo et al., 2019). These methods improve tracking and

detection in a complementary manner, but their SOT model and detector are independent, so high model complexity is usually incurred. Instead of the model cascade, we design SOS to extend detection methods toward SOT tasks, and produce an SOT-by-detection framework.

3 Domain effect

3.1 Preliminaries of the data domain and detector based on visual restoration

3.1.1 Domain generation

The dataset is publicly available for underwater object detection, i.e., Underwater Robotic Picking Contest 2018 (URPC2018, http://en.cnurpc.org/), which was collected on the natural seabed at Zhangzidao, Dalian, China. URPC2018 is composed of 2901 aquatic images for training and 800 samples for testing. In addition, it contains four categories: trepang, echinus, shell, and starfish.

Based on URPC2018, three data domains are generated: (1) domain-*O*—the original dataset with the *train* set for training and the *test* set for testing; (2) domain-*F*—all samples are processed by filtering-based restoration (FRS), producing the *train-F* set for training and the *test-F* set for testing; (3) domain-*G*—all samples are restored by generative adversarial network (GAN) based restoration (GAN-RS), generating the *train-G* set for training and the *test-G* set for testing. Mixed *train*, *train-F*, and *train-G* are denoted as *train-all*. As shown in Fig. 2, domain-*O* has strong color distortion, haziness, and low contrast. The degraded visual samples are effectively restored in domain-*F* and domain-*G*.

3.1.2 Domain analysis

According to Chen XY et al. (2019b), Lab color space has good ability to describe underwater image properties. The bias between the distribution center and balance point (i.e., (128, 128)) means strong color distortion, and the concentrated distribution indicates strong haziness. In addition, the underwater color image quality evaluation metric (UCIQE) (Yang and Sowmya, 2015) and underwater image quality measures (including underwater image colorfulness measure (UICM), underwater image sharpness measure (UISM), underwater image contrast measure (UIConM), and underwater image quality measure (UIQM)) (Panetta et al., 2016) are used to describe domain quality, and high indicators represent high quality. Referring to Fig. 3 and Table 1, for domain quality, we define domain-*G*>domain-*F*>domain-*O*.

3.1.3 Detector

Because of the ability to induce both high accuracy and real-time inference, we use single-stage detectors to perform underwater offline/online object detection. In detail, this paper investigates SSD (Liu W et al., 2016), RetinaNet (Lin et al., 2017), RefineDet (Zhang et al., 2018), and DRNet (Chen XY et al., 2019a). All these detectors are trained based on *train*, *train-F*, *train-G*, or *train-all*. The backbone networks we have used (VGG16, MobileNet, and ResNet101) are pre-trained on the ImageNet dataset for better extraction of category-agnostic basic visual features. For evaluation, mean average precision (mAP) is employed to describe detection accuracy. Eight NVIDIA GeForce RTX 1080Ti GPUs are used for both training and testing.

Table 1 Quality assessment for the data domain

Domain	UCIQE	UICM	UISM	UIConM	UIQM
Domain-*O*	0.39	0.20	3.86	0.12	1.58
Domain-*F*	**0.56**	**3.38**	12.87	0.17	4.51
Domain-*G*	0.53	2.27	**13.81**	**0.18**	**4.78**

The best results are in bold. UCIQE: underwater color image quality evaluation metric; UICM: underwater image colorfulness measure; UISM: underwater image sharpness measure; UIConM: underwater image contrast measure; UIQM: underwater image quality measure

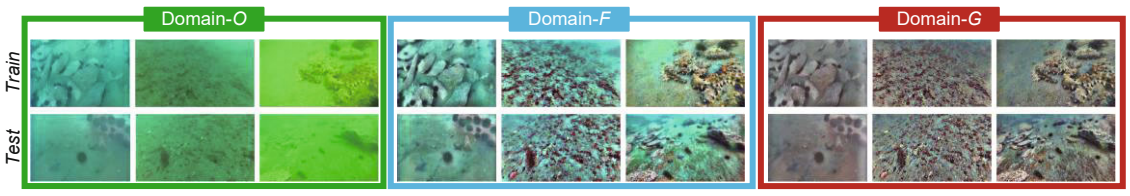

Fig. 2 Typical samples in domain-*O*, domain-*F*, and domain-*G*

3.2 Within-domain performance

In this test, detectors' training and evaluation are based on an identical data domain. The following analysis will unveil two points: (1) Domain quality has an ignorable effect on detection performance; (2) Restoration is a thankless method for improving within-domain detection performance, because of the problem of low recall efficiency. Note that low recall efficiency means low precision under the condition of the same recall rate (Chi et al., 2019).

3.2.1 Numerical analysis

Two sets of training and evaluation are performed: (1) SSD with different input sizes (i.e., 320 and 512) and backbones (i.e., VGG16 (Simonyan and Zisserman, 2014), MobileNet (Howard et al., 2017), and ResNet101 (He et al., 2016)); (2) RetinaNet512, RefineDet512, and DRNet512 with VGG. As shown in Tables 2 and 3, in terms of mAP, detection accuracy is negatively correlated with domain quality. However, mAP cannot reflect accuracy details, so for the following analysis we will continue investigating within-domain performance.

3.2.2 Visualization of convolutional representation

Fig. 4 demonstrates multi-scale features in SSD and DRNet. These features serve as the input for detection heads, so they are the final convolutional features for detection. Referring to Fig. 4, despite domain diversity, there is relatively little difference in object saliency in multi-scale feature maps. It is seen that convolution is able to capture salient object representation from the low-quality data domain. Hence, in terms of object saliency, domain quality has an ignorable effect on convolutional representation.

3.2.3 Precision–recall analysis

As shown in Fig. 5, precision–recall curves are employed for further analysis of detection performance. It can be seen that precision–recall curves have two typical appearances. On one hand, the

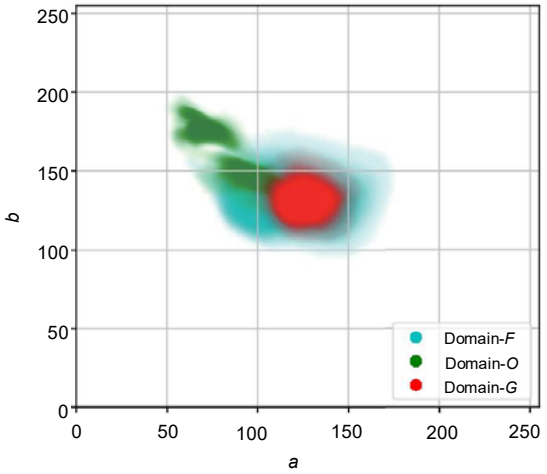

Fig. 3 Domain visualization in the Lab color space. The *a*-*b* distribution of domain-*O* is concentrated and has color bias. In contrast, distributions of domain-*F* and domain-*G* are more scattered and have smaller biases. Color transparency indicates the distribution probability (References to color refer to the online version of this figure)

Table 2 SSD detection results under conditions of different input sizes or backbones

Method	Training data	Test data	mAP (%)	Precision (%)			
				Trepang	Echinus	Shell	Starfish
SSD320-VGG16	*Train*	*Test*	**69.3**	67.8	**84.9**	**44.7**	**79.7**
	Train-F	*Test-F*	67.8	**68.9**	82.3	42.2	78.0
	Train-G	*Test-G*	65.9	65.4	82.3	39.0	76.9
SSD512-VGG16	*Train*	*Test*	**72.9**	**70.2**	**87.1**	**50.8**	**83.5**
	Train-F	*Test-F*	71.3	68.9	85.8	48.5	82.1
	Train-G	*Test-G*	69.5	67.2	84.7	45.3	80.9
SSD512-MobileNet	*Train*	*Test*	**70.7**	**65.3**	**87.1**	**47.5**	**82.8**
	Train-F	*Test-F*	68.9	63.7	85.1	45.4	81.7
	Train-G	*Test-G*	67.4	61.5	84.9	42.6	80.5
SSD512-ResNet101	*Train*	*Test*	**67.0**	59.8	**86.3**	**41.7**	**80.3**
	Train-F	*Test-F*	65.6	**61.1**	84.7	37.5	79.1
	Train-G	*Test-G*	64.6	60.1	83.7	38.6	76.2

The best results are in bold. mAP: mean average precision

Table 3 Detection results based on RetinaNet, RefineDet, and DRNet

Method	Training data	Test data	mAP (%)	Precision (%)			
				Trepang	Echinus	Shell	Starfish
RetinaNet512-VGG16	*Train*	*Test*	**74.0**	**69.8**	**88.1**	**54.7**	**83.4**
	Train-F	*Test-F*	72.5	69.1	87.1	50.7	82.9
	Train-G	*Test-G*	71.0	67.3	86.9	48.9	81.1
RefineDet512-VGG16	*Train*	*Test*	**76.0**	**73.8**	**90.2**	**54.1**	**85.8**
	Train-F	*Test-F*	72.9	72.0	88.6	46.4	84.6
	Train-G	*Test-G*	72.0	71.4	88.4	46.3	81.8
DRNet512-VGG16	*Train*	*Test*	**77.1**	**75.6**	**91.1**	**55.1**	**86.7**
	Train-F	*Test-F*	75.4	73.6	89.8	52.7	85.6
	Train-G	*Test-G*	73.8	72.0	89.8	49.9	83.5

The best results are in bold. mAP: mean average precision

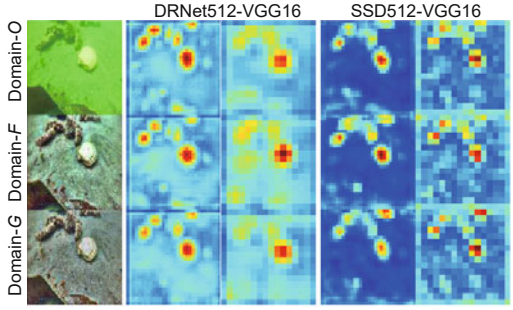

Fig. 4 Visualization of convolutional representation of objects. These features are associated with 64- or 128-size anchors, matching objects in this image. All features are processed using the L2 norm across channels, and are then normalized for visualization. For a fair comparison, the same normalization factor is used for scale-identical features (References to color refer to the online version of this figure)

high-precision part contains high-confidence detection results, and here domain-related curves are highly overlapped. That is, when detecting high-confidence objects, domain difference is negligible for detection accuracy. On the other hand, curves are separated in the low-precision part. In detail, the curve of domain-F is usually below that of domain-O, while the curve of domain-G is usually below that of domain-F. That is, when detecting hard objects (i.e., low-confidence detection results), false positives increase with better domain quality. Therefore, recall efficiency is gradually reduced with increasing restoration intensity.

Based on the above analysis, it can be concluded that visual restoration impairs recall efficiency and is unfavorable for improving within-domain detection. In addition, because domain-related mAPs are relatively close and high-confident recall is far more important than low-confidence recall in robotic perception, we conclude that domain quality has an ignorable effect on within-domain object detection.

3.3 Cross-domain performance

In this test, detectors are trained and evaluated on different data domains. The following analysis presents three viewpoints: (1) It is widely accepted that domain shift induces a significant drop in accuracy; (2) For cross-domain inference, learning based on the low-quality domain is more generalizable than that based on the high-quality domain; (3) In domain-mixed learning, the low-quality domain makes a smaller contribution, so low-quality samples cannot be well learned.

3.3.1 Cross-domain evaluation

We use domain-O and domain-G for evaluation of direction-related domain shift. That is, we train detectors on *train* and evaluate them on *test-G*, or vice versa. As shown in Table 4, the mAPs of all categories seriously decline. As a result, if *train* and *test-G* are employed, SSD512-VGG16 suffers from a 17.4% mAP drop, whereas DRNet512-VGG16 encounters 15.9% decrease in mAP. However, if *train-G* and *test* are adopted, SSD and DRNet would suffer from a more dramatic accuracy deterioration, i.e., mAP decrease of 49.4% and 56.3%, respectively. According to different degrees of accuracy drop caused by direction-opposite domain shift, it is seen that the generalization of *train* to *test-G* is better than that of *train-G* to *test*. Therefore, it can be concluded that compared to the high-quality domain, the low-quality domain induces better cross-domain generalizability.

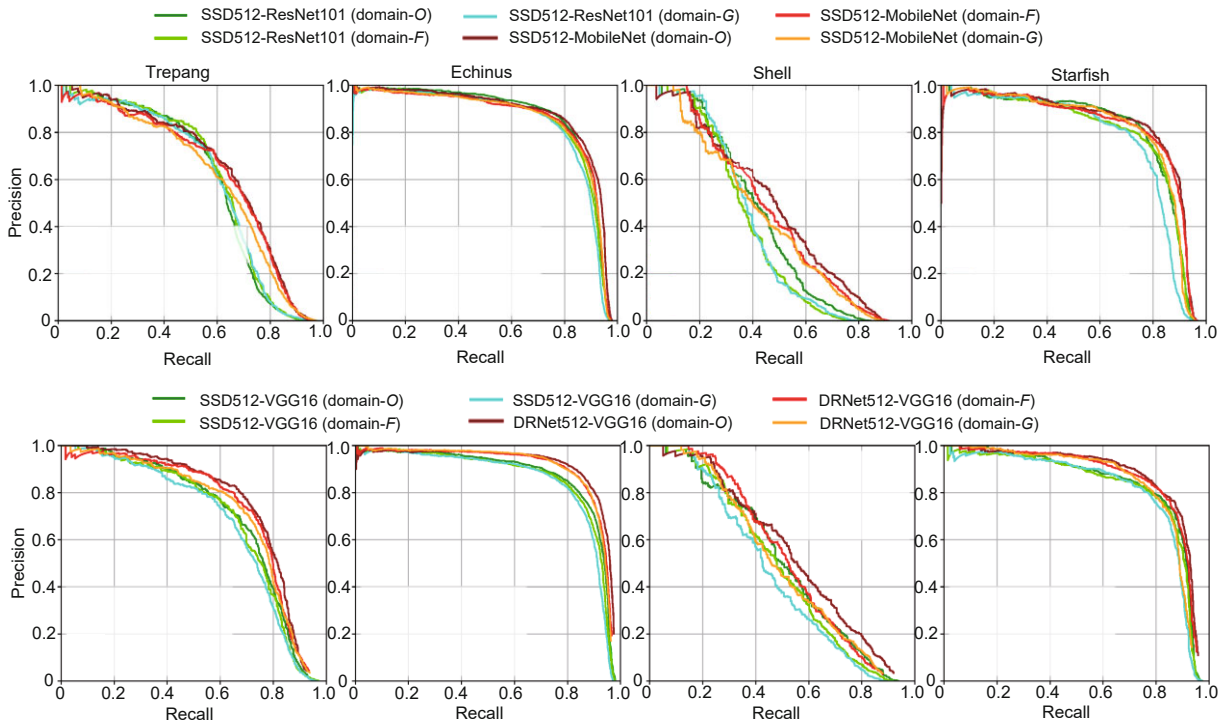

Fig. 5 Precision–recall curves: for high precision (e.g., > 0.9), domain difference has an ignorable effect on detection performance. Overall, domain-*F* and domain-*G* reduce recall efficiency, resulting in a lower AP. AP: average precision (References to color refer to the online version of this figure)

Table 4 Cross-domain evaluation

Method	Training data	Test data	mAP (%)	Precision (%)			
				Trepang	Echinus	Shell	Starfish
SSD512-VGG16	*Train*	*Test-G*	52.1 ↓ 17.4	42.5 ↓ 24.7	70.2 ↓ 14.5	36.6 ↓ 8.7	59.0 ↓ 21.9
	Train-G	*Test*	23.5 ↓ 49.4	15.5 ↓ 54.7	42.3 ↓ 44.8	12.9 ↓ 37.9	23.3 ↓ 60.2
DRNet512-VGG16	*Train*	*Test-G*	57.9 ↓ 15.9	53.7 ↓ 18.3	74.2 ↓ 15.6	40.0 ↓ 9.9	63.7 ↓ 19.8
	Train-G	*Test*	20.8 ↓ 56.3	7.5 ↓ 68.1	44.5 ↓ 46.6	13.6 ↓ 41.5	17.3 ↓ 69.4

↓ means decrease with respect to within-domain performance of the same test set. mAP: mean average precision

3.3.2 Cross-domain training

To explore the detection performance with mixed domain learning, we use *train-all* to train detectors, and then evaluate them on *test*, *test-F*, and *test-G*. Referring to Table 5, on *test-F* and *test-G*, SSD512-VGG16 and DRNet512-VGG16 perform on par with their within-domain performances. However, both SSD512-VGG16 and DRNet512-VGG16 exhibit dramatically worse accuracy on *test*, i.e., greater than a 20% mAP drop. With the same training settings, within-domain performances can be similarly produced on high-quality domain-*F* and domain-*G*, but low-quality domain-*O* suffers from a significant decline in accuracy. That is, when *train-all* is adopted, samples in *train* lose their effects to some extent. Thus, we conclude that cross-domain training is useless in improving detection performance. Moreover, a quality-diverse data domain makes different contributions to the learning process, so that low-quality samples cannot be well learned if mixed with high-quality samples.

Table 5 Cross-domain training

Method	Training data	Test data	mAP (%)	Precision (%)			
				Trepang	Echinus	Shell	Starfish
SSD512-VGG16	*Train-all*	*Test*	51.0 ↓ 21.9	34.5 ↓ 35.7	75.6 ↓ 11.5	40.9 ↓ 9.9	53.1 ↓ 30.4
		Test-F	71.4 ↑ 0.1	69.2 ↑ 0.3	85.4 ↓ 0.4	48.4 ↓ 0.1	82.4 ↑ 0.3
		Test-G	67.3 ↓ 2.2	63.8 ↓ 3.4	83.0 ↓ 1.7	45.5 ↑ 0.2	76.9 ↓ 4.0
DRNet512-VGG16	*Train-all*	*Test*	52.0 ↓ 25.1	34.5 ↓ 41.1	75.6 ↓ 15.5	40.9 ↓ 14.2	53.1 ↓ 33.6
		Test-F	75.8 ↑ 0.4	75.0 ↑ 1.4	89.8 0	53.1 ↑ 0.4	85.3 ↓ 0.3
		Test-G	72.2 ↓ 1.6	70.5 ↓ 1.5	86.6 ↓ 3.2	51.1 ↑ 1.2	80.7 ↓ 2.8

↓ and ↑ respectively mean decrease and increase with respect to within-domain performance of the same test set. mAP: mean average precision

3.4 Domain effect on online detection

We collect real-world aquatic videos on the seabed, called "online data," to reveal the domain effect on robotic perception: How does visual restoration contribute to object detection?

3.4.1 Online object detection in aquatic scenes

Based on online data, we use DRNet512-VGG16 to detect underwater objects. According to different training domains, we denote detection methods as DRNet512-VGG16-*O*, DRNet512-VGG16-*F*, and DRNet512-VGG16-*G*, which are trained on *train*, *train-F*, and *train-G*, respectively. If DRNet512-VGG16-*F* or DRNet512-VGG16-*G* is employed, corresponding visual restoration (i.e., FRS or GAN-RS) should also be adopted to cope with online data. As shown in Fig. 6, DRNet512-VGG16-*O* almost completely loses its effect on object perception. In addition, DRNet512-VGG16-*F* and FRS have difficulty in detecting underwater objects. In contrast, DRNet512-VGG16-*G* and GAN-RS have higher recall rate and detection precision in this real-world task. Because the same detection method and training data are used, the huge performance gap should be caused by the training domain.

3.4.2 Online domain analysis

As shown in Fig. 7, there is a huge discrepancy between the online domain and domain-*O*. Thus, DRN512-VGG16-*O* suffers from serious degradation of detection accuracy. Domain shift is moderated by

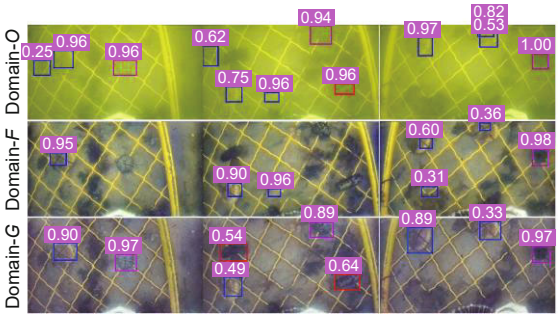

Fig. 6 Demonstration of online detection. DRNet512-VGG16-*O* and DRNet512-VGG16-*F* are hardly qualified for this online detection. By suppressing the problem of domain shift, DRNet512-VGG16-*G* and GAN-RS perform better in this field underwater scene. "Trepang," "echinus," and "shell" are detected in red, purple, and blue boxes, respectively. Confidence scores are presented on the top of boxes. GAN-RS: GAN-based restoration (References to color refer to the online version of this figure)

FRS, but FRS is not sufficient to preserve detection performance in this scenario. On the contrary, GAN-RS has higher restoration intensity. As a result, processed by GAN-RS, online domain and domain-*G* are highly overlapped as illustrated in Fig. 7. Therefore, DRN512-VGG16-*G* and GAN-RS are able to perform this detection task well. It can be seen that the problem of domain shift is gradually solved with increasing restoration intensity. In addition, underwater scene domains are manifold (Fig. 2), so domain-diverse data collection is unattainable. Therefore, contributing to domain shift suppression, visual restoration is essential for object detection in underwater environments.

In brief, visual restoration essentially changes the domain of data. When training data and test data have the same domain, visual restoration has an ignorable effect. However, for real-world marine operations, optical data collected in real time by the robot must have different domains from the training data because of varying degrees of degradation. The resulting domain shift will cause great damage to the detection accuracy, but visual restoration can effectively suppress this problem. Therefore, visual restoration plays a crucial role in real-time underwater visual perception.

4 Detection continuity/stability and underwater robotic visual perception framework

4.1 Non-reference assessments

Considering that the data collected by robots in practical application scenarios usually lack annotations, we propose non-reference assessments that rely on MOT rather than ground-truth labels. Our assessments follow a reasonable assumption: object motion is smooth across time without high-frequency location jitter or change of existence. We leverage a detector and an MOT module to recall all object tracklets in a video. Specifically, any detector to be evaluated can be used for VID, and we employ the IoU-based MOT tracker reported by Chen XY et al. (2020) to associate detected boxes. Unlike label-based evaluation, we focus only on detected tracklets, because a totally missed tracklet does not impact continuity or stability.

As described in Fig. 8, a detector locates and classifies objects at each frame f. Each object has confidence score (s), box center (c_x, c_y), and size (w, h). N tracklets $\{\mathcal{T}_n | n = 1, 2, ..., N\}$ are produced after the whole video is processed by VID and MOT. The video duration and tracklet duration are denoted as t_v and t_n, respectively. The vertical axis of Fig. 8 describes c_x, c_y, w, or h.

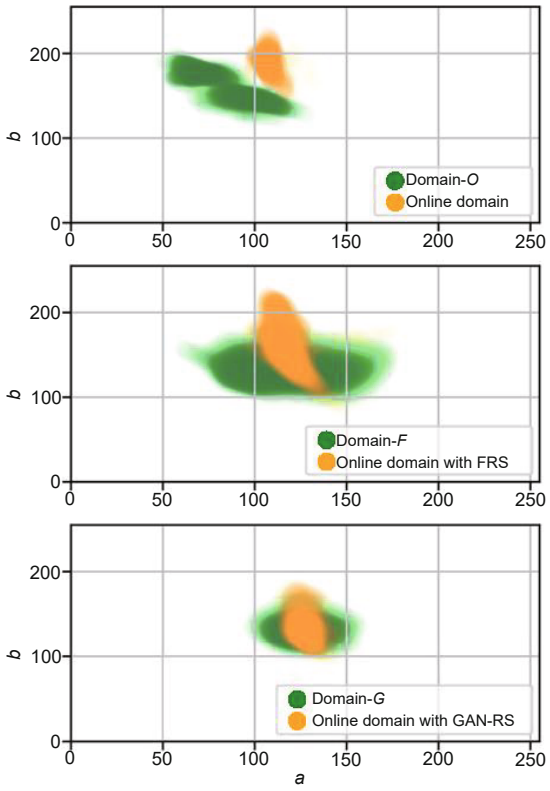

Fig. 7 Comparison of online domain and training domains in the Lab color space. Color transparency indicates the distribution probability (References to color refer to the online version of this figure)

4.1.1 Recall continuity

As for recall continuity, we consider the impact of short tracklet duration and tracklet fragments. Referring to Fig. 1a and $\mathcal{T}_2$ in Fig. 8, tracklets with short duration frequently appear in VID. To capture them, we design an extremely short duration error (ESDE) and short duration error (SDE) with various duration thresholds as

$$\mathrm{ESDE\ (SDE)} = \frac{1}{t_\mathrm{v}} \sum_{n=1}^{N} t_n, \tag{1}$$

where $t_n = 0$ if $t_n \geq S$ (S represents the duration threshold). In this study, $S = 3$ for ESDE and $S = 10$ for SDE, which describe different degrees of the short duration problem.

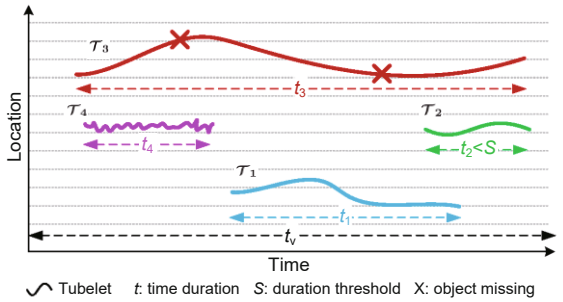

Fig. 8 Problem formation: object tracklets could suffer from short tracklet duration (e.g., $\mathcal{T}_2$), fragments (e.g., $\mathcal{T}_3$), and location jitter (e.g., $\mathcal{T}_4$)

Fig. 1a and $\mathcal{T}_3$ in Fig. 8 represent the tracklet fragment problem. Some MOT algorithms end a tracklet after a recall failure. Conversely, we count the number of continuous recall failures with S_{lost}, and leave an $S_{\text{lost}}^{\max}$-frame life duration for each tracklet. That is, if a recall-failed tracklet is re-matched by a box in the consequent $S_{\text{lost}}^{\max}$ frames, the tracklet can be retrained. In this way, the total number of missing objects (mo) in the whole tracklet can be captured, forming a tracklet fragment error (TFE) and fragmental tracklet ratio (FTR) as follows:

$$\begin{aligned}\text{TFE} &= \sum_{n=1}^{N} \text{mo}_n \Big/ \sum_{n=1}^{N} t_n,\\ \text{FTR} &= \frac{1}{N}\sum_{n=1}^{N} \text{ft}_n,\end{aligned} \tag{2}$$

where mo_n is the number of missing objects in the n^{th} tracklet and ft_n represents whether there are missing objects in the n^{th} tracklet. Here, if $\text{mo}_n > 0$, $\text{ft}_n = 1$; otherwise, $\text{ft}_n = 0$.

TFE describes the ratio of missing objects to all objects in tracklets, while FTR gives the ratio of faulty tracklets to all tracklets. They are complementary for the tracklet fragment problem; that is, a better VID result should have lower TFE and FTR in the meantime. That is, there is a small number of missing objects, and the missing objects are concentrated in a small number of tracklets. Note that the amount of calculation is numerically small, so a log transformation is used to enhance the contrast, i.e., $\log_{100}(1 + 99\alpha)$, where α represents ESDE, SDE, TFE, or FTR. Finally, the recall continuity error (RCE) is defined as RCE = ESDE + SDE + TFE + FTR.

4.1.2 Localization stability

Object tracklets should be smooth in localization, and box center/size jitter causes damage to localization stability (Fig. 1b and $\mathcal{T}_4$ in Fig. 8). We evaluate temporal stability in the Fourier domain so that our approach could work without labels. Time-domain data p can be transformed into the Fourier domain by $P = \mathcal{F}(p)$, where p represents c_x, c_y, w, or h. Thus, P contains frequency information of p, and we extract frequency-related amplitude with $\widetilde{P} = \text{Abs}(P)$ ($\text{Abs}(\cdot)$ denotes the magnitude of a complex number). Note that each tracklet produces different frequency components because of variable data length (i.e., tracklet duration). That is, $\widetilde{P} = \{(q_k^p, A_k^p)\}, q_k^p = k/t, k = 0, 1, ..., \lfloor t/2 \rfloor$ ($\lfloor\cdot\rfloor$ denotes the rounding down operation). Here, q is the frequency set, t is the tracklet duration, and A denotes the frequency-related amplitude. Based on the Fourier analysis, the center jitter error (CJE) and size jitter error (SJE) are designed as

$$\begin{aligned}\text{CJE} &= \left(\sum_{n=1}^{N}\sum_{p\in\{c_x,c_y\}}\sum_{k=1}^{\lfloor t_n/2\rfloor} q_{n,k}^p A_{n,k}^p\right) \Big/ \sum_{n=1}^{N} t_n,\\ \text{SJE} &= \left(\sum_{n=1}^{N}\sum_{p\in\{w,h\}}\sum_{k=1}^{\lfloor t_n/2\rfloor} q_{n,k}^p A_{n,k}^p\right) \Big/ \sum_{n=1}^{N} t_n.\end{aligned} \tag{3}$$

Ultimately, localization jitter error LJE = CJE + SJE.

4.2 Online tracklet refinement

To enhance recall continuity and localization stability, we refine VID results based on tracklets. A new attribute is used to describe tracklets: current duration S_{dur}. Therefore, a tracklet can be formulated as $\mathcal{T} = \{\mathcal{D}, \text{ID}, S_{\text{lost}}, S_{\text{dur}}\}$, where S_{dur} records tracklet duration at each timestamp, S_{lost} was explained in Section 4.1, ID denotes the tracklet identity, $\mathcal{D}$ is the object set in the tracklet (i.e., $\{(s, c_x, c_y, w, h)\}$), and the length of $\mathcal{D}$ (i.e., S_{obj}) cannot exceed $S_{\text{obj}}^{\max} = 5$. That is, if $S_{\text{dur}} > S_{\text{obj}}^{\max}$, only the latest $S_{\text{obj}}^{\max}$ objects are preserved in $\mathcal{D}$.

4.2.1 Short tracklet suppression

To suppress short tracklets and enhance ESDE (SDE), we define a tracklet as a reliable tracklet if $S_{\text{dur}} > S_{\text{SDE}}$, and then boxes in unreliable tracklets are suppressed. This method is beneficial to continuity, and it has two-fold effects on accuracy. First, false positives can be suppressed, because their recall is usually non-consecutive across time, so a reliable tracklet is hard to form. Second, false negatives can be produced because an object is not reported until it forms an S_{SDE}-length tracklet.

4.2.2 Fragment filling

In terms of the fragment issue and TFE (FTR), we imagine a missing object in a tracklet based on a reasonable assumption; i.e., the object motion is uniform in an extremely short duration (e.g., $S_{\text{obj}}^{\max}$). When a tracklet suffers from a

recall failure at the f^{th} frame, its previous boxes $\left\{\left(c_x^{f-i}, c_y^{f-i}, w^{f-i}, h^{f-i}\right)|i=1,2,...,S_{\text{obj}}\right\}$ can be used to predict its current location. In detail, we first estimate the velocity v_p, i.e., $v_p = \sum_{i=1}^{S_{\text{obj}}-1}\left(p^{f-i}-p^{f-i-1}\right)\Big/(S_{\text{obj}}-1)$, and then the current location can be given as $p = p^{f-1}+v_p$, where p denotes c_x, c_y, w, or h.

4.2.3 Temporal location fusion

For location stability and CJE (SJE), we add the object to its tracklet, and then produce a new location by merging $\{p|p \in \mathcal{D},\ \mathcal{D} \in \mathcal{T}\}$. We try four filters, i.e., median, mean, weighted mean, and Kalman filters. Median and mean filters use the median and mean of the last S_{obj} locations as the current location, respectively. As for the weighted mean filter, a geometric progression is contrasted with $\Omega = \{\omega^l|l = 1,...,0.1\}$, where l is an S_{obj}-length arithmetic progression. The normalized Ω is used as the weights of the weighted mean filter, and the updated location can be formulated as $\hat{p}^f = \sum_{i=0}^{S_{\text{obj}}} \Omega_i p^{f-i}$. A Kalman filter (Kalman, 1960) can fuse predicted and observed values with their calculated weighted coefficients. A filter that has better temporal information fusion should have high localization stability, which can also verify the validity of the proposed LJE.

4.3 Robotic visual perception framework

4.3.1 Small-overlap suppression

We promote a VID model to generate the SOT result by propagating the previous location $b^{f-1} = (c_x^{f-1}, c_y^{f-1}, w^{f-1}, h^{f-1})$ before non-maximum suppression (NMS). Taking inspiration from NMS, we leverage IoU-based suppression to this end. Referring to Algorithm 1, after selection by the confidence threshold, the IoU between candidate boxes and b^{f-1} is calculated. Then candidate boxes with small IoUs (e.g., $< U^{\text{sos}}$) are discarded. Next, a tracking failure would be reported if all boxes are suppressed by SOS. Subsequently, NMS is performed on the remaining boxes. Finally, we select a box with the maximum IoU as the current SOT result b^f. Compared with the method of IoU-based re-scoring, the SOS does not affect confidence scores. In our opinion, the confidence score and IoU are two different properties of objects that describe the object category and object motion, respectively. Therefore, SOS-NMS is based on alternating the confidence score and IoU, i.e., (1) discarding obviously incorrect candidates with the confidence and IoU threshold, (2) discarding candidates without the maximum local confidence score, and (3) generating a single-object location with the maximum IoU. Note that SOS-NMS has a speed advantage over NMS because a significant number of candidate boxes are suppressed by computationally efficient SOS.

Algorithm 1 SOS-NMS

Input: After selection by the confidence threshold, boxes $\mathcal{B} = \{b_1, b_2, ..., b_m\}$, confidence scores $\mathcal{S} = \{s_1, s_2, ..., s_m\}$; previous tracked box b^{f-1}; SOS threshold U^{sos}; NMS threshold U^{nms} // SOS based on IoU; // $\cup$ and $\setminus$ denote element addition and element removal, // respectively

Output: Tracked box b^f

$\mathcal{B}^{\text{sos}} = \mathcal{B}; \mathcal{S}^{\text{sos}} = \mathcal{S}; \mathcal{O}^{\text{sos}} = \text{iou}(b^{f-1}, \mathcal{B})$
while $(b_i, s_i, o_i) \in (\mathcal{B}^{\text{sos}}, \mathcal{S}^{\text{sos}}, \text{IoU}^{\text{sos}})$ **do**
  if $o_i < U^{\text{sos}}$ **then**
    $\mathcal{B}^{\text{sos}} \setminus b_i; \mathcal{S}^{\text{sos}} \setminus s_i; \mathcal{O}^{\text{sos}} \setminus o_i$
  end if
end while // Inspection of the tracking failure
if $\mathcal{B}^{\text{sos}} =$ empty **then**
  return $b^f =$ empty
end if // NMS based on the confidence score
$\mathcal{B}^{\text{nms}} = \{\}; \mathcal{S}^{\text{nms}} = \{\}; \mathcal{O}^{\text{nms}} = \{\}$
while $\mathcal{B}^{\text{sos}} \neq$ empty **do**
  $\text{idx} = \text{argmax}\,\mathcal{S}^{\text{sos}}$
  $b = \mathcal{B}^{\text{sos}}_{\text{idx}}; s = \mathcal{S}^{\text{sos}}_{\text{idx}}; o = \mathcal{O}^{\text{sos}}_{\text{idx}}$
  $\mathcal{B}^{\text{nms}} \cup \{b\}; \mathcal{S}^{\text{nms}} \cup \{s\}; \mathcal{O}^{\text{nms}} \cup \{o\}$
  $\mathcal{B}^{\text{sos}} \setminus b; \mathcal{S}^{\text{sos}} \setminus s; \mathcal{O}^{\text{sos}} \setminus o$
  while $(b_i, s_i) \in (\mathcal{B}^{\text{sos}}, \mathcal{S}^{\text{sos}})$ **do**
    if $\text{iou}(b, b_i) > U^{\text{nms}}$ **then**
      $\mathcal{B}^{\text{sos}} \setminus b_i; \mathcal{S}^{\text{sos}} \setminus s_i; \mathcal{O}^{\text{sos}} \setminus o_i$
    end if
  end while
end while // Selection of a single box with IoU
$\text{idx} = \text{argmax}\mathcal{O}^{\text{nms}}$
return $b^f = \mathcal{B}^{\text{nms}}_{\text{idx}}$

4.3.2 Overall framework with SOT-by-detection

As shown in Fig. 9, the proposed robotic visual perception framework adopts only a visual restoration model and a detection model. We apply GAN-RS (Chen XY et al., 2019b) to obtain restored images and use them as detection model inputs, which can suppress domain shift. The MOT branch and SOT-by-detection branch can identify the initial object location and track the object to be grasped. We first define the condition of the MOT-SOT switch: (1) MOT is initially performed; (2) When reliable tracklets (i.e., $S_{\text{dur}} > S_{\text{SDE}}$ captured by OTR) are

found, the SOT-by-detection branch is activated to track the tracklet with the highest confidence score; (3) The MOT branch is re-activated after an SOT failure is captured by SOS. Note that for SOT-by-detection, OTR processes only the tracked tracklet. Remarkably, the SOT-by-detection branch is faster than the MOT branch because (1) SOS is able to significantly reduce the NMS computational cost and (2) data association in the MOT branch is usually time-consuming. The proposed robotic visual perception framework has two-fold advantages in object search and grasping: (1) there is no need for a man-determined initial location and (2) a complex detection-tracking cascade is avoided. Benefiting from the above-mentioned design, the proposed framework can achieve high underwater object detection precision and robust object tracking, and provide a good foundation for underwater robot operation.

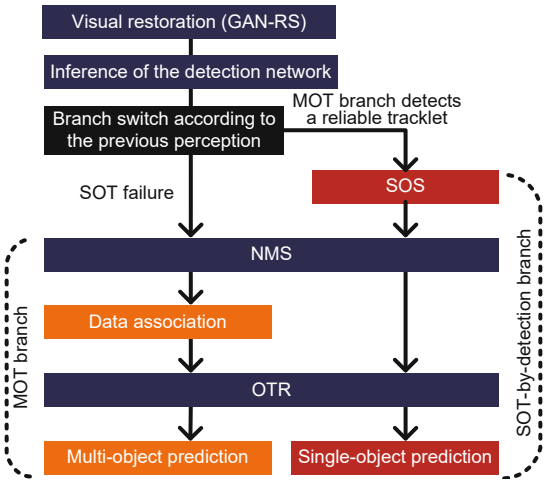

Fig. 9 Robotic visual perception framework with our proposed SOS and OTR. The MOT branch is designed to search the initial box for SOT-by-detection. A reliable tracklet is captured by OTR, while the SOT failure is captured by SOS. If switch conditions are not met, the previous behavior is continuously performed. GAN-RS: GAN-based restoration; MOT: multi-object tracking; NMS: non-maximum suppression; OTR: online tracklet refinement; SOT: single-object tracking; SOS: small-overlap suppression

5 Experiments and analyses

5.1 Validation of non-reference assessments

5.1.1 Preliminaries

We analyzed real-time online detectors, i.e., SSD (Liu W et al., 2016), RetinaNet (Lin et al., 2017), RefineDet (Zhang et al., 2018), DRNet (Chen XY et al., 2019a), temporal SSD (TSSD) (Chen XY et al., 2020), temporal refinement networks (TRNet), and temporal dual refinement networks (TDRNet) (Chen XY et al., 2021). The first four are static detectors, whereas the last three are temporal methods. These detectors are closely related and represent an evolutionary path, so we present the effects of their designs on temporal performance based on our metrics; these evaluations can verify the effectiveness of our assessments. SSD detects objects in a single-stage manner (Liu W et al., 2016). Based on SSD, RetinaNet adopts feature pyramid networks (FPNs) to enhance a shadow-layer receptive field (Lin et al., 2017). Based on RetinaNet, RefineDet introduces a two-step regression to the single-stage pipeline (Zhang et al., 2018). Based on RefineDet, DRNet performs joint anchor-feature refinement for detection (Chen XY et al., 2019a). Referring to Section 2, there are five types of VID approaches, but post-processing and tracking-based methods actually adopt static detectors, and batch-frame approaches can hardly work in real-world scenes, so we analyzed the methods with feature aggregation or temporally sustained proposal. Based on SSD, TSSD uses attentional long short-term memory (LSTM) to aggregate visual features across time (Chen XY et al., 2020). As temporally sustained proposal approaches, TRNet and TDRNet propagate refined anchors and feature offsets across time based on RefineDet and DRNet respectively (Chen XY et al., 2021). All these detectors were trained and evaluated on the ImageNet VID dataset (Russakovsky et al., 2015). Both the confidence threshold and NMS threshold were fixed at 0.5. Ultimately, the merit of SOT-by-detection with TDRNet as the detector was verified in a real-world robotic grasping task. All training and tests were conducted on eight NVIDIA GeForce RTX 1080Ti GPUs.

5.1.2 Tracklet visualization

As shown in Fig. 10a, we used a VID case with nine object instances to visualize SSD detection. Referring to Fig. 10b, SSD suffered from serious continuity and stability problems. At the beginning of this video, a vast number of missing objects (i.e., "×" on curves) and short tracklets (i.e., short curves and scattered points) appeared due to motion blur.

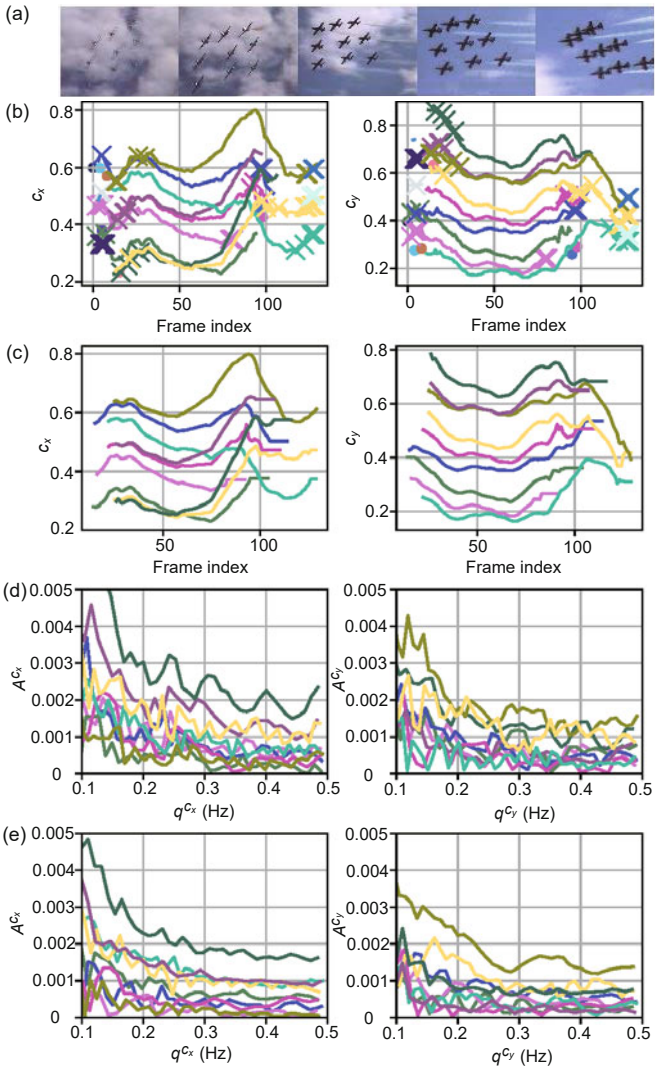

Fig. 10 Tracklet visualization for SSD: (a) video snippet; (b) original tracklets; (c) refined tracklets with OTR; (d) original Fourier results; (e) Fourier results with OTR. Colors differentiate ID and "×" denotes missing objects at a time-stamp. A denotes the frequency-related amplitude and q is the frequency set. OTR: online tracklet refinement (References to color refer to the online version of this figure)

Then, continuity problems appeared again at the end of the video because of occlusion. For localization stability, Fig. 10d plots the amplitude of high-frequency components (>0.1 Hz) in the Fourier domain. Numerically, ESDE = 0.448, SDE = 0.715, TFE=0.410, FTR = 0.619, RCE = 2.192, CJE = 0.264, SJE = 0.183, LJE = 0.447.

OTR is able to refine the SSD results from the tracklet perspective. As shown in Fig. 10c, OTR eliminated all short tracklets and fragments, and the refined tracklets were smoother. Referring to Fig. 10e, the high-frequency amplitude in the Fourier domain was suppressed to some degree. As a result, ESDE = 0, SDE = 0, TFE = 0, FTR = 0, RCE = 0, CJE = 0.219, SJE = 0.142, LJE = 0.361.

5.1.3 Validation of LJE

To verify our LJE, we applied four filters mentioned in Section 4.2 in temporal location fusion based on TDRNet. The mean filter fuses temporal information, so it should work better than the median filter that selects only single-moment location information. With weights set manually, the weighted mean filter can better fuse the temporal information than the mean filter. The Kalman filter fuses predicted and measured values using calculated

weights, so it should have the best LJE among these filters. The results are shown in Table 6. With short tracklet suppression and fragment filling to generate more valid tracklets, the LJE of "None" (Table 6) was better than the OTR-free value (TDRNet) in Table 7. In addition, the lower LJE represents higher localization stability, and the experimental results agreed with the theoretical analysis of the four filters used here. This proved the effectiveness of our proposed LJE. Note that we did not verify the RCE, because it is a statistic metric.

5.1.4 VID evaluation of continuity and stability

Detectors were evaluated with our proposed non-reference assessments on the VID validation set (Table 7). The results showed that there was a low correlation between accuracy and continuity/stability. From static SSD (without FPN) and RetinaNet (with FPN), we observed that FPN improved localization stability because of spatial feature fusion, but reduced continuity because it can detect hard objects (e.g., small objects) that easily produce continuity problems. In addition, using anchor refinement, RefineDet had a higher accuracy but lower continuity and stability than RetinaNet because anchor-feature mis-alignment was deteriorated. Finally, DRNet conducted joint anchor-feature refinement to relieve RefineDet's drawback; i.e., features were relatively accurate in describing refined anchors. Thus, DRNet performed better than RefineDet on almost all metrics. Based on OTR, all metrics can be effectively improved for all tested approaches. Because different thresholds (i.e., 0.5 vs. 0.01) were used for MOT and AP evaluations, AP cannot be reported with OTR.

For temporal approaches, we draw readers' attention to three detector pairs, i.e., TRNet vs. RefineDet, TDRNet vs. DRNet, and TSSD vs. SSD, where the temporal detector is extended from the static detector. For the first two comparisons, temporal detectors achieved performance that is on par with, and sometimes even worse than those of static approaches. Therefore, the design of temporally sustained proposal has an ignorable effect on detection continuity and stability. In contrast, TSSD performed better than SSD by a large margin on all metrics, which validated the effectiveness of temporal feature aggregation. That is, TSSD can smooth visual features across time, and produces more temporally consistent results.

Table 6 Stability evaluation of TDRNet with different filters based on the proposed non-reference metrics

Filter	Localization stability		
	CJE	SJE	LJE
None	0.200	0.291	0.491
Median	0.194	0.259	0.453
Mean	0.178	0.224	0.402
Weighted mean	0.170	0.218	0.388
Kalman	0.153	0.187	0.340

CJE: center jitter error; SJE: size jitter error; LJE: localization jitter error

5.2 Robotic object search and grasping tasks

5.2.1 Object search and grasping

This framework has been applied to a remotely operated vehicle (ROV) and fully autonomous grasping was achieved. The test venue was a five-meter-deep natural unstructured seabed, located at Jinshitan, Dalian, China. As shown in Fig. 11, an ROV with a soft robotic arm was developed for grasping marine products (e.g., echinus and shell). For better integration, we used NVIDIA Xavier to implement our visual framework, which was placed inside the ROV. On NVIDIA Xavier, our framework can run at about 16 frames/s, meeting the requirements for autonomous grasping. Based on our analysis that visual restoration can effectively alleviate the domain shift, restored images were used as inputs for detection and tracking. Specifically, the MOT branch operated first for object search, that is, perception of an object group. After a reliable tracklet was detected, the SOT-by-detection branch operated for detailed perception of an object instance. Then the ROV was adjusted to put the target in ROV's grasping area and began to grasp. In this task, the proposed framework was competent in detecting and tracking objects for robotic perception and provided flexible perception ability for object groups and instances. Visual restoration can improve the precision of underwater object detection and the proposed OTR can make tracking more robust, thus obtaining consistent and stable localization results. Our efficient SOS further improved the real-time performance of our framework. Based on this framework, our robot was able to efficiently approach and grasp targets. Please see Video S1 in supplementary materials for the real-world experiment.

Table 7 Continuity and stability evaluation of several existing detectors based on the proposed non-reference metrics

Method	mAP	Recall continuity					Localization stability		
		ESDE	SDE	TFE	FTR	RCE	CJE	SJE	LJE
w/o OTR static method									
SSD (Liu W et al., 2016)	0.630	0.062	0.234	0.320	0.246	0.862	0.242	0.334	0.576
RetinaNet (Lin et al., 2017)	0.656	0.060	0.250	0.350	0.283	0.943	0.236	0.317	0.553
RefineDet (Zhang et al., 2018)	0.669	0.126	0.350	0.391	0.306	1.173	0.257	0.362	0.619
DRNet (Chen XY et al., 2019a)	**0.694**	0.114	0.330	0.389	0.312	1.145	0.248	0.346	0.594
w/o OTR temporal method									
TRNet (Chen XY et al., 2021)	0.665	0.120	0.334	0.375	0.265	1.094	0.252	0.346	0.598
TDRNet (Chen XY et al., 2021)	0.673	0.116	0.345	0.388	0.297	1.146	0.247	0.360	0.607
TSSD (Chen XY et al., 2020)	0.654	**0.059**	**0.206**	**0.257**	**0.240**	**0.762**	**0.210**	**0.253**	**0.463**
w/ OTR (weighted mean)									
SSD	–	0.003	0.026	0.0	0.0	0.029	0.169	0.208	0.377
RetinaNet	–	0.003	0.023	0.0	0.0	0.026	0.168	0.204	0.372
RefineDet	–	0.004	0.037	0.0	0.0	0.041	0.173	0.212	0.385
DRNet	–	0.003	0.036	0.0	0.0	0.039	0.172	0.208	0.380
TRNet	–	0.003	0.030	0.0	0.0	0.033	0.171	0.209	0.380
TDRNet	–	0.004	0.031	0.0	0.0	0.035	0.170	0.218	0.388
TSSD	–	0.003	0.029	0.0	0.0	0.032	0.159	0.180	0.339

The best results are in bold (for only w/o OTR methods). mAP: mean average precision; ESDE: extremely short duration error; SDE: short duration error; TFE: tracklet fragment error; FTR: fragmental tracklet ratio; RCE: recall continuity error; CJE: center jitter error; SJE: size jitter error; LJE: localization jitter error; OTR: online tracklet refinement

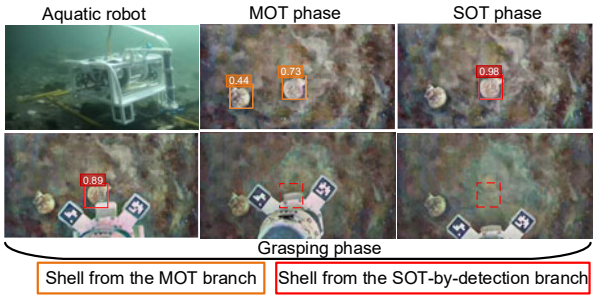

Fig. 11 Our robotic visual perception framework in the object grasping task. The proposed framework can provide robust visual information with visual restoration and flexible detection and tracking for robotic search and grasping

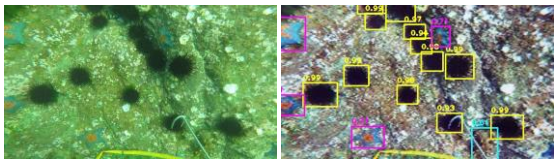

Fig. 12 Underwater detection of intensive objects: (a) raw underwater image; (b) restored image and detection results

5.2.2 Detection of intensive objects

To further verify the effectiveness of our visual framework, we conducted the experiment of detecting intensive objects. The results are shown in Fig. 12. It can be seen that although the objects in the image are intensive, our visual framework can still detect all objects accurately, suggesting the superiority of our framework.

5.3 Discussion

This paper examined domain-related detection learning phenomena. Detection continuity and stability were also evaluated and enhanced from the MOT perspective, making the underwater search and tracking more robust. Further, we discuss the following points to inspire future works:

1. Recall efficiency. In within-domain tests, a higher-quality domain induces lower detection performance, because of low recall efficiency. Thus, a high-quality domain incurs more false positives. However, object candidates that can bring about false positives exist in both the training and testing phases. We can see that in these conditions, the learning of these candidates is insufficient. Therefore, we advocate further research on how these candidates separately impact training and inference for exploring more efficient learning methods.

2. CNN domain selectivity. In cross-domain training, low-quality samples are less effective, so

accuracy drops on the test set. We saw that the CNN learning is characterized by domain selectivity. That is, sample contributions are different in CNN-based detection learning. Therefore, we advocate further research into CNN domain selectivity for building more robust real-world detectors.

3. Detection continuity and stability. Our evaluation indicates that temporal performance benefits from feature aggregation. On one hand, spatial smoothing and scale smoothing are effective (see RetinaNet vs. SSD); on the other hand, temporal fusion is more efficient (TSSD vs. SSD). Thus, we advocate investigating fusion approaches for improving continuity and stability of the detector itself. For example, Bertasius et al. (2018) leveraged deformable convolutions to construct robust temporal features.

6 Conclusions and future work

In this paper, we performed domain analysis and revealed how visual restoration contributes to object detection in aquatic scenes. In addition, object detection recall continuity and localization stability have been analyzed in a novel way for robotic perception. Finally, an underwater robotic visual perception framework has been proposed for underwater object search and grasping. We presented original viewpoints and proposed novel methods as follows: (1) Although visual restoration has an ignorable effect on within-domain convolutional representation and detection accuracy, it is essential in online robotic perception because it can suppress domain shift to improve ROV detection accuracy in real-world sea scenarios. (2) New non-reference assessments have been proposed to reflect temporal continuity and stability, and OTR has been designed to improve recall continuity and localization stability. (3) SOS has been proposed to extend VID methods to SOT tasks, and the robotic visual perception framework with SOT-by-detection has been developed. As a result, our conclusions and methods have been verified on datasets, and underwater autonomous object search and grasping have been achieved in a real sea area.

In the future, we will enhance temporal performance with feature fusion and improve SOT-by-detection using online learning.

Contributors

Yue LU designed the research. Yue LU, Xingyu CHEN, and Li WEN proposed the methods. Yue LU conducted the experiments. Junzhi YU processed the data. Zhengxing WU participated in the visualization. Yue LU drafted the paper. Zhengxing WU and Li WEN helped organize the paper. Xingyu CHEN and Junzhi YU revised and finalized the paper.

Compliance with ethics guidelines

Yue LU, Xingyu CHEN, Zhengxing WU, Junzhi YU, and Li WEN declare that they have no conflict of interest.

References

Bernardin K, Stiefelhagen R, 2008. Evaluating multiple object tracking performance: the clear MOT metrics. *EURASIP J Image Video Process*, 2008:246309.

Bertasius G, Torresani L, Shi JB, 2018. Object detection in video with spatiotemporal sampling networks. Proc 15th European Conf on Computer Vision, p.342-357. https://doi.org/10.1007/978-3-030-01258-8_21

Cai MX, Wang Y, Wang S, et al., 2020. Grasping marine products with hybrid-driven underwater vehicle-manipulator system. *IEEE Trans Autom Sci Eng*, 17(3):1443-1454. https://doi.org/10.1109/TASE.2019.2957782

Chen XY, Yang XY, Kong SH, et al., 2019a. Dual refinement network for single-shot object detection. Proc Int Conf on Robotics and Automation, p.8305-8310. https://doi.org/10.1109/ICRA.2019.8793816

Chen XY, Yu JZ, Kong SH, et al., 2019b. Towards real-time advancement of underwater visual quality with GAN. *IEEE Trans Ind Electron*, 66(12):9350-9359. https://doi.org/10.1109/TIE.2019.2893840

Chen XY, Yu JZ, Wu ZX, 2020. Temporally identity-aware SSD with attentional LSTM. *IEEE Trans Cybern*, 50(6):2674-2686. https://doi.org/10.1109/TCYB.2019.2894261

Chen XY, Yu JZ, Kong SH, et al., 2021. Joint anchor-feature refinement for real-time accurate object detection in images and videos. *IEEE Trans Circ Syst Video Technol*, 31(2):594-607. https://doi.org/10.4324/9781003144281-4

Chen YH, Li W, Sakaridis C, et al., 2018. Domain adaptive faster R-CNN for object detection in the wild. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.3339-3348. https://doi.org/10.1109/CVPR.2018.00352

Chi C, Zhang SF, Xing JL, et al., 2019. Selective refinement network for high performance face detection. Proc AAAI Conf on Artificial Intelligence, p.8231-8238. https://doi.org/10.1609/aaai.v33i01.33018231

Everingham M, van Gool L, Williams CKI, et al., 2010. The PASCAL visual object classes (VOC) challenge. *Int J Comput Vis*, 88(2):303-338. https://doi.org/10.1007/s11263-009-0275-4

Feichtenhofer C, Pinz A, Zisserman A, 2017. Detect to track and track to detect. Proc IEEE Int Conf on Computer Vision, p.3057-3065. https://doi.org/10.1109/ICCV.2017.330

Gong ZY, Cheng JH, Chen XY, et al., 2018. A bio-inspired soft robotic arm: kinematic modeling and hydrodynamic experiments. *J Bion Eng*, 15(2):204-219. https://doi.org/10.1007/s42235-018-0016-x

He KM, Zhang XY, Ren SQ, et al., 2016. Deep residual learning for image recognition. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.770-778. https://doi.org/10.1109/CVPR.2016.90

Howard AG, Zhu ML, Chen B, et al., 2017. MobileNets: efficient convolutional neural networks for mobile vision applications. https://arxiv.org/abs/1704.04861

Inoue N, Furuta R, Yamasaki T, et al., 2018. Cross-domain weakly-supervised object detection through progressive domain adaptation. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.5001-5009. https://doi.org/10.1109/CVPR.2018.00525

Kalman RE, 1960. A new approach to linear filtering and prediction problems. *J Bas Eng*, 82(1):35-45. https://doi.org/10.1115/1.3662552

Kalogeiton V, Ferrari V, Schmid C, 2016. Analysing domain shift factors between videos and images for object detection. *IEEE Trans Patt Anal Mach Intell*, 38(11):2327-2334. https://doi.org/10.1109/TPAMI.2016.2551239

Kang K, Li HS, Yan JJ, et al., 2018. T-CNN: tubelets with convolutional neural networks for object detection from videos. *IEEE Trans Circ Syst Video Technol*, 28(10):2896-2907. https://doi.org/10.1109/TCSVT.2017.2736553

Khodabandeh M, Vahdat A, Ranjbar M, et al., 2019. A robust learning approach to domain adaptive object detection. Proc IEEE/CVF Int Conf on Computer Vision, p.480-490. https://doi.org/10.1109/ICCV.2019.00057

Kim HU, Kim CS, 2016. CDT: cooperative detection and tracking for tracing multiple objects in video sequences. Proc 14th European Conf on Computer Vision, p.851-867. https://doi.org/10.1007/978-3-319-46466-4_51

Kristan M, Leonardis A, Matas J, et al., 2018. The sixth visual object tracking VOT2018 challenge results. Proc European Conf on Computer Vision, p.3-53. https://doi.org/10.1007/978-3-030-11009-3_1

Li B, Xu YX, Fan SS, et al., 2018. Underwater docking of an under-actuated autonomous underwater vehicle: system design and control implementation. *Front Inform Technol Electron Eng*, 19(8):1024-1041. https://doi.org/10.1631/FITEE.1700382

Li CY, Guo JC, Cong RM, et al., 2016. Underwater image enhancement by dehazing with minimum information loss and histogram distribution prior. *IEEE Trans Image Process*, 25(12):5664-5677. https://doi.org/10.1109/tip.2016.2612882

Lin TY, Goyal P, Girshick R, et al., 2017. Focal loss for dense object detection. Proc IEEE Int Conf on Computer Vision, p.2999-3007. https://doi.org/10.1109/ICCV.2017.324

Liu RS, Fan X, Zhu M, et al., 2020. Real-world underwater enhancement: challenges, benchmarks, and solutions under natural light. *IEEE Trans Circ Syst Video Technol*, 30(12):4861-4875. https://doi.org/10.1109/TCSVT.2019.2963772

Liu W, Anguelov D, Erhan D, et al., 2016. SSD: single shot multibox detector. Proc 14th European Conf on Computer Vision, p.21-37. https://doi.org/10.1007/978-3-319-46448-0_2

Lowe DG, 2004. Distinctive image features from scale-invariant keypoints. *Int J Comput Vis*, 60(2):91-110. https://doi.org/10.1023/B:VISI.0000029664.99615.94

Luo H, Xie WX, Wang XG, et al., 2019. Detect or track: towards cost-effective video object detection/tracking. Proc AAAI Conf on Artificial Intelligence, p.8803-8810. https://doi.org/10.1609/aaai.v33i01.33018803

Panetta K, Gao C, Agaian S, 2016. Human-visual-system-inspired underwater image quality measures. *IEEE J Ocean Eng*, 41(3):541-551. https://doi.org/10.1109/JOE.2015.2469915

Raj A, Namboodiri VP, Tuytelaars T, 2015. Subspace alignment based domain adaptation for RCNN detector. Proc British Machine Vision Conf, p.166.1-166.11.

Russakovsky O, Deng J, Su H, et al., 2015. ImageNet large scale visual recognition challenge. *Int J Comput Vis*, 115(3):211-252. https://doi.org/10.1007/s11263-015-0816-y

Schechner YY, Karpel N, 2004. Clear underwater vision. Proc IEEE Computer Society Conf on Computer Vision and Pattern Recognition, p.536-543. https://doi.org/10.1109/CVPR.2004.1315078

Simonyan K, Zisserman A, 2014. Very deep convolutional networks for large-scale image recognition. https://arxiv.org/abs/1409.1556

Xu JL, Ramos S, Vázquez D, et al., 2014. Domain adaptation of deformable part-based models. *IEEE Trans Patt Anal Mach Intell*, 36(12):2367-2380. https://doi.org/10.1109/TPAMI.2014.2327973

Yang M, Sowmya A, 2015. An underwater color image quality evaluation metric. *IEEE Trans Image Process*, 24(12):6062-6071. https://doi.org/10.1109/TIP.2015.2491020

Zhang SF, Wen LY, Bian X, et al., 2018. Single-shot refinement neural network for object detection. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.4203-4212. https://doi.org/10.1109/CVPR.2018.00442

Zhou XY, Zhuo JC, Krähenbühl P, 2019. Bottom-up object detection by grouping extreme and center points. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.850-859. https://doi.org/10.1109/CVPR.2019.00094

Zhu DQ, Qu Y, Yang SX, 2019. Multi-AUV SOM task allocation algorithm considering initial orientation and ocean current environment. *Front Inform Technol Electron Eng*, 20(3):330-341. https://doi.org/10.1631/FITEE.1800562

Zhu YS, Zhao CY, Guo HY, et al., 2019. Attention CoupleNet: fully convolutional attention coupling network for object detection. *IEEE Trans Image Process*, 28(1):113-126. https://doi.org/10.1109/TIP.2018.2865280

List of electronic supplementary materials

Video S1 Underwater autonomous object search and grasping in real sea areas