Adaptive Dynamic Dehazing via Instruction-Driven and Task-Feedback Closed-Loop Optimization for Diverse Downstream Task Adaptation

Yafei Zhang[1], Shuaitian Song[1], Huafeng Li[1]*, Shujuan Wang[1], Yu Liu[2]

[1]Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China

[2]School of Instrument Seicence and Opto-electronics Engineering, Hefei University of Technology, Hefei 230009 , China

zyfeimail@163.com, songshuaitiann@163.com, hfchina99@163.com, wangsj@kust.edu.cn, yuliu@hfut.edu.cn

Abstract

In real-world vision systems, haze removal is required not only to enhance image visibility but also to meet the specific needs of diverse downstream tasks. To address this challenge, we propose a novel adaptive dynamic dehazing framework that incorporates a closed-loop optimization mechanism. It enables feedback-driven refinement based on downstream task performance and user instruction–guided adjustment during inference, allowing the model to satisfy the specific requirements of multiple downstream tasks without retraining. Technically, our framework integrates two complementary and innovative mechanisms: (1) a task feedback loop that dynamically modulates dehazing outputs based on performance across multiple downstream tasks, and (2) a text instruction interface that allows users to specify high-level task preferences. This dual-guidance strategy enables the model to adapt its dehazing behavior after training, tailoring outputs in real time to the evolving needs of multiple tasks. Extensive experiments across various vision tasks demonstrate the strong effectiveness, robustness, and generalizability of our approach. These results establish a new paradigm for interactive, task-adaptive dehazing that actively collaborates with downstream applications.

Code — https://github.com/songshuaitian/ADeT-Net

Introduction

Haze is a common atmospheric condition that severely impairs image visibility and scene understanding. Consequently, image dehazing has become a critical preprocessing step in real-world systems like autonomous driving and surveillance. Early approaches mainly aimed to enhance visual quality using handcrafted priors or supervised models trained on synthetic data. However, improving visual appearance alone does not guarantee better performance in downstream vision tasks. In practical scenarios, the dehazed output often serves as the input to task-specific models, and misalignment between the objectives of dehazing and those of downstream tasks can lead to suboptimal or even detrimental outcomes. To address this issue, recent studies (Zhang et al. 2020; Sun et al. 2022) have explored integrating downstream tasks into the dehazing pipeline by jointly training the dehazing model with a specific task network. While such integration can enhance performance for a single task, it also introduces significant limitations: these methods require retraining for each new task and lack flexibility to adapt to different tasks once deployed. Despite these efforts, achieving a generalizable solution that supports diverse downstream tasks remains an open challenge.

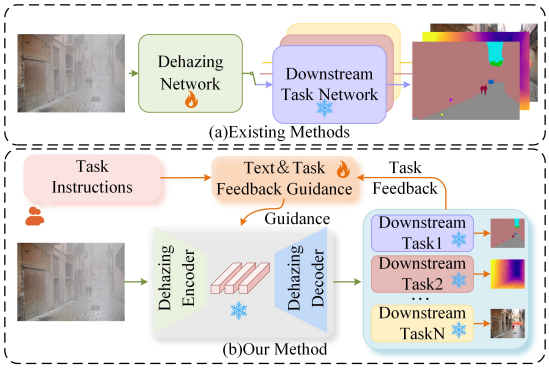

Figure 1: Comparison between existing methods and ours.

To address these limitations, we propose a closed-loop optimization framework for adaptive and dynamic image dehazing. Unlike traditional approaches that produce static outputs regardless of task context (As shown in Figure 1), our method is designed to optimize dehazing results not only in terms of visual clarity but also with respect to the specific requirements of downstream tasks. The key innovation lies in a dual-guidance mechanism that integrates both feedback from downstream task performance and high-level user intent expressed in textual instructions. By jointly leveraging these two sources of guidance, the model can adjust its dehazing behavior in real time during inference, without requiring any retraining or task-specific fine-tuning. This architecture holds promise for serving as a flexible and interactive dehazing module adaptable to various downstream tasks.

To realize the above design, we construct an adaptive dehazing framework centered on a closed-loop optimization mechanism. The process begins with training an Initial Dehazing Network (IDN) on synthetic hazy data to ac-

*Corresponding author.

quire general image restoration capabilities. During inference, the dehazing results are refined in real time through a dual-guidance strategy. Specifically, feedback from downstream task performance is used to guide feature modulation based on how well the current dehazed image supports the task. Meanwhile, user-provided textual instructions are interpreted to capture high-level intent and semantic preferences. These two signals—task-driven feedback and instruction-based semantics—jointly inform the modulation of intermediate features within the dehazing network, enabling dynamic and context-aware adaptation without requiring retraining. This framework introduces a novel paradigm that bridges low-level image restoration with high-level task guidance, allowing the dehazing process to become interactive, controllable, and task-specific. This design offers a generalizable solution for real-world vision systems where adaptability and compatibility with multiple tasks are essential. In summary, the main contributions of this work are as follows:

- We propose a novel closed-loop dehazing framework that enables dynamic, task-aware, and instruction-driven refinement during inference, achieving real-time adaptation without any model retraining or fine-tuning. This design significantly improves the flexibility and deployment efficiency of dehazing model in dynamic, multi-task environments.
- We present a dual-guidance mechanism that combines downstream task feedback and semantic-level textual instructions to support dynamic, task-specific adaptation of the dehazing process. This mechanism is instantiated through two modules—Task Feedback-Guided Adaptation (TFGA) and Instruction-Guided Modulation (IGM)—which collaboratively enable real-time and fine-grained refinement of dehazing outputs according to varied task objectives, without requiring retraining.
- We conduct extensive experiments on diverse downstream tasks including detection, segmentation, and depth estimation. Results show consistent performance gains over both traditional and task-aware baselines, demonstrating the effectiveness and adaptability of our approach, and offering a promising paradigm for interactive and task-driven dehazing in real-world applicaiton.

Related Work

Typical Image Dehazing

In image dehazing, atmospheric scattering model-based methods are widely used. These approaches typically estimate the transmission map and atmospheric light to reconstruct clear images, and can be categorized into three types: end-to-end joint parameter estimation (E2E-JPE) methods (Cai et al. 2016; Zhang and Patel 2018; Ren et al. 2020), physical model simplification (PMS) methods (Li et al. 2017; Zhang and Tao 2020), and physical prior-guided (PPG) methods (He, Sun, and Tang 2011; Zhang, Wang, and Wang 2021). E2E-JPE methods, such as DehazeNet (Cai et al. 2016) and DCPDN (Zhang and Patel 2018), treat transmission and atmospheric light as learnable parameters and substitute them into the scattering model for restoration. They provide high physical interpretability but are sensitive to parameter errors. PMS methods, like AOD-Net (Li et al. 2017) and FAMED-Net (Zhang and Tao 2020), simplify the model by combining parameters, but at the cost of physical fidelity. PPG methods utilize priors (e.g., dark/bright channel (He, Sun, and Tang 2011; Zhang, Wang, and Wang 2021)) to guide estimation, offering generalization with limited data but prone to artifacts when assumptions fail.

Beyond model-based approaches, end-to-end deep learning methods directly map hazy images to clear outputs. They are typically CNN-based (Li et al. 2023; Zheng et al. 2023; Li et al. 2022), GAN-based (Shao et al. 2020; Lan et al. 2025), or Transformer-based (Guo et al. 2022; Song et al. 2023). CNN-based methods exploit local features without modeling physical processes, while GAN-based approaches enhance realism via adversarial training but may suffer from instability. Transformer-based methods capture long-range dependencies through self-attention, effectively modeling fog patterns. However, most of these methods overlook the adaptability of dehazed results to downstream tasks.

Downstream Task-Driven Image Dehazing

Downstream task-driven image dehazing has garnered increasing attention, yet conventional methods often focus solely on enhancing visual quality while overlooking their impact on downstream tasks. To address this issue, researchers have explored joint optimization strategies that couple dehazing with high-level vision tasks. Methods such as UDnD (Zhang et al. 2020), ADAM-Dehaze (Sun et al. 2022), and MS-FODN (Wan et al. 2025) integrate object detection into the dehazing process, while VRD-IR (Yang et al. 2023) incorporates object recognition. Although these approaches consider the influence of dehazing on downstream tasks, they are typically tailored to a single task and lack the flexibility to generalize across multiple tasks. Furthermore, they do not utilize performance feedback or semantic guidance from downstream tasks to regulate the dehazing process, often leading to suboptimal outputs.

Methodology

Overview

Figure 2 illustrates the framework of the proposed method. To adapt the dehazing results to various downstream tasks without requiring task-specific fine-tuning, a closed-loop optimization strategy is introduced, combining text instruction guidance and feedback from downstream tasks. In this framework, the initial dehazed images are input into downstream task models aligned with the objectives described in the text instructions. Feedback from these tasks, together with semantic information extracted from the instructions, jointly informs adjustments to the dehazing network. This enables the network to generate results better suited to the current task's requirements. Two key modules support this dynamic adjustment: the TFGA module, which adapts the network's feature outputs based on task performance, and the IGM module, which interprets the instructions' semantic content to guide dehazing adjustments. Together, these

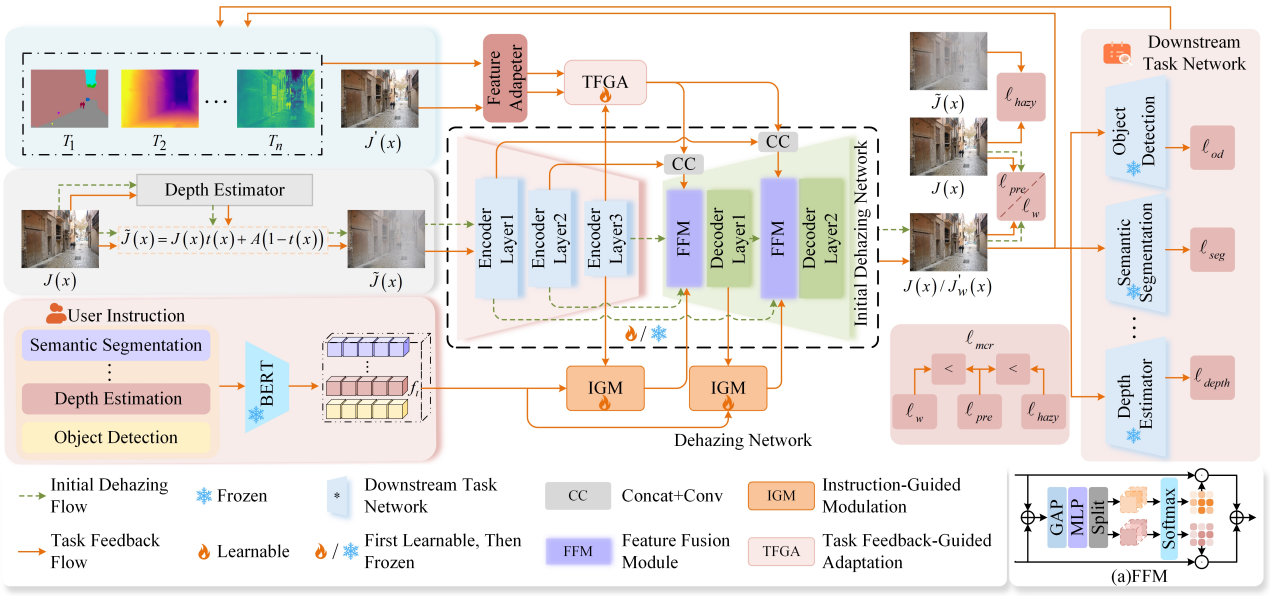

Figure 2: Overview of the proposed method. The method constructs a closed-loop regulation mechanism jointly guided by semantic task instructions and task performance feedback. It leverages the semantic features of text instructions extracted by BERT, the initial dehazed images, and feedback from downstream tasks to collaboratively adjust dehazing features via the IGM and TFGA modules, enabling adaptive optimization across diverse downstream scenario.

modules form a closed-loop system that connects dehazing outputs and task execution feedback, thereby enhancing the adaptability of the dehazing results across diverse tasks.

Initial Dehazing Netowrk

As shown in Figure 2, based on the atmospheric scattering model, we add haze to the clear image $J(x)$ with the assistance of a depth estimator to generate a hazy image $\tilde{J}(x)$. $\tilde{J}(x)$ is then fed into the IDN encoder to extract deep features. These features are then passed to the decoder for image reconstruction, yielding an initial dehazed result $J'(x)$. The IDN, based on Transformer architecture and following the U-Net encoder-decoder paradigm, includes three encoder layers for processing features at different scales. The decoder consists of two layers and two FFMs. Features at the same scale from the encoder and decoder are combined via residual connections and then passed to the FFM for fusion.

During IDN training, the FFM integrates features from the encoder and decoder at corresponding scales, reducing information loss during transmission. In the closed-loop optimization stage, the FFM further receives task semantics from the TFGA and IGM modules. This enables the model to dynamically regulate the dehazing process according to task requirements and instruction content. To ensure strong restoration performance, we jointly optimize the IDN using both l_1-loss and contrastive loss:

$$\ell_{predeh} = \|J'(x) - J(x)\|_1 + \lambda \sum_{v=1}^{n} \beta_v \cdot \frac{\|\mathrm{VGG}_v(J(x)) - \mathrm{VGG}_v(J'(x))\|_1}{\|\mathrm{VGG}_v(J'(x)) - \mathrm{VGG}_v(\tilde{J}(x))\|_1} \quad (1)$$

where VGG_v denotes the output from the v-th layer of the VGG19 (Simonyan and Zisserman 2014), and β_v is the corresponding weighting coefficient, as defined in (Wu et al. 2021). In addition, λ is a hyperparameter set to 0.1.

Task Feedback-Guided Adaptation

Given that the decoder in a dehazing network primarily performs image reconstruction and detail generation, regulating it enables precise optimization of image quality with relatively low complexity and risk due to its relatively independent structure. Therefore, this paper focuses the regulation on the decoder to directly guide detail recovery during image reconstruction, thereby producing dehazed images better aligned with the requirements of downstream tasks. As illustrated in Figure 3, the TFGA is mainly composed of a bidirectional cross-attention mechanism and two Channel-wise Feature Fusion Blocks (CFFB). The input features, $\boldsymbol{F}_{id}$ and $\boldsymbol{F}_{down}$, are the outputs of the Feature Adapter applied to $J'(x)$ and downstream task feedback, respetively. For tasks such as semantic segmentation or depth estimation, $\boldsymbol{F}_{down}$ is the output of the downstream task applied to the dehazed image $J'(x)$. For object detection, $\boldsymbol{F}_{down}$ refers to the intermediate features extracted by the feature extraction network of the downstream task from $J'(x)$.

To enhance the representational capacity of useful information in $\boldsymbol{F}_{id}$, the bidirectional cross-attention mechanism performs interaction modeling in two directions. In the first direction, $\boldsymbol{F}_{id}$ and $\boldsymbol{F}_{down}$ are concatenated and passed through a convolutional layer, followed by a linear transformation to obtain the fused feature $\boldsymbol{F}_{id,down}$, which serves as the query $\boldsymbol{Q}$ for the attention module. Meanwhile, $\boldsymbol{F}_{id}$ is linearly projected to generate the key-value pairs $\boldsymbol{K}$ and $\boldsymbol{V}$.

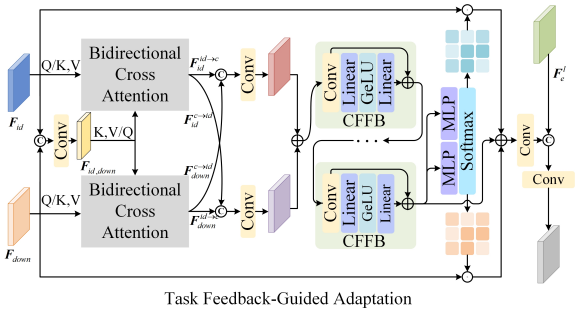

Figure 3: Structure of the TFGA.

The resulting cross-attention output is denoted as $\boldsymbol{F}_{id}^{id\to c}$. In the reverse direction, $\boldsymbol{F}_{id,down}$ is linearly projected to form $\boldsymbol{K}$ and $\boldsymbol{V}$, which are then used to attend to $\boldsymbol{F}_{id}$, producing the output $\boldsymbol{F}_{id}^{c\to id}$. Symmetrically, $\boldsymbol{F}_{down}$ undergoes the same procedure to yield $\boldsymbol{F}_{down}^{id\to c}$ and $\boldsymbol{F}_{down}^{c\to id}$.

To further explore the structural information in $\boldsymbol{F}_{id}$ and $\boldsymbol{F}_{down}$, we concatenate $\boldsymbol{F}_{id}^{id\to c}$ with $\boldsymbol{F}_{down}^{c\to id}$, and $\boldsymbol{F}_{id}^{c\to id}$ with $\boldsymbol{F}_{down}^{id\to c}$. These concatenated features are then processed through convolutional layers to extract deeper semantic representations. The outputs are summed and passed through two CFFBs to obtian $\boldsymbol{F}_{idd}$, which is then fed into two Multi-Layer Perceptrons (MLPs). Finally, a Softmax function is applied to obtain the regulation weight matrices $\boldsymbol{Q}_{id}$ and $\boldsymbol{Q}_{down}$ for $\boldsymbol{F}_{id}$ and $\boldsymbol{F}_{down}$. The fused feature incorporating downstream task feedback is computed as:

$$\boldsymbol{F}_{id,dow} = \mathrm{Conv}\left(\boldsymbol{F}_{id} \odot \boldsymbol{Q}_{id} + \boldsymbol{F}_{down} \odot \boldsymbol{Q}_{down} + \boldsymbol{F}_{idd}\right) \quad (2)$$

We concatenate $\boldsymbol{F}_{id,dow}$ with the output $\boldsymbol{F}_e^l$ from the final Transformer layer in the encoder, and further process them with a convolutional layer to obtain the modulated result that reflects the downstream task feedback. This result, along with $\boldsymbol{F}_e^l$ and the output of the previous encoder layer, is input into the FFM for integration. The integrated feature is then used in the decoding stage.

In the aforementioned process, the performance of the downstream task network on the initial dehazed result $J'(x)$ is fed back to the dehazing network. This feedback then guides the feature adjustment during the dehazing process, enabling the representations extracted at the decoder end to be more aligned with the requirements of specific tasks. Through this feedback mechanism, the dehazing network not only possesses the capability to restore clear images but also generates feature representations that are more conducive to downstream task.

Instruction-Guided Modulation

Although downstream task feedback can make the restoration results more task-specific, the optimization process may become blind and inefficient without semantic guidance. Therefore, this paper further introduces textual instruction information to incorporate semantic guidance alongside task feedback, establishing a complementary relationship between semantic intention parsing and performance-driven

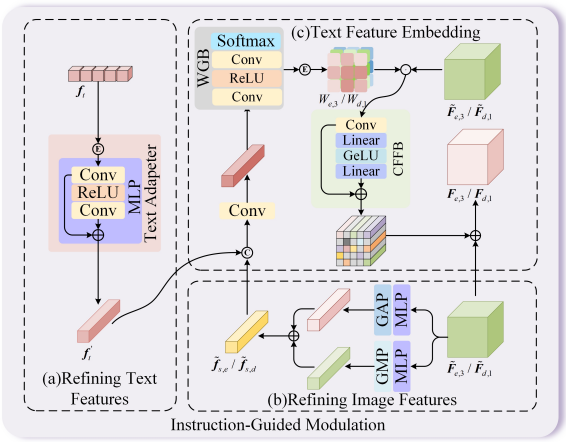

Figure 4: Structure of the IGM.

regulation. Specifically, the IGM interprets the task semantics from the text, while the TFGA handles feedback from downstream tasks. This approach forms a robust closed-loop optimization mechanism, improving the adaptability of dehazed images to diverse downstream tasks.

To more directly leverage the textual instructions for enhancing the controllability and task adaptability of the model, the output of the IGM is applied exclusively to the decoder of the dehazing network. Specifically, the task instructions (`text`) provided by the user are fed into a pretrained BERT model to extract the instruction feature vector $\boldsymbol{f}_t$. Let $\tilde{\boldsymbol{F}}_{e,3}$ denote the features obtained from the third layer of the encoder in the IDN, and let $\tilde{\boldsymbol{F}}_{d,1}$ represent the features reconstructed by the first layer of the decoder. In our method, two IGM modules are introduced to gradually inject task semantics into the decoding process, thereby equipping the dehazing network with a degree of semantic controllability. The first IGM takes $\boldsymbol{f}_t$ and $\tilde{\boldsymbol{F}}_{e,3}$ as inputs to guide the adjustment of decoder features, while the second IGM uses $\boldsymbol{f}_t$ and $\tilde{\boldsymbol{F}}_{d,1}$ to further enhance the embedding of task-specific information. As illustrated in Figure 4, the IGM module comprises three key components: text feature refinement, image feature refinement, and text feature embedding.

The core objective of text feature refinement is to refine a task-oriented representation from the BERT-extracted feature $\boldsymbol{f}_t$. This process is implemented using a text adapter composed of an expand operation and a MLP. The adapter not only enhances the representation of task-related semantics but also projects $\boldsymbol{f}_t$ from the text semantic space into the image feature space, facilitating cross-modal alignment and feature modulation. The output of the adapter is denoted as $\boldsymbol{f}_t'$. The image feature refinement block comprises two branches, each consisting of an MLP followed by either GMP or GAP, respectively. The outputs of these two branches are summed to obtain the refined image feature representation. When the inputs to the IGM are $\boldsymbol{f}_t$ and $\tilde{\boldsymbol{F}}_{e,3}$ or $\tilde{\boldsymbol{F}}_{d,1}$, the corresponding refined results are denoted as $\tilde{\boldsymbol{f}}_{s,e}$ and $\tilde{\boldsymbol{f}}_{s,d}$, respectively.

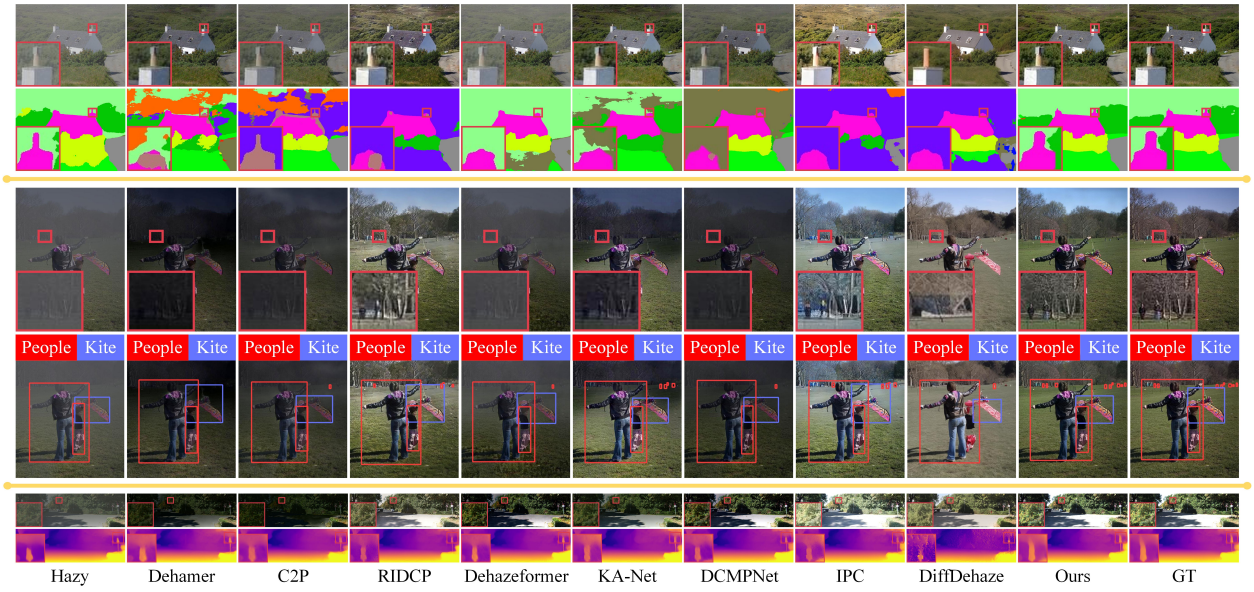

Figure 5: Visual comparison with state-of-the-art methods on Setting 1. Each part includes two rows: dehazing results (rows 1) and corresponding downstream task outputs (rows 2). Input images are taken from ADE20K, COCO, and KITTI.

The text feature embedding block consists of a Weight Generation Block (WGB) and a CFFB, as shown in Figure 4. During embedding, $\boldsymbol{f}_t'$ and $\tilde{\boldsymbol{f}}_{s,e}$ / $\tilde{\boldsymbol{f}}_{s,d}$ are input into the WGB, which generates modulation parameters $\boldsymbol{W}_{e,3}$ ($\boldsymbol{W}_{d,1}$) through an expand operation. These parameters modulate the input image features $\tilde{\boldsymbol{F}}_{e,3}$ ($\tilde{\boldsymbol{F}}_{d,1}$):

$$\bar{\boldsymbol{F}}_{e,3} = \boldsymbol{W}_{e,3} \odot \tilde{\boldsymbol{F}}_{e,3}, \quad \bar{\boldsymbol{F}}_{d,1} = \boldsymbol{W}_{d,1} \odot \tilde{\boldsymbol{F}}_{d,1} \tag{3}$$

The modulated features $\bar{\boldsymbol{F}}_{e,3}$ ($\bar{\boldsymbol{F}}_{d,1}$) are subsequently fed into the CFFB. The output is added to the original image features to obtain the final modulated representations:

$$\boldsymbol{F}_{e,3} = \text{CFFB}(\bar{\boldsymbol{F}}_{e,3}) + \tilde{\boldsymbol{F}}_{e,3}, \quad \boldsymbol{F}_{d,1} = \text{CFFB}(\bar{\boldsymbol{F}}_{d,1}) + \tilde{\boldsymbol{F}}_{d,1} \tag{4}$$

In this process, the CFFB further enhances the structural expressiveness of the visual features $\tilde{\boldsymbol{F}}_{e,3}$ ($\tilde{\boldsymbol{F}}_{d,1}$), thereby highlighting information beneficial for image restoration. Furthermore, since the semantic information from the textual instructions is injected into $\boldsymbol{F}_{e,3}$ and $\boldsymbol{F}_{d,1}$, the visual features can be dynamically adapted to downstream task instructions.

To ensure that the dehazed results modulated by the TFGA and the IGM better align with the requirements of downstream tasks, we use both l_1-loss and contrastive loss:

$$\begin{aligned} \ell_{dehaze} = &\|J_w'(x) - J(x)\|_1 + \\ &\lambda \sum_{v=1}^{n} \beta_v \frac{\|VGG_v(J(x)) - VGG_v(J_w'(x))\|_1}{\|VGG_v(J_w'(x)) - VGG_v(\tilde{J}(x))\|_1} \end{aligned} \tag{5}$$

where $J_w'(x)$ denotes the dehazed result after modulation. We expect this modulated result, guided by downstream task feedback and text instructions, to outperform the intermediate and initial dehazed outputs. Accordingly, the reconstruction loss is expected to satisfy the following inequality:

$$\|J_w'(x) - J(x)\|_1 < \|J'(x) - J(x)\|_1 < \|\tilde{J}(x) - J(x)\|_1 \tag{6}$$

To enforce the relative quality constraint between different dehazing results, we propose a Multi-level Contrastive Ranking Loss (ℓ_{mcr}), defined as:

$$\ell_{mcr} = \max(\ell_w - \ell_p + \beta_1, 0) + \max(\ell_w - \ell_h + \beta_2, 0) \tag{7}$$

where β_1 and β_2 are hyperparameters, empirically set to 0.1 and 0.3, respectively, with $\beta_1 < \beta_2$. The individual loss terms are defined as follows: $\ell_w = \|J_w'(x) - J(x)\|_1$, $\ell_p = \|J'(x) - J(x)\|_1$, and $\ell_h = \|\tilde{J}(x) - J(x)\|_1$.

This paper aims to adapt the restoration results of a dehazing model to various downstream tasks without retraining it. Thus, we introduce a task-specific loss term ℓ_{down} to ensure the performance of the downstream tasks:

$$\ell_{down} = \ell_{task}(\tilde{y}_{gt}(x), \tilde{y}'(x)) \tag{8}$$

where ℓ_{task} denotes the loss function of the downstream task network. $\tilde{y}_{gt}(x)$ represents the ground truth, and $\tilde{y}'(x)$ denotes the output of $J_w'(x)$ appilied to the downstream task network. For semantic segmentation, object detection, and depth estimation, we adopt the same loss functions as those used in SegFormer (Xie et al. 2021), YOLOv5 [1], and RADepth (He et al. 2022), respectively. Accordingly, the overall loss function is defined as:

$$\ell_{total} = \ell_{dehaze} + \ell_{mcr} + \gamma \ell_{down} \tag{9}$$

where γ is a weighting coefficient for the downstream task loss, empirically set to 0.01.

Closed-Loop Optimization

During model inference, the core advantage of our method lies in its ability to dynamically adjust the initial dehazed results through a closed-loop optimization mechanism, thereby achieving better alignment with specific

[1]https://github.com/ultralytics/yolov5

Methods	ADE20K			COCO			KITTI		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Dehamer (Guo et al. 2022)	22.63	0.8964	0.1538	23.26	0.9280	0.1369	24.87	0.9212	0.1029
C2P (Zheng et al. 2023)	22.26	0.8842	0.1594	23.02	0.9044	0.1322	25.01	0.9342	0.0926
RIDCP (Wu et al. 2023)	<u>27.21</u>	0.9207	0.0702	**28.18**	0.8990	0.0974	27.18	0.9441	0.1242
Dehazeformer (Song et al. 2023)	23.76	0.8958	0.1332	22.41	0.8639	0.1412	23.36	0.9160	0.1163
KA-Net (Feng et al. 2024)	24.57	0.9153	0.0914	24.92	0.9241	0.1291	26.73	0.9413	0.0968
DCMPNet (Zhang, Zhou, and Li 2024)	25.64	<u>0.9571</u>	<u>0.0326</u>	25.39	<u>0.9422</u>	0.0892	28.52	0.9617	0.0384
DiffDehaze (Wang et al. 2025)	25.43	0.9194	0.0853	24.68	0.9127	0.1048	26.91	0.9568	0.0454
IPC (Fu et al. 2025)	26.23	0.9432	0.0618	26.32	0.9381	<u>0.0837</u>	<u>28.94</u>	<u>0.9632</u>	<u>0.0296</u>
Ours	**27.47**	**0.9701**	**0.0293**	<u>27.14</u>	**0.9587**	**0.0698**	**30.50**	**0.9740**	**0.0176**

Table 1. Comparison of dehazing performance. Best and runner-up values are highlighted in bold and underlined, respectively.

Methods	SS	OD		DE (Error Metric ↓)				DE (Accuracy Metric ↑)		
	mIoU↑	mAP↑	mAP50-95↑	AbsRel	SqRel	RMSE	RMSElog	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Dehamer	44.04	49.8	30.4	0.115	0.860	4.802	0.195	0.874	0.956	0.980
C2P	46.66	52.3	33.5	0.123	0.888	4.933	0.206	0.855	0.943	0.972
RIDCP	46.15	51.7	33.0	<u>0.100</u>	<u>0.680</u>	<u>4.307</u>	<u>0.175</u>	<u>0.898</u>	<u>0.966</u>	**0.984**
Dehazeformer	45.84	52.0	33.2	0.105	0.711	4.431	0.182	0.889	0.964	<u>0.983</u>
KA-Net	45.47	50.5	31.2	0.108	0.749	4.554	0.187	0.884	0.960	0.982
DCMPNet	<u>46.92</u>	<u>53.1</u>	<u>34.2</u>	0.108	0.725	4.579	0.190	0.883	0.959	0.980
DiffDehaze	41.84	42.1	26.1	0.120	0.849	5.400	0.204	0.849	0.950	0.980
IPC	45.90	52.3	33.9	**0.099**	0.688	**4.273**	**0.174**	**0.902**	**0.967**	**0.984**
Ours	**50.34**	**54.7**	**35.7**	**0.099**	**0.662**	4.310	**0.174**	<u>0.898</u>	**0.967**	**0.984**

Table 2. Comparison of downstream task performance across semantic segmentation (SS), object detection (OD), and depth estimation (DE) on Setting 1. Best and runner-up values are highlighted in bold and underlined, respectively.

downstream tasks. Unlike traditional dehazing models that operate in a static manner, our approach regulates the dehazing process in real time by leveraging feedback from downstream task performance and user instructions. Such a mechanism not only enhances the capacity of the model for task-oriented collaboration but also establishes a novel paradigm for building interactive and controllable intelligent visual systems.

Experiments

Experimental Settings

Datasets. We train our model using 5,000 randomly sampled images from each of the ADE20K(Zhou et al. 2017), COCO(Lin et al. 2014), and KITTI(Geiger et al. 2013) datasets, where hazy-clear image pairs are synthesized based on the atmospheric scattering model. For evaluation, we use the original test sets, which consist of 2,000, 5,000, and 697 images, respectively.

Implementation Details. We first train the IDN and freeze its parameters upon convergence. Then, only the TFGA and IGM modules are trained. Both stages adopt the Adam optimizer (Adam et al. 2014) with $\beta_1 = 0.9$, $\beta_2 = 0.999$, an initial learning rate of 1.0×10^{-4}, and a cosine annealing schedule. To enhance sample diversity, we apply data augmentation strategies similar to those used in SegFormer, YOLOv5, and RADepth across all three training sets. The initial dehazing stage is trained for 300 epochs, followed by 100 epochs for the task feedback stage. All experiments are conducted using PyTorch 1.8.0 on an NVIDIA GeForce RTX 3090 GPU with 24 GB of memory.

Evaluation Metrics. Following existing methods, we adopt PSNR, SSIM, and LPIPS (Zhang et al. 2018) to evaluate dehazing quality. To further demonstrate the effectiveness of our method for downstream tasks, we use the following metrics: Mean Intersection-over-Union (mIoU) (Yu et al. 2021) for semantic segmentation, Mean Average Precision (mAP) and mAP@[0.5:0.95) (He and Todorovic 2022) for object detection, and four error metrics (AbsRel, SqRel, RMSE, and RMSElog) and along with three accuracy metrics ($\delta < 1.25$, $\delta < 1.25^2$, and $\delta < 1.25^3$) (Eigen, Puhrsch, and Fergus 2014) for depth estimation.

Comparison with State-of-the-arts

We compare the proposed method with eight state-of-the-art approaches: Dehamer, C2P, RIDCP, Dehazeformer, KA-Net, DCMPNet, DiffDehaze, and IPC. Specifically, we design three experimental settings: **Setting 1** trains specific downstream tasks using the dehazing results from each comparison method; **Setting 2** directly inputs the dehazed results into downstream task networks for testing; **Setting 3** finetunes the dehazing network based on the performance feedback from the downstream tasks. The results for Setting 1 are shown in Figure 5, Table 1, and Table2, while those for **Settings 2** and **3** are in the **supplementary materials**.

As shown in Figure 5, our method demonstrates clear advantages in detail preservation, color fidelity, and brightness restoration. The dehazed images produced by our method exhibit visual quality closer to the ground truth (GT). Moreover, the results for downstream tasks in Figure 5 show that our method matches or even surpasses the performance of models specifically trained for each task. This demonstrates

Figure 6: Visual comparison of dehazing results from ablation studies.

Methods	ADE20K			COCO			KITTI		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Baseline	23.47	0.9255	0.0791	22.83	0.9058	0.1151	25.37	0.9091	0.0701
Baseline + FFM	25.27	0.9522	0.0460	24.73	0.9368	0.0860	27.08	0.9333	0.0353
Baseline + FFM + IGM	26.36	0.9646	0.0367	25.89	0.9515	0.0749	29.00	0.9664	0.0241
Baseline + FFM + TFGA	26.62	0.9672	0.0333	26.42	0.9550	0.0741	29.30	0.9688	0.0223
Full Model (All Modules)	**27.47**	**0.9701**	**0.0293**	**27.14**	**0.9587**	**0.0698**	**30.50**	**0.9740**	**0.0176**

Table 3. Ablation studies on the dehazing performance.

Methods	SS	OD		DE (Error Metric ↓)				DE (Accuracy Metric ↑)		
	mIoU↑	mAP↑	mAP50-95↑	AbsRel	SqRel	RMSE	RMSElog	δ<1.25	δ<1.25^2	δ<1.25^3
Baseline	49.42	53.5	34.8	0.113	0.753	4.539	0.172	0.874	0.941	0.982
Baseline + FFM	50.25	54.6	35.6	0.100	0.666	4.321	0.175	0.897	0.966	**0.984**
Baseline + FFM + IGM	50.28	54.6	35.5	0.100	0.664	4.318	0.175	**0.898**	0.966	**0.984**
Baseline + FFM + TFGA	50.29	54.6	**35.7**	0.100	0.663	4.316	0.175	**0.898**	**0.967**	**0.984**
Full Model (All Modules)	**50.34**	**54.7**	**35.7**	**0.099**	**0.662**	**4.310**	**0.174**	**0.898**	**0.967**	**0.984**

Table 4. Ablation experiments on downstream tasks.

its strong adaptability to the diverse requirements of various downstream tasks. We further conduct quantitative evaluations using objective metrics, as summarized in Table 1. Our method achieves the best performance on most metrics, validating its superiority. Quantitative results in Table 2, obtained by applying downstream task models to the dehazed outputs of different methods, further demonstrate the strong adaptability of our method.

Ablation Studies

The proposed method consists of three modules: TFGA, IGM, and FFM. To evaluate the contribution of each module to overall model performance, we conduct ablation studies using the ADE20K, COCO, and KITTI datasets. Specifically, the baseline model is obtained by removing both TFGA and IGM from the complete model and replacing the FFM module with a simple summation. The ablation experiment results are presented in Figure 6 and Tables 3 and 4.

Effectiveness of FFM: As shown in Table 3 and 4, introducing FFM into the baseline model results in the noticeable gains across all evaluation metrics. These results demonstrate the effectiveness of FFM in feature fusion.

Effectiveness of IGM: To assess the impact of IGM on model performance, we replace it with a conventional structure consisting of feature concatenation followed by convolution. The results in Table 4 show that incorporating IGM into the baseline+FFM configuration improves performance across all downstream tasks, confirming its effectiveness in enhancing task adaptability.

Effectiveness of TFGA: To verify the effectiveness of TFGA, we replace it with a simple addition operation. As indicated in Table 3 and 4, adding TFGA to the Baseline+FFM setting further improves performance on various metrics, demonstrating its effectiveness in task-guided optimization. In addition, as shown in Figure 6, the introduction of each module also improves the quality of the dehazed image, further confirming the effectiveness of each module.

Conclusion

This paper presents a novel adaptive dehazing framework that overcomes the limitations of conventional methods through a dynamic, task-aware, and instruction-driven design. Unlike existing methods that generate static dehazing results, our approach supports real-time adaptation without model retraining, making it well suited for deployment in dynamic, multi-task environments. The framework incorporates a closed-loop optimization strategy supported by two complementary modules: TFGA and IGM. Together, they form a dual-guidance mechanism that enables the model to adaptively tailor its outputs to the specific requirements of diverse downstream tasks during inference. Extensive experiments on object detection, semantic segmentation, and depth estimation demonstrate the effectiveness, robustness, and generalizability of our approach. While our method supports joint modeling and adaptation across multiple downstream tasks, it has only been evaluated on a fixed set of tasks. In real-world scenarios, task types and requirements may change dynamically, posing greater adaptation challenges. Future work will focus on exploring the model's generalization and adaptability to evolving task compositions.

Acknowledgments

This work was supported in part by the National Science Foundation of China under Grant 62571222 and Grant 62161015, the Yunnan Fundamental Research Projects under Grant 202301AV070004 and Grant 202501AS070123, the Major Science and Technology Special Projects of Yunnan Province under Grant 202502AD080006.

References

Adam, K. D. B. J.; et al. 2014. A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Cai, B.; Xu, X.; Jia, K.; Qing, C.; and Tao, D. 2016. DehazeNet: An End-to-End System for Single Image Haze Removal. *IEEE Transactions on Image Processing*, 25(11): 5187–5198.

Eigen, D.; Puhrsch, C.; and Fergus, R. 2014. Depth map prediction from a single image using a multi-scale deep network. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, volume 2, 2366–2374.

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*, 46(12): 9303–9320.

Fu, J.; Liu, S.; Liu, Z.; Guo, C.-L.; Park, H.; Wu, R.; Wang, G.; and Li, C. 2025. Iterative Predictor-Critic Code Decoding for Real-World Image Dehazing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 12700–12709.

Geiger, A.; Lenz, P.; Stiller, C.; and Urtasun, R. 2013. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11): 1231–1237.

Guo, C.; Yan, Q.; Anwar, S.; Cong, R.; Ren, W.; and Li, C. 2022. Image Dehazing Transformer with Transmission-Aware 3D Position Embedding. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5802–5810.

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

He, L.; and Todorovic, S. 2022. DESTR: Object Detection with Split Transformer. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9367–9376.

He, M.; Hui, L.; Bian, Y.; Ren, J.; Xie, J.; and Yang, J. 2022. Ra-depth: Resolution adaptive self-supervised monocular depth estimation. In *European Conference on Computer Vision*, 565–581. Springer.

Lan, Y.; Cui, Z.; Liu, C.; Peng, J.; Wang, N.; Luo, X.; and Liu, D. 2025. Exploiting diffusion prior for real-world image dehazing with unpaired training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 4455–4463.

Li, B.; Peng, X.; Wang, Z.; Xu, J.; and Feng, D. 2017. AOD-Net: All-in-One Dehazing Network. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 4780–4788.

Li, H.; Gao, J.; Zhang, Y.; Xie, M.; and Yu, Z. 2022. Haze transfer and feature aggregation network for real-world single image dehazing. *Knowledge-Based Systems*, 251: 109309.

Li, S.; Zhou, Y.; Ren, W.; and Xiang, W. 2023. PFONet: A Progressive Feedback Optimization Network for Lightweight Single Image Dehazing. *IEEE Transactions on Image Processing*, 32: 6558–6569.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European Conference*, 740–755. Springer.

Ren, W.; Pan, J.; Zhang, H.; Cao, X.; and Yang, M.-H. 2020. Single image dehazing via multi-scale convolutional neural networks with holistic edges. *International Journal of Computer Vision*, 128: 240–259.

Shao, Y.; Li, L.; Ren, W.; Gao, C.; and Sang, N. 2020. Domain Adaptation for Image Dehazing. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2805–2814.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Song, Y.; He, Z.; Qian, H.; and Du, X. 2023. Vision Transformers for Single Image Dehazing. *IEEE Transactions on Image Processing*, 32: 1927–1941.

Sun, S.; Ren, W.; Wang, T.; and Cao, X. 2022. Rethinking Image Restoration for Object Detection. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, volume 35, 4461–4474.

Wan, Y.; Li, J.; Lin, L.; Yuan, Q.; and Shen, H. 2025. Collaboration of Dehazing and Object Detection Tasks: A Multi-Task Learning Framework for Foggy Image. *IEEE Transactions on Geoscience and Remote Sensing*, 63: 5615413.

Wang, R.; Zheng, Y.; Zhang, Z.; Li, C.; Liu, S.; Zhai, G.; and Liu, X. 2025. Learning Hazing to Dehazing: Towards Realistic Haze Generation for Real-World Image Dehazing. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 23091–23100.

Wu, H.; Qu, Y.; Lin, S.; Zhou, J.; Qiao, R.; Zhang, Z.; Xie, Y.; and Ma, L. 2021. Contrastive Learning for Compact Single Image Dehazing. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10546–10555.

Wu, R.-Q.; Duan, Z.-P.; Guo, C.-L.; Chai, Z.; and Li, C. 2023. RIDCP: Revitalizing Real Image Dehazing via High-Quality Codebook Priors. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 22282–22291.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, volume 34, 12077–12090.

Yang, Z.; Huang, J.; Chang, J.; Zhou, M.; Yu, H.; Zhang, J.; and Zhao, F. 2023. Visual recognition-driven image restoration for multiple degradation with intrinsic semantics recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14059–14070.

Yu, C.; Gao, C.; Wang, J.; Yu, G.; Shen, C.; and Sang, N. 2021. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *International journal of computer vision*, 129: 3051–3068.

Zhang, H.; and Patel, V. M. 2018. Densely Connected Pyramid Dehazing Network. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3194–3203.

Zhang, J.; and Tao, D. 2020. FAMED-Net: A Fast and Accurate Multi-Scale End-to-End Dehazing Network. *IEEE Transactions on Image Processing*, 29: 72–84.

Zhang, L.; Wang, S.; and Wang, X. 2021. Single image dehazing based on bright channel prior model and saliency analysis strategy. *IET Image Processing*, 15(5): 1023–1031.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 586–595.

Zhang, Y.; Zhou, S.; and Li, H. 2024. Depth Information Assisted Collaborative Mutual Promotion Network for Single Image Dehazing. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2846–2855.

Zhang, Z.; Zhao, L.; Liu, Y.; Zhang, S.; and Yang, J. 2020. Unified density-aware image dehazing and object detection in real-world hazy scenes. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*.

Zheng, Y.; Zhan, J.; He, S.; Dong, J.; and Du, Y. 2023. Curricular Contrastive Regularization for Physics-Aware Single Image Dehazing. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5785–5794.

Zhou, B.; Zhao, H.; Puig, X.; Fidler, S.; Barriuso, A.; and Torralba, A. 2017. Scene Parsing through ADE20K Dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5122–5130.