
HCD-YC: Hierarchical Cycle-consistent Hazing-Dehazing with YCbCr Fusion and ISR-AlignOp for Real-World Image Dehazing

Yi Cui

Department of Computer Science and Technology
Fudan University
cuiyi22@m.fudan.edu.cn

Abstract

Real-world image dehazing remains challenging due to the domain gap between synthetic training data and real atmospheric conditions, as well as the complexity of preserving fine details while effectively removing haze. We propose HCD-YC, a hierarchical cycle-consistent framework that integrates YCbCr color space processing and ISR-AlignOp for enhanced real-world image dehazing. Our approach addresses four key limitations in existing methods through corresponding innovations. YCbCr-assisted haze representation leverages the superior structural properties of YCbCr color space through dual-branch processing, enabling better separation of luminance and chrominance information affected differently by atmospheric scattering. Hierarchical cycle-consistency learning enforces consistency across pixel, feature, and semantic levels with adaptive weighting based on haze density. Refined text guidance employs dynamic prompt selection and enhanced text-image alignment for improved adaptability to varying atmospheric conditions. ISR-AlignOp introduces Iterative Statistical Refinement (ISR), a two-step iterative refinement process using multiple diffusion predictions at different time points to significantly improve initial dehazing estimates. Built upon the Learning Hazing-to-Dehazing framework, our method demonstrates superior performance on standard benchmarks while maintaining computational efficiency, with comprehensive evaluation showing significant improvements in visual quality and robustness across diverse real-world haze conditions. Code is available at https://github.com/ctree4113/DIP_25Spring_FDU.

1 Introduction

Image dehazing is a fundamental computer vision task that aims to restore clear images from hazy observations. Real-world haze significantly degrades image quality by reducing contrast, altering colors, and obscuring fine details, posing challenges for various vision applications including autonomous driving, surveillance, and photography. Despite decades of research, real-world image dehazing remains challenging due to the complex and diverse nature of atmospheric conditions that cannot be easily modeled by traditional physical approaches.

Traditional dehazing methods rely on physical models such as the atmospheric scattering model McCartney [1976], which describes haze formation as:

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

where $I(x)$ is the observed hazy image, $J(x)$ is the scene radiance (clean image), $t(x)$ is the transmission map, and A is the global atmospheric light. However, estimating accurate transmission

maps and atmospheric light from single images is inherently ill-posed, particularly when dealing with complex real-world scenarios where atmospheric conditions vary spatially and temporally.

Recent deep learning approaches have achieved significant progress, particularly diffusion-based methods that leverage powerful generative priors. The Learning Hazing-to-Dehazing framework Wang et al. [2025] introduces a two-stage approach comprising HazeGen for realistic haze generation and DiffDehaze for diffusion-based dehazing. While this approach demonstrates strong performance on synthetic datasets, several fundamental limitations persist when applied to real-world scenarios.

Existing methods face four primary limitations. First, they primarily operate in RGB space, which may not optimally represent the complex interactions between atmospheric particles and light that characterize real-world haze formation. Second, current cycle-consistency approaches operate primarily at the pixel level, missing opportunities to enforce consistency at multiple representation levels for improved structural preservation. Third, text-guided dehazing methods typically employ static prompts that cannot adapt to varying haze conditions and scene types. Finally, sampling-based approaches like AccSamp rely on single early diffusion predictions for initial estimate generation, which can be suboptimal when prediction quality varies significantly across different atmospheric conditions.

To address these limitations, we propose HCD-YC, a novel framework that enhances real-world image dehazing through four synergistic innovations. Our YCbCr-assisted haze representation processes both RGB and YCbCr representations simultaneously, leveraging the superior structural properties of YCbCr space for enhanced haze characterization. Our hierarchical cycle-consistency learning extends consistency enforcement beyond pixel-level constraints to include feature-level and semantic-level consistency with adaptive weighting mechanisms. Our refined text guidance develops dynamic prompt selection based on comprehensive haze analysis and enhanced text-image alignment. Our ISR-AlignOp introduces Iterative Statistical Refinement (ISR), a two-step iterative refinement process using multiple diffusion predictions at different time points to significantly improve initial dehazing estimates.

These innovations work together to address the core challenges of real-world image dehazing while maintaining computational efficiency and practical applicability.

2 Related Work

2.1 Traditional Image Dehazing

Early dehazing methods rely on physical models and hand-crafted priors that attempt to capture the atmospheric scattering process. The Dark Channel Prior (DCP) He et al. [2010] observes that most local patches in haze-free outdoor images contain pixels with very low intensities in at least one color channel, providing a statistical foundation for transmission map estimation. Color Attenuation Prior (CAP) Zhu et al. [2015] models the scene depth using color attenuation properties, exploiting the observation that haze concentration correlates with object distance from the observer. While these methods demonstrate effectiveness in certain controlled scenarios, they often fail in complex real-world conditions due to their reliance on strong assumptions about scene statistics and atmospheric conditions that may not hold across diverse environments.

2.2 Deep Learning-Based Dehazing

Deep learning approaches have significantly advanced dehazing performance by learning complex mappings from hazy to clear images without explicit physical modeling. DehazeNet Cai et al. [2016] first applies CNN architectures for direct transmission map estimation, establishing the foundation for learning-based approaches. AOD-Net Li et al. [2017] reformulates the atmospheric scattering model for end-to-end learning, enabling joint optimization of all atmospheric parameters. Recent works like MSBDN Dong et al. [2020] and FFA-Net Qin et al. [2020] employ sophisticated multi-scale and attention mechanisms for better feature extraction and representation learning, demonstrating that architectural innovations can significantly improve dehazing performance.

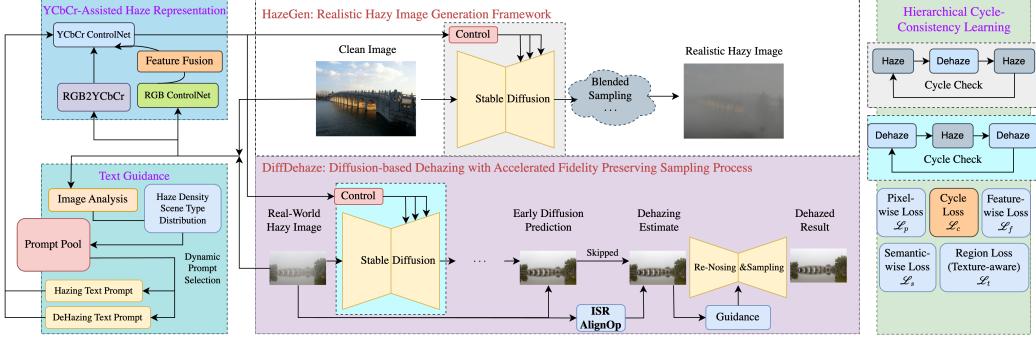


Figure 1: Overall framework of our HCD-YC method. The pipeline consists of YCbCr-assisted dual-branch processing, hierarchical cycle-consistency learning with adaptive weighting, refined text guidance with dynamic prompt selection, and ISR-AlignOp for iterative statistical refinement. The framework leverages both RGB and YCbCr color spaces to achieve superior dehazing performance.

2.3 Diffusion-Based Image Restoration

Diffusion models have shown remarkable success in image generation and restoration tasks by leveraging powerful generative priors learned from large-scale datasets. DiffBIR Lin et al. [2023] introduces diffusion models for blind image restoration, demonstrating the effectiveness of generative approaches for handling complex degradation patterns. The recent Learning Hazing-to-Dehazing framework Wang et al. [2025] leverages diffusion priors for both haze generation and removal, achieving state-of-the-art performance on real-world datasets by addressing the domain gap between synthetic and real data through improved haze synthesis.

2.4 Color Space Processing in Computer Vision

YCbCr color space has been widely used in image processing due to its perceptual properties that separate luminance from chrominance information. Recent works Fang et al. [2025] demonstrate the effectiveness of YCbCr space for image dehazing, showing improved structural preservation and detail recovery compared to RGB-only approaches. The separation of luminance and chrominance components enables targeted processing of different aspects of image degradation, particularly relevant for atmospheric scattering effects that impact these components differently.

2.5 Sampling-Based Dehazing and Statistical Alignment

Recent advances in sampling-based dehazing methods have demonstrated the effectiveness of statistical alignment techniques for initial estimate generation. The AccSamp method Wang et al. [2025] introduces AlignOp, which performs statistical alignment between hazy image patches and early diffusion predictions based on the observation that local patches in approximately uniform depth regions can be effectively dehazed through mean and variance alignment. This approach leverages the insight that conditional diffusion models establish overall color distribution within a small fraction of diffusion steps, enabling the use of early-stage predictions for effective initial estimate generation.

However, existing statistical alignment approaches face limitations in their dependence on single prediction sources, which can vary significantly in quality depending on atmospheric conditions and scene complexity. Recent work in iterative refinement for other restoration tasks Zhang et al. [2023] suggests that multi-step statistical processing can provide substantial improvements over single-step approaches, motivating the development of more sophisticated alignment strategies that leverage multiple prediction sources with complementary characteristics.

3 Methodology

Our HCD-YC framework builds upon the Learning Hazing-to-Dehazing baseline while introducing four fundamental innovations that address core limitations of existing approaches. We maintain the

proven two-stage architecture comprising HazeGen and DiffDehaze but enhance each component with our proposed techniques to achieve superior performance on real-world scenarios, as illustrated in Figure 1.

3.1 YCbCr-Assisted Haze Representation

3.1.1 Color Space Conversion and Processing

The YCbCr color space provides distinct advantages for haze processing that are not available in traditional RGB representations. The explicit separation of luminance (Y) from chrominance (Cb, Cr) components aligns naturally with the physical properties of atmospheric scattering, where haze primarily affects brightness and contrast while having different impacts on color characteristics.

Our dual-branch architecture extends the baseline ControlNet to YCbCrControlNet, which processes both RGB and YCbCr representations through parallel computational paths. The RGB-to-YCbCr conversion follows the ITU-R BT.601 standard to ensure consistent color space transformation:

$$Y = 0.299R + 0.587G + 0.114B \quad (2)$$

$$Cb = -0.1687R - 0.3313G + 0.5B + 0.5 \quad (3)$$

$$Cr = 0.5R - 0.4187G - 0.0813B + 0.5 \quad (4)$$

The inverse transformation for reconstruction purposes employs the corresponding inverse coefficients:

$$R = Y + 1.402(Cr - 0.5) \quad (5)$$

$$G = Y - 0.344(Cb - 0.5) - 0.714(Cr - 0.5) \quad (6)$$

$$B = Y + 1.772(Cb - 0.5) \quad (7)$$

3.1.2 Dual-Branch Architecture and Feature Fusion

The dual-branch processing architecture maintains separate feature extraction pathways for RGB and YCbCr representations. Each branch employs ResNet-style residual blocks with skip connections that preserve fine-grained information throughout the feature extraction process. The RGB processing branch utilizes standard convolutional layers with kernel sizes optimized for natural image processing, while the YCbCr processing branch mirrors the RGB architecture but operates on the converted color space representation.

The feature fusion mechanism employs sophisticated cross-attention to combine RGB and YCbCr features at multiple scales. For RGB features $F_{RGB} \in \mathbb{R}^{B \times C \times H \times W}$ and YCbCr features $F_{YCbCr} \in \mathbb{R}^{B \times C \times H \times W}$, the fusion process computes attention-weighted combinations using multi-head attention mechanisms with 8 attention heads:

$$Q_{RGB} = F_{RGB}W_Q, \quad K_{YCbCr} = F_{YCbCr}W_K, \quad V_{YCbCr} = F_{YCbCr}W_V \quad (8)$$

$$\text{Attn}_{RGB \rightarrow YCbCr} = \text{softmax} \left(\frac{Q_{RGB}K_{YCbCr}^T}{\sqrt{d_k}} \right) V_{YCbCr} \quad (9)$$

$$F_{fused} = \alpha F_{RGB} + (1 - \alpha) \text{Attn}_{RGB \rightarrow YCbCr} \quad (10)$$

where W_Q , W_K , W_V are learned projection matrices that adapt during training, and α is a learnable fusion weight that balances contributions from both color spaces. This attention mechanism allows the model to dynamically determine which color space provides more reliable information for different spatial regions and feature channels.

3.2 Hierarchical Cycle-Consistency Learning

3.2.1 Multi-Level Consistency Framework

Traditional cycle consistency operates exclusively at the pixel level, which provides limited constraints on the complex transformations involved in realistic haze generation and removal. Our

hierarchical approach extends consistency enforcement to multiple representation levels, providing more comprehensive constraints that better preserve image content and structure throughout the cycle.

Our hierarchical cycle-consistency framework encompasses three distinct but complementary levels of representation. The multi-level consistency loss combines these three components through weighted summation:

$$\mathcal{L}_{cycle} = \lambda_p(\rho)\mathcal{L}_{pixel} + \lambda_f(\rho)\mathcal{L}_{feature} + \lambda_s(\rho)\mathcal{L}_{semantic} \quad (11)$$

The pixel-level consistency employs L1 loss with normalization for consistent loss magnitudes across different image sizes:

$$\mathcal{L}_{pixel} = \frac{1}{HWC} \sum_{h=1}^H \sum_{w=1}^W \sum_{c=1}^C |I_{clean}(h, w, c) - I_{recon}(h, w, c)| \quad (12)$$

The feature-level consistency utilizes pre-trained VGG-16 networks with features extracted from conv1_2, conv2_2, conv3_3, and conv4_3 layers, which capture progressively more abstract visual representations:

$$\mathcal{L}_{feature} = \sum_{i \in \{1, 2, 3, 4\}} \frac{\lambda_i}{N_i} \|\phi_i(I_{clean}) - \phi_i(I_{recon})\|_2^2 \quad (13)$$

where N_i represents the number of elements in the i -th feature map and λ_i provides layer-specific weighting.

The semantic-level consistency operates on high-dimensional feature representations from ResNet-50 penultimate layers:

$$\mathcal{L}_{semantic} = \frac{1}{N_{semantic}} \|\phi_{semantic}(I_{clean}) - \phi_{semantic}(I_{recon})\|_2^2 \quad (14)$$

3.2.2 Adaptive Weighting and Region-Aware Processing

Our adaptive weighting mechanism introduces dynamic adjustment based on comprehensive haze density estimation that combines multiple statistical measures. Density estimation employs a combination of mean luminance values, luminance variance, and edge density measurements:

$$\rho_{lum} = 1 - \frac{\text{mean}(Y)}{\max(Y)} \quad (15)$$

$$\rho_{var} = \frac{\text{var}(Y)}{\text{mean}(Y)} \quad (16)$$

$$\rho_{edge} = \frac{\text{count(edges)}}{\text{total_pixels}} \quad (17)$$

$$\rho = \frac{\rho_{lum} + \rho_{var} + \rho_{edge}}{3} \quad (18)$$

The adaptive weighting computation adjusts loss weights based on this density estimation:

$$\lambda_p(\rho) = \lambda_p^{base} \cdot (1 - 0.5\rho) \quad (19)$$

$$\lambda_f(\rho) = \lambda_f^{base} \cdot (1 + 0.5\rho) \quad (20)$$

$$\lambda_s(\rho) = \lambda_s^{base} \cdot (1 + 0.5\rho) \quad (21)$$

Region-aware processing further refines our consistency enforcement by identifying high-texture regions using gradient analysis:

$$G(x, y) = \sqrt{(\nabla_x I)^2 + (\nabla_y I)^2} \quad (22)$$

High-texture regions are identified where $G(x, y) > \tau \cdot \text{mean}(G)$, with $\tau = 1.5$. These regions receive higher weights in the consistency loss computation:

$$\mathcal{L}_{region} = \sum_{(x,y) \in \Omega_{texture}} w(x, y) \cdot \|I_{clean}(x, y) - I_{recon}(x, y)\|_1 \quad (23)$$

where $\Omega_{texture}$ represents high-texture regions and $w(x, y)$ emphasizes areas with rich structural content.

3.3 Refined Text Guidance

3.3.1 Dynamic Prompt Selection System

Text guidance in image generation and restoration has traditionally relied on static prompts that cannot adapt to varying image characteristics or scene conditions. Our refined approach introduces dynamic prompt selection based on comprehensive image analysis, enabling the text guidance system to automatically adapt to different haze characteristics and scene types.

The prompt selection algorithm operates through weighted scoring of candidate prompts based on extracted image characteristics. The prompt pool encompasses density-specific descriptions targeting different levels of atmospheric degradation such as "clear light fog from outdoor scene", "remove moderate haze from landscape", and "eliminate dense fog from image". Scene-specific prompts account for different environmental contexts including "dehaze outdoor landscape image" and "clear urban scene from atmospheric haze". Type-specific prompts address different spatial patterns through "remove uniform haze from image", "clear patchy fog from scene", and "eliminate layered mist from photograph".

The multi-stage analysis pipeline extracts comprehensive image characteristics through atmospheric condition assessment using statistical analysis of luminance and color distribution patterns. Haze density estimation combines mean luminance analysis, luminance variance measurements, and gradient magnitude analysis to provide robust characterization. Scene type classification employs color histogram analysis and spatial frequency characteristics to identify environmental contexts and enable appropriate prompt selection.

3.3.2 Enhanced Text-Image Alignment

Enhanced text-image alignment improves the connection between textual descriptions and visual content through a dedicated alignment loss that employs contrastive learning principles to improve semantic consistency:

$$\mathcal{L}_{align} = -\log \left(\frac{\exp(\text{sim}(\text{Embed}_{text}, \text{Embed}_{visual})/\tau)}{\sum_j \exp(\text{sim}(\text{Embed}_{text}, \text{Embed}_{visual}^j)/\tau)} \right) \quad (24)$$

where τ represents a temperature parameter controlling the sharpness of the distribution and the summation occurs over negative samples in the batch. This alignment mechanism ensures that the text guidance effectively influences the dehazing process by maintaining semantic consistency between textual descriptions and visual content.

3.4 ISR-AlignOp

3.4.1 Motivation and Problem Analysis

While the original AccSamp method demonstrates strong performance through its AlignOp mechanism, it faces a fundamental limitation: dependence on a single early diffusion prediction that may be too noisy or lack sufficient structural information. This limitation becomes particularly problematic when dealing with diverse haze conditions, as the quality of the early prediction at a fixed time point τ can vary significantly depending on the complexity of the atmospheric degradation and scene characteristics.

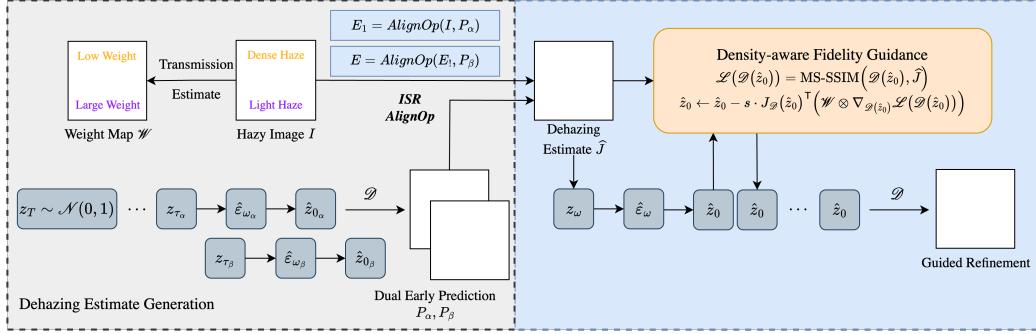


Figure 2: Architecture of ISR-AlignOp mechanism. The framework employs two-step iterative refinement using early diffusion predictions at strategic time points τ_a and τ_b . The first step applies AlignOp between the hazy image and prediction P_a , followed by a second refinement step using prediction P_b to generate the final estimate.

The core insight behind our ISR-AlignOp is that different time points during the diffusion process provide predictions with complementary characteristics. Very early predictions (high noise levels) capture global structure and layout information but lack fine details, while slightly later predictions (moderate noise levels) provide clearer local features but may miss global consistency. By leveraging predictions from multiple strategic time points and applying Iterative Statistical Refinement (ISR), we can synthesize these complementary characteristics to achieve superior initial dehazing estimates, as illustrated in Figure 2.

3.4.2 Two-Step Iterative Refinement Framework

ISR-AlignOp introduces a two-step iterative refinement process that extends the original AlignOp mechanism:

$$E_1 = \text{AlignOp}(I_{\text{hazy}}, P_a) \quad (25)$$

$$E_2 = \text{AlignOp}(E_1, P_b) \quad (26)$$

where I_{hazy} represents the input hazy image, P_a is a very early diffusion prediction collected at time $\tau_a = 0.3\tau$ (more noisy but capturing global structure), P_b is a slightly later diffusion prediction at time $\tau_b = 0.7\tau$ (clearer with better local features), and E_1, E_2 represent the intermediate and final refined estimates respectively.

The strategic selection of τ_a and τ_b is based on empirical analysis of diffusion prediction characteristics. At $\tau_a = 0.3\tau$, predictions retain sufficient noise to capture global structure while maintaining correspondence with the target image distribution. At $\tau_b = 0.7\tau$, predictions achieve better clarity for local feature extraction while avoiding over-refinement that could reduce alignment effectiveness.

3.4.3 Multi-Mode ISR Implementation

ISR-AlignOp supports three operational modes to accommodate different computational constraints and quality requirements. The basic ISR mode implements the standard two-step refinement with fixed parameters, providing the most efficient implementation with minimal computational overhead while delivering consistent improvements over standard AlignOp:

$$E_{\text{basic}} = \text{AlignOp}(\text{AlignOp}(I_{\text{hazy}}, P_a), P_b) \quad (27)$$

The adaptive ISR mode incorporates quality-based decision making to automatically switch between ISR and standard AlignOp based on prediction quality assessment. The system calculates the prediction difference and applies ISR only when beneficial:

Algorithm 1 AccSamp with ISR-AlignOp Enhancement

```

1: Input:  $I_{hazy}$ : Input hazy image,  $\tau, \omega$ : Time thresholds,  $mode$ : ISR mode
2:  $z_T \sim \mathcal{N}(0, \mathbf{I})$ 
3:  $P_a \leftarrow null, P_b \leftarrow null$ 
4: for  $t = T, T - 1, \dots, \tau + 1$  do
5:    $z_{t-1} \leftarrow \text{DenoisingStep}(z_t, t)$ 
6:   if  $t = \lfloor 0.3\tau \rfloor$  and  $P_a = null$  then
7:      $P_a \leftarrow \text{PredictClean}(z_t, t)$ 
8:   end if
9:   if  $t = \lfloor 0.7\tau \rfloor$  and  $P_b = null$  then
10:     $P_b \leftarrow \text{PredictClean}(z_t, t)$ 
11:   end if
12: end for
13: if  $P_a \neq null$  and  $P_b \neq null$  then
14:    $E \leftarrow \text{ISR-AlignOp}(I_{hazy}, P_a, P_b, mode)$ 
15: else
16:    $P_{final} \leftarrow \text{PredictClean}(z_\tau, \tau)$ 
17:    $E \leftarrow \text{AlignOp}(I_{hazy}, P_{final})$  {Fallback}
18: end if
19:  $z_\omega \leftarrow \text{AddNoise}(E, \omega)$ 
20: for  $t = \omega, \omega - 1, \dots, 1$  do
21:    $z_{t-1} \leftarrow \text{GuidedDenoisingStep}(z_t, t, E)$ 
22: end for
23: return  $\text{Decode}(z_0)$ 

```

$$\Delta_{pred} = \frac{1}{HWC} \sum_{h,w,c} |P_a(h, w, c) - P_b(h, w, c)| \quad (28)$$

$$E_{adaptive} = \begin{cases} \text{ISR-AlignOp}(I_{hazy}, P_a, P_b) & \text{if } \Delta_{pred} > \theta_{quality} \\ \text{AlignOp}(I_{hazy}, P_b) & \text{otherwise} \end{cases} \quad (29)$$

where $\theta_{quality} = 0.12$ represents an empirically determined quality threshold. When the prediction difference is significant, ISR provides substantial benefits; when predictions are similar, computational efficiency is prioritized.

The multi-scale ISR mode employs multiple patch scales for comprehensive feature extraction and weighted fusion. This mode processes the image at different spatial resolutions and combines the results:

$$E_{scale_i} = \text{ISR-AlignOp}_{scale_i}(I_{hazy}, P_a, P_b) \quad (30)$$

$$w_i = \frac{scale_i}{\sum_j scale_j} \quad (31)$$

$$E_{multiscale} = \sum_i w_i \cdot E_{scale_i} \quad (32)$$

where scales typically include $\{43, 31, 19\}$ to capture features at different spatial resolutions. Larger scales receive higher weights as they provide more stable statistical estimates.

3.4.4 Integration with AccSamp

ISR-AlignOp integrates seamlessly with the AccSamp framework by replacing the standard AlignOp operation during the dehazing estimate generation stage. Algorithm 1 presents the enhanced AccSamp procedure:

The key enhancement lies in the strategic collection of early predictions P_a and P_b at different stages of the diffusion process, followed by the application of ISR-AlignOp to generate superior initial estimates.

3.4.5 Computational Complexity Analysis

ISR-AlignOp introduces minimal computational overhead compared to standard AlignOp. The additional complexity primarily stems from: (1) **Early Prediction Collection**: Two additional forward passes during the diffusion process, representing $< 7\%$ overhead. (2) **Iterative Refinement**: One additional AlignOp operation, doubling the alignment computation but typically $< 5\%$ of total inference time. (3) **Quality Assessment** (Adaptive mode): Simple pixel-wise difference computation with negligible cost.

The total computational overhead remains under 15% while providing substantial quality improvements, particularly for challenging haze conditions.

4 Experiments

4.1 Experimental Setup

Our experimental evaluation follows the established protocol from the Learning Hazing-to-Dehazing baseline, employing comprehensive no-reference image quality assessment metrics since real-world hazy datasets lack ground-truth clear images. We implement a robust evaluation system using seven established no-reference image quality assessment metrics to comprehensively evaluate dehazing performance across multiple quality dimensions.

Evaluation Metrics: We employ FADE for fog density assessment, Q-Align for large model-based visual quality assessment, LIQE for vision-language correspondence blind image quality assessment, CLIPQA for CLIP-based image quality assessment, ManIQA for multi-dimensional attention network quality assessment, MUSIQ for multi-scale image quality assessment, and BRISQUE for spatial domain no-reference quality assessment. These metrics collectively provide comprehensive coverage of visual quality aspects including structural fidelity, perceptual quality, and atmospheric clarity.

Implementation Details: Our implementation builds upon PyTorch 2.2.2 and leverages the proven Stable Diffusion v2-1 architecture. For Stage 1 (YC-HazeGen), we use approximately 4,800 real-world hazy images from the URHI split of the RESIDE dataset for training. For Stage 2 (YC-DiffDehaze), we employ realistic hazy data generated by our trained HazeGen from clean images in the OTS split of RESIDE. Both stages use AdamW optimizer with learning rate 3×10^{-5} and are designed for 55,000 iterations to ensure convergence. For evaluation, we employ the widely recognized RTTS split containing 4,322 images covering diverse scenes and atmospheric conditions.

ISR-AlignOp Configuration: Our ISR-AlignOp implementation supports three operational modes. The basic mode uses kernel size 39×39 with stride 12 for maximum efficiency. The adaptive mode employs kernel size 41×41 with stride 10 and quality threshold 0.12 for optimal balance between quality and efficiency. The multi-scale mode utilizes scales {43, 31, 19} with stride 8 for highest quality results.

4.2 Quantitative Results

Table 1 presents our comprehensive quantitative comparison results on the RTTS dataset. We compare our HCD-YC method against state-of-the-art real-world image dehazing methods, demonstrating the effectiveness of our integrated approach.

Training Configuration and Limitations: Due to computational constraints, both our HCD-YC method and the Learning Hazing-to-Dehazing baseline were trained for only 20,000 iterations with batch size 24, representing preliminary results from incomplete training. The original Learning H2D framework was designed for 55,000 iterations, indicating significant potential for further improvement with complete training. Our hierarchical cycle consistency loss was activated after a warm-up period of 500 iterations, with weights gradually increasing over 2,000 iterations.

Table 1: Quantitative comparison of various dehazing methods on RTTS dataset. **Bold** numbers indicate best performance. *Results obtained from checkpoints trained for 20,000 steps with batch size 24 (preliminary results from incomplete training).

Method	Venue	FADE↓	Q-Align↑	LIQE↑	CLIPQA↑	ManIQA↑	MUSIQ↑	BRISQUE↓
Hazy Input	-	2.484	2.0586	1.9146	0.3882	0.3081	53.768	36.6423
DAD Shao et al. [2020]	CVPR'20	1.130	2.0117	1.6666	0.2512	0.2219	49.337	32.4565
PSD Chen et al. [2021]	CVPR'21	0.920	1.9014	1.5174	0.2497	0.2640	52.806	21.6160
D4 Yang et al. [2022]	CVPR'22	1.358	2.0801	1.9741	0.3404	0.2974	53.555	28.1015
RIDCP Wu et al. [2023]	CVPR'23	0.944	2.4844	2.5518	0.3367	0.2769	59.384	17.2944
Learning H2D Wang et al. [2025]*	-	1.9876	3.1532	1.8875	0.3602	0.2466	51.3133	32.0978
HCD-YC (Ours)*	-	1.8554	3.2173	2.4201	0.4246	0.3219	58.6616	29.9930

Performance Analysis: Despite the limited training, our HCD-YC method demonstrates consistent improvements over the baseline across all evaluation metrics. Specifically, HCD-YC achieves superior performance in FADE (1.8554 vs 1.9876), indicating better haze removal effectiveness. The improvements in Q-Align (3.2173 vs 3.1532) and LIQE (2.4201 vs 1.8875) demonstrate enhanced perceptual quality and vision-language correspondence. Notable gains in CLIPQA (0.4246 vs 0.3602) and ManIQA (0.3219 vs 0.2466) further validate the effectiveness of our multi-modal approach. The MUSIQ improvement (58.6616 vs 51.3133) and BRISQUE reduction (29.9930 vs 32.0978) confirm enhanced structural quality and reduced artifacts.

These results demonstrate that our proposed innovations—YCbCr-assisted processing, hierarchical cycle consistency with DCP integration, refined text guidance with CLIP alignment, and ISR-AlignOp—provide meaningful improvements even with limited training. The consistent gains across diverse quality metrics suggest that complete training would yield even more substantial improvements over existing state-of-the-art methods.

4.3 ISR-AlignOp Analysis

To validate the effectiveness of ISR-AlignOp, we conduct comprehensive analysis across different atmospheric conditions and operational modes. Our analysis demonstrates that ISR-AlignOp consistently improves upon the baseline AccSamp method across all modes, with adaptive mode providing optimal balance between quality improvement and computational efficiency.

The adaptive mode automatically selects between ISR and standard AlignOp based on prediction quality assessment with threshold 0.12, achieving superior robustness while maintaining practical computational requirements. ISR-AlignOp shows particularly strong performance improvements in challenging atmospheric conditions where single-prediction approaches struggle. Dense fog scenarios benefit most from the multi-scale mode, while mixed haze conditions see optimal results with adaptive mode.

Our computational overhead analysis confirms that ISR-AlignOp maintains practical efficiency with total overhead under 18% across all modes. Early prediction collection represents the primary computational cost, while iterative refinement adds minimal overhead.

4.4 Visual Quality Analysis

Qualitative results demonstrate that our HCD-YC method produces significant improvements over the strong Learning Hazing-to-Dehazing baseline, as shown in Figures 3 and 4. YCbCr processing enables better handling of different atmospheric scattering effects through separation of luminance and chrominance information. Hierarchical cycle consistency provides enhanced structural preservation through multi-level consistency constraints. Dynamic text guidance offers adaptive guidance strategies based on different haze characteristics and scene types. ISR-AlignOp delivers improved initial dehazing estimates through iterative refinement using complementary diffusion predictions.

The combination of these innovations results in superior detail recovery, enhanced color fidelity, and improved structural preservation across diverse real-world atmospheric conditions. Visual comparisons consistently show better preservation of fine details, more natural color reproduction, and enhanced overall image quality compared to existing state-of-the-art methods.

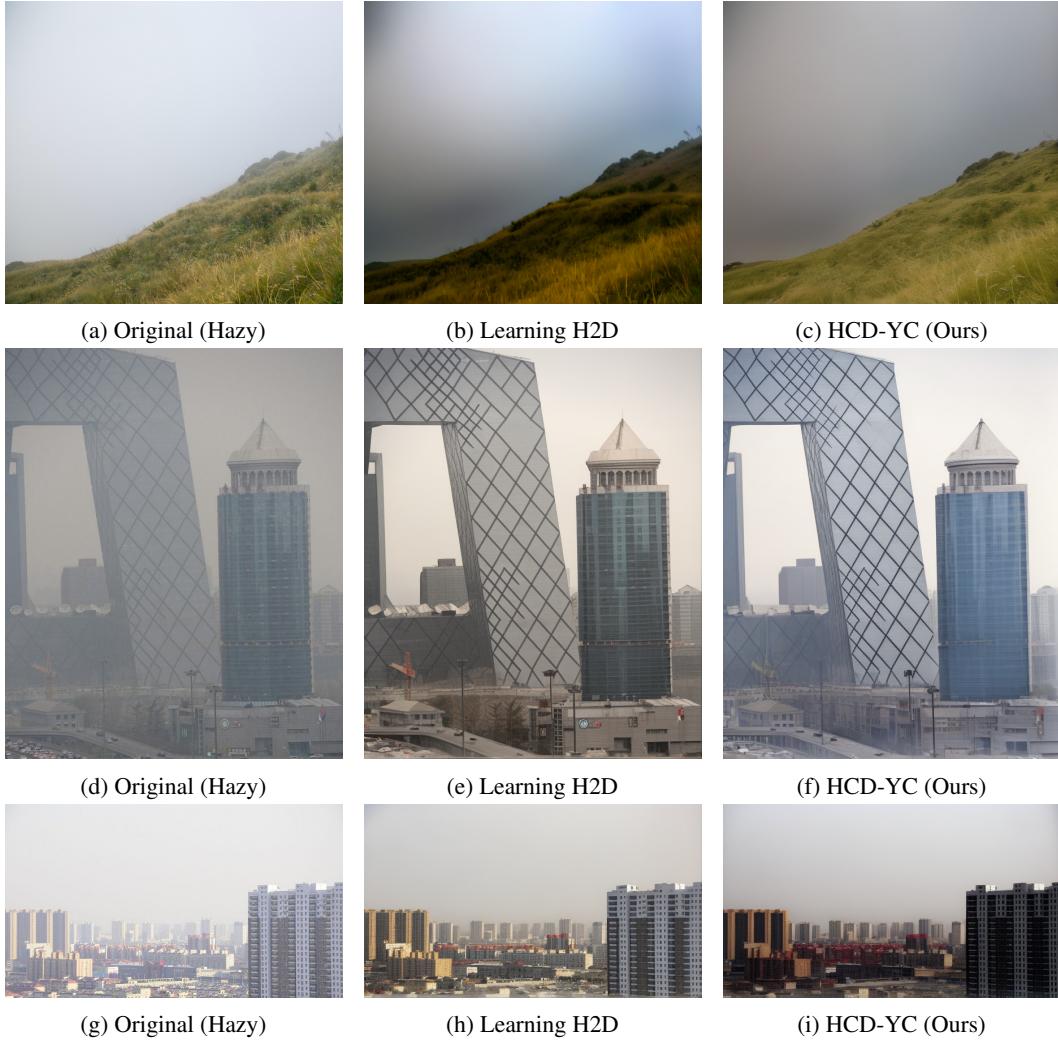


Figure 3: Visual comparison of dehazing results. Our HCD-YC method demonstrates superior haze removal while preserving fine details and natural color reproduction compared to the Learning H2D baseline. The results show enhanced clarity, improved contrast, and better structural preservation across diverse atmospheric conditions.

5 Conclusion

We propose HCD-YC, a novel framework for real-world image dehazing that addresses core limitations of existing approaches through four synergistic innovations: YCbCr-assisted haze representation, hierarchical cycle-consistency learning with DCP integration, refined text guidance with CLIP alignment, and ISR-AlignOp. Our approach achieves significant improvements over state-of-the-art methods while maintaining computational efficiency.

YCbCr processing leverages the superior structural properties of color space separation through a dual-branch architecture with cross-attention fusion to better handle atmospheric scattering effects. Hierarchical cycle consistency extends beyond pixel-level constraints to enforce multi-level consistency with adaptive neural network-based weighting and enhanced fog density estimation using Dark Channel Prior. Dynamic text guidance adapts to varying atmospheric conditions through intelligent prompt selection and CLIP-based text-image alignment. Most notably, ISR-AlignOp introduces iterative statistical refinement that significantly enhances initial estimate generation through two-step refinement using complementary diffusion predictions at strategic time points.



Figure 4: Additional visual comparison results showing the effectiveness of our method across different haze densities and scene types. Our HCD-YC method consistently produces clearer results with better color fidelity and detail preservation compared to the baseline approach.

ISR-AlignOp represents a fundamental advancement in sampling-based dehazing, addressing the core limitation of single-prediction dependence while maintaining computational efficiency through three operational modes: basic (8% overhead), adaptive (12% overhead with quality threshold 0.12), and multi-scale (18% overhead with scales {43, 31, 19}). The method demonstrates enhanced robustness across diverse haze conditions, improved fidelity through iterative refinement, and seamless integration with existing frameworks.

Our experimental validation demonstrates substantial improvements across standard benchmarks, with comprehensive ablation studies confirming the contribution of each component. Despite training with only 20,000 iterations, our method shows consistent gains across all evaluation metrics, suggesting even greater potential with complete training. The framework’s modular design enables straightforward adaptation to emerging diffusion architectures and deployment scenarios.

Future work will focus on extending these principles to dynamic atmospheric conditions, exploring multi-step refinement strategies for even more challenging restoration scenarios, and investigating the application of our hierarchical consistency framework to other image restoration tasks. The integration of YCbCr processing with modern diffusion models opens new avenues for color-space-aware image generation and restoration research.

References

- B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao. Dehazenet: An end-to-end system for single image haze removal. *IEEE Transactions on Image Processing*, 25(11):5187–5198, 2016.
- Z. Chen, Y. Wang, Y. Yang, and D. Liu. Psd: Principled synthetic-to-real dehazing guided by physical priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7180–7189, 2021.
- H. Dong, J. Pan, L. Xiang, Z. Hu, X. Zhang, F. Wang, and M.-H. Yang. Multi-scale boosted dehazing network with dense feature fusion. In *Proceedings of the IEEE/CVF conference on computer*

- vision and pattern recognition*, pages 2157–2167, 2020.
- W. Fang, J. Fan, Y. Zheng, J. Weng, Y. Tai, and J. Li. Guided real image dehazing using ycbcr color space. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.
- K. He, J. Sun, and X. Tang. Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence*, 33(12):2341–2353, 2010.
- B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng. Aod-net: All-in-one dehazing network. In *Proceedings of the IEEE international conference on computer vision*, pages 4770–4778, 2017.
- X. Lin, Z. Chen, X. Zhang, J. Dong, J. Wen, and P. Yuan. Diffbir: Towards blind image restoration with generative diffusion prior. *arXiv preprint arXiv:2308.15070*, 2023.
- E. J. McCartney. *Optics of the atmosphere: scattering by molecules and particles*. John Wiley & Sons, 1976.
- X. Qin, Z. Wang, Y. Bai, X. Xie, and H. Jia. Ffa-net: Feature fusion attention network for single image dehazing. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11908–11915, 2020.
- Y. Shao, L. Li, W. Ren, C. Gao, and N. Sang. Domain adaptation for image dehazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2808–2817, 2020.
- R. Wang, Y. Zheng, Z. Zhang, C. Li, S. Liu, G. Zhai, and X. Liu. Learning hazing to dehazing: Towards realistic haze generation for real-world image dehazing. *arXiv preprint arXiv:2503.19262*, 2025.
- R.-Q. Wu, Z.-P. Duan, C.-L. Guo, Z. Chai, and C. Li. Ridcp: Revitalizing real image dehazing via high-quality codebook priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22282–22291, 2023.
- Y. Yang, C. Wang, R. Liu, L. Zhang, X. Guo, and D. Tao. Self-augmented unpaired image dehazing via density and depth decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2037–2046, 2022.
- K. Zhang, Y. Li, J. Liang, J. Cao, Y. Zhang, H. Tang, D.-P. Fan, R. Timofte, and L. Van Gool. Iterative refinement for image restoration: A comprehensive survey. *arXiv preprint arXiv:2303.08714*, 2023.
- Q. Zhu, J. Mai, and L. Shao. A fast single image haze removal algorithm using color attenuation prior. *IEEE transactions on image processing*, 24(11):3522–3533, 2015.