Degradation-Agnostic Clarity Learning for Unpaired Image Dehazing

Zhiyuan Song[1], Hannan Lu[2], Haiqian Han[3], Bo Li[4], Chang Liu[3⋆], Pengxu Wei[1,5], Xiangyang Ji[3], and Liang Lin[1,5]

[1] Sun Yat-sen University, Guangzhou, China
[2] Harbin Institute of Technology, Harbin, China
[3] Department of Automation and BNRist, Tsinghua University, Beijing, China
[4] Changan Automobile, Beijing, China
[5] Peng Cheng Laboratory, Shenzhen, China
songzhy29@mail2.sysu.edu.cn, liuchang2022@tsinghua.edu.cn

Abstract. Unpaired image dehazing models cast the restoration process as unsupervised domain translation, effectively avoiding the rigid physical assumptions of traditional methods. However, these methods frequently suffer from severe semantic hallucinations and content distortion. This stems from the significant semantic gap between unpaired clean and hazy domains, forcing the discriminator to overfit to high-level semantics rather than low-level clarity. In this paper, we propose *Degradation-Agnostic Clarity Learning (DCL)*, a novel paradigm that steers the network to learn genuine low-level clarity. We introduce two core insights: (1) Explicitly injecting low-level degradations provides strong *surface statistical shortcuts* that effectively suppress high-level semantic interference; (2) Clarity is a relative concept, and various degraded variants of a clean image share a common intrinsic *clarity direction*. To realize this, we innovatively formulate discriminator training as a *Multi-Task Learning (MTL) problem*. First, we employ Adversarial AutoAugment to dynamically mine unlearned, diverse degraded clean samples by maximizing the discriminator's loss. Second, we utilize Pareto Optimization to extract the shared common descent direction across degraded variants, thereby preventing the model from overfitting to any specific degradation type. Furthermore, we extend our core hypothesis to an Enhance-to-Resist self-supervised texture enhancement paradigm that benefits both dehazing and low-light enhancement tasks. Extensive experiments on real-world datasets demonstrate that DCL significantly suppresses semantic hallucinations and achieves state-of-the-art performance in unpaired dehazing.

Keywords: Degradation-invariant feature learning · Multi-Task Learning · Self-Supervised Image Enhancement

1 Introduction

Image dehazing aims to remove degradations caused by atmospheric haze, such as reduced contrast, distorted colors, and obscured textures, beneficial for visual

⋆ Corresponding author.

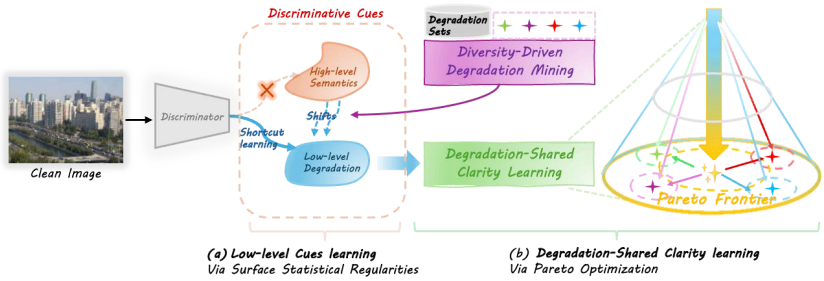


Fig. 1: Conceptual overview of our Degradation-Agnostic Clarity Learning (DCL). In the unpaired domain translation setting, we explicitly introduce low-level degradations to construct surface statistical shortcuts for the discriminator. This effectively suppresses the network's reliance on high-level semantics, thereby preventing semantic hallucinations. Simultaneously, we employ Pareto optimization to extract the shared, intrinsic clarity direction across multiple degraded variants of clear images.

enhancement as well as downstream vision tasks [57] (autonomous driving, video surveillance, etc.). However, in real-world conditions, paired hazy–clean data are rarely available, which complicates supervised training and objective evaluation. Physical-prior methods [4, 7, 8, 13, 19, 33, 35, 36, 49, 51, 52, 60] approach dehazing from the atmospheric-scattering perspective and commonly rely on synthetically generated pairs for training. However, the rigid assumptions behind these priors often fail to hold in the wild, limiting generalization and yielding color artifacts or incomplete dehazing.

Recent works [39] cast dehazing as unpaired domain translation, learning to map hazy images into the clean domain without paired supervision. By avoiding reliance on simplifying physical assumptions, these models can better accommodate diverse scattering encountered in real scenes. Despite these advances, unpaired domain translation approaches often hallucinate textures and distort scene content, especially in heavily hazy regions where structures are ambiguous. As illustrated in Fig. 2, this phenomenon fundamentally arises from the significant semantic gap between the clear and hazy domains in unpaired datasets. Driven by the domain-alignment objective, the generator fabricates false semantic content to match the discriminator's expected distribution.[6] Generally, there are two paradigms to mitigate this: manually adjusting the semantic distribution of the dataset, or steering the discriminator to focus strictly on our desired target—low-level clarity. We explore the latter. The former is cumbersome and lacks scalability, whereas the latter remains largely unexplored. We aim to provide a principled, generalizable solution for the dehazing community. However, steering the model towards low-level clarity in an unpaired setting poses two primary challenges. First is the **semantic interference**: we must force the discriminator to focus on low-level statistical shifts rather than abstract high-level semantics. Second is the

[6] Additional analysis can be found in the Appendix B.

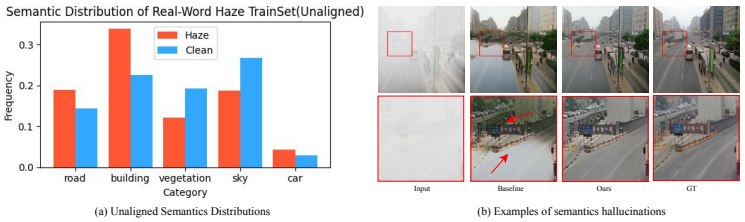


Fig. 2: The semantic gap in real-world unpaired dehazing and inherent issues of unpaired training on synthetic data. We replicate the standard unpaired translation process on our synthetic dataset, illustrating typical semantic hallucinations.

ambiguity of clarity: without paired supervision, defining a clear objective for clarity is non-trivial. To address these, we propose two fundamental hypotheses:

1. **Hypothesis 1 (Semantic Suppression via Degradation).** If we explicitly construct additional low-level differences (degradations) as surface statistical regularities [22] while weakening high-level semantic cues, the discriminator will naturally exploit these low-level statistics as shortcuts, thereby ignoring semantic interference. [11, 14, 37, 47]
2. **Hypothesis 2 (Relative Clarity).** Low-level clarity is a relative concept. Different degraded variants of a single clean image represent distinct shifts from the clean distribution. Conversely, these different degraded variants inherently share the same intrinsic information direction towards clarity.

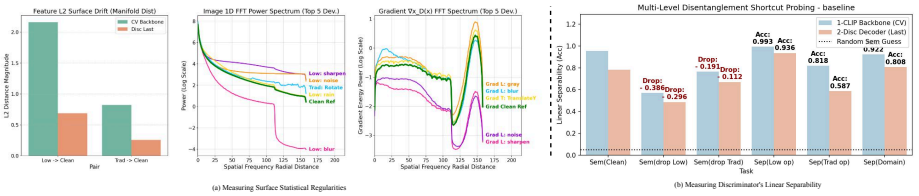


Fig. 3: Measurement of superficial statistical patterns and impact on discriminator feature learning. We compare explicit low-level degradations against traditional (non-low-level) augmentations. Low-level degradations exhibit significantly stronger surface statistical shifts, which effectively perturb and suppress the discriminator's ability to extract high-level semantic features compared to traditional augmentations while preserving the linearly separable low-level structure associated with image clarity and degradation types.

Remarkably, as illustrated in Fig. 1, we find a structural synergy: *explicitly constructing low-level degradation cues not only weakens semantic interference but also provides the variants needed to learn the shared clarity direction.* This clarity direction also serves as a key superficial statistical shortcut for the discriminator to distinguish hazy images from clear ones. As illustrated in Fig. 3, compared to traditional augmentations, our injected low-level degradations exhibit strong surface statistical biases. These biases significantly suppress the semantic features learned by the discriminator, validating Hypothesis 1. As illustrated in Fig. 4,

although different degradations are generated, they all guide the image back to clarity, preliminarily supporting Hypothesis 2. The core step, however, lies

Fig. 4: Low-level degradations induce stronger surface-statistical shortcuts while preserving image clarity than traditional augmentations. We compare different augmentation strategies applied to target-domain clear images. Traditional augmentations introduce only weak surface-statistical changes, which may still allow semantic hallucinations. In contrast, explicit low-level degradations create stronger low-level variations and encourage the discriminator to focus on degradation-related cues. Despite these injected degradations, the images largely preserve the underlying characteristics of the clear domain.

in ensuring that the discriminator learns the shared clarity information across variants rather than overfitting to a single degradation type. We innovatively formulate this objective as a Multi-Task Learning (MTL) problem, decomposed into two sub-goals: 1. **Degradation-Diversified Sample Mining**: We aim to expose the discriminator to under-fitted, non-similar degradations. Considering the dynamic nature of the decision boundary, we employ an Adversarial AutoAugment scheme [24, 56]. By combining the maximization of the discriminator's loss as a reward signal, the model actively explores diverse samples under varying degradation probabilities and intensities. 2. **Shared Clarity Extraction**: To avoid the trade-off effect caused by naively weighting different degradation losses, we utilize Pareto Optimization [43]. This ensures the model finds a common descent direction (Pareto front) among various degradations, fundamentally preventing overfitting to specific degradation semantics.

Furthermore, based on Hypothesis 2, we extend this framework to an **Enhance-to-Resist self-supervised learning paradigm for texture enhancement**. By applying high-frequency degradations (e.g., noise) to generated images, we force the model to reconstruct texture details against these perturbations. This self-supervised objective strengthens high-frequency clarity components, which is highly beneficial for both dehazing and low-light enhancement.

In summary, our main contributions are as follows:

- **A New Paradigm for Unpaired Dehazing**: We reveal the semantic hallucination issue in unpaired translation and propose Degradation-Agnostic Clarity Learning (DCL). DCL provides fresh insights into unsupervised clarity learning by exploiting the shared information among degraded clean images.
- **Methodological Innovation (DCL as MTL)**: We uniquely formulate degradation-agnostic learning as a Multi-Task Learning problem. Specifically, we employ an Adversarial AutoAugment mechanism to dynamically mine unlearned diverse degradations, and a Pareto Optimization strategy to extract the true, shared clarity direction without degradation-specific overfitting.
- **Broad Applicability via Self-Supervision**: We deduce an Enhance-to-Resist self-supervised texture enhancement paradigm based on our relative clarity hypothesis, extending the benefits of our approach beyond dehazing.
- **State-of-the-Art Performance**: Extensive experiments demonstrate that DCL significantly suppresses semantic hallucinations, achieving competitive performance in real-world image dehazing.

2 Related Work

2.1 Unpaired Image Dehazing

To address the bottleneck of paired training, recent studies have increasingly focused on unpaired training strategies to improve generalization. These methods can be broadly divided into two main categories. The first category continues to leverage physical priors, integrating them into an unpaired framework. For instance, D4 [52], D4+ [51] incorporates the atmospheric scattering model directly into its generator, and others like Diff-Dehazer [28] and DeHazeSB [29] respectively use pre-trained diffusion models or optimal transport theory, both guided by physical priors, to provide supervisory signals for dehazing. In stark contrast, the second category, exemplified by methods like CycleGAN-Turbo [39], relies purely on powerful generative capabilities to translate between domains. While these approaches are not constrained by the fragile assumptions of physical priors and thus offer greater potential for generalization, the direction of their learning process is often problematic. Lacking explicit guidance, they are prone to producing semantic hallucinations, particularly on datasets with significant distributional mismatch, where textures from the target domain are incorrectly painted onto the source image.

2.2 Unpaired Image-to-image Translation

Unsupervised image-to-image translation has made significant progress through techniques such as cycle consistency [59], shared latent space modeling [30], content-preserving losses [44], and contrastive learning [38]. Recent efforts have also explored diffusion-based models for unpaired domain transfer [39, 41]. However, translating across domains with mismatched semantic distributions often

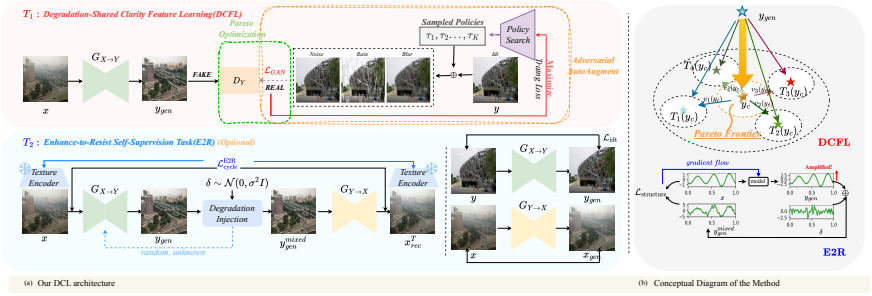


Fig. 5: Overview of our DCL. As illustrated in (a), we optimize DCFL(T_1) to learn degradation-shared clean features and E2R(T_2) to obtain degradation-resistant enhanced texture features, and The haze synthesis translation follows the baseline [39] and is omitted for brevity. At inference, dehazing is performed using the generator $G_{X\to Y}$ only. (b) presents key ideas underlying the two core components of our method.

leads to semantic flipping [18, 21]—a known limitation of GAN-based frameworks that align global statistics while ignoring local structure. Several approaches attempt to address this by focusing on salient regions [45], improving geometric consistency [12], or introducing contrastive or mutual information constraints [18, 23]. Nonetheless, most methods still struggle under severe degradation (e.g., haze), where input features are corrupted or missing. This limits their ability to achieve semantically stable and detail-preserving translations in real-world applications.

3 Methodology

3.1 Overview

Our Degradation-Agnostic Clarity Learning (DCL) framework follows an unpaired image-to-image translation paradigm [39]. We use two generators, $G_{X\to Y}$ and $G_{Y\to X}$, and two discriminators, D_Y and D_X. The generator $G_{X\to Y}$ is the target dehazing network that maps hazy images from domain X to clean images in domain Y, while $G_{Y\to X}$ maps clean images back to the hazy domain for cycle-consistent training. However, standard adversarial learning may encourage the generator to bridge the domain gap by hallucinating high-level semantic content. To alleviate this issue, DCL guides D_Y to focus on low-level clarity cues rather than semantic shortcuts. As shown in Fig. 5, DCL contains two main components. First, **Degradation-Shared Clarity Feature Learning** formulates degradation-guided adversarial learning as a multi-task optimization problem, where Adversarial Reinforcement Learning (ARL) mines informative degradation variants and Pareto Optimization extracts their shared clarity direction. Second, **Enhance-to-Resist (E2R) Self-Supervised Texture Enhancement** imposes a degradation-resistant feature reconstruction constraint to improve texture structure without paired supervision.

3.2 Degradation-Shared Clarity Feature Learning

Motivation. According to Hypothesis 2, different degraded variants of a clean image $y_c \in \mathcal{Y}$ share an intrinsic clarity component. Directly training D_Y with degraded images, however, may cause overfitting to degradation-specific artifacts, such as particular noise patterns or blur kernels, as shown in Fig. 4. We therefore cast clarity learning as a multi-task learning problem, where each degradation type defines a distinct task and the objective is to extract their shared clarity direction.

Diversity-Driven Degradation Mining via Adversarial AutoAugment
Motivation. The decision boundary of D_Y changes during training. A static degradation set can be quickly fitted, reducing its ability to provide effective low-level cues. Inspired by Adversarial AutoAugment [56], we dynamically mine hard and under-learned degradation variants to keep challenging D_Y and encourage cancellation among degradation-specific signals.
Formulation. We introduce a policy network $\pi(\tau|\theta_\pi)$ to sample a degradation policy τ from a degradation space $\mathcal{T}$, including degradation type, probability, and magnitude. For a clean image y_c, the degraded variant is denoted as $T_\tau(y_c)$. The policy is trained to maximize the discriminator loss on the degraded sample:

$$R(\tau) = \mathcal{L}_{D_Y}(T_\tau(y_c); \theta_{D_Y}). \tag{1}$$

The policy network is optimized to maximize the expected reward:

$$\max_{\theta_\pi} \mathbb{E}_{\tau\sim\pi(\cdot|\theta_\pi)}[R(\tau)]. \tag{2}$$

This adversarial mining process produces a set of informative degradation variants $T_1, T_2, \ldots, T_K$, preventing D_Y from overfitting to a fixed set of artifacts.

Degradation-Shared Clarity Learning via Pareto Optimization Motivation. Given K mined degradation tasks, a simple weighted sum of losses may bias training toward dominant degradation types. To extract degradation-agnostic clarity information, we formulate the update of D_Y as a multi-objective optimization problem.

Formulation. Let $\mathcal{L}_k(\theta_{D_Y}) = \mathcal{L}_{D_Y}(T_k(y_c); \theta_{D_Y})$ be the loss associated with the k-th degradation task. Our goal is to find a Pareto stationary point where no individual task's loss can be decreased without increasing another. We employ the Multiple Gradient Descent Algorithm (MGDA) [43] to find a common descent direction. Specifically, we solve for a set of optimal task weights $\alpha = [\alpha_1, \ldots, \alpha_K]^T$ in the unit simplex Δ^{K-1} that minimizes the norm of the combined gradient:

$$\min_{\alpha\in\Delta^{K-1}} \left\| \sum_{k=1}^{K} \alpha_k \nabla_{\theta_{D_Y}} \mathcal{L}_k(\theta_{D_Y}) \right\|_2^2. \tag{3}$$

The resulting gradient update $\Delta\theta_{D_Y} = \sum_{k=1}^{K} \alpha_k \nabla_{\theta_{D_Y}} \mathcal{L}_k(\theta_{D_Y})$ represents the shared trajectory towards intrinsic clarity, strictly avoiding any degradation-specific biases.

Theoretical Analysis We provide a formal justification for why ARL and Pareto Optimization help extract degradation-agnostic clarity information.

Assumptions. Let the discriminator's feature extractor be $\phi(\cdot;\theta_\phi) : \mathbb{R}^{H\times W\times C} \to \mathbb{R}^d$. The full discriminator is $D(y) = h(\phi(y))$, with parameters $\theta_{D_Y} = \{\theta_\phi, \theta_h\}$. We posit the following:

Assumption 1 (Local Approximate Feature Decomposition) *In the local neighborhood induced by low-level degradation operators around a clean image y_c, the discriminator feature admits an approximate shared-private decomposition:*

$$\phi(T_k(y_c)) = \phi_c(y_c) + \mathbf{v}_k(y_c) + \mathbf{r}_k(y_c), \tag{4}$$

where $\phi_c(y_c)$ denotes the degradation-shared clarity component, $\mathbf{v}_k(y_c)$ denotes the degradation-specific component introduced by the k-th degradation T_k, and $\mathbf{r}_k(y_c)$ is a small approximation residual. For the pristine clean image, i.e., T_0 being the identity transformation, we have $\mathbf{v}_0(y_c) \approx \mathbf{0}$ and $\phi(y_c) \approx \phi_c(y_c)+\mathbf{r}_0(y_c)$. We further assume that the shared clarity component is approximately invariant under the considered low-level degradations, i.e.,

$$\phi_c(T_k(y_c)) \approx \phi_c(y_c), \tag{5}$$

This is consistent with the local linearity of piecewise-linear neural networks [5] and shared representations in multi-task feature learning [2].

Assumption 2 (Zero-Mean degradation Features) *The set of degradations $\mathcal{T}$ is sufficiently diverse such that the corresponding degradation feature vectors have no systematic directional bias. Formally, their expectation is the zero vector:*

$$\mathbb{E}_{T_i\sim\mathcal{T}}[\mathbf{v}_i(y_c)] = \lim_{K\to\infty} \frac{1}{K}\sum_{i=1}^{K} \mathbf{v}_i(y_c) \approx \mathbf{0}. \tag{6}$$

Without sufficient diversity, training may collapse to a few dominant degradations and violate Eq. 6. ARL encourages exploration of hard degradation variants, making the empirical degradation distribution more diverse. According to Assumption 1, the gradient of the k-th task can be decomposed into $\nabla\mathcal{L}_k \approx \nabla_{\theta_{D_Y}}\ell(\phi_c) + \nabla_{\theta_{D_Y}}\ell(\mathbf{v}_k)$. MGDA then finds weights that reduce conflicts among task gradients. Since degradation-specific components tend to cancel under Assumption 2, the final update is dominated by the shared clarity component $\nabla\theta_{D_Y}\ell(\phi_c)$.

3.3 Enhance-to-Resist Self-Supervised Texture Enhancement

Motivation. Unpaired dehazing models often struggle to recover fine textures. Inspired by Hypothesis 2, we extend it and further use degradation as a self-supervised signal for texture enhancement. The key idea is to perturb the generated clean image with high-frequency degradation and require its texture

representation to remain reconstructable. This encourages the generator to produce stronger, input-supported texture details. A formal theoretical analysis can be found in Appendix B.2.

Formulation. We introduce a pre-trained texture encoder E, our generator G, and a random high-frequency degradation operator T_{hf} drawn from a specific degradation space $\mathcal{T}_{HF}$. For a given hazy input $x \sim p_{\text{haze}}$, the generator produces an enhanced image $y = G(x)$.

Our core idea is to train G such that its output, even after being heavily corrupted by an unknown high-frequency degradation T_{hf}, can still be mapped back to the intrinsic texture features of the original hazy image $E(x)$. This forms an adversarial-like feature consistency constraint, formulated as the Enhance-to-Resist ($\mathcal{L}_{\text{E2R}}$) loss:

$$\mathcal{L}_{\text{E2R}} = \mathbb{E}_{x \sim p_{\text{haze}}, T_{hf} \sim \mathcal{T}_{HF}} \left[\left\| E(T_{hf}(G(x))) - E(x) \right\|_2^2 \right]. \tag{7}$$

Minimizing Eq. 7 compels the generator $G(x)$ to embed robust, highly distinguishable texture details that can survive the destructive transformation T_{hf}, thereby achieving self-supervised detail enhancement without requiring clean reference pairs. This strategy proves highly effective not only for dehazing but also for generalized low-light enhancement.

Integration into the Cycle Framework We embed our Enhancement-to-Resist (E2R) strategy into a CycleGAN-Turbo backbone [39], elevating the conventional pixel-level cycle-consistency to a structure-aware, latent-space objective. The key innovation is a self-supervised learning task in which the generator is challenged to produce outputs robust to unforeseen degradation.

Specifically, we leverage a pre-trained Variational Autoencoder (VAE [42]), E, as a fixed texture feature extractor. The VAE is chosen for its proven ability to learn a meaningful latent space that captures the essential factors of variation in texture and structure. During training, a stochastic degradation operator, T, is applied to the generator's output $y_{\text{gen}} = G_{X \to Y}(x)$ prior to the backward mapping. The final reconstruction is given by $x_{\text{rec}}^T = G_{Y \to X}(T(G_{X \to Y}(x)))$. We then formulate an E2R loss in the VAE feature space:

$$\mathcal{L}_{cycle}^{\text{E2R}} = \mathbb{E}_{x \sim p_{\text{haze}}, T \sim p(\mathcal{T})} \left[\left\| E(x) - E(x_{rec}^T) \right\|_2^2 \right]$$

Minimizing this objective elegantly transforms the cycle-consistency loop into a powerful, self-supervised driver for targeted texture enhancement.

To preserve color and stabilize training, we retain the identity loss:

$$\mathcal{L}_{\text{idt}} = \mathbb{E}_{x \sim p_{\text{haze}}} \| G_{Y \to X}(x) - x \| + \mathbb{E}_{y \sim p_{\text{clean}}} \| G_{X \to Y}(y) - y \|. \tag{8}$$

4 Experiments

4.1 Datasets and Evaluation metrics

Training Datasets. We follow existing works [51, 58] and select the widely used unpaired real-world dehazing dataset, RESIDE-Unpaired dataset [58], which

includes 3577 real outdoor clean images and 2903 hazy images with higher quality from the RESIDE, as our training dataset.
Testing Datasets. RTTS dataset [31], which contains over $4,000$ real hazy images with diverse scenes, resolutions, and degradation issues. Additionally, we further conduct comparisons on Fattal's dataset [8], Foggy Driving [21].
Metrics. To comprehensively evaluate dehazing effectiveness, we combine no-reference measures. HazDesNet [16, 54] is used to estimate haze removal extent, while CLIPIQA [46], MUSIQ [25] and MANIQA [50] provide perceptual quality assessment.

4.2 Implementation Details

Our framework is trained on one NVIDIA RTX 4090 GPU. To augment the training data, we directly resize images with a size of 256×256 and apply random rotations of 90, 180, 270 degrees, as well as the horizontal and vertical flip. For all experiments, we use the AdamW optimizer with a learning rate of 1e-5, a batch size of 1 and train the model for $30k$ steps. Traditional augmentations used for qualitative comparision (excluding low-level degradations) consist of TranslateX, Cutout, Rotate, Color, and AutoContrast, covering geometric transformations, pixel distribution shifts, color perception, and occlusion-induced information loss. The adversarial autoaugmentation sampling pool includes identity, noise, rain, blur, sharpen, gray, dark, jpeg and candidate batch size is 3. And the loss hyperparameters for our full mehthod are set as $\lambda_{cycle} = 1$ for $\mathcal{L}1$, $\lambda_{cycle} = 10$ for $\mathcal{L}_{cycle}^{E2R}$. $\lambda_{idt} = 1$, $\lambda_{GAN} = 0.5$. Efficiency Analysis can be found in Supplementary.

4.3 Comparison with State-of-the-Art Methods

We evaluated our method against SOTA approaches on real-world datasets (RTTS, Fattal, FoggyDriving). To ensure fairness, we retrained unpaired methods [28, 29, 39, 51] on a common dataset [58] and used official pre-trained models for paired methods [10, 17]. Table 1, 2 and Figure 6 shows that our method significantly outperforms existing works. We attribute this to the distinct failure modes of prior approaches: Physical-prior methods [10, 17, 29, 51] break down in real scenes because of their brittle assumptions, whereas GAN-based methods tend to hallucinate semantics [39]. Diff-Dehazer [28], as a hybrid of both, ultimately provides only a compromise, inheriting limitations from each side. Our framework directly addresses these issues. The degradation-guided discriminator prevents semantic distortion by focusing on low-level statistics, and our enhancement strategy robustly restores fine-grained textures. Consequently, our method achieves marked improvements across both quantitative metrics and visual quality. More results can be found in Appendix D.

4.4 Ablation Study

We first ablate the overall architecture to verify its effectiveness. We then delve into each component, uncovering several intriguing properties and confirming

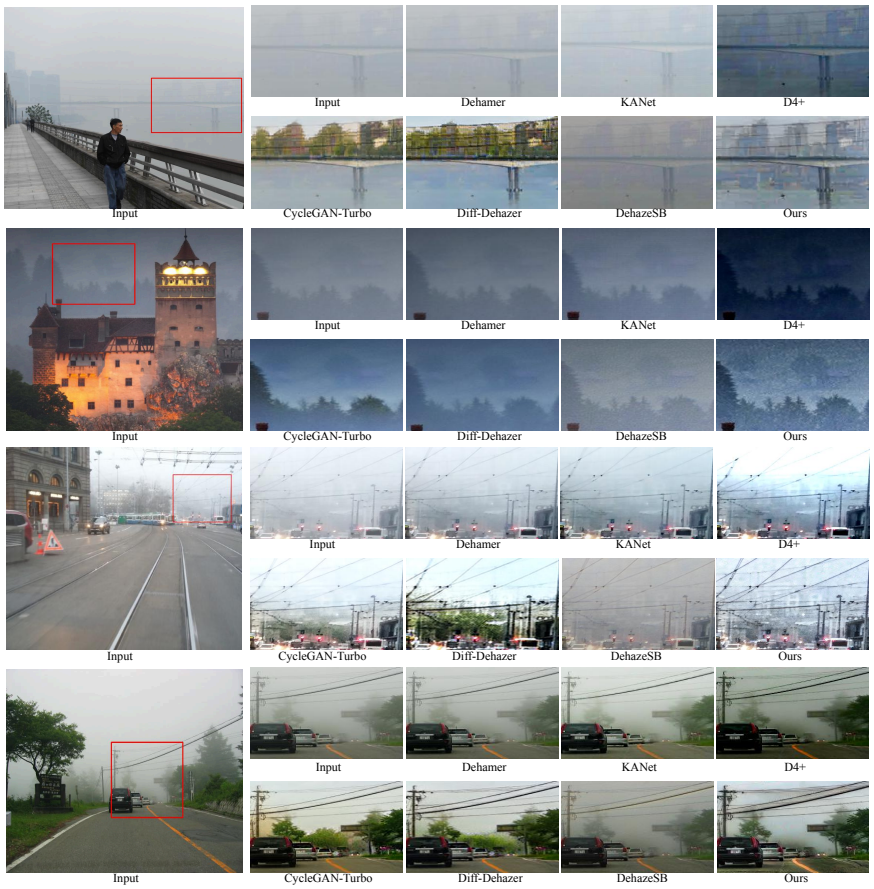


Fig. 6: Qualitative comparisons on real-world dataset from RTTS (first row), Fattal's dataset (second row), and FoggyDriving dataset (last two rows). **Zoom-in for best view.**

their significance. To isolate the effect of each component, every ablation is trained with only its dedicated loss. Specifically, GAN loss is excluded when studying E2R, and E2R-related losses are excluded when studying DCFL.

Effectiveness of Major Components in DCL We first ablate the core components of DCL. As shown in Tab. 3 and Fig. 7, traditional augmentations introduce only weak surface-statistical changes, leaving the discriminator prone to high-level semantic cues and thus semantic hallucinations. In contrast, explicit low-level degradations provide stronger degradation-related cues, which help suppress semantic interference. However, using low-level degradations alone may overfit to specific degradation artifacts. Simply summing losses from diverse degradations also fails to isolate their shared clarity direction, allowing certain degradation patterns to dominate training. By incorporating Pareto Optimization, the full DCL extracts a common descent direction across degradations, effectively

Table 1: Quantitative comparison on real-world image dehazing datasets: RTTS and Fattal's data.

	RTTS				Fattal's data			
method	HazDesNet↓	CLIPIQA↑	MUSIQ↑	MANIQA↑	HazDesNet↓	CLIPIQA↑	MUSIQ↑	MANIQA↑
Dehamer [17]	0.319	0.347	50.55	0.279	0.124	0.475	61.43	0.348
KANet [10]	0.372	0.28	54.53	0.257	0.146	0.53	65.11	0.351
D4+ [51]	0.333	0.305	53.29	**0.289**	0.115	0.546	63.2	0.38
CycleGAN-Turbo [39]	0.372	0.289	56.63	0.286	0.163	0.53	66.65	0.36
Diff-Dehazer [28]	0.321	0.356	**57.8**	0.271	0.154	0.467	61.66	0.311
DehazeSB [29]	0.438	**0.377**	53.09	0.263	0.253	0.527	63.74	0.33
Ours	**0.278**	0.325	57.47	0.278	**0.098**	**0.587**	**68.83**	**0.438**

Table 2: Quantitative comparison on real-world image dehazing datasets: FoggyDriving subsets.

	FoggyDriving (Person_Public)				FoggyDriving (Web)			
method	HazDesNet↓	CLIPIQA↑	MUSIQ↑	MANIQA↑	HazDesNet↓	CLIPIQA↑	MUSIQ↑	MANIQA↑
Dehamer [17]	0.257	0.299	54.17	0.3	0.355	0.393	51.93	0.311
KANet [10]	0.329	0.32	54.6	0.273	0.372	0.357	54.05	0.276
D4+ [51]	0.257	0.332	53.23	0.283	0.292	0.35	51.58	0.286
CycleGAN-Turbo [39]	0.344	0.306	56.06	0.298	0.387	0.367	58	0.307
Diff-Dehazer [28]	0.346	**0.46**	44.6	0.262	0.348	0.44	56.46	0.308
DehazeSB [29]	0.371	0.343	52.52	0.282	0.46	0.415	51.77	0.286
Ours	**0.193**	0.408	**59.42**	**0.322**	**0.24**	**0.46**	**60.39**	**0.312**

suppressing semantic hallucinations while recovering intrinsic low-level clarity. In addition, the E2R module further improves texture structures.

Effectiveness of Sampling Size on Pareto Optimization. We study the effect of the sampled pool size K, i.e., the number of degradation variants mined by ARL for Pareto Optimization. Repeated degradation types are rejected and re-sampled to encourage diversity. As shown in Fig. 8, when $K = 1$, Pareto Optimization degenerates to a single-gradient update and lacks cross-degradation constraints, leading to degradation-specific over-enhancement. With $K = 3$, three diverse adversarially mined gradients provide sufficient mutual constraints, enabling variant-specific artifacts to be suppressed while enhancing the shared clarity direction. Further increasing the pool to $K = 5$ introduces redundant or easy degradations that have already been well fitted by the discriminator, which disturbs the balanced Pareto direction and slightly increases artifacts. We therefore set $K = 3$ by default to balance degradation diversity and gradient mutual cancellation.

Effectiveness of texture fidelity in DCL We evaluate the texture fidelity and dehazing ability of DCL on unpaired heavy-haze synthetic datasets and further validate semantic preservation on FoggyDriving. We compare DCL with state-of-the-art unpaired dehazing methods [28, 29, 34, 39, 48] and semantic-preserving image translation methods [18, 21]. Following common practice, we use a pre-trained DeepLabV3 [3] model to compute segmentation maps and report pxAcc,

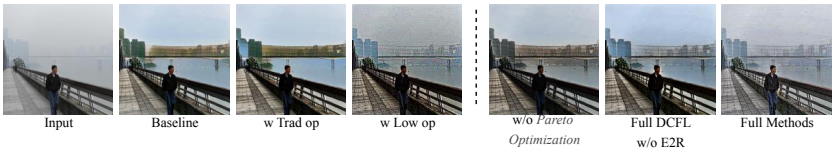


Fig. 7: Ablation results of DCL.

Table 3: Ablation study on the RTTS dataset of DCL.

ARL	PO	E2R	HazDesNet↓	CLIPIQA↑	MUSIQ↑	MANIQA↑
✗	✗	✗	0.372	0.289	<u>56.63</u>	**0.286**
✓	✗	✗	0.282	0.292	53.38	0.272
✓	✓	✗	**0.264**	<u>0.299</u>	54.67	0.243
✓	✓	✓	<u>0.278</u>	**0.325**	**57.47**	<u>0.278</u>

Fig. 8: The effect of different sampled pool sizes in our DCFL.

clsAcc, and mIoU. As shown in Tabs. 4 and 5, DCL achieves better PSNR and segmentation performance than prior-based methods [28, 29, 48], frequency-based methods [34], and pixel- or semantic-constrained translation methods [18, 21]. This improvement mainly comes from suppressing semantic cues and avoiding over-reliance on source-domain appearance or semantic constraints, which may incorrectly preserve haze as content under severe degradation. More visualizations are provided in Appendix E.1.

How Does DCL Remove Degradation-Specific Components? We analyze whether DCL suppresses degradation-specific information. We define degradation components as the feature differences between degraded-clean images and their clean counterparts, and remove the cross-degradation mean to reduce shared clarity leakage. Fig. 9 shows that different degradation types are widely distributed in PCA space and exhibit low or negative pairwise cosine similarities, confirming the diversity and conflict among degradation-specific directions and supporting Assumption 2. We further compare MGDA with direct gradient summation in Fig. 10. MGDA produces a much lower projection ratio onto degradation-specific gradient components, indicating that it adaptively suppresses conflicting degradation cues and preserves degradation-agnostic information.

Analysis of the Learned Clarity Direction. We further examine whether the learned discriminator direction is truly clarity-related. We construct controlled image sets with varying haze intensity, semantic category [9], and degradation type, and compare the loss sensitivity of different discriminators with an Oracle trained on paired synthetic dehazing data. As shown in Fig. 11, DCL maintains haze-level sensitivity close to the Oracle, whereas the low-level-only variant is affected by non-haze degradation cues. Fig. 12 further shows that DCL is less sensitive to semantic and degradation variations than the baseline and low-level-only discriminator. These results suggest that DCL learns a degradation-agnostic clarity direction instead of relying on semantic or degradation-specific shortcuts.

Table 4: Quantitative comparisons of dehazing performance and texture fidelity on our constructed unpaired synthetic dataset.

Method	PSNR↑	SSIM↑	pxAcc↑	clsAcc↑	mIoU↑
Hazy image	8.11	0.51	67.3	35.6	26.7
Baseline [39]	13.76	0.71	62.0	38.9	25.7
DehazeSB [29]	10.67	0.555	63.0	30.0	22.1
Baseline+SCC [18]	11.9	0.679	61.8	36.2	23.3
Baseline+SRUNIT [21]	13.96	0.721	67.0	40.3	28.8
Ours	14.51	0.746	67.8	41.9	29.5
Baseline(Aligned/Oracle)	**16.33**	**0.767**	**70.6**	**45.8**	**32.3**

Table 5: Performance comparison on semantic segmentation on Foggy-Driving.

Method	pxAcc↑	clsAcc↑	mIoU↑
Baseline [39]	84.4	47.2	34.9
Diff-Dehazer [28]	83.6	46.3	33.3
DehazeSB [29]	81.2	47.1	34.0
FrDiff [34]	52.9	18.3	12.3
ZRID-Net [48]	81.2	42.7	31.4
Ours	**85.5**	**50.9**	**37.7**

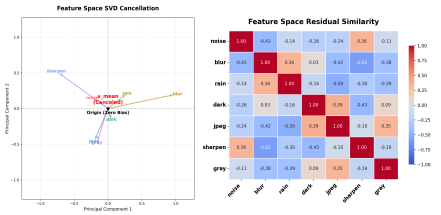


Fig. 9: Diversity and conflict of degradation-specific components.

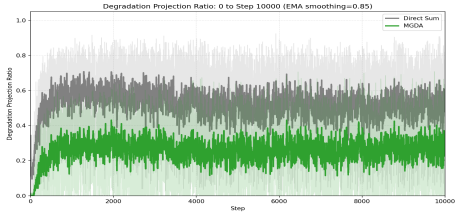


Fig. 10: Suppression of degradation-specific optimization.

When is E2R Effective for Texture Enhancement? E2R is effective when haze weakens but does not completely erase texture cues. In moderately degraded regions, its degradation-based reconstruction objective encourages the recovery of input-supported high-frequency structures, improving both dehazing quality and texture fidelity, as shown in Tab. 6. In severely degraded RTTS samples with strong sensor noise or JPEG artifacts, E2R may also preserve some pre-existing artifacts while enhancing frequency cues. Thus, E2R is best viewed as a texture fidelity enhancer for recoverable structures.

Analysis of Different Degradation Strength and Different Texture Encoder. We study how degradation strength affects texture reconstruction, using noise and different texture encoders(VGG [55], VAE [42]). As noise increases, we observe that enhancement first improves and then declines (Figure 13a). Too little noise offers weak supervision, while excessive noise overwhelms the signal, which prevents the model from extracting any meaningful signal for enhancement. While all enable enhanced texture robustness, VAE encoding with strong texture diversity delivers superior results than others under same degradation strength, highlighting the importance of rich texture representations for effective enhancement.

Table 6: E2R ablation on FoggyDriving(Person_Public33).

Model	HazDesNet↓	CLIPIQA↑	MUSIQ↑	MANIQA↑
Ours w/o E2R	0.2449	0.244	49.62	0.247
Ours w/ E2R	**0.1498**	**0.325**	**53.69**	**0.291**

Generalization to Low-Light Image Enhancement. Given its universal modeling and minimal assumptions, we explore its capability boundaries by applying it to the task of low-light image enhancement. Our algorithm substan-

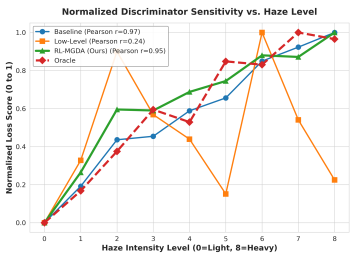


Fig. 11: Discriminator Sensitivity to Haze Intensity.

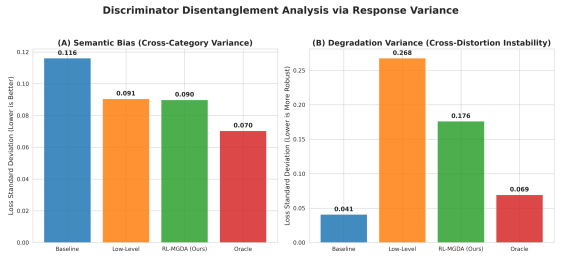
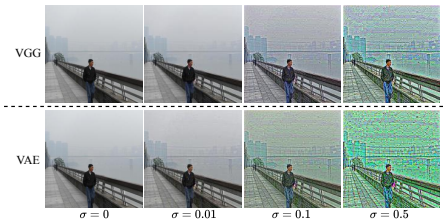
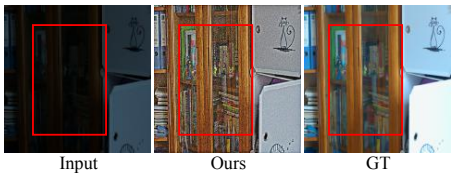


Fig. 12: Bias Sensitivity to Semantic and Degradation Variations.



(a) Visualizing the impact of degradation strength and Texture Encoder on our E2R.



(b) Generalization of E2R to Low-Light Enhancement.

Fig. 13: The Broad Applicability of E2R self-supervised texture enhancement paradigm.

tially improves image quality and adaptively enhances all texture and structure, regardless of initial brightness (Figure 13).

5 Discussion

Conclusion. We presented Degradation-Agnostic Clarity Learning (DCL) to address semantic hallucinations in unpaired image dehazing. DCL exploits explicit low-level degradations to weaken semantic cues and recover a shared clarity direction across diverse degraded variants. By casting this process as multi-task optimization, it combines adversarial degradation mining with Pareto optimization to extract clarity-oriented direction, and further extends to self-supervised texture enhancement. Experiments on real-world dehazing and low-light enhancement validate the effectiveness and generalization of our approach.

6 Acknowledgment

We thank Haowei Li for helpful discussions and valuable feedback on this work. This work is supported in part by Guangdong Provincial General Fund No. 2024A1515010208, Guangzhou Science and Technology Program Project No.2025A04J5465, National Natural Science Foundation of China (NSFC) under Grant No.62376292 and No.62406167.

References

1. Ancuti, C.O., Ancuti, C., Timofte, R., De Vleeschouwer, C.: 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. Argyriou, A., Evgeniou, T., Pontil, M.: Convex multi-task feature learning. Machine learning **73**(3), 243–272 (2008)
3. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
4. Chen, Z., Wang, Y., Yang, Y., Liu, D.: Psd: Principled synthetic-to-real dehazing guided by physical priors. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7180–7189 (2021)
5. Chu, L., Hu, X., Hu, J., Wang, L., Pei, J.: Exact and consistent interpretation for piecewise linear neural networks: A closed form solution. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining. pp. 1244–1253 (2018)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
7. Fang, C., He, C., Xiao, F., Zhang, Y., Tang, L., Zhang, Y., Li, K., Li, X.: Real-world image dehazing with coherence-based label generator and cooperative unfolding network. arXiv preprint arXiv:2406.07966 (2024)
8. Fattal, R.: Dehazing using color-lines. ACM transactions on graphics (TOG) **34**(1), 1–14 (2014)
9. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In: 2004 conference on computer vision and pattern recognition workshop. pp. 178–178. IEEE (2004)
10. Feng, Y., Ma, L., Meng, X., Zhou, F., Liu, R., Su, Z.: Advancing real-world image dehazing: Perspective, modules, and training. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(12), 9303–9320 (2024)
11. Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., Holz, T.: Leveraging frequency analysis for deep fake image recognition. In: International conference on machine learning. pp. 3247–3258. PMLR (2020)
12. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Zhang, K., Tao, D.: Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2427–2436 (2019)
13. Fu, J., Liu, S., Liu, Z., Guo, C.L., Park, H., Wu, R., Wang, G., Li, C.: Iterative predictor-critic code decoding for real-world image dehazing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
14. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In: International conference on learning representations (2018)
15. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)

16. Guan, T., Li, C., Zheng, Y., Wu, X., Bovik, A.C.: Dual-stream complex-valued convolutional network for authentic dehazed image quality assessment. IEEE Transactions on Image Processing **33**, 466–478 (2023)
17. Guo, C.L., Yan, Q., Anwar, S., Cong, R., Ren, W., Li, C.: Image dehazing transformer with transmission-aware 3d position embedding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5812–5820 (2022)
18. Guo, J., Li, J., Fu, H., Gong, M., Zhang, K., Tao, D.: Alleviating semantics distortion in unsupervised low-level image-to-image translation via structure consistency constraint. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18249–18259 (2022)
19. 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)
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
21. Jia, Z., Yuan, B., Wang, K., Wu, H., Clifford, D., Yuan, Z., Su, H.: Semantically robust unpaired image translation for data with unmatched semantics statistics. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 14273–14283 (2021)
22. Jo, J., Bengio, Y.: Measuring the tendency of cnns to learn surface statistical regularities. arXiv preprint arXiv:1711.11561 (2017)
23. Jung, C., Kwon, G., Ye, J.C.: Exploring patch-wise semantic relation for contrastive learning in image-to-image translation tasks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18260–18269 (2022)
24. Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T.: Training generative adversarial networks with limited data. In: Proc. NeurIPS (2020)
25. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5148–5157 (2021)
26. Kong, X., Dong, C., Zhang, L.: Towards effective multiple-in-one image restoration: A sequential and prompt learning strategy. arXiv preprint arXiv:2401.03379 (2024)
27. Kumari, N., Zhang, R., Shechtman, E., Zhu, J.Y.: Ensembling off-the-shelf models for gan training. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10651–10662 (2022)
28. Lan, Y., Cui, Z., Liu, C., Peng, J., Wang, N., Luo, X., Liu, D.: 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)
29. 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)
30. Lee, H.Y., Tseng, H.Y., Huang, J.B., Singh, M., Yang, M.H.: Diverse image-to-image translation via disentangled representations. In: Proceedings of the European conference on computer vision (ECCV). pp. 35–51 (2018)
31. Li, B., Ren, W., Fu, D., Tao, D., Feng, D., Zeng, W., Wang, Z.: Benchmarking single-image dehazing and beyond. IEEE Transactions on Image Processing **28**(1), 492–505 (2018)

32. Li, H., Wei, P., Zhang, D., Lin, L.: Learning heterogeneous degradation representation for real-world super-resolution. In: The Fourteenth International Conference on Learning Representations
33. Li, L., Dong, Y., Ren, W., Pan, J., Gao, C., Sang, N., Yang, M.H.: Semi-supervised image dehazing. IEEE Transactions on Image Processing **29**, 2766–2779 (2019)
34. Liu, C., Qi, L., Pan, J., Qian, X., Yang, M.H.: Frequency domain-based diffusion model for unpaired image dehazing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2025)
35. Liu, W., Hou, X., Duan, J., Qiu, G.: End-to-end single image fog removal using enhanced cycle consistent adversarial networks. IEEE Transactions on Image Processing **29**, 7819–7833 (2020)
36. McCartney, E.J.: Optics of the atmosphere: scattering by molecules and particles. New York (1976)
37. Mescheder, L., Geiger, A., Nowozin, S.: Which training methods for gans do actually converge? In: International conference on machine learning. pp. 3481–3490. PMLR (2018)
38. Park, T., Efros, A.A., Zhang, R., Zhu, J.Y.: Contrastive learning for unpaired image-to-image translation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16. pp. 319–345. Springer (2020)
39. Parmar, G., Park, T., Narasimhan, S., Zhu, J.Y.: One-step image translation with text-to-image models. arXiv preprint arXiv:2403.12036 (2024)
40. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
41. Sasaki, H., Willcocks, C.G., Breckon, T.P.: Unit-ddpm: Unpaired image translation with denoising diffusion probabilistic models. arXiv preprint arXiv:2104.05358 (2021)
42. Sauer, A., Lorenz, D., Blattmann, A., Rombach, R.: Adversarial diffusion distillation. In: European Conference on Computer Vision. pp. 87–103. Springer (2024)
43. Sener, O., Koltun, V.: Multi-task learning as multi-objective optimization. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 31, pp. 525–536. Curran Associates, Inc. (2018), http://papers.nips.cc/paper/7334-multi-task-learning-as-multi-objective-optimization.pdf
44. Taigman, Y., Polyak, A., Wolf, L.: Unsupervised cross-domain image generation. arXiv preprint arXiv:1611.02200 (2016)
45. Tang, H., Liu, H., Xu, D., Torr, P.H., Sebe, N.: Attentiongan: Unpaired image-to-image translation using attention-guided generative adversarial networks. IEEE transactions on neural networks and learning systems **34**(4), 1972–1987 (2021)
46. Wang, J., Chan, K.C., Loy, C.C.: Exploring clip for assessing the look and feel of images. In: AAAI (2023)
47. Wang, S.Y., Wang, O., Zhang, R., Owens, A., Efros, A.A.: Cnn-generated images are surprisingly easy to spot... for now. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8695–8704 (2020)
48. Wang, S., Ren, W., Gao, P., Yu, J., Liu, J.: Zrid-net: Zero-reference real-world image dehazing framework via deep self-decoupling and reverse knowledge transfer. IEEE Transactions on Circuits and Systems for Video Technology **36**(2), 2034–2052 (2026). https://doi.org/10.1109/TCSVT.2025.3609735

49. Wu, R.Q., Duan, Z.P., Guo, C.L., Chai, Z., Li, C.: Ridcp: Revitalizing real image dehazing via high-quality codebook priors. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22282–22291 (2023)
50. Yang, S., Wu, T., Shi, S., Lao, S., Gong, Y., Cao, M., Wang, J., Yang, Y.: 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)
51. Yang, Y., Wang, C., Guo, X., Tao, D.: Robust unpaired image dehazing via density and depth decomposition. International Journal of Computer Vision **132**(5), 1557–1577 (2024)
52. Yang, Y., Wang, C., Liu, R., Zhang, L., Guo, X., Tao, D.: 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)
53. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13. pp. 818–833. Springer (2014)
54. Zhang, J., Min, X., Zhu, Y., Zhai, G., Zhou, J., Yang, X., Zhang, W.: Hazdesnet: An end-to-end network for haze density prediction. IEEE transactions on intelligent transportation systems **23**(4), 3087–3102 (2020)
55. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: 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)
56. Zhang, X., Wang, Q., Zhang, J., Zhong, Z.: Adversarial autoaugment. arXiv preprint arXiv:1912.11188 (2019)
57. Zhao, Q., Deng, W., Wei, P., Dong, Z., Lu, H., Ji, X., Lin, L.: Delving into cascaded instability: a lipschitz continuity view on image restoration and object detection synergy. Advances in Neural Information Processing Systems **38**, 164856–164881 (2026)
58. Zhao, S., Zhang, L., Shen, Y., Zhou, Y.: Refinednet: A weakly supervised refinement framework for single image dehazing. IEEE Transactions on Image Processing **30**, 3391–3404 (2021)
59. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)
60. Zhu, Q., Mai, J., Shao, L.: Single image dehazing using color attenuation prior. In: BMVC. vol. 4, pp. 1674–1682. Citeseer (2014)