Exploiting Diffusion Prior for Real-World Image Dehazing with Unpaired Training

Yunwei Lan[1,2], **Zhigao Cui**[1 *], **Chang Liu**[2], **Jialun Peng**[2], **Nian Wang**[1], **Xin Luo**[2], **Dong Liu**[2 *]

[1]Rocket Force University of Engineering
[2]University of Science and Technology of China
yunweilan@mail.ustc.edu.cn, cuizg10@126.com, lc980413@mail.ustc.edu.cn, pjl@mail.ustc.edu.cn,
nianwang04@outlook.com, xinluo@mail.ustc.edu.cn, dongeliu@ustc.edu.cn

Abstract

Unpaired training has been verified as one of the most effective paradigms for real scene dehazing by learning from unpaired real-world hazy and clear images. Although numerous studies have been proposed, current methods demonstrate limited generalization for various real scenes due to limited feature representation and insufficient use of real-world prior. Inspired by the strong generative capabilities of diffusion models in producing both hazy and clear images, we exploit diffusion prior for real-world image dehazing, and propose an unpaired framework named Diff-Dehazer. Specifically, we leverage diffusion prior as bijective mapping learners within the CycleGAN, a classic unpaired learning framework. Considering that physical priors contain pivotal statistics information of real-world data, we further excavate real-world knowledge by integrating physical priors into our framework. Furthermore, we introduce a new perspective for adequately leveraging the representation ability of diffusion models by removing degradation in image and text modalities, so as to improve the dehazing effect. Extensive experiments on multiple real-world datasets demonstrate the superior performance of our method.

Introduction

Under hazy conditions, the quality of images is severely degraded. Such degradation significantly affects the visual appeal of images, and causes information loss, further restricting their performance in other downstream tasks, e.g., object detection (Huang, Le, and Jaw 2020). Therefore, image dehazing, which aims to restore clear images from hazy ones, has been extensively studied in the past decade.

Following conventional studies, we normally formulate the hazy image with the Atmospheric Scattering Model (ASM), which can be written as:

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

where I represents the hazy image and J refers to its corresponding clear image. A and t denote the atmospheric light and transmission map, where both of them are usually unknown. Following the formulation in Eq. (1), early dehazing methods (He, Sun, and Tang 2010; Meng et al. 2013) try to estimate A and t with hand-crafted physical priors, and restore hazy images into clear ones by reversing the ASM. Even so, these methods normally obtain over-saturated results, since hand-crafted physical priors are not universally compatible with all scenes. With the advancement of deep learning, prevailing methods (Zheng et al. 2023; Song et al. 2023; Qiu et al. 2023) aim to design neural networks to model physical parameters, or directly restore images in an end-to-end manner, as illustrated in Fig. 1 (a). These methods primarily utilize synthetic paired data for training since obtaining real-world hazy and clear image pairs is virtually impossible. Although improved performance is observed, such paired training paradigm often fails to generalize to real-world dehazing scenarios, due to the ill-presenting performance of trained networks that lack real-world information from hazy images.

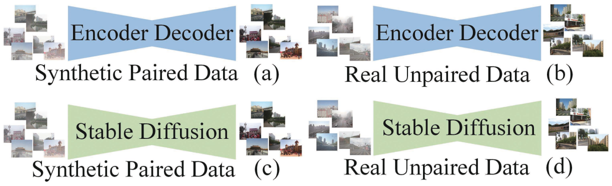

Figure 1: (a) Previous dehazing methods with paired training. (b) Previous CycleGAN-based dehazing methods with unpaired training. (c) Existing stable diffusion-based dehazing methods with paired training. (d) Our stable diffusion-based dehazing method with unpaired training.

To tackle the bottleneck of paired training, recent studies (Zhao et al. 2021; Yang et al. 2022) consider the paradigm of unpaired training for image dehazing, so as to discover available information from real-world hazy images and model the particular mapping between real-world hazy and clear images. The key problem of unpaired training is how to impose a structure-consistent constraint between hazy and dehazed images, so that the influence of misaligned information caused by unpaired real clear images can be suppressed. CycleGAN (Zhu et al. 2017) offers a classic solution by leveraging a framework to maintain the consistency between the mapped domain and the original one. Such paradigm, as shown in Fig. 1 (b), is latter adopted by several studies, e.g., CycleDehaze (Engin, Genç, and Kemal Ekenel 2018), D4

[*]Corresponding Author

(Yang et al. 2022), and ODCR (Wang et al. 2024), along with promising real-world dehazing performance. However, these methods struggle to achieve effective representation by a limited number of training images, resulting in sub-optimal performance. Inspired by the strong representation capabilities of diffusion models, some methods (Liu et al. 2024c; Lin et al. 2024) retort to the pre-training and fine-tuning mode, which leverages diffusion prior to improve the dehazing effect, as illustrated in Fig. 1 (c). Nevertheless, these methods still have limitations in several real-world scenarios due to their reliance on synthetic paired training data that fails to simulate real scenes.

To address the aforementioned problems, we propose an effective paradigm for real-world image dehazing, namely Diff-Dehazer. As shown in Fig. 1 (d), we improve image dehazing with both diffusion prior and the unpaired training paradigm. We argue that simply inheriting the unpaired training framework will exhibit sub-optimal dehazing performance without accounting for the physical properties of real-world hazy scenes. Consequently, we conduct a comprehensive investigation on the physical background of image dehazing and integrate physical priors into our framework. Furthermore, We observe that text description can improve the resulting image by offering enriched high-level semantics, which has validated its effectiveness in previous studies. Therefore, we adopt a multi-modal paradigm, incorporating text and image to improve the dehazing effect.

Particularly, our contributions are four-fold: 1) We adopt a pre-trained stable diffusion as the foundation of our CycleGAN framework, so as to leverage its strong representation ability in modeling real-world data. 2) We boost the generalization ability of image dehazing with physical priors, whose potentials are greatly neglected in previous studies, and propose Physics-Aware Guidance (PAG) in our framework. 3) We utilize the enriched high-level semantics stored in text modality, and propose Text-Aware Guidance (TAG) to bootstrap textual guidance for image dehazing. 4) To facilitate further studies in this topic, we construct an unpaired real-world dataset, consisting of $6,519$ hazy images and $11,293$ clear images. Extensive experiments on existing benchmark datasets illustrate the superior performance of our method compared to state-of-the-art methods.

Related Works

Image Dehazing

Early dehazing methods extract physical priors from the statistical properties of natural images, and then restore hazy images into clear ones via the ASM. For example, DCP (He, Sun, and Tang 2010) proposes the dark channel prior to estimate the transmission map and atmospheric light. BCCR (Meng et al. 2013) explores inherent boundary constraint and L1-norm-based contextual regularization to optimize the transmission map. Nevertheless, these methods are heuristic for real-world image dehazing. They usually lead to over-enhanced results, since hand-crafted priors fail to fit the complexity and diversity of real-world haze. With the development of deep learning, latter studies concentrate on designing networks for image dehazing. For instance, C2PNet (Zheng et al. 2023) customizes a circular learning strategy for image dehazing. Some methods (Guo et al. 2022; Song et al. 2023; Qiu et al. 2023) perform image dehazing based on the Transformer (Vaswani et al. 2017) architecture. Even so, all the aforementioned methods still struggle to handle several dehazing cases, especially the ones under real-world scenarios, due to their reliance on synthetic paired data.

To address these limitations, RefineDNet (Zhao et al. 2021) designs a weakly supervised two-stage dehazing framework. Based on CycleGAN (Zhu et al. 2017), D4 (Yang et al. 2022) introduces a self-augmented dehazing framework to decompose the ASM. Differently, ODCR (Wang et al. 2024) proposes orthogonal decoupling contrastive regularization to improve the dehazing results. Although improved performance is observed by these methods, they often fail to demonstrate effectiveness in real-world image dehazing due to insufficient feature representation of conventional encoder-decoder architecture.

Diffusion Model-Based Image Restoration

Recent advancements in diffusion models (Song, Meng, and Ermon 2021; Rombach et al. 2022; Liu et al. 2024a; Chang Liu and Dong Liu 2023; Liu et al. 2024b) have shown superior performance in various vision tasks, including Text-to-Image (T2I) generation (Ruiz et al. 2023) and conditional image generation (Zhang, Rao, and Agrawala 2023). Meanwhile, there is an increasing trend towards the application of diffusion models in low-level vision tasks. For example, StableSR (Wang, Chan, and Loy 2023) utilizes prior knowledge encapsulated in the stable diffusion for blind image restoration. PASD (Yang et al. 2023) introduces a pixel-aware cross-attention module to enable the stable diffusion to perceive local image structures for image super-resolution. Relying on the attribute prior in the pre-trained model, PTG-RM (Xu et al. 2024) designs an additional lightweight module to refine the results of a target restoration network. For image dehazing, RSHazeDiff (Xiong et al. 2024) proposes a unified Fourier-aware diffusion model for remote sensing image dehazing based on DDPM (Ho, Jain, and Abbeel 2020). Diff-Plugin (Liu et al. 2024c) proposes a lightweight task plugin to provide task-specific priors, guiding the diffusion process for image restoration.

Despite efforts to leverage off-the-shelf features from diffusion models, the exploration of image dehazing is still insufficient, since all aforementioned methods rely on the use of synthetic data and neglect the vitalness of real-world ones, which fail to deal with real-world hazy images.

Method

Following the CycleGAN, we establish a hazing-dehazing cycle for unpaired training on real-world data, along with hazing and dehazing processes. During the training, the real hazy image x is transformed into a fake clear image and then transformed back into a cycle hazy image $\hat{x}$, as depicted in Fig. 2. A similar process is applied to the real clear image y in reverse order. Specifically, the hazing process comprises a hazing backbone and Text-Aware Guidance (TAG). The dehazing process consists of three components: the dehazing

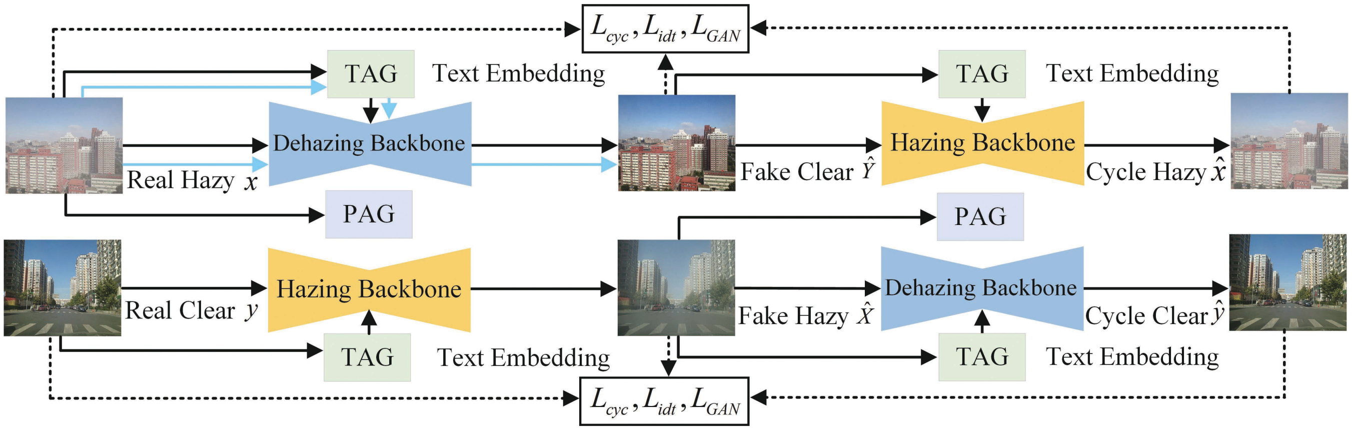

Figure 2: Overview of our method. Black arrows represent the training process and blue arrows denote the inference process.

backbone, Physics-Aware Guidance (PAG), and TAG. We train the framework using cycle-consistent constraints. After training, we can obtain a clear image from a hazy one using only the dehazing process (see the blue arrows in Fig. 2). Details of the aforementioned components are illustrated in the following parts.

Backbone Network

We use SD Turbo (v2.1) (Sauer et al. 2023; Parmar et al. 2024), a distilled version of Stable Diffusion (SD) 2.1 as hazing and dehazing backbones, since it allows us to generate a large number of high-quality images within one step. As shown in Fig. 3 (a), this network comprises three main components: the encoder and decoder of VAE (Kingma and Welling 2013), and the U-Net (Ronneberger, Fischer, and Brox 2015). To efficiently leverage the diffusion prior encapsulated in the SD Turbo, we train it using LoRA (Hu et al. 2022) adapters instead of starting from scratch. Specifically, we only update the input layer of the backbone network as well as the additional LoRA adapters while keeping the remaining model parameters frozen.

Previous SD-based image restoration methods typically begin with mapping the image from pixel to latent space via pre-trained VAEs (Esser, Rombach, and Ommer 2021), where the restoration process is then performed in the VAE latent space. Once the VAE features are generated by the diffusion model, the restored image is converted back to pixel space via the decoder of VAE. However, these methods are not directly applicable to image dehazing since performing dehazing in a highly compressed space inevitably leads to significant loss of image information. Consequently, the final restored images exhibit noticeably lower fidelity and significant discrepancies from the original ones in multiple aspects, e.g., regional details and textures. To preserve the details of the source image during the dehazing process, we implement a skipped connection between the encoder and decoder of the VAE.

Text-Aware Guidance

Textual feature is proven to offer enriched high-level semantics for generative models, and has already demonstrated its effectiveness in stable diffusion (Rombach et al. 2022). Based on this finding, we expect to further integrate the textual information into our framework, so as to enhance the dehazing process with textual features. In doing so, we propose Text-Aware Guidance (TAG), with its illustration shown in Fig. 3 (b). Specifically, we first employ a pre-trained image captioner, i.e., BLIP-2 (Li et al. 2023), to produce a caption for the input image, where the extracted caption is then utilized in further processes. Note that users can manually customize the text description according to their needs when inferencing any real-world hazy images.

Different from the explicit use of text descriptions in current methods, we introduce TAG from the following two perspectives. On the one hand, we refine these captions by eliminating haze-related terms, e.g., "*haze*", "*fog*", etc., so as to explicitly facilitate image dehazing in the text modality. On the other hand, we leverage both positive and negative prompts via classifier-free guidance (Ho and Salimans 2021) to improve the quality of the dehazed image, with the refined caption serving as the positive prompt. Considering that the hazing attribute might be too complicated to describe with explicit words, we learn a prompt by textual inversion (Gal et al. 2022) from thousands of real-world hazy images to holistically depict the hazing attribute implicitly. With acquired positive and negative prompts, we achieve more comprehensive semantic guidance via the text encoder of CILP and further boost the image quality. As for the hazing process, we first obtain the image caption directly through BLIP-2, and then combine the haze-related terms with the extracted caption, considering it as a positive prompt. For the negative prompt, we assign an empty text.

Physics-Aware Guidance

Previous studies have demonstrated that using physical priors is more likely to remove haze, particularly in real-world scenarios since these priors are statistical laws from a large number of real-world clear images. Therefore, we integrate physical priors into our framework and introduce Physics-Aware Guidance (PAG) for image dehazing, as illustrated in Fig. 3 (c). Different from previous methods, we investigate various physical priors and integrate two well-performing

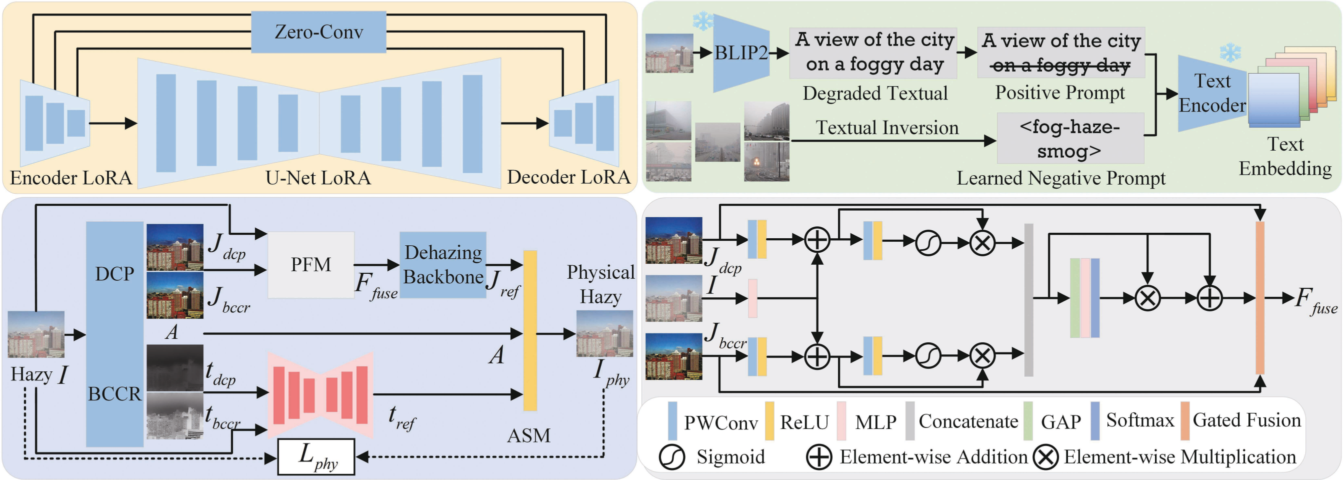

Figure 3: Orange Area: Backbone network of our framework. Note that hazing and dehazing backbones share the same U-Net and employ two individual VAEs. Green Area: Text-Aware Guidance (TAG). Blue Area: Physics-Aware Guidance (PAG). Dark Channel Prior (DCP) and Boundary Constraint and Contextual Regularization (BCCR) are two physical prior-based dehazing methods. ASM is the Atmospheric Scattering Model. Gray Area: The structure of Perceptual Fusion Model (PFM).

(i.e., DCP and BCCR) into the framework. After obtaining preliminary dehazed result J, atmospheric light A, and transmission map t in Eq. (1), we reconstruct the hazy image via the ASM and design a physical loss. By fine-tuning the dehazing backbone, the model adheres to the underlying physical principles, achieving more effective dehazing and physical awareness. We provide more details on how we obtain preliminary results in our supplementary materials.

Reconstruction of Hazy Image To leverage the physical priors encapsulated in clear images dehazed by DCP and BCCR, we treat them as clear images for hazy image reconstruction, thereby compelling the model to learn more about the physical properties of real-world haze. Observing that either J_{dcp} or J_{bccr} is better than the other in some regions under various application scenarios, we further refine them rather than applying them directly. Specifically, for t_{dcp} and t_{bccr}, we concatenate and feed them into a U-Net to obtain a refined transmission map t_{ref}. Given that the input image contains source information, we take it as input and feed it into the U-Net along with the transmission map. To combine the advantages of J_{dcp} and J_{bccr}, we introduce the Perceptual Fusion Model (PFM) to effectively fuse them, resulting in a composite image F_{fuse}. Within our framework, the dehazing backbone is capable of not only restoring clear images from hazy inputs but also enhancing the image quality of clear inputs. Consequently, we consider the dehazing backbone as a refined network and feed J_{fuse} into it to obtain a refined clear image J_{ref} with enhanced natural details and physical awareness. In doing so, we can reconstruct a more qualified hazy image I_{phy} using refined J_{ref}, t_{ref}, and A, imposing a more reasonable physical constraint.

The structure of the PFM is shown in Fig. 3 (d). We utilize point-wise convolution layers, ReLU, and MLP to extract latent features from J_{dcp}, J_{bccr}, and I. To preserve any information potentially lost in the initial dehazed images, we add the feature of I to those of J_{dcp} and J_{bccr}. Guided features are obtained using point-wise convolution layers, ReLU, Sigmoid, and residual connections. Subsequently, we concatenate and re-weight them using global average pooling, MLP, and Softmax. The concatenated features are then multiplied by the obtained weight, followed by a residual connection to produce coarse-fused features. Finally, we re-weight the proportions of J_{dcp} and J_{bccr} via a gated fusion (Chen et al. 2019), resulting in fine-fused results J_{fuse}.

Loss Function

Following CycleGAN (Zhu et al. 2017), we design the training loss for the proposed model as:

$$L = L_{cyc} + \lambda_{phy} L_{phy} + \lambda_{idt} L_{idt} + \lambda_{GAN} L_{GAN}, \quad (2)$$

where L_{cyc}, L_{phy}, L_{idt}, and L_{GAN} represent the cycle consistency loss, prior loss, identity loss, and GAN loss, respectively. λ_{phy}, λ_{idt}, and λ_{GAN} are the corresponding hyperparameters to control the weight of L_{phy}, L_{idt}, and L_{GAN}. We present more details of L_{cyc}, L_{idt}, and L_{GAN} in our supplementary materials since they are similar to CycleGAN.

Physical Loss For our model, due to the use of physical priors, we design the physical loss as:

$$L_{phy} = L_{rec}(I, I_{phy}), \quad (3)$$

where $I_{phy} = J_{ref} t_{ref} + A(1 - t_{ref})$ refers to the reconstructed hazy image derived by the ASM. I denotes the overall notation involving both real/fake hazy images. L_{rec} is the combined distance of L_1 and LPIPS (Zhang et al. 2018).

Experiments and Discussions

Datasets

We collect over $7,000$ real-world hazy images from RESIDE (Li et al. 2018), and select over $10,000$ clear images

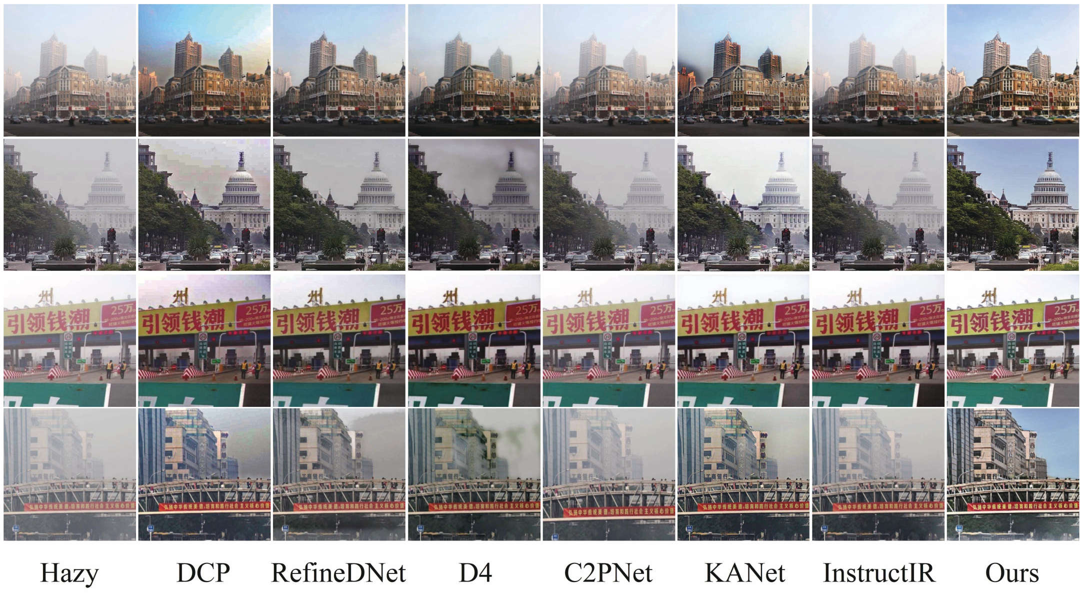

Figure 4: Visual comparison of samples from Haze2020 and RTTS. Our method can effectively remove haze and generate high-quality images with natural color and realistic contrast. More visual results are presented in our supplementary materials.

	RTTS				Haze2020			
	FID ↓	NIQE ↓	MUSIQ ↑	CLIPIQA ↑	FID ↓	NIQE ↓	MUSIQ ↑	CLIPIQA ↑
DCP	73.017	4.271	52.935	0.276	91.948	3.827	54.375	0.364
BCCR	75.984	4.289	52.429	0.270	92.798	3.765	54.438	0.362
RefineDNet (TIP2021)	65.076	4.012	56.117	0.257	88.229	3.972	54.779	0.350
PSD (CVPR2021)	73.847	4.017	56.430	0.272	91.922	3.868	58.522	0.257
Dehamer (CVPR2022)	66.208	4.882	53.787	0.365	84.416	4.118	54.528	0.419
D4 (CVPR2022)	69.400	4.637	58.445	0.351	87.536	3.971	53.458	0.309
DehazeFormer (TIP2023)	68.636	4.693	53.692	0.352	82.842	4.208	55.590	0.418
C2PNet(CVPR2023)	66.117	5.037	53.960	**0.390**	83.959	4.206	54.565	0.423
KANet(TPAMI2024)	64.963	4.339	54.513	0.285	84.888	3.742	56.411	0.376
InstructIR (ECCV2024)	66.278	4.902	54.464	0.369	83.561	4.164	55.080	0.422
Diff-Plugin (CVPR2024)	65.787	5.334	50.740	0.361	80.965	4.407	52.752	0.424
Ours	**52.344**	**3.943**	**59.207**	0.328	**71.984**	**3.571**	**62.273**	**0.439**

Table 1: Quantitative results on RTTS and Haze2020. The best results are denoted in **bold**.

from ADE20K (Zhou et al. 2019) as well as OTS (a subset of RESIDE) to construct hazy and clear images as our training set, respectively.

To evaluate the effectiveness of our proposed method, we conduct qualitative and quantitative experiments on various real-world image datasets, including URHI, RTTS, Haze2020, OHAZE (Ancuti et al. 2018), NHAZE (Ancuti, Ancuti, and Timofte 2020), and the proposed dataset in Fattal (2014). Specifically, URHI and RTTS are two subsets of RESIDE that contain over $4,000$ real-world hazy images. Haze2020 consists of over $1,000$ hazy images selected by DA (Shao et al. 2020). OHAZE and NHAZE contain 45 and 55 pairs of outdoor hazy images, respectively, which are artificially generated using a haze machine. The dataset proposed in Fattal (2014) contains 37 real-world hazy images.

Comparisons with State-of-the-Art Methods

We compare the performance of our method with several state-of-the-art methods, including DCP (He, Sun, and Tang 2010), BCCR (Meng et al. 2013), RefineDNet (Zhao et al. 2021), PSD (Chen et al. 2021), Dehamer (Guo et al. 2022), D4 (Yang et al. 2022), DehazeFormer (Song et al. 2023), C2PNet (Zheng et al. 2023), KANet (Feng et al. 2024), InstructIR (Conde, Geigle, and Timofte 2024), and Diff-Plugin (Liu et al. 2024c). Notably, DCP and BCCR are physical prior-based methods. RefineDNet and D4, along with our method, do not require paired data for training and are categorized as weakly supervised methods. The remaining methods are fully supervised methods and necessitate paired data for training. For images with available ground truth, we evaluate these methods using full-reference metrics such as PSNR, SSIM (Wang et al. 2004), LPIPS, and VSI (Zhang,

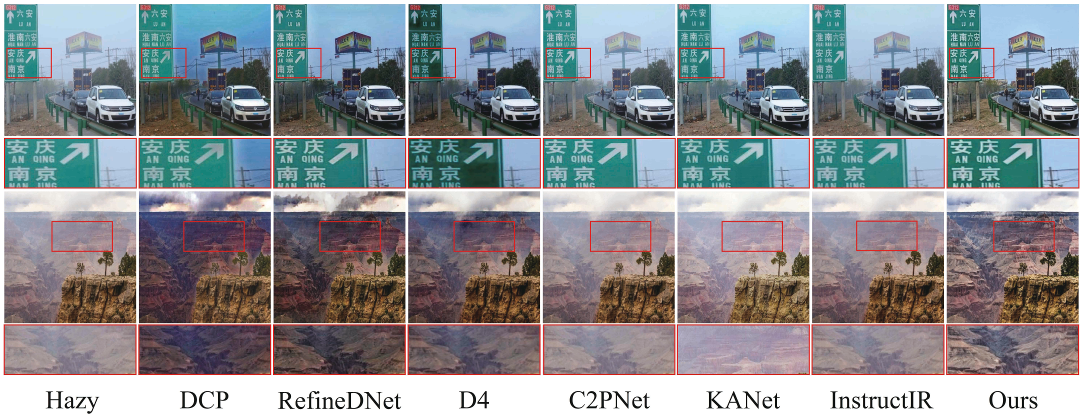

Figure 5: Visual comparison of samples from Haze2020. Areas where our method works better are boxed out and zoomed in. Our method can generate clear images with high fidelity and discriminative textures.

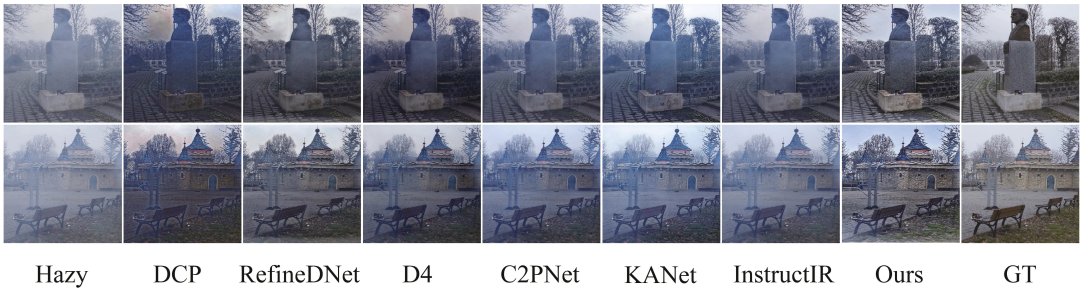

Figure 6: Visual comparison of samples from OHAZE.

	OHAZE			
	PSNR ↑	SSIM ↑	LPIPS ↓	VSI ↑
DCP	17.005	0.819	0.242	0.945
BCCR	15.487	0.753	0.261	0.943
RefineDNet	18.693	0.755	0.260	**0.954**
PSD	14.727	0.717	0.316	0.930
Dehamer	17.827	0.688	0.330	0.927
D4	16.767	0.690	0.290	0.939
DehazeFormer	15.243	0.672	0.339	0.911
C2PNet	18.014	0.708	0.317	0.929
KANet	17.713	0.814	0.265	0.939
InstructIR	18.616	0.716	0.321	0.931
Diff-Plugin	16.470	0.526	0.364	0.920
Ours	**19.539**	**0.825**	**0.195**	**0.954**

Table 2: Quantitative results on OHAZE. The best results are denoted in **bold**.

Shen, and Li 2014). In cases where ground truth is unavailable, we evaluate them by no-reference metrics, including FID (Heusel et al. 2017), CLIPIQA (Wang, Chan, and Loy 2023), NIQE (Mittal, Soundararajan, and Bovik 2012), and MUSIQ (Ke et al. 2021).

Results on RTTS and Haze2020. Fig. 4 shows a qualitative comparison of the results on RTTS and Haze2020. The physical prior-based method DCP effectively processes the images. However, due to the inherent limitations of physical priors, it results in the inevitable over-enhancement of the images, producing darker results with lower visual quality. Unpaired dehazing methods, such as RefineDNet and D4, can dehaze to a certain extent as they do not require paired synthetic datasets for training. However, these methods fail to thoroughly dehaze due to limited feature representation and insufficient use of physical priors, resulting in images with residual artifacts and insufficient details. MB-TaylorFormer and C2PNet demonstrate exceptional performance on synthetic images through well-designed networks and training strategies, but they still face challenges with real-world hazy images and even fall short of the efficacy of a physical prior-based method. Similarly, recently proposed all-in-one image restoration methods have only been demonstrated to be effective on synthetic images. When dealing with real-world hazy images, these methods (e.g., Instruct IR) are almost ineffective. The proposed method outperforms other state-of-the-art methods and can generate more realistic and natural images with rich details and textures, as well as high contrast.

The quantitative results of RTTS and Haze2020 are shown in Table 1. In RTTS, our method exhibits superior performance in FID, NIQE, and MUSIQ scores. For CLIPIQA, our method is ranked lower. This can be attributed to the integration of physical priors. However, our intention of using

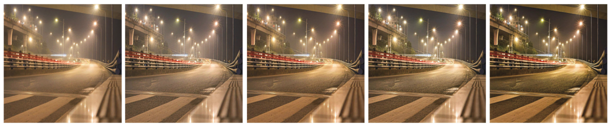

Figure 7: Ablation study of each component. From left to right, respectively, the hazy image, the result of **Backbone**, the result of **Backbone+PAG**, the result of **Backbone+TAG**, and the result of **Ours**.

Backbone	✓	✓	✓	✓
PAG		✓		✓
TAG			✓	✓
FID	74.575	73.096	74.100	**71.984**
NIQE	3.783	3.762	3.574	**3.571**
MUSIQ	60.606	60.571	61.882	**62.273**
CLIPIQA	0.427	0.420	0.436	**0.439**

Table 3: Ablation study of each component.

physical priors is to limit the stochastics of the stable diffusion to some extent, ensuring that the resulting dehazed image is more natural while not generating something that does not match the original hazy image. Therefore, it is justifiable to sacrifice a certain amount of CLIPIQA. In Haze2020, our method surpasses others across various metrics.

To further illustrate the superior performance of our proposed method, we zoom in on local details within selected images and show them in Fig. 5. As we can see, our method not only effectively restores the textures of images but also maintains the image fidelity.

Results on OHAZE In Table 2 and Fig. 6, we present the qualitative and quantitative results of OHAZE. Note that we do not need to retrain the network when evaluating this dataset. Our method exhibits superior performance in visualization and quantitative metrics compared to other methods, further highlighting its strong generalization ability, as it can effectively process real-world hazy images with various types and haze densities, producing high-quality results.

The results of NHAZE, URHI, and the dataset proposed in Fattal (2014) are presented in our supplementary materials. Moreover, We compare model efficiency and object detection accuracy in our supplementary materials.

Ablation Study

We conduct a series of ablation experiments to verify the effectiveness of each key component. Specifically, we discuss the backbone network, PAG, and TAG. We first establish four variants: **Backbone:** Removing TAG and PAG. **Backbone + PAG:** Removing TAG. **Backbone + TAG:** Removing PAG. **Ours:** The full model of our method. As shown in Table 3 and Fig. 7, the proposed method achieves the best performance in metrics and visual appeals, which validates that each component plays a critical role in our framework. Moreover, we conduct an ablation study to analyze the effectiveness of the backbone network supplementary materials.

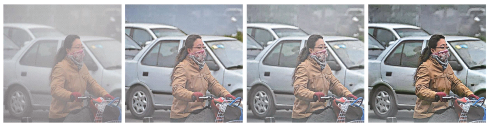

Figure 8: Ablation study of positive and negative prompts. From left to right, respectively, the hazy image, the result with no prompt, the result with only a positive prompt, and the result with both positive and negative prompts.

Baseline	✓	✓	✓	✓
DCP		✓		✓
BCCR			✓	✓
FID	74.100	76.768	74.999	**71.984**
NIQE	3.574	**3.534**	3.574	3.571
MUSIQ	61.882	61.585	61.399	**62.273**
CLIPIQA	0.436	0.414	0.420	**0.439**

Table 4: Ablation study of physical priors.

Analysis of Text-Aware Guidance We conduct an ablation study to validate how TAG works within our method. Firstly, we evaluate the effectiveness of positive and negative prompts. As illustrated in Fig. 8, using a positive text prompt alleviates the generation of mismatched details during the dehazing process, which is caused by the inherent stochastics of the stable diffusion. The employment of a negative text prompt can further improve the dehazing effect. Additionally, we demonstrate the impact of the guidance scale and present the results in our supplementary materials.

Analysis of Physics-Aware Guidance We demonstrate how physical priors in our framework affect the results. We establish some variants as follows: **Baseline:** the full model of our method without PAG. **Baseline + DCP:** Removing the use of BCCR. **Baseline + BCCR:** Removing the use of DCP. **Ours:** We adopt both DCP and BCCR to constrain the network training. Table 4 demonstrates that our method yields superior results from a comprehensive perspective. Additionally, we validate the impact of the weight of physical loss and provide the results in our supplementary materials.

Conclusion

In this paper, we delve into the potential of prior encapsulated in the stable diffusion and physical priors derived from natural images, thereby proposing an unpaired framework for real-world image dehazing named Diff-Dehazer. Furthermore, by leveraging enriched high-level semantics contained in text, we perform dehazing in text and image modalities to get more qualified results. Extensive experiments have validated the superiority of our method. However, due to our fine-tuning paradigm, our method may suffer from the inherent problems of diffusion models, which interact with noises and may produce diversified images. This may cause misalignment in some cases with severe haze. Despite this, we perform stably on most cases.

References

Ancuti, C. O.; Ancuti, C.; and Timofte, R. 2020. NH-HAZE: An image dehazing benchmark with non-homogeneous hazy and haze-free images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 444–445.

Ancuti, C. O.; Ancuti, C.; Timofte, R.; and De Vleeschouwer, C. 2018. 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*, 754–762.

Chang Liu and Dong Liu. 2023. Late-Constraint Diffusion Guidance for Controllable Image Synthesis. arXiv:2305.11520.

Chen, D.; He, M.; Fan, Q.; Liao, J.; Zhang, L.; Hou, D.; Yuan, L.; and Hua, G. 2019. Gated context aggregation network for image dehazing and deraining. In *2019 IEEE winter conference on applications of computer vision (WACV)*, 1375–1383. IEEE.

Chen, Z.; Wang, Y.; Yang, Y.; and Liu, D. 2021. PSD: Principled synthetic-to-real dehazing guided by physical priors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7180–7189.

Conde, M. V.; Geigle, G.; and Timofte, R. 2024. High-quality image restoration following human instructions. *arXiv preprint arXiv:2401.16468*.

Engin, D.; Genç, A.; and Kemal Ekenel, H. 2018. Cycle-dehaze: Enhanced cyclegan for single image dehazing. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 825–833.

Esser, P.; Rombach, R.; and Ommer, B. 2021. Taming Transformers for High-Resolution Image Synthesis. In *CVPR*, 12873–12883.

Fattal, R. 2014. Dehazing using color-lines. *ACM transactions on graphics (TOG)*, 34(1): 1–14.

Feng, Y.; Ma, L.; Meng, X.; Zhou, F.; Liu, R.; and Su, Z. 2024. Advancing real-world image dehazing: perspective, modules, and training. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Guo, C.-L.; Yan, Q.; Anwar, S.; Cong, R.; Ren, W.; and Li, C. 2022. Image dehazing transformer with transmission-aware 3d position embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5812–5820.

He, K.; Sun, J.; and Tang, X. 2010. Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence*, 33(12): 2341–2353.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Ho, J.; and Salimans, T. 2021. Classifier-Free Diffusion Guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*.

Hu, E. J.; yelong shen; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Huang, S.-C.; Le, T.-H.; and Jaw, D.-W. 2020. DSNet: Joint semantic learning for object detection in inclement weather conditions. *IEEE transactions on pattern analysis and machine intelligence*, 43(8): 2623–2633.

Ke, J.; Wang, Q.; Wang, Y.; Milanfar, P.; and Yang, F. 2021. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 5148–5157.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Li, B.; Ren, W.; Fu, D.; Tao, D.; Feng, D.; Zeng, W.; and Wang, Z. 2018. Benchmarking single-image dehazing and beyond. *IEEE Transactions on Image Processing*, 28(1): 492–505.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Lin, J.; Zhang, Z.; Wei, Y.; Ren, D.; Jiang, D.; Tian, Q.; and Zuo, W. 2024. Improving image restoration through removing degradations in textual representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2866–2878.

Liu, C.; Li, R.; Zhang, K.; Lan, Y.; and Liu, D. 2024a. StableV2V: Stablizing Shape Consistency in Video-to-Video Editing. arXiv:2411.11045.

Liu, C.; Xu, S.; Peng, J.; Zhang, K.; and Liu, D. 2024b. Towards Interactive Image Inpainting via Robust Sketch Refinement. *TMM*, 9973–9987.

Liu, Y.; Ke, Z.; Liu, F.; Zhao, N.; and Lau, R. W. 2024c. Diff-Plugin: Revitalizing Details for Diffusion-based Low-level Tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4197–4208.

Meng, G.; Wang, Y.; Duan, J.; Xiang, S.; and Pan, C. 2013. Efficient image dehazing with boundary constraint and contextual regularization. In *Proceedings of the IEEE international conference on computer vision*, 617–624.

Mittal, A.; Soundararajan, R.; and Bovik, A. C. 2012. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3): 209–212.

Parmar, G.; Park, T.; Narasimhan, S.; and Zhu, J.-Y. 2024. One-step image translation with text-to-image models. *arXiv preprint arXiv:2403.12036*.

Qiu, Y.; Zhang, K.; Wang, C.; Luo, W.; Li, H.; and Jin, Z. 2023. Mb-taylorformer: Multi-branch efficient transformer expanded by taylor formula for image dehazing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12802–12813.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 234–241. Springer.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 22500–22510.

Sauer, A.; Lorenz, D.; Blattmann, A.; and Rombach, R. 2023. Adversarial diffusion distillation. *arXiv preprint arXiv:2311.17042*.

Shao, Y.; Li, L.; Ren, W.; Gao, C.; and Sang, N. 2020. Domain adaptation for image dehazing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2808–2817.

Song, J.; Meng, C.; and Ermon, S. 2021. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.

Song, Y.; He, Z.; Qian, H.; and Du, X. 2023. Vision transformers for single image dehazing. *IEEE Transactions on Image Processing*, 32: 1927–1941.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, J.; Chan, K. C.; and Loy, C. C. 2023. Exploring CLIP for Assessing the Look and Feel of Images. In *AAAI*, volume 37, 2555–2563.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Wang, Z.; Zhao, H.; Peng, J.; Yao, L.; and Zhao, K. 2024. ODCR: Orthogonal Decoupling Contrastive Regularization for Unpaired Image Dehazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 25479–25489.

Xiong, J.; Yan, X.; Wang, Y.; Zhao, W.; Zhang, X.-P.; and Wei, M. 2024. RSHazeDiff: A Unified Fourier-aware Diffusion Model for Remote Sensing Image Dehazing. *arXiv preprint arXiv:2405.09083*.

Xu, X.; Kong, S.; Hu, T.; Liu, Z.; and Bao, H. 2024. Boosting Image Restoration via Priors from Pre-trained Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2900–2909.

Yang, T.; Ren, P.; Xie, X.; and Zhang, L. 2023. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. *arXiv preprint arXiv:2308.14469*.

Yang, Y.; Wang, C.; Liu, R.; Zhang, L.; Guo, X.; and Tao, D. 2022. Self-augmented unpaired image dehazing via density and depth decomposition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2037–2046.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.

Zhang, L.; Shen, Y.; and Li, H. 2014. VSI: A visual saliency-induced index for perceptual image quality assessment. *IEEE Transactions on Image processing*, 23(10): 4270–4281.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhao, S.; Zhang, L.; Shen, Y.; and Zhou, Y. 2021. RefineDNet: A weakly supervised refinement framework for single image dehazing. *IEEE Transactions on Image Processing*, 30: 3391–3404.

Zheng, Y.; Zhan, J.; He, S.; Dong, J.; and Du, Y. 2023. Curricular contrastive regularization for physics-aware single image dehazing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5785–5794.

Zhou, B.; Zhao, H.; Puig, X.; Xiao, T.; Fidler, S.; Barriuso, A.; and Torralba, A. 2019. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127: 302–321.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232.