Yue LU, Xingyu CHEN, Zhengxing WU, Junzhi YU, Li WEN, 2022. A novel robotic visual perception framework for underwater operation. *Frontiers of Information Technology & Electronic Engineering,* 23(11):1602-1619.
https://doi.org/10.1631/FITEE.2100366

A novel robotic visual perception framework for underwater operation

Key words: Underwater operation; Robotic perception; Visual restoration; Video object detection

Corresponding author: Junzhi YU
E-mail: junzhi.yu@ia.ac.cn
ORCID: https://orcid.org/0000-0002-6347-572X

Motivation

1. Different from general object detection, there are two problems in underwater robotic visual perception, i.e., domain shift (discordance between the training domain and testing domain) and detection continuity and stability.
2. Underwater visual signals usually suffer from degeneration and form low-quality images and videos. The relationship between image quality and convolutional representation remains unclear. How does visual restoration contribute to object detection?
3. Detection continuity and stability are important for robotic perception. Uneven or discontinuous detection results cannot be ignored because they could cause jitter and even an error in control of the robot.

Method: domain effect

1. Based on URPC2018, three data domains are generated: (1) domain-*O*—the original dataset; (2) domain-*F*—all samples are processed by filtering based restoration (FRS); (3) domain-*G*—all samples are restored by GAN-based restoration (GAN-RS). Based on the underwater image quality measures and Lab color space, for domain quality, we define domain-*G*> domain-*F*>domain-*O*.

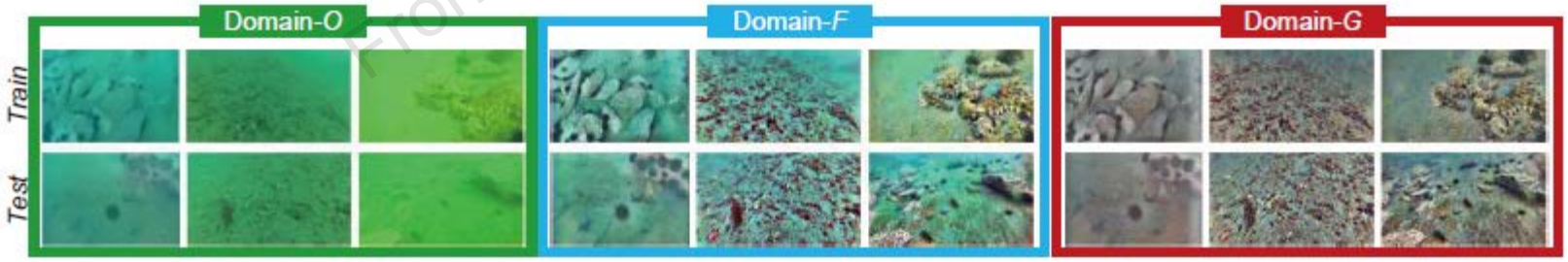

Fig. 2 Typical samples in domain-O, domain-F, and domain-G

Method: domain effect (Cont'd)

2. Based on visualization of convolutional representation and precision–recall 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.

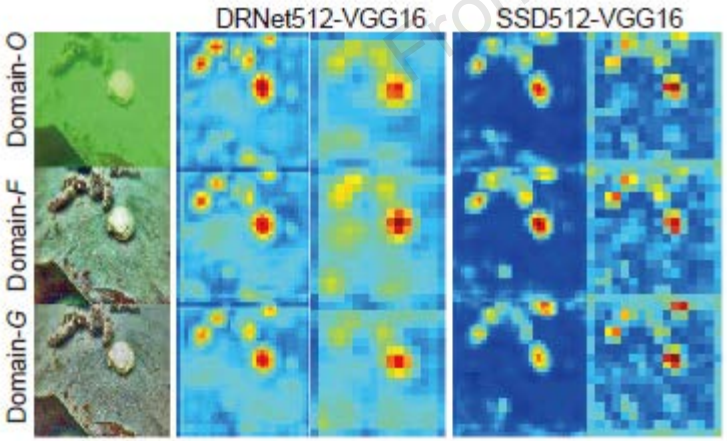

Fig. 4 Visualization of convolutional representation of objects. These features are associated with 64- or 128-size anchors, matching objects in this image. All features are processed using the L2 norm across channels, and are then normalized for visualization. For a fair comparison, the same normalization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

Method: domain effect (Cont'd)

3. We use domain-*O* and domain-*G* for evaluation of direction-related domain shift. The results show that compared to the high-quality domain, the low-quality domain induces better cross-domain generalizability.

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

Method: domain effect (Cont'd)

4. As for the real sea areas, the 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.

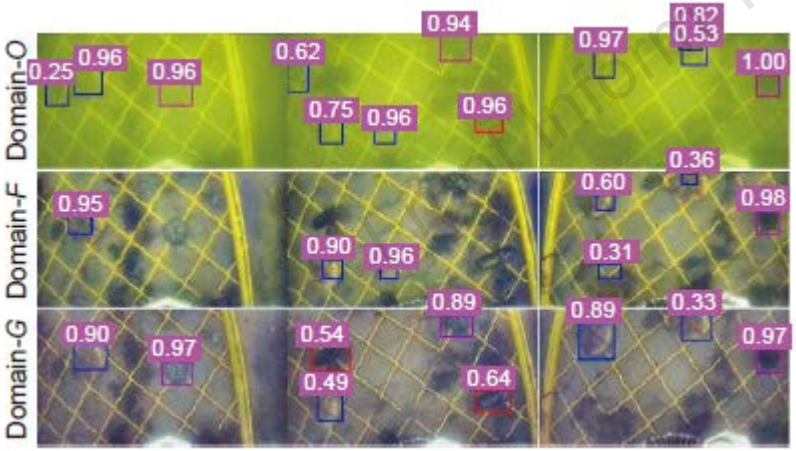

Fig. 6 Demonstration of online detection. DRNet512-VGG16-O and DRNet512-VGG16-F are hardly qualified for this online detection. By suppressing the problem of domain shift, DRNet512-VGG16-G and GAN-RS perform better in this field underwater scene. "Trepang," "echinus," and "shell" are detected in red, 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)

Method: detection continuity/stability

1. 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. For recall continuity, extremely short duration error (ESDE), short duration error (SDE), tracklet fragment error (TFE), and fragmental tracklet ratio (FTR) are proposed. For localization stability, the center jitter error (CJE) and size jitter error (SJE) are designed. Thus, the recall continuity error (RCE) is defined as RCE=ESDE+SDE+TFE+FTR, and localization jitter error (LJE) is defined as LJE=CJE+SJE.

Method: detection continuity/stability (Cont’d)

2. To enhance recall continuity and localization stability, we refine VID results based on tracklets, and propose online tracklet refinement (OTR). The OTR consists of three parts: (1) short tracklet suppression which can suppress false positives and produce false negatives; (2) fragment filling which can reduce tracklet fragment error; (3) temporal location fusion.

Method: underwater robotic visual perception framework

1. The proposed robotic visual perception framework adopts a visual restoration model and a detection model.

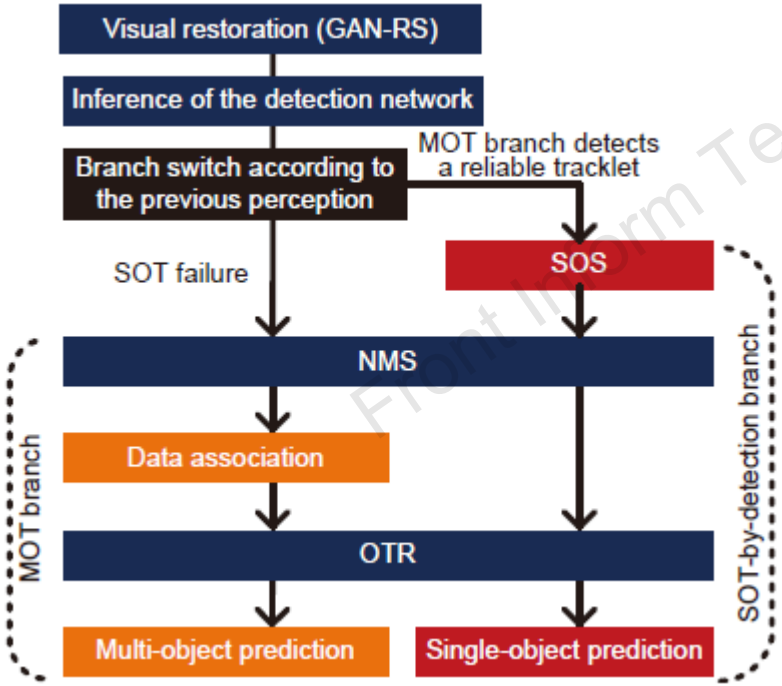

Fig. 9 Robotic visual perception framework with our proposed SOS and OTR. The MOT branch is designed to search the initial box for SOT-by-detection. A reliable tracklet is captured by OTR, while the SOT failure is captured by SOS. If switch conditions are not met, the previous behavior is continuously 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

Method: underwater robotic visual perception framework (Cont'd)

2. Taking inspiration from NMS, we leverage IoU-based suppression and design SOS-NMS. SOS-NMS is based on alternating the confidence score and IoU, and it 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 $\mathcal{B}^{\text{sos}} = \text{empty}$ **then**
  return $b^f = \text{empty}$
end if // NMS based on the confidence score
$\mathcal{B}^{\text{nms}} = \{\}; \mathcal{S}^{\text{nms}} = \{\}; \mathcal{O}^{\text{nms}} = \{\}$
while $\mathcal{B}^{\text{sos}} \neq \text{empty}$ **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 $b^f = \mathcal{B}^{\text{nms}}_{\text{idx}}$

Major results

Validation of LJE

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

Major results (Cont'd)

Continuity and stability evaluation

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

Major results (Cont'd)

Autonomous object search and grasping in real sea areas

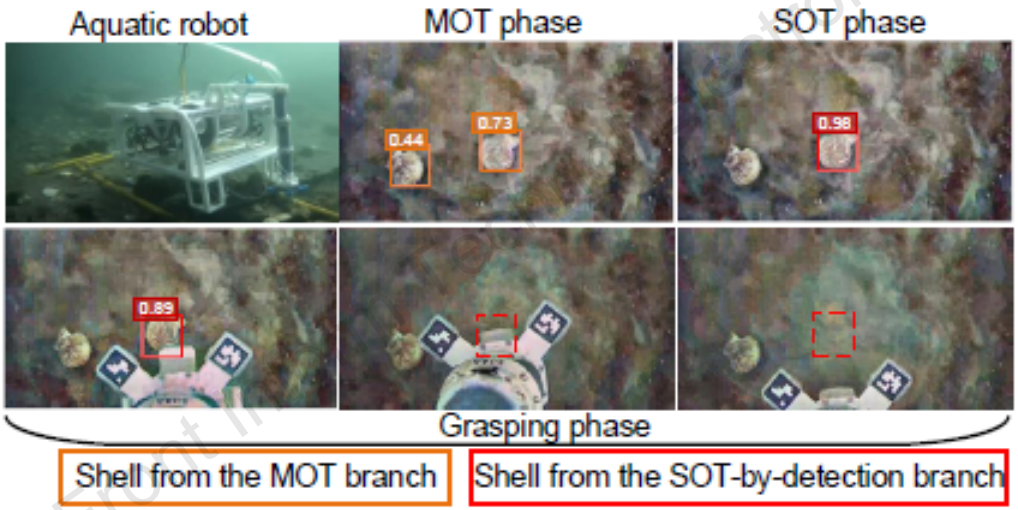

Fig. 11 Our robotic visual perception framework in the object grasping task. The proposed framework can provide robust visual information with visual restoration and flexible detection and tracking for robotic search and grasping

Conclusions

1. We performed domain analysis and revealed how visual restoration contributes to object detection in aquatic scenes.
2. Object detection recall continuity and localization stability have been analyzed in a novel way for robotic perception.
3. An underwater robotic visual perception framework has been proposed for underwater object search and grasping.
4. As a result, our conclusions and methods have been verified on datasets, and underwater autonomous object search and grasping have been achieved in real sea area.

Yue LU received the B.E. degree in electrical information science and technology from the College of Electronic Science and Engineering, Jilin University, Changchun, China, in 2018. He is currently pursuing the Ph.D. degree in control theory and control engineering with the Institute of Automation, Chinese Academy of Sciences, Beijing, China. His research interests include computer vision, deep learning, and underwater robotics.

Xingyu CHEN received the B.E. degree in electrical engineering and automation from the College of Nuclear Technology and Automation Engineering, Chengdu University of Technology, Chengdu, China, in 2015, and the Ph.D. degree in control theory and control engineering from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2020. He is currently a Researcher with Ytech, Kuaishou Technology. His research interests lie in the joint field of robotics and computer vision, including but not limited to scene perception, 3D geometry understanding, and human–machine interaction.

Zhengxing WU received the B.E. degree in logistics engineering from the School of Control Science and Engineering, Shandong University, Jinan, China, in 2008, and the Ph.D. degree in control theory and control engineering from the Institute of Automation, Chinese Academy of Sciences (IACAS), Beijing, China, in 2015. He is currently a Professor with the State Key Laboratory of Management and Control for Complex Systems, IACAS. His current research interests include bioinspired robots and intelligent control systems.

Junzhi YU received the B.E. degree in safety engineering and the M.E. degree in precision instruments and mechanology from the North University of China, Taiyuan, China, in 1998 and 2001, respectively, and the Ph.D. degree in control theory and control engineering from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2003. From 2004 to 2006, he was a Post-Doctoral Research Fellow with the Center for Systems and Control, Peking University, Beijing. He was an Associate Professor with the Institute of Automation, Chinese Academy of Sciences, in 2006, where he was a Full Professor in 2012. In 2018, he joined the College of Engineering, Peking University, as a Tenured Full Professor. His current research interests include intelligent robots, motion control, and intelligent mechatronic systems.

Li WEN received the B.E. degree in mechatronics engineering from the Beijing Institute of Technology, Beijing, China, in 2005, and the Ph.D. degree in mechanical engineering from Beihang University, Beijing, China, in 2011. From 2011 to 2013, he was a Postdoctoral Fellow with George Lauder Laboratory, Harvard University. He is currently a Professor with Beihang University. His research interests include bio-inspired robotics, soft robotics, smart materials, and comparative biomechanics. Dr. WEN is an Associate Editor for *IEEE Robotics and Automation Letters*.