Towards a Hippocampus–Neocortex Inspired Two-Stage Diffusion Framework for Single-Image Dehazing

Xinyue Liu (xyliu@dlut.edu.cn)[1], Jiayu Liu (liujiayu@mail.dlut.edu.cn)[1], Junchi Wu[1], Linlin Zong (llzong@dlut.edu.cn)[1] & Wenxin Liang[*] (wxliang@dlut.edu.cn)[1]

[1]School of Software, Dalian University of Technology, [*] Corresponding author

Abstract

Modeling perceptual restoration under severe degradation remains a core challenge in vision science. Inspired by the hierarchical organization of human episodic memory, we propose a cognitively grounded two-stage diffusion framework for single image dehazing. Analogous to neocortical–hippocampal division of labor, the first stage encodes a coarse, stable global structure, while the second stage progressively restores fine-grained details conditioned on this intermediate representation. This staged reconstruction mirrors memory consolidation, where global context is stabilized before perceptual detail is refined, enabling a more efficient and cognitively plausible coarse-to-fine inference process. By decomposing restoration into functionally distinct stages, our framework avoids inefficient multi-scale optimization and promotes stable, coherent convergence through target-conditioned diffusion dynamics. Extensive experiments demonstrate that the proposed model achieves state-of-the-art performance, exhibiting strong robustness, interpretability, and practical applicability in complex real-world dehazing scenarios.

Keywords: Diffusion Model; Iterative Refinement; Single Image Dehazing

Introduction

Perceptual restoration under severe environmental degradation is not merely an image processing problem, but a fundamental challenge for both vision science and artificial intelligence. Human perception exhibits a remarkable ability to recover coherent and stable representations from highly corrupted sensory input, suggesting the existence of structured, hierarchical inference mechanisms that support perceptual stability and refinement. In contrast, current computational models for visual restoration remain fragmented, often emphasizing either local feature extraction or global relational modeling, but rarely integrating both within a unified, cognitively plausible framework.

Existing CNN-based approaches primarily operate as feed-forward local processors, effectively capturing fine-grained visual patterns but lacking mechanisms for constructing globally consistent perceptual hypotheses. For example, DehazeNet (Cai et al., 2016) pioneered the use of neural networks to estimate atmospheric scattering model (ASM) parameters, while AOD-Net (B. Li et al., 2017), GridDehazeNet (X. Liu et al., 2019), and MsCNN (Kundu & Ari, 2020) abandoned explicit physical modeling and instead learned haze-to-clear image mappings in an end-to-end manner. From a cognitive perspective, these models resemble early sensory processing stages that emphasize local receptive field responses, yet struggle to support higher-level perceptual organization, often resulting in incomplete structural recovery and inconsistent visual coherence (Blau & Michaeli, 2018; Z. Chen et al., 2024; Delbracio et al., 2021; Freirich et al., 2021; T. Li et al., 2025; Yuan et al., 2023).

Transformer-based models, by contrast, prioritize global contextual integration, analogous to higher-order cortical processing, but often exhibit insufficient sensitivity to subtle low-level details, leading to degraded perceptual fidelity. Vision Transformers (ViTs) have been introduced into dehazing networks. Leveraging the self-attention mechanism for capturing long-range contextual relationships (N. Zhang et al., 2023), ViT-based models such as SAN (Liang et al., 2019), SAD-Net (Sun et al., 2022), and Dehamer (Guo et al., 2022) enhance the representational capacity of dehazing networks by integrating channel and spatial attention for refined feature learning. However, the intrinsic patch-based tokenization of ViTs weakens their sensitivity to fine-grained local textures (Z. Chen et al., 2021; H. Wu, Xiao, et al., 2021), leading to locally incoherent or over-smoothed details in dehazed outputs (Yuan et al., 2023). To leverage the complementary strengths of CNNs and Transformers, hybrid architectures have recently been proposed across multiple vision domains (B. Chen et al., 2024; J. Liu et al., 2023; N. Zhang et al., 2023). For instance, LHNet (Yuan et al., 2023) integrates a CNN backbone with a Transformer sub-network and an adaptive aggregation module to jointly exploit local and global cues. This hybrid design effectively enhances performance beyond pure CNN or ViT counterparts. Nevertheless, the fusion of local and global representations remains nontrivial due to their inherent differences in scale, semantic granularity, and feature abstraction levels.

Recently, diffusion models (Ho et al., 2020) have drawn intensive attention due to their strong ability to generate high-quality images both unconditionally (Dhariwal & Nichol, 2021; Y. Song et al., 2021) and conditionally (Saharia et al., 2023; Whang et al., 2022). Diffusion family offers a compelling analogue to human perceptual inference: they iteratively refine noisy sensory representations into coherent percepts, mirroring the coarse-to-fine, hierarchical integration processes observed in human vision. Each step in the diffusion process can be seen as a constrained update of a perceptual hypothesis, akin to predictive coding or probabilistic inference, where global context and local detail are gradually

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

reconciled. This interpretability and structured generative trajectory distinguish diffusion models from other generative paradigms, suggesting a closer alignment with biologically plausible mechanisms of perceptual reconstruction and stability. However, diffusion models face difficulties when directly used for single image dehazing, since the huge distribution difference between dense-haze and clear images makes diffusion models struggle with weak and deviate distribution guidance, and the high variance with respect to predicting the spatial mean of the score function leads to color shift (Deck & Bischoff, 2023; Yu et al., 2023).

DehazeDDPM (Yu et al., 2023) is the first diffusion model specifically designed for the dehazing task, which applies physical priors to compensate information of different image regions. Nevertheless, the physical priors have limited effect in dense haze scenarios, and the reliance of ASM limits the model's transferability. Moreover, it is admitted that DehazeDDPM still suffers from the color shift problem (Deck & Bischoff, 2023). Experimental results reported in the literatures (Yu et al., 2023) demonstrate that DehazeDDPM does not outperform other methods like LNNet(Yuan et al., 2023) and MITNet(Shen et al., 2023), this indicates that relying on a diffusion model alone is inadequate to address the dehazing problem well. DiffLI2D(Yang et al., 2025) exploits the semantic latent space of frozen pre-trained diffusion models, revealing that haze and content information can be disentangled across diffusion time steps and used as guidance for dehazing. This line of work demonstrates the promise of diffusion-based representations for dehazing, but remains limited to latent feature extraction rather than full generative reconstruction.

In this work, we propose D2-Dehaze, a two-stage diffusion framework for single-image dehazing that mirrors the complementary roles of the neocortex and hippocampus in human cognition. The first stage captures long-range global structures by down-sampling hazy images and targets, then applying a diffusion process to iteratively refine a coarse intermediate dehazed result. This resembles how the hippocampus rapidly encodes global contextual information as a stable scaffold. The second stage focuses on high-frequency local details, treating the intermediate output as a condition for super-resolution-like refinement, akin to the neocortex gradually consolidating and enriching perceptual representations.

By decomposing dehazing into these functionally distinct stages, D2-Dehaze avoids the inefficiencies of one-stage diffusion and effectively reconciles global and local information without explicit feature fusion. Supervision from the ground-truth target consistently guides each stage toward accurate structural and textural manifolds, enhancing stability and generalization. Extensive evaluations on Dense-Haze, NH-HAZE, and synthetic datasets demonstrate that D2-Dehaze achieves state-of-the-art performance across both perceptual (FID, LPIPS) and distortion (PSNR, SSIM) metrics, while producing visually consistent, high-fidelity results in diverse real-world scenarios.

Our Method

Overall Framework

We introduce a new single-image dehazing framework, D2-Dehaze, based on two-stage diffusion refinement, as illustrated in Fig. 1. The framework is composed of two distinct stages inspired by the complementary roles of the hippocampus and neocortex.

The first stage, diffusion for global dehazing (DGD), corresponds to hippocampal-style rapid abstraction of global structure. In DGD, each pair of hazy image X and target image Y are down-sampled to smaller-sized images x and y. The x and y pairs are used as conditions and inputs to train a diffusion model. DGD produces an intermediate dehazed image I_m corresponding to an input X.

The second stage, diffusion for detail recovery (DDR), mirrors neocortical-style gradual refinement of fine details. In DDR, each intermediate image I_m is up-sampled to an image Z with the same size of the corresponding input X, then the Z and Y pairs are used as conditions and inputs to train another diffusion model. DDR obtains the final dehazed results.

Stage 1: Diffusion for Global Dehazing

Down-sample We use average-pooling to down-sample each input hazy image X and its corresponding target image Y to smaller sized x and y. With a parameter γ as the scaling factor and the constant 2 as the sliding step size, $2^\gamma * 2^\gamma$ pixels from the input hazy image are average-pooled to a single pixel in the down-sampled image.

Down-sampling reduces local variability and noise while capturing essential global structure, analogous to the hippocampus rapidly encoding coarse, stable representations of a scene. By averaging features over localized regions, this process enables diffusion models to efficiently extract global context, suppress high-frequency noise, and stabilize subsequent reconstruction, while also reducing computational complexity for large-scale processing.

Diffusion With x as the condition and y as the input, we train a diffusion model D_1 to denoise from noise to y. Down-sampling plays a hippocampal role by compressing visual input into a coarse memory trace that preserves global structure while suppressing high-frequency noise. Rather than performing detailed reconstruction, D_1 acts as a hippocampo-cortical interface, transforming compressed traces into stable global prototypes through stochastic replay. By fitting noise sequences under degraded-image guidance, it approximates the clear-image distribution, generalizes to unseen samples, and preserves global semantics while filtering local detail.

The noise addition process during training is illustrated in Fig. 2 and Algorithm 1. Starting from $t = 0$, y_t denotes the image corrupted by noise at time t. The conditional distribution of y_t given y_0, obtained by marginalizing intermediate states, is defined in Eq. 1.

$$y_t = \sqrt{\bar{\alpha}_t} y_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \epsilon \sim \mathcal{N}(0, I), \quad (1)$$

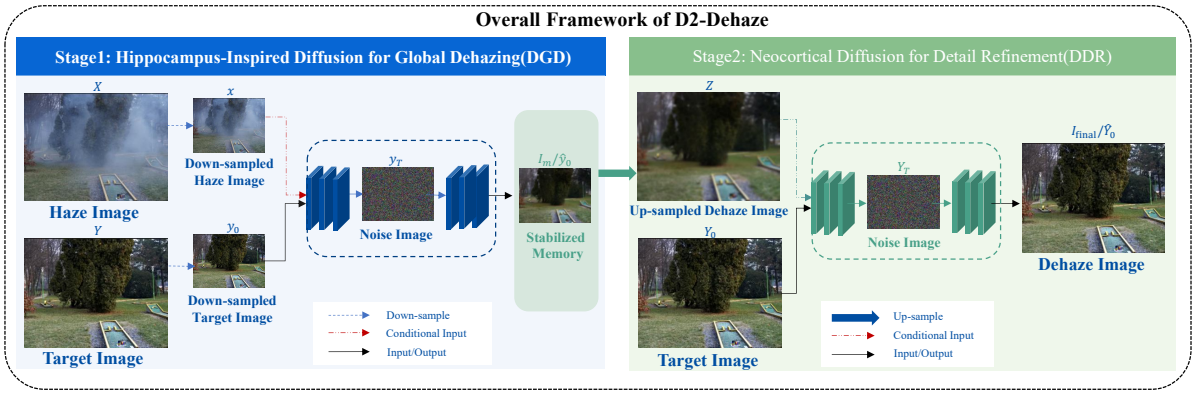


Figure 1: Overall framework of our proposed D2-Dehaze.

Algorithm 1 Training of diffusion for global dehazing

repeat
  $(x, y_0) \sim p(x, y)$
  $t \sim Uniform(\{1, \ldots, T\})$
  $\epsilon \sim \mathcal{N}(0, I)$
  Take gradient descent step on
    $\nabla_\theta \left\| \epsilon_\theta(x, y_t, \bar{\alpha}_t) - \epsilon \right\|^2$
until converged

$t \in [0, T]$, where $T > 0$ is a fixed constant. The noise term ϵ follows a normal distribution, $\epsilon \sim \mathcal{N}(0, I)$. The cumulative product of the coefficients is defined as $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, where the hyperparameters $\alpha_{1:T}$ satisfy $0 < \alpha_t < 1$. These parameters control the variance of the noise introduced at each iteration. Furthermore, y_0 is scaled by $\sqrt{\bar{\alpha}_t}$, which progressively reduces its contribution. This scaling ensures that the overall variance of the random variables remains bounded as $t \to \infty$. Finally a pure noise image y_T is got under the condition x.

In each noise addition step, t and y_t are fed to a U-net network for training, which learns to estimate the noise under the constraint of L_2 norm. Specifically, the U-Net is trained on the following variant of the variational bound:

$$L(\theta) := E_{t, y_0, \epsilon}\left[\left\|\epsilon_\theta\left(x, \sqrt{\bar{\alpha}_t} y_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \bar{\alpha}_t\right) - \epsilon\right\|^2\right], \tag{2}$$

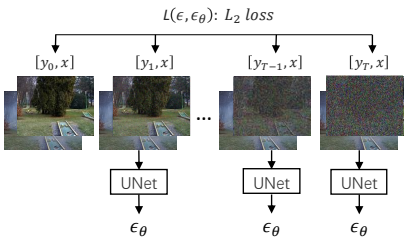


Figure 2: Training of diffusion for global dehazing.

Algorithm 2 Inference of diffusion for global dehazing

$\epsilon_\theta \sim \mathcal{N}(0, I)$
for $t = T, \ldots, 1$ **do**
  $\epsilon \sim \mathcal{N}(0, I)$ if $t > 1$, else $\epsilon = 0$
  $y_{t-1} \leftarrow \frac{1}{\sqrt{\alpha_t}}\left(y_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x, y_t, \bar{\alpha}_t)\right)$
  $+ \sqrt{(1-\alpha_t)\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}}\epsilon$
end for
return $\hat{y}_0$

where $\epsilon \sim \mathcal{N}(0, I)$, ϵ_θ describes the trained U-Net network.

The *inference process* from noise to hazy-free distribution is shown in Fig.3 and Algorithm 2 .

Firstly, D_1 samples pure noise, denoted as y_T, from a Gaussian distribution. y_T and x are then input to the trained U-Net network. The U-Net network produces an estimate of the current noise, ϵ_θ. As a condition, x provides fundamental distribution guidance, transforming the standard normal distribution into a more clear data distribution.

Given y_t and $\epsilon_\theta(x, y_t, \bar{\alpha}_t)$, the posterior distribution of y_{t-1} can be obtained as follows:

$$y_{t-1} \leftarrow \frac{1}{\sqrt{\alpha_t}}\left(y_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x, y_t, \bar{\alpha}_t)\right) + g_1(t)\epsilon, \tag{3}$$

$$g_1(t) \leftarrow \sqrt{(1-\alpha_t)\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}}, \tag{4}$$

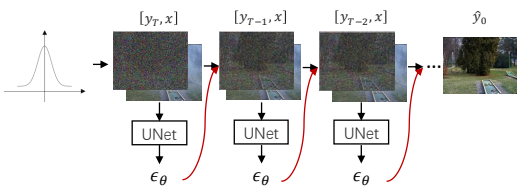


Figure 3: Inference of diffusion for global dehazing.

where $\epsilon \sim \mathcal{N}(0, I)$. This can be interpreted as one step of Langevin dynamics, where ϵ_θ acts as an approximation to the gradient of the data log-density. Notably, to reduce computation, when estimating noise ϵ_θ at the current moment, we simultaneously pass $\bar{\alpha}_t$ along with the indispensable x and y_t. y_{t-1} is then iteratively input to the U-Net to continue noise estimation and denoising.

After T step iterations, D_1 obtains the fitted distribution $\hat{y}_0$ for y_0. $\hat{y}_0$ is intermediate result image and is denoted as I_m, which will be used as another condition in the second stage.

Stage 2: Diffusion for Detail Recovery

Up-sample The goal of this stage is to recover fine details from the up-sampled intermediate image I_m. Serving as a hippocampus-like memory trace that preserves global structure, I_m is up-sampled via bicubic interpolation to produce Z, where each pixel aggregates information from 16 neighbors using a distance-weighted sum, analogous to the neocortex gradually refining detailed perceptual representations based on stabilized global context.

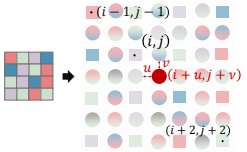


Figure 4: Up-sampling interpolation. The original pixels are in square shapes and the interpolated pixels are in circle shapes

Diffusion With Z as the condition and Y as the input, we train another diffusion model D_2 to learn the denoising process from noise to Y.

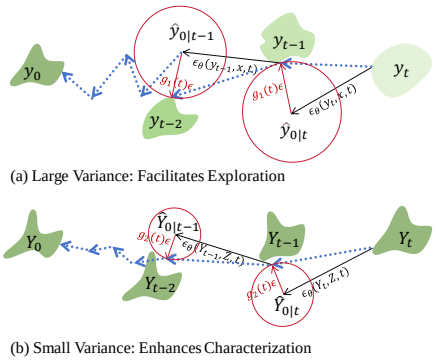
The training procedure of D_2 follows the same diffusion learning paradigm as D_1, reflecting a shared statistical learning mechanism. However, their inference behaviors differ in variance scheduling, mirroring the functional distinction between the hippocampus and neocortex. The first-stage model adopts a larger variance to enable broad stochastic exploration, analogous to hippocampal encoding of uncertain global structure. In contrast, the second-stage model employs a smaller variance to support fine-grained refinement under strong structural constraints, resembling neocortical consolidation and detail integration based on stabilized memory traces Z.

To make the variance smaller, we replace the variance in Eq.4 with Eq.5, i.e,

$$g_2(t) \leftarrow \sqrt{(1-\alpha_t)}. \tag{5}$$

Thus, in D_2, the posterior distribution of Y_{t-1} is represented by Eq.6:

$$Y_{t-1} \leftarrow \frac{1}{\sqrt{\alpha_t}}\left(Y_t - \frac{1-\alpha_t}{\sqrt{1-\bar{a}_t}}\epsilon_\theta\left(Z, Y_t, \bar{\alpha}_t\right)\right) + g_2(t)\,\epsilon. \tag{6}$$



(a) Large Variance: Facilitates Exploration

(b) Small Variance: Enhances Characterization

Figure 5: Reverse diffusion with variance modulation: large variance in D_1 explores uncertain global structure, small variance in D_2 refines precise local details.

Through iterative noise estimation, we ultimately obtain the final dehazed result $I_{final} = \hat{Y}$.

To reveal the underlying significance of this operation, the denoising formula can be formally decomposed. First, starting from Y_t, we estimate Y_0 by rearranging the terms in Eq.1 as follows:

$$\hat{Y}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(Y_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta\left(Z, Y_t, \bar{\alpha}_t\right)\right). \tag{7}$$

Additionally, through algebraic manipulation and completing the square, we can express Y_{t-1} in terms of (Y_0, Y_t) as

$$Y_{t-1} = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}Y_t + \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t}Y_0 + \sqrt{\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}(1-\alpha_t)}\epsilon. \tag{8}$$

In Eq.8, Y_0 is explicitly expressed. Equation 8 aligns with Equation 3 when parameterizing the reverse process and deriving the variational lower bound for the log-likelihood of the reverse chain. By substituting Y_0 in Eq.8 with its expression from Eq.7, Eq.3 can be derived.

In Eq. 8, Y_{t-1} is determined by Y_t, Y_0, and the uncertainty term ϵ. By rearranging (Eq. 7), we estimate a distribution of Y_0 through "denoising" Y_t, which is initially biased. The uncertainty term perturbs this hippocampus-like intermediate trace $\hat{Y}_{0|t}$, introducing stochastic exploration akin to the hippocampus probing multiple possible memory trajectories.

The magnitude of this perturbation is illustrated by the red circles in Fig. 5. In D_1, repeated oscillatory perturbations guide the model to recover global structure under high uncertainty. In D_2, however, the up-sampled condition map Z provides a neocortex-like stable scaffold, so smaller, carefully tuned variance suffices to refine local details with minimal stochasticity, enabling precise reconstruction as shown in Fig. 5(b).

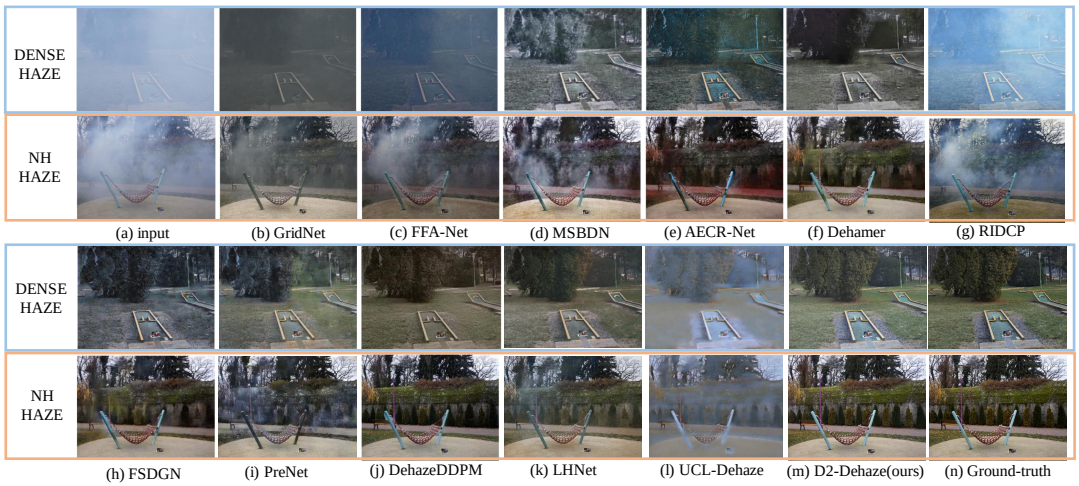


Figure 6: Comparison of visual results on NH-HAZE and DENSE-HAZE dataset.

Results and Discussion

Experimental Settings

Datasets To evaluate the performance of our method, we adopt four datasets. In real-world challenging scenes, we adopt two benchmark real-world datasets: Dense-Haze (Ancuti et al., 2019) and NH-HAZE (Ancuti et al., 2020). In order to enlarge the datasets and enhance the sensitivity of the model to color, we follow previous studies by performing data augmentation in HSV space (Yuan et al., 2023).

Implementation Details The entire network is trained on four NVIDIA TITAN RTX 24G in PyTorch. We employ the ADAM optimizer with a learning rate of 0.00002. The batch size is configured to 1, and the patch dimensions are set to 64 × 64. The scaling factor γ is assigned a value of 2.

Comparison with Baseline Methods

Quantity Comparison As shown in Table 1, hybrid CNN–ViT models such as LHNet and MITNet outperform earlier approaches, confirming the importance of integrating global context with local detail in dehazing. Diffusion-based DehazeDDPM further surpasses CNN-, ViT-, and GAN-based methods, highlighting the strong generative capacity of diffusion models. Our proposed D2-Dehaze achieves the best overall performance across nearly all metrics, demonstrating more faithful texture recovery, fewer artifacts, and competitive PSNR/SSIM scores.

Visual Comparison We further provide visual comparisons between our proposed method and several baseline approaches on real hazy images sampled from Dense-Haze and NH-HAZE datasets, as shown in Fig. 6. Across both Dense-Haze and NH-HAZE datasets, most existing methods leave residual haze, blur, or global color shifts, whereas our D2-Dehaze effectively removes haze, restores structural details, and produces results that are visually closest to the ground truth.

These results suggest that perceptual quality depends not only on representational power but also on the efficiency of the processing pathway. By decomposing the task into two cognitively efficient stages, D2-Dehaze avoids the inefficiencies of single-stage optimization and achieves a more balanced and robust reconstruction, reflecting how adaptive processing strategies shape effective perception across varying conditions.

Table 1: Quantitative comparison of different dehazing algorithms. ↓ means lower is better and ↑ denotes higher is better. Best results are in bold, second-best underlined.

	DENSE-HAZE			NH-HAZE		
Method	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑
AOD-Net(ICCV2017)	0.5991	13.14	0.4144	0.4947	15.40	0.5693
GridNet(ICCV2019)	0.5102	13.31	0.3681	0.3046	13.80	0.5370
FFA-Net(AAAI2020)	0.4976	14.39	0.4524	0.3653	19.87	0.6915
MSBDN(CVPR2020)	0.5358	15.37	0.4858	0.2918	19.23	0.7056
AECR-Net(CVPR2021)	0.5368	15.80	0.4660	0.2782	19.88	0.7073
Dehamer(CVPR2022)	0.4796	16.62	0.5602	0.2296	20.66	0.6844
FSDGN(ECCV2022)	0.4190	16.91	0.5806	0.2248	19.99	0.7106
RIDCP(CVPR2023)	0.5507	8.08	0.4173	0.3578	12.27	0.4996
DehazeDDPM(arXiv2023)	0.2994	19.04	0.5922	0.1624	22.28	0.7309
LHNet(MM2023)	0.4442	18.87	0.5608	0.1917	22.09	0.6940
MITNet(MM2023)	0.4596	19.97	0.6056	0.1982	21.26	0.7122
SLP(TIP2023)	0.5532	17.17	0.5651	0.3643	19.33	0.6258
UCL-Dehaze(TIP2024)	0.4879	16.31	0.5261	0.3927	16.13	0.7045
DEA-Net(TIP2024)	0.5167	17.24	0.5993	0.2092	20.56	0.7052
DiffLI2D(ECCV2024)	0.4060	16.97	0.5840	0.2170	20.29	0.7380
SGDN (AAAI2025)	-	**22.03**	0.6423	-	24.51	0.7973
D2-Dehaze (ours)	**0.1627**	21.19	**0.6503**	**0.1438**	**25.58**	**0.8252**

Ablation Study

Effectiveness of Two-stage Diffusion To evaluate the effectiveness of the two-stage diffusion strategy, we compare D2-Dehaze with the one-stage DDPM using PSNR and SSIM. As shown in Table 2, D2-Dehaze achieves significantly better performance.

Visual comparisons in Fig. 7 further show that, under the same number of diffusion steps, the one-stage model produces blurred and structurally incomplete results, whereas our two-stage approach yields clearer structures and finer de-

Table 2: Comparison of one-stage diffusion and two-stage diffusion.

Method	DENSE-HAZE		NH-HAZE	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑
DDPM	11.32	0.2135	12.12	0.2569
D2-Dehaze (ours)	**21.19**	**0.6503**	**25.58**	**0.8252**

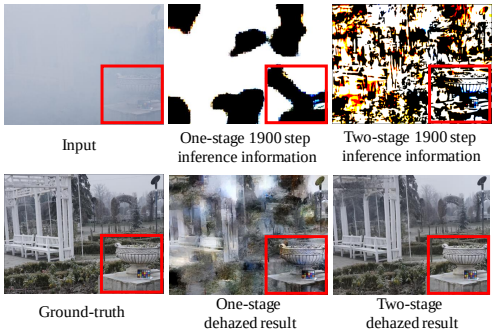


Figure 7: Comparison of one-stage diffusion and two-stage diffusion at different iterative steps.

tails. These results highlight how structuring the generative process into more efficient stages reduces representational inefficiencies and improves perceptual outcomes, reflecting how adaptive processing pathways shape more optimal and reliable perception.

Effectiveness of DGD The first stage of D2-Dehaze is a dehazing process on the down-sampled image. To verify the contribution of our DGD in this stage to the whole dehazing task, we compare D2-Dehaze (DGD+DDR) with methods LHNet+DDR, MITNet+DDR and UCL+DDR by replacing DGD with dehazing algorithms LHNet, MITNet and UCL. We use the DENSE-HAZE dataset, the results are reported in Table 3. It can be seen that, comparing with other combinations, D2-Dehaze (DGD+DDR) achieves better results on all the three metrics.

Table 3: Comparison of different methods for the two-stage D2-Dehaze framework.

Method	LPIPS ↓	PSNR ↑	SSIM ↑
First Stage: Dehazing			
LHNet+DDR	0.3321	19.29	0.6388
MITNet+DDR	0.4293	19.37	0.6413
UCL+DDR	0.5382	18.53	0.6223
Second Stage: Super-Resolution			
DGD+SRDiff	0.2575	20.56	0.6029
DGD+RankSRGAN	0.2548	20.57	0.6036
D2-Dehaze (ours)	**0.1627**	**21.19**	**0.6503**

Effectiveness of DDR The second stage detail recovery task is similar to an image super-resolution problem. To demonstrate the effectiveness of our DDR in this stage, we replace it with two image super-resolution algorithms, SRDiff (H. Li et al., 2022) and RankSRGAN (W. Zhang et al., 2019), and get two image dehazing methods, DGD+SRDiff and DGD+RankSRGAN. The comparison of D2-Dehaze and these two is shown in Table 3. It can be seen that D2-Dehaze (DGD+DDR) performs better than the other two combinations.

Investigation of parameter γ

The setting of the scaling factor parameter γ is crucial to the performance of our proposed D2-Dehaze method. To investigate the impact of the scaling parameter γ, we perform experiments on the NH-HAZE dataset using different γ values. We use distortion metrics PSNR and SSIM, and color shift metric CDiff for evaluation.

Figure 8: Visual results of D2-Dehaze with different factors.

In the experiment, $\gamma \in 0, 1, 2, 3$ controls the down-sampling factor, where $\gamma = 0$ denotes no down-sampling and higher values indicate scaling by 2^{γ}. As shown in Fig. 8, small γ values $(0, 1)$ lead to insufficient global abstraction and residual artifacts, while large γ (3) causes information loss and blurring. In contrast, $\gamma = 2$ achieves the best balance between global structure and local detail.

These findings highlight that perceptual performance depends not only on processing capacity but also on the efficiency of the processing pathway itself: both overly coarse and overly fine strategies can lead to perceptual inefficiencies or distortions. By regulating the abstraction scale, the model demonstrates a dynamic trade-off between efficiency and fidelity, mirroring how human cognition adapts processing strategies across tasks and contexts to redefine what constitutes "optimal" perception.

Conclusion

We present D2-Dehaze, a two-stage diffusion framework for single-image dehazing, inspired by neocortex–hippocampus interactions. The first stage abstracts global structure akin to hippocampal replay, while the second refines local details using a stable neocortical scaffold, avoiding one-stage inefficiencies. Experiments show consistent gains in perceptual and distortion metrics, robust color fidelity, and generalization. This framework illustrates how efficiency- and exploration-driven pathways shape perceptual reconstruction, offering a cognitively grounded view of "optimal" perception.

Acknowledgments

This work was supported by National Natural Science Foundation of China (No. 62476040).

References

Ancuti, C. O., Ancuti, C., Sbert, M., & Timofte, R. (2019). Dense-haze: A benchmark for image dehazing with dense-haze and haze-free images. *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 1014–1018. https://doi.org/10.1109/ICIP.2019.8803046

Ancuti, C. O., Ancuti, C., & Timofte, R. (2020). Nh-haze: An image dehazing benchmark with non-homogeneous hazy and haze-free images. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1798–1805. https://doi.org/10.1109/CVPRW50498.2020.00230

Berman, D., Treibitz, T., & Avidan, S. (2016). Non-local image dehazing. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1674–1682. https://doi.org/10.1109/CVPR.2016.185

Blau, Y., & Michaeli, T. (2018). The perception-distortion tradeoff. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6228–6237. https://doi.org/10.1109/CVPR.2018.00652

Cai, B., Xu, X., Jia, K., Qing, C., & Tao, D. (2016). Dehazenet: An end-to-end system for single image haze removal. *IEEE Transactions on Image Processing*, *25*(11), 5187–5198. https://doi.org/10.1109/TIP.2016.2598681

Chen, B., Zou, X., Zhang, Y., Li, J., Li, K., Xing, J., & Tao, P. (2024). Leformer: A hybrid cnn-transformer architecture for accurate lake extraction from remote sensing imagery. *Proceedings of ICASSP 2024 – IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5710–5714. https://doi.org/10.1109/ICASSP48485.2024.10446785

Chen, Y., Li, W., Sakaridis, C., Dai, D., & Van Gool, L. (2018). Domain adaptive faster r-cnn for object detection in the wild. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3339–3348. https://doi.org/10.1109/CVPR.2018.00352

Chen, Z., He, Z., & Lu, Z. (2024). Dea-net: Single image dehazing based on detail-enhanced convolution and content-guided attention. *IEEE Transactions on Image Processing*, *33*, 1002–1015. https://doi.org/10.1109/TIP.2024.3354108

Chen, Z., Zhu, Y., Zhao, C., Hu, G., Zeng, W., Wang, J., & Tang, M. (2021). Dpt: Deformable patch-based transformer for visual recognition. *Proceedings of the 29th ACM International Conference on Multimedia (MM)*, 2899–2907. https://doi.org/10.1145/3474085.3475467

Cong, X., Gui, J., Miao, K., Zhang, J., Wang, B., & Chen, P. (2020). Discrete haze level dehazing network. *Proceedings of the 28th ACM International Conference on Multimedia (MM '20)*, 1828–1836. https://doi.org/10.1145/3394171.3413876

Deck, K., & Bischoff, T. (2023). Easing color shifts in score-based diffusion models. *arXiv preprint arXiv:2306.15832*. https://doi.org/10.48550/arXiv.2306.15832

Delbracio, M., Talebei, H., & Milanfar, P. (2021). Projected distribution loss for image enhancement. *2021 IEEE International Conference on Computational Photography (ICCP)*, 1–12. https://doi.org/10.1109/ICCP51581.2021.9466271

Dhariwal, P., & Nichol, A. (2021). Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems (NeurIPS 34)*, 8780–8794. https://proceedings.neurips.cc/paper/2021/hash/878088f08a0c63efe193a9512023b907-Abstract.html

Dong, H., Pan, J., Xiang, L., Hu, Z., Zhang, X., Wang, F., & Yang, M. (2020). Multi-scale boosted dehazing network with dense feature fusion. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2154–2164. https://doi.org/10.1109/CVPR42600.2020.00223

Dong, J., & Pan, J. (2020). Physics-based feature dehazing networks. *Computer Vision – ECCV 2020*, 188–204. https://doi.org/10.1007/978-3-030-58577-8_11

Freirich, D., Michaeli, T., & Meir, R. (2021). A theory of the distortion-perception tradeoff in wasserstein space. *Advances in Neural Information Processing Systems (NeurIPS 34)*, 25661–25672. https://proceedings.neurips.cc/paper/2021/hash/2566125672f9e0c9a8c0f08a8d3a8d20-Abstract.html

Guo, C., Yan, Q., Anwar, S., Cong, R., Ren, W., & Li, C. (2022). Image dehazing transformer with transmission-aware 3d position embedding. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5802–5810. https://doi.org/10.1109/CVPR52688.2022.00572

He, K., Sun, J., & Tang, X. (2011). Single image haze removal using dark channel prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, *33*(12), 2341–2353. https://doi.org/10.1109/TPAMI.2010.168

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems (NeurIPS 30)*, 6626–6637. https://proceedings.neurips.cc/paper/2017/hash/66266637f24a8a0321284b8648d7a907-Abstract.html

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems (NeurIPS 33)*, 6840–6851. https://proceedings.neurips.cc/paper/2020/hash/68406851f08a0c63efe193a9512023b9-Abstract.html

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems (NeurIPS 25)*, 1106–1114. https://proceedings.neurips.cc/

paper/2012/hash/11061114f08a0c63efe193a9512023b9-Abstract.html

Kundu, S., & Ari, S. (2020). Mscnn: A deep learning framework for p300-based brain–computer interface speller. *IEEE Transactions on Medical Robotics and Bionics*, *2*(1), 86–93. https://doi.org/10.1109/TMRB.2019.2959559

Li, B., Peng, X., Wang, Z., Xu, J., & Feng, D. (2017). Aod-net: All-in-one dehazing network. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 4780–4788. https://doi.org/10.1109/ICCV.2017.511

Li, B., Peng, X., Wang, Z., Xu, J., & Feng, D. (2018). End-to-end united video dehazing and detection. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, *32*, 7016–7023. https://doi.org/10.1609/aaai.v32i1.12287

Li, B., Ren, W., Fu, D., Tao, D., Feng, D., Zeng, W., & Wang, Z. (2019). Benchmarking single-image dehazing and beyond. *IEEE Transactions on Image Processing*, *28*(1), 492–505. https://doi.org/10.1109/TIP.2018.2867951

Li, H., Yang, Y., Chang, M., Chen, S., Feng, H., Xu, Z., Li, Q., & Chen, Y. (2022). Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, *479*, 47–59. https://doi.org/10.1016/j.neucom.2022.01.029

Li, T., Liu, Y., Ren, W., Shiri, B., & Lin, W. (2025). Single image dehazing using fuzzy region segmentation and haze density decomposition. *IEEE Transactions on Circuits and Systems for Video Technology*, *35*(10), 9964–9978. https://doi.org/10.1109/TCSVT.2025.3558232

Liang, X., Li, R., & Tang, J. (2019). Selective attention network for image dehazing and deraining. *Proceedings of the 1st ACM International Conference on Multimedia in Asia (ACM MM in Asia)*, 1–6. https://doi.org/10.1145/3338533.3366688

Ling, P., Chen, H., Tan, X., Jin, Y., & Chen, E. (2023). Single image dehazing using saturation line prior. *IEEE Transactions on Image Processing*, *32*, 3238–3253. https://doi.org/10.1109/TIP.2023.3279980

Liu, J., Sun, H., & Katto, J. (2023). Learned image compression with mixed transformer-cnn architectures. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14388–14397. https://doi.org/10.1109/CVPR52729.2023.01383

Liu, X., Ma, Y., Shi, Z., & Chen, J. (2019). Griddehazenet: Attention-based multi-scale network for image dehazing. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 7313–7322. https://doi.org/10.1109/ICCV.2019.00741

Mittal, A., Moorthy, A. K., & Bovik, A. C. (2012). No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, *21*(12), 4695–4708. https://doi.org/10.1109/TIP.2012.2214050

Mittal, A., Soundararajan, R., & Bovik, A. C. (2013). Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, *20*(3), 209–212. https://doi.org/10.1109/LSP.2012.2227726

Qin, X., Wang, Z., Bai, Y., Xie, X., & Jia, H. (2020). Ffa-net: Feature fusion attention network for single image dehazing. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, *34*, 11908–11915. https://doi.org/10.1609/aaai.v34i07.6865

Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D. J., & Norouzi, M. (2023). Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(4), 4713–4726. https://doi.org/10.1109/TPAMI.2022.3204461

Sakaridis, C., Dai, D., Hecker, S., & Van Gool, L. (2018). Model adaptation with synthetic and real data for semantic dense foggy scene understanding. *Proceedings of the European Conference on Computer Vision (ECCV)*, 707–724. https://doi.org/10.1007/978-3-030-01261-8_42

Shen, H., Zhao, Z., Zhang, Y., & Zhang, Z. (2023). Mutual information-driven triple interaction network for efficient image dehazing. *Proceedings of the 31st ACM International Conference on Multimedia (MM)*, 7–16. https://doi.org/10.1145/3581783.3612299

Song, J., Meng, C., & Ermon, S. (2021). Denoising diffusion implicit models. *International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=St1giarCHLP

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., & Poole, B. (2021). Score-based generative modeling through stochastic differential equations. *International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=PxTIG12RRHS

Sun, Z., Zhang, Y., Bao, F., Wang, P., Yao, X., & Zhang, C. (2022). Sadnet: Semi-supervised single image dehazing method based on an attention mechanism. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, *18*(2), Art. 58, 1–23. https://doi.org/10.1145/3478457

Wang, C., Shen, H., Fan, F., Shao, M., Yang, C., Luo, J., & Deng, L. (2021). Eaa-net: A novel edge assisted attention network for single image dehazing. *Knowledge-Based Systems*, *228*, 107279. https://doi.org/10.1016/j.knosys.2021.107279

Wang, Y., Yan, X., Wang, F. L., Xie, H., Yang, W., Zhang, X., Qin, J., & Wei, M. (2024). Ucl-dehaze: Towards real-world image dehazing via unsupervised contrastive learning. *IEEE Transactions on Image Processing*, *33*, 1361–1374. https://doi.org/10.1109/TIP.2024.3362153

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, *13*(4), 600–612. https://doi.org/10.1109/TIP.2003.819861

Whang, J., Delbracio, M., Talebi, H., Saharia, C., Dimakis, A. G., & Milanfar, P. (2022). Deblurring via stochastic refinement. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16272–16282. https://doi.org/10.1109/CVPR52688.2022.01581

Wu, H., Qu, Y., Lin, S., Zhou, J., Qiao, R., Zhang, Z., Xie, Y., & Ma, L. (2021). Contrastive learning for compact single image dehazing. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10546–10555. https://doi.org/10.1109/CVPR46437.2021.01041

Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., & Zhang, L. (2021). Cvt: Introducing convolutions to vision transformers. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 22–31. https://doi.org/10.1109/ICCV48922.2021.00009

Wu, R., Duan, Z., Guo, C., Chai, Z., & Li, C. (2023). Ridcp: Revitalizing real image dehazing via high-quality codebook priors. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 22282–22291. https://doi.org/10.1109/CVPR52729.2023.02134

Yang, Z., Yu, H., Li, B., Zhang, J., Huang, J., & Zhao, F. (2025). Unleashing the potential of the semantic latent space in diffusion models for image dehazing. *Proceedings of the European Conference on Computer Vision (ECCV)*, 371–389. https://doi.org/10.1007/978-3-031-72784-9_21

Yu, H., Huang, J., Zheng, K., Zhou, M., & Zhao, F. (2023). High-quality image dehazing with diffusion model. *arXiv preprint arXiv:2308.11949*. https://doi.org/10.48550/arXiv.2308.11949

Yu, H., Zheng, N., Zhou, M., Huang, J., Xiao, Z., & Zhao, F. (2022). Frequency and spatial dual guidance for image dehazing. *European Conference on Computer Vision (ECCV)*, 181–198. https://doi.org/10.1007/978-3-031-19800-7_11

Yuan, S., Chen, J., Li, J., Jiang, W., & Guo, S. (2023). Lhnet: A low-cost hybrid network for single image dehazing. *Proceedings of the 31st ACM International Conference on Multimedia (MM)*, 7706–7717. https://doi.org/10.1145/3581783.3612594

Zhang, N., Nex, F., Vosselman, G., & Kerle, N. (2023). Lite-mono: A lightweight cnn and transformer architecture for self-supervised monocular depth estimation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18537–18546. https://doi.org/10.1109/CVPR52729.2023.01778

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 586–595. https://doi.org/10.1109/CVPR.2018.00068

Zhang, W., Liu, Y., Dong, C., & Qiao, Y. (2019). Ranksrgan: Generative adversarial networks with ranker for image super-resolution. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 3096–3105. https://doi.org/10.1109/ICCV.2019.00319

Zheng, Z., Ren, W., Cao, X., Hu, X., Wang, T., Song, F., & Jia, X. (2021). Ultra-high-definition image dehazing via multi-guided bilateral learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16180–16189. https://doi.org/10.1109/CVPR46437.2021.01592

Zhu, H., Peng, X., Chandrasekhar, V. R., Li, L., & Lim, J. (2018). Dehazegan: When image dehazing meets differential programming. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI)*, 1234–1240. https://doi.org/10.24963/ijcai.2018/172

Zhu, Q., Mai, J., & Shao, L. (2014). Single image dehazing using color attenuation prior. *Proceedings of the British Machine Vision Conference (BMVC)*, 1–12. https://doi.org/10.5244/C.28.114