LEARNING TO TRAIN UNDERWATER IMAGE ENHANCERS WITH REINFORCEMENT-GUIDED UPDATE SCHEDULING

Anonymous authors
Paper under double-blind review

ABSTRACT

Underwater images often suffer from color distortion and scattering-related detail degradation, which arise from different physical processes and exhibit distinct optimization dynamics. Existing methods that jointly optimize color correction and detail restoration typically rely on fixed loss weights or predefined update schedules, implicitly treating their relative importance as constant throughout training. However, the two objectives often progress asynchronously, making static coordination unable to accommodate their stage-dependent optimization demands. We propose a learning-to-train framework that formulates update scheduling as a Markov decision process. A PPO-based controller observes the current training state and dynamically selects among alternating, joint, and weighted updates for a cascaded enhancement network. We further construct a frequency-aware state representation from recent reconstruction quality trends and low/high frequency alignment to track the relative progress of global color recovery and local detail restoration. By maximizing cumulative rewards that account for reconstruction improvement, loss reduction, and frequency balance, the controller learns stage-adaptive update policies. Comprehensive qualitative and quantitative evaluations on multiple benchmark datasets, together with extensive ablation studies, demonstrate the effectiveness and robustness of the proposed method.

1 INTRODUCTION

High-quality underwater imagery provides essential visual support for a wide range of underwater applications, including object detection (Chen et al., 2024b; Zhou et al., 2025), robotic navigation (Liu et al., 2025a;b), and marine ecological monitoring (Lin et al., 2025; Yu et al., 2025). However, wavelength-dependent absorption and light scattering often cause color distortion, contrast attenuation, and detail degradation (Fan et al., 2026), thereby reducing visual quality and downstream perception reliability.

Absorption and scattering produce different degradations: the former mainly affects global color appearance, while the latter reduces visibility and local structural details (Zhang et al., 2024; Tang et al., 2024; Jiang et al., 2025; She et al., 2025). Accordingly, underwater enhancement involves coupled objectives of global color recovery and local detail restoration. Since these objectives often progress asynchronously during training, their relative optimization demands vary across stages, making a fixed optimization emphasis suboptimal.

Existing methods typically coordinate multiple restoration objectives using fixed loss weights, predefined update rules, or architectural adaptation. Recent approaches improve model-level flexibility through degradation-aware experts and conditional branches (Zamfir et al., 2025; Guo et al., 2024), or prompt-based conditioning (Lai et al., 2025; Potlapalli et al., 2023). However, their optimization procedures are generally fixed throughout training and thus cannot adapt to restoration objectives that progress at different rates. This motivates us to make the training schedule itself learnable.

We therefore revisit underwater image enhancement from a learning-to-train perspective and formulate update scheduling as a Markov Decision Process (MDP) (Ding et al., 2025; Meggendorfer et al., 2025). As illustrated in Figure 1, three representative training trajectories exhibit different interactions between color recovery and detail restoration. Detail restoration may converge ear-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

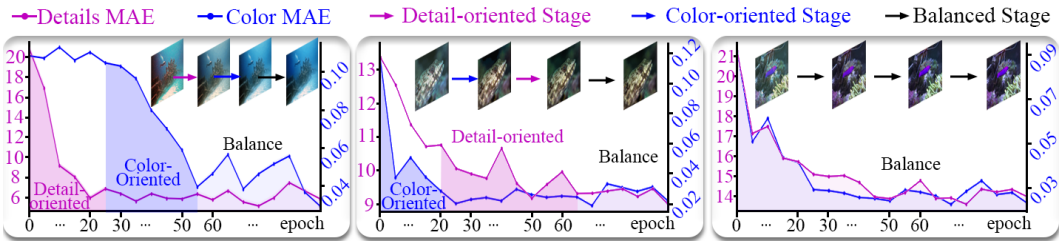


Figure 1: **Motivation for state-dependent update scheduling.** Color correction and detail restoration converge asynchronously, resulting in color-dominant, detail-dominant, and balanced optimization demands at different training stages. This motivates the controller to select among alternating, joint, and weighted updates according to the evolving training state.

lier than color recovery, color recovery may precede detail restoration, or the two objectives may evolve in a relatively balanced manner. These sample- and stage-dependent convergence patterns result in changing optimization demands. Since each update modifies the model parameters and thereby affects subsequent training states, update scheduling naturally becomes a sequential decision problem. The controller therefore selects $a_t \in \mathcal{A}$ according to the feedback state s_t, forming a state-adaptive sequence of alternating, joint, and weighted updates. These strategies adjust module participation and relative update strengths to accommodate the evolving optimization demands. To characterize training progress, we construct a frequency-aware state from recent reconstruction trends and low/high-frequency alignment, where the latter serves as a proxy for the relationship between global appearance recovery and local detail restoration. Guided by rewards based on reconstruction improvement, loss reduction, and frequency alignment, the controller learns to select suitable predefined update strategies across training stages. In summary, the main contributions of this work are as follows:

- We introduce a learning-to-train framework for underwater image enhancement that formulates update scheduling as a Markov decision process. Instead of relying on fixed loss weights or predefined schedules, a PPO-based controller dynamically selects among alternating, joint, and weighted updates according to the evolving training state.
- We develop a frequency-aware training state representation that combines recent reconstruction-quality trends with low/high frequency alignment. This representation provides observable feedback on the relative optimization progress of global color recovery and local detail restoration, enabling stage-adaptive scheduling decisions.
- Extensive experiments on four underwater image benchmarks demonstrate that the learned scheduler provides a better overall trade-off than fixed training strategies and achieves the highest PSNR among the compared methods on all three paired datasets.

2 Related Works

Underwater Image Enhancement. Underwater image enhancement jointly addresses color distortion and scattering-related visibility and detail degradation. Recent restoration methods improve adaptability mainly through conditional representations, high-level planning, and specialized architectures. PromptIR (Potlapalli et al., 2023) and InstructIR (Conde et al., 2024) condition restoration features using learned prompts or instructions, while RestoreAgent (Chen et al., 2024a) and SnowMaster (Lai et al., 2025) organize restoration operations through high-level feedback or planning. Expert-based methods, including Uni-MoE (Li et al., 2025), AdaptIR (Guo et al., 2024), and Complexity Experts (Zamfir et al., 2025), improve degradation specialization through expert routing or difficulty-aware computation. These approaches primarily provide model-level adaptability, whereas their optimization procedures generally remain based on fixed loss weights or predefined update rules.

Adaptive Optimization Strategies. Adaptive training methods modify optimization according to observed training behavior. Existing approaches dynamically adjust objective weights using loss,

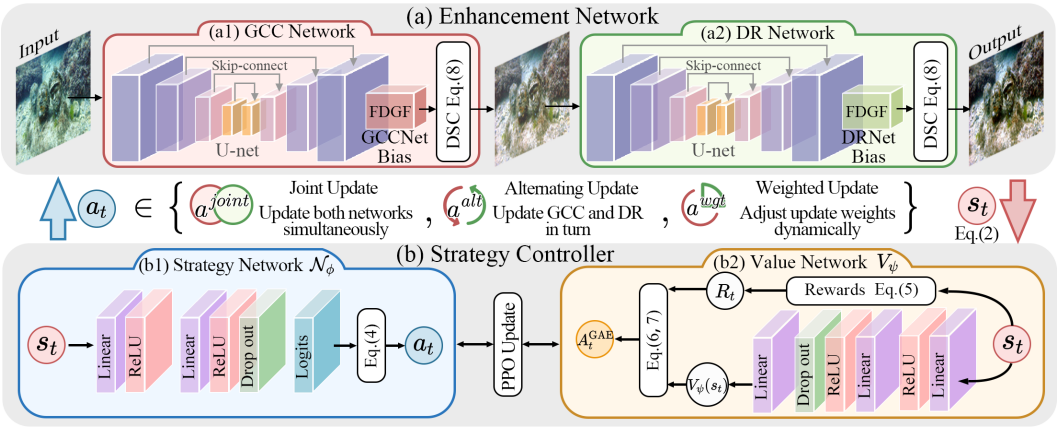


Figure 2: **Reinforcement-guided update scheduling.** (a) The cascaded GCC–DR network enhances degraded images. FDGF (Figure 3) performs frequency-aware fusion, while MBB promotes complementary frequency preferences between the two modules. (b) Given the state s_t, the strategy network $\mathcal{N}_\phi$ selects an update action a_t, and the value network V_ψ supports controller optimization.

gradient, or uncertainty information (Hayder et al., 2026), regulate sample order or difficulty through curriculum learning (Zhou et al., 2026), or learn optimizer and update decisions from training feedback (Ye et al., 2026). Their decision variables, however, are typically limited to scalar weights, sample selection, or optimizer configurations. In contrast, we study module-level update scheduling for underwater enhancement. A PPO-based controller observes reconstruction-quality trends and low/high-frequency alignment, and selects among alternating, joint, and weighted updates to coordinate global color recovery and local detail restoration. Thus, our method complements architectural adaptation by making the training schedule itself learnable.

3 Methodology

Our framework consists of a cascaded underwater image enhancer and a reinforcement-guided update scheduler. At each decision step, the scheduler observes recent optimization trends, selects an update action, and applies the corresponding update rule to the enhancement network. The resulting parameter update produces a new training state and an instantaneous reward, forming a closed-loop learning-to-train process. In this way, the coordination of global color recovery and local detail restoration is learned from training feedback rather than specified by a fixed schedule.

3.1 Markov Decision Process Modeling

Global color recovery and local detail restoration may progress at different rates during training. Therefore, the appropriate update rule should vary with the evolving optimization state rather than remain fixed throughout the entire training process. An update action changes the parameters of the enhancement network and consequently affects both the next observable state and the subsequent optimization trajectory. Update scheduling is thus a sequential decision-making problem in which each decision is conditioned on the current training state but may have delayed effects on the final enhancement performance. We formulate this process as a Markov Decision Process (MDP):

$$\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, P, R, \gamma \rangle, \tag{1}$$

where $\mathcal{S}$ denotes the training-state space, which summarizes the recent optimization progress and its evolving trend; $\mathcal{A}$ represents the set of available training strategies, characterizing the strategy decision executed at each step; P represents the state-transition dynamics induced by applying an update action to the enhancement network; R evaluates whether the selected action improves the current optimization trajectory; and $\gamma \in (0, 1]$ is the discount factor that balances immediate training improvement against long-term return. The transition dynamics are not modeled explicitly; instead, they are realized through the parameter updates of the enhancer.

3.2 REINFORCEMENT-GUIDED UPDATE SCHEDULER

Under the MDP formulation in Eq. (1), the update scheduler shown in Figure 2(b) observes the current training state s_t and selects an update action a_t at each decision step. The selected action determines how the Global Color Correction (GCC) and Detail Restoration (DR) modules are updated, and therefore affects both the immediate optimization result and the subsequent training trajectory.

Training-state representation. To provide the scheduler with observable feedback on the evolving optimization process, the state is constructed from recent reconstruction and frequency-alignment histories:

$$s_t = \left[m_{t-w+1:t},\, d_{t-w+1:t}\right]^\top \in \mathbb{R}^{2w}, \tag{2}$$

where $m_{t-w+1:t}$ and $d_{t-w+1:t}$ denote the recent histories of PSNR and frequency-alignment error over a sliding window of w steps, respectively. Retaining a short history allows the scheduler to capture not only the current quality level but also its recent evolution, making the state representation less sensitive to transient fluctuations. The PSNR history characterizes the overall reconstruction progress, whereas the frequency-alignment history serves as an observable proxy for the relative balance between global appearance recovery and local detail restoration.

To compute d_t, low-pass filtering and residual decomposition are applied to the output image $\mathbf{I}$. Specifically, $\mathbf{I}$ is decomposed into a low-frequency component $\mathbf{I}_{\text{low}} = \text{AvgPool}(\mathbf{I})$ and a complementary high-frequency residual $\mathbf{I}_{\text{high}} = \mathbf{I} - \mathbf{I}_{\text{low}}$. The former mainly captures global appearance variations, while the latter provides complementary information about local structures and details. This decomposition is used to construct an observable optimization-state indicator rather than to physically separate absorption and scattering. The frequency-alignment error is defined as

$$d_t = \left| \log \frac{\mathbb{E}[(\mathbf{I}_{\text{high}})^2] \cdot \mathbb{E}[(\mathbf{I}^{\text{gt}}_{\text{low}})^2]}{\mathbb{E}[(\mathbf{I}_{\text{low}})^2] \cdot \mathbb{E}[(\mathbf{I}^{\text{gt}}_{\text{high}})^2]} \right|, \tag{3}$$

where $\mathbb{E}[\cdot]$ denotes the spatial expectation. This metric measures the deviation between the output and ground truth in their relative high-to-low-frequency energy ratios. A smaller d_t indicates that the output exhibits a frequency balance closer to that of the ground truth, whereas a larger value suggests an uneven recovery of global appearance and local structural details.

Scheduling policy and action space. We define the discrete action space here formally as $\mathcal{A} = \{a^{\text{alt}}, a^{\text{joint}}, a^{\text{wgt}}\}$, corresponding to alternating, joint, and weighted updates, respectively. Given the state representation above, a policy network $\mathcal{N}_\phi$ parameterized by ϕ (Figure 2(b1)) maps s_t to a probability distribution over $\mathcal{A}$:

$$P(a_t \mid s_t; \phi) = \frac{\exp\big(\mathcal{N}_\phi(s_t)_{a_t}\big)}{\sum_{a \in \mathcal{A}} \exp\big(\mathcal{N}_\phi(s_t)_a\big)}, \tag{4}$$

where $\mathcal{N}_\phi(s_t)_a$ denotes the logit associated with candidate action $a \in \mathcal{A}$. The action space contains three complementary update modes: alternating update, joint update, and weighted update. Alternating update applies module-wise updates to the GCC and DR modules in an alternating manner. Joint update optimizes both modules within the same step using the default objective configuration. Weighted update also optimizes both modules jointly but uses an asymmetric weighting configuration to modify the relative emphasis of the color- and detail-oriented objectives. These actions allow the scheduler to regulate both the update participation of the two modules and their relative optimization emphasis.

Reward design. Once an action a_t is sampled, it is translated into a concrete update rule for the GCC and DR modules. Applying the selected rule for one optimization step produces a new state and an instantaneous reward:

$$R(s_t, a_t) = \lambda_m \frac{\Delta m_t}{\sigma_m} - \lambda_{\mathcal{L}} \frac{\Delta \mathcal{L}_t}{\sigma_{\mathcal{L}}} - \lambda_d \frac{\Delta d_t}{\sigma_d}, \tag{5}$$

where λ_m, $\lambda_{\mathcal{L}}$, and λ_d control the contributions of reconstruction improvement, loss reduction, and frequency balance, respectively. The operator $\Delta(\cdot)$ denotes the signed change between two consecutive decision steps, and $\sigma_{(\cdot)}$ denotes the standard deviation of the corresponding changes

within a sliding window. The standard-deviation normalization reduces scale differences among the three feedback signals. Here, m_t, $\mathcal{L}_t$, and d_t denote PSNR, the training loss of the enhancement network, and the frequency-alignment error, respectively. Under the convention $\Delta x_t = x_t - x_{t-1}$, an increase in PSNR or a decrease in training loss and frequency-alignment error contributes positively to the reward. The resulting reward encourages update actions that improve overall reconstruction quality while avoiding a persistent imbalance between global appearance recovery and local detail restoration.

Long-term policy optimization. Because the effect of an update action may emerge only after several subsequent network updates, a value network V_ψ (Figure 2(b2)) is introduced to estimate the expected discounted return from the current state:

$$V_\psi(s_t) = \mathbb{E}\left[\lim_{N\to\infty}\sum_{l=0}^{N}\gamma^l R_{t+l}\right], \tag{6}$$

where R_{t+l} denotes the instantaneous reward at step $t+l$, and $\gamma \in (0,1)$ balances immediate and future returns. Based on this value estimate, the generalized advantage estimate $\hat{\mathrm{A}}_t$ is computed using the temporal-difference (TD) error:

$$\begin{aligned}\hat{\mathrm{A}}_t &= \sum_{l=0}^{T-t-1}(\gamma\lambda)^l\delta_{t+l},\\ \delta_t &= R(s_t,a_t) + \gamma V_\psi(s_{t+1}) - V_\psi(s_t),\end{aligned} \tag{7}$$

where δ_t denotes the one-step TD error, and $\lambda = 0.95$ is the generalized-advantage-estimation smoothing factor that trades off bias and variance. A positive $\hat{\mathrm{A}}_t$ indicates that the sampled action yields a higher return than expected under the current policy, encouraging a larger selection probability in similar states. A negative value has the opposite effect. Finally, the policy network is optimized using proximal policy optimization (PPO) (Schulman et al., 2017), whose clipped policy updates improve training stability while allowing the scheduler to progressively favor update actions that provide higher long-term returns.

3.3 Cascaded Enhancement Network

To support update scheduling over global color recovery and local detail restoration, we construct a cascaded network consisting of Global Color Correction (GCC) and Detail Restoration (DR) modules, as shown in Figure 2(a). The two modules provide functionally differentiated parameter groups whose updates can be coordinated by the scheduler. Image enhancement is performed using the DSC curve (Xue et al., 2024):

$$y = [1 + \exp(-(k \odot x + b))]^{-1}, \tag{8}$$

in which x and y denote the input and output, respectively, and b and k are the predicted bias and scaling parameters.

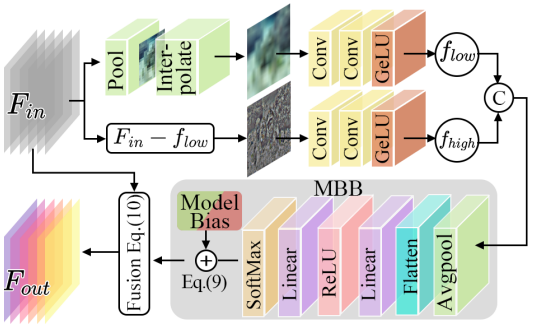


Figure 3: **FDGF module.** Low-frequency features and high-frequency residuals are adaptively fused; MBB applies complementary biases favoring low-frequency responses in GCC and high-frequency responses in DR.

Frequency-aware feature fusion. The Frequency Decomposition Gated Fusion (FDGF) module in Figure 3 decomposes the input feature F_{in} into a low-frequency component f_{low} and a high-frequency residual $f_{high} = F_{in} - f_{low}$. The former mainly captures global appearance variations, while the latter preserves local structures and details. This decomposition provides complementary frequency cues rather than a strict physical separation of absorption and scattering.

The two components are concatenated and fed into the Model Bias Block (MBB), which predicts content-dependent gating coefficients:

$$[\alpha_{low}, \alpha_{high}]^\top = \mathrm{MBB}([\mathbf{f}_{low}; \mathbf{f}_{high}]) + \boldsymbol{\beta}, \tag{9}$$

Table 1: Quantitative comparisons on the LSUI, EUVP, UIEB, and U45 datasets. Full-reference metrics include PSNR, SSIM, PCQI, and LPIPS, while no-reference metrics include UIQM, UCIQE, and NIQE. ↑ indicates that higher values are better, and ↓ indicates that lower values are better. **Red bold** denotes the best result, and blue denotes the second-best result.

Methods	LSUI				EUVP				UIEB				U45		
	PSNR↑	SSIM↑	LPIPS↓	PCQI↑	PSNR↑	SSIM↑	LPIPS↓	PCQI↑	PSNR↑	SSIM↑	LPIPS↓	PCQI↑	UIQM↑	UCIQE↑	NIQE↓
U-color	21.25	0.839	0.232	0.727	20.81	0.837	0.232	0.772	23.24	0.903	**0.100**	0.816	4.371	0.580	4.713
U-Shape	23.81	0.840	0.180	0.704	24.18	0.851	0.178	0.752	21.79	0.789	0.175	0.636	4.262	0.567	4.628
UWCNN	17.05	0.736	0.364	0.542	20.53	0.822	0.272	0.700	14.10	0.690	0.325	0.504	4.303	0.514	4.602
TACL	20.17	0.789	0.230	0.752	19.14	0.775	0.220	0.784	23.22	0.848	0.138	0.803	4.390	0.625	**4.040**
HLRP	12.49	0.179	0.418	0.233	11.43	0.182	0.408	0.284	13.00	0.225	0.377	0.219	4.106	**0.643**	4.897
ICSP	11.36	0.520	0.502	0.768	11.76	0.539	0.391	0.800	11.22	0.531	0.526	0.763	2.446	0.529	9.653
HUPE	20.79	0.830	0.275	0.791	19.37	0.804	0.289	0.827	23.38	**0.916**	0.131	0.871	4.237	0.611	4.487
Ours	**26.49**	**0.900**	**0.127**	**0.804**	**29.54**	**0.923**	**0.068**	**0.852**	**23.64**	0.910	0.118	**0.922**	**4.501**	0.612	4.295

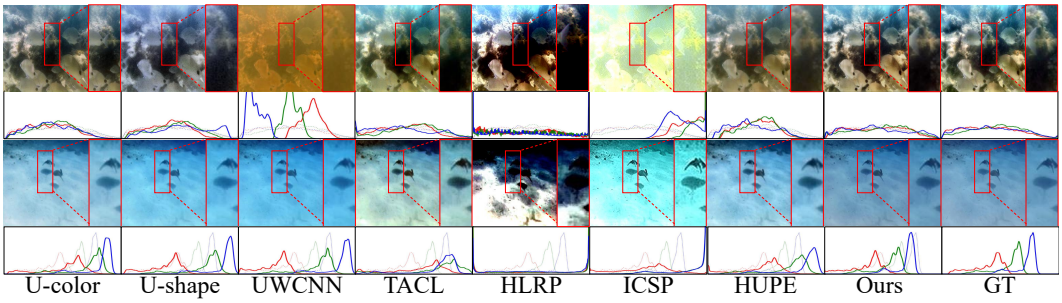


Figure 4: **RGB histogram analysis on the UIEB and EUVP datasets.** Each sample includes the enhanced result, a locally zoomed region highlighted by a red box, and the corresponding RGB histograms (horizontal axis: pixel intensity in the range 0–255; vertical axis: frequency). Solid curves denote the enhanced distributions, while dashed curves indicate the ground truth distributions. A higher degree of curve overlap implies more accurate color restoration.

where $[\cdot;\cdot]$ denotes channel-wise concatenation. Complementary biases are used to encourage different frequency preferences: the GCC branch adopts $\boldsymbol{\beta}_{\text{color}} = [+0.3, -0.3]^\top$, while the DR branch adopts $\boldsymbol{\beta}_{\text{haze}} = [-0.3, +0.3]^\top$. These biases provide initialization preferences, whereas the final coefficients remain content-adaptive.

The output feature is obtained through residual fusion:

$$F_{out} = F_{in} + \alpha_{low}\text{Conv}(f_{low}) + \alpha_{high}\text{Conv}(f_{high}). \tag{10}$$

Enhancement objective. The enhancement network is optimized using reconstruction, color, and detail-oriented objectives:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \lambda_c\mathcal{L}_c + \lambda_s\mathcal{L}_s, \tag{11}$$

where $\mathcal{L}_{\text{rec}}$ combines MSE and SSIM, $\mathcal{L}_c$ constrains global RGB statistics for color consistency, and $\mathcal{L}_s$ applies a dark-channel-prior-based sparsity constraint to reduce scattering-related veiling and improve local structures. Together with the cascaded architecture and complementary frequency biases, these objectives encourage functional specialization between global color correction and local detail restoration rather than enforcing a strict separation.

4 Experiments

4.1 Experimental Settings

Datasets: LSUI (Peng et al., 2023), EUVP (Islam et al., 2020), UIEB (Li et al., 2020b), and U45 (Wang et al., 2020). Compared methods: U-color (Li et al., 2021), U-shape (Peng et al., 2023), UWCNN (Li et al., 2020a), TACL (Liu et al., 2022), HLRP (Zhuang et al., 2022), ICSP (Hou et al., 2024), HUPE (Zhang et al., 2025). Evaluation metrics: PSNR, SSIM, LPIPS, and PCQI are adopted as full-reference metrics. UIQM (Panetta et al., 2016), UCIQE (Yang & Sowmya, 2015),

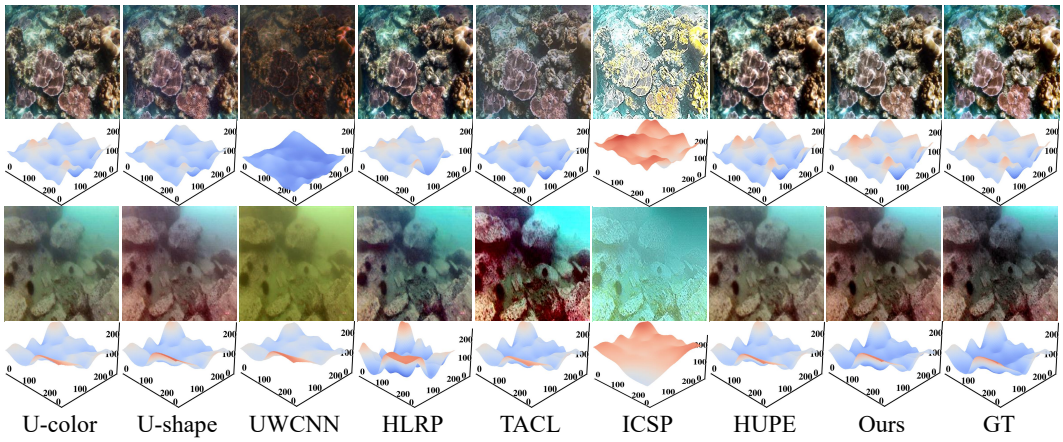


Figure 5: **Comparison of spatial luminance distributions on the UIEB and LSUI datasets.** The upper row shows the enhanced results produced by different methods, while the lower row presents the corresponding luminance surface plots. The X and Y axes represent pixel coordinates, and the Z axis denotes pixel luminance values.

and NIQE (Mittal et al., 2013) are used as no-reference metrics. Training configuration: All models are trained on an NVIDIA GeForce RTX 3090 GPU. A cosine annealing learning-rate scheduler with a period of 200 is employed. The initial learning rate is set to 0.0002, and the total number of training epochs is 200.

4.2 Evaluation of Underwater Image Datasets

Different underwater image enhancement methods are evaluated on LSUI, EUVP, UIEB, and U45 datasets, with results reported in Table 1. Figure 6 further presents PSNR percentile curves to characterize performance distributions beyond mean values. On EUVP and LSUI, our method consistently outperforms the compared methods across all percentiles, including the lower tail (P10–P25), indicating fewer failure cases on challenging images with severe color cast, low contrast, or strong haze. On UIEB, it remains competitive in the mid-to-high percentile range, showing strong performance on moderately degraded samples. Overall, the percentile analysis demonstrates that the proposed method delivers consistent dataset-level gains and improved robustness, particularly on difficult inputs.

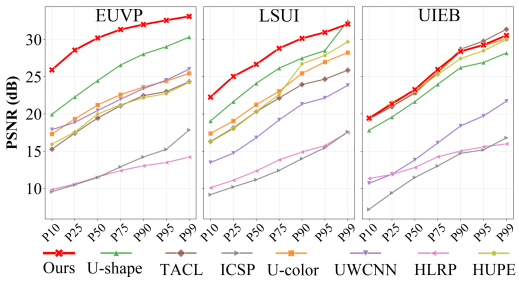


Figure 6: **PSNR percentile comparison.** The curves show PSNR values at different percentiles, with red denoting ours.

Figure 4 compares the RGB-channel distributions produced by different methods. In contrast to the distribution shifts observed in other methods, the proposed method exhibits close agreement with the ground truth in both overall shape and peak locations. This behavior is attributed to the MDP-based strategy controller, which adjusts restoration trajectories according to degradation-related states and reduces statistical bias induced by fixed schedules. In addition, the model bias in MBB incorporates priors of underwater degradations, guiding the network toward feature representations consistent with natural-image statistics. As shown in Figure 5, the luminance surfaces produced by other methods contain prominent spikes and depressions, suggesting residual scattering and local over-enhancement. With dynamic scheduling by the strategy controller, the proposed method better constrains the joint optimization of global luminance and local details, yielding smoother and more continuous intensity surfaces that are closer to the ground truth.

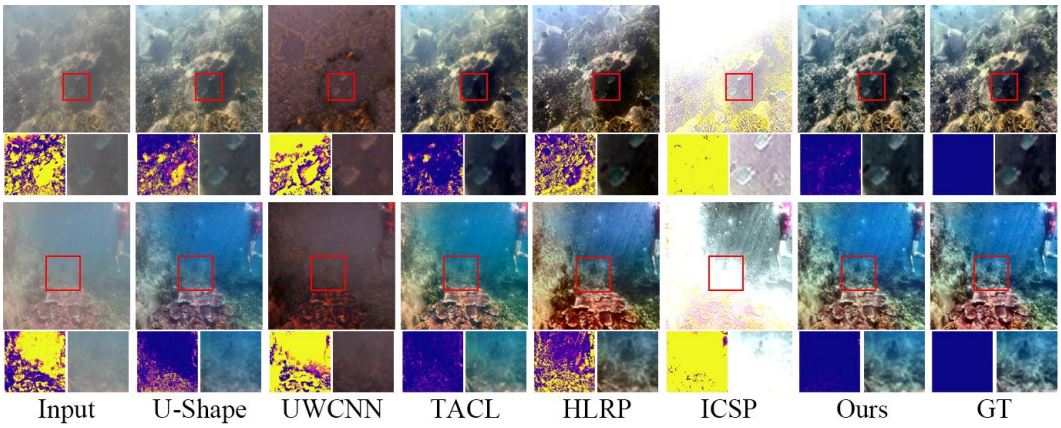


Figure 7: **Pixel-wise difference comparison on UIEB.** Lower-left heatmaps show ground-truth deviations (warmer colors indicate larger errors), while lower-right regions magnify error areas.

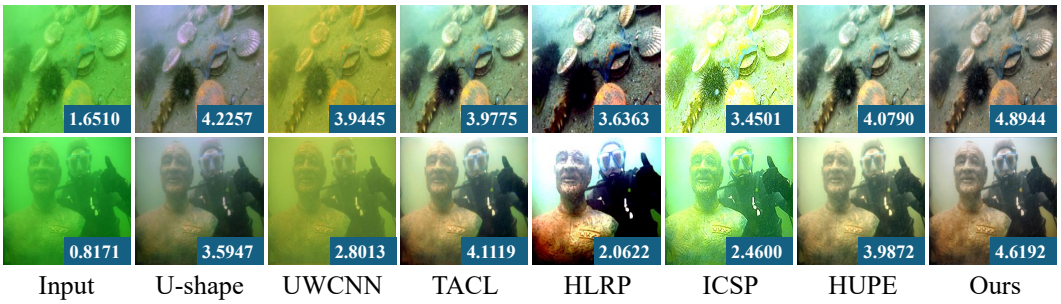


Figure 8: **Visual quality comparison on the U45 dataset.** The values shown in the lower-right corner of each image indicate the corresponding UIQM scores.

Figure 7 focuses on scattering-dominated scenes, where severe veiling reduces visibility and weakens local structures. Compared with competing methods, our method more effectively suppresses scattering artifacts and recovers object boundaries and fine textures. Its weaker error responses and magnified regions closer to the ground truth demonstrate superior detail restoration under strong scattering degradation. Figure 8 presents U45 images dominated by severe green color casts. Our method more effectively removes global chromatic bias and restores natural, balanced colors, whereas competing methods retain casts or introduce over-correction. Together, these results support the complementary roles of GCC and DR and the scheduler's coordination during optimization.

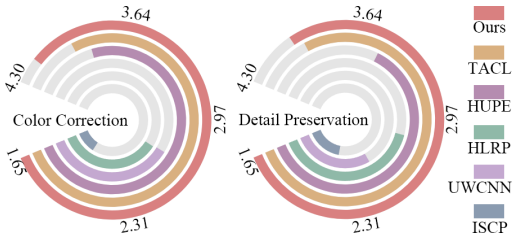


Figure 9: **User study results.** Our method leads in color correction and detail preservation.

4.3 User Study

We also conducted a user study to evaluate the perceptual quality of the enhanced images. We randomly selected 30 underwater images from the UIEB and EUVP datasets and invited six image processing experts to rate the results of each method. The experts assessed color correction and detail preservation on a five-point scale, where 1 indicates the poorest performance and 5 the best. As shown in Figure. 9, our method achieved the highest average scores on both criteria, supporting its effectiveness in underwater image enhancement.

Table 2: Ablation studies of reward functions, training strategies, and model modules. The best results within each group are shown in bold.

(a) Reward ablation

Reward weights	PSNR ↑	SSIM ↑	PCQI ↑
$(\lambda_m, 0, 0)$	23.55	0.909	0.915
$(0, 0, \lambda_d)$	23.50	0.909	0.919
$(0, \lambda_{\mathcal{L}}, 0)$	23.32	0.903	0.918
$(\lambda_m, 0, \lambda_d)$	23.51	**0.910**	0.917
$(\lambda_m, \lambda_{\mathcal{L}}, 0)$	23.52	0.909	0.908
$(0, \lambda_{\mathcal{L}}, \lambda_d)$	23.58	**0.910**	0.913
$(\lambda_m, \lambda_{\mathcal{L}}, \lambda_d)$	**23.64**	**0.910**	**0.922**

(b) Strategy ablation

Strategy	PSNR ↑	CCF ↑	UCIQE ↑
Joint	23.07	25.74	**0.621**
Alternate	23.05	25.78	0.615
Weighted	23.37	24.62	0.612
Ours	**23.64**	**26.88**	**0.621**

(c) Module ablation

Model	PSNR ↑	SSIM ↑	PCQI ↑
w/o FDGF	23.62	0.905	0.920
w/o MBB	23.33	0.909	0.918
Full model	**23.64**	**0.910**	**0.922**

4.4 Ablation Study

Ablation experiments are further conducted on the UIEB dataset. Table 2(b) shows that state-conditioned scheduling better adapts module participation and relative update emphasis than fixed alternating, joint, or weighted strategies. Figure 10 highlights and enlarges local detail regions using differently colored rectangles to enable direct visual comparison, with red denoting our method. Other strategies introduce noticeable red-channel artifacts, whereas our method preserves clearer structures and a more natural color appearance. Consistently, our method ranks best on PSNR, CCF, and UCIQE, indicating a better balance between reconstruction and perceptual quality.

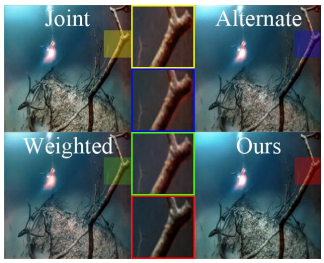


Figure 10: **Strategy ablation.** Colored boxes highlight magnified image across strategies.

Table 2(a) evaluates the reward components weighted by λ_m, $\lambda_{\mathcal{L}}$, and λ_d. The full reward ranks first in PSNR and PCQI while sharing the best SSIM result. Across the reward configurations, each signal provides useful supervision, but no single component dominates every criterion. The reconstruction term favors fidelity, the loss term reflects optimization progress, and the frequency-alignment term penalizes imbalanced low- and high-frequency recovery. Their complementary behavior explains the advantage of combining all three signals. Compared with the setting using only λ_m and $\lambda_{\mathcal{L}}$, adding the λ_d-weighted frequency-alignment term improves all three metrics, supporting its contribution to the combined reward.

Table 2(c) confirms the effectiveness of MBB and FDGF. MBB breaks the symmetry between the two branches and encourages complementary global-appearance and local-detail specialization, while FDGF enables frequency-aware decomposition and fusion. The results show removing either component reduces performance. Without FDGF, frequency cues are not explicitly separated and recombined; without MBB, the two branches lose their complementary biases. This supports using both modules. Their combination produces the best and most stable restoration results.

5 Conclusion

This paper presents a learning-to-train framework for underwater image enhancement. We formulate module update scheduling as an MDP and use a PPO-based controller to select alternating, joint, or weighted updates according to reconstruction and frequency-alignment trends. Together with a cascaded enhancer comprising GCC and DR modules, as well as a multi-signal reward, the scheduler learns stage-adaptive policies coordinating color recovery and detail restoration. Experiments on four benchmarks show a better overall trade-off than fixed strategies and the highest PSNR among the compared methods on all three paired datasets.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, translation. They were not used to develop the research methodology, design or conduct experiments, analyze data, interpret results, or formulate scientific claims. All AI-assisted text and formatting changes were reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

REFERENCES

Haoyu Chen, Wenbo Li, Jinjin Gu, Jingjing Ren, Sixiang Chen, Tian Ye, Renjing Pei, Kaiwen Zhou, Fenglong Song, and Lei Zhu. Restoreagent: Autonomous image restoration agent via multimodal large language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 110643–110666, 2024a.

Long Chen, Yunzhou Xie, Yaxin Li, Qi Xu, and Junyu Dong. Cwscnet: Channel-weighted skip connection network for underwater object detection. *IEEE Transactions on Image Processing*, 33:5206–5218, 2024b.

Marcos V. Conde, Gregor Geigle, and Radu Timofte. Instructir: High-quality image restoration following human instructions. In *Proceedings of the European Conference on Computer Vision*, pp. 1–21, 2024.

Yimian Ding, Jingzehua Xu, Yiyuan Yang, Guanwen Xie, Xinqi Wang, and Shuai Zhang. Aoi-mdp: An aoi optimized markov decision process dedicated in the underwater task. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 29348–29350, 2025.

Guodong Fan, Yu Zhou, Jingchun Zhou, Yakun Ju, Guang-Yong Chen, Jinjiang Li, and Alex C. Kot. Dcd-uie: Decoupled chromatic diffusion model for underwater image enhancement. *IEEE Transactions on Image Processing*, 35:449–464, 2026.

Hang Guo, Tao Dai, Yuanchao Bai, Bin Chen, Xudong Ren, Zexuan Zhu, and Shu-Tao Xia. Parameter efficient adaptation for image restoration with heterogeneous mixture-of-experts. In *Advances in Neural Information Processing Systems*, 2024.

Zeeshan Hayder, Ali Cheraghian, Lars Petersson, Mehrtash Harandi, and Richard Hartley. Dto-kd: Dynamic trade-off optimization for effective knowledge distillation. In *Proceedings of the Fourteenth International Conference on Learning Representations*, 2026.

Guojia Hou, Nan Li, Peixian Zhuang, Kunqian Li, Haihan Sun, and Chongyi Li. Non-uniform illumination underwater image restoration via illumination channel sparsity prior. *IEEE Transactions on Circuits and Systems for Video Technology*, 34:799–814, 2024.

Md Jahidul Islam, Youya Xia, and Junaed Sattar. Fast underwater image enhancement for improved visual perception. *IEEE Robotics and Automation Letters*, 5:3227–3234, 2020.

Qiuping Jiang, Yuese Gu, Zongwei Wu, Chongyi Li, Huan Xiong, Feng Shao, and Zhihua Wang. Deep underwater image quality assessment with explicit degradation awareness embedding. *IEEE Transactions on Image Processing*, 34:1297–1310, 2025.

Jianyu Lai, Sixiang Chen, Yunlong Lin, Tian Ye, Yun Liu, Song Fei, Zhaohu Xing, Hongtao Wu, Weiming Wang, and Lei Zhu. Snowmaster: Comprehensive real-world image desnowing via mllm with multi-model feedback optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4302–4312, 2025.

Chongyi Li, Saeed Anwar, and Fatih Porikli. Underwater scene prior inspired deep underwater image and video enhancement. *Pattern Recognition*, 98:107038, 2020a.

Chongyi Li, Chunle Guo, Wenqi Ren, Runmin Cong, Junhui Hou, Sam Kwong, and Dacheng Tao. An underwater image enhancement benchmark dataset and beyond. *IEEE Transactions on Image Processing*, 29:4376–4389, 2020b.

Chongyi Li, Saeed Anwar, Junhui Hou, Runmin Cong, Chunle Guo, and Wenqi Ren. Underwater image enhancement via medium transmission-guided multi-color space embedding. *IEEE Transactions on Image Processing*, 30:4985–5000, 2021.

Yunxin Li, Shenyuan Jiang, Baotian Hu, Longyue Wang, Wanqi Zhong, Wenhan Luo, Lin Ma, and Min Zhang. Uni-moe: Scaling unified multimodal llms with mixture of experts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5), 2025. doi: 10.1109/TPAMI.2025.3532688.

Hongkai Lin, Dingkang Liang, Zhenghao Qi, and Xiang Bai. A unified image-dense annotation generation model for underwater scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 961–970, 2025.

Risheng Liu, Zhiying Jiang, Shuzhou Yang, and Xin Fan. Twin adversarial contrastive learning for underwater image enhancement and beyond. *IEEE Transactions on Image Processing*, 31:4922–4936, 2022.

Rui Liu, Sheng Fan, Wenguan Wang, and Yi Yang. Underwater visual slam with depth uncertainty and medium modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 970–980, 2025a.

Yi Liu, Qiuping Jiang, Xingbo Li, Ting Luo, and Wenqi Ren. Toward better than pseudo-reference in underwater image enhancement. *IEEE Transactions on Image Processing*, 34:6168–6179, 2025b.

Tobias Meggendorfer, Maximilian Weininger, and Patrick Wienhöft. Solving robust markov decision processes: Generic, reliable, efficient. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20:209–212, 2013.

Karen Panetta, Chen Gao, and Sos Agaian. Human-visual-system-inspired underwater image quality measures. *IEEE Journal of Oceanic Engineering*, 41:541–551, 2016.

Lintao Peng, Chunli Zhu, and Liheng Bian. U-shape transformer for underwater image enhancement. *IEEE Transactions on Image Processing*, 32:3066–3079, 2023.

Vaishnav Potlapalli, Syed Waqas Zamir, Salman H Khan, and Fahad Shahbaz Khan. Promptir: Prompting for all-in-one image restoration. In *Advances in Neural Information Processing Systems*, volume 36, pp. 71275–71293, 2023.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017.

Mengkun She, Felix Seegräber, David Nakath, Patricia Schöntag, and Kevin Köser. Relative illumination fields: Learning medium and light independent underwater scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 29110–29119, 2025.

Yunkai Tang, Chengxuan Zhu, Renjie Wan, Chao Xu, and Boxin Shi. Neural underwater scene representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11780–11789, 2024.

Zheng'ao Wang, Tanlin Li, and Wen Qu. Underwater image enhancement using dual adversarial network. In *Proceedings of the 4th International Conference on Graphics and Signal Processing*, pp. 16–21, 2020.

Xinwei Xue, Tianjiao Ma, Yidong Han, Long Ma, and Risheng Liu. Learning deep scene curve for fast and robust underwater image enhancement. *IEEE Signal Processing Letters*, 31:6–10, 2024.

Miao Yang and Arcot Sowmya. An underwater color image quality evaluation metric. *IEEE Transactions on Image Processing*, 24:6062–6071, 2015.

Xiaowen Ye, Xianxin Song, Yi Wu, Hao Xu, and Jun Zhang. Integrated sensing and communication for underwater acoustic networks based on deep reinforcement learning. *IEEE Transactions on Mobile Computing*, 25:14482–14497, 2026.

Meng Yu, Liquan Shen, Yihan Yu, Yu Zhang, and Rui Le. Task-driven underwater image enhancement via hierarchical semantic refinement. *IEEE Transactions on Image Processing*, 35:42–56, 2025.

Eduard Zamfir, Zongwei Wu, Nancy Mehta, Yuedong Tan, Danda Pani Paudel, Yulun Zhang, and Radu Timofte. Complexity experts are task-discriminative learners for any image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12753–12763, 2025.

Fan Zhang, Shaodi You, Yu Li, and Ying Fu. Atlantis: Enabling underwater depth estimation with stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11852–11861, 2024.

Zengxi Zhang, Zhiying Jiang, Long Ma, Jinyuan Liu, Xin Fan, and Risheng Liu. Hupe: Heuristic underwater perceptual enhancement with semantic collaborative learning. *International Journal of Computer Vision*, pp. 1–19, 2025.

Jingchun Zhou, Zongxin He, Dehuan Zhang, Siyuan Liu, Xianping Fu, and Xuelong Li. Spatial residual for underwater object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47:4996–5013, 2025.

Wujie Zhou, Yiben Li, Qiuping Jiang, Liang Liao, Runmin Cong, and Weisi Lin. Decouple-then-synergize: A self-paced collaborative learning network for rgb-t snowy urban scene parsing. *IEEE Transactions on Image Processing*, 35:5758–5772, 2026.

Peixian Zhuang, Jiamin Wu, Fatih Porikli, and Chongyi Li. Underwater image enhancement with hyper-laplacian reflectance priors. *IEEE Transactions on Image Processing*, 31:5442–5455, 2022.