The Visual Computer (2023) 39:5199–5208
https://doi.org/10.1007/s00371-022-02654-4

ORIGINAL ARTICLE

SSoB: searching a scene-oriented architecture for underwater object detection

Wanqi Yuan[1] · Chenping Fu[2] · Risheng Liu[1,3,4] · Xin Fan[1,3]

Accepted: 8 August 2022 / Published online: 10 September 2022

Abstract
Underwater object detection (UOD) suffers from low detection accuracy because of environmental degradations, such as haze-like effects, color distortions, and imaging noises. Therefore, we commit to resolving the issue of object detection with compounded environmental degradations that greatly challenges existing deep learning-based detectors. We propose a neural architecture search -based deep learning network to realize the UOD task, which can automatically discover the scene-oriented feature representation. Our network is accomplished through a unified macro-detector and a novel mixed anti-aliasing block (MAaB)-based search space. The macro-detector targets to learn intrinsic feature representations automatically from underwater images containing various environmental degradations and complete the subsequent detection tasks. The novel MAaB-based search space is proposed toward complex underwater scenes. The candidate operator MAaB has multiple kernels and anti-aliased convolutions in a single block for boosting the contextual representation capacity and the robustness of degraded factors. Finally, we use the differential search strategy guides the whole learning process to obtain the scene-friendly results. Extensive experiments demonstrate that our method outperforms the state-of-the-art detectors by a large margin. More importantly, in cases where environmental degradation is severely disturbed, our method is also superior to other popular detectors.

Keywords Object detection · Underwater scenes · Neural architecture search · Deep learning

1 Introduction

As an exponential increase in the availability of underwater imagery currently, deep learning-based underwater object detection (UOD) shows potentially unprecedented research opportunities for many halobios [1,2]. However, UOD suffers from low detection accuracy because of various environmental degradations. *First are haze-like effects.* The water medium scatters the light causing low-contrast and haze-like phenomena in the underwater photography [3]. *Second are color distortions.* Wavelength absorption usually causes a color reduction in the captured image, which leads to bluish or greenish underwater images [3,4]. *Third is imaging noise.* Electronics and sediments affect high dimensional imaging, causing noises in the underwater image. These environmental degradations greatly interfere with the imaging process, which makes UOD difficult.

The main difficulty of UOD is that the structural and statistical properties of objects in the underwater image are obstructed by various environmental degradations. Therefore, it is necessary to design appropriate detection structures for better feature representation. In a typical deep learning-based object detector, a backbone network plays an important role in extracting basic features for detecting [5–7]. Not surprisingly, if a backbone can extract more useful features, its corresponding detector will perform better. Hence, starting from AlexNet [8], more powerful backbones have been developed, such as ResNet [9], ResNetXT [10], MobileNetV2 [6], CBNet [5], and YOLOX [11]. While promising, they consume expensive computational costs for case-by-case design. In addition, since most of these existing backbones are originally designed for classification or

✉ Xin Fan
xin.fan@dlut.edu.cn ; xin.fan@ieee.org

1 International School of Information Science and Engineering, Dalian University of Technology, Dalian, China

2 School of Software Technology, Dalian University of Technology, Dalian, China

3 Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, Dalian, China

4 Peng Cheng Laboratory, Shenzhen, China

general detection tasks, directly using them to extract features for UOD may lead to suboptimal performance. Indeed, some researchers attempt to design specific backbones for underwater scenes [12–15]. However, these backbones heavily rely on abundant architecture engineering and subtle adjustments experiences. Besides, environmental degradation information exists in underwater images, while these heuristic manners hardly acquire these information from extensive images.

Recently, neural architectural search (NAS)-based methods [16–18] for computer vision tasks (e.g., classification and general object detection) have been introduced and applied well. The representative gradient-based architecture search methods [7,19,20] relaxed the non-differential architecture as a continuous weighted network for achieving differential search. Unfortunately, primitive search space (e.g., separable convolutions), is still a challenge to search optimal architecture for extracting deep features in underwater scenes with various degradation factors.

To alleviate the aforementioned issues, this paper focuses on a deep learning-based method that aims to search scene-oriented backbones (SSoB) and to embed a mixed anti-aliasing block (MAaB)-based search space, for solving UOD task. First, we develop NAS technology to discover the underwater scene-oriented backbone. As a result, our network can extract typical features under the interference of various environmental degradations. Then, we formulate a novel search space, which is more robust and stable to environmental degradations such as haze-like effects and imaging noises. Finally, with the MAaB-based search space, we employ the differentiable search strategy guides search processes, generating a scene-friendly result. Thus, our contributions can be distilled as threefold as follows:

- Different from existing heuristic backbones for UOD that heavily depend on engineering experiences, we construct a novel scene-oriented backbones learning model around environmental degradations from the differential NAS perspective.
- Toward the complex underwater scene, we propose new blocks as the candidate operations of a search space, i.e., MAaB. MAaB has multiple kernels in a single block to boost the contextual representation capacity and introduces anti-aliased convolutions to enhance the robustness of degraded factors.
- Extensive experiments are conducted on a popular underwater dataset URPC2020[1]. As shown in Fig. 1, our searched scene-oriented architecture significantly outperforms other state-of-the-art methods (including CNN-/transformer-based detectors) by a large margin.

[1] http://www.urpc.org.cn/index.html.

2 Related works

2.1 Underwater object detection

UOD aims at determining what and where an object is in an underwater image. Generally, deep learning-based detectors generally consist of four parts: a backbone that extracts feature from an image, a neck followed backbone that fuses multi-level features, a region proposal network (not necessarily part) followed the extracted features that generates prediction candidates, and a head for classification and localization prediction. In recent years, various methods in literatures have been proposed to tackle with UOD tasks. The common solution for UOD is to re-train existing detectors including CNN- and transformer-based detectors. Among them, some also attempt to redesign structures based on these existing detectors for UOD. Here we briefly review some of the recent detectors:

The state-of-the-art detectors can be briefly categorized into two major branches. The first branch contains CNN-based methods such as YOLO [21], SSD [22], RetinaNet [23], FSAF [24], YOLOX [11], Free-Anchor [25], FoveBox [26], Faster RCNN [27], FPN [28], Mask RCNN [29], Grid RCNN [30], Cascade RCNN [31], and Guided Anchoring [32]. The other branch contains transformer-based methods such as DETR [33], Swin Transformer [34], and PVTv1 [35]. Besides, some researchers also attempt to improve the feature extraction and representation capacity of structure based on these popular detectors for UOD [12,14,15].

2.2 Backbone for underwater object detection

Backbone play a vital role in detectors to extract basic object features for detection. UOD detectors generally adopt existing backbones, and most of these backbone are designed for classification or general detection. Meanwhile, some researchers also attempt to design specific backbones for UOD-based existing backbones. Here we briefly introduce these backbones:

The original works RCNN [36] and OverFeat [37] are pioneers for deep learning-based detectors. After them, almost all of current detectors use the pretraining and fine-tuning paradigm, that is, directly adopt the networks that are pretrained for ImageNet [38] classification task as their detection backbone. For instance, VGG [39], ResNet [9], and ResNetXT [10] are classification backbones, but they are widely used by the state-of-the-art detectors. Recently, CBNet [5], Darknet53 [40], and MoblieNetv2 [6] are designed for general detection. Obviously, UOD directly adopts these backbones may lead to suboptimal performance. In addition, there are some works design specific backbones for UOD [12,15]. However, these handcrafted methods requires much manpower and computation cost.

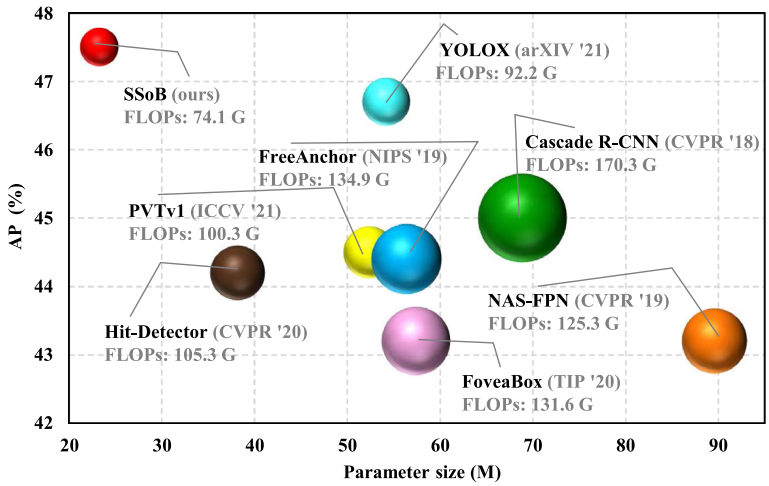


Fig. 1 Accuracy-speed-size trade-off accurate models on URPC2020 dataset for our method and other state-of-the-art detectors

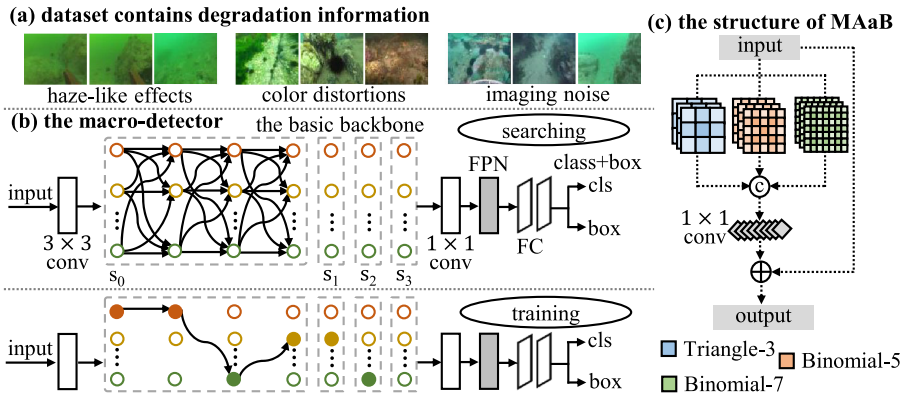


Fig. 2 Workflow of our method. As for the searching stage, we define a macro-detector to contain the basic backbone, FPN, and class+box. The basic backbone contains 20 layers to be searched, and each layer chooses a block from the MAaB-based search space. After the searching stage, we can derive the final structure by selecting one of optimal blocks at each layer. In this way, we can train this searched network, aiming to extract scene-oriented features from underwater images

More importantly, underwater datasets contains more environmental degradation information, but handcrafted methods hardly acquire these information from extensive images.

2.3 Neural architecture search

Neural architecture search is a automatic manner of learning architectures from data distribution that outperforms human expertise [41–43]. NAS for classification has attracted great attention recently. Some works [44–46] adopt reinforcement learning-based methods to use a RNN controller to generate a cell-based structure. Some works [47–49] use evolutionary algorithm-based manners to form architectures by mutating current ones. To speed up searching process, some works [7,50] adopt gradient-based paradigm to form a continuous relaxation search space, which allow the the differentiable optimization during the whole search phase.

Some recent works attempt to develop NAS for object detection. Early works attempt to [16] adopt evolutionary strategy to search a better backbone for detection tasks. Another group of approaches [17,18,51] use reinforcement learning algorithms to train a controller to generate potential components of detectors. Unfortunately, these two kinds of methods are too resources demanding, causing inefficient search. Recently, [20,52] formulate a detection supernet into differential form with a set of architecture and weight parameters, so that they can perform search in a gradient descent manner and reduce search cost of several hours. However, existing NAS methods have not yet been explored in the detection of underwater environments. Besides, the search space is designed by previously built blocks and might be naive for complex underwater scenarios.

3 The proposed approach

3.1 The scene-oriented architecture learning model

Existing manually designed detection backbones mainly depend on engineering skills. They are too resource demand-

ing for case-by-case redesigns. More importantly, we know that underwater images contain rich information of environmental degradations. Obviously, heuristic manners hardly acquire these information from extensive images. To overcome these problems, from the NAS perspective, we raise a differentiable search optimization strategy to design our UOD backbone SSoB, which can be formulated as:

$$\begin{aligned} &\min_{\alpha} \ \mathcal{L}(\boldsymbol{\omega}_{\alpha}^{*}, \alpha; \mathcal{D}_{\mathtt{val}}) \\ &s.t., \quad \boldsymbol{\omega}_{\alpha_{\mathtt{u}}}^{*} = \arg\min_{\boldsymbol{\omega}_{\alpha}} \mathcal{L}(\boldsymbol{\omega}_{\alpha}, \alpha; \mathcal{D}_{\mathtt{tr}}), \end{aligned} \tag{1}$$

where $\mathcal{L}(\cdot)$ is the loss function of detectors. $\mathcal{D}_{\mathtt{val}}$ and $\mathcal{D}_{\mathtt{tr}}$ are training and validation datasets, respectively. As shown in Fig. 2a, both of $\mathcal{D}_{\mathtt{val}}$ and $\mathcal{D}_{\mathtt{tr}}$ contain various underwater degradations. The search approach seeks to find a backbone α that minimizes the validation loss $\mathcal{L}(\boldsymbol{\omega}_{\alpha}^{*}, \alpha; \mathcal{D}_{\mathtt{val}})$ with the trained weights $\boldsymbol{\omega}_{\alpha}^{*}$. The weights $\boldsymbol{\omega}_{\alpha}^{*}$ associated with the backbone are obtained by minimizing the training loss $\mathcal{L}(\boldsymbol{\omega}_{\alpha}, \alpha; \mathcal{D}_{\mathtt{tr}})$.

As shown in Fig. 2b, we propose a macro-detector framework to solve problem in Eq. 1. The macro-detector is decoupled into three main principled parts, i.e., the basic backbone, FPN, and class+box. The basic backbone, extracting features of images, contains a 3×3 convolution with stride of 2, four stages that contain 20 blocks to be searched, and another 1×1 convolution with stride of 1. According to practical experience, the channel of each stage is set to {48, 192, 384, 768}, respectively. Then we send the features into FPN to fuse these features from different stages. After FPN, class+box is used to predict object classification and bounding box.

3.2 MAaB-based search space

To begin with, according to the latest NAS method [7,20], we configure to define a block as the smallest module. To this end, the macro-detector searching space comprises a layer-level search, which allows us to explore the whole network from a block perspective. We adopt the fundamental routine to design the layer-wise search space: a search space includes a number of candidate blocks (operations). Each layer to be searched can choose a different block from candidate blocks.

How to construct a layer-level search space plays a vital role in NAS technique, existing NAS-based approaches for classification or general detection [7,20] are mainly designed primitive operators (e.g., separable convolutions), and these unsophisticated operators may pose a touch issue for optimizing the backbone architectures. For this purpose, we consider requirements of contextual representation capacity and degradation robustness in underwater scenes for constructing our search space. The search space is consisted of novel blocks MAaB which is specifically designed for underwater scenes.

There are two main aspects to consider for extracting more typical features from complex underwater scenes. For one thing, a backbone needs to extract multi-scale features as much as possible. Many approaches [53,54] choose to fuse features after backbones, while this means more layers are needed. For another thing, detecting objects from cloudy images requires high robustness of a detector. However, common downsampling operations (such as the convolution with stride 2, MaxPooling) do not have the capacity to anti-alias, which may cause damage to robustness [55]. To overcome these issues, we design the MAaB blocks inspired by [55,56]. MAaB has multiple different sizes of kernels in one block, which can easily fuse multi-scale features without extra layers. Besides, it introduces anti-aliased operations (i.e., Convblupool) in a block, which can enhance the robustness of degraded factors. Figure 2c show the structure of MAaB block, which is composed of several convblurpool with stride 2 and one 1×1 convolution with stride 1. For an input, it is split into N groups along channel axis. Then each group is processed by an independent convblurpool. The outputs of these parallel branches are concatenated and then fused by the final 1×1 convolution to reduce output channels. If the input and output have the same dimension, we use a skip operation to add them.

Convblupool [55] is an convolution operation with a normalized Guassian filter with stride 2 to downsample an input. The Convblupool use blur kernels. In the paper, we set N to [1,2,3] to construct our search space. $N = 1$, there is one convblurpool with blur kernel Triangle-3, Binomial-5, or Binomial-7. $N = 2$, there are two convblurpool operation, and their kernel sizes are [Triangle-3,Binomial-5], [Triangle-3,Binomial-7], or [Binomial-5,Binomial-7]. $N = 3$, there are three convblurpool operations, and their kernel sizes are [Triangle-3,Binomial-5,Binomial-7]. In detail, the value of Triangle-3, Binomial-5, and Binomial-7 are [1, 2, 1], [1, 4, 6, 4, 1], and [1, 6, 15, 20, 15, 6, 1]. The weights are normalized, and the filters are the outer product of the following vectors with themselves. Specifically, the eight candidate blocks are given in the following. Note that we have a popular block called "skip," which allows us to reduce the depth of the backbone network.

- Triangle-3, group=1, MAaB (T3)
- Triangle-5, group=1, MAaB (B5)
- Triangle-7, group=1, MAaB (B7)
- Triangle-3, Binomial-5, group=2, MAaB (T3-B5)
- Triangle-3, Binomial-7, group=2, MAaB (T3-B7)
- Triangle-3, Binomial-7, group=2, MAaB (B5-B7)
- Triangle-3, Binomial-5, Binomial-7, group=3, MAaB (T3-B5-B7)
- Skip

3.3 The differentiable search algorithm

We adopt the differentiable manner proposed in [19] to solve Eq.(1). In searching phase, the output of each intermediate layer is computed with a weighted sum based on all candidate blocks. For backbones, the output of i-th layer is formulated as

$$x_i = \sum_{b \in \mathcal{B}} \frac{\exp(\alpha_i^b)}{\sum_{b' \in \mathcal{B}} \exp(\alpha_i^{b'})} b(x_{i-1}), \tag{2}$$

where x_i is the output of the ith layer, α_i^b is the parameter for block $b(\cdot)$, and it can be simply perceived as the scores of b-th block in ith layer. And $\mathcal{B}$ denotes the search space as described in the above subsection. The continuous relaxation of Eq.(2) makes the entire problem Eq.(1) differentiable to both weights and architecture parameters, so we can search the backbone in an end-to-end manner. In the training phase, we choose a block with highest scores for each layer to build our backbone.

At last, the loss function used in Eq. (1) is defined as follows:

$$\mathcal{L}(\alpha_{\mathrm{u}}, \boldsymbol{\omega}_{\alpha_{\mathrm{u}}}) = \mathcal{L}_{det}(\alpha_{\mathrm{u}}, \boldsymbol{\omega}_{\alpha_{\mathrm{u}}}) + \gamma \mathcal{L}_{flo}(\alpha) \tag{3}$$

The first term $\mathcal{L}_{\mathrm{det}}(\cdot)$ denotes the loss of detectors, which is the classification and localization loss. As underwater detectors are often deployed to mobile CPUs, we introduce the second term to guarantee detection efficiency. $\mathcal{L}_{\mathrm{flo}}(\cdot)$ indicates FLOPs of the backbone part and can be decomposed as linear sum of each operations. The two terms are weighted by a balancing parameter γ. It is clear that the loss function (3) is differentiable due to the continuous relaxation of Eq.(2). Thus $\{\alpha_{\mathrm{u}}, \boldsymbol{\omega}_{\alpha_{\mathrm{u}}}\}$ can be optimized jointly using SGD.

4 Experiments

4.1 Experimental configurations

We conduct experiments on URPC2020 dataset which consists of 6575 underwater images. The dataset is split into trainval set with 5260 images and test set with 1315 images. The dataset has 4 object categories including echinus, holothurian, scallop, and starfish. We analyze our method by numerous comparison experiments. For all experiments, the input image is resized to the default size of the respective methods, and implementation is based on mmdetection [2] and Pytorch framework.

SSoB searching. We first initialize the basic backbone with kaiming_init. Then we search the backbone on URPC2020 trainval set. We use SGD optimizer with a batch size of 2 images, and search for 12 epochs. In each iteration, we update $\boldsymbol{\omega}_\alpha$ and α alternately. We set learning rate, momentum and balancing parameter γ being 0.04, 0.9 and 0.01, respectively.

Detection training. We choose blocks with highest scores for each layer to build SSoB. We first pretrain SSoB on Imagenet for 150 epochs. Then we fine-tune the whole detector on URPC2020 trainval dataset for 24 epochs with SGD optimizer, and 1× schedule. We set the initial learning rate being 0.04 which is divided by 10 at the 7th and 10th epoch. The weight decay , momentum and batch size are 0.0001, 0.9 and 2, respectively.

4.2 Main results

Comparisons with handcrafted methods We replace the backbone in FPN [28] with other excellent backbone, i.e., Darknet53, ResNetXT101, and MobileNetv2, and form three competitors accordingly. As shown in Table 1. SSoB surpasses these competitors by a large margin with less parameters. Specifically, SSoB is 6.8% higher on AP compared to the Darknet53 based detector with less than one half of the parameters. Compared with ResNetXT101, the similar excellent phenomenon is also existed. In addition, we outperform MobileNetv2 by 9.1% on AP with less parameters. These experimental results demonstrate that our methods can design a better backbone than handcrafted methods.

Comparisons with NAS-based methods. As shown in Table 2, we compare SSoB with detectors that adopt NAS based model. FBNet [7] is searched on ImageNet dataset for classification tasks, and we directly apply it as the backbone of a detector. Unfortunately, its performance on detection is disappointing. NS-FPN [17]and hit detector [20] are designed for detection tasks, and we thus re-search the architecture on URPC2020. NAS-FPN aims to discover a new feature pyramid architecture for detectors while leaving the backbone unchanged. Our method outperforms NAS-FPN by 4.3% with less parameters and FLOPs. Hit detector discovers architectures for all components (i.e. backbone, neck, and head) of detectors while its search space is consisted of common blocks (such as separable block). Our method also surpasses hit detector. These experiment results indicate that it is important to design specific backbones with efficient search space toward underwater scenes in an detector.

Comparisons with state-of-the-art methods. We compare our methods with other state-of-the-art methods on URPC2020, the results are summarized in Table 3. SSoB only applies simple data augmentation and 1X training scheme, which achieves 47.5% AP without bells and whistles. Our method has fewer parameters and performs better than CNN-based detectors. Specifically, CSAM and FERNet are designed for UOD tasks. Both develop sophisticated deep

[2] https://github.com/open-mmlab/mmdetection.

Table 1 Comparisons with handcrafted models on dataset URPC2020

Model	# Params	# FLOPs	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
Darknet53 [40]	50.7M	124.8G	40.7	76.3	39.0	15.0	36.4	45.0
ResNetXT101 [10]	60.8M	134.9G	43.1	78.5	42.7	22.8	40.0	48.5
MoblieNetv2 [6]	25.8M	**58.7G**	38.4	73.3	36.0	14.8	34.2	42.4
SSoB	**23.3M**	74.1G	**47.5**	**82.8**	**50.3**	**25.2**	**42.2**	**52.8**

The best result is in bold

Table 2 Comparisons with NAS-based methods on dataset URPC2020

Model	Modified			# Params	# FLOPs	AP	AP_{50}	AP_{75}
	B	N	H					
FBNet-C [7]	✓	-	-	25.4M	109.5G	33.7	68.1	28.5
NAS-FPN [17]	-	✓	-	89.6M	125.3G	43.2	79.3	42.2
Hit detector [20]	✓	✓	✓	38.2M	105.3G	44.2	77.8	45.9
SSoB	✓	-	-	**23.3M**	**74.1G**	**47.5**	**82.8**	**50.3**

B: Backbone, N: Neck, H: Head The best result is in bold

Table 3 Comparison with state-of-the-art methods on URPC2020

Model	Backbone	# Params	# FLOPs	FPS	AP	AP_{50}	AP_{75}
Free-anchor [25]	ResNetXT101	56.4M	134.9G	2.9	44.4	80.0	44.6
FoveaBox [26]	ResNet101	57.4M	131.6G	5.3	43.2	79.3	42.8
YOLOX [11]	YOLOX-l	54.2M	92.2G	**7.2**	46.7	81.2	49.4
Grid RCNN [30]	ResNetXT101	122.0M	255.9G	2.6	43.5	78.4	44.1
Cascade RCNN [30]	ResNet50	68.9M	170.3G	5.8	45.0	78.8	47.5
CSAM [14]	DarkNet-53	66.4M	165.5G	5.5	46.4	79.2	41.1
FERNet [12]	VGG16	232.5M	326.6G	3.2	44.2	76.7	42.3
PVTv1 [35]	PVT-Medium	52.4M	100.3G	3.1	44.5	80.1	44.7
DETR [33]	ResNet50	41.3M	**43.6G**	4.4	22.8	55.5	13.3
SSoB	Searched	**23.3M**	74.1G	6.7	**47.5**	**82.8**	**50.3**

The best result is in bold

architectures to improve the feature representation capacity. SSoB outperforms them by a large margin. For AP, SSoB is 1.1% higher than CSAM and 3.3% than FERNet. In addition, SSoB also outperforms transformer-based methods. Although DETR has fewer FLOPs than SSoB, SSoB outperforms DETR by 24.7% in AP, 27.3% in AP_{50}, and 37.0% in AP_{75}. DETR fails to detect small objects very well, so its performance is poor on underwater datasets with numerous small object. These experiment results further demonstrate that our methods can design a better architecture than existing popular detectors.

we also compare the proposed methods in terms of speed, as shown in Table 3. Our method can achieve 6.7 FPS, which is similar to YOLOX. Compared with other methods, our method can be well qualified for real-time detection tasks.

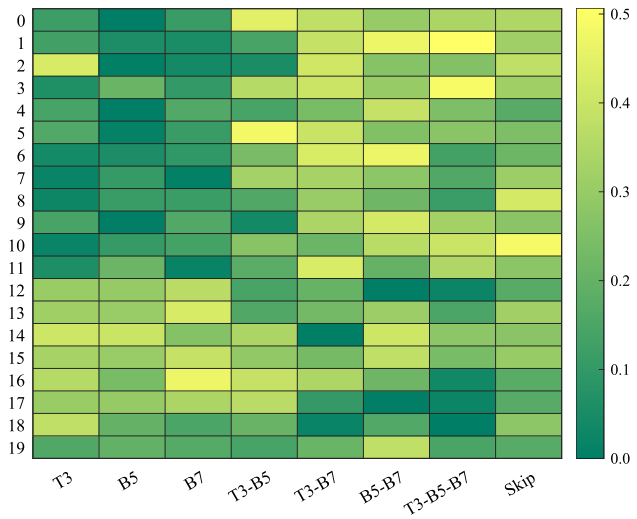


Fig. 3 The heatmaps of final candidate operators (i.e., α), where 20 layers to be searched of the backbone are plotted orderly. yellow boxes indicate the final choice

4.3 Performance analysis

Searching space analysis. Figure 3 plots the heatmaps toward the final searched backbone. Visual inspection shows that

Table 4 Comparison of performance of SSoB on different detectors. (URPC2020)

Model	Backbone	AP	AP_{50}	AP_{75}
Cascade RCNN [30]	ResNet50	45.0	78.8	47.5
	SSoB	**46.9**	**80.7**	**48.0**
YOLOX [11]	YOLOX-l	46.7	81.2	49.4
	SSoB	**47.2**	**81.6**	**50.1**
FoveaBox [26]	ResNet101	43.2	79.3	42.8
	SSoB	**45.5**	**80.2**	**47.8**
Ours	SSoB	**47.5**	**82.8**	**50.3**

The result of detectors with SSoB is in bold

multi-kernel (such as T3-B5-B7, B5-B7) operations occupy the main position in the first 12 layers, which indicates that the first half of the backbone paying more attention to extract more image information. For the rear 8 layers, single-kernel (such as T3, B5, and B7) operations are got relatively high scores, and it demonstrates that the rear half of the backbone relaxes the requirements for feature intensity. In addition, almost all blocks have been selected, and it demonstrate that single- and multi- kernel are necessary for construction of underwater backbones.

Various detectors. To validate the generalization ability of SSoB, we combine SSoB with different detectors. We select popular detectors like Cascade RCNN, YOLOX, and FoveaBox in this analysis. As demonstrated in Table 4, performances of these detectors are improved prominently (for AP, 1.9% in Cascade RCNN, 0.5% in YOLOX, and 2.3% in FoveaBox). SSoB shows the strong generalization capacity on different detectors. However, the best performance is achieved by our original methods. The search process is based on a macro-detector. Therefore, the searched network combined with other detectors may result in suboptimal performance.

The robustness of SSoB on other datasets. In order to verify that SSoB also has an effective performance improvement on other datasets, we carry out comparative experiments with currently proposed methods. The comparative experiments are carried out on the experimental dataset UODD [14] proposed by CSAM. UODD contains 3 types of underwater objects, *i.e.,* holothurian, echinus,and scallop. We take 2560 images for training and 502 images for testing. As shown in Table 5, the detection accuracy of SSoB comprehensively outperforms these detectors. For instance, SSoB surpasses FoveaBox by 5.1%, YOLOX by 1.9%, Grid RCNN by 2.8%, Cascade RCNN by 1.6%, and CSAM by 1.6% in AP.

Finally, we also compare the proposed methods in terms of speed, as shown in Table 5. Our method can achieve 12.0 FPS, which is similar to YOLOX. Compared to other methods, our method can be well qualified for real-time detection tasks.

Study on various environmental degradations Figure 4 exhibits some qualitative examples of various environmental degradations. For color distortions, most popular backbones fail to complete detection, there are error and missed detec-

Table 5 The comparison results on UODD dataset

Model	Backbone	AP	AP_{50}	AP_{75}	FPS
FoveaBox [26]	ResNet101	45.6	85.1	43.5	7.3
YOLOX [11]	YOLOX-l	48.8	86.3	51.7	**12.3**
Grid RCNN [30]	ResNetXT101	47.9	85.8	52.1	6.6
Cascade RCNN [30]	ResNet50	49.1	87.9	52.1	10.0
CSAM [14]	DarkNet-53	49.1	88.4	48.3	10.4
SSoB	searched	**50.7**	**89.7**	**53.8**	12.0

The result of detectors with SSoB is in bold

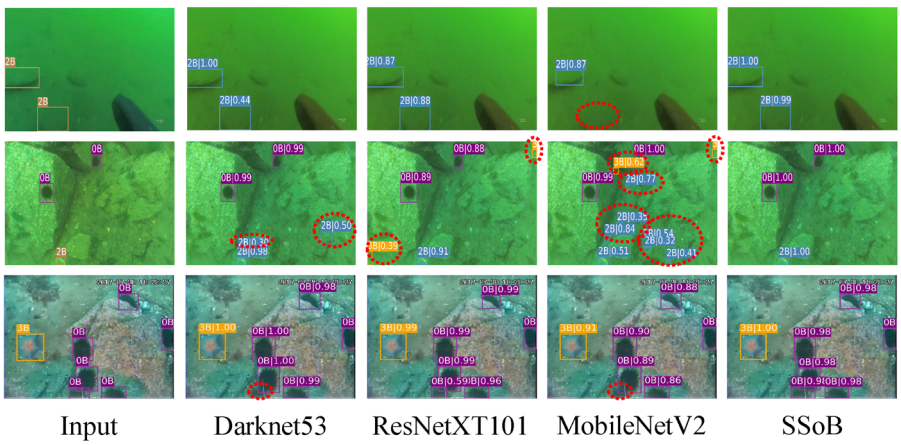


Fig. 4 Some qualitative examples on URPC2020. The example from top to bottom is haze-like effects, color distortions, and imaging noise, respectively. Both error and missed detection are marked with a red dotted box

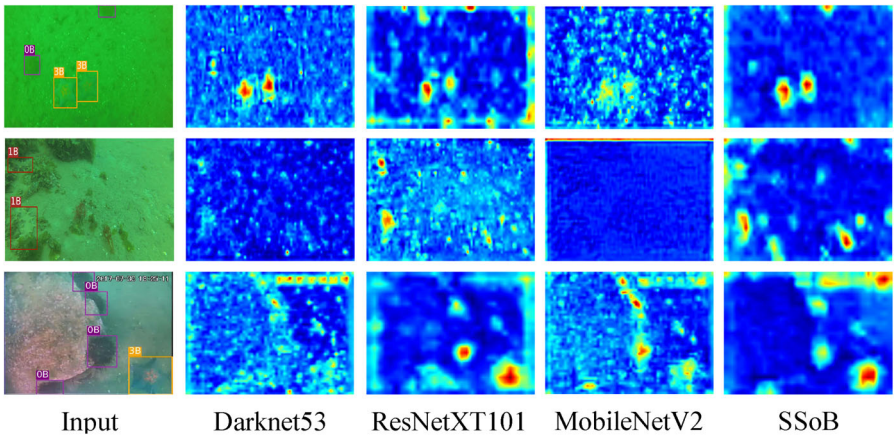


Fig. 5 Examples of visualization of the feature maps on URPC2020. The example from top to bottom is haze-like effects, color distortions, and imaging noise, respectively

tion phenomena in these backbones. However, our SSoB completes the detection task very well. For haze-like effects and imaging noise, some and our methods can complete the detection task very well. Some methods still have error and missed detection phenomena, for example, MobileNetV2 and Darknet53 have missed detection. The qualitative results demonstrate that SSoB can overcome the obstacles that degradation poses to feature extractions. Figure 5 shows the feature visualization results of various environmental degradations. For color distortions, the feature response of MobileNetV2 and Darknet53 is relatively weak. For haze-like effects and imaging noise, the amplitude of feature response of Darknet53, ResNetXT101, and MobileNetV2 is attenuated inconsistently. But on various environmental degradations, our SSoB significant boots the feature response on discriminative region while suppressing the interference. These feature response results further demonstrate that SSoB does perform well on various environmental degradations.

5 Conclusion

In this paper, we propose an automatically scene-oriented feature extraction module for solving the UOD task. Based on NAS technology, we fully discover the potential and inherent information of different underwater images; thus, our backbone can comprehensively extract deep features. Meanwhile, we also formulate a MAaB-based search space that can further improve the performance of our methods. Both qualitative and quantitative experimental results demonstrate that our SSoB has great superiority over the state-of-the-art methods.

Declarations

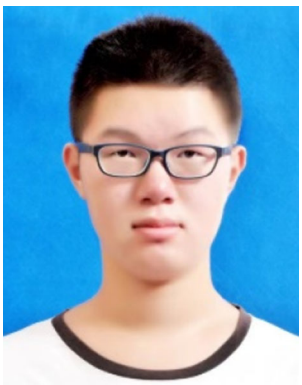
Conflict of interest We declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work, there is no professional or other personal interest of any nature or kind in any product, service and/or company that could be construed as influencing the position presented in, or the review of, the manuscript entitled, "SSoB: Searching a Scene-Oriented Architecture for Underwater Object Detection."

References

1. Pang, Y., Wu, C., Wu, H., Yu, X.: Over-sampling strategy-based class-imbalanced salient object detection and its application in underwater scene. Vis, Comput (2022)
2. Mhala, N.C., Pais, A.R.: A secure visual secret sharing (vss) scheme with cnn-based image enhancement for underwater images. Vis. Comput. **37**, 2097 (2021)
3. Liang, P., Dong, P., Wang, F., Ma, P., Bai, J., Wang, B., Li, C.: Learning to remove sandstorm for image enhancement. Vis, Comput (2022)
4. Lin, R., Liu, J., Liu, R., Fan, X.: Global structure-guided learning framework for underwater image enhancement. Vis, Comput (2021)
5. Liu, Y., Wang, Y., Wang, S., Liang, T., Zhao, Q., Tang, Z., Ling, H.: Cbnet: a novel composite backbone network architecture for object detection. In: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI. pp. 11 653–11 (2020)
6. Sandler, M., Howard, A.G., Zhu, M., Zhmoginov, A., Chen, L.: Mobilenetv2: inverted residuals and linear bottlenecks. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR. (2018), pp. 4510–4520
7. Wu, B., Dai, X., Zhang, P., Wang, Y., Sun, F., Wu, Y., Tian, Y., Vajda, P., Jia, Y., Keutzer, K.: Fbnet: hardware-aware efficient convnet design via differentiable neural architecture search. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR. pp. 10 734–10 (2019)
8. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Commun. ACM **60**(6), 84–90 (2017)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR. pp. 770–778 (2016)
10. Xie, S., Girshick, R.B., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR. pp. 5987–5995. (2017)
11. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: YOLOX: exceeding YOLO series in 2021. In: CoRR. vol. abs/2107.08430, (2021). [Online]. Available: https://arxiv.org/abs/2107.08430

12. Fan, B., Chen, W., Cong, Y., Tian, J.: Dual refinement underwater object detection network. In: Computer Vision - ECCV - 16th European Conference, Glasgow, UK, August 23–28,: Proceedings. Part XX **12365**(2020), 275–291 (2020)
13. Lin, W., Zhong, J., Liu, S., Li, T.H., Li, G.: ROIMIX: proposal-fusion among multiple images for underwater object detection. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP. pp. 2588–2592 (2020)
14. Jiang, L., Wang, Y., Jia, Q., Xu, S., Liu, Y., Fan, X., Li, H., Liu, R., Xue, X., Wang, R.: Underwater species detection using channel sharpening attention. In: ACM Multimedia Conference, pp. 4259–4267 (2021)
15. Liu, C., Wang, Z., Wang, S., Tang, T., Tao, Y., Yang, C., Li, H., Liu, X., Fan, X.: A new dataset, poisson gan and aquanet for underwater object grabbing. IEEE Trans. Circuits Syst. Video Technol. **32**(5), 2831–2844 (2022)
16. Chen, Y., Yang, T., Zhang, X., Meng, G., Xiao, X., Sun, J.: Detnas: backbone search for object detection. In: Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems, NeurIPS. pp. 6638–6648. (2019)
17. Ghiasi, G., Lin, T., Le, Q.V.: NAS-FPN: learning scalable feature pyramid architecture for object detection. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR. pp. 7036–7045. (2019)
18. Wang, N., Gao, Y., Chen, H., Wang, P., Tian, Z. Shen, C., Zhang, Y.: NAS-FCOS: fast neural architecture search for object detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR. pp. 11 940–11. (2020)
19. Liu, H., Simonyan, K., Yang, Y.: DARTS: differentiable architecture search. In: 7th International Conference on Learning Representations, ICLR. (2019)
20. Guo, J., Han, K., Wang, Y., Zhang, C., Yang, Z., Wu, H., Chen, X., Xu, C.: Hit-detector: Hierarchical trinity architecture search for object detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR. pp. 11 402. (2020)
21. Redmon, J., Divvala, S.K., Girshick, R.B., Farhadi, A.: You only look once: Unified, real-time object detection. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR. pp. 779–788 (2016)
22. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S.E., Fu, C., Berg, A.C.: SSD: single shot multibox detector. In: Computer Vision - ECCV 2016–14th European Conference, Amsterdam, The Netherlands, October 11–14,: Proceedings. Part I **9905**(2016), pp. 21–37. (2016)
23. Lin, T., Goyal, P., Girshick, R.B., He, K., Dollár, P.: Focal loss for dense object detection. IEEE Trans. Pattern Anal. Mach. Intell. **42**(2), 318–327 (2020)
24. Zhu, C., He, Y., Savvides, M.: Feature selective anchor-free module for single-shot object detection. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR. pp. 840–849. (2019)
25. Zhang, X., Wan, F., Liu, C., Ji, R., Ye, Q.: Freeanchor: learning to match anchors for visual object detection. In: Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems, NeurIPS. pp. 147–155. (2019)
26. Kong, T., Sun, F., Liu, H., Jiang, Y., Li, L., Shi, J.: Foveabox: Beyound anchor-based object detection. IEEE Trans. Image Process. **29**, 7389–7398 (2020)
27. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems, NeurIPS
28. Lin, T., Dollár, P., Girshick, R.B., He, K., Hariharan, B., Belongie, S.J.: Feature pyramid networks for object detection. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR. pp. 36–944. (2017)
29. He, K., Gkioxari, G., Dollár, P., Girshick, R.B.: Mask R-CNN. IEEE Trans. Pattern Anal. Mach. Intell. **42**(2), 386–397 (2020)
30. Lu, X., Li, B., Yue, Y., Li, Q., Yan, J.: Grid R-CNN. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR. pp. 7363–7372. (2019)
31. Cai, Z., Vasconcelos, N.: Cascade R-CNN: delving into high quality object detection. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR. pp. 6154–6162. (2018)
32. Wang, J., Chen, K., Yang, S., Loy, C.C., Lin, D.: Region proposal by guided anchoring. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR. pp. 2965–2974. (2019)
33. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Computer Vision - ECCV - 16th European Conference, Glasgow, UK, August 23–28. Proceedings, Part **I**(12346), 213–229 (2020)
34. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: hierarchical vision transformer using shifted windows In: IEEE/CVF International Conference on Computer Vision, ICCV. pp. 9992–10 002. (2021)
35. Wang, W., Xie, E., Li, X., Fan, D., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: IEEE/CVF International Conference on Computer Vision, ICCV. pp. 548–558 (2021)
36. Girshick, R.B., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR. pp. 580–587 (2014)
37. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y.: Overfeat: integrated recognition, localization and detection using convolutional networks. In: 2nd International Conference on Learning Representations, ICLR. (2014)
38. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR
39. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: 3rd International Conference on Learning Representations, ICLR. (2015)
40. Redmon, J., Farhadi, A.: Yolov3: an incremental improvement. CoRR. vol. abs/1804.02767, (2018). [Online]. Available: http://arxiv.org/abs/1804.02767
41. Liu, R., Ma, L., Zhang, J., Fan, X., Luo, Z.: Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR. pp. 10 561–10 570. (2021)
42. Xu, Y., Xie, L., Zhang, X., Chen, X., Qi, G., Tian, Q., Xiong, H.: PC-DARTS: partial channel connections for memory-efficient architecture search. In: 8th International Conference on Learning Representations, ICLR. (2020)
43. Ma, L., Jin, D., Liu, R., Fan, X., Luo, Z.: Joint over and under exposures correction by aggregated retinex propagation for image enhancement. IEEE Signal Process. Lett. **27**, 1210–1214 (2020)
44. Cai, H., Chen, T., Zhang, W., Yu, Y., Wang, J.: Efficient architecture search by network transformation. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, AAAI. S. A. McIlraith and K. Q. Weinberger, Eds., pp. 2787–2794. (2018)
45. Liu, C., Zoph, B., Neumann, M., Shlens, J., Hua, W., Li, L., Fei-Fei, L., Yuille, A.L., Huang, J., Murphy, K., "Progressive neural architecture search," in Computer Vision - ECCV 2018–15th European Conference, Munich, Germany, September 8–14,: Proceedings. Part I **11205**(2018), 19–35 (2018)
46. Zoph, B., Vasudevan, V., Shlens, J., Le, Q.V.: Learning transferable architectures for scalable image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR. pp. 8697–8710. (2018)

47. Real, E. Aggarwal, A., Huang, Y., Le, Q.V.: Regularized evolution for image classifier architecture search. In: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI. pp. 4780–4789. (2019)
48. Yang, Z., Wang, Y., Chen, X., Shi, B., Xu, C., Xu, C., Tian, Q., Xu, C.: CARS: continuous evolution for efficient neural architecture search. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR. pp. 1826–1835.(2020)
49. Guo, Z., Zhang, X., Mu, H., Heng, W., Liu, Z., Wei, Y., Sun, J.: Single path one-shot neural architecture search with uniform sampling. In: Computer Vision - ECCV - 16th European Conference, Glasgow, UK, August 23–28,: Proceedings. Part XVI **12361**(2020), 544–560 (2020)
50. Xue, C., Yan, J., Yan, R., Chu, S.M., Hu, Y., Lin, Y.: Transferable automl by model sharing over grouped datasets. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR. pp. 9002–9011. (2019)
51. Du, X., Lin, T., Jin, P., Ghiasi, G., Tan, M., Cui, Y., Le, Q.V., Song, X.: Sinenet: Learning scale-permuted backbone for recognition and localization. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR. pp. 11 589–11 598. (2020)
52. Xu, H., Yao, L., Li, Z., Liang, X., Zhang, W.: Auto-fpn: Automatic network architecture adaptation for object detection beyond classification. In: IEEE/CVF International Conference on Computer Vision, ICCV. pp. 6648–6657. (2019)
53. Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell. **40**(4), 834–848 (2018)
54. Li, H., Xiong, P., An, J., Wang, L.: Pyramid attention network for semantic segmentation. In: British Machine Vision Conference, BMVC. p. 285. (2018)
55. Zhang, R.: Making convolutional networks shift-invariant again. In: Proceedings of the 36th International Conference on Machine Learning, ICML. vol. **97**, pp. 7324–7334. (2019)
56. Tan, M., Le, Q.V.: Mixconv: mixed depthwise convolutional kernels. In: 30th British Machine Vision Conference, BMVC. p. 74. (2019)



Wanqi Yuan is currently an undergraduate student in DUT-RU International School of Information Science & Engineering, Dalian University of Technology. His research interests include machine learning and underwater object detection.

Chenping Fu received the M.S. degree in computer science from Liaoning University, Shenyang, China, in 2019. She is currently pursuing the Ph.D. degree in software engineering with the Dalian University of Technology, Dalian. Her research interests include computer vision, object detection, and deep learning.

Risheng Liu (Member, IEEE) received the B.S. and Ph.D. degrees both in mathematics from the Dalian University of Technology in 2007 and 2012. He was a visiting scholar in the Robotic Institute of Carnegie Mellon University from 2010 to 2012. He served as Hong Kong Scholar Research Fellow at the Hong Kong Polytechnic University from 2016 to 2017. He is currently a professor with DUT-RU International School of Information Science & Engineering, Dalian University of Technology. He was awarded the "Outstanding Youth Science Foundation" of the National Natural Science Foundation of China. His research interests include machine learning, optimization and computer vision.

Xin Fan (Senior Member, IEEE) was born in 1977. He received the B.E. and Ph.D. degrees in Information and Communication Engineering from Xi'an Jiaotong University, Xi'an, China, in 1998 and 2004, respectively. He was with Oklahoma State University at Stillwater, Stillwater, OK, USA, from 2006 to 2007, as a postdoctoral research fellow. He joined the School of Software, Dalian University of Technology, Dalian, China, in 2009. His current research interests include computational geometry and machine learning, and their applications to low-level image processing and diffusion tensor imaging magnetic resonance image analysis.