SelfPromer: Self-Prompt Dehazing Transformers with Depth-Consistency

Cong Wang[1]*, Jinshan Pan[2], Wanyu Lin[1], Jiangxin Dong[2], Wei Wang[3], Xiao-Ming Wu[1]

[1]Department of Computing, The Hong Kong Polytechnic University
[2]School of Computer Science and Engineering, Nanjing University of Science and Technology
[3]International School of Information Science and Engineering, Dalian University of Technology

Abstract

This work presents an effective depth-consistency Self-Prompt Transformer, terms as SelfPromer, for image dehazing. It is motivated by an observation that the estimated depths of an image with haze residuals and its clear counterpart vary. Enforcing the depth consistency of dehazed images with clear ones, therefore, is essential for dehazing. For this purpose, we develop a prompt based on the features of depth differences between the hazy input images and corresponding clear counterparts that can guide dehazing models for better restoration. Specifically, we first apply deep features extracted from the input images to the depth difference features for generating the prompt that contains the haze residual information in the input. Then we propose a prompt embedding module that is designed to perceive the haze residuals, by linearly adding the prompt to the deep features. Further, we develop an effective prompt attention module to pay more attention to haze residuals for better removal. By incorporating the prompt, prompt embedding, and prompt attention into an encoder-decoder network based on VQGAN, we can achieve better perception quality. As the depths of clear images are not available at inference, and the dehazed images with one-time feed-forward execution may still contain a portion of haze residuals, we propose a new continuous self-prompt inference that can iteratively correct the dehazing model towards better haze-free image generation. Extensive experiments show that our SelfPromer performs favorably against the state-of-the-art approaches on both synthetic and real-world datasets in terms of perception metrics including NIQE, PI, and PIQE. The source codes will be made available at https://github.com/supersupercong/SelfPromer.

Introduction

Recent years have witnessed advanced progress in image dehazing due to the development of deep dehazing models. Mathematically, the haze process is usually modeled by an atmospheric light scattering model (He, Sun, and Tang 2011) formulated as:

$$\mathrm{I}(x) = \mathrm{J}(x)\mathrm{T}(x) + (1 - \mathrm{T}(x))\mathrm{A}, \quad (1)$$

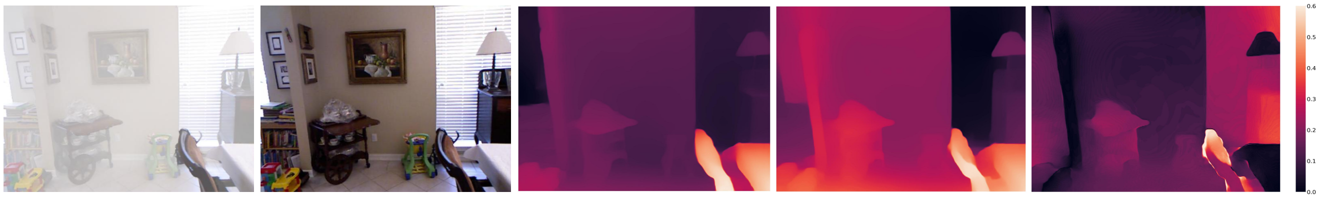
where I and J denote a hazy and haze-free image, respectively, and A denotes the global atmospheric light, x denotes the pixel index, and the transmission map T is usually modeled as $\mathrm{T}(x) = e^{-\beta \mathrm{d}(x)}$ with the scene depth $\mathrm{d}(x)$, and the scattering coefficient β reflects the haze density.

Most existing works develop various variations of deep Convolutional Neural Networks (CNNs) for image dehazing (Liu et al. 2019a; Dong et al. 2020; Dong and Pan 2020; Chen et al. 2021; Jin et al. 2022, 2023; Wang et al. 2023). They typically compute a sequence of features from the hazy input images and directly reconstruct the clear ones based on the features, which have achieved state-of-the-art results on benchmarks (Li et al. 2019) in terms of PSNRs and SSIMs. However, as dehazing is ill-posed, very small errors in the estimated features may degrade the performance. Existing works propose to use deep CNNs as image priors and then restore the clear images iteratively. However, they cannot effectively correct the errors or remove the haze residuals in the dehazed images as these models are fixed in the iterative process (Liu et al. 2019b). It is noteworthy that the human visual system generally possesses an intrinsic correction mechanism that aids in ensuring optimal results for a task. This phenomenon has been a key inspiration behind the development of a novel dehazing approach incorporating a correction mechanism that guides deep models toward better haze-free results generation.

Specifically, if a dehazed result exists haze residuals, a correction mechanism can localize these regions and guide the relevant task toward removing them. Notably, NLP-based text prompt learning has shown promise in guiding the models by correcting the predictions (Liu et al. 2023). However, text-based prompts may not be appropriate for tasks that require solely visual inputs without accompanying text. Recent works (Herzig et al. 2022; Gan et al. 2023) attempted to address this issue by introducing text-free prompts into vision tasks. For instance, PromptonomyViT (Herzig et al. 2022) evaluates the adaptation of multi-task prompts such as depth, normal, and segmentation to improve the performance of the video Transformers. Nevertheless, these prompts may not be suitable for image dehazing tasks, as they could not capture the haze-related content.

To better guide the deep model for better image dehazing, this work develops an effective self-prompt dehazing Transformer. Specifically, it explores the depth consistency of hazy images and their corresponding clear ones as a prompt. In particular, our study is motivated by the substan-

[*] supercong94@gmail.com.



(a) Hazy (b) Clear (c) Hazy depth map (d) Clear depth map (e) Difference map

Figure 1: Haze residuals pose a significant challenge to accurately estimating the depth of clear images, creating inconsistencies compared to hazy images. A difference map (e) is utilized to locate haze residuals on the estimated depth, while minimal haze residuals will result in consistent estimates. By analyzing the difference map, we can identify the impact of haze residuals, leading to the development of improved dehazing models to mitigate this effect and enhance the quality of dehazed images. The difference map (e) is derived by |hazy depth − clear depth| with equalization for better visualization.

tial difference between the estimated depths of hazy images and their clear counterparts, i.e., the same scene captured in the same location should be consistent regarding depth. Depth is typically related to the transmission map in the atmospheric light scattering model as shown in Eq. (1). Thus, if the dehazed images can be reconstructed accurately, their estimated depths should be close to those of their clear counterparts at large. However, haze residuals often degrade the accuracy of depth estimation, resulting in significant differences between hazy and clear images, as illustrated in Fig. 1(e). Yet, the difference map of estimated depths from images with haze residuals and clear images often points to the regions affected by haze residuals.

Based on the above observation, we design a prompt to guide the deep models for perceiving and paying more attention to haze residuals. Our prompt is built upon the estimated feature-level depth differences, of which the inconsistent regions can reveal haze residual locations for deep model correction. On top of the prompt, we introduce a prompt embedding module that linearly combines input features with the prompt to better perceive haze residuals. Further, we propose a prompt attention module that employs self-attention guided by the prompt to pay more attention to haze residuals for better haze removal. Our encoder-decoder architecture combines these modules using VQGAN (Esser, Rombach, and Ommer 2021) to enhance the perception quality of the results, as opposed to relying solely on PSNRs and SSIMs metrics for evaluation.

As the depths of clear images suffer from unavailability at inference and dehazed images obtained via one-time feed-forward execution may have haze residuals, we introduce a continuous self-prompt inference to address these challenges. Specifically, our proposed approach feeds the hazy input image to the model and sets the depth difference as zero to generate clearer images that serve as the clear counterpart. The clear image participates in constructing the prompt to conduct prompt dehazing. The inference operation is continuously conducted as the depth differences can keep correcting the deep dehazing models toward better clean image generation.

This paper makes the following contributions:

- We make the first attempt to formulate the prompt by considering the cues of the estimated depth differences between the image with haze residuals and its clear counterpart in the image dehazing task.
- We propose a prompt embedding module and a prompt attention module to respectively perceive and pay more attention to haze residuals for better removal.
- We propose a new continuous self-prompt inference approach to iteratively correct the deep models toward better haze-free image generation.

The Proposed SelfPromer

Our SelfPromer comprises two branches: the prompt branch and the self-prompt dehazing Transformer branch. The prompt branch generates a prompt by using the deep depth difference and deep feature extracted from the hazy input. The other branch exploits the generated prompt to guide the deep model for image dehazing. We incorporate a prompt embedding module and prompt attention module to perceive and pay more attention to the haze residuals for better removal. The proposed modules are formulated into an encoder-decoder architecture based on VQGAN for better perception quality (Zhou et al. 2022; Chen et al. 2022).

Overall Framework

Fig. 2 illustrates our method at the training stage. Given a hazy images I, we first utilize trainable encoder $Enc(\cdot)$ to extract features:

$$\mathbf{F}_{\text{Enc}} = Enc(\text{I}). \tag{2}$$

Then, we compute the depth difference of the hazy image I and its corresponding clear image J in feature space:

$$\text{D}_1 = DE(\text{I});\ \text{D}_2 = DE(\text{J}), \tag{3a}$$

$$\mathbf{F}_{\text{D}_1} = Enc_{\text{pre}}^{\text{frozen}}(\text{D}_1);\ \mathbf{F}_{\text{D}_2} = Enc_{\text{pre}}^{\text{frozen}}(\text{D}_2), \tag{3b}$$

$$\mathbf{F}_{\text{D}_{\text{diff}}} = |\mathbf{F}_{\text{D}_1} - \mathbf{F}_{\text{D}_2}|, \tag{3c}$$

where $DE(\cdot)$ denotes the depth estimator[1] (Ranftl et al. 2022). $Enc_{\text{pre}}^{\text{frozen}}(\cdot)$ denotes the pre-trained VQGAN encoder which is frozen when training our dehazing models.

[1] We chose DPT_Next_ViT_L_384 to balance accuracy, speed, and model size: https://github.com/isl-org/MiDaS.

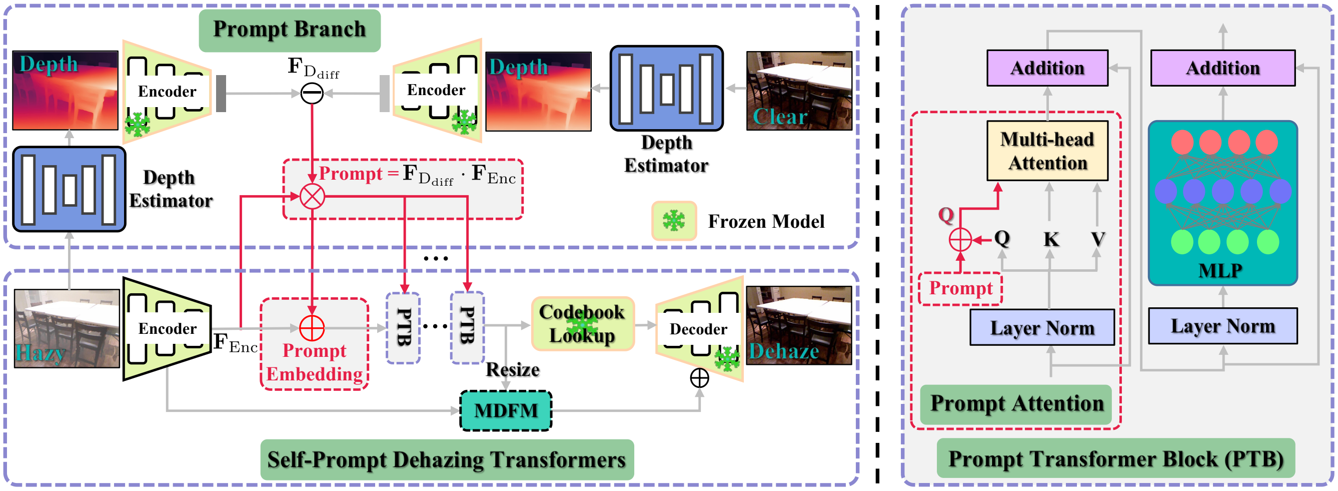


Figure 2: SelfPromer at training stage. Our method comprises two branches: prompt branch and self-prompt dehazing Transformer branch. The prompt branch generates a prompt by using the deep depth difference and deep feature extracted from the hazy input. The other branch exploits the generated prompt to guide the deep model for image dehazing. We incorporate a prompt embedding module and prompt attention module to perceive and pay more attention to the haze residuals for better removal. The proposed modules are formulated into an encoder-decoder architecture based on VQGAN for better perception quality. MDFM is detailed in Eq. (11). The inference is illustrated in Fig. 3.

Next, we exploit $\mathbf{F}_{\mathrm{D_{diff}}}$ to build the Prompt, and develop a prompt embedding module and a prompt attention module in Transformers, i.e., *PTB* to better generate haze-aware features:

$$\mathrm{Prompt} = \mathbf{F}_{\mathrm{D_{diff}}} \cdot \mathbf{F}_{\mathrm{Enc}}, \quad \text{\# Prompt} \tag{4a}$$

$$\mathbf{F}_{\mathrm{ProEmbed}} = \mathrm{Prompt} + \mathbf{F}_{\mathrm{Enc}}, \quad \text{\# Prompt Embedding} \tag{4b}$$

$$\mathbf{F}_{\mathrm{PTB}} = \mathit{PTB}(\mathrm{Prompt}, \mathbf{F}_{\mathrm{ProEmbed}}), \quad \text{\# Prompt Transformer} \tag{4c}$$

where $\mathbf{F}_{\mathrm{ProEmbed}}$ means the features of prompt embedding.

The generated feature $\mathbf{F}_{\mathrm{PTB}}$ is further matched with the learned haze-free $\texttt{Codebook}$ at the pre-trained VQGAN stage by the *Lookup* method (Esser, Rombach, and Ommer 2021; Zhou et al. 2022):

$$\mathbf{F}_{\mathrm{mat}} = \mathit{Lookup}(\mathbf{F}_{\mathrm{PTB}}, \texttt{Codebook}). \tag{5}$$

Finally, we reconstruct the dehazing images $\bar{\mathrm{J}}$ from the matched features $\mathbf{F}_{\mathrm{mat}}$ by decoder of pre-trained VQGAN $\mathit{Dec}_{\mathrm{pre}}^{\mathrm{frozen}}(\cdot)$ with residual learning (Chen et al. 2022) by mutual deformable fusion module *MDFM*:

$$\bar{\mathrm{J}} = \mathit{Dec}_{\mathrm{pre}}^{\mathrm{frozen}}(\mathbf{F}_{\mathrm{mat}}) + \mathit{MDFM}\Big(\mathbf{F}_{\mathrm{Enc}}^{\mathrm{s}}, \mathbf{F}_{\mathrm{PTB}}^{\mathrm{s,u}}\Big), \tag{6}$$

where $\mathbf{F}_{\mathrm{Enc}}^{\mathrm{s}}$ means the encoder features at s scale, while $\mathbf{F}_{\mathrm{PTB}}^{\mathrm{s,u}}$ denotes the s× upsampling features of *PTB*. We conduct the residual learning with MDFM in $\{1, 1/2, 1/4, 1/8\}$ scales between the encoder and decoder like FeMaSR (Chen et al. 2022). Here, $\mathbf{F}_{\mathrm{Enc}}^{1/8}$ denotes the $\mathbf{F}_{\mathrm{Enc}}$ in Eq. (2).

Loss Functions. We use pixel reconstruction loss $\mathcal{L}_{\mathrm{rec}}$, codebook loss $\mathcal{L}_{\mathrm{code}}$, perception loss $\mathcal{L}_{\mathrm{per}}$, and adversarial loss $\mathcal{L}_{\mathrm{adv}}$ to measure the error between the dehazed images $\bar{\mathrm{J}}$ and the corresponding ground truth J:

$$\mathcal{L} = \mathcal{L}_{\mathrm{rec}} + \lambda_{\mathrm{code}}\mathcal{L}_{\mathrm{code}} + \lambda_{\mathrm{per}}\mathcal{L}_{\mathrm{per}} + \lambda_{\mathrm{adv}}\mathcal{L}_{\mathrm{adv}}, \tag{7}$$

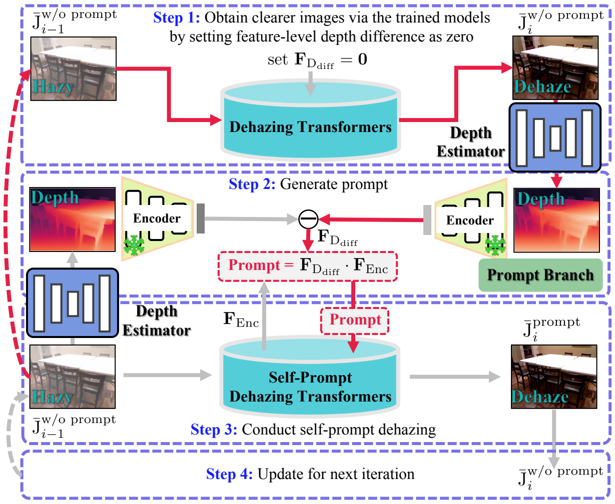

$\bar{\mathrm{J}}_{i-1}^{\text{w/o prompt}}$
Step 1: Obtain clearer images via the trained models by setting feature-level depth difference as zero
set $\mathbf{F}_{\mathrm{D_{diff}}} = \mathbf{0}$
$\bar{\mathrm{J}}_{i}^{\text{w/o prompt}}$
Hazy
Dehazing Transformers
Dehaze
Depth Estimator
Step 2: Generate prompt
Depth
Encoder
$\mathbf{F}_{\mathrm{D_{diff}}}$
Encoder
Depth
Prompt = $\mathbf{F}_{\mathrm{D_{diff}}} \cdot \mathbf{F}_{\mathrm{Enc}}$
Prompt Branch
Depth Estimator
$\mathbf{F}_{\mathrm{Enc}}$
Prompt
$\bar{\mathrm{J}}_{i}^{\text{prompt}}$
Hazy
Self-Prompt Dehazing Transformers
Dehaze
$\bar{\mathrm{J}}_{i-1}^{\text{w/o prompt}}$
Step 3: Conduct self-prompt dehazing
Step 4: Update for next iteration
$\bar{\mathrm{J}}_{i}^{\text{w/o prompt}}$

Figure 3: Continuous Self-Prompt Inference. i^{th} prompt inference contains four steps: Sequential execution from top to bottom. The magenta line describes the 'self' process that builds the prompt from the hazy image itself.

where

$$\mathcal{L}_{\mathrm{rec}} = ||\bar{\mathrm{J}} - \mathrm{J}||_1 + \lambda_{\mathrm{ssim}}\big(1 - \mathit{SSIM}(\bar{\mathrm{J}}, \mathrm{J})\big), \tag{8a}$$

$$\mathcal{L}_{\mathrm{code}} = ||\bar{z}_{\mathbf{q}} - z_{\mathbf{q}}||_2^2, \tag{8b}$$

$$\mathcal{L}_{\mathrm{per}} = ||\Phi(\bar{\mathrm{J}}) - \Phi(\mathrm{J})||_2^2, \tag{8c}$$

$$\mathcal{L}_{\mathrm{adv}} = \mathbb{E}_{\mathrm{J}}[\log \mathcal{D}(\mathrm{J})] + \mathbb{E}_{\bar{\mathrm{J}}}[1 - \log \mathcal{D}(\bar{\mathrm{J}})], \tag{8d}$$

where $SSIM(\cdot)$ denotes the structural similarity (Wang et al. 2004) for better structure generation. $z_{\mathbf{q}}$ is the haze-free codebook features by feeding haze-free images J to pre-trained VQGAN while $\bar{z}_{\mathbf{q}}$ is the reconstructed codebook features. $\Phi(\cdot)$ denotes the feature extractor of VGG19 (Simonyan and Zisserman 2015). $\mathcal{D}$ is the discriminator (Zhu et al. 2017). $\lambda_{\text{code}}, \lambda_{\text{per}}, \lambda_{\text{adv}}$, and λ_{ssim} are weights.

For **inference**, we propose a new self-prompt inference approach as our training stage involves the depth of clear images to participate in forming the prompt while clear images are not available at testing.

Self-Prompt Transformers

The self-prompt Transformer contains the prompt generated by the prompt branch, a prompt embedding module, and a prompt attention module which is contained in the prompt Transformer block. In the following, we introduce the definition of the prompt, prompt embedding module, and prompt attention module, and prompt Transformer block in detail.

Prompt (Definition). The prompt is based on the estimated depth difference between the input image and its clear counterpart. It is defined in Eq. (4a) which can better contain haze residual features as $\mathbf{F}_{\text{D}_{\text{diff}}}$ with higher response value reveals inconsistent parts which potentially correspond to the haze residuals in the input hazy image.

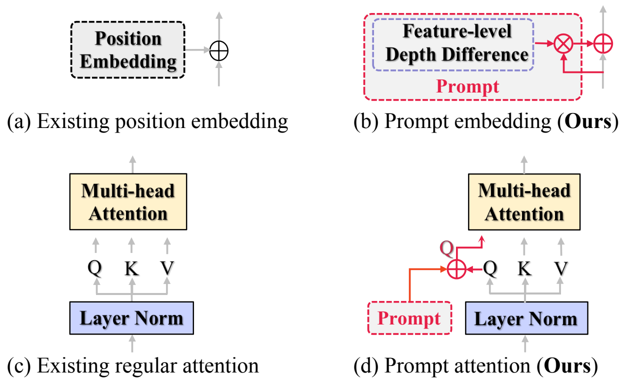
Prompt Embedding. Existing Transformers (Zheng et al. 2022) usually use the position embedding method (Fig. 4(a)) to represent the positional correlation, which does not contain haze-related information so that it may not effectively perceive the haze residual information well. Moreover, image restoration requires processing different input sizes at inference while the position embedding is defined with fixed parameters at training (Zheng et al. 2022). Hence, position embedding may be not a good choice for image dehazing. To overcome these problems, we propose prompt embedding which is defined in Eq. (4b). By linearly adding the extracted features $\mathbf{F}_{\text{Enc}}$ with Prompt, the embedded feature $\mathbf{F}_{\text{ProEmbed}}$ perceives the haze residual features as Prompt extracts the haze residual features. Note that as $\mathbf{F}_{\text{ProEmbed}}$ has the same size as $\mathbf{F}_{\text{Enc}}$, it does not require fixed sizes like position embedding.

Prompt Attention. Existing Transformers usually extract Query $\mathbf{Q}$, Key $\mathbf{K}$, and Value $\mathbf{V}$ from input features to estimate scaled-dot-product attention shown in Fig. 4(c). Although Transformers are effective for feature representation, the standard operation may be not suitable for image dehazing. To ensure the Transformers pay more attention to haze residuals for better removal, we propose prompt attention $ProAtt(\cdot)$ by linearly adding the query with Prompt:

$$\mathbf{Q} = \mathbf{Q} + \mathrm{Prompt}, \tag{9a}$$

$$ProAtt(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = Softmax\Big(\frac{\mathbf{Q}\mathbf{K}^{\mathrm{T}}}{\sqrt{d_{\text{head}}}}\Big)\mathbf{V}, \tag{9b}$$

where d_{head} means the number of the head. We set d_{head} as 8 in this paper. Fig. 4(d) illustrates the proposed prompt attention. Note that as $\mathbf{Q}$ in attention is to achieve the similarity relation for expected inputs (Ding et al. 2021), our prompt attention by linearly adding the prompt Prompt with the Query $\mathbf{Q}$ can pay more attention to haze residuals for better removal.



(a) Existing position embedding (b) Prompt embedding (**Ours**)

(c) Existing regular attention (d) Prompt attention (**Ours**)

Figure 4: (a)-(b) Existing position embedding vs. Prompt embedding (Ours). Our prompt embedding can better perceive haze information. (c)-(d) Existing regular attention vs. Prompt attention (Ours). Our prompt attention can pay more attention to haze residuals.

Prompt Transformer Block. According to the above attention design, our prompt Transformer block (PTB) can be sequentially computed as:

$$\mathbf{Q}, \mathbf{K}, \mathbf{V} = LN(\mathbf{X}^{l-1}), \tag{10a}$$

$$\hat{\mathbf{X}}^{l} = ProAtt(\mathbf{Q}, \mathbf{K}, \mathbf{V}) + \mathbf{X}^{l-1}, \tag{10b}$$

$$\mathbf{X}^{l} = MLP\Big(LN(\hat{\mathbf{X}}^{l})\Big) + \hat{\mathbf{X}}^{l}, \tag{10c}$$

where $\mathbf{X}^{l-1}$ and $\mathbf{X}^{l}$ mean the input and output of the l^{th} prompt Transformer block. Specially, $\mathbf{X}^{0}$ is the $\mathbf{F}_{\text{ProEmbed}}$. LN and MLP denote the layer normalization and multilayer perceptron. The PTB is shown in the right part of Fig. 2.

It is worth noting that our prompt embedding and prompt attention are flexible as we can manually set $\mathbf{F}_{\text{D}_{\text{diff}}} = \mathbf{0}$, the network thus automatically degrade to the model without prompt, which will be exploited to form our continuous self-prompt inference.

Mutual Deformable Fusion Module

As VQGAN is less effective for preserving details (Gu et al. 2022; Chen et al. 2022), motivated by the deformable models (Dai et al. 2017; Zhu et al. 2019) that can better fuse features, we propose a mutual deformable fusion module (MDFM) by fusing features mutually to adaptively learn more suitable offsets for better feature representation:

$$\text{off}_1 = Conv\Big(\mathcal{C}[\mathbf{F}^{\text{s}}_{\text{Enc}}, \mathbf{F}^{\text{s,u}}_{\text{PTB}}]\Big); \text{off}_2 = Conv\Big(\mathcal{C}[\mathbf{F}^{\text{s,u}}_{\text{PTB}}, \mathbf{F}^{\text{s}}_{\text{Enc}}]\Big), \tag{11a}$$

$$\mathbf{Y}_1 = DMC(\mathbf{F}^{\text{s}}_{\text{Enc}}, \text{off}_1); \mathbf{Y}_2 = DMC(\mathbf{F}^{\text{s,u}}_{\text{PTB}}, \text{off}_2), \tag{11b}$$

$$\mathbf{F}_{\text{MDFM}} = Conv\Big(\mathcal{C}[\mathbf{Y}_1, \mathbf{Y}_2]\Big), \tag{11c}$$

where $Conv(\cdot)$, $\mathcal{C}[\cdot]$, and $DMC(\cdot)$ respectively denote the 1×1 convolution, concatenation, and deformable convolution. off_k $(k = 1, 2.)$ denotes the estimated offset.

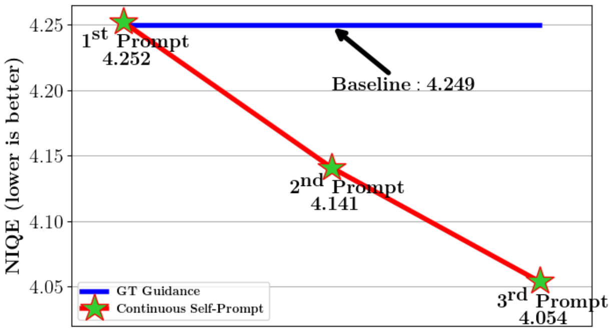


Figure 5: Continuous self-prompt inference vs. GT guidance (Baseline) on the SOTS-indoor dataset. GT guidance means we use the GT image to participate in forming the prompt at inference like the process of the training stage, which serves as the baseline.

Continuous Self-Prompt Inference

Our model requires the depth of clear images during training, but these images are unavailable at inference. Additionally, dehazed images generated by a one-time feed-forward execution may still contain some haze residuals. To address these issues, we propose a continuous self-prompt inference approach that leverages prompt embedding and prompt attention through linear addition. By setting feature-level depth difference $\mathbf{F}_{\mathrm{D_{diff}}}$ to zero, we can feed hazy images to our trained network and obtain clearer dehazed results which participate in building the prompt to conduct prompt dehazing. The iterative inference is conducted to correct the deep models to ensure the deep models toward better haze-free image generation:

$$\bar{\mathbf{J}}_i^{\text{w/o prompt}} = \mathcal{N}^{\text{w/o prompt}}(\bar{\mathbf{J}}_{i-1}^{\text{w/o prompt}}),\ \text{set}\ \mathbf{F}_{\mathrm{D_{diff}}} = \mathbf{0},\ \text{\# Step 1} \quad (12a)$$

$$\text{Prompt} = \mathbf{F}_{\mathrm{D_{diff}}} \cdot \mathbf{F}_{\mathrm{Enc}}; \mathbf{F}_{\mathrm{Enc}} = \mathit{Enc}(\bar{\mathbf{J}}_{i-1}^{\text{w/o prompt}}),\ \text{\# Step 2} \quad (12b)$$

$$\bar{\mathbf{J}}_i^{\text{prompt}} = \mathcal{N}^{\text{prompt}}(\bar{\mathbf{J}}_{i-1}^{\text{w/o prompt}}, \text{Prompt}),\ \text{\# Step 3} \quad (12c)$$

$$\bar{\mathbf{J}}_i^{\text{w/o prompt}} = \bar{\mathbf{J}}_i^{\text{prompt}},\quad (i = 1, 2, \cdots),\ \text{\# Step 4} \quad (12d)$$

where $\mathcal{N}^{\text{w/o prompt}}$ denotes our trained network without prompt by setting $\mathbf{F}_{\mathrm{D_{diff}}}$ as zero, while $\mathcal{N}^{\text{prompt}}$ means our trained network with prompt. $\mathbf{F}_{\mathrm{D_{diff}}} = |\mathit{Enc}_{\text{pre}}^{\text{frozen}}(\mathit{DE}(\bar{\mathbf{J}}_{i-1}^{\text{w/o prompt}})) - \mathit{Enc}_{\text{pre}}^{\text{frozen}}(\mathit{DE}(\bar{\mathbf{J}}_i^{\text{w/o prompt}}))|$. $\bar{\mathbf{J}}_0^{\text{w/o prompt}}$ denotes the original hazy images, while $\bar{\mathbf{J}}_{i-1}^{\text{w/o prompt}}$ is regarded as the image with haze residuals and $\bar{\mathbf{J}}_i^{\text{w/o prompt}}$ in Eq. (12a) is regarded as the clear counterpart of $\bar{\mathbf{J}}_i^{\text{w/o prompt}}$. $\bar{\mathbf{J}}_i^{\text{prompt}}$ means the i^{th} prompt dehazing results.

According to Eq. (12), the inference is a **continuous self-prompt** scheme, i.e., we get the clear images from the hazy image itself by feeding it to $\mathcal{N}^{\text{w/o prompt}}$ to participate in producing the prompt and the inference is continuously conducted. Fig. 3 better illustrates the inference process.

Fig. 5 shows our continuous self-prompt at 2nd and 3rd prompts outperforms the baseline which uses ground-truth (GT) to participate in forming the prompt like the process of the training stage.

Experiments

In this section, we evaluate the effectiveness of our method against state-of-the-art ones (SOTAs) on commonly used benchmarks and illustrate the effectiveness of the key components in the proposed method.

Implementation Details. We use 10 PTBs, i.e., $l = 10$, in our model. The details about the VQGAN are presented in the supplementary materials. We crop an image patch of 256 × 256 pixels. The batch size is 10. We use ADAM (Kingma and Ba 2015) with default parameters as the optimizer. The initial learning rate is 0.0001 and is divided by 2 at 160K, 320K, and 400K iterations. The model training terminates after 500K iterations. The weight parameters λ_{code}, λ_{per}, λ_{adv}, and λ_{ssim} are empirically set as 1, 1, 0.1, and 0.5. Our implementation is based on the PyTorch.

Synthetic Datasets. Following the protocol of (Yang et al. 2022), we use the RESIDE ITS (Li et al. 2019) as our training dataset and the SOTS-indoor (Li et al. 2019) and SOTS-outdoor (Li et al. 2019) as the testing datasets.

Real-world Datasets. Li et al. (2019) collect large-scale real-world hazy images, called UnannotatedHazyImages. We use these images as a real-world hazy dataset.

Evaluation Metrics. As we mainly aim to recover images with better perception quality, we use widely-used **NIQE** (Mittal, Soundararajan, and Bovik 2013), **PI** (Ma et al. 2017), and **PIQE** (N. et al. 2015) to measure restoration quality. Since the distortion metrics PSNR and SSIM (Wang et al. 2004) cannot model the perception quality well, we use them for reference only. Notice that all metrics are re-computed for fairness. We use the grayscale image to compute the PSNR and SSIM. We compute NIQE and PI by the provided metrics at https://pypi.org/project/pyiqa/. The PIQE is computed via https://github.com/buyizhiyou/NRVQA.

Main Results

Results on Synthetic Datasets. Tab. 1 and Tab. 2 respectively report the comparison results with SOTAs on the SOTS-indoor and SOTS-outdoor datasets (Li et al. 2019). Our method achieves better performance in terms of NIQE, PI, and PIQE, indicating the generated results by our method possess higher perception quality. Fig. 6 and Fig. 7 show that our method restores much clearer images while the evaluated approaches generate the results with haze residual or artifacts. As we train the network with a one-time feed-forward process, PSNRs and SSIMs are naturally decreased (**SelfPromer**$_1$ vs. **SelfPromer**$_3$ in Tabs. 1 and 2) when inference is conducted iteratively. We argue distortion metrics including PSNRs and SSIMs are not good measures for image dehazing as Figs. 6 and 7 have shown methods with higher PSNR and SSIMs cannot recover perceptual results, e.g., Dehamer (Guo et al. 2022) and D4 (Yang et al. 2022), while our method with better perception metrics can generate more realistic results.

Results on Real-World Datasets. Tab. 3 summarises the comparison results on the real-world datasets (Li et al. 2019), where our method performs better than the evaluated methods. Fig. 8 illustrates that our method generates an image with vivid color and finer details.

Methods		GridNet	PFDN	UHD	PSD	Uformer	Restormer	D4	Dehamer	**SelfPromer$_1$**	**SelfPromer$_3$**
Perception	NIQE ↓	4.239	4.412	4.743	4.828	4.378	4.321	4.326	4.529	4.252	**4.054**
	PI ↓	3.889	4.143	4.962	4.567	3.967	3.936	3.866	4.035	3.926	**3.857**
	PIQE ↓	28.924	32.157	39.204	35.174	29.806	29.384	30.480	32.446	30.596	**27.927**
Distortion	PSNR ↑	32.306	33.243	16.920	13.934	33.947	36.979	19.142	36.600	35.960	34.467
	SSIM ↑	0.9840	0.9827	0.7831	0.7160	0.9846	0.9900	0.8520	0.9865	0.9877	0.9852

Table 1: Comparisons on SOTS-indoor dataset. Our method achieves better performance in terms of NIQE, PI, and PIQE. The best results are marked in bold. ↓ (↑) denotes lower (higher) is better. **SelfPromer$_i$** means the i^{th} prompt results.

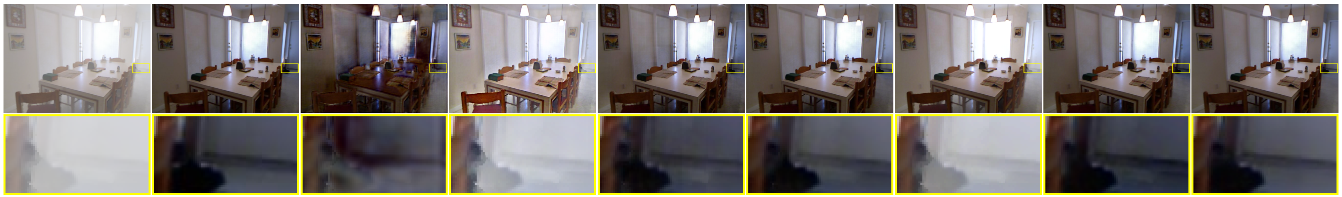

(a) Input (b) GT (c) UHD (d) PSD (e) Uformer (f) Restormer (g) D4 (h) Dehamer (i) **SelfPromer**

Figure 6: Visual comparisons on SOTS-indoor. SelfPromer generates clearer results, even than the GT image.

Methods		GridNet	PFDN	UHD	PSD	Uformer	Restormer	D4	Dehamer	**SelfPromer$_1$**	**SelfPromer$_3$**
Perception	NIQE ↓	2.844	2.843	3.756	2.884	2.903	2.956	2.917	3.164	**2.646**	2.685
	PI ↓	2.070	2.326	3.381	2.392	2.241	2.254	2.137	2.251	**2.003**	2.027
	PIQE ↓	6.547	6.732	10.891	8.937	6.748	6.904	7.567	6.458	6.577	**6.151**
Distortion	PSNR ↑	16.327	16.872	11.758	15.514	19.618	18.337	26.138	21.389	18.471	16.954
	SSIM ↑	0.8016	0.8532	0.6074	0.7488	0.8798	0.8634	0.9540	0.8926	0.8771	0.8288

Table 2: Comparisons on SOTS-outdoor. SelfPromer achieves better perception metrics including NIQE, PI, and PIQE, suggesting that the proposed method has a better generalization ability to unseen images for more natural results generation.

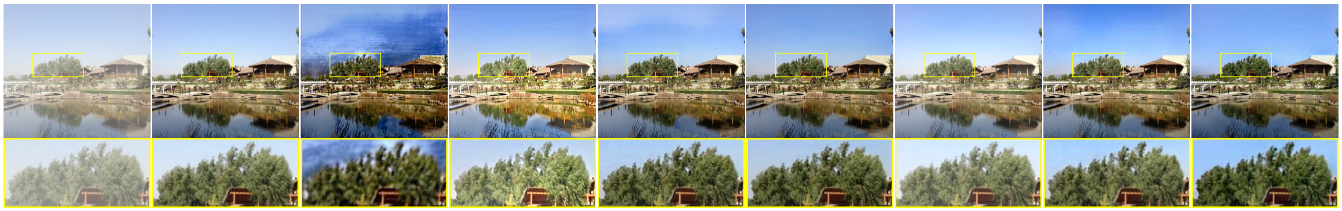

(a) Input (b) GT (c) UHD (d) PSD (e) Uformer (f) Restormer (g) D4 (h) Dehamer (i) **SelfPromer**

Figure 7: Visual comparisons on SOTS-outdoor. SelfPromer is able to generate more natural results. Note that our method produces more consistent colors in the sky region, while the others generate inconsistent colors and the D4 (Yang et al. 2022) leaves extensive haze.

Analysis and Discussion

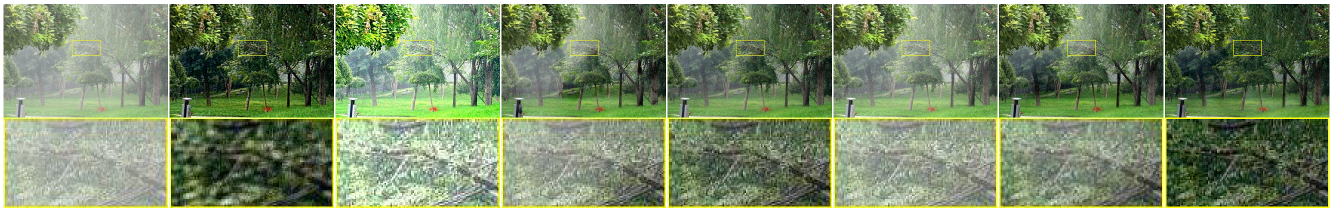
We further analyze the effectiveness of the proposed method and understand how it works on image dehazing. The results in this section are obtained from the SOTS-indoor dataset if not further mentioned. Our results are from the 1st prompt inference for fair comparisons, i.e., $i = 1$ in Eq. (12) if not further specifically mentioned.

Effectiveness of prompt. Initially, we assess the effect of the prompt on image dehazing. Notably, various prospective prompt candidates exist, such as image-level depth difference as the input of the VQGAN encoder or concatenation between deep features extracted from the input and depth features as the input of the Transformers. Our proposed prompt is compared with these candidates, as illustrated in Tab. 4(b) and 4(c), demonstrating that none of these candidates outperforms our proposed prompt.

Note that our method without prompt leads to a similar model with CodeFormer (Zhou et al. 2022) which directly inserts regular Transformers into VQGAN. Tab. 4 shows prompt help yield superior perception quality than the model without prompt (Tab. 4(a)). The efficacy of our model with the prompt is further affirmed by Fig. 9, indicating that the

Methods		GridNet	PFDN	UHD	PSD	Uformer	Restormer	D4	Dehamer	**SelfPromer$_1$**	**SelfPromer$_3$**
Perception	NIQE ↓	4.341	4.917	4.515	4.199	4.214	4.213	4.257	4.248	4.161	**4.062**
	PI ↓	3.685	3.736	3.858	3.521	3.429	3.436	3.414	3.495	3.477	**3.391**
	PIQE ↓	14.699	17.874	23.168	15.851	16.787	17.176	18.678	15.909	16.252	**14.026**

Table 3: Comparisons on real-world dataset. SelfPromer achieves better performance, indicating that our method is more robust to real-world scenarios for realistic results generation.

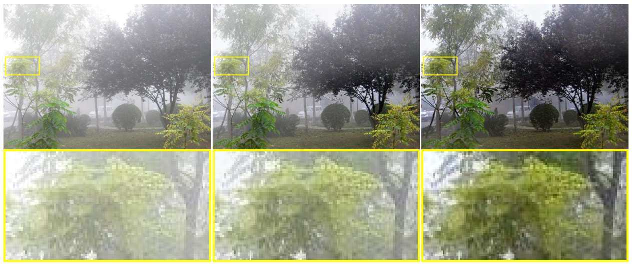


(a) Input (b) UHD (c) PSD (d) Uformer (e) Restormer (f) D4 (g) Dehamer (h) **SelfPromer**

Figure 8: Visual comparisons on real-world dataset. Our SelfPromer is able to generate much clearer results.

Experiments	NIQE ↓	PI ↓	PIQE ↓
(a) Without the prompt	4.258	3.937	31.904
(b) Image-level depth difference	4.901	4.343	32.141
(c) Concat of image and depth features	4.362	4.077	34.107
(d) Feature-level depth difference (**Ours**)	4.252	3.926	30.596

Table 4: Effect of the proposed prompt. Feature-level depth difference is a better prompt formalization. while the concatenation in image-level and feature-level between the input image and its depth is not as well as ours.



(a) Input (b) w/o prompt (c) w/ prompt (**Ours**)

Figure 9: Visual comparisons of the model without prompt (b) and with prompt (c) on real-world scenarios.

Experiments	NIQE ↓	PI ↓	PIQE ↓
(a) Without embedding	4.410	4.113	32.193
(b) Position embedding	4.267	3.992	31.877
(c) Regular attention	4.300	4.102	31.486
(d) Proposed (**Ours**)	4.252	3.926	30.596

Table 5: Effectiveness of prompt embedding and attention.

model with the prompt generates better results, while the model without prompt fails to remove haze effectively.

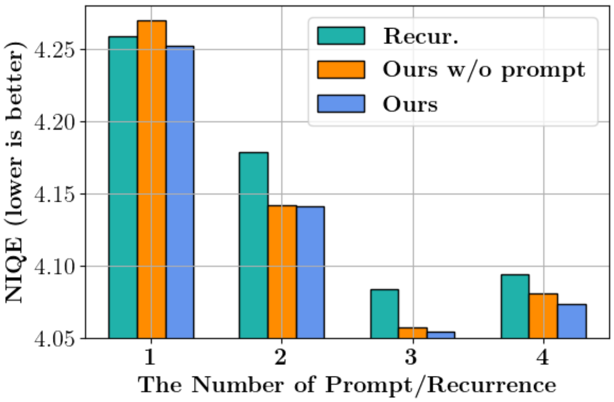
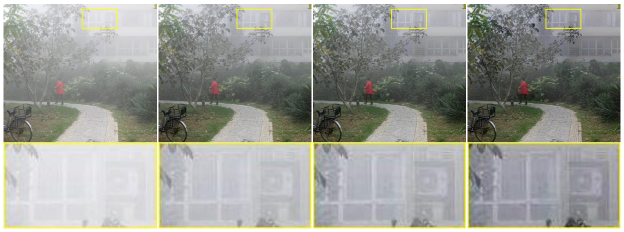


Figure 10: Effectiveness of continuous self-prompt (Ours) vs. recurrent dehazing (Recur.). 'Ours w/o prompt' means the results of Eq. (12a).

Effectiveness of prompt embedding and prompt attention. One might ponder the relative efficacy of our prompt embedding and attention in contrast to the prevalent technique of position embedding and regular attention. In this regard, we assess the effect of these embedding approaches in Tab. 5. The table reveals that our prompt embedding proves more advantageous over the position embedding since the former is associated with haze residual information. Tab. 5 indicates that our prompt attention yields better results as compared to commonly used attention methods. These findings signify that incorporating prompts in enhancing Query estimation accounts for the haze information, thereby culminating in more effective image dehazing results.

Effect of the number of steps in continuous self-prompt. The inference stage involves several steps to generate the prompt for better image dehazing. We thus examine the effect of the number of steps in the continuous self-prompt. Fig. 10 reveals that the optimal performance is achieved with



(a) Input (b) 1st prompt (c) 2nd prompt (d) 3rd prompt

Figure 11: Visual improvement of continuous self-prompt inference on a real-world example.

a number of steps equal to 3 in the continuous self-prompts (i.e., $i = 3$ in Eq. (12)), in terms of NIQE. Notably, additional prompts do not improve the dehazing performance any further. One real-world example in Fig. 11 demonstrates that our continuous self-prompt method can gradually enhance dehazing quality.

Continuous self-prompt vs. recurrent dehazing. We use the continuous self-prompt approach to restore clear images progressively at inference. To determine whether a recurrent method that is training our model without prompt achieves similar or better results, we compare our proposed method with it in Fig. 10, demonstrating that the recurrent method is not as good as our continuous self-prompt.

Continuous self-prompt vs. GT guidance. Fig. 5 compares the NIQE performance of ground truth (GT) guidance with that of the continuous self-prompt algorithm. Results show that while GT guidance performs better than the 1st prompt, it falls short of the effectiveness of the 2nd and 3rd prompts. This is likely due to GT guidance's limited ability to handle haze residuals which may still exist in the dehazed images, which are addressed by the self-prompt's ability to exploit residual haze information to progressively improve dehazing quality over time. Moreover, as GT is not available in the real world, these findings may further support the use of self-prompt as a more practical alternative.

Depth-consistency. Fig. 12 shows heat maps of depth differences obtained by the continuous self-prompt inference with different prompt steps. The results demonstrate both image-level and feature-level depth differences decrease as the number of prompt steps increases, indicating the depths obtained with the prompt, i.e., Eq. (12c), become increasingly consistent with those obtained without it, i.e., Eq. (12a).

Applications to Low-Light Image Enhancement

Furthermore, we extend the application of our method, SelfPromer, to the domain of low-light image enhancement. To assess its performance, we conduct a comparative analysis with current state-of-the-art methods, SNR (Xu et al. 2022) and LLFlow (Wang et al. 2022). All these methods are trained using the widely adopted LOL dataset (Wei et al. 2018). Fig. 13 shows several visual examples sourced from real-world benchmarks (Lee, Lee, and Kim 2013; Guo, Li, and Ling 2017). These illustrative results effectively underscore our approach is capable of generating results that are notably more true-to-life, with colors that appear more natural. On the contrary, existing state-of-the-art methods tend to yield results that suffer from under-/over-exposure issues.

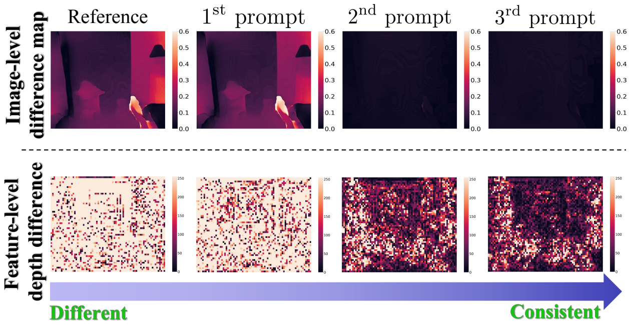
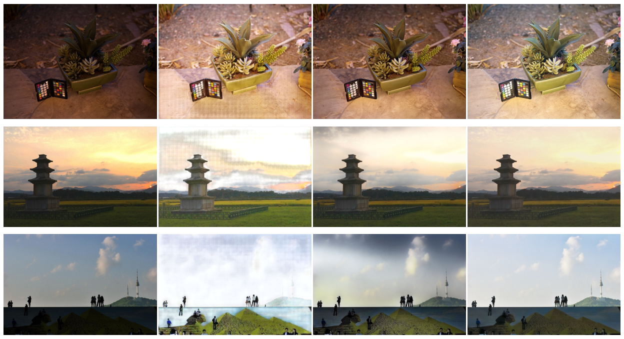


Figure 12: Illustration of continuous depth-consistency. Reference means the depth difference between the input hazy image and GT. The input haze image is Fig. 1(a).



(a) Input (b) SNR (c) LLFlow (d) **SelfPromer**

Figure 13: Applications to low-light image enhancement on challenging real-world examples.

Conclusion

We have proposed a simple yet effective self-prompt Transformer for image dehazing by exploring the prompt built on the estimated depth difference between the image with haze residuals and its clear counterpart. We have shown that the proposed prompt can guide the deep model for better image dehazing. To generate better dehazing images at the inference stage, we have proposed continuous self-prompt inference, where the proposed prompt strategy can remove haze progressively. We have shown that our method generates results with better perception quality in terms of NIQE, PI, and PIQE.

References

Chen, C.; Shi, X.; Qin, Y.; Li, X.; Han, X.; Yang, T.; and Guo, S. 2022. Real-World Blind Super-Resolution via Feature Matching with Implicit High-Resolution Priors. In *ACM MM*, 1329–1338.

Chen, Z.; Wang, Y.; Yang, Y.; and Liu, D. 2021. PSD: Principled Synthetic-to-Real Dehazing Guided by Physical Priors. In *CVPR*, 7180–7189.

Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; and Wei, Y. 2017. Deformable Convolutional Networks. In *ICCV*, 764–773.

Ding, H.; Liu, C.; Wang, S.; and Jiang, X. 2021. Vision-Language Transformer and Query Generation for Referring Segmentation. In *ICCV*, 16301–16310.

Dong, H.; Pan, J.; Xiang, L.; Hu, Z.; Zhang, X.; Wang, F.; and Yang, M. 2020. Multi-Scale Boosted Dehazing Network With Dense Feature Fusion. In *CVPR*, 2154–2164.

Dong, J.; and Pan, J. 2020. Physics-Based Feature Dehazing Networks. In *ECCV*, 188–204.

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

Gan, Y.; Ma, X.; Lou, Y.; Bai, Y.; Zhang, R.; Shi, N.; and Luo, L. 2023. Decorate the Newcomers: Visual Domain Prompt for Continual Test Time Adaptation. In *AAAI*.

Gu, Y.; Wang, X.; Xie, L.; Dong, C.; Li, G.; Shan, Y.; and Cheng, M. 2022. VQFR: Blind Face Restoration with Vector-Quantized Dictionary and Parallel Decoder. In *ECCV*, 126–143.

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 *CVPR*, 5812–5820.

Guo, X.; Li, Y.; and Ling, H. 2017. LIME: Low-Light Image Enhancement via Illumination Map Estimation. *IEEE TIP*, 26(2): 982–993.

He, K.; Sun, J.; and Tang, X. 2011. Single Image Haze Removal Using Dark Channel Prior. *IEEE TPAMI*, 33(12): 2341–2353.

Herzig, R.; Abramovich, O.; Ben-Avraham, E.; Arbelle, A.; Karlinsky, L.; Shamir, A.; Darrell, T.; and Globerson, A. 2022. PromptonomyViT: Multi-Task Prompt Learning Improves Video Transformers using Synthetic Scene Data. *CoRR*, abs/2212.04821.

Jin, Y.; Lin, B.; Yan, W.; Yuan, Y.; Ye, W.; and Tan, R. T. 2023. Enhancing visibility in nighttime haze images using guided apsf and gradient adaptive convolution. In *ACM MM*, 2446–2457.

Jin, Y.; Yan, W.; Yang, W.; and Tan, R. T. 2022. Structure Representation Network and Uncertainty Feedback Learning for Dense Non-Uniform Fog Removal. In *ACCV*, 2041–2058.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.

Lee, C.; Lee, C.; and Kim, C. 2013. Contrast Enhancement Based on Layered Difference Representation of 2D Histograms. *IEEE TIP*, 22(12): 5372–5384.

Li, B.; Ren, W.; Fu, D.; Tao, D.; Feng, D.; Zeng, W.; and Wang, Z. 2019. Benchmarking Single-Image Dehazing and Beyond. *IEEE TIP*, 28(1): 492–505.

Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; and Neubig, G. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9): 1–35.

Liu, X.; Ma, Y.; Shi, Z.; and Chen, J. 2019a. GridDehazeNet: Attention-Based Multi-Scale Network for Image Dehazing. In *ICCV*, 7313–7322.

Liu, Y.; Pan, J.; Ren, J.; and Su, Z. 2019b. Learning Deep Priors for Image Dehazing. In *ICCV*, 2492–2500.

Ma, C.; Yang, C.; Yang, X.; and Yang, M. 2017. Learning a no-reference quality metric for single-image super-resolution. *CVIU*, 158: 1–16.

Mittal, A.; Soundararajan, R.; and Bovik, A. C. 2013. Making a "Completely Blind" Image Quality Analyzer. *IEEE SPL*, 20(3): 209–212.

N., V.; D., P.; Bh., M. C.; Channappayya, S. S.; and Medasani, S. S. 2015. Blind image quality evaluation using perception based features. In *NCC*, 1–6.

Ranftl, R.; Lasinger, K.; Hafner, D.; Schindler, K.; and Koltun, V. 2022. Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer. *IEEE TPAMI*, 44(3): 1623–1637.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*.

Wang, C.; Pan, J.; Wang, W.; Dong, J.; Wang, M.; Ju, Y.; and Chen, J. 2023. PromptRestorer: A Prompting Image Restoration Method with Degradation Perception. In *NeurIPS*.

Wang, Y.; Wan, R.; Yang, W.; Li, H.; Chau, L.; and Kot, A. C. 2022. Low-Light Image Enhancement with Normalizing Flow. In *AAAI*, 2604–2612.

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

Wei, C.; Wang, W.; Yang, W.; and Liu, J. 2018. Deep Retinex Decomposition for Low-Light Enhancement. In *BMVC*, 155.

Xu, X.; Wang, R.; Fu, C.-W.; and Jia, J. 2022. SNR-Aware Low-Light Image Enhancement. In *CVPR*, 17714–17724.

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 *CVPR*, 2037–2046.

Zheng, Z.; Yue, X.; Wang, K.; and You, Y. 2022. Prompt Vision Transformer for Domain Generalization. *CoRR*, abs/2208.08914.

Zhou, S.; Chan, K. C. K.; Li, C.; and Loy, C. C. 2022. Towards Robust Blind Face Restoration with Codebook Lookup Transformer. In *NeurIPS*.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks. In *ICCV*, 2242–2251.

Zhu, X.; Hu, H.; Lin, S.; and Dai, J. 2019. Deformable ConvNets V2: More Deformable, Better Results. In *CVPR*, 9308–9316.