

# Empowering Semantic-Sensitive Underwater Image Enhancement with VLM

Guodong Fan<sup>1</sup>, Shengning Zhou<sup>1</sup>, Genji Yuan<sup>1</sup>, Huiyu Li<sup>2</sup>, Jingchun Zhou<sup>3</sup>, Jinjiang Li<sup>1</sup>\*

<sup>1</sup>Shandong Technology and Business University, Yantai, China

<sup>2</sup>Shandong University of Finance and Economics, Jinan, China

<sup>3</sup>Dalian Maritime University, Dalian, China

fgd96@outlook.com, 2023420043@sdtbu.edu.cn, 202414175@sdtbu.edu.cn,  
huiyuoy@163.com, zhoujingchun03@qq.com, lijjiang@gmail.com

## Abstract

In recent years, learning-based underwater image enhancement (UIE) techniques have rapidly evolved. However, distribution shifts between high-quality enhanced outputs and natural images can hinder semantic cue extraction for downstream vision tasks, thereby limiting the adaptability of existing enhancement models. To address this challenge, this work proposes a new learning mechanism that leverages Vision-Language Models (VLMs) to empower UIE models with semantic-sensitive capabilities. To be concrete, our strategy first generates textual descriptions of key objects from a degraded image via VLMs. Subsequently, a text-image alignment model remaps these relevant descriptions back onto the image to produce a spatial semantic guidance map. This map then steers the UIE network through a dual-guidance mechanism, which combines cross-attention and an explicit alignment loss. This forces the network to focus its restorative power on semantic-sensitive regions during image reconstruction, rather than pursuing a globally uniform improvement, thereby ensuring the faithful restoration of key object features. Experiments confirm that when our strategy is applied to different UIE baselines, significantly boosts their performance on perceptual quality metrics as well as enhances their performance on detection and segmentation tasks, validating its effectiveness and adaptability.

## Introduction

Underwater image processing is a critical technology in fields such as ocean exploration, biological monitoring, and underwater robotics (González-Sabbagh and Robles-Kelly 2023; Wang et al. 2019). Due to severe degradation from light absorption and scattering, UIE has emerged as an essential pre-processing step. While recent deep learning-based methods have achieved remarkable success in producing visually pleasing results for human observers (Jiang et al. 2023), a significant challenge has surfaced: a disconnect between perceptual image quality and the performance of machine cognition tasks.

We observe that state-of-the-art (SOTA) enhancement does not consistently translate to improved performance on tasks like object detection or segmentation (Tolie, Ren, and Elyan 2024; Zhu et al. 2019), as illustrated in Fig. 1(b). This

is due to the "task-agnostic" or "semantic-blind" nature of current UIE methods. In their pursuit of global, uniform enhancement, they often introduce imperceptible artifacts or cause a distribution shift that is misaligned with the data expectations of downstream models (Liu et al. 2019b; Fu et al. 2022). This reveals their fundamental limitation: an inability to distinguish between semantic focal points (e.g., marine creatures, artifacts) and non-focal points (e.g., background water), leading to the corruption of semantic cues vital for machine understanding.

To solve this problem, researchers have begun to explore semantic-guided enhancement strategies. Early works leveraged semantic segmentation maps to guide feature learning or pixel fusion (Wu et al. 2023). However, these methods depend heavily on high-quality, pixel-level annotations, which are notoriously scarce in underwater domains, risking the introduction of erroneous priors.

More recently, with the rise of VLMs, some have employed global, style-driven text prompts to guide image restoration (Pei, Huang, and Zhang 2020). For instance, using a prompt like "a clear underwater photo" can improve overall image quality. While innovative, this approach is still essentially a "one-size-fits-all" strategy. It does not pay attention to the specific semantic content within the image, and thus cannot achieve the fine-grained, object-centric processing required for a truly robust enhancement. A clear gap remains in achieving targeted, content-aware enhancement.

To this end, we propose a powerful semantic-sensitive UIE strategy, the core of which is to use the open-world understanding ability of VLMs to empower the UIE model to perceive and focus on key semantic regions. Specifically, our mechanism first utilizes a VLM to generate textual descriptions of core objects from the input degraded image. Then, through a text-image alignment model, we accurately map these high-level semantic descriptions back to the 2D space of the image, thus generating a spatial semantic guidance map. Finally, this map is efficiently injected into the decoder of the UIE network through a dual-guidance mechanism, combining cross-attention and an explicit alignment loss. This guides the network to allocate more resources to semantically vital regions during image reconstruction. This design makes the enhancement process no longer a blind pursuit of global visual improvement, but an intelligent, content-aware restoration process that ensures the structure

\*Corresponding author.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

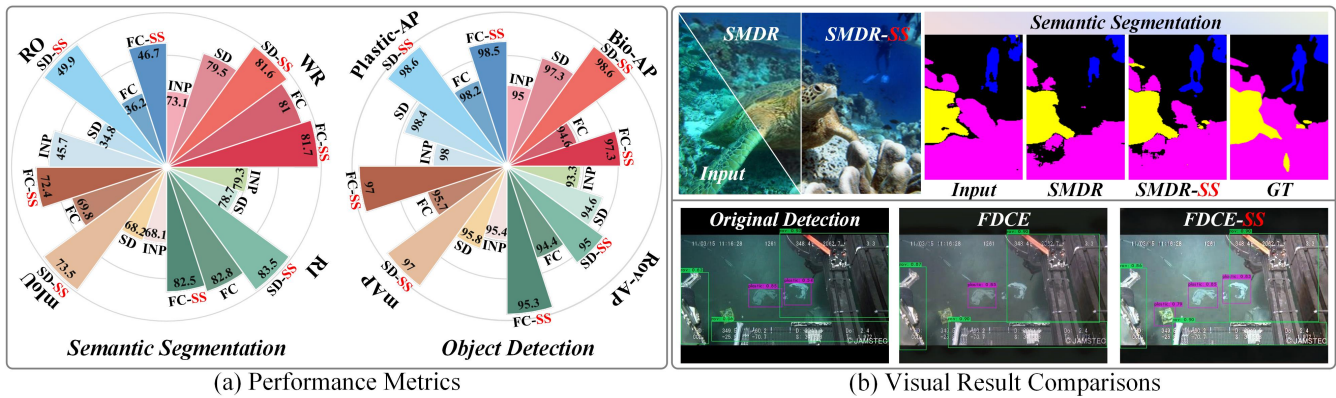


Figure 1: Impact of our semantic-sensitive strategy on downstream tasks. The radar charts on the left demonstrate consistent quantitative improvements in semantic segmentation and object detection when baseline models are empowered by our -SS strategy (in Red). Correspondingly, the qualitative examples on the right illustrate these benefits: our method leads to segmentation results closer to the Ground Truth (top) and enables more confident and accurate object detection (bottom).

and texture of key objects are faithfully restored and enhanced. The main contributions of this paper can be summarized as follows:

- We propose a VLM-driven, semantic-sensitive learning strategy to address the semantic blindness of traditional UIE, aiming to produce results that are robust for both human and machine perception.
- We design a dual-guidance mechanism to operationalize this strategy. The semantic map is used to simultaneously: structurally guide the network’s information flow via a cross-attention injection module, and explicitly regularize its intermediate features via a new semantic alignment loss.
- Experimental validation demonstrates that our strategy not only improves perceptual quality but also boosts performance on subsequent machine cognition tasks, proving the effectiveness and adaptability of our strategy.

## Related Work

In this section, we review prior image enhancement in three main areas: UIE techniques, downstream-aware image enhancement, and existing semantic-guided enhancement.

### Underwater Image Enhancement Techniques

UIE aims to correct image degradation caused by absorption and scattering of water bodies. Early methods are mainly based on physical models. These methods recover clear images by inversely modeling the physical process of underwater imaging (Soni and Kumare 2020; Wang et al. 2019). However, due to the diversity and complexity of the underwater environment, the physical prior assumptions of these methods are prone to lead to color distortion or noise amplification. In recent years, learning-based convolutional neural network (CNN) methods have become the mainstream in the field of UIE. These methods demonstrate powerful degradation feature extraction and image reconstruction capabilities

by performing end-to-end learning on pairwise datasets (Zamir et al. 2020). Learning methods have achieved remarkable success on traditional image quality metrics (Yang et al. 2023; Ju et al. 2025).

### Image Enhancement for Downstream Tasks

With the evolution of UIE technology, the academic community has increasingly recognized the discrepancy between enhancement goals centered on human vision and the requirements of downstream machine vision tasks (Liu et al. 2023a). Directly applying SOTA UIE methods as a pre-processing step for object detection or segmentation tasks yields inconsistent performance improvements, sometimes even causing a decline (Zhou et al. 2023; Chen et al. 2020). This “enhancement paradox” has spurred a new research direction: downstream-aware image enhancement. Research in this area is still in its nascent stages. Early explorations have primarily focused on joint training, where an enhancement network and a downstream task network are optimized simultaneously (Yu, Shi, and Han 2023; Kneubuehler et al. 2020). Although this approach enhances task relevance, it typically requires customizing the model for a specific downstream task, thus lacking flexibility and generality. Other works have attempted to design new loss functions aimed at preserving more features beneficial to downstream tasks during training (Liu et al. 2019a; Huang et al. 2023). These efforts collectively point in a clear direction: for enhancement to be genuinely beneficial to downstream tasks, the process must transcend the pixel level and incorporate higher-dimensional information.

### Semantic-guided Image Enhancement

Leveraging semantic information to guide image enhancement is becoming a popular strategy. Liao et al. (Liao et al. 2020) used semantic segmentation maps to iteratively refine structural priors. Yan et al. (Yan et al. 2019) utilized semantic segmentation to guide pixel fusion by assigning scaling factors to different regions. However, these methods rely

heavily on high-quality, pixel-level annotated data to train the segmentation models. Such data is particularly scarce in underwater scenarios, which can lead to the introduction of erroneous semantic priors that misguide the enhancement process.

With the rise of VLMs, researchers have begun to exploit their powerful open-world understanding capabilities to guide image restoration (Zhou et al. 2025; Huang, Yang, and Guibas 2024; Wang et al. 2023). In the UIE domain, Liu et al. (Liu et al. 2024) employed the CLIP model as a discriminator, using global text prompts to guide a generative model for image recovery. While this approach cleverly utilizes the cross-modal capabilities of VLMs, its guidance mechanism remains fundamentally global and style-driven. It does not concern itself with the specific objects present within the image and is therefore incapable of performing fine-grained, content-specific processing.

## Methodology

In this section, we introduce our strategy to empower UIE networks with semantic sensitivity, as shown in Fig. 2. We begin by defining the generation of a guidance map from VLM-derived text, creating an object-centric prior. We then elaborate on dual-guidance strategy: a cross-attention module provides architectural guidance, while a semantic alignment loss enforces feature-level supervision.

### Generation of the Semantic Guidance Map

To overcome the "task-agnostic" nature of UIE models, the network must first be made aware of what to focus on. We leverage VLMs to generate a reliable semantic prior that is specific to the content of each image, avoiding the need for scarce, densely annotated data required by traditional semantic segmentation. We first generate a semantic guidance map, denoted as  $M_{sem} \in \mathbb{R}^{H \times W}$ , for each degraded underwater image  $I_d \in \mathbb{R}^{H \times W \times 3}$ . This map is designed to precisely quantify the semantic relevance between each spatial location in the image and the key objects of interest.

**Cross-Modal Feature Alignment** Given the input image  $I_d$  and its corresponding textual description  $T$  generated by a VLM (adopting LLaVA (Liu et al. 2023b)), we leverage the visual encoder  $\Phi_v$  and text encoder  $\Phi_t$  of a pre-trained vision-language alignment model (adopting BLIP (Li et al. 2022)) to extract features.  $\Phi_v$  encodes  $I_d$  into a set of patch features  $F_v = \{f_v^1, f_v^2, \dots, f_v^N\} \in \mathbb{R}^{N \times C}$ , while  $\Phi_t$  encodes  $T$  into a global textual feature vector  $f_t \in \mathbb{R}^C$ . After their respective projection layers and normalization, we obtain unit feature vectors  $\hat{v}_i$  and  $\hat{t}$ .

**Semantic Similarity Calculation and Sharpening** We compute the initial semantic relevance scores  $s_i = \hat{v}_i^\top \hat{t}$  by taking the cosine similarity between each patch feature and the text feature. However, the raw similarity distribution can be too smooth to provide definitive guidance for the model (Zhao et al. 2024). To accentuate high-relevance regions and suppress irrelevant background, we design a semantic sharpening function,  $\Psi_{sharp}$ , which combines a power-law trans-

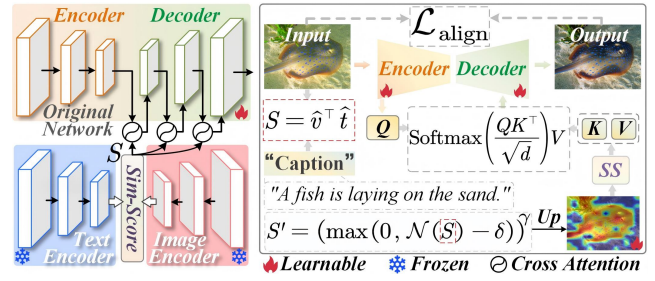


Figure 2: Overview of our semantic-sensitive learning strategy. A VLM-generated guidance map steers the UIE network via a dual-guidance approach: cross-attention and explicit supervision via an alignment loss.

formation with a thresholding operation:

$$s'_i = \Psi_{sharp}(s_i; \gamma, \delta) = (\max(0, \mathcal{N}(s_i) - \delta))^\gamma, \quad (1)$$

where  $\mathcal{N}(\cdot)$  denotes min-max normalization to the range  $[0, 1]$ ,  $\delta$  is a threshold to filter out low-relevance noise responses, and  $\gamma > 1$  is the power-law coefficient used to non-linearly expand the score gap to focus attention. Finally, we reshape the 1D score sequence  $S' = \{s'_1, \dots, s'_N\}$  and upsample it to the original image dimensions via interpolation, yielding the final single-channel semantic guidance map  $M_{sem}$ .

### Semantically-Guided Image Reconstruction

With the guidance map generated, the next step is to integrate it into the UIE network. We propose a synergistic approach that combines structural guidance and explicit supervision to inject the semantic map into the decoder. Injecting semantic guidance here directly influences the decoding and reconstruction process, ensuring that as details are being rebuilt, priority is given to semantically important regions. This allows the semantic prior to be robustly and effectively utilized throughout the entire reconstruction process.

**Cross Attention Injection Mechanism** We inject semantic information into the UIE network's decoder via cross attention, allowing it to take effect at various decoding stages  $l$ . At stage  $l$ , the decoder feature  $d_l$  acts as the query  $Q_l$  by the linear projection to attend to the skip-connection feature  $e_l$  from the encoder, which has been modulated by  $M_{sem}$ . Specifically,  $M_{sem}$  is downsampled to the same spatial resolution as  $e_l$ , yielding  $\tilde{M}^{(l)}$ . This map is then used to element-wise weight  $e_l$ , which is then linearly projected to generate the key  $K_l$  and value  $V_l$ . Based on this, the decoder can preferentially extract information from the semantically "illuminated" encoder features:

$$d'_l = \text{softmax} \left( \frac{Q_l K_l^\top}{\sqrt{d_k}} \right) V_l. \quad (2)$$

This design empowers the network with semantic-sensitivity and provides structural guidance.

**Explicit Semantic Alignment Loss** While cross-attention provides structural guidance, its effect is implicit. To provide

Methods	UIEB			U45		Challenge60	
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	UIQM $\uparrow$	UCIQE $\uparrow$	UIQM $\uparrow$	UCIQE $\uparrow$
PUIE	21.05	0.869	0.117	3.187	0.388	2.686	0.376
<b>PUIE-SS</b>	23.20(+2.15)	0.884(+0.015)	0.092(-0.025)	3.210(+0.023)	0.392(+0.004)	2.705(+0.019)	0.379(+0.004)
SMDR	22.44	0.899	0.106	3.108	0.422	2.727	0.415
<b>SMDR-SS</b>	23.28(+0.84)	0.909(+0.010)	0.099(-0.007)	3.139(+0.031)	0.436(+0.014)	2.852(+0.125)	0.433(+0.018)
UIR	22.89	0.885	0.124	2.874	0.374	2.458	0.371
<b>UIR-SS</b>	24.62(+1.73)	0.901(+0.016)	0.113(-0.011)	3.016(+0.142)	0.378(+0.004)	2.659(+0.201)	0.386(+0.015)
PFormer	23.53	0.877	0.113	3.105	0.437	2.426	0.426
<b>PFormer-SS</b>	24.97(+1.44)	<b>0.933(+0.056)</b>	<b>0.087(-0.026)</b>	3.126(+0.021)	<b>0.449(+0.012)</b>	2.456(+0.030)	<b>0.437(+0.011)</b>
FDCE	23.66	0.909	0.111	3.258	0.434	3.015	0.418
<b>FDCE-SS</b>	24.63(+0.97)	0.927(+0.018)	0.093(-0.018)	<b>3.267(+0.009)</b>	0.436(+0.002)	<b>3.079(+0.064)</b>	0.421(+0.003)

Table 1: Quantitative comparison of various UIE methods on the UIEB, U45, and Challenge60 datasets. -SS denotes empowering the baseline model with our proposed semantic-sensitive mechanism. Bold indicates the best performance among all methods, while Red values show the performance gain achieved by our mechanism.

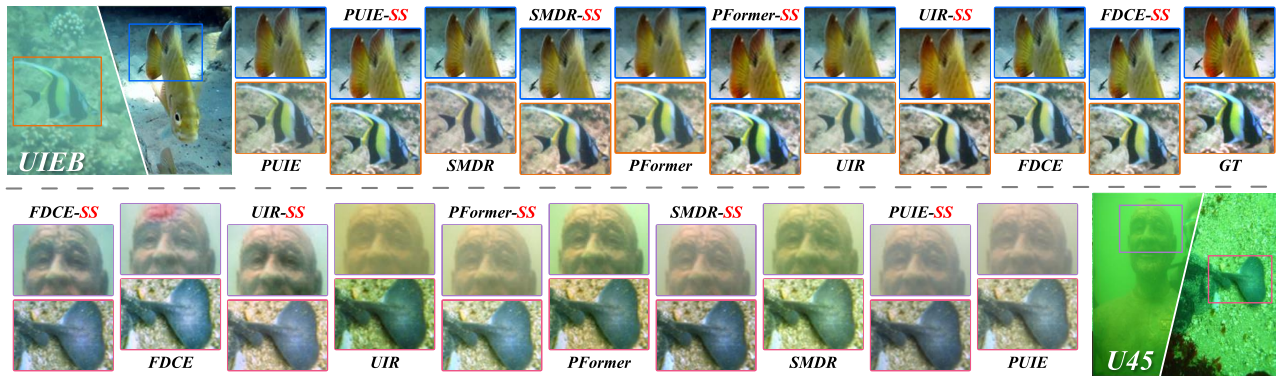


Figure 3: Visual comparison of enhancement results. Our -SS models produce images with better color fidelity and sharper details on key objects on the UIEB dataset (top). In challenging U45 scenes (bottom), our method effectively restores natural colors while avoiding artifacts introduced by baseline models.

a more direct and measurable supervisory signal, we introduce an explicit semantic alignment loss, denoted as  $\mathcal{L}_{\text{align}}$ . It ensures the network is not just structurally guided but is also explicitly penalized for producing feature representations that deviate from the semantic prior.  $\mathcal{L}_{\text{align}}$  operates directly on the intermediate feature maps of the decoder, forcing their spatial distribution to align with the semantic guidance map.

For the feature map  $\mathbf{F}^{(l)}$  produced at stage  $l$  of the decoder, our proposed  $\mathcal{L}_{\text{align}}^{(l)}$  is composed of two terms:

$$\mathcal{L}_{\text{align}}^{(l)}(\mathbf{F}^{(l)}, \tilde{M}^{(l)}) = \underbrace{\|\mathbf{F}^{(l)} \odot (1 - \tilde{M}^{(l)})\|_F^2}_{\text{Background}} - \underbrace{\eta \langle \mathbf{F}^{(l)}, \tilde{M}^{(l)} \rangle}_{\text{Foreground}}, \quad (3)$$

where  $\|\cdot\|_F^2$  is the squared Frobenius norm,  $\langle \cdot, \cdot \rangle$  denotes the inner product, and  $\eta$  is a hyperparameter balancing the two terms. For background suppression, this term penalizes the network for producing unnecessarily strong activations in

non-key regions (defined by the mask  $1 - \tilde{M}^{(l)}$ ) by minimizing the energy of the feature map  $\mathbf{F}^{(l)}$ , thereby suppressing background noise. For the foreground enhancement term, it rewards the network for producing strong responses in key object regions that are consistent with the semantic guidance by maximizing the correlation between  $\mathbf{F}^{(l)}$  and  $\tilde{M}^{(l)}$ .

### Overall Training Objective

To produce a final result that is both visually pleasing and semantically sound, the training objective must balance low-level pixel fidelity, high-level perceptual quality, and our proposed semantic alignment.

The final objective function,  $\mathcal{L}_{\text{total}}$ , is a weighted sum of a reconstruction loss,  $\mathcal{L}_{\text{recon}}$ , and our proposed semantic alignment loss,  $\mathcal{L}_{\text{align}}$ :

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{recon}} + \lambda_{\text{align}} \sum_{l \in L} \mathcal{L}_{\text{align}}^{(l)}, \quad (4)$$

Methods	Object detection				Semantic Segmentation					
	Plastic-AP $\uparrow$	Bio-AP $\uparrow$	Rov-AP $\uparrow$	mAP $\uparrow$	RO $\uparrow$	RI $\uparrow$	FV $\uparrow$	WR $\uparrow$	HD $\uparrow$	mIoU $\uparrow$
<i>Original</i>	98.00	94.95	93.33	95.43	45.68	79.31	66.44	73.11	75.97	68.10
PUIE	97.90	95.25	93.07	95.40	33.31	80.04	68.61	78.55	70.49	66.20
<b>PUIE-SS</b>	97.84(-0.06)	97.13(+1.88)	93.86(+0.79)	96.28(+0.88)	38.84(+5.53)	82.36(+2.32)	75.33(+6.72)	81.31(+2.76)	76.15(+5.66)	70.80(+4.60)
UIR	96.49	92.43	94.22	94.37	34.54	82.97	74.78	79.19	71.08	68.52
<b>UIR-SS</b>	97.22(+0.73)	93.96(+1.53)	94.77(+0.55)	95.31(+0.94)	42.03(+7.49)	<b>84.78(+1.81)</b>	74.91(+0.12)	79.06(-0.13)	71.49(+0.41)	70.45(+1.93)
PFormer	98.20	96.29	94.97	95.50	36.23	83.83	75.58	80.90	70.12	69.34
<b>PFormer-SS</b>	98.48(+0.28)	96.99(+0.70)	95.15(+0.18)	96.87(+1.37)	<b>51.52(+15.29)</b>	83.34(-0.49)	<b>77.64(+2.06)</b>	<b>83.43(+2.53)</b>	<b>77.80(+7.67)</b>	<b>74.75(+5.41)</b>
SMDR	98.39	97.25	94.61	95.76	34.78	78.66	73.08	79.47	74.92	68.18
<b>SMDR-SS</b>	<b>98.59(+0.20)</b>	<b>98.59(+1.34)</b>	95.01(+0.40)	96.98(+1.22)	49.94(+15.17)	83.48(+4.82)	76.99(+3.91)	81.55(+2.09)	75.58(+0.67)	73.51(+5.33)
FDCE	98.17	94.55	94.44	95.72	36.24	82.81	76.11	80.96	72.80	69.78
<b>FDCE-SS</b>	98.50(+0.33)	97.27(+2.72)	<b>95.27(+0.83)</b>	<b>97.01(+1.29)</b>	46.71(+10.47)	82.54(-0.27)	75.65(-0.46)	81.73(+0.78)	75.17(+2.37)	72.36(+2.58)

Table 2: Quantitative evaluation on downstream tasks, comparing UIE baselines with and without our -SS strategy. The tasks include object detection, and semantic segmentation. Red denotes improvement over the baseline, while Green indicates a decline.

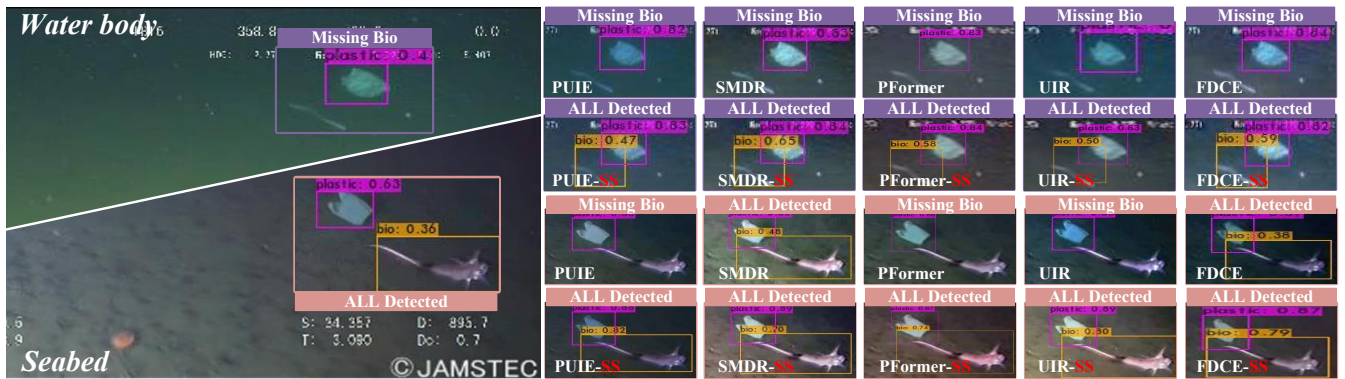


Figure 4: Visual comparison on the object detection task. Our -SS enhancement significantly improves the detection of small, low-contrast objects in both murky water body (top) and complex seabed (bottom) environments, effectively mitigating the missed detection issue prevalent in baseline methods.

where  $l$  represents the set of decoder stages where the semantic alignment loss is applied, and  $\lambda_{\text{align}}$  is a key hyperparameter that balances the two loss terms and we empirically set  $\lambda_{\text{align}}$  to 0.1.

The reconstruction loss,  $\mathcal{L}_{\text{recon}}$ , is itself a composite of a L1 loss and a perceptual loss,  $\mathcal{L}_{\text{percep}}$ . While the L1 loss ensures basic pixel fidelity, the perceptual loss guarantees structural and textural similarity by comparing the enhanced and reference images in the feature space of the pre-trained VGG-19 network (Simonyan and Zisserman 2014):

$$\mathcal{L}_{\text{recon}} = \|I_e - I_{gt}\|_1 + \lambda_{\text{percep}} \sum_j \|\phi_j(I_e) - \phi_j(I_{gt})\|_1, \quad (5)$$

where  $\phi_j$  denotes the feature extractor from the  $j$ -th convolutional layer of the VGG-19, and  $\lambda_{\text{percep}}$  is its weight. By optimizing this objective, the empowered network learns to restore accurate pixel values while preserving perceptual features and ensuring its internal representations align with the semantic prior. This holistic training strategy results in

an enhancement that is visually superior and more robust for both human and machine perception.

## Experiments

In this section, we conduct comprehensive experiments to validate the effectiveness of our strategy. We present our experimental setup (datasets, metrics, implementation), followed by extensive quantitative and qualitative comparisons on both perceptual quality and downstream vision tasks. Finally, ablation studies analyze the contributions of proposed strategy.

### Datasets and Evaluation Metrics

**Datasets for UIE Tasks.** To train and evaluate our models, we utilize a combination of standard UIE benchmarks. We use the UIEB dataset (Li et al. 2019) for supervised training, leveraging its training set of 790 paired images. For full-reference evaluation, we test on its official test set

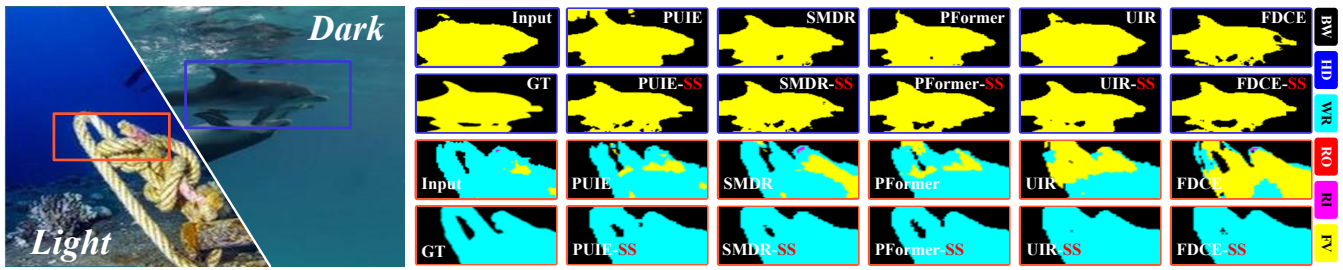


Figure 5: Visual comparison on the semantic segmentation task. Our semantic-sensitive enhancement preserves object boundaries and reduces background confusion in both dark (top row) and light (bottom row) scenes, leading to more accurate segmentation masks compared to baseline methods.

of 100 paired images. To assess the generalization capability and enhancement quality in real-world scenarios without ground-truth, we test the models trained on UIEB on two additional no-reference datasets: U45 (Li, Li, and Wang 2019) and Challenge60 (Li et al. 2019).

**Datasets for Downstream Vision Tasks.** To validate that our enhancement produces semantically robust images that are more conducive to machine understanding, we evaluate its impact on two distinct downstream tasks using their benchmark datasets: Trash-ICRA19 dataset (Fulton et al. 2019) for object detection, SUIM dataset (Islam et al. 2020) for semantic segmentation.

**Evaluation Metrics.** Our evaluation protocol is comprehensive, covering both perceptual quality and downstream task performance. For UIE tasks, we measure full-reference quality on the UIEB test set using the peak signal-to-noise ratio (PSNR), the structural similarity index measure (SSIM) and the learned perceptual image patch similarity (LPIPS) (Zhang et al. 2018). For no-reference evaluation on U45 and Challenge60, we employ the underwater image quality measure (UIQM) (Panetta, Gao, and Aghaian 2015) and the underwater color image quality evaluation (UCIQE) (Yang and Sowmya 2015). For downstream vision tasks, we report the average precision (AP) for three distinct categories (Plastic, Bio, Rov) and the overall mean average precision (mAP) for object detection; we report the intersection over union (IoU) for five categories: Robots (RO), Reefs and invertebrate (RI), Fish and vertebrates (FV), Wrecks or ruins (WR), Human divers (HD) and the overall mean intersection over union (mIoU) for semantic segmentation.

### Implementation Details

Our semantic-sensitive mechanism is designed to be a versatile and pluggable module that can be integrated into various UIE architectures. To demonstrate its broad applicability, we select five recent, SOTA UIE models that feature an encoder-decoder structure as baselines: PUIE (Fu et al. 2022), SMDR (Zhang et al. 2024), UIR (Guo et al. 2025), PFormer (Khan et al. 2025), and FDCE (Cheng et al. 2025).

**Training for UIE Task.** For our primary enhancement experiments, we first train these five baseline models on the UIEB training set to establish their original performance

and obtain their official weights. Subsequently, we integrate our semantic-sensitive mechanism into each of these models. We then retrain these empowered models, denoted with the suffix -SS (for Semantic-Sensitive), on the same UIEB dataset. All original and -SS models are then evaluated on the test sets of UIEB, U45, and Challenge60.

**Evaluation on Downstream Tasks.** For downstream task evaluation, we use pre-trained weights of both the original baseline models and our -SS variants as image pre-processors. For each task, the input images from the respective test set are first enhanced by five original and five -SS models. The resulting enhanced images are then fed into a unified downstream network for evaluation. This rigid setup ensures that any observed performance difference is strictly attributable to the intrinsic quality and semantic integrity of the enhancement provided by the low-level model.

### Quantitative Evaluation

**Performance on UIE Tasks.** Table 1 presents the quantitative results on the UIE datasets. For UIEB, our mechanism shows improvements in both PSNR and SSIM across all five baselines. This indicates that by focusing on semantically important regions, our method can achieve higher fidelity in image reconstruction. For U45 and Challenge60, our method also performs excellently on the UIQM and UCIQE metrics. Although there are fluctuations, the overall trend is positive. This demonstrates that the semantic-guided approach does not overfit to a specific data distribution and can maintain better color balance and clarity in challenging scenes.

**Performance on Downstream Tasks.** Table 2 summarizes the evaluation results on downstream vision tasks. The consistent outperformance of our empowered versions serves as strong evidence that our strategy produces images with more robust features for machine analysis. Notably, while some baselines struggle to outperform the Original unenhanced images, our method consistently delivers significant gains over both. The improvement is particularly pronounced in semantic segmentation, where our method provides cleaner object structures for pixel-level classification, and also consistent in object detection, making targets more identifiable.

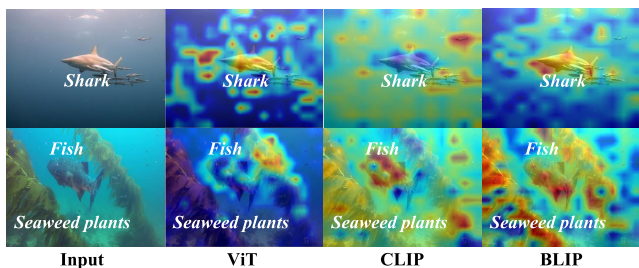


Figure 6: Visual comparison of guidance maps from different models.

## Qualitative Analysis

**Comparisons on UIE Tasks.** Fig. 3 presents a visual comparison of the enhancement results on the UIEB and U45 datasets. For UIEB, while all baseline methods improve the overall brightness and color, they exhibit suboptimal performance in preserving the fine details of the key object (the fish). PUIE and UIR tend to result in over-saturated colors, whereas SMDR and PFormer sometimes flatten the textures. For U45, the original images suffer from severe color cast and low visibility. The baseline methods either fail to sufficiently remove the greenish color cast or introduce unnatural artifacts. In contrast, our empowered models focus their restoration capabilities on semantically salient regions. Consequently, they exhibit sharper details, more natural color transitions, and a clearer separation from the background, consistently producing results that are more visually pleasing. confirms that our enhancement strategy makes it easier for downstream models to identify and locate key targets.

**Comparisons on Downstream Tasks.** Fig. 4 and Fig. 5 visually demonstrate the performance improvements on various downstream tasks. Fig. 4 presents a marine debris detection scenario featuring small, low-contrast objects against a complex seabed background. When using the original baseline enhancer, the detection model frequently fails. However, after applying our semantic-sensitive mechanism, the same object is successfully detected in all cases. This indicates that the empowered model provides targeted local contrast enhancement that is sufficient to identify previously missed objects. Fig. 5 illustrates the impact on semantic segmentation. In the dark scene, both PUIE-SS and SMDR-SS models yield superior segmentation results by effectively enhancing the contrast between the dolphin and the dark body of water. Conversely, in the bright scene, baseline methods tend to over-enhance the entire scene, which leads to mis-segmentation errors. In contrast, our empowered models, by selectively enhancing the target, maintain a clear distinction between the foreground object and the background.

## Ablation Study

**Analysis of Text-Image Alignment for Semantic Guidance.** To justify our choice of model for generating the semantic guidance map, a critical precursor in our strategy. As shown in Fig. 6, we compare guidance maps generated by three different vision-language alignment ap-

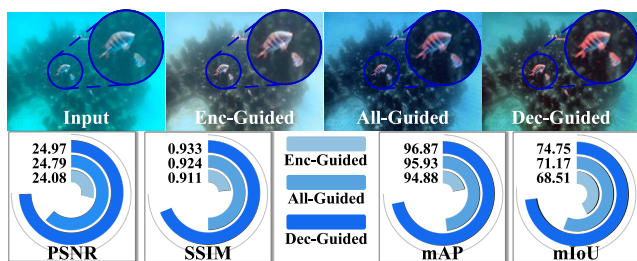


Figure 7: Visual comparison of the semantic guidance injection stage.

proaches: a standard Vision Transformer (ViT) using class attention, CLIP, and BLIP. The class-based attention of the ViT produces coarse and diffuse heatmaps, failing to precisely localize objects or disentangle concepts like "fish" and "seaweed." While contrastive learning endows CLIP with sharper attention, it also induces noise-prone spurious activations in background regions. In stark contrast, BLIP's fusion-based alignment strategy yields superior results. It generates clean, sharply-defined, and spatially accurate maps that correctly highlight the objects described in the text prompt with minimal noise. This high-fidelity spatial guidance is crucial for the effectiveness of our subsequent enhancement stages, validating our selection of BLIP.

**Analysis of Semantic Guidance Injection Stage.** To validate our strategy of injecting guidance into the decoder, we ablate three configurations: injecting into the Encoder only, All stages, or the Decoder only. As shown in Fig. 7, our decoder-only approach is demonstrably superior. Qualitatively, it produces the most visually pleasing result with vibrant colors and sharp object details. Quantitatively, it achieves the highest scores across all metrics, including perceptual quality and machine cognition. This confirms our hypothesis that injecting semantic guidance during the decoder's reconstruction process is the most effective strategy, as it directly steers image formation, whereas injecting it during the encoder's feature extraction stage is less effective and can even be counter-productive.

## Conclusion

This paper addresses the semantic blindness of traditional UIE methods, a core limitation that hinders their performance in both human perception and machine cognition. We propose a novel semantic-sensitive learning strategy that empowers UIE models with content-aware capabilities. Our strategy first leverages a VLM to generate a spatial semantic guidance map. It then steers the enhancement process through a dual-guidance mechanism, which synergistically combines a cross-attention module for structural guidance and an explicit semantic alignment loss for feature-level supervision. Extensive experiments demonstrate that our strategy is both effective and adaptable. When applied to various UIE baselines, it produces visually superior results and significantly boosts performance on subsequent detection and segmentation tasks.

## Acknowledgments

This work is supported by the National Natural Science Foundation of China under grant Nos. U24A20219, 62272281, 62576193, 62202268, and 62301105; special funds of Taishan Scholars Project of Shandong Province under grant Nos. tsqn202306274 and tsqn202507240; Natural Science Foundation of Shandong Province under grant Nos. ZR2025QC712, ZR2025QC695 and ZR2025MS985.

## References

- Chen, L.; Jiang, Z.; Tong, L.; Liu, Z.; Zhao, A.; Zhang, Q.; Dong, J.; and Zhou, H. 2020. Perceptual underwater image enhancement with deep learning and physical priors. *IEEE Transactions on Circuits and Systems for Video Technology*, 3078–3092.
- Cheng, Z.; Fan, G.; Zhou, J.; Gan, M.; and Chen, C. P. 2025. FDCE-Net: underwater image enhancement with embedding frequency and dual color encoder. *IEEE Transactions on Circuits and Systems for Video Technology*, 1728–1744.
- Fu, Z.; Wang, W.; Huang, Y.; Ding, X.; and Ma, K.-K. 2022. Uncertainty inspired underwater image enhancement. In *European Conference on Computer Vision*, 465–482. Springer.
- Fulton, M.; Hong, J.; Islam, M. J.; and Sattar, J. 2019. Robotic detection of marine litter using deep visual detection models. In *2019 International Conference on Robotics and Automation*, 5752–5758. IEEE.
- González-Sabbagh, S. P.; and Robles-Kelly, A. 2023. A survey on underwater computer vision. *ACM Computing Surveys*, 1–39.
- Guo, X.; Dong, Y.; Chen, X.; Chen, W.; Li, Z.; Zheng, F.; and Pun, C.-M. 2025. Underwater image restoration via polymorphic large kernel cnns. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, 1–5. IEEE.
- Huang, I.; Yang, G.; and Guibas, L. 2024. Blenderalchemy: Editing 3d graphics with vision-language models. In *European Conference on Computer Vision*, 297–314. Springer.
- Huang, Y.; Yuan, F.; Xiao, F.; Lu, J.; and Cheng, E. 2023. Underwater image enhancement based on zero-reference deep network. *IEEE Journal of Oceanic Engineering*, 903–924.
- Islam, M. J.; Edge, C.; Xiao, Y.; Luo, P.; Mehtaz, M.; Morse, C.; Enan, S. S.; and Sattar, J. 2020. Semantic segmentation of underwater imagery: Dataset and benchmark. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 1769–1776. IEEE.
- Jiang, Q.; Kang, Y.; Wang, Z.; Ren, W.; and Li, C. 2023. Perception-driven deep underwater image enhancement without paired supervision. *IEEE Transactions on Multimedia*, 4884–4897.
- Ju, Y.; Xiao, J.; Zhang, C.; Xie, H.; Luo, A.; Zhou, H.; Dong, J.; and Kot, A. C. 2025. Towards marine snow removal with fusing Fourier information. *Information Fusion*, 102810.
- Khan, M. R.; Negi, A.; Kulkarni, A.; Phutke, S. S.; Vipparthi, S. K.; and Murala, S. 2025. Phaseformer: Phase-based attention mechanism for underwater image restoration and beyond. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision*, 9618–9629. IEEE.
- Kneubuehler, D.; Gu, S.; Van Gool, L.; and Timofte, R. 2020. Flexible example-based image enhancement with task adaptive global feature self-guided network. In *European Conference on Computer Vision*, 343–358. Springer.
- Li, C.; Guo, C.; Ren, W.; Cong, R.; Hou, J.; Kwong, S.; and Tao, D. 2019. An underwater image enhancement benchmark dataset and beyond. *IEEE Transactions on Image Processing*, 4376–4389.
- Li, H.; Li, J.; and Wang, W. 2019. A fusion adversarial underwater image enhancement network with a public test dataset. *arXiv preprint arXiv:1906.06819*.
- Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 12888–12900. PMLR.
- Liao, L.; Xiao, J.; Wang, Z.; Lin, C.-W.; and Satoh, S. 2020. Guidance and evaluation: Semantic-aware image inpainting for mixed scenes. In *European Conference on Computer Vision*, 683–700. Springer.
- Liu, H.; Jin, F.; Zeng, H.; Pu, H.; and Fan, B. 2023a. Image enhancement guided object detection in visually degraded scenes. *IEEE Transactions on Neural Networks and Learning Systems*, 14164–14177.
- Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023b. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36: 34892–34916.
- Liu, P.; Wang, G.; Qi, H.; Zhang, C.; Zheng, H.; and Yu, Z. 2019a. Underwater image enhancement with a deep residual framework. *IEEE Access*, 94614–94629.
- Liu, R.; Hou, M.; Liu, J.; Fan, X.; and Luo, Z. 2019b. Compounded layer-prior unrolling: A unified transmission-based image enhancement framework. In *2019 IEEE International Conference on Multimedia and Expo*, 538–543. IEEE.
- Liu, S.; Li, K.; Ding, Y.; and Qi, Q. 2024. Underwater image enhancement by diffusion model with customized clip-classifier. *arXiv preprint arXiv:2405.16214*.
- Panetta, K.; Gao, C.; and Agaian, S. 2015. Human-visual-system-inspired underwater image quality measures. *IEEE Journal of Oceanic Engineering*, 541–551.
- Pei, Y.; Huang, Y.; and Zhang, X. 2020. Consistency guided network for degraded image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 2231–2246.
- Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Soni, O. K.; and Kumare, J. S. 2020. A survey on underwater images enhancement techniques. In *2020 IEEE 9th International Conference on Communication Systems and Network Technologies*, 333–338. IEEE.

Tolie, H. F.; Ren, J.; and Elyan, E. 2024. DICAM: Deep inception and channel-wise attention modules for underwater image enhancement. *Neurocomputing*, 127585.

Wang, C.; Pan, J.; Wang, W.; Dong, J.; Wang, M.; Ju, Y.; and Chen, J. 2023. Promptrestorer: A prompting image restoration method with degradation perception. *Advances in Neural Information Processing Systems*, 8898–8912.

Wang, Y.; Song, W.; Fortino, G.; Qi, L.-Z.; Zhang, W.; and Liotta, A. 2019. An experimental-based review of image enhancement and image restoration methods for underwater imaging. *IEEE Access*, 140233–140251.

Wu, Y.; Pan, C.; Wang, G.; Yang, Y.; Wei, J.; Li, C.; and Shen, H. T. 2023. Learning semantic-aware knowledge guidance for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1662–1671.

Yan, B.; Niu, X.; Bare, B.; and Tan, W. 2019. Semantic segmentation guided pixel fusion for image retargeting. *IEEE Transactions on Multimedia*, 676–687.

Yang, M.; and Sowmya, A. 2015. An underwater color image quality evaluation metric. *IEEE Transactions on Image Processing*, 6062–6071.

Yang, X.; Li, H.; Li, X.; and Li, T. 2023. HIFGAN: A high-frequency information-based generative adversarial network for image super-resolution. *ACM Transactions on Multimedia Computing, Communications and Applications*, 1–19.

Yu, N.; Shi, H.; and Han, Y. 2023. Joint correcting and refinement for balanced low-light image enhancement. *IEEE Transactions on Multimedia*, 6310–6324.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; Yang, M.-H.; and Shao, L. 2020. Learning enriched features for real image restoration and enhancement. In *European Conference on Computer Vision*, 492–511. Springer.

Zhang, D.; Zhou, J.; Guo, C.; Zhang, W.; and Li, C. 2024. Synergistic multiscale detail refinement via intrinsic supervision for underwater image enhancement. In *Proceedings of the AAAI conference on Artificial Intelligence*, 7033–7041.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhao, C.; Wang, K.; Zeng, X.; Zhao, R.; and Chan, A. B. 2024. Gradient-based visual explanation for transformer-based clip. In *International Conference on Machine Learning*, 61072–61091. PMLR.

Zhou, H.; Dong, W.; Liu, X.; Zhang, Y.; Zhai, G.; and Chen, J. 2025. Low-light image enhancement via generative perceptual priors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 10752–10760.

Zhou, J.; Pang, L.; Zhang, D.; and Zhang, W. 2023. Underwater image enhancement method via multi-interval subhistogram perspective equalization. *IEEE Journal of Oceanic Engineering*, 474–488.

Zhu, X.; Pang, J.; Yang, C.; Shi, J.; and Lin, D. 2019. Adapting object detectors via selective cross-domain alignment. In

*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 687–696.