Physics Meets Perception: A Reinforcement Learning Framework for Unpaired Real-World Image Dehazing

Yunwei Lan[1,2], Zhigao Cui[1] ✉, Chang Liu[2], Menglin Zhang[2], Nian Wang[1], Cong Zhang[1], and Dong Liu[2] ✉

[1] Rocket Force University of Engineering, Xi'an 710025, China
[2] MOE Key Laboratory of Brain-Inspired Intelligent Perception and Cognition, University of Science and Technology of China, Hefei 230093, China
(ywlan, lc980413, zhangmenglin)@mail.ustc.edu.cn, dongeliu@ustc.edu.cn, cuizg10@126.com, nianwang04@outlook.com, zhangc1022@126.com

Abstract. Real-world image dehazing remains a challenging problem due to the significant domain gap between synthetic training data and natural haze, often leading to poor generalization. While unsupervised Generative Adversarial Networks (GANs) attempt to bridge this gap, they frequently suffer from training instability and hallucinatory artifacts. Recently, Reinforcement Learning (RL) has emerged as a promising alternative; however, current RL-based restoration paradigms predominantly rely on diffusion models, leading to prohibitive computational costs and ineffective exploration. To address these bottlenecks, we propose Dehaze-RL, an efficient framework tailored for unpaired real-world dehazing. Bypassing the expensive iterative sampling of diffusion models, we design an efficient policy network to predict hybrid actions in a Physics Hybrid Action Space. By explicitly estimating physical parameters, enhancement factors, and a gating policy map, we anchor the agent's exploration in reliable physical priors, effectively resolving the exploration dilemma. These components are dynamically integrated via a gating fusion mechanism. Furthermore, we introduce a Multi-Granularity Reward Mechanism to provide comprehensive feedback, seamlessly aligning the outputs with human perception without structural distortions. Extensive experiments demonstrate that Dehaze-RL outperforms state-of-the-art methods in both visual fidelity and quantitative metrics.

Keywords: Image dehazing · Reinforcement learning · Physical prior

1 Introduction

Image dehazing plays a pivotal role in computer vision, aiming to restore visibility and color fidelity from degraded observations. It serves as an essential prerequisite for numerous high-level vision tasks, such as autonomous driving

✉ Corresponding authors.

Fig. 1: Pipeline comparison and qualitative dehazing results. (a) and (b) contrast traditional dehazing paradigms with our proposed Dehaze-RL framework. Visual results in (c) demonstrate that our RL refinement significantly surpasses the warmed-up baseline in both synthetic and real-world hazy environments.

and video surveillance, where clear visual input is imperative for robust performance. Mathematically, the haze degradation process is typically formulated by the Atmospheric Scattering Model (ASM) [27]:

$$I(x) = J(x)t(x) + A(1 - t(x)), \tag{1}$$

where $I(x)$ and $J(x)$ denote the observed hazy image and the latent clean scene, respectively. A represents the global atmospheric light, and $t(x)$ describes the transmission map. Early prior-based methods [7,12,53] primarily relied on handcrafted priors to estimate the unknown parameters (t and A), subsequently obtaining the dehazed result J via the ASM. However, they often struggle in complex real-world scenarios, leading to inevitable color distortions and artifacts.

With the advent of deep learning [13–15], data-driven neural networks have significantly advanced image dehazing. They generally follow two strategies: estimating the physical parameters of the ASM (e.g., DCPDN [47]) or learning a direct mapping to the clear image (e.g., AOD-Net [19], DEANet [4], C2PNet [51], and HazeFlow [34]). Despite achieving impressive results, these methods predominantly rely on large-scale synthetic datasets for training. Consequently, they often falter in real-world scenarios due to the significant domain gap between simplistic synthetic models and the complex degradation.

To break free from the dependence on synthetic paired data and achieve robust restoration in real-world scenarios, the community has pivoted towards unpaired training paradigms [21,38,44,50]. Without pixel-wise supervision, these methods typically employ Generative Adversarial Networks (GANs) [11] to align generated outputs with the distribution of clear images. However, despite generating visually plausible results, GAN-based approaches suffer from inherent drawbacks: (1) Training Instability: The adversarial min-max game is notoriously difficult to stabilize and prone to mode collapse; (2) Hallucinations: Generators often introduce unrealistic artifacts or alter semantic content to fool the discriminator; (3) Misaligned Objectives: The discriminator optimizes for distribution matching rather than perceptual quality, often failing to distinguish between a naturally clear image and a dehazed but distinctively noisy one.

Reinforcement Learning (RL) has achieved transformative success in aligning Large Language Models (LLMs) with human preferences. Algorithms such as PPO [32] and the recently popularized GRPO [22] have established a robust paradigm for optimizing non-differentiable rewards, a success now extending to complex generative tasks (e.g., FlowGRPO [25]). Inspired by this, recent efforts [24, 39, 41] have attempted to migrate RL to image restoration by treating the iterative denoising steps of Diffusion Models as a sequential decision-making process. However, directly transplanting this strategy to image dehazing is inherently problematic: (i) the multi-step sampling incurs prohibitive computational overhead; (ii) exploration based on abstract latent noise is often inefficient for conditional restoration where the output is strictly constrained by the hazy input observation. To bridge this gap, IRPO [40] pioneers the integration of GRPO into restoration fine-tuning. While efficient, it relies on generic frequency modulation as its action space, which lacks physical interpretability for dehazing and struggles to disentangle physical parameters. Furthermore, IRPO remains confined to a supervised paradigm requiring paired ground-truth references, severely limiting its generalization to diverse real-world scenarios.

To address these challenges, we propose Dehaze-RL, a novel framework that leverages RL for efficient, unpaired real-world dehazing. Bypassing the computationally expensive iterative sampling of diffusion models, our Dehaze-RL employs an efficient policy network that acts as an intelligent agent. As illustrated in Figs. 1(a) and (b), instead of learning a direct mapping from hazy to clear images, our agent predicts hybrid actions within a Physics Hybrid Action Space: estimating physical variables (atmospheric light A, transmission map t) and Image Signal Processing (ISP) enhancement factors (contrast C, saturation S), seamlessly fused via a dynamic gating map (g). By explicitly anchoring the agent's exploration in reliable physical priors, we effectively resolve the inefficient exploration dilemma. Furthermore, to guide the agent in the absence of ground truth, we introduce a Multi-Granularity Reward Mechanism encompassing semantic, perceptual, and structural consistency rewards. This comprehensive system effectively aligns the agent's policy with human visual perception. As visually demonstrated in Fig. 1(c), our Dehaze-RL significantly enhances clarity and artifact suppression across both synthetic and real-world scenarios when compared with the warmed-up baseline. Our main contributions are as follows:

(1) We propose Dehaze-RL, a novel and efficient reinforcement learning framework tailored for unpaired real-world image dehazing.

(2) We design a Physics Hybrid Action Space that tightly integrates physical scattering models with ISP enhancement factors, effectively resolving the ineffective exploration dilemma in conditional dehazing tasks.

(3) We introduce a Multi-Granularity Reward Mechanism to provide comprehensive visual quality feedback, achieving robust alignment between the unsupervised optimization objective and human visual perception.

(4) Extensive experiments demonstrate that Dehaze-RL significantly outperforms state-of-the-art methods.

2 Related Work

2.1 Image Dehazing

Prior-based Dehazing Methods. Early dehazing algorithms primarily rely on hand-crafted priors to estimate physical parameters. Representative methods include the Dark Channel Prior (DCP) [12], based on minimum intensity statistics, and the Color-Line Prior [7], leveraging the RGB collinearity of local patches. While interpretable, these methods often falter in complex scenarios, leading to significant color distortion and artifacts.
Supervised Learning-based Methods. Deep learning approaches generally follow two paradigms. ASM-based networks (e.g., DCPDN [47]) embed the physical model to regress transmission and atmospheric light independently. Conversely, end-to-end networks learn a direct mapping from hazy to clear images. Representative models range from early CNNs like AOD-Net [19] to recent advanced designs such as DEANet [4], C2PNet [51], and DehazeFormer [35]. Despite their success, these methods rely heavily on synthetic paired data, often suffering from performance drops when facing real-world degradation.
Unsupervised Dehazing. To tackle the domain gap, unsupervised methods [17, 18] learn directly from unpaired real-world data. Early methods like CycleDehaze [5] utilize cycle-consistency. Subsequent works incorporate physical constraints (e.g., RefineDNet [50]) or latent space regularization via contrastive learning (UCL [38]) and self-augmented consistency (D4+ [44]). However, these approaches predominantly rely on Generative Adversarial Networks (GANs). Consequently, they inherit the inherent flaws of adversarial min-max optimization, frequently suffering from training instability, mode collapse, and hallucinatory color artifacts. In stark contrast, our Dehaze-RL employs a stable, reward-driven reinforcement learning framework to achieve visually faithful restoration.

2.2 Reinforcement Learning in Vision

Early RL for Image Restoration. Early RL-based restoration treated the task as a sequential decision-making process using discrete actions. For instance, PixelRL [10] employs a multi-agent strategy where each pixel selects filters from a predefined bank, while RL-Restore [46] learns to dynamically choose lightweight CNNs from a toolbox. While effective for optimizing non-differentiable metrics, these methods rely on discrete selections or simple adjustments, lacking the capability to explicitly model the complex physical degradation inherent in dehazing.
RL for Generation and Restoration. Recent advancements have integrated RL to align deep models with human preferences. Generative RL algorithms, such as FlowGRPO [25] and DanceGRPO [42], optimize the iterative generative process (e.g., flow trajectories or diffusion chains) by fine-tuning the policy network. Inspired by this, several works [39, 41] adapt these paradigms to image restoration. However, noise-based exploration in conditional tasks is often ineffective and computationally expensive due to iterative sampling. To improve efficiency, IRPO [40] introduces Group Relative Policy Optimization (GRPO) to

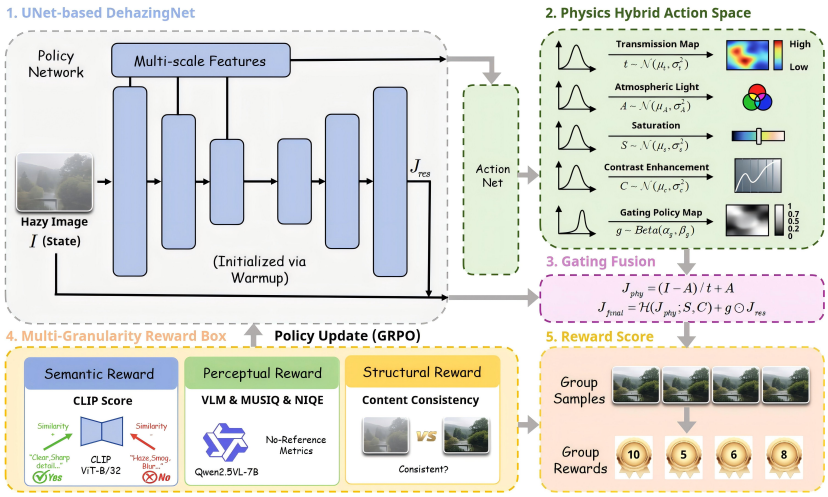

Fig. 2: Architectural framework of Dehaze-RL. The GRPO-optimized Policy Network utilizes multi-scale features to predict explicit parameters (t, A, S, C, g) within a Physics Hybrid Action Space. These actions are dynamically integrated via Gating Fusion. Finally, a Multi-Granularity Reward Mechanism guides the policy update to seamlessly align the unsupervised restoration with human visual perception.

fine-tune backbones via adaptive frequency modulation. However, IRPO remains limited by its supervised nature (requiring paired ground truth) and an abstract, non-physical action space. In contrast, our Dehaze-RL predicts hybrid actions (t, A, S, C, g) within a Physics Hybrid Action Space. By operating in an unpaired manner guided by a Multi-Granularity Reward Mechanism, Dehaze-RL enables efficient optimization tailored for real-world dehazing.

3 Method

3.1 Overview and Problem Formulation

We propose Dehaze-RL, a hybrid reinforcement learning framework tailored for real-world unpaired image dehazing. As illustrated in Fig. 2, we formulate the dehazing task as a Contextual Bandit problem. Our complete RL formulation is characterized by four core components, denoted as $\langle \mathcal{S}, \mathcal{A}, \pi, \mathcal{R} \rangle$:

One-Shot Decision Process. Distinct from traditional sequential RL tasks (e.g., robotics or diffusion denoising) that require trajectory optimization over a horizon $T > 1$, our framework treats dehazing as a one-shot decision process. The agent observes the hazy input (State), executes an instantaneous action (Parameter Prediction), and receives an immediate reward. This formulation eliminates the complexity of estimating future returns, allowing us to leverage the efficient GRPO algorithm for policy optimization.

State Space ($\mathcal{S}$). The state $s \in \mathcal{S}$ represents the context of the current instance. In our formulation, the state is strictly the observed hazy image $I \in \mathbb{R}^{H\times W\times 3}$.
Action Space ($\mathcal{A}$). We define a Physics Hybrid Action Space to ensure physical interpretability. The action $a \in \mathcal{A}$ is a composite vector $a = [t, A, S, C, g]$ comprising three facets: (1) Physical Parameters (t, A): the transmission map and global atmospheric light; (2) ISP Enhancement Factors (S, C): global saturation and contrast for visual refinement; (3) Gating Policy Map (g): a spatially variant map that dynamically fuses the physics-based restoration and the learned residual to achieve optimal visual quality.
Policy (π). Initialized by a warmed-up baseline, the policy $\pi_\theta(a|s)$ parameterized by our policy network with learnable weights θ, maps the hazy context to the probability distributions of these optimal parameters. During training, the agent samples actions to explore diverse restoration strategies; during inference, it employs the deterministic means to ensure stable outputs.
Reward ($\mathcal{R}$). The optimization objective is to maximize the immediate reward $R(s, a)$. We design a Multi-Granularity Reward Mechanism comprising three distinct components: (1) Semantic Reward (CLIP) for text-driven optimization direction; (2) Perceptual Reward combining high-level naturalness feedback from a Vision-Language Model (Qwen2.5-VL-7B [3]) with no-reference perceptual metrics (MUSIQ [16], NIQE [26]); and (3) Structural Reward to enforce content consistency. This composite reward completely replaces paired ground truth, directly guiding the GRPO policy update toward human-aligned visual realism.

3.2 Physics Hybrid Action Space

Unlike standard generative RL methods that explore within an abstract latent noise space, we design a continuous Physics Hybrid Action Space to ensure physical interpretability and efficient exploration. The agent predicts a composite action $a = [t, A, S, C, g]$. To enable effective exploration during training, we model these parameters as stochastic distributions rather than deterministic scalars.
Network Architecture. The policy network π_θ employs a U-Net-like backbone [31]. Specifically, multi-scale features extracted from the encoder are fused and routed into two specialized action heads: a local branch utilizing lightweight convolutions to predict spatially-variant maps (t, g), and a global branch leveraging pooling and Multi-Layer Perceptrons (MLPs) to regress scene-level scalar parameters (A, S, C). Concurrently, the decoder of the backbone outputs a learned residual feature map J_{res}, which compensates for complex high-frequency details during the final gating fusion. **Detailed network architectures are deferred to the supplementary material.**

Crucially, to stabilize early exploration, we incorporate the Dark Channel Prior (DCP) [12] as a physical anchor. Rather than predicting physical parameters from scratch, the network learns residual corrections (Δ_t, Δ_A) relative to the DCP estimates: $\mu_t = \text{logit}(t_{dcp}) + \Delta_t$ and $\mu_A = \text{logit}(A_{dcp}) + \Delta_A$. This strong physical inductive bias initializes the agent in a scientifically plausible state, significantly accelerating RL convergence.

Stochastic Action Sampling. To ensure physical validity and stable exploration, we employ specific probability distributions to enforce mathematical constraints on our continuous actions:

(1) Gaussian Distributions (t, A, S, C): We predict the mean μ and standard deviation σ to model these parameters as $\mathcal{N}(\mu, \sigma^2)$. Following the reparameterization trick, we sample latent variables and apply custom activations to bound their ranges: $t \in [0.1, 1.0]$ and $A \in [0.25, 1.0]$ via scaled Sigmoid functions (preventing division-by-zero and matching illumination priors), while S and C are constrained via Tanh functions to enable bidirectional relative enhancements.

(2) Beta Distribution (g): Since the fusion weight g must be strictly bounded within $(0, 1)$, simple clipping would introduce severe gradient truncation. Instead, we predict concentration parameters α, β to model $g \sim \text{Beta}(\alpha, \beta)$. Crucially, we enforce $\alpha, \beta > 1.1$ via a shifted Softplus activation. This guarantees a unimodal distribution, preventing the spatial gating map from collapsing to binary extremes and ensuring stable gradient flow.

Restoration Process. The sampled actions reconstruct the final clear image J through a unified sequential pipeline. Initially, we invert the atmospheric scattering model to obtain the physically dehazed base $J_{phy} = (I - A)/t + A$. To correct potential color shifts and elevate perceptual realism, we subsequently apply a differentiable ISP function $J_{isp} = \mathcal{H}(J_{phy}; S, C)$, where the contrast factor C scales pixel deviations from the grayscale mean, and saturation S interpolates between the image and its luminance. Finally, the gating map g acts as a spatial confidence mask to dynamically fuse the enhanced image J_{isp} with a learned residual feature J_{res} extracted from the backbone, yielding the ultimate restoration $J_{final} = J_{isp} + g \odot J_{res}$.

3.3 Optimization via GRPO

To circumvent the training instability of GANs and the prohibitive computational overhead, we adopt Group Relative Policy Optimization (GRPO) [22]. GRPO perfectly fits our one-shot Bandit formulation by leveraging group-based feedback. However, naively applying it to image dehazing is highly non-trivial. Standard RL exploration in an unconstrained pixel or latent space exacerbates the ill-posed nature of dehazing, inevitably leading to color shifts and physical artifacts. To overcome this challenge, rather than using GRPO out-of-the-box, we structurally ground its exploration within our Physics Hybrid Action Space and steer its updates via the Multi-Granularity Reward Mechanism, ensuring the optimization remains both physically plausible and visually faithful.

Group Sampling and Advantage. For each hazy input I, the policy π_θ generates a group of G candidate actions $\{a_1, \ldots, a_G\}$ by sampling from the predicted action distributions. These actions are executed to produce G restored images, which are then evaluated to obtain a reward set $R = \{r_1, \ldots, r_G\}$. The advantage A_i for the i-th action is estimated relative to its peers via Eq. 2:

$$A_i = \frac{r_i - \text{mean}(R)}{\text{std}(R) + \epsilon}, \tag{2}$$

where ϵ is a small constant ensuring numerical stability.

Optimization Objective. The policy is optimized to maximize the expected advantage. While standard GRPO typically employs a KL-divergence term to prevent large policy shifts, we empirically observed that such generic constraints can dominate gradients and suppress reward-driven optimization in dehazing tasks. Consequently, we replace the KL loss with our Action Regularization $\mathcal{L}_{reg}$. The total loss $\mathcal{L}_G$ is defined as:

$$\mathcal{L}_G = -\frac{1}{G}\sum_{i=1}^{G}\left(A_i \cdot \log \pi_\theta(a_i|I)\right) + \mathcal{L}_{reg}. \tag{3}$$

Action Regularization ($\mathcal{L}_{reg}$). To prevent the predicted actions from severely deviating from valid initializations and to restrict trivial solutions, we introduce a composite regularization term:

$$\mathcal{L}_{reg} = \lambda_{anc} \cdot \underbrace{\mathbb{E}\left[\text{ReLU}(|t - t_{ref}| - \delta)\right]}_{\text{Anchor Loss}} + \lambda_{res} \cdot \underbrace{\left(\|\Delta_t\|_2^2 + \|\Delta_A\|_2^2\right)}_{\text{Residual Norm}} + \lambda_g \cdot \underbrace{\|g\|_2^2}_{\text{Gating Penalty}}, \tag{4}$$

where t is the transmission, t_{ref} is the reference from the warmed-up baseline, and λ_{anc}, λ_{res}, λ_g are balancing weights. Specifically, the Anchor Loss (margin $\delta = 0.1$) permits local exploration of the transmission t while strictly penalizing severe deviations from the warm-up priors. The Residual Norm minimizes the L_2 norm of parameter residuals (Δ_t, Δ_A), enforcing a minimal action principle that prioritizes the strong physical prior (DCP). Crucially, the Gating Penalty restricts the gating map g to prevent the network from trivially bypassing the physical scattering model and over-relying on the residual compensator.

3.4 Multi-Granularity Reward Mechanism

In the absence of ground-truth references for real-world scenarios, we design a Multi-Granularity Reward Mechanism (as illustrated in Fig. 2) to guide policy optimization. This mechanism evaluates the restoration quality across three explicit visual dimensions: semantic, perceptual, and structural.

Semantic Reward (R_{sem}). To ensure the dehazed output accurately reflects a haze-free scene without semantic ambiguity, we utilize a CLIP-based relative similarity reward. Leveraging a pre-trained CLIP [30] (`ViT-B/32`), R_{sem} computes the cosine similarity between the restored image J and two opposing sets of text prompts: positive prompts $\mathcal{T}_{pos}$ (e.g., "clear photo", "sharp details") and negative prompts $\mathcal{T}_{neg}$ (e.g., "foggy photo", "haze"). Specifically, the reward is defined as the margin between these similarities: $R_{sem} = \cos(\phi(J), \bar{e}_{pos}) - \cos(\phi(J), \bar{e}_{neg})$, where $\phi(\cdot)$ is the visual encoder and $\bar{e}$ denotes the averaged text embeddings. This contrastive formulation directs the exploration toward semantically plausible regions to prevent unnatural color shifts.

Perceptual Reward (R_{per}). This component aligns the restoration with human visual preferences by combining Vision-Language Models (VLMs) with statistical metrics: (1) VLM Feedback: We leverage Qwen2.5-VL-7B [3] as a high-level visual quality critic. By prompting the model to evaluate clarity, color

vividness, and naturalness, we extract a normalized numerical score R_{VLM} from its reasoning. (2) No-Reference (NR) Metrics: To provide fine-grained supervision, we incorporate MUSIQ [16] and NIQE [26]. These metrics reward sharp textures while penalizing unnatural noise, forming the combined perceptual reward: $R_{per} = \lambda_v R_{VLM} + \lambda_m R_{MUSIQ} - \lambda_n R_{NIQE}$.

Structural Reward (R_{str}). To preserve the essential details and prevent content distortion, we enforce a dual-level consistency constraint: (1) Feature-level Consistency: Implemented via PatchNCE [28], we maximize the mutual information between corresponding patches of the hazy input I and the restored output J in a projected feature space. (2) Pixel-level Physical Reversibility: To ensure the dehazing process remains faithful to the original scene, we re-synthesize the hazy observation using the predicted parameters via the ASM: $I_{rec} = J \odot t + A(1-t)$. By penalizing the reconstruction error $R_{phy} = -\|I_{rec} - I\|_1$, we force the agent to find a physically reversible solution that strictly preserves the input's structural identity.

Summary of Reward Function. The total reward for GRPO training is formulated as a weighted sum: $R = \omega_1 R_{sem} + \omega_2 R_{per} + \omega_3 R_{str}$. This composite objective enables the agent to balance semantic clarity, perceptual quality, and structural fidelity in an entirely unsupervised manner. The training process is further stabilized by the action regularization $\mathcal{L}_{reg}$. Notably, our framework replaces unstable adversarial gaming with these rewards, achieving significantly more stable convergence for real-world dehazing. **Further details of the Qwen2.5-VL prompts and reward design are provided in the supplementary material.**

4 Experiments and Discussion

4.1 Experimental Setup

Datasets. Following the unpaired training paradigm, our framework utilizes the RESIDE dataset [20], sampling 7,000 real-world hazy images from the URHI and RTTS subsets as the hazy domain, and 10,000 high-quality unpaired images from SOTS and ADE20K [52] as the clear domain. To evaluate real-world performance, we test on real-world benchmarks including URHI (over 4,000 images), RTTS (over 4,000 images), Haze2020 (over 1,000 images) [33], and Fattal's dataset (over 30 images) [7]. Moreover, we evaluate on the O-HAZE [1] (45 pairs) and I-HAZE [2] (35 pairs) datasets without retraining. These datasets are generated by professional haze machines, serving as robust proxies to validate our method's content preservation.

Implementation Details. Dehaze-RL is implemented in PyTorch. Our training pipeline consists of two phases: (1) Warmup Pre-training: To establish a stable physical mapping and avoid cold-start instability in RL, we first pre-train the policy network using standard unpaired translation constraints (e.g., PatchNCE [28], adversarial loss, and identity loss) to learn a robust initialization. (2) RL Fine-tuning: Initialized with the warmup weights, we perform full-parameter optimization of the entire policy network using the GRPO algorithm.

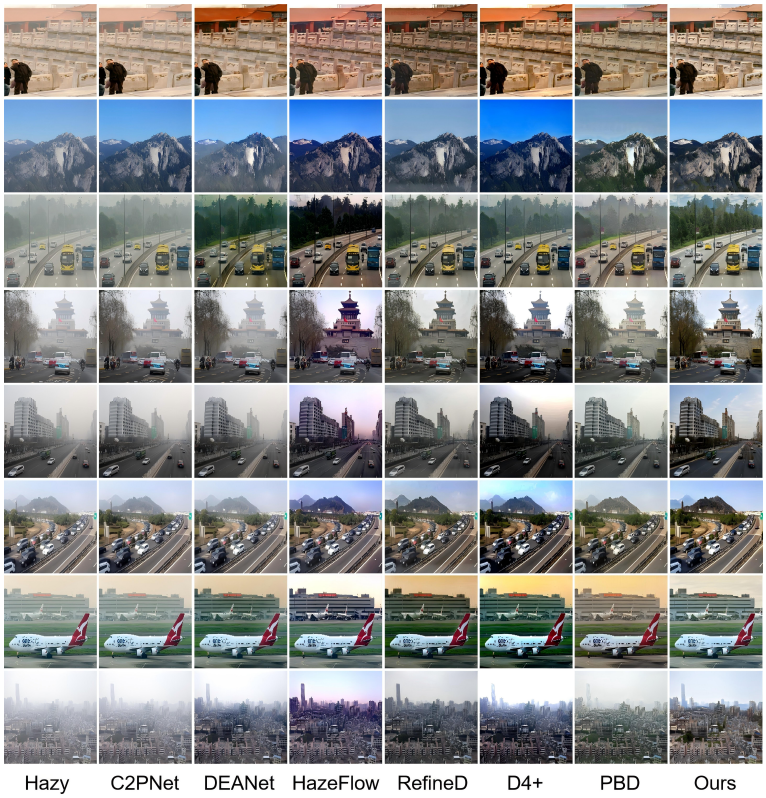

Fig. 3: Comparison on Fattal, RTTS, URHI, and Haze2020 benchmarks. From top to bottom, every two rows show representative results from each dataset. Compared to SOTA methods, Dehaze-RL achieves superior visual fidelity in real-world scenarios.

During the RL phase, the policy network (group size $G = 6$) is optimized via Adam (lr=1×10^{-5}). Training spans three RTX 3090 GPUs: two for policy updates and one hosting Qwen2.5-VL-7B [3] as a dedicated reward server. The hyperparameters are set as follows: total reward weights $(\omega_1, \omega_2, \omega_3) = (1.0, 1.0, 1.0)$, perceptual coefficients $(\lambda_v, \lambda_m, \lambda_n) = (1.0, 0.5, 0.5)$, and action regularization $(\lambda_{anc}, \lambda_{res}, \lambda_g) = (1.0, 0.1, 0.5)$.

For real-world unpaired data, we evaluate the dehazed outputs using no-reference metrics: CLIPIQA [36], MUSIQ [16], MANIQA [43], and NIQE [26]. Conversely, for datasets with paired ground truth, we measure using full-reference metrics: PSNR, SSIM, LPIPS [49], and VSI [48].

4.2 Comparisons with State-of-the-Arts

We compare Dehaze-RL against state-of-the-art methods, categorized into three groups: (1) Prior-based method: DCP [12]; (2) Paired (Supervised) methods:

Table 1: Quantitative comparison on real-world RTTS and URHI datasets. Best results are **bold**, and second-best are underlined. For NIQE, lower is better; for CLIPIQA, MUSIQ, and MANIQA, higher is better.

	RTTS				URHI			
Method	NIQE	CLIPIQA	MUSIQ	MANIQA	NIQE	CLIPIQA	MUSIQ	MANIQA
DCP [12]	4.271	0.276	52.935	0.119	4.129	0.396	55.996	0.143
AODNet [19]	4.697	0.358	55.205	0.145	4.359	0.463	56.678	0.162
MBFormer [29]	4.874	0.383	53.576	0.139	4.174	0.454	57.293	0.163
C2PNet [51]	5.037	0.390	53.960	0.142	4.368	0.464	56.542	0.161
KANet [8]	4.339	0.285	54.513	0.110	3.839	0.404	57.774	0.144
DEANet [4]	4.805	0.379	53.628	0.132	4.233	0.446	56.306	0.155
SGDN [6]	5.716	0.297	44.858	0.081	5.225	0.328	46.817	0.100
IPCDehaze [9]	4.122	0.422	59.424	0.167	3.729	0.451	62.941	0.178
HazeFlow [34]	4.745	**0.431**	63.057	0.160	4.295	**0.469**	63.428	0.181
YOLY [21]	4.553	0.375	51.847	0.129	4.289	0.429	54.896	0.147
RefineD [50]	4.012	0.257	56.117	0.108	3.668	0.384	58.540	0.138
D4 [45]	4.637	0.351	58.445	0.139	4.372	0.454	55.958	0.160
D4+ [44]	4.274	0.346	53.599	0.131	3.942	0.387	55.807	0.147
PBD [37]	4.043	0.252	53.194	0.101	3.565	0.366	56.954	0.139
FrDiff [23]	6.088	0.325	46.935	0.106	5.639	0.340	47.901	0.121
Ours	**3.641**	0.427	**63.921**	**0.168**	**3.505**	0.462	**64.987**	**0.182**

AODNet [19], MBFormer [29], C2PNet [51], KANet [8], DEANet [4], SGDN [6], IPCDehaze [9], and HazeFlow [34]; (3) Unpaired (Unsupervised) methods: YOLY [21], RefineDNet [50], D4 [45], D4+ [44], PBD [37], and FrDiff [23].

Comparison Results on Real-World Datasets. Fig. 3 presents visual comparisons across diverse real-world scenarios. Fully supervised methods (e.g., C2PNet, DEANet) struggle with domain shifts, often leaving dense residual haze in distant backgrounds such as cityscapes and highways. Conversely, unpaired baselines aggressively remove haze but severely compromise color fidelity due to a lack of physical constraints; RefineD introduces unnatural shifts, while D4+ hallucinates over-saturated sky artifacts. In stark contrast, guided by our Physics Hybrid Action Space and multi-granularity rewards, Dehaze-RL effectively clears thick, non-homogeneous haze while flawlessly preserving natural lighting, accurate colors, and crisp structural details.

Table 1 reports comprehensive comparisons on the RTTS and URHI datasets. Dehaze-RL establishes a new state-of-the-art, outperforming competitive unpaired baselines and rivaling recent fully supervised models. Specifically, it achieves the best NIQE scores (3.641 on RTTS / 3.505 on URHI), proving that our physics-grounded RL introduces far fewer statistical distortions than traditional adversarial training. Furthermore, Dehaze-RL dominates perceptual metrics with the highest MUSIQ (63.921 / 64.987) and MANIQA (0.168 / 0.182), validating that our rewards generate visually faithful results. Notably, despite being unpaired, our method achieves highly competitive CLIPIQA scores (0.427 / 0.462), comparable to HazeFlow and C2PNet. This confirms our framework successfully preserves the underlying semantic and structural integrity.

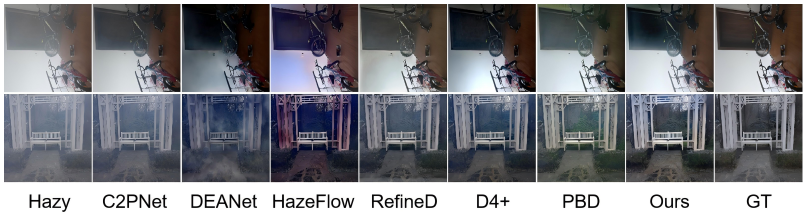

Fig. 4: Qualitative comparison on haze machine-generated datasets. The first row displays visual results from the I-HAZE dataset, while the second row showcases an example from the O-HAZE dataset.

Comparison Results on Machine-Generated Haze. Fig. 4 presents visual comparisons on the I-HAZE and O-HAZE datasets, where haze is simulated using professional haze machines. As observed, prior methods struggle to balance thorough dehazing with structural and color fidelity. DEANet introduces severe structural artifacts, while HazeFlow suffers from prominent unnatural color shifts (e.g., a purple/pink tint). Other baselines, such as RefineD and PBD, tend to leave persistent residual haze. In contrast, our method effectively handles the dense, machine-generated haze while faithfully restoring crisp structural details. **Detailed quantitative evaluations and complexity analysis are provided in the supplementary material.**

4.3 Ablation Study

We conduct extensive ablation studies to validate the effect of our core components: the RL Optimization Strategy, the Physics Hybrid Action Space, and the Multi-Granularity Reward Mechanism.

Ablation on RL Optimization. To isolate the performance gains introduced by our RL fine-tuning, we conduct a cross-paradigm validation across two distinct initialization strategies:

(1) Gain over Supervised Baseline (Synthetic): Standard L_1 regression typically yields overly smooth outputs. Our RL stage effectively breaks this bottleneck (Table 2), boosting MUSIQ (from 53.667 to 55.994) and CLIPIQA (from 0.304 to 0.355). As shown in Fig. 1(c), the rewards guide the agent to restore sharper and more vibrant details.

(2) Gain over Unsupervised Baseline (Real-world): When warmed-up via unpaired adversarial learning (PatchNCE + GAN), RL fine-tuning yields a massive perceptual leap. It boosts MUSIQ (58.927 to 63.921), reduces NIQE (4.226 to 3.641), and effectively corrects baseline color deviations while enhancing high-frequency details (Fig. 1(c)).

These consistent gains demonstrate the universality of our framework. It acts as a paradigm-agnostic booster, enabling the policy network to transcend the inherent perceptual limits of its pre-training objectives.

Ablation on Action Space and Gating Mechanism. We analyze the contribution of the components within our Physics Hybrid Action Space and the

Table 2: Cross-paradigm validation of our RL optimization framework. Whether initialized via supervised regression (Synthetic) or unsupervised adversarial learning (Real-world), our RL fine-tuning consistently improves all perceptual metrics. Notably, real-world warm-up outperforms the synthetic one by bridging the domain gap, providing a superior initial policy that elevates the perceptual ceiling.

Domain	Warmup Paradigm	Method	NIQE ↓	CLIPIQA ↑	MUSIQ ↑	MANIQA ↑
Synthetic	Supervised (L_1)	Warmup Baseline	4.694	0.304	53.667	0.114
		After RL (Ours)	**4.583**	**0.355**	**55.994**	**0.128**
Real-world	Unpaired (GAN)	Warmup Baseline	4.226	0.382	58.927	0.144
		After RL (Ours)	**3.641**	**0.427**	**63.921**	**0.168**

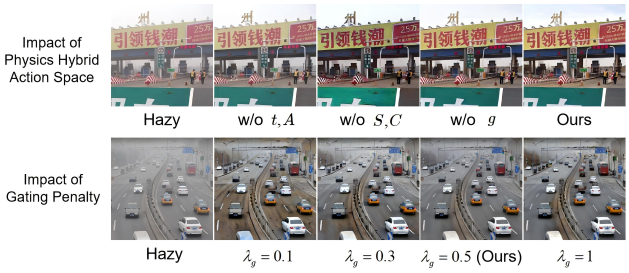

Fig. 5: Visual ablation on the action space and gating mechanism. The top row illustrates the impact of specific action components (t, A, S, C, and g). The bottom row presents a sensitivity analysis of the gating penalty coefficient λ_g, demonstrating that our method ($\lambda_g = 0.5$) achieves the best visual balance.

gating fusion mechanism by constructing three variants: (1) w/o Physical Parameters (t, A), removing the transmission map and atmospheric light; (2) w/o Enhancement Actions (S, C), removing the saturation and contrast adjustments; and (3) w/o Gating Policy Map (g), disabling the spatially adaptive fusion.

As shown in the top row of Fig. 5, without the physical parameters, the model struggles to handle thick haze layers. Without the enhancement actions, the results appear dim and lack visual vibrancy. Without the gating map g, the network fails to adaptively balance the physical and residual branches.

Furthermore, we investigate the impact of the gating penalty coefficient λ_g, which is crucial for preventing the agent from trivially over-relying on the residual branch. As shown in the bottom row of Fig. 5, a relaxed penalty ($\lambda_g \in \{0.1, 0.3\}$) provides insufficient regularization, resulting in severe noise and over-exposure artifacts on the road surface. Conversely, an excessively strict penalty ($\lambda_g = 1.0$) overly suppresses the residual compensator, leading to suboptimal dehazing with residual haze. Our default setting ($\lambda_g = 0.5$) strikes the optimal balance, yielding visually realistic results.

Ablation on Multi-Granularity Reward. To verify the necessity of our multi-granularity reward mechanism, we construct four variants: (1) w/o Semantic Reward, removing the text-image similarity guidance; (2) w/o NR Perceptual

Table 3: Quantitative ablation study of the Multi-Granularity Reward Mechanism.

	RTTS			
Method	NIQE ↓	CLIPIQA ↑	MUSIQ ↑	MANIQA ↑
w/o Semantic Reward (CLIP Score)	3.925	0.406	62.212	0.162
w/o NR Perceptual Metrics (MUSIQ, NIQE)	3.977	0.393	61.575	<u>0.165</u>
w/o VLM Feedback (Qwen2.5-VL)	<u>3.776</u>	0.422	62.680	0.161
w/o Structural Reward (Content Consistency)	3.789	<u>0.425</u>	<u>63.129</u>	0.159
Full Reward (Ours)	**3.641**	**0.427**	**63.921**	**0.168**

Metrics, disabling no-reference image quality feedback; (3) w/o VLM Feedback, removing the high-level evaluation; and (4) w/o Structural Reward, removing the dual-level feature consistency and physical reversibility constraints.

As shown in Table 3, removing any individual component degrades performance. Without the Semantic Reward, the agent lacks text-driven direction, yielding universally suboptimal scores. Disabling the NR Perceptual Metrics causes the most severe drop in basic image quality (e.g., the lowest MUSIQ score). Furthermore, removing the VLM Feedback deprives the model of high-level naturalness guidance, preventing top-tier perceptual restoration.

Notably, ablating the Structural Reward triggers a severe reward hacking phenomenon: the policy overfits to perceptual scores—maintaining deceptively high CLIPIQA and MUSIQ metrics—but suffers from catastrophic generative hallucinations and structural collapse. This collapse is captured by the MANIQA metric, which drops to a minimum (0.159). Ultimately, our full reward strikes the optimal balance across all perceptual and structural dimensions. **Sensitivity analyses for reward hyperparameters are provided in the supplementary material.**

5 Conclusion

We propose Dehaze-RL, a novel framework reformulating real-world dehazing as a Contextual Bandit problem. By leveraging a Physics Hybrid Action Space and GRPO with Multi-Granularity Rewards, our method overcomes the instability of adversarial training and the dependency on paired data. Extensive experiments demonstrate that Dehaze-RL achieves state-of-the-art perceptual quality and robust generalization across diverse real-world scenarios.

Limitations and Future Work. Despite its efficacy, Dehaze-RL faces two challenges. First, in extreme haze where structural information is almost entirely lost, current rewards may provide insufficient guidance. Future work will explore more comprehensive reward functions and advanced exploration strategies to handle such severe degradations. Second, the optimization remains sensitive to warmup initialization. We plan to investigate more robust paradigms, such as progressive warm-up, to further enhance the framework's stability.

6 Acknowledgment

This work was supported by the Open Fund of the MOE Key Laboratory of Equipment Intelligent Utilization under Grant 2024-AAIE-KF04-02. We acknowledge the support of GPU cluster built by MCC Lab of Information Science and Technology Institution, USTC.

References

1. Ancuti, C.O., Ancuti, C., Timofte, R., et al.: O-HAZE: a dehazing benchmark with real hazy and haze-free outdoor images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 754–762 (2018)
2. Ancuti, C., Ancuti, C.O., Timofte, R., et al.: I-HAZE: A dehazing benchmark with real hazy and haze-free indoor images. In: Advanced Concepts for Intelligent Vision Systems. pp. 620–631. Springer (2018)
3. Bai, S., Chen, K., Liu, X., et al.: Qwen2.5-VL technical report. arXiv preprint arXiv:2502.13923 (2025)
4. Chen, Z., He, Z., Lu, Z.M.: DEA-Net: Single image dehazing based on detail-enhanced convolution and content-guided attention. IEEE Transactions on Image Processing **33**, 1002–1015 (2024)
5. Engin, D., Genç, A., Kemal Ekenel, H.: Cycle-Dehaze: Enhanced CycleGAN for single image dehazing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 825–833 (2018)
6. Fang, W., Fan, J., Zheng, Y., et al.: Guided real image dehazing using YCbCr color space. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 2906–2914 (2025)
7. Fattal, R.: Dehazing using color-lines. ACM Transactions on Graphics (TOG) **34**(1), 1–14 (2014)
8. Feng, Y., Ma, L., Meng, X., et al.: Advancing real-world image dehazing: perspective, modules, and training. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(12), 9303–9320 (2024)
9. Fu, J., Liu, S., Liu, Z., et al.: Iterative predictor-critic code decoding for real-world image dehazing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12700–12709 (2025)
10. Furuta, R., Inoue, N., Yamasaki, T.: PixelRL: Fully convolutional network with reinforcement learning for image processing. IEEE Transactions on Multimedia **22**(7), 1704–1719 (2019)
11. Goodfellow, I., Pouget-Abadie, J., Mirza, M., et al.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
12. He, K., Sun, J., Tang, X.: Single image haze removal using dark channel prior. IEEE Transactions on Pattern Analysis and Machine Intelligence **33**(12), 2341–2353 (2010)
13. Hu, T., Wu, L., Dong, W., Wu, P., Sun, J., Xu, X., Yan, Q., Zhang, Y.: Boosting HDR image reconstruction via semantic knowledge transfer. IEEE Transactions on Image Processing **35**, 1910–1922 (2026)
14. Hu, T., Yan, Q., Jin, J., Dong, W., Wu, P., Qi, Y., Lin, W., Zhang, Y.: Ghost-free HDR imaging via latent low-frequency priors and deformable attention alignment. IEEE Transactions on Image Processing **35**, 2684–2698 (2026)
15. Hu, T., Yan, Q., Qi, Y., Zhang, Y.: Generating content for hdr deghosting from frequency view. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 25732–25741 (2024)
16. Ke, J., Wang, Q., Wang, Y., et al.: MUSIQ: Multi-scale image quality transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5148–5157 (2021)
17. Lan, Y., Cui, Z., Liu, C., et al.: Exploiting diffusion prior for real-world image dehazing with unpaired training. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 4455–4463 (2025)

18. Lan, Y., Cui, Z., Luo, X., Liu, C., Wang, N., Zhang, M., Su, Y., Liu, D.: When schrodinger bridge meets real-world image dehazing with unpaired training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8756–8765 (2025)
19. Li, B., Peng, X., Wang, Z., et al.: AOD-Net: All-in-one dehazing network. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4770–4778 (2017)
20. Li, B., Ren, W., Fu, D., et al.: Benchmarking single-image dehazing and beyond. IEEE Transactions on Image Processing **28**(1), 492–505 (2018)
21. Li, B., Gou, Y., Gu, S., et al.: You only look yourself: Unsupervised and untrained single image dehazing neural network. International Journal of Computer Vision **129**, 1754–1767 (2021)
22. Liu, A., Feng, B., Xue, B., et al.: DeepSeek-V3 technical report. arXiv preprint arXiv:2412.19437 (2025)
23. Liu, C., Qi, L., Pan, J., et al.: Frequency domain-based diffusion model for unpaired image dehazing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7538–7547 (2025)
24. Liu, F., Xu, J., Hu, X.: Real-world adverse weather image restoration via dual-level reinforcement learning with high-quality cold start. In: Advances in Neural Information Processing Systems. vol. 38, pp. 112525–112599 (2025)
25. Liu, J., Liu, G., Liang, J., et al.: Flow-GRPO: Training flow matching models via online rl. In: Advances in Neural Information Processing Systems. vol. 38, pp. 40783–40818 (2025)
26. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a "completely blind" image quality analyzer. IEEE Signal Processing Letters **20**(3), 209–212 (2012)
27. Nayar, S.K., Narasimhan, S.G.: Vision in bad weather. In: Proceedings of the IEEE International Conference on Computer Vision. vol. 2, pp. 820–827. IEEE (1999)
28. Park, T., Efros, A.A., Zhang, R., et al.: Contrastive learning for unpaired image-to-image translation. In: Computer Vision – ECCV. pp. 319–345. Springer (2020)
29. Qiu, Y., Zhang, K., Wang, C., et al.: MB-TaylorFormer: Multi-branch efficient transformer expanded by Taylor formula for image dehazing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12802–12813 (2023)
30. Radford, A., Kim, J.W., Hallacy, C., et al.: Learning transferable visual models from natural language supervision. In: Proceedings of the International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)
31. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI. pp. 234–241. Springer (2015)
32. Schulman, J., Wolski, F., Dhariwal, P., et al.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
33. Shao, Y., Li, L., Ren, W., et al.: Domain adaptation for image dehazing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2808–2817 (2020)
34. Shin, J., Chung, S., Yang, Y., et al.: HazeFlow: Revisit haze physical model as ODE and non-homogeneous haze generation for real-world dehazing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6263–6272 (2025)
35. Song, Y., He, Z., Qian, H., et al.: Vision transformers for single image dehazing. IEEE Transactions on Image Processing **32**, 1927–1941 (2023)

36. Wang, J., Chan, K.C., Loy, C.C.: Exploring CLIP for assessing the look and feel of images. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 2555–2563 (2023)
37. Wang, N., Cui, Z., Su, Y., et al.: Weakly supervised image dehazing via physics-based decomposition. IEEE Transactions on Circuits and Systems for Video Technology **36**(1), 637–652 (2026)
38. Wang, Y., Yan, X., Wang, F.L., et al.: UCL-Dehaze: Towards real-world image dehazing via unsupervised contrastive learning. IEEE Transactions on Image Processing **33**, 1361–1374 (2024)
39. Wu, R., Sun, L., Zhang, Z., et al.: DP^2O-SR: Direct perceptual preference optimization for real-world image super-resolution. In: Advances in Neural Information Processing Systems. vol. 38, pp. 172015–172041 (2025)
40. Xu, H., Liu, Y., Jiang, B., et al.: IRPO: Boosting image restoration via post-training GRPO. arXiv preprint arXiv:2512.00814 (2025)
41. Xu, X., Chu, R., Wang, J., et al.: Enhancing diffusion-based restoration models via difficulty-adaptive reinforcement learning with IQA reward. arXiv preprint arXiv:2511.01645 (2025)
42. Xue, Z., Wu, J., Gao, Y., et al.: DanceGRPO: Unleashing GRPO on visual generation. arXiv preprint arXiv:2505.07818 (2025)
43. Yang, S., Wu, T., Shi, S., et al.: MANIQA: Multi-dimension attention network for no-reference image quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1191–1200 (2022)
44. Yang, Y., Wang, C., Guo, X., et al.: Robust unpaired image dehazing via density and depth decomposition. International Journal of Computer Vision **132**(5), 1557–1577 (2024)
45. Yang, Y., Wang, C., Liu, R., et al.: Self-augmented unpaired image dehazing via density and depth decomposition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2037–2046 (2022)
46. Yu, K., Dong, C., Lin, L., et al.: Crafting a toolchain for image restoration by deep reinforcement learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2443–2452 (2018)
47. Zhang, H., Patel, V.M.: Densely connected pyramid dehazing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3194–3203 (2018)
48. Zhang, L., Shen, Y., Li, H.: VSI: A visual saliency-induced index for perceptual image quality assessment. IEEE Transactions on Image Processing **23**(10), 4270–4281 (2014)
49. Zhang, R., Isola, P., Efros, A.A., et al.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 586–595 (2018)
50. Zhao, S., Zhang, L., Shen, Y., et al.: RefineDNet: A weakly supervised refinement framework for single image dehazing. IEEE Transactions on Image Processing **30**, 3391–3404 (2021)
51. Zheng, Y., Zhan, J., He, S., et al.: Curricular contrastive regularization for physics-aware single image dehazing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5785–5794 (2023)
52. Zhou, B., Zhao, H., Puig, X., et al.: Semantic understanding of scenes through the ADE20K dataset. International Journal of Computer Vision **127**, 302–321 (2019)
53. Zhu, Q., Mai, J., Shao, L.: A fast single image haze removal algorithm using color attenuation prior. IEEE Transactions on Image Processing **24**(11), 3522–3533 (2015)