



山东大学
SHANDONG UNIVERSITY



2023世界海洋科技大会
海洋信息与大数据分会场

水下智能感知与理解

UNDERWATER INTELLIGENT PERCEPTION & UNDERSTANDING

报告人：丛润民

山东大学控制科学与工程学院
机器智能与系统控制教育部重点实验室

2023-09-18

► Outline

- Introduction
- Technical Methods
 - An Underwater Image Enhancement Benchmark Dataset And Beyond (TIP'20)
 - PUGAN: Physical Model-guided Underwater Image Enhancement Using GAN With Dual-Discriminators (TIP'23)
 - WaterMask: Instance Segmentation For Underwater Imagery (ICCV'23)
- Future work



习近平总书记在二十大报告中指出——
(四) 促进区域协调发展。发展海洋经济，保护海洋生态环境，加快建设海洋强国。



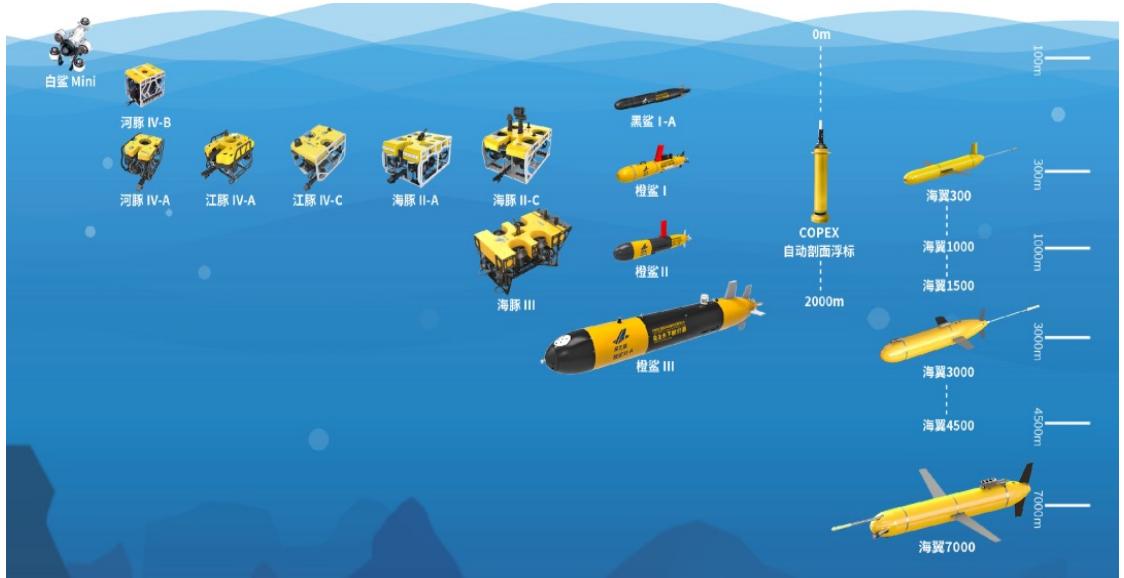
中华人民共和国中央人民政府

《中华人民共和国国民经济和社会发展第十四个五年规划和2035年远景目标纲要》

积极拓展海洋经济发展空间。聚焦新一代信息技术、生物技术、新能源、新材料、高端装备、新能源汽车、绿色环保以及航空航天、海洋装备等战略性新兴产业，加快关键核心技术创新应用，增强要素保障能力，培育壮大产业发展新动能。深化军民科技协同创新，加强海洋、空天、网络空间、生物、新能源、人工智能、量子科技等领域军民统筹发展，推动军地科研设施资源共享，推进军地科研成果双向转化应用和重点产业发展。

智慧海洋工程是全面提升经略海洋能力的整体解决方案。

Introduction



Equipment



Underwater imaging



水下内容感知增强

Underwater Content Perception & Enhancement

Underwater Image Enhancement



Underwater image enhancement methods improve the visibility of underwater images, eliminate color deviation and stretch contrast, and effectively improve the visual quality of images.



Hot Paper



Highly Cited Paper

Top 50 Popular Documents in IEEE TIP

An Underwater Image Enhancement Benchmark Dataset and Beyond

*Chongyi Li, Chunle Guo, Wenqi Ren, Runmin Cong, Junhui Hou,
Sam Kwong, and Dacheng Tao*

IEEE Transactions on Image Processing, 2020

https://li-chongyi.github.io/proj_benchmark.html



Motivation

- Despite the prolific work, both the comprehensive study and insightful analysis of underwater image enhancement algorithms remain largely due to [the lack of a publicly available real-world underwater image dataset](#).
- Lacking sufficient and effective training data, the performance of deep learning-based underwater image enhancement algorithms does not match the success of recent deep learning-based high-level and low-level vision problems.

Contributions

1. We construct **a large-scale real-world underwater image enhancement benchmark** (i.e., UIEB) which contains 950 real underwater images. Moreover, the corresponding reference images for 890 images are provided according to well-designed pairwise comparisons.
2. With the constructed UIEB, we conduct **a comprehensive study of the state-of-the-art single underwater image enhancement algorithms** ranging from qualitative to quantitative evaluations.
3. We propose a **CNN model (i.e., Water-Net) trained by the UIEB for underwater image enhancement**, which demonstrates the generalization of the constructed UIEB and the advantages of our Water-Net, and also motivates the development of deep learning-based underwater image enhancement.

UIEB Dataset

- Underwater images with their corresponding reference images



✓ How to generate the corresponding reference images?

- ◆ It is practically impossible to simultaneously photograph a real underwater scene and the corresponding ground truth image for different water types.



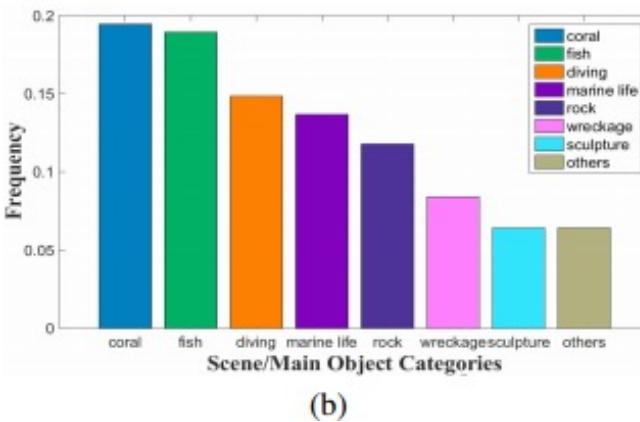
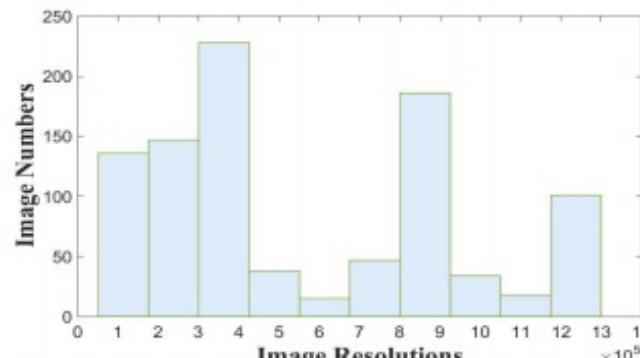
● Data Collection

There are three objectives for underwater image collection

- A diversity of underwater scenes, different characteristics of quality degradation, and a broad range of image content should be covered.
- The amount of underwater images should be large.
- The corresponding high-quality reference images should be provided so that pairs of images enable fair image quality evaluation and end-to-end learning.

UIEB Dataset

- Data Collection



- ◆ These underwater images are collected from Google, YouTube, related papers, and our self-captured videos.
- ◆ We mainly retain the underwater images which meet the first objective.
- ◆ After data refinement, most of the collected images are weeded out, and 950 candidate images are remaining.

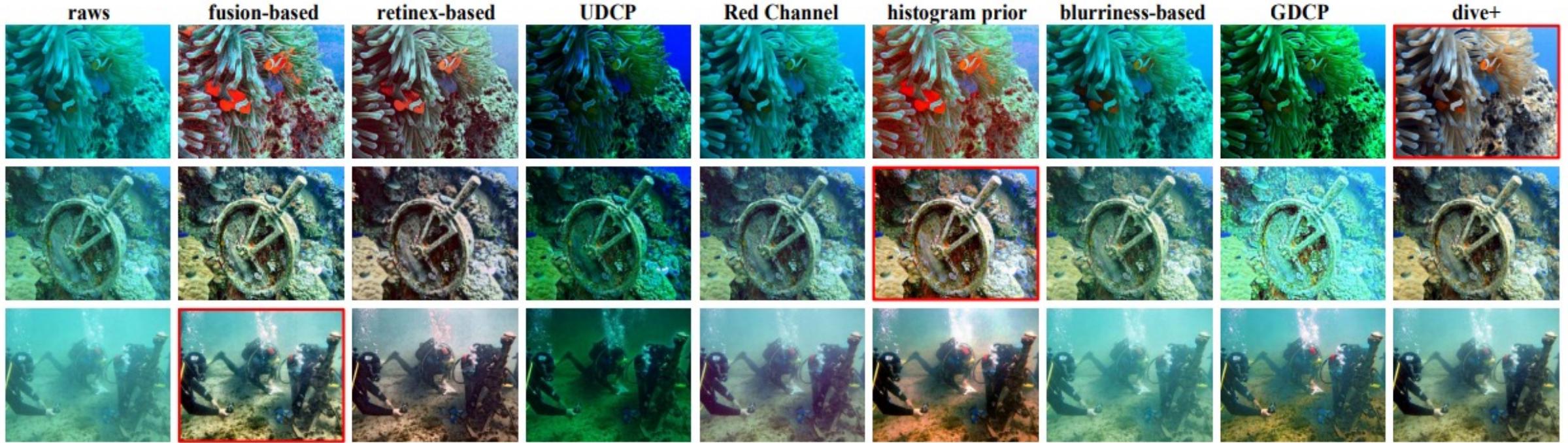


● Reference Image Generation

- ◆ With the candidate underwater images, the potential reference images are generated by [12 image enhancement methods](#), including 9 underwater image enhancement methods, 2 image dehazing methods, 1 commercial application for enhancing underwater images (i.e., dive+).
- ◆ With raw underwater images and the enhanced results, we invite [50 evaluator](#) to perform pairwise comparisons among the 12 enhanced results of each raw underwater image under the same monitor.
- ◆ The reference image for a raw underwater image is first selected by [majority voting](#) after pairwise comparisons.
- ◆ After that, if the selected reference image has [greater than half the number of votes labeled dissatisfaction](#), its corresponding raw underwater image is treated as a [challenging image](#) and the reference image is discarded.
- ◆ We totally achieve [890 available reference images](#) which have higher quality than any individual methods and [a challenging set](#) including 60 underwater images.

UIEB Dataset

- Reference Image Generation



UIEB Dataset

- Reference Image Generation

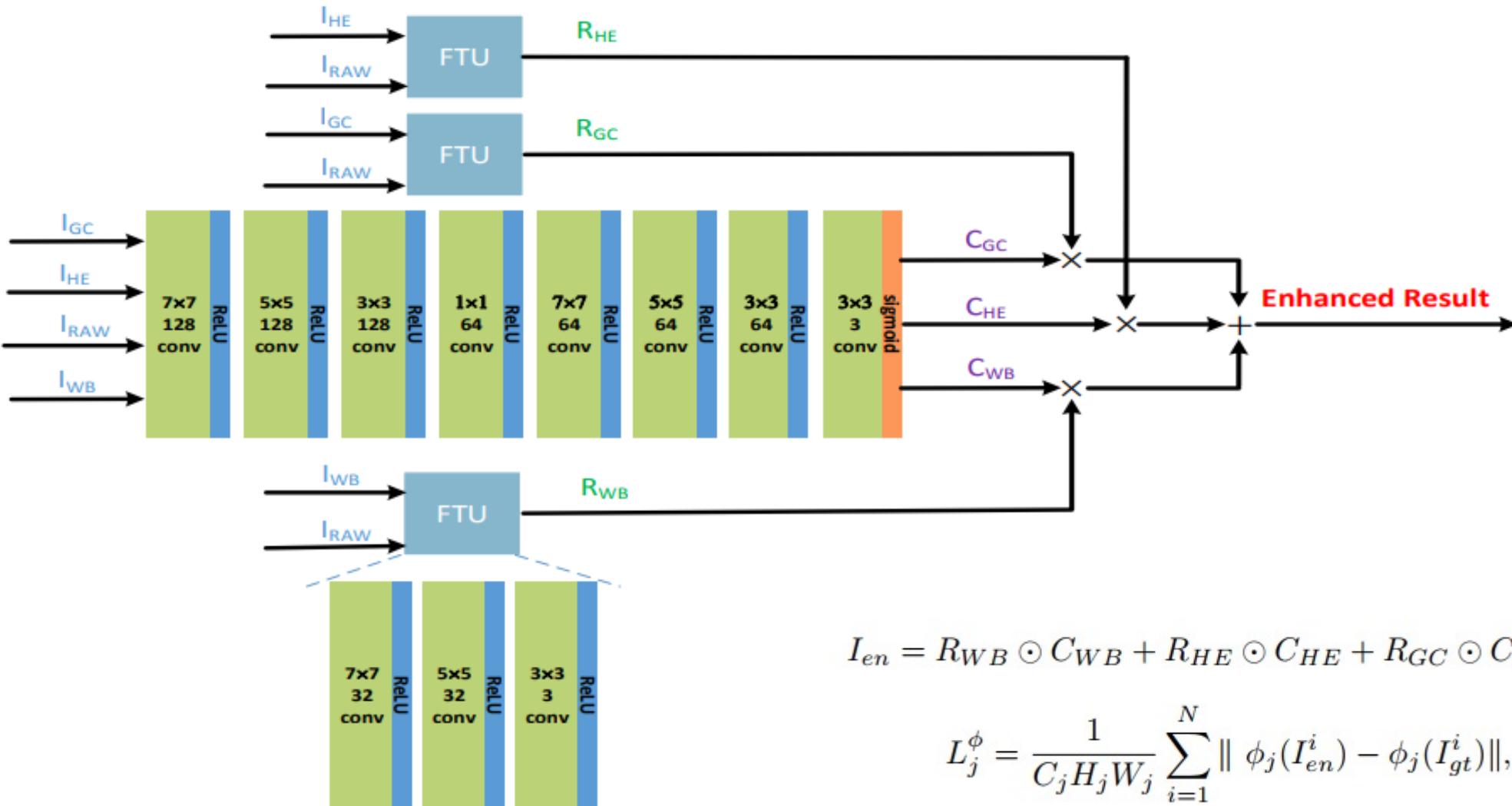
Method	Percentage (%)
fusion-based [31]	24.72
two-step-based [32]	7.30
retinex-based [33]	0.22
DCP [35]	2.58
UDCP [37]	0.00
regression-based [39]	1.80
GDCP [40]	0.34
Red Channel [42]	0.90
histogram prior [45]	13.37
blurriness-based [46]	3.48
MSCNN [66]	0.90
dive+	43.93



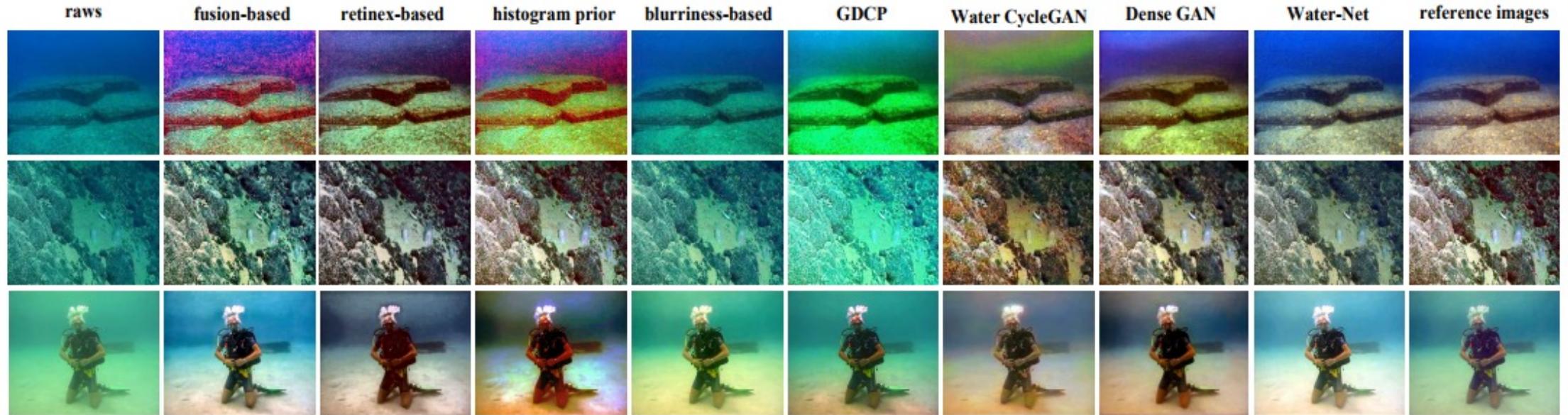
Water-Net

- ◆ The purpose of the proposed Water-Net as a baseline is to call for the development of deep learning-based underwater image enhancement, and demonstrate the generalization of the UIEB for training CNNs.
- ◆ Gated-fusion network: Water-Net
 - ✓ In general, the fusion-based method achieves decent results, which benefits from the inputs derived by multiple pre-processing operations and a fusion strategy. In the proposed Water-Net, we also employ such a manner.
 - ✓ Based on the characteristics of underwater image degradation, we generate three inputs by respectively applying White Balance (WB), Histogram Equalization (HE), and Gamma Correction (GC) algorithms to an underwater image.

Water-Net



Performance



FULL-REFERENCE IMAGE QUALITY ASSESSMENT IN TERMS OF MSE,
PSNR, AND SSIM ON TESTING SET.

Method	MSE ($\times 10^3$) ↓	PSNR (dB) ↑	SSIM ↑
fusion-based [31]	1.1280	17.6077	0.7721
retinex-based [33]	1.2924	17.0168	0.6071
GDCP [40]	4.0160	12.0929	0.5121
histogram prior [45]	1.7019	15.8215	0.5396
blurriness-based [46]	1.9111	15.3180	0.6029
Water CycleGAN [53]	1.7298	15.7508	0.5210
Dense GAN [55]	1.2152	17.2843	0.4426
Water-Net	0.7976	19.1130	0.7971



Conclusion and Future Work

- As analyzed in qualitative and quantitative evaluations, there is no method which always wins in terms of full- and no-reference metrics. In addition, effective non-reference underwater image quality evaluation metrics are highly desirable.
- The existing algorithms follow inaccurate image formation models or assumptions, which inherently limit the performance of underwater image enhancement.
- Note that the use of inaccurate imaging models is a major problem which keeps the field of underwater computer vision at standstill.

It's time for learning-based solutions!



Conclusion and Future Work

- Extending the constructed dataset towards more challenging underwater images and underwater videos.
- Trying to design a range map estimation network. With the estimated range maps, we will make full use of such key prior information to further improve the performance of underwater image enhancement network.
- Re-organizing the selection of the reference images from more reliable results and also further train the volunteers on what the degrading effects of attenuation and backscatter are, and what it looks like when either is improperly corrected.
- Providing multiple reference images for an underwater image and defining the image quality level of their reference images when we re-organize the selection of the reference images.

PUGAN: Physical Model-Guided Underwater Image Enhancement Using GAN with Dual-Discriminators

*Runmin Cong, Wenyu Yang, Wei Zhang, Chongyi Li, Chun-Le Guo,
Qingming Huang, and Sam Kwong*

IEEE Transactions on Image Processing, 2023

https://rmcong.github.io/proj_PUGAN.html



Traditional method

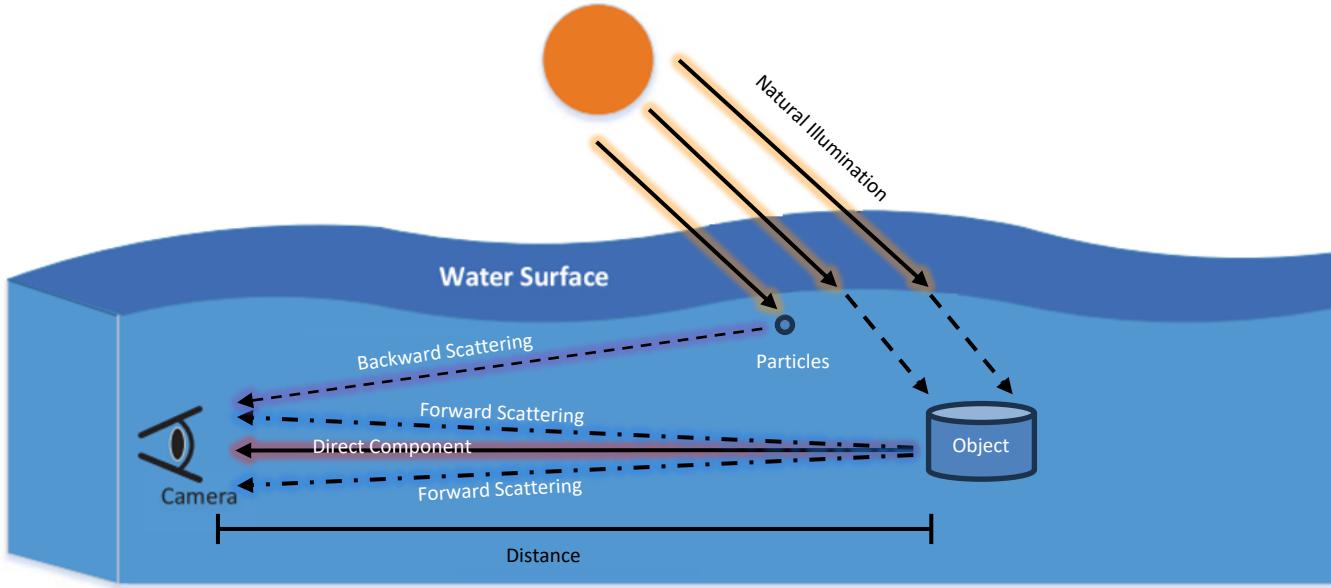
Non physical model method

Some of these methods directly apply existing image enhancement methods to underwater image data, and there are also specialized algorithms designed **specifically** for the characteristics of underwater images.

Physical models-based method

Mathematical modeling of the degradation process of underwater images, parameter estimation based on the model, and then inversion to obtain clear underwater images.

Traditional method



$$I(x) = J(x)t(x) + A(1 - t(x))$$

where I is the observed underwater image, J denotes the restored image, A represents the background light, and t is the transmission map, describing the portion of the light that is not scattered and reaches the camera.



Motivation

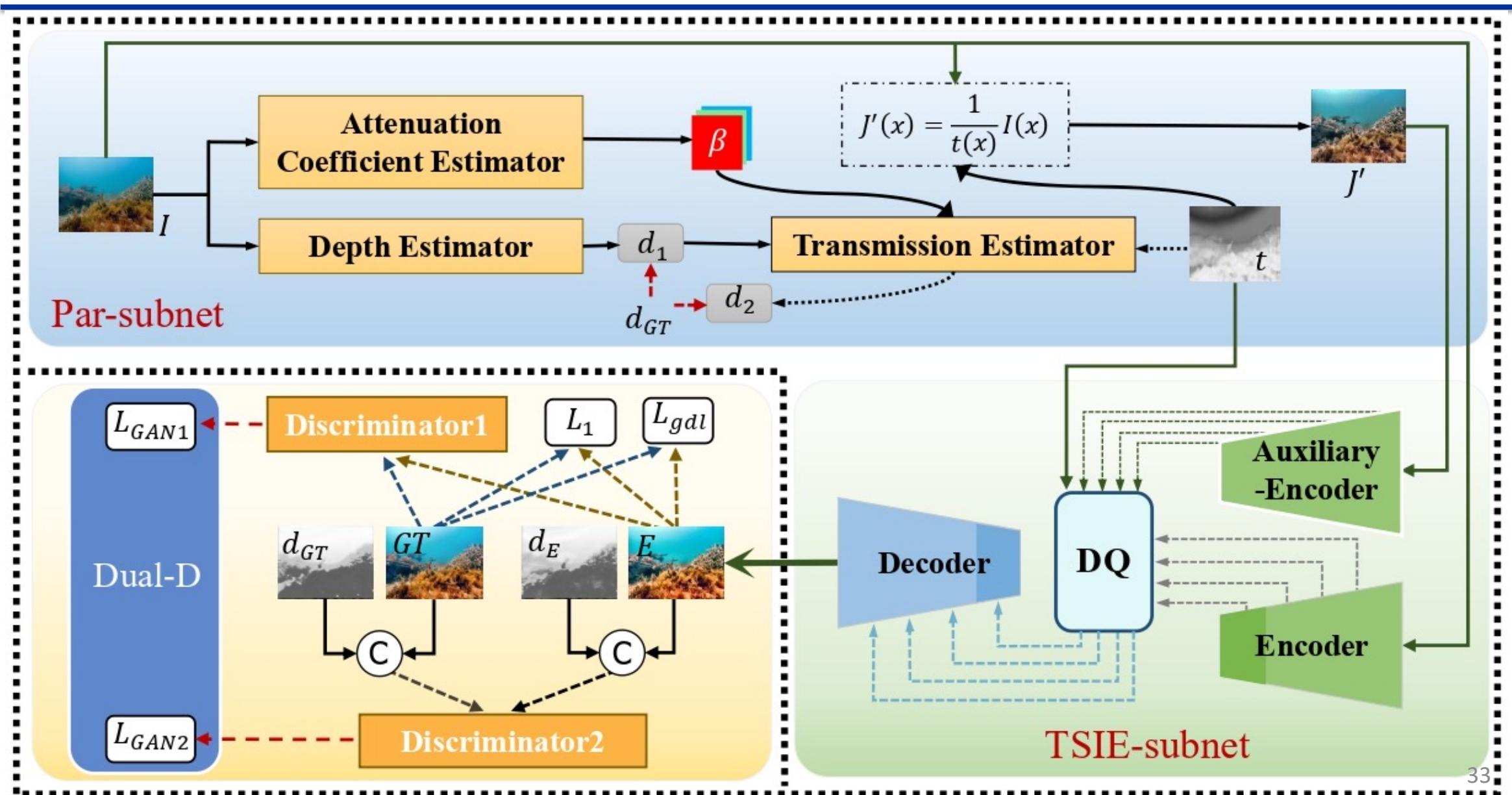
- Traditional methods based on **non-physical models largely rely on handmade feature design**, which makes them prone to over or under enhancement, thereby affecting the overall visual effect. Although modeling the underwater imaging process is beneficial for solving the unique visual problems of underwater images, **relying solely on physical models is not reliable because it is difficult to simulate a universal model to cope with complex underwater environments.**
- The deep learning method utilizes the powerful learning ability of deep networks and can achieve good results in certain situations. However, underwater environments are often complex and diverse, and **relying solely on network learning may distort the enhanced results.**

Therefore, we hope to design a network architecture that can effectively combine them to play to complementary advantages and collaborative promotion.

Contributions

- Considering the respective advantages of the physical model and the GAN model for the UIE task, we propose a **Physical Model-Guided framework using GAN with Dual-Discriminators (PUGAN)**, consisting of a Phy-G and a Dual-D. Extensive experiments on three benchmark datasets demonstrate that our PUGAN outperforms state-of-the-art methods in both qualitative and quantitative metrics.
- We design **two subnetworks in the Phy-G**, including the **Par-subnet** and the **TSIE-subnet**, for the parameter estimation of physical model and the physical model guided CNN-based enhancement, respectively. On the one hand, we introduce an intermediate variable in the Par-subnet, i.e., depth, to enable effective estimation of the transmission map. On the other hand, we propose a **DQ module in TSIE-subnet** to quantify the distortion degrees and achieve targeted encoder feature reinforcing.
- In addition to the pixel-level global similarly loss and perceptual loss, we design **the style-content adversarial loss in the Dual-D** to constrain the style and content of the enhanced underwater image to be realistic.

Our Method



Par-subnet

The physical model of the underwater imaging process:

$$I(x) = J(x)t(x) + A(1 - t(x))$$

$$t(x) = e^{-\beta d(x)}$$

t is the transmission map, describing the portion of the light that is not scattered and reaches the camera, β is the attenuation coefficient of the water, and d is the depth of scene. Therefore, the depth can also reflect the attenuation of the scene to a certain extent.

We can inversely derive the calculation formula of the enhanced image J' as:

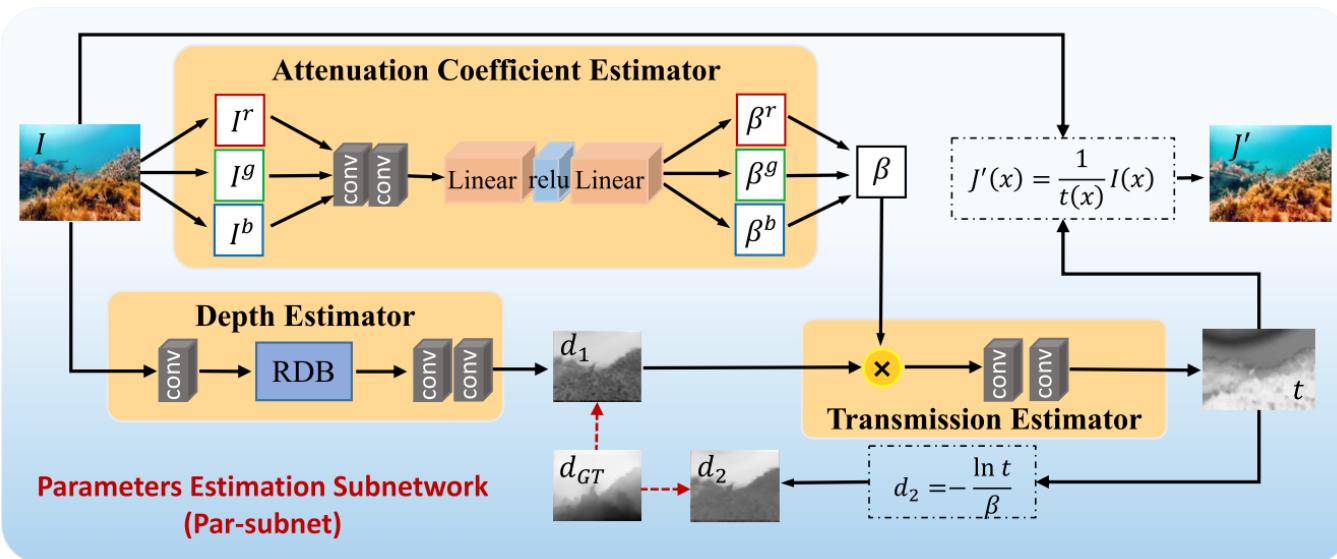
$$J(x) = \frac{1}{t(x)} I(x) - A\left(\frac{1}{t(x)} - 1\right)$$

correct the color

remove the influence of background light

obtaining color-corrected underwater images through physical model inversion during the first stage

$$J'(x) = \frac{1}{t(x)} I(x)$$



$$\beta^c = \text{linear}(\text{relu}(\text{linear}(\text{conv.p.r}(I^c))))$$

$$\beta = \text{cat}(\beta^r, \beta^g, \beta^b)$$

$$d_1 = \sigma(\text{conv}(\text{conv.b.r}(\text{RDB}(\text{conv.b.r}(I)))))$$

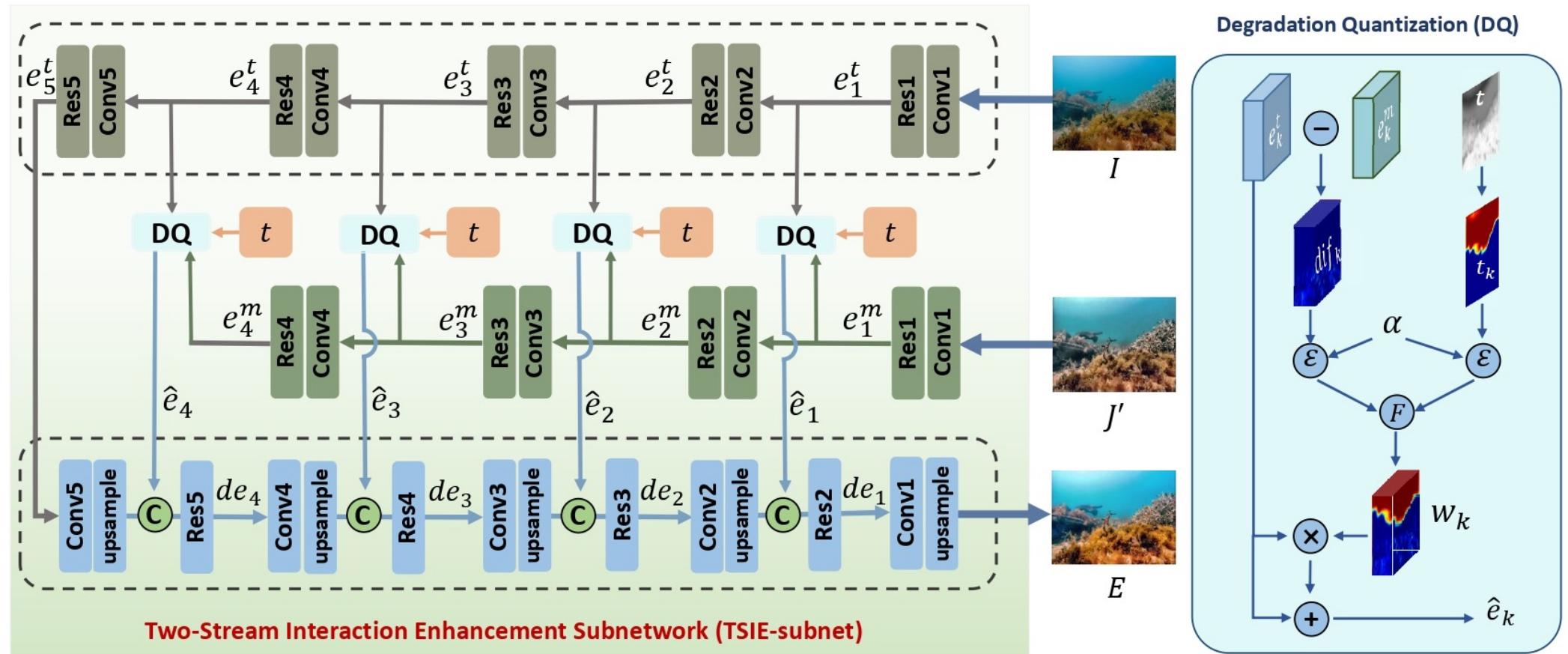
$$t = \sigma(\text{conv}(\text{conv.b.r}(d_1 \cdot \beta)))$$

$$d_2 = -\frac{\ln t}{\beta}$$

TSIE-subnet

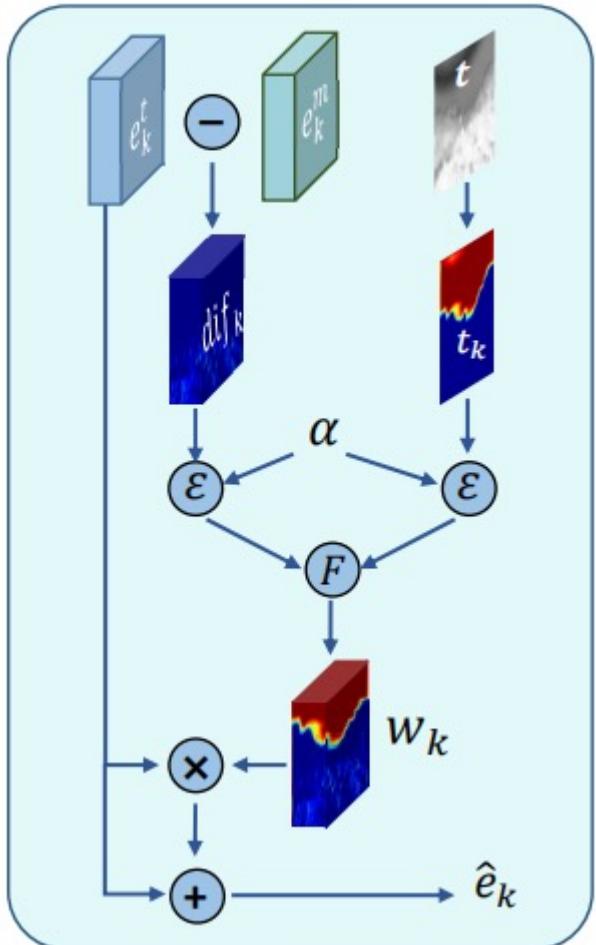
In the first stage, we invert color-enhanced underwater images with better interpretability using the learned physical model parameters. But as mentioned before, **the enhancement effect is not perfect due to the exclusion of background light.**

Therefore, we re-enhance the underwater images under the CNN network architecture in the second stage guided by the color-enhanced images, thereby forming a two-stream architecture to realize the interaction of multi-source information.



TSIE-subnet

Degradation Quantization (DQ)



On the one hand, we can locate severely degraded regions by directly comparing the difference between the color enhanced image features and the original image features, which can be described as:

$$dif_k = conv.b.r(|e_k^t - e_k^m|) \cdot \varepsilon(conv.b.r(|e_k^t - e_k^m|) - \alpha)$$

On the other hand, the degree of degradation of underwater images is negatively correlated with the transmission characteristics. Therefore, we can also identify some regions that are prone to degradation from the transmission map:

$$t_k = (1 - maxpool(t)) \cdot \varepsilon(1 - maxpool(t) - \alpha)$$

Combining these two aspects, the final weights can be defined as follows:

$$w_k = \sigma(conv((conv.b.r(t_k + dif_k))))$$

Subsequently, these weights are applied to the input features e^k to generate the updated features \hat{e}^k through the residual connection:

$$\hat{e}_k = e_k^t + e_k^t \otimes w_k$$

Loss Function

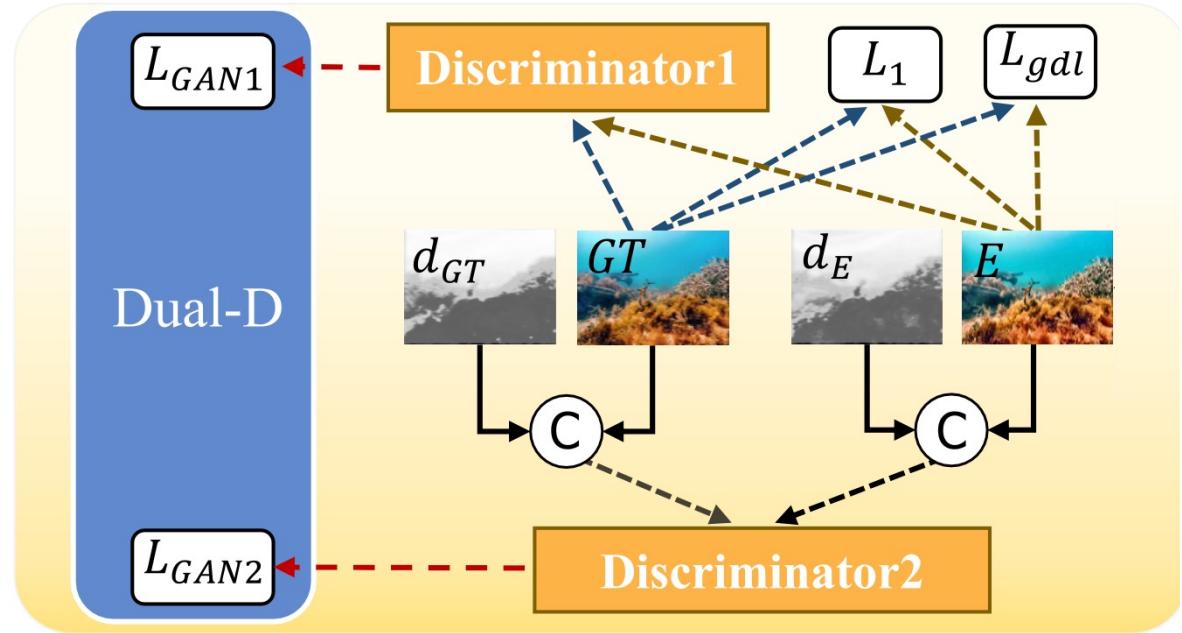
For Par-subnet:

We first train the attenuation coefficient estimator and then freeze their parameters to train the depth estimator and transmission estimator. To control the accuracy of the transmission map, we use the transmission map and attenuation coefficient to compute the depth map again. Therefore, the loss of Par-subnet is defined as follows:

$$L_p = \frac{1}{H \times W} \left[\sum_{m=1}^H \sum_{n=1}^W (|d(m, n) - d_1(m, n)|) + \sum_{m=1}^H \sum_{n=1}^W (|d(m, n) - d_2(m, n)|) \right] + \frac{1}{3} \sum_{c=1}^3 (|\hat{\beta}^c - \beta^c|)$$

For Phy-G:

In order to make the generated image as visually pleasing as possible while maintaining its authenticity of the image, we use global similarly loss, perceptual loss and adversarial loss to compose the final loss:

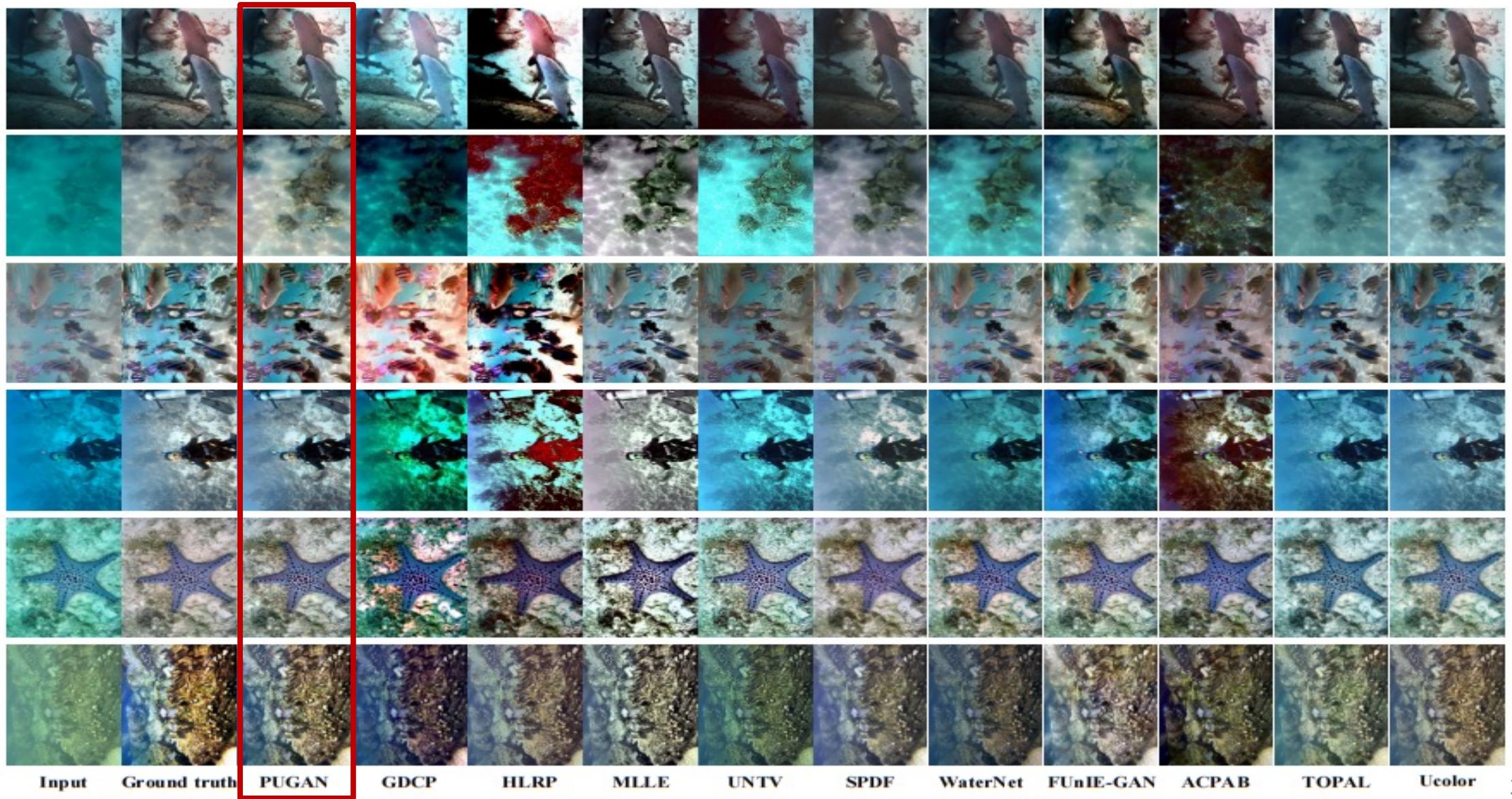


$$\begin{aligned} L = & \lambda_1 \cdot \arg \min_G \max_{D_1} L_{GAN1}(G, D_1) \\ & + \lambda_2 \cdot \arg \min_G \max_{D_2} L_{GAN2}(G, D_2) \\ & + \lambda_3 \cdot L_1(E, Y) + \lambda_4 \cdot L_{gdl}(E, Y) \end{aligned}$$

$$\arg \min_G \max_{D_1} L_{GAN1}(G, D_1) = \mathbb{E}_{\{I, Y\}} [\log D_1(Y)] + \mathbb{E}_{\{I, Y\}} [\log(1 - D_1(E))]$$

$$\arg \min_G \max_{D_2} L_{GAN2}(G, D_2) = \mathbb{E}_{\{I, Y, d\}} [\log D_2(Y, d_Y)] + \mathbb{E}_{\{I, Y, d\}} [\log(1 - D_2(E, d_E))]$$

Experiments



Experiments

Datasets	Test-UIEB		Test-UFO		Test-EUVP	
Methods	PSNR↑	MSE↓	PSNR↑	MSE↓	PSNR↑	MSE↓
GDCP [4]	13.72	3.37	14.33	2.87	13.35	3.58
ACDE [25]	16.85	1.67	14.31	2.83	15.03	2.35
HLRP [49]	12.17	4.24	11.69	4.66	11.32	5.08
MLLE [50]	18.82	1.12	15.05	2.45	15.06	2.32
UNTV [51]	16.57	1.88	17.12	1.42	17.50	1.39
SPDF [52]	19.85	0.92	17.57	1.37	18.82	1.09
deep-sesr [41]	15.77	2.08	23.22	0.38	23.22	0.35
FUnIE-GAN [5]	18.07	1.78	22.97	0.41	23.53	0.41
WaterNet [2]	19.81	1.02	19.63	0.83	20.58	0.71
UWCNN [28]	13.26	4.00	16.41	1.98	17.72	1.40
JI-Net [38]	18.21	2.46	16.54	1.78	–	–
ACPAB [35]	15.20	2.52	17.04	1.73	18.06	1.40
TOPAL [37]	19.85	0.93	19.31	0.83	19.98	0.75
Ucolor [6]	20.61	0.78	19.45	0.85	20.08	0.76
PUGAN	21.67	0.54	23.70	0.32	24.05	0.34

Experiments

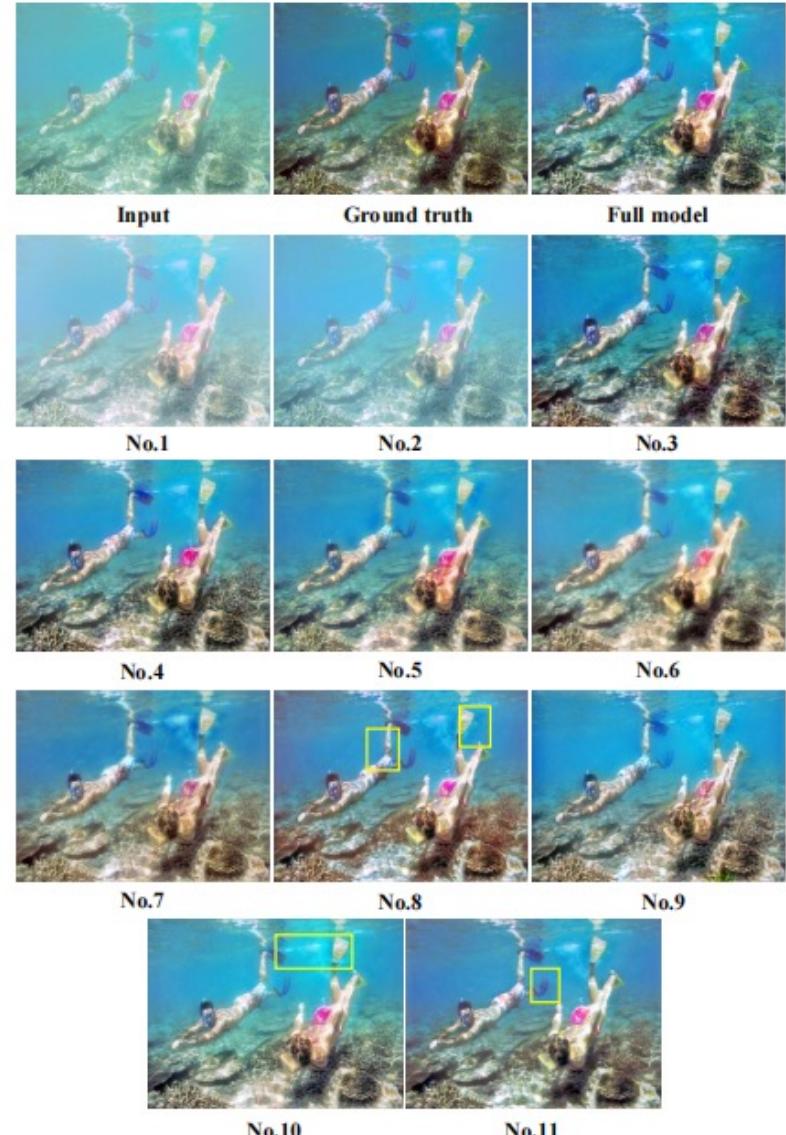
Datasets	Test-UIEB				Test-UFO				Test-EUVP			
Methods	UIQM↑	FDUM↑	UICQE↑	CCF↑	UIQM↑	FDUM↑	UICQE↑	CCF↑	UIQM↑	FDUM↑	UICQE↑	CCF↑
input	2.69	0.36	0.52	19.59	2.48	0.48	0.56	30.03	2.49	0.45	0.55	30.27
Ground truth	3.01	0.55	0.62	27.34	2.88	0.67	0.60	28.53	2.88	0.62	0.58	31.11
GDCP [4]	2.67	0.84	0.61	47.28	2.10	0.81	0.66	62.83	2.43	0.87	0.63	57.92
ACDE [25]	3.41	0.49	0.56	29.05	3.35	0.51	0.57	33.44	3.30	0.43	0.56	33.38
HLRP [49]	1.99	0.81	0.66	55.25	2.47	0.81	0.67	63.23	2.41	0.75	0.65	64.56
MLLE [50]	2.65	0.66	0.61	40.12	2.39	0.76	0.62	56.43	2.28	0.69	0.61	60.31
UNTV [51]	2.94	0.72	0.59	26.37	2.60	0.80	0.62	38.81	2.47	0.77	0.62	40.78
SPDF [52]	3.08	0.44	0.56	17.46	3.18	0.50	0.56	22.96	3.19	0.27	0.55	24.54
deep-sesr [41]	2.97	0.41	0.53	15.97	3.07	0.61	0.59	23.90	3.10	0.54	0.57	24.34
FUnIE-GAN [5]	3.34	0.68	0.56	21.38	2.97	0.58	0.60	27.85	2.99	0.56	0.59	30.10
WaterNet [2]	3.04	0.44	0.58	16.68	3.08	0.53	0.59	25.60	3.06	0.50	0.58	27.17
UWCNN [28]	2.21	0.28	0.48	10.65	2.93	0.28	0.52	15.91	2.96	0.39	0.52	19.02
JI-Net [38]	2.67	0.57	0.59	25.98	3.17	0.54	0.59	28.70	3.24	0.67	0.58	27.38
ACPAB [35]	2.92	0.56	0.58	33.66	3.06	0.51	0.58	33.78	2.98	0.45	0.58	35.90
TOPAL [37]	3.08	0.48	0.57	22.82	3.02	0.36	0.61	28.85	3.01	0.32	0.43	28.50
Ucolor [6]	3.30	0.43	0.57	17.65	3.14	0.52	0.59	24.53	3.12	0.49	0.58	26.51
PUGAN	3.28	0.68	0.62	27.94	2.85	0.64	0.60	33.49	2.94	0.53	0.60	30.34

Experiments

Input	Ground Truth	GDCP	HLRP	MLLE	UNTV	PUGAN
						
UIQM / UICQE 3.24 / 0.45	UIQM / UICQE 3.51 / 0.61	UIQM / UICQE 3.11 / 0.61	UIQM / UICQE 1.83 / 0.68	UIQM / UICQE 3.41 / 0.59	UIQM / UICQE 3.53 / 0.53	UIQM / UICQE 3.62 / 0.62
FDUM / CCF 0.25 / 12.84	FDUM / CCF 0.55 / 30.77	FDUM / CCF 0.82 / 45.38	FDUM / CCF 0.86 / 78.75	FDUM / CCF 0.54 / 32.15	FDUM / CCF 0.54 / 18.78	FDUM / CCF 0.64 / 33.42
						
UIQM / UICQE 2.61 / 0.41	UIQM / UICQE 3.42 / 0.62	UIQM / UICQE 2.71 / 0.53	UIQM / UICQE 2.36 / 0.68	UIQM / UICQE 3.32 / 0.60	UIQM / UICQE 3.25 / 0.49	UIQM / UICQE 3.45 / 0.61
FDUM / CCF 0.22 / 8.51	FDUM / CCF 0.56 / 24.90	FDUM / CCF 0.57 / 26.96	FDUM / CCF 0.81 / 76.09	FDUM / CCF 0.55 / 23.55	FDUM / CCF 0.47 / 12.09	FDUM / CCF 0.63 / 25.15

Ablation Study

		PSNR↑	MSE↓
Full model (E)		21.67	0.54
Par-subnet	No.1	J'	18.59
	No.2	J'^*	19.00
	No.3	E^*	21.48
	No.4	w/o Estimator (t)	21.08
TSIE-subnet	No.5	single-stream with I	19.87
	No.6	single-stream with J'	20.03
	No.7	w/o DQ	19.88
	No.8	w/o dif_k	20.71
	No.9	w/o t_k	20.08
Dual-D	No.10	w/o L_{GAN_1}	21.00
	No.11	w/o L_{GAN_2}	20.93





Conclusion

- In this paper, we propose a **physical model-guided GAN model** for underwater image enhancement.
- In the phy-G, we fully combine the physical model and the CNN-based model, where the **Par-subnet** generates the color enhanced underwater image by physical inversion, and the **TSIE-subnet** equipped with a DQ module aims to generate the final enhanced image through the regional and differential feature learning.
- In addition, we design a novel **Dual-D structure** to judge the reconstruction results of the generator, following a style-content synergy mechanism.
- Our extensive experiments on different benchmarks demonstrate the superiority of this method and the effectiveness of each module.



水下环境内容理解

Underwater Environment Content Understanding

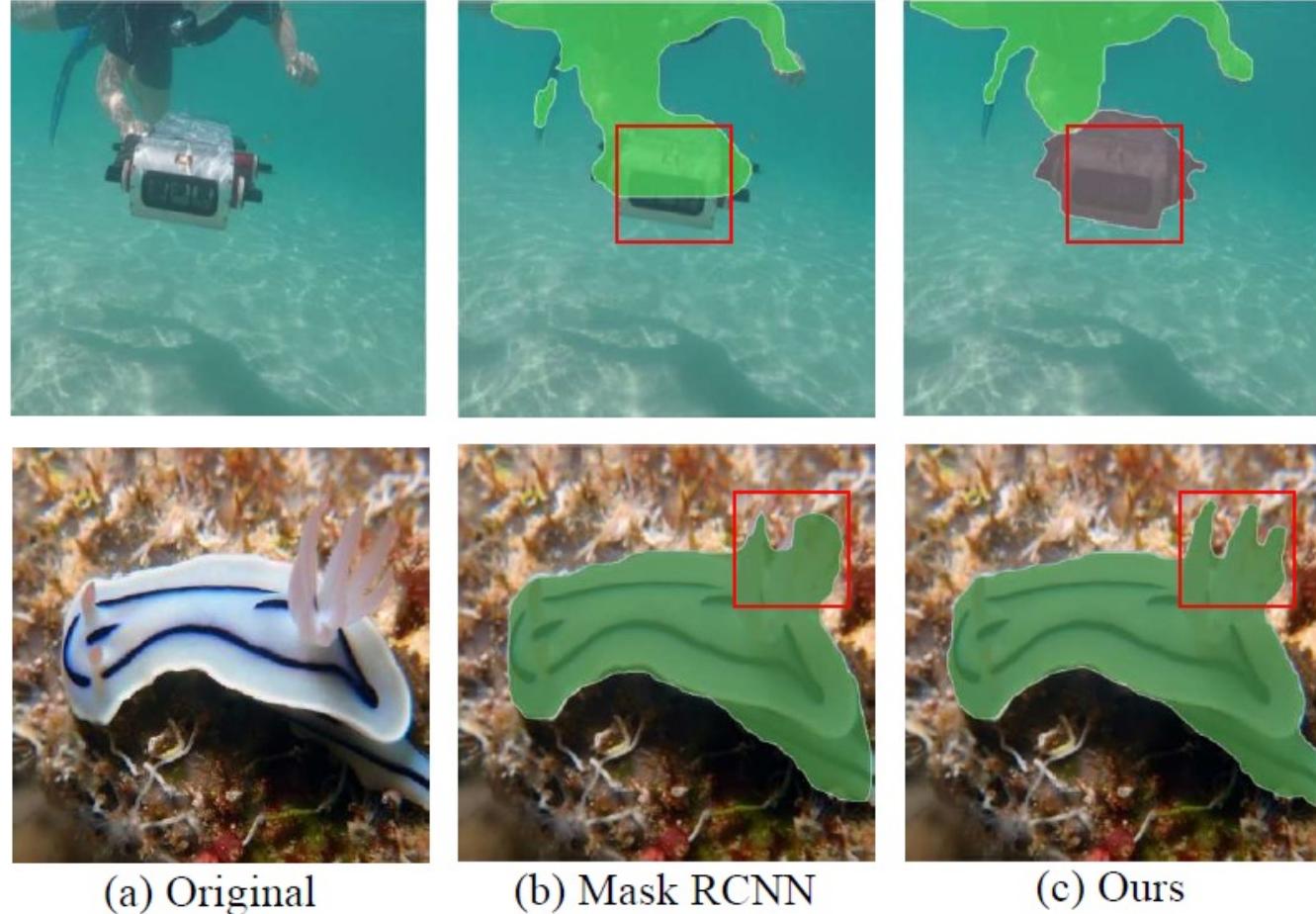
WaterMask: Instance Segmentation for Underwater Imagery

Shijie Lian, Hua Li, Runmin Cong, Suqi Li, Wei Zhang, and Sam Kwong

IEEE International Conference on Computer Vision, 2023

<https://github.com/LiamLian0727/WaterMask>

Introduction



- Since instance segmentation is valuable in estimating object interactions and inferring scene geometry, it is of great use in many underwater vision scenarios such as underwater robot vision and underwater vehicle autopilot.
- However, the segmentation of image instances for general underwater scenes has not been thoroughly explored. The results of directly applying natural image segmentation models to underwater images with generally degraded quality are often unsatisfactory!



Motivation

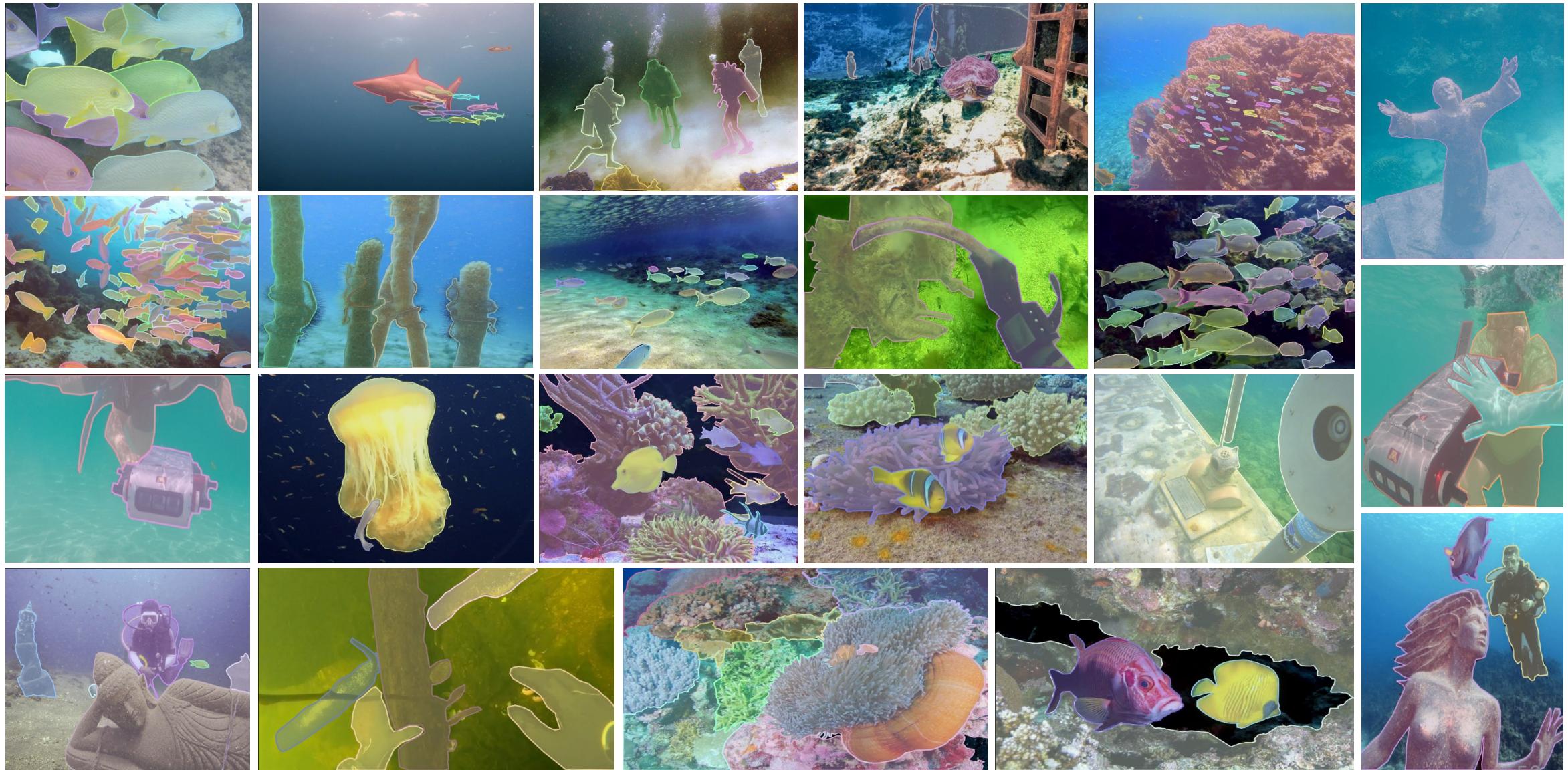
- On the one hand, there is **no general underwater image instance segmentation dataset** to promote training and evaluation of instance segmentation models. On the other hand, quality degradation of underwater images is inevitable due to wavelength and distance-related attenuation and scattering. **Low-quality images** often lead to the **failure of current segmentation methods**.
- To alleviate this issue, we propose the **first underwater image instance segmentation (UIIS) dataset**, aiming to promote the development of instance segmentation for underwater tasks.
- Simultaneously, we propose **WaterMask** for multi-object underwater image instance segmentation according to the **intrinsic characteristics of underwater imagery**.



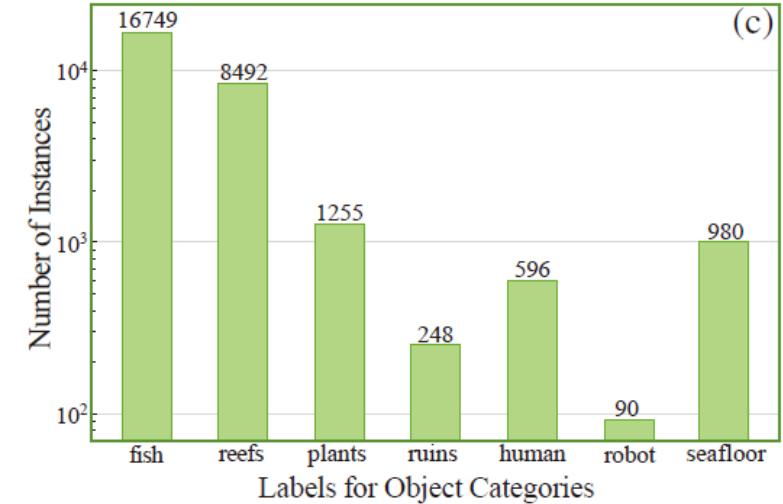
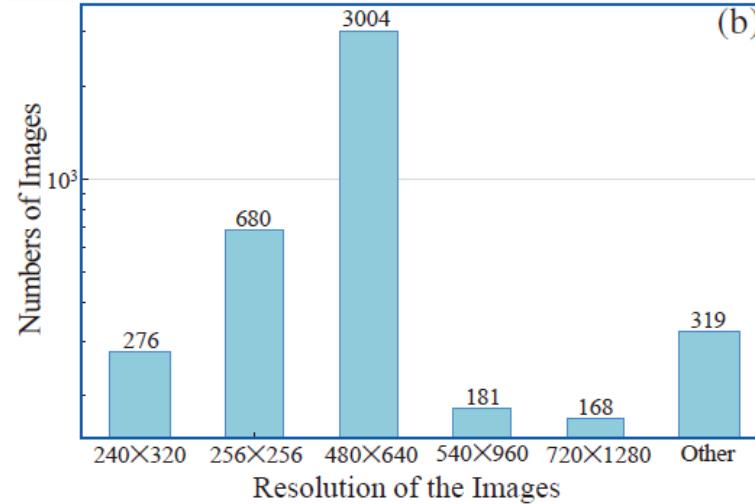
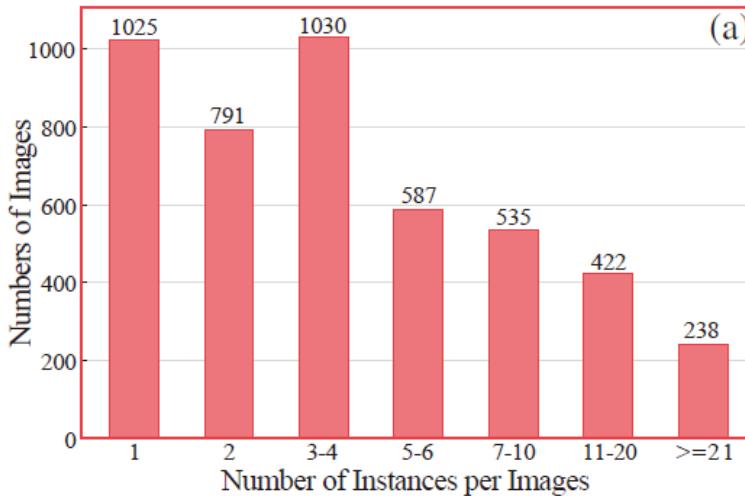
Contributions

- a) We construct the first general underwater image instance segmentation (UIIS) dataset **containing 4,628 images for 7 categories with pixel-level annotations** for underwater instance segmentation task.
- b) We propose the **first underwater instance segmentation model WaterMask**, as far as we know. In WaterMask, we devise **DSGAT and MFRM modules to reconstruct and refine** the image features with underwater imaging degradation, and Boundary Mask Strategy with boundary learning loss to **optimize the boundaries** of underwater clustered instances.
- c) Extensive experiments on popular evaluation criteria demonstrate the **effectiveness** of the proposed UIIS dataset and WaterMask.

Our UIIS Dataset



Dataset Statistic and Challenges



➤ Challenge in the number of instance.

We counted the number of instances in the dataset and the scenes with more than 5 instances accounted for 38.5% of the total and more than 10 instances accounted for 14.2%, in which the image with the most instances had 162 instances.

➤ Challenges in small or large instances.

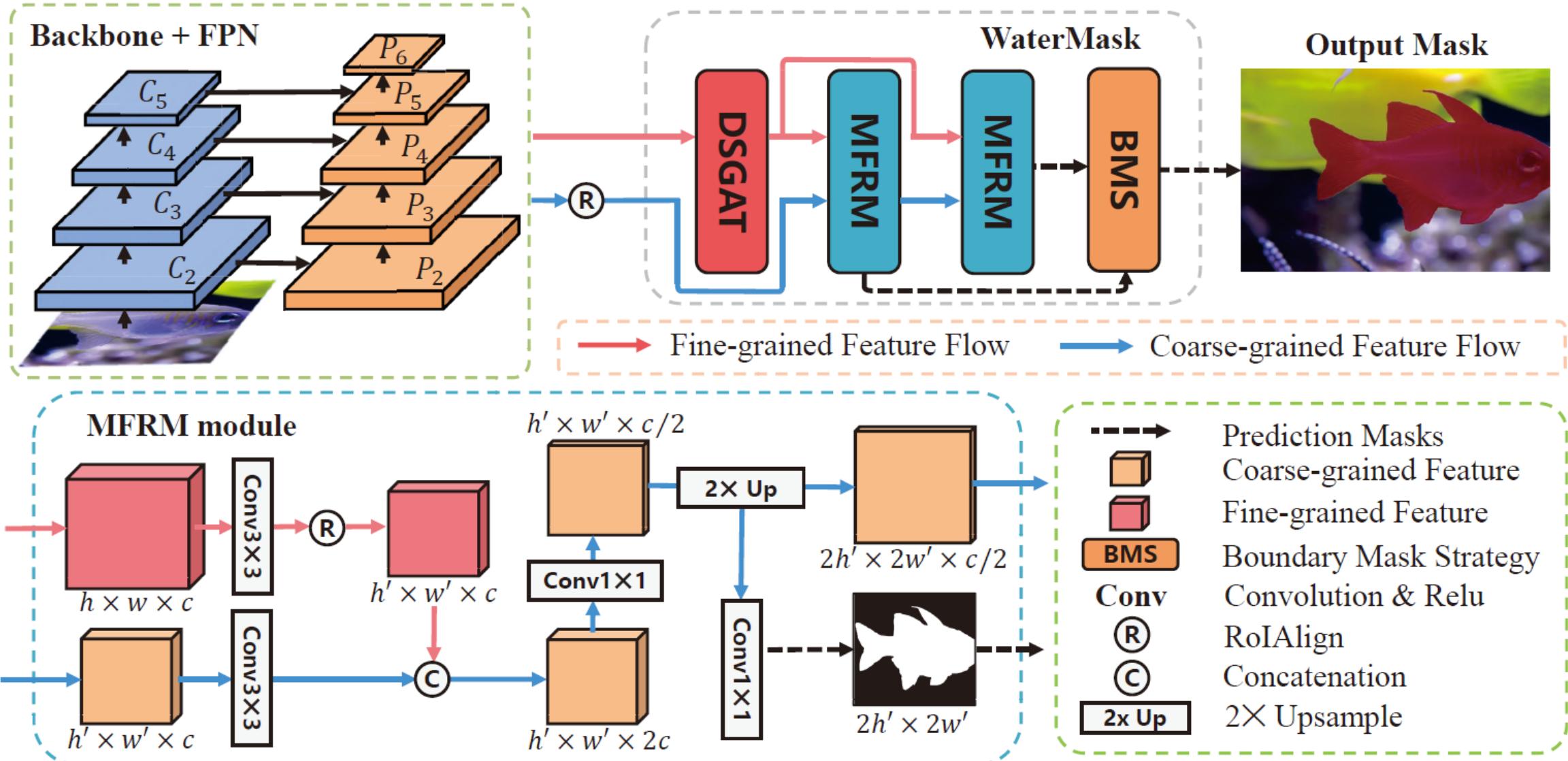
UIIS dataset have 3319 instances less than 14×14 pixels, accounting for 11.7% of the total, in addition to 6485 instances of size larger than 128x128 pixels, accounting for 22.8% of the total.

➤ Challenges in various image resolutions and image scenarios.

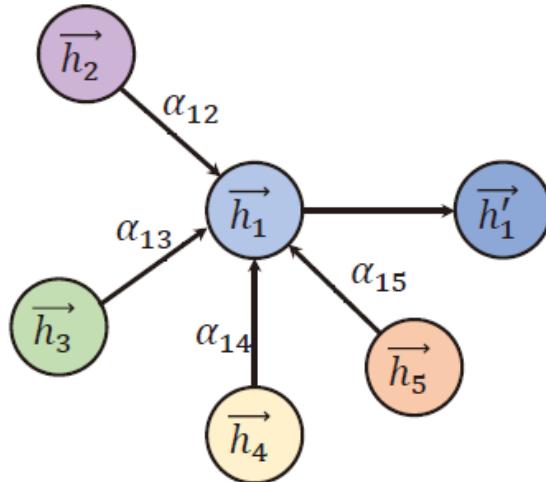
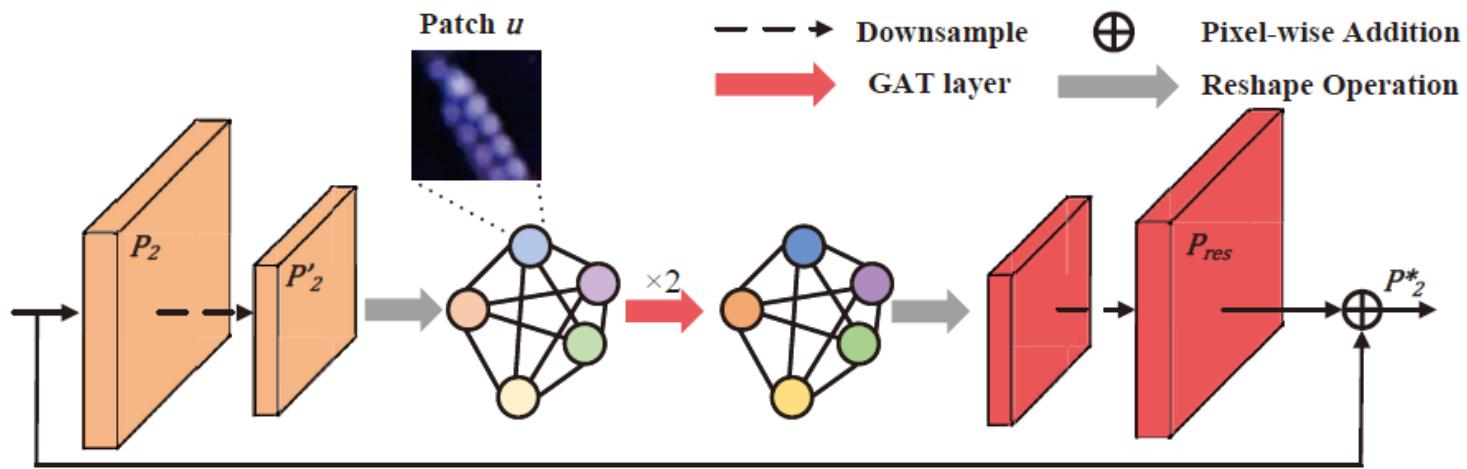
Contains images of various resolutions to match handheld camera shots or industrial equipment shots.

Contains images with significantly degraded quality, high saturation or high contrast images to evaluate the performance of the network in different ocean scenarios.

WaterMask



Difference Similarity Graph Attention Module



Although underwater images generally suffer from quality degradation, underwater instances are **mostly clustered**, which makes it possible for underwater images to **have similar visual information in multiple places**, retaining different degraded details under different water and lighting conditions. Therefore, we propose DSGAT for collecting this similar visual information by computing the attention between image patches so that each patch can be **complemented** by the **visual information of multiple other similar patches**, and reconstructing the image details by extracting and combining information through GAT operations.

$$a_{ij} = \frac{\exp(\sigma(l^T [W\vec{h}_i \| W\vec{h}_j]))}{\sum_{n \in \mathcal{N}_i} \exp(\sigma(l^T [W\vec{h}_i \| W\vec{h}_n]))},$$

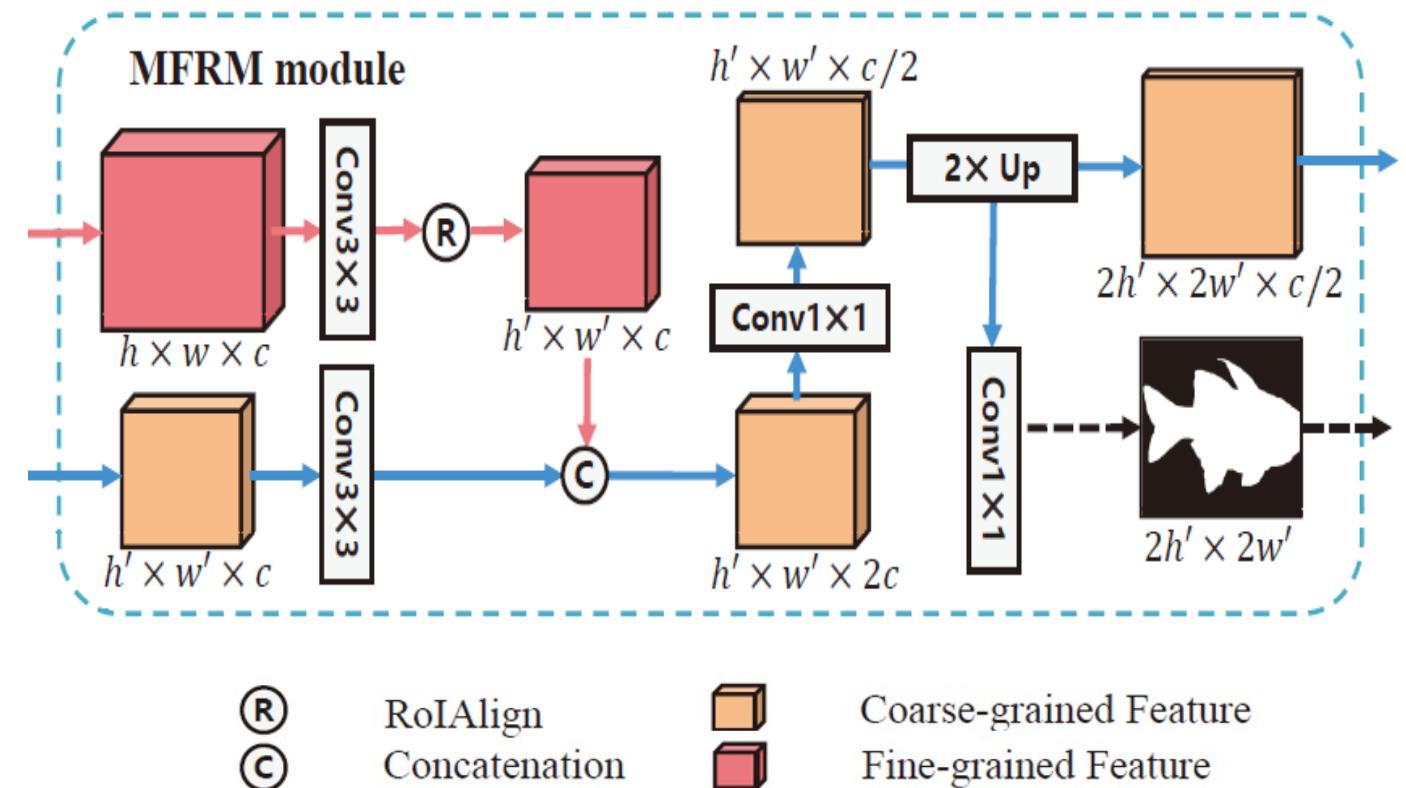
$$\vec{h}'_i = \delta \left(\sum_{n \in \mathcal{N}_i} a_{ij} W\vec{h}_j \right).$$

Multi-level Feature Refinement Module

We then design the **Multi-level Feature Refinement Module (MFRM)**, which infers different resolution masks by supplementing the degradation information so that **higher resolution features can be utilized** to fully predict the boundaries.

MFRM sends the features extracted from the feature pyramid by 14×14 RoIAlign operation to two 3×3 convolutional layers to generate the **initial instance feature F1**. After that, we utilize the fine-grained features generated by DSGAT to iteratively refine the initial F1 by MFRM.

The MFRM will be executed twice, outputting features F2 and F3, which will be used as foreground and boundary predictions, respectively.





Boundary Mask Strategy

We feed the features F_2 and F_3 into the 1×1 convolution layer to generate instance masks M_2 and M_3 with different resolutions. The pixels in F_2 have a large receptive field and contain rich high-level information, which is beneficial for predicting the approximate location of the instance mask, but because of the low feature resolution, the boundaries of the prediction results tend to be rough. Conversely, F_3 , while the high-resolution mask reduces the boundary error, also causes the network to overpredict other pixels of the mask. Therefore, we use M_2 and M_3 to splice our output together, with $B_{2\times} = f_{2\times}(B(M_2))$ and $R_{2\times} = f_{2\times}(B(M_2) \vee B(G_2))$ in the following equation.

Boundary Learning Loss

The boundaries of underwater instances are often blurred, and the pixels used for training boundary classification are much smaller than those used for mask classification, leading to the fact that the commonly used BCE loss is not effective in helping the network to learn information from the boundary. We design the Boundary Learning Loss (BLL) to assign more weights to the boundary regions, thus forcing the network to pay more attention to the classification within the boundary pixels and thus make more accurate predictions.

Calculated Output:

$$p_{ij} = \begin{cases} b^2 - 1, & \text{if } i = j = \frac{b-1}{2} \\ -1, & \text{otherwise,} \end{cases} \quad B(M) = \begin{cases} 1, & \text{if } |\nabla^2 p(M)| \leq \mu b^2 \\ 0, & \text{otherwise,} \end{cases}$$

$$M_{out} = f_{2\times}(M_2) \odot B_{2\times} + M_3 \odot (1 - B_{2\times})$$

Loss Function:

$$\mathcal{L}_B = \frac{\sum_i^{H \times W} R_{2\times}^i \cdot BCE(M_3^i, G_3^i)}{\sum_i^{H \times W} R_{2\times}^i} \quad \mathcal{L}_{mask} = \mathcal{L}_B + \sum_{k \in [1,2]} \lambda_k \mathcal{L}_{BCE}(M_k, G_k)$$

Experiments

Method	Backbone	Schedule	mAP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L	AP _f	AP _h	AP _r
Mask R-CNN	R50-FPN	1×	21.7	39.5	21.0	8.2	18.3	29.9	42.0	42.0	16.6
WaterMask R-CNN	R50-FPN	1×	23.3	39.7	24.8	8.2	19.2	33.7	43.8	46.5	14.4
Mask R-CNN [‡]	R50-FPN	3×	23.5	42.3	23.7	7.8	19.3	34.9	44.3	46.4	15.8
WaterMask R-CNN[‡]	R50-FPN	3×	26.4	43.6	28.8	9.1	21.1	38.1	46.9	54.0	18.2
Mask R-CNN	R101-FPN	1×	22.3	40.2	24.5	8.0	19.7	30.7	42.8	46.3	16.7
WaterMask R-CNN	R101-FPN	1×	25.6	41.7	27.9	8.8	21.3	36.0	45.3	53.9	19.0
Mask R-CNN [‡]	R101-FPN	3×	23.4	40.9	25.3	9.3	19.8	32.5	43.6	49.0	18.0
WaterMask R-CNN[‡]	R101-FPN	3×	27.2	43.7	29.3	9.0	21.8	38.7	46.3	54.8	20.9
Cascade Mask R-CNN [‡]	R101-FPN	3×	25.5	42.8	27.8	7.5	20.1	35.0	43.9	52.9	22.3
Cascade WaterMask R-CNN[‡]	R101-FPN	3×	27.1	42.9	30.4	8.3	21.0	38.9	47.0	55.8	22.5

Table 2: Comparison with Mask R-CNN and Cascade Mask R-CNN on UIIS dataset. Models with [‡] were trained with 3× schedule using multi-scale training.

Experiments

Method	Backbone	mAP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L	AP _f	AP _h	AP _r	Params
Mask R-CNN [‡] [13]	ResNet-101	23.4	40.9	25.3	9.3	19.8	32.5	43.6	49.0	18.0	63M
Mask Scoring R-CNN [‡] [14]	ResNet-101	24.6	41.9	26.5	8.4	20.0	34.3	44.2	52.8	16.0	79M
Cascade Mask R-CNN [‡] [3]	ResNet-101	25.5	42.8	27.8	7.5	20.1	35.0	43.9	52.9	22.3	88M
BMask R-CNN [‡] [7]	ResNet-101	22.1	36.2	24.4	5.8	17.5	35.0	40.7	50.0	17.7	66M
Point Rend [20]	ResNet-101	24.8	41.7	25.4	7.8	21.6	34.2	44.8	50.4	18.6	75M
Point Rend [‡] [20]	ResNet-101	25.9	43.4	27.6	8.2	20.2	38.6	43.3	54.1	20.6	75M
R ³ -CNN [‡] [28]	ResNet-101	24.9	40.5	27.8	9.7	21.4	33.6	45.4	52.2	20.2	77M
SOLOv2 [29]	ResNet-101	24.5	40.9	25.1	5.6	19.4	37.6	36.4	48.3	20.6	65M
QueryInst [‡] [10]	ResNet-101	26.0	42.8	27.3	8.2	21.7	35.1	43.3	54.1	20.6	191M
Mask Transfiner [‡] [19]	ResNet-101	24.6	42.1	26.0	7.2	19.4	36.1	43.8	26.3	19.8	63M
Mask2Former [‡] [6]	ResNet-101	25.7	38.0	27.7	6.3	18.9	38.1	41.1	51.9	23.1	63M
WaterMask R-CNN	ResNet-101	25.6	41.7	27.9	8.8	21.3	36.0	45.3	53.9	19.0	67M
WaterMask R-CNN[‡]	ResNet-101	27.2	43.7	29.3	9.0	21.8	38.9	46.3	54.8	20.9	67M
Cascade WaterMask R-CNN[‡]	ResNet-101	27.1	42.9	30.4	8.3	21.0	38.9	47.0	55.8	22.5	107M

Table 3: Comparison with the State-of-the-art Methods on UIIS. Models with [‡] were trained with 3× schedule using multi-scale training. The data marked in red are the best, and those in blue are the second best.

Experiments

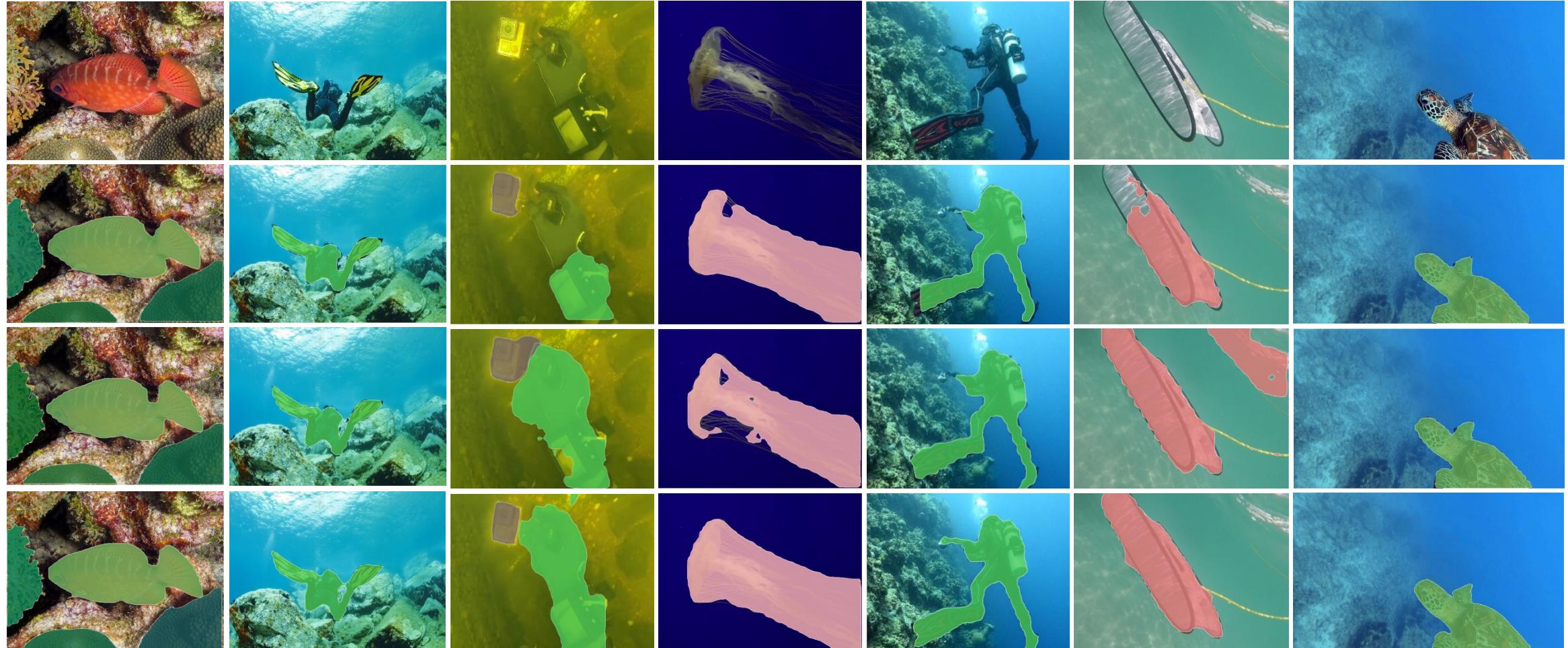


Figure 1: Qualitative comparison on the UIIS dataset. The first row represents the original image, and the second, third and fourth rows represent the results of Mask R-CNN, QueryInst and ours, respectively.

Ablation Study

Methods	mAP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L
w/o DSGAT	24.2	40.2	25.7	8.2	20.9	33.3
w/o MFRM	23.1	38.4	24.6	8.4	20.1	31.8
w/o BMS	22.5	41.2	23.1	8.4	19.0	31.1
w/o BLL	23.9	40.7	25.4	8.7	20.9	32.9
WaterMask	25.6	41.7	27.9	8.8	21.3	36.0

Table 4: Effectiveness of each component in WaterMask.
ResNet-101-FPN and 1× training schedule is adopted.

k	mAP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L
5	23.1	39.2	23.8	8.6	19.9	31.8
7	24.0	40.3	25.2	8.0	21.1	31.8
9	24.9	42.2	26.6	8.3	21.2	34.9
11	25.6	41.7	27.9	8.8	21.3	36.0
13	25.5	41.4	27.3	8.1	20.9	36.3

Table 5: Different value of k. k is the number of farthest nodes to be connected.

Patch	mAP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L
8×8	-	-	-	-	-	-
12×12	25.6	41.7	27.9	8.8	21.3	36.0
16×16	24.2	40.6	25.8	8.4	21.6	32.0
20×20	23.5	38.1	25.2	8.7	20.1	32.5

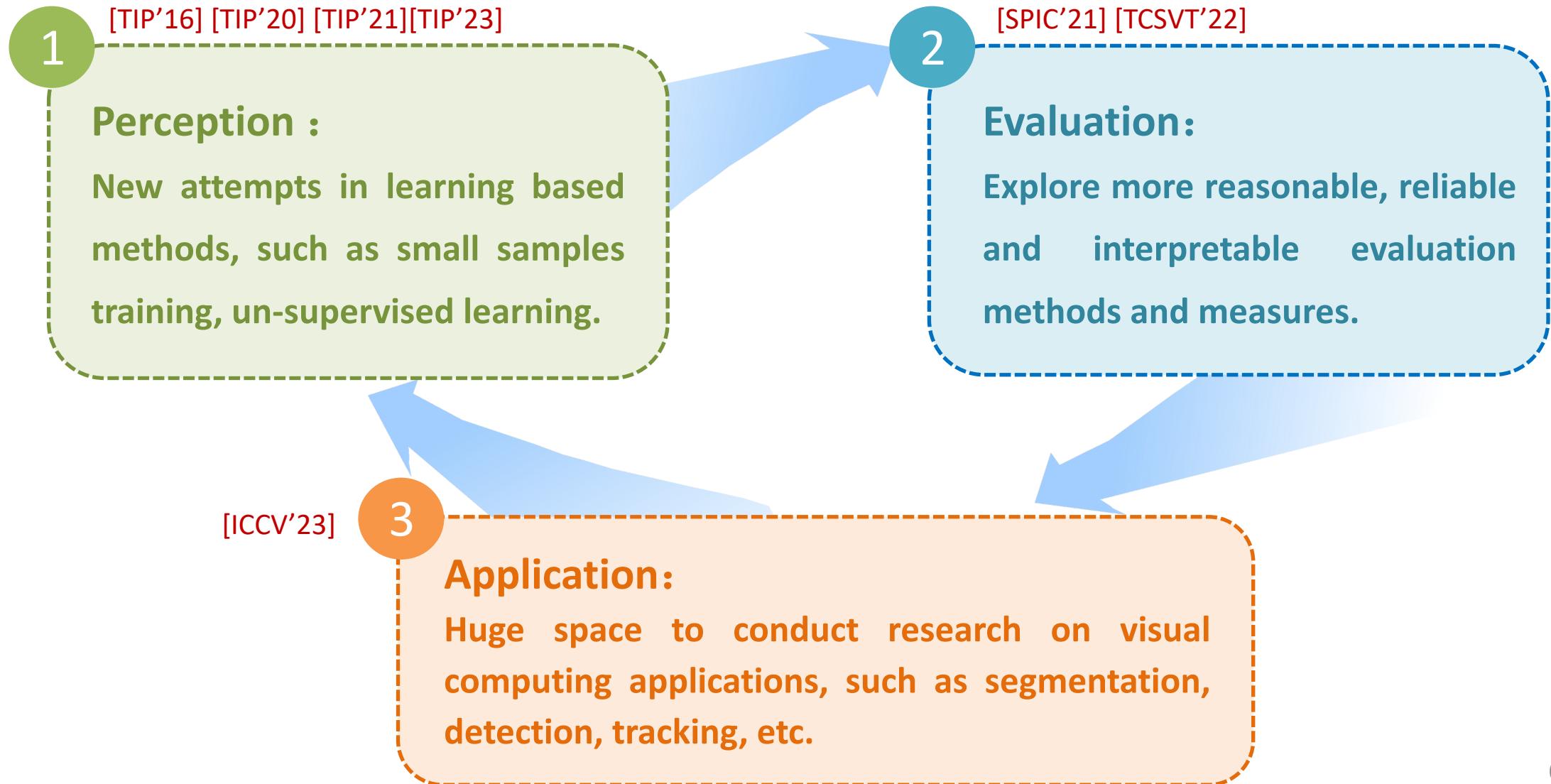
Table 6: Different Size of Patch. Each graph node corresponds to a 4s × 4s patch, where s is downsampling stride. When s = 2, the memory required by the model has exceeded the upper limit of the device.



Conclusion and Future Work

- In this paper, we have constructed the **first general underwater image instance segmentation dataset** with pixel-level annotations, which enables us to comprehensively explore the underwater instance segmentation task.
- According to the intrinsic characteristics of underwater imagery, we have proposed WaterMask for underwater instance segmentation. Extensive experiments have demonstrated the effectiveness of the proposed UIIS dataset and WaterMask.
- In future work, we plan to extend the UIIS datasets to **broader and more challenging underwater images and underwater videos**.

Future work





THANKS FOR WATCHING



为天下储人才



为国家图富强



机器智能与系统控制教育部重点实验室

Key Laboratory of Machine Intelligence & System Control, Ministry of Education

视觉感知与智能系统实验室成立于2013年9月，依托控制科学与技术国家A类学科，致力于智能系统感知、决策与应用领域的研究，团队包括国家级特聘教授1人、国家四青人才3人、泰山学者6人、中国科协青年托1人。目前承担国家、省部级各类科研经费3000余万元，获得国内外学术奖励10余次。



张伟
长江学者特聘教授
控制学院副院长



宋然 教授/博导
青年拔尖人才



元辉 教授/博导
国家优青



丛润民 教授/博导
中国科协青年托



贾潇 教授/博导
青年泰山



张敬林 教授/博导
青年泰山



李腾 教授/博导
山东省优青



李帅 教授/博导
齐鲁青年学者



李晓磊 副教授/硕导
加州伯克利分校 博后



李振华 副教授/硕导
卡尔加里大学 博后



鲁威志 副教授/硕导
法国国立科学院 博士



程吉禹 副教授/硕导
香港中文大学 博士



Pourya 副教授/硕导
上海交通大学 博士

