Real-world Image Dehazing with Coherence-based Pseudo Labeling and Cooperative Unfolding Network

Chengyu Fang[1,*], **Chunming He**[1,3,*,†], **Fengyang Xiao**[1,2], **Yulun Zhang**[4,†], **Longxiang Tang**[1], **Yuelin Zhang**[5], **Kai Li**[6], **Xiu Li**[1,†]
[1]Shenzhen International Graduate School, Tsinghua University, [2]Sun Yat-sen University,
[3]Duke University, [4]Shanghai Jiao Tong University,
[5]The Chinese University of Hong Kong, [6] Meta Reality Labs
✉ *chengyufang.thu@gmail.com*

Abstract

Real-world Image Dehazing (RID) aims to alleviate haze-induced degradation in real-world settings. This task remains challenging due to the complexities in accurately modeling real haze distributions and the scarcity of paired real-world data. To address these challenges, we first introduce a cooperative unfolding network that jointly models atmospheric scattering and image scenes, effectively integrating physical knowledge into deep networks to restore haze-contaminated details. Additionally, we propose the first RID-oriented iterative mean-teacher framework, termed the Coherence-based Label Generator, to generate high-quality pseudo labels for network training. Specifically, we provide an optimal label pool to store the best pseudo-labels during network training, leveraging both global and local coherence to select high-quality candidates and assign weights to prioritize haze-free regions. We verify the effectiveness of our method, with experiments demonstrating that it achieves state-of-the-art performance on RID tasks. Code will be available at https://github.com/cnyvfang/CORUN-Colabator.

1 Introduction

Real-world image dehazing (RID) is a challenging task that aims to restore images affected by complex haze in real-world scenarios. The goal is to generate visual-appealing results while enhancing the performance of downstream tasks [1, 2]. The atmospheric scattering model (ASM), providing a physical framework for real-world dehazing, is formulated as follows:

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

where $P(x)$ and $J(x)$ are the hazy image and the haze-free counterpart. A signifies the global atmospheric light. $t(x)$ characterizes the transmission map reflecting varying degrees of haze visibility across different regions.

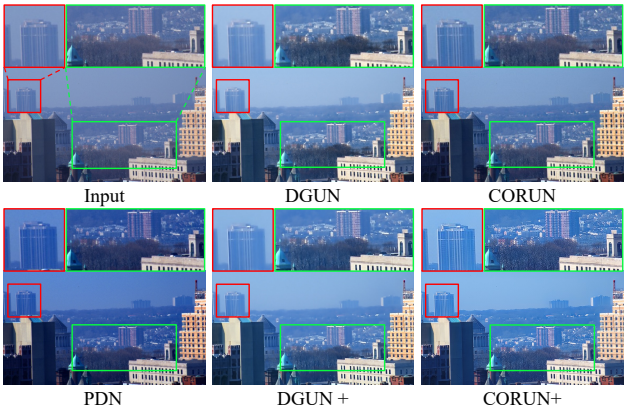

Figure 1: Results of cutting-edge methods. Our CORUN better restores hazy-contaminated details. Furthermore, techniques optimized by our Colabator framework, indicated by a "+" suffix, exhibit strong generalization in haze removal and color correction.

*Equal Contribution, † Corresponding Author, ✉ Email Address

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Conventional methods [3, 4] are limited by fixed feature extractors, which struggle to handle the complexities of real haze. Although existing deep learning-based methods [5–9] demonstrate improved performance, they face two significant challenges: (1) These methods do not accurately model the complex distribution of haze, leading to color distortion, as illustrated in fig. 1 DGUN [10]. (2) Real-world settings lack sufficient paired data for network training while optimizing the network with synthesized data brings a domain gap, limiting the generalizability of the models.

To overcome the first challenge, PDN [11] first introduces unfolding network [12, 13] to the RID field. In specific, PDN unfolds the iterative optimization steps of an ASM-based solution into a deep network for end-to-end training, incorporating physical information into the deep network. However, PDN does not effectively leverage the complementary information between the dehazed image and the transmission map, bringing overfitting problems and resulting in detail blurring (see fig. 1).

In this paper, we introduce the COopeRative Unfolding Network (CORUN), also derived from the ASM-based function, to address PDN's limitations and better model real hazy distribution. CORUN cooperatively models the atmospheric scattering and image scene by incorporating Transmission and Scene Gradient Descent Modules at each stage, corresponding to each iteration of the traditional optimization algorithm. To prevent overfitting, we introduce a global coherence loss, which constrains the entire pipeline to adhere to physical laws while alleviating constraints on the intermediate layers. These design choices collectively ensure that CORUN effectively integrates physical information into deep networks, thereby excelling in restoring haze-contaminated details, as depicted in fig. 1.

To enhance generalizability in real-world scenarios, we introduce the first RID-oriented iterative mean-teacher framework, named Coherence-based label generator (Colabator), designed to generate high-quality dehazed images as pseudo labels for training dehazing methods. Specifically, Colabator employs a teacher network, a dehazing network pretrained on synthesized datasets, to generate dehazed images on label-free real-world datasets. These restored images are stored in a dynamically updated label pool as pseudo labels for training the student network, which shares the same structure as the teacher network but with distinct weights. During network training, the teacher network generates multiple pseudo labels for a single real-world hazy image. We propose selecting the best labels to store in the label pool based on visual fidelity and dehazing performance.

To achieve this, we design a compound image quality assessment strategy tailored to the dehazing task, evaluating the global coherence of the dehazed images and selecting the most visually appealing ones without distortions for inclusion in the label pool. Additionally, we propose a patch-level certainty map to encourage the network to focus on well-restored regions of the dehazed pseudo labels, effectively constraining the local coherence between the outputs of the student model and the teacher model. As shown in fig. 1, Colabator, generating high-quality pseudo labels for network training, enhances the student dehazing network's capacity for haze removal and color correction.

Our contributions are summarized as follows:

(1) We propose a novel dehazing method, CORUN, to cooperatively model the atmospheric scattering and image scene, effectively integrating physical information into deep networks.

(2) We propose the first iterative mean-teacher framework, Colabator, to generate high-quality pseudo labels for network training, enhancing the network's generalization in haze removal.

(3) We evaluate our CORUN with the Colabator framework on real-world dehazing tasks. Abundant experiments demonstrate that our method achieves state-of-the-art performance.

2 Related Works

2.1 Real-world Image Dehazing

The dissonance between synthetic and real haze distributions often hinders existing Learning-based dehazing methods [14–18] from effectively dehazing real-world images. Consequently, there's a growing emphasis on tackling challenges specific to real-world dehazing [19–23].

Given the characteristics of real haze, RIDCP [7] and Wang *et al.* [24] proposed novel haze synthesis pipelines. However, relying solely on synthetic data limits models' robustness in real-world dehazing scenarios. Recognizing the distributional disparities between synthetic and real haze, methods like CDD-GAN [25], D4 [26], Shao *et al.* [27], and Li *et al.* [28] have utilized CycleGAN [29] for

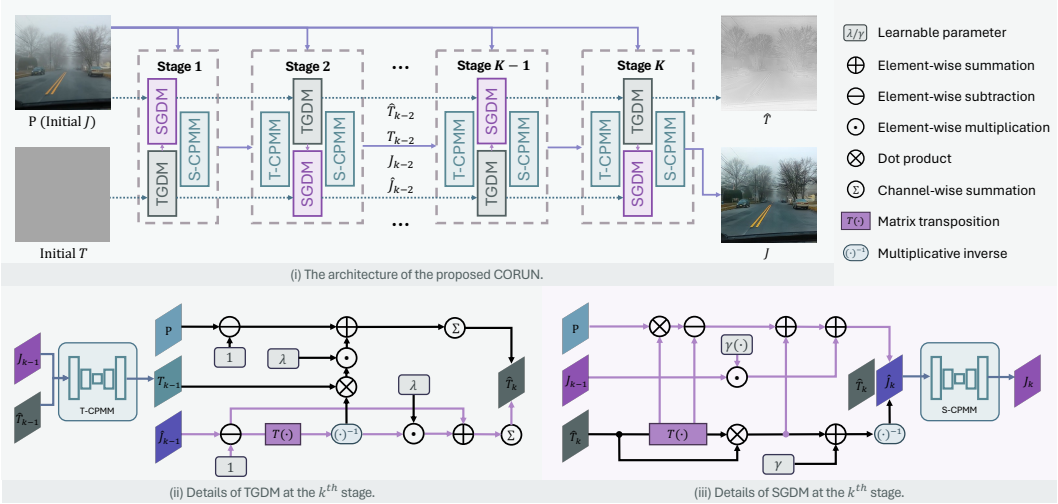

Figure 2: The architecture of the proposed CORUN with the details at k^{th} stage.

dehazing. Despite this, the challenges inherent in GAN [30] training often result in artifacts. Some approaches combine synthetic and real-world data, applying unsupervised loss to supervise real-world dehazing learning [19]. However, these losses lack sufficient precision, leading to suboptimal results. Other methods leverage pseudo-labels [31, 32], but the erroneous pseudo-labels cause degrade quality.

To address these challenges, we introduce a coherence-based pseudo labeling method termed Colabator. Our approach selectively identifies and prioritizes high-quality regions within pseudo labels, leading to enhanced robustness and superior generation quality for real-world image dehazing.

2.2 Deep Unfolding Image Restoration

Deep Unfolding Networks (DUNs) integrate model-based and learning-based approaches [33, 34] and thus offer enhanced interpretability and flexibility compared to traditional learning-based methods. Increasingly, DUNs are being utilized for various image tasks, including image super-resolution [35], compressive sensing [36], and hyperspectral image reconstruction [37]. DGUN [10] proposes a general form of proximal gradient descent to learn degradation. However, it fails to decouple prior knowledge, relying solely on single-path DUN to model degradation and construct mappings, posing challenges in comprehending complex degradation. Yang and Sun first introduced DUNs to the image dehazing field and proposed PDN [11]. However, PDN does not exploit the complementary information between the dehazed image and the transmission map, resulting in detail blurring. Our CORUN optimizes the atmospheric scattering model and the image scene feature through dual proximal gradient descent, thus preventing overfitting and facilitating detail restoration.

3 Methodology

3.1 Cooperative Unfolding Network

We propose the Cooperative Unfolding Network (CORUN), the first Deep Unfolding Network (DUN) method utilizing Proximal Gradient Descent (PGD) to optimize image dehazing performance by levaraging the Atmospheric Scattering Model (ASM) and neural image reconstruction in a cooperative manner. Each stage of CORUN includes Transmission and Scene Gradient Descent Modules (T&SGDM) paired with Cooperative Proximal Mapping Modules (T&S-CPMM). These modules work together to model atmospheric scattering and image scene features, enabling the adaptive capture and restoration of global composite features within the scene.

According to eq. (1), given a hazy image $\mathbf{P} \in \mathbb{R}^{H\times W\times 3}$, we initialize a transmission map $\mathbf{T} \in \mathbb{R}^{H\times W\times 1}$. In gradient descent, we simplify the atmospheric light $A \in \mathbb{R}^3$ and implicitly estimate it in the CORUN pipeline to focus on the detailed characterization of the scene and the relationship between volumetric haze and scene. Hence, eq. (1) can be rewrite as

$$\mathbf{P} = \mathbf{J} \cdot \mathbf{T} + \mathbf{I} - \mathbf{T}, \tag{2}$$

Where $\mathbf{J}$ means the clear image without hazy, $\mathbf{I}$ is the all-one matrix. Based on eq. (2), we can define our cooperative dehazing energy function like

$$L(\mathbf{J}, \mathbf{T}) = \frac{1}{2}\|\mathbf{P} - \mathbf{J} \cdot \mathbf{T} + \mathbf{T} - \mathbf{I}\|_2^2 + \psi(\mathbf{J}) + \phi(\mathbf{T}), \tag{3}$$

where $\psi(\mathbf{J})$ and $\phi(\mathbf{T})$ are regularization terms on $\mathbf{T}$ and $\mathbf{J}$. We introduce two auxiliary variables $\hat{\mathbf{T}}$ and $\hat{\mathbf{J}}$ to approximate $\mathbf{T}$ and $\mathbf{J}$, respectively. This leads to the following minimization problem:

$$\{\hat{\mathbf{J}}, \hat{\mathbf{T}}\} = \arg\min_{\mathbf{J},\mathbf{T}} L(\mathbf{J}, \mathbf{T}). \tag{4}$$

Transmission optimization. Give the estimated coarse transmission map $\mathbf{T}$ and dehazed image $\hat{\mathbf{J}}_{k-1}$ at iteration $k-1$, the variable $\mathbf{T}$ can be updated as:

$$\mathbf{T}_k = \arg\min_{\mathbf{T}} \frac{1}{2}\left\|\mathbf{P} - \hat{\mathbf{J}}_{k-1} \cdot \mathbf{T} + \mathbf{T} - \mathbf{I}\right\|_2^2 + \phi(\mathbf{T}). \tag{5}$$

We construct the proximal mapping between $\hat{\mathbf{T}}$ and $\mathbf{T}$ by a encoder-decoder like neural network which we named T-CPMM and denoted as prox_ϕ:

$$\mathbf{T}_k = \text{prox}_\phi(\mathbf{J}_{k-1}, \hat{\mathbf{T}}_k), \tag{6}$$

the auxiliary variables $\hat{\mathbf{T}}$, which we calculate by our proposed TGDM can be formulated as:

$$\hat{\mathbf{T}}_k = \sum_{c\in\{R,G,B\}} (\mathbf{I} - \hat{\mathbf{J}}^c_{k-1} + \frac{\lambda_k}{(\mathbf{I} - \hat{\mathbf{J}}^c_{k-1})^\top})^{-1} \cdot (\mathbf{I} - \mathbf{P}^c + \frac{\lambda_k \mathbf{T}_{k-1}}{(\mathbf{I} - \hat{\mathbf{J}}^c_{k-1})^\top}). \tag{7}$$

The variable λ_k is a learnable parameter, we enable CORUN to learn this parameter at each stage during the end-to-end learning process, allowing the network to adaptively control the updates in iteration.

Scene optimization. Give $\hat{\mathbf{T}}_k$ and $\mathbf{J}$, the variable $\mathbf{J}$ can be updated as:

$$\mathbf{J}_k = \arg\min_{\mathbf{J}} \frac{1}{2}\|\mathbf{P} - \mathbf{J} \cdot \hat{\mathbf{T}}_k + \hat{\mathbf{T}}_k - \mathbf{I}\|_2^2 + \psi(\mathbf{J}). \tag{8}$$

Same as the proximal mapping process in the transmission optimization, S-CPMM has the similar structure as T-CPMM but different inputs, we denote S-CPMM as prox_ψ:

$$\mathbf{J}_k = \text{prox}_\psi(\hat{\mathbf{J}}_k, \hat{\mathbf{T}}_k), \tag{9}$$

where the $\hat{\mathbf{J}}_k$ we process by our SGDM can be presented as:

$$\hat{\mathbf{J}}_k = (\hat{\mathbf{T}}_k^\top \hat{\mathbf{T}}_k + \mu_k \mathbf{I})^{-1} \cdot (\hat{\mathbf{T}}_k^\top \mathbf{P} + \hat{\mathbf{T}}_k^\top \hat{\mathbf{T}}_k - \hat{\mathbf{T}}_k^\top + \mu_k \mathbf{J}_{k-1}), \tag{10}$$

as the λ_k in transmission optimization, μ_k is also a learnable parameter to bring more generalization capabilities to the network.

Details about CPMM. T-CPMM and S-CPMM share the same structure for improved mapping quality. Each CPMM block uses a 4-channel convolution to embed $\mathbf{T}$ and $\mathbf{J}$ into a 30-dimensional feature map. The distinction between T-CPMM and S-CPMM lies in their outputs: T-CPMM produces a 1-channel result to aid TGDM in predicting a scene-compliant transmission map, whereas S-CPMM generates a 3-channel RGB image. This enables S-CPMM to learn additional scene feature information, such as atmospheric light and blur, assisting SGDM in generating higher-quality dehazed results with more details. For more efficient computation, each CPMM comprises only 3 layers with $[1, 1, 1]$ blocks, doubling the dimensions with increasing depth.

3.2 Coherence-based Pseudo Labeling by Colabator

We generate and select pseudo labels using our proposed plug-and-play coherence-based label generator, Colabator. Colabator consists of a teacher network with weights θ_{tea} shared with the student network θ_{stu} via exponential moving average (EMA). It employs a tailored mean-teacher strategy with a trust weighting process and an optimal label pool to generate high-quality pseudo labels, addressing the scarcity of real-world data. Figure 3 illustrates the pipeline of our Colabator.

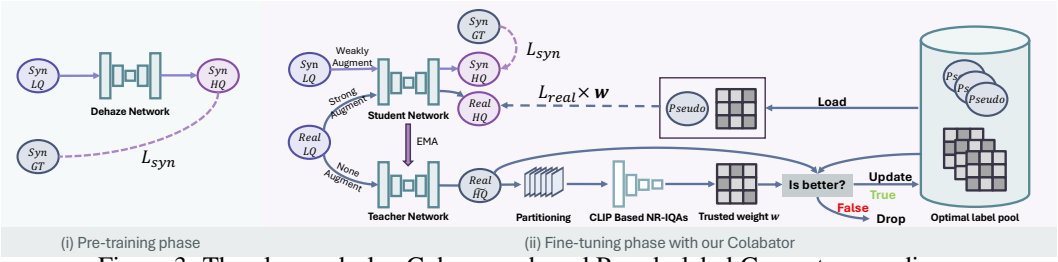

Figure 3: The plug-and-play Coherence-based Pseudo-label Generator paradigm.

Iterative mean-teacher dehazing. Given a real hazy image $\mathbf{P}^R_{LQ} \in \mathbb{R}^{H\times W\times 3}$, we initially apply augmentations to generate corresponding strongly degraded data using a strong augmentor $\mathcal{A}_s(\cdot)$, which randomly applies adjustments such as contrast, brightness, posterize, sharpness, JPEG compression, and Gaussian blur. Unlike the common mean-teacher strategy, we omit functions like solarize, equalize, shear, and translate to prevent unnecessary degradation that might mislead model learning. We use the non-augmented image as the input for the teacher network and the strongly augmented image for the student network, generating the following results:

$$\mathbf{P}^R_{\widetilde{HQ}}, \mathbf{T}^R_{\widetilde{HQ}} = f_{\theta_{tea}}(\mathbf{P}^R_{LQ}), \quad \mathbf{P}^R_{HQ}, \mathbf{T}^R_{HQ} = f_{\theta_{stu}}(\mathcal{A}_s(\mathbf{P}^R_{LQ})), \tag{11}$$

where $\mathbf{P}^R_{\widetilde{HQ}} \in \mathbb{R}^{H\times W\times 3}$ is the result from the teacher network using the non-augmented input, and $\mathbf{P}^R_{HQ} \in \mathbb{R}^{H\times W\times 3}$represents the result from the student network by strong augment input and $\mathbf{T}^R_{\widetilde{HQ}}$, $\mathbf{T}^R_{HQ}$ are the corresponding transmission map. The different degrees of data augmentation lead to varying dehazing results, typically resulting in $\mathbf{P}^R_{\widetilde{HQ}}$ having better quality than $\mathbf{P}^R_{HQ}$.

This approach ensures the model descends in the correct direction and helps mitigate the overfitting issues often associated with direct pseudo-label learning methods. By iterating, our teacher network generates increasingly high-quality pseudo-labels, providing more reliable supervision.

Label trust weighting. To better leverage the pseudo-dehazed images $\mathbf{P}^R_{\widetilde{HQ}}$ generated by the teacher network for model supervision, we designed a composite image quality assessment strategy for further processing these pseudo-dehazed images and get the trusted weight w which means the reliability of each location of an image. Our composite strategy primarily consists of a haze density evaluator $\mathcal{D}(\cdot)$ based on pre-trained CLIP [38] model and fixed text feature, and a non-reference image quality evaluator $\mathcal{Q}(\cdot)$. We partition $\mathbf{P}^R_{\widetilde{HQ}}$ into an sequence $\mathbf{S}^R_{\widetilde{HQ}} \in \mathbb{R}^{N\times N\times 3\times (H/N)\times (W/N)}$ and use $\mathcal{D}(\cdot)$ and $\mathcal{Q}(\cdot)$ to predict the density score and quality score. The final trusted weight w we can get from:

$$w = \Psi(\text{norm}(\mathcal{D}(\mathbf{S}^R_{\widetilde{HQ}})) \cdot \text{norm}(\mathcal{Q}(\mathbf{S}^R_{\widetilde{HQ}})), \tag{12}$$

where Ψ is compose sequence to map and resize as $\mathbf{P}^R_{\widetilde{HQ}}$, norm$(\cdot)$ means normalize scores from 0 to 1, that higher score means lower haze density and better image quality.

Optimal label pool. To ensure the use of optimal pseudo-labels and avoid domain adaptation collapse due to instability during training, we proposed an optimal label pool $\mathcal{P}$ to maintain the pseudo-labels in their optimal state. The overall procedure of our optimal label pool process is summarized in algorithm 1, compare pseudo-dehazed image $\mathbf{P}^R_{\widetilde{HQ}_i}$ with previous pseudo-label $\mathbf{P}^R_{Psei}$ and update pseudo-dehazed image as pseudo-label if it better than previous. To summarize the algorithm 1 and eq. (11), the overall process of Colabator can be formalize as:

$$\mathbf{P}^R_{HQ}, \mathbf{T}^R_{HQ}, \mathbf{P}^R_{Pse}, \mathbf{T}^R_{Pse}, w_{pse} = \mathcal{C}(\mathbf{P}^R_{LQ}, \theta_{tea}, \theta_{stu}, \mathcal{A}_s, \mathcal{D}(\cdot), \mathcal{Q}(\cdot), \mathcal{P}), \tag{13}$$

where $\mathcal{C}$ is our Colabator framework, $\mathbf{P}^R_{Pse}$ is the paired pseudo label of $\mathcal{A}_s(\mathbf{P}^R_{LQ})$, $\mathbf{T}^R_{Pse}$ is the corresponding pesudo transmission map, w_{pse} means the trusted weight of the pseudo label.

Weights update. The teacher network's weights θ_{tea} are updated by exponential moving average (EMA) of the student network's weights θ_{stu}, which is denoted as follows:

$$\theta_{tea} = \eta\theta_{tea} + (1-\eta)\theta_{stu}, \tag{14}$$

where η is momentum and $\eta \in (0, 1)$. Using this update strategy, the teacher model can aggregate previously learned weights immediately after each training step, ensuring updating stability.

Algorithm 1 Optimal label pool process

Require: Haze density evaluator $\mathcal{D}(\cdot)$ and image quality evaluator $\mathcal{Q}(\cdot)$;
Optimal label pool $\mathcal{P}$;
Sample a batch of real hazy images $\{\mathbf{P}^R_{LQ_i}\}^b_{i=1}$;
for each $\mathbf{P}^R_{LQ_i}$ **do**
  Get teacher network prediction: $\mathbf{P}^R_{\widetilde{HQ}_i}, \mathbf{T}^R_{\widetilde{HQ}_i} = f_{\theta_{tea}}(\mathbf{P}^R_{LQ_i})$;
  Partition $\mathbf{P}^R_{\widetilde{HQ}_i}$ into $N \times N$ and get $\mathbf{S}^R_{\widetilde{HQ}_i}$;
  Compute score map of $\mathbf{S}^R_{\widetilde{HQ}_i}$: $d_i = \text{norm}(\mathcal{D}(\mathbf{S}^R_{\widetilde{HQ}_i}))$, and $q_i = \text{norm}(\mathcal{Q}(\mathbf{S}^R_{\widetilde{HQ}_i}))$;
  Load $\mathbf{P}^R_{Psei}, \mathbf{T}^R_{Psei}, w_{Psei}, d_{Psei}, q_{Psei} = \mathcal{P}(i)$
  if $d_i > d_{Psei}$ and $q_i > q_{Psei}$ **then**
    Compute trusted weight: $w_i = \Psi(d_i + q_i)$
    Update $\mathcal{P}(i) = (\mathbf{P}^R_{\widetilde{HQ}_i}, \mathbf{T}^R_{\widetilde{HQ}_i}, w_i, d_i, q_i)$
    Return $\mathbf{P}^R_{\widetilde{HQ}_i}, \mathbf{T}^R_{\widetilde{HQ}_i}, w_i$ as pesudo label.
  else
    Return $\mathbf{P}^R_{Psei}, \mathbf{T}^R_{Psei}, w_{pse_i}$ as pesudo label.
  end if
end for

3.3 Semi-supervised Real-world Image Dehazing

To achieve success in real-world dehazing, we designed several loss functions for our CORUN and Colabator to constrain their learning process. We introduce a reconstruction loss using the L_1 norm $\|\cdot\|_1$. To enhance visual perception, we employ contrastive and common perceptual regularization to ensure the consistency of the reconstruction results with the ground truth in terms of features at different levels. The perceptual loss is defined as follows:

$$L^{common}_{Rec}(\mathbf{P}_{HQ}, \mathbf{P}_{GT}) = \|\mathbf{P}_{GT}, \mathbf{P}_{HQ}\|_1 + \beta_c \sum_{i=1}^{n} \tau_i \|\varphi_i(\mathbf{P}_{GT}), \varphi_i(\mathbf{P}_{HQ})\|_1 \tag{15}$$

$$L^{contra}_{Rec}(\mathbf{P}_{LQ}, \mathbf{P}_{HQ}, \mathbf{P}_{GT}) = \|\mathbf{P}_{GT}, \mathbf{P}_{HQ}\|_1 + \beta_c \sum_{i=1}^{n} \tau_i \frac{\|\varphi_i(\mathbf{P}_{GT}), \varphi_i(\mathbf{P}_{HQ})\|_1}{\|\varphi_i(\mathbf{P}_{LQ}), \varphi_i(\mathbf{P}_{HQ})\|_1}, \tag{16}$$

where $\mathbf{P}_{HQ}$ is the dehazed result, $\varphi_i(\cdot)$ means the i_{th} hidden layer of pre-trained VGG-19 [39], τ_i is the weight coefficient. Besides, to constrain the entire pipeline to obey physical laws while alleviating constraints on the intermediate layers, and prevent overfitting, we introduce a global coherence loss:

$$L_{Coh}(\mathbf{P}_{LQ}, \mathbf{P}_{HQ}, \mathbf{T}_{HQ}) = \|(\mathbf{P}_{HQ} \odot \mathbf{T}_{HQ} + (\mathbf{I} - \mathbf{T}_{HQ})) - \mathbf{P}_{LQ}\|_1, \tag{17}$$

where $\odot$ is the Hadamard product, $\mathbf{I}$ means the all-ones matrix as the same size of $\mathbf{P}^S_{LQ}$. The global coherence loss ensures that CORUN can more efficiently integrate physical information into the deep network to facilitate the recovery of more physically consistent details. In addition, we introduce a density loss L_{dens} based on $\mathcal{D}(\cdot)$ to score and constraint the model to dehaze in the semantic domain:

$$L_{Dens}(\mathbf{P}) = \mathcal{D}(\mathbf{P}). \tag{18}$$

Pre-training phase. To ensure the capacity in dehazing and transmission map estimation, we pre-trained CORUN on synthetic paired datasets which contained clear image $\mathbf{P}^S_{GT} \in \mathbb{R}^{H \times W \times 3}$ and synthetic hazy image $\mathbf{P}^S_{LQ} \in \mathbb{R}^{H \times W \times 3}$. Setting $\mathbf{P}^S_{LQ}$ as input, we can get the result by

$$\mathbf{P}^S_{HQ}, \mathbf{T}^S_{HQ} = f_{\theta_{stu}}(\mathcal{A}_w(\mathbf{P}^S_{LQ})), \tag{19}$$

where $\mathcal{A}_w$ means weakly geometric data augment, $\mathbf{P}^S_{HQ}$ means the dehazed result of synthetic hazy image, and $\mathbf{T}^S_{HQ}$ is the corresponding transmission map. In the pre-training phase, our CORUN is

Metrics	Hazy	PDN [11]	MBDN [14]	DH [15]	DAD [27]	PSD [19]	D4 [26]	RIDCP [7]	DGUN [10]	Ours
FADE↓	2.484	**0.876**	1.363	1.895	1.130	0.920	1.358	0.944	1.111	**0.824**
BRISQUE↓	36.642	30.811	27.672	33.862	32.241	27.713	33.210	**17.293**	27.968	**11.956**
NIMA↑	4.483	4.464	4.529	4.522	4.312	4.598	4.484	**4.965**	4.653	**5.342**

Table 1: Quantitative results on RTTS dataset. **Red** and **blue** indicate the best and the second best.

Figure 4: Visual comparison on RTTS[40]. Please zoom in for a better view.

optimized end-to-end using two supervised loss functions. The overall loss of the pre-training phase:

$$\begin{aligned} L_{pre} =& \rho_r L_{Rec}^{contra}(\mathcal{A}_w(\mathbf{P}_{LQ}^S), \mathbf{P}_{HQ}^S, \mathbf{P}_{GT}^S) \\ &+ \rho_c L_{Coh}(\mathcal{A}_w(\mathbf{P}_{LQ}^S), \mathbf{P}_{HQ}^S, \mathbf{T}_{HQ}^S) + L_{Dens}(\mathbf{P}_{HQ}^S), \end{aligned} \tag{20}$$

where ρ_r is the trade-off weight of L_{Rec}^{contra}, ρ_c is the trade-off weight of L_{Coh}.

Fine-tuning phase. In fine-tuning phase, we adapt our CORUN pre-trained on synthetic data to the real-world domain by our Colabator framework. For more steady learning, in this phase, we train with both synthetic and real-world data. As eq. (13), we generate $\mathbf{P}_{HQ}^R, \mathbf{T}_{HQ}^R, \mathbf{P}_{Pse}^R, \mathbf{T}_{Pse}^R, w_{pse}$ from $\mathbf{P}_{LQ}^R$, and we get $\mathbf{P}_{HQ}^S, \mathbf{T}_{HQ}^S$ use the eq. (19). The overall loss of the fine-tuning phase:

$$\begin{aligned} L_{fine} =& w\rho_r L_{Rec}^{contra}(\mathcal{A}_s(\mathbf{P}_{LQ}^R), \mathbf{P}_{HQ}^R, \mathbf{P}_{Pse}^R) + \rho_r L_{Rec}^{common}(\mathbf{P}_{HQ}^S, \mathbf{P}_{GT}^S) \\ &+ w\rho_c L_{Coh}(\mathcal{A}_s(\mathbf{P}_{LQ}^R), \mathbf{P}_{HQ}^R, \mathbf{T}_{HQ}^R) + L_{Dens}(\mathbf{P}_{HQ}^S) + wL_{Dens}(\mathbf{P}_{HQ}^R). \end{aligned} \tag{21}$$

4 Experiments

4.1 Experimental Setup

Data Preparation. We use RIDCP500 [7] dataset, comprising 500 clear images with depth maps estimated by [41], and follow the same way of RIDCP [7] for generating paired data. During the fine-tuning phase, we incorporate the URHI subset of RESIDE dataset [40], which only consists of 4,807 real hazy images, for generating pseudo-labels and fine-tuning the network. We evaluate our framework qualitatively and quantitatively on the RTTS subset, which comprises over 4,000 real hazy images featuring diverse scenes, resolutions, and degradation. Fattal's dataset [42], comprising 31 classic real hazy cases, serves as a supplementary source for cross-dataset visual comparison.

Implementation Details. Our framework is implemented using PyTorch [43] and trained on four NVIDIA RTX 4090 GPUs. During the pre-training phase, we train the network for 30K iterations,

Figure 5: Visual comparison on Fattal's data[42].

Datasets	Metrics	w/o Colabator DGUN	w/ Colabator DGUN	w/o Colabator CORUN	w/ Colabator CORUN (Ours)
RTTS	FADE↓	1.111	0.857	1.091	0.824
	BRISQUE↓	25.085	20.731	16.541	11.956
	NIMA↑	4.813	5.190	4.856	5.342

Table 2: Generalization and Effect of our Colabator.

Datasets	Metrics	w/o Mean-teacher	w/o Trusted weight	w/o Optimal label pool
RTTS	FADE↓	0.912	0.827	0.846
	BRISQUE↓	15.728	16.606	15.707
	NIMA↑	4.921	4.867	5.285

Table 3: Module's Effect of our Colabator.

Datasets	Metrics	Stages 1	2	4 (Ours)	6
RTTS	FADE↓	0.785	0.808	0.824	0.839
	BRISQUE↓	15.520	15.151	11.956	16.227
	NIMA↑	5.228	5.281	5.342	5.187

Table 4: Effect of stage number.

optimizing it with AdamW [44] using momentum parameters $(\beta_1 = 0.9, \beta_2 = 0.999)$ and an initial learning rate of 2×10^{-4}, gradually reduced to 1×10^{-6} with cosine annealing. In Colabator, the initial learning rate is set to 5×10^{-5} with only 5K iterations. Following [7], we employ random crop and flip for synthetic data augmentation. We use DA-CLIP [45] as our haze density evaluator and MUSIQ [46] as the image quality evaluator. Our CORUN consists of 4 stages and the trade-off parameters in the loss are set to β_c, ρ_r, ρ_c are set to $0.2, 5, 10^{-2}$, respectively.

Metrics. We utilize the Fog Aware Density Evaluator (FADE) [47] to assess the haze density in various methods. However, FADE focuses on haze density exclusively, overlooking other crucial image characteristics such as color, brightness, and detail. To address this limitation, we also employ Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [48], and Neural Image Assessment(NIMA) [49] for a more comprehensive evaluation of image quality and aesthetic. Higher NIMA scores, along with lower FADE and BRISQUE scores, indicate better performance. We use PyIQA [50] for BRISQUE and NIMA calculations, and the official MATLAB code for FADE calculations. All of these metrics are non-reference because there is no ground-truth in RTTS [40].

4.2 Comparative Evaluation

We compare our method with 8 state-of-the-art methods: PDN [11], MBDN [14], DH [15], DAD [27], PSD [19], D4 [26], RIDCP [7], DGUN [10]. The quantitative results, presented in table 1, show that our method achieved the highest performance, outperforming the second-best method (RIDCP) by 17.0%. Specifically, our method improved FADE, BRISQUE, and NIMA scores by 12.7%, 30.8%, and 7.6%, respectively. This demonstrates that our method surpasses current state-of-the-art techniques in both dehazing capability and the quality, and aesthetics of the generated images.

The visual comparisons of our proposed method and state-of-the-art algorithms are shown in figs. 4 and 5. We can observe that these methods have demonstrated some effectiveness in real-world dehazing tasks, but when images containing white objects, sky, or extreme haze, the results from PDN, DAD, PSD, and RIDCP exhibited varying degrees of dark patches and contrast inconsistencies. Conversely, D4 caused an overall reduction in brightness, leading to detail loss in darker areas. Under these conditions, DGUN produced relatively aesthetically pleasing results but lost significant local detail, impairing overall visual quality. Notably, PSD achieved higher brightness but suffered from

Dataset	PDN [11]	MBDN [14]	DH [15]	DAD [27]	PSD [19]	D4 [26]	RIDCP [7]	DGUN [10]	Ours
RTTS[40]	4.52	3.47	3.23	4.35	3.90	4.66	**7.14**	6.04	**7.76**
Fattal's[42]	4.85	3.33	3.19	4.80	4.28	4.38	**7.28**	6.33	**8.04**

Table 5: User study scores on RTTS[40] and Fattal's[42].

Class(AP)	Hazy	PDN [11]	MBDN [14]	DH [15]	DAD [27]	PSD [19]	D4 [26]	RIDCP [7]	DGUN [10]	Ours
Bicycle	0.51	0.55	0.54	0.47	0.52	0.52	0.54	**0.57**	0.55	**0.59**
Bus	0.25	0.29	0.27	0.23	0.29	0.25	0.28	**0.32**	**0.31**	**0.31**
Car	0.61	0.65	0.63	0.51	0.65	0.63	0.64	**0.67**	0.66	**0.68**
Motor	0.38	0.45	0.43	0.37	0.38	0.42	0.42	**0.47**	0.46	**0.49**
Person	0.73	**0.76**	0.75	0.69	0.74	0.74	0.75	**0.76**	**0.76**	**0.77**
Mean	0.50	0.54	0.52	0.45	0.52	0.51	0.53	**0.56**	0.55	**0.57**

Table 6: Object detection results on RTTS[40].

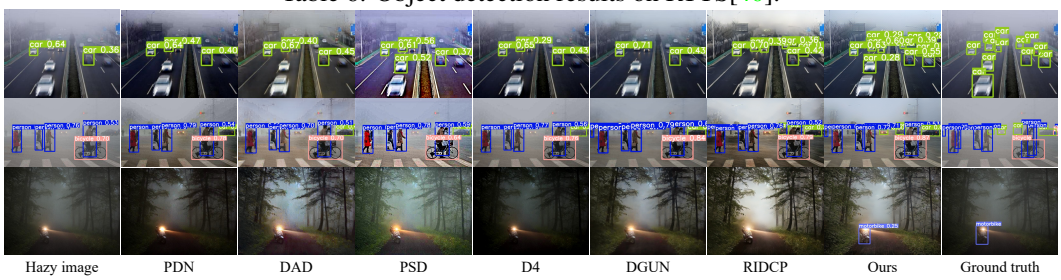

Figure 6: Visual comparison of object detection on RTTS [40].

severe oversaturation. CORUN+ consistently outperforms others by producing clearer images with natural colors and better contrast, effectively removing haze while preserving image details.

4.3 Ablation Study

Generalization and Effect of Colabator. We evaluates the performance and the impact of our proposed Colabator framework across different metrics. As shown in table 2, removing the fine-tuning phase of Colabator led to significant performance drops, highlighting its critical role in the dehazing process. To evaluate the generalizability of Colabator, we conducted additional experiments by replacing our CORUN with the DGUN [10], while maintaining consistent training settings. Results in table 2 and fig. 1 indicate that Colabator substantially enhances DGUN's performance, demonstrating its effectiveness as a plug-and-play paradigm with strong generalization capabilities.

Effect of Colabator. We validate the effect of our Colabator. In table 3, we systematically removed critical components, such as mean-teacher, trusted weighting, and the optimal label pool, from the model architecture. The outcomes indicate the performance deteriorates when these components are removed, highlighting their essential role in the system.

Ablations on stage number. The number of stages in a deep unfolding network significantly impacts its efficiency and performance. To investigate this, we experimented with different stage numbers for CORUN+, specifically choosing k values from the set $\{1, 2, 4, 6\}$. The results detailed in table 4, indicate that CORUN+ achieves high-quality dehazing with 4 stages. Notably, increasing the number of stages does not necessarily improve outcomes. Excessive stages can increase the network's complexity, hinder convergence, and potentially introduce errors in the results.

4.4 User Study and Downstream Task

User Study. We conducted a user study to evaluate the human subjective visual perception of our proposed method against other methods. We invited five experts with an image processing background and 16 naive observers as testers. These testers were instructed to focus on three primary aspects: (i) Haze density compared to the original hazy image, (ii) Clarity of details in the dehazed image, and (iii) Color and aesthetic quality of the dehazed image. The results for each method, along with the corresponding hazy images, were presented to the testers anonymously. They scored each method on a scale from 1 (worst) to 10 (best). The hazy images were selected randomly, with a total of 225 images from RTTS[51] and 54 images from Fattal's[42] dataset. The user study scores are reported in table 5, showing that our method achieved the highest average score.

Figure 7: Failure cases. Our results show low quality texture details.

Downstream Task Evaluation. The performance of high-level vision tasks, *e.g.* object detection and semantic segmentation, is greatly affected by image quality, with severely degraded images often leading to erroneous results [52, 53]. To address this performance degradation, some methods have incorporated image restoration as a preprocessing step for high-level vision tasks. To validate the effectiveness of our approach for high-level vision, we utilized pretrained YOLOv3 [54], and tested it on the RTTS [40] dataset, and evaluated the results using the mean Average Precision (mAP) metric. As shown in table 6 and fig. 6, our method demonstrates a substantial advantage over existing methods, verifying our efficacy in facilitating high-level vision understanding.

5 Limitations and Future Work

In fig. 7, our CORUN+ model struggles to maintain result quality and preserve texture details when dealing with severely degraded inputs, such as strong compression and extreme high-density haze. This challenge persists across existing methods and remains unresolved. We attribute this difficulty to the model's struggle in reconstructing scenes from dense haze, where information is often severely lacking or entirely lost, affecting the reconstruction of both haze-free and low haze density areas. Moreover, the model solely focuses on dehazing and lacks the capability to address other image degradations, such as image deblurring[55] and low-light image enhancement [56, 57], limiting its ability to achieve high-quality reconstruction results from complex degraded images. To address this limitation in future research, we propose not only focusing on environmental degradation but also considering additional information about image degradation when solving real-world dehazing problems. In addition to this, we can integrate robust generative methods to improve the network's ability to restore dense haze regions [58–62], synthesize haze that matches real-world distributions [63–66], and introduce more modalities as supplements to RGB images, enhancing the model's ability to effectively recover details [67].

6 Conclusions

In this paper, we introduce CORUN to cooperatively model atmospheric scattering and image scenes and thus incorporate physical information into deep networks. Furthermore, we propose Colabator, an iterative mean-teacher framework, to generate high-quality pseudo-labels by storing the best-ever results with global and local coherence in a dynamic label pool. Experiments demonstrate that our method achieves state-of-the-art performance in real-world image dehazing tasks, with Colabator also improving the generalization of other dehazing methods. The code will be released.

Acknowledgments and Disclosure of Funding

This work was supported by the STI 2030-Major Projects under Grant 2021ZD0201404. The authors thank the NeurIPS committee for granting us the NeurIPS 2024 Scholar Award, which has helped us participate in the conference.

References

[1] Chunming He, Kai Li, Guoxia Xu, Jiangpeng Yan, and Longxiang Tang. Hqg-net: Unpaired medical image enhancement with high-quality guidance. *IEEE Trans. Neural Netw. Learn. Syst.*, 2023. 1

[2] Christos Sakaridis, Dengxin Dai, Simon Hecker, and Luc Van Gool. Model adaptation with synthetic and real data for semantic dense foggy scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 687–704, 2018. 1

[3] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 33(12):2341–2353, 2010. 2

[4] Mingye Ju, Chunming He, Can Ding, Wenqi Ren, Lin Zhang, and Kai-Kuang Ma. All-inclusive image enhancement for degraded images exhibiting low-frequency corruption. *Trans. Circuits Syst. Video Technol.*, 2024. 2

[5] Yeying Jin, Wending Yan, Wenhan Yang, and Robby T. Tan. Structure representation network and uncertainty feedback learning for dense non-uniform fog removal. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pages 2041–2058, December 2022. 2

[6] Chunming He, Yuqi Shen, Chengyu Fang, Fengyang Xiao, Longxiang Tang, Yulun Zhang, Wangmeng Zuo, Zhenhua Guo, and Xiu Li. Diffusion models in low-level vision: A survey. *arXiv preprint arXiv:2406.11138*, 2024.

[7] Ruiqi Wu, Zhengpeng Duan, Chunle Guo, Zhi Chai, and Chongyi Li. Ridcp: Revitalizing real image dehazing via high-quality codebook priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2, 7, 8, 9, 18

[8] Shenghai Yuan, Jijia Chen, Jiaqi Li, Wenchao Jiang, and Song Guo. Lhnet: A low-cost hybrid network for single image dehazing. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7706–7717, 2023.

[9] Shenghai Yuan, Jijia Chen, Wenchao Jiang, Zhiming Zhao, and Song Guo. Lhnetv2: A balanced low-cost hybrid network for single image dehazing. *IEEE Transactions on Multimedia*, 2024. 2

[10] Chong Mou, Qian Wang, and Jian Zhang. Deep generalized unfolding networks for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17399–17410, 2022. 2, 3, 7, 8, 9

[11] Dong Yang and Jian Sun. Proximal dehaze-net: A prior learning-based deep network for single image dehazing. In *Proceedings of the european conference on computer vision (ECCV)*, pages 702–717, 2018. 2, 3, 7, 8, 9

[12] Chunming He, Kai Li, Guoxia Xu, Yulun Zhang, Runze Hu, Zhenhua Guo, and Xiu Li. Degradation-resistant unfolding network for heterogeneous image fusion. In *ICCV*, pages 12611–12621, 2023. 2

[13] Guoxia Xu, Chunming He, Hao Wang, Hu Zhu, and Weiping Ding. Dm-fusion: Deep model-driven network for heterogeneous image fusion. *IEEE Trans. Neural Netw. Learn. Syst.*, 2023. 2

[14] Hang Dong, Jinshan Pan, Lei Xiang, Zhe Hu, Xinyi Zhang, Fei Wang, and Ming-Hsuan Yang. Multi-scale boosted dehazing network with dense feature fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2157–2167, 2020. 2, 7, 8, 9

[15] Chun-Le Guo, Qixin Yan, Saeed Anwar, Runmin Cong, Wenqi Ren, and Chongyi Li. Image dehazing transformer with transmission-aware 3d position embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5812–5820, 2022. 7, 8, 9

[16] Xiaohong Liu, Yongrui Ma, Zhihao Shi, and Jun Chen. Griddehazenet: Attention-based multi-scale network for image dehazing. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 7314–7323, 2019. 18, 19

[17] Xu Qin, Zhilin Wang, Yuanchao Bai, Xiaodong Xie, and Huizhu Jia. Ffa-net: Feature fusion attention network for single image dehazing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07), 2020. 18, 19

[18] Tian Ye, Yunchen Zhang, Mingchao Jiang, Liang Chen, Yun Liu, Sixiang Chen, and Erkang Chen. Perceiving and modeling density for image dehazing. In *European conference on computer vision*, pages 130–145. Springer, 2022. 2

[19] Zeyuan Chen, Yangchao Wang, Yang Yang, and Dong Liu. Psd: Principled synthetic-to-real dehazing guided by physical priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7180–7189, 2021. 2, 3, 7, 8, 9

[20] Yuwei Qiu, Kaihao Zhang, Chenxi Wang, Wenhan Luo, Hongdong Li, and Zhi Jin. Mb-taylorformer: Multi-branch efficient transformer expanded by taylor formula for image dehazing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12802–12813, 2023.

[21] Haiyan Wu, Yanyun Qu, Shaohui Lin, Jian Zhou, Ruizhi Qiao, Zhizhong Zhang, Yuan Xie, and Lizhuang Ma. Contrastive learning for compact single image dehazing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10551–10560, 2021.

[22] Yu Zheng, Jiahui Zhan, Shengfeng He, Junyu Dong, and Yong Du. Curricular contrastive regularization for physics-aware single image dehazing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5785–5794, 2023. 18, 19

[23] Yafei Zhang, Shen Zhou, and Huafeng Li. Depth information assisted collaborative mutual promotion network for single image dehazing. *arXiv preprint arXiv:2403.01105*, 2024. 2

[24] Jing Wang, Songtao Wu, Zhiqiang Yuan, Qiang Tong, and Kuanhong Xu. Frequency compensated diffusion model for real-scene dehazing. *Neural Networks*, 175:106281, 2024. 2

[25] Xiang Chen, Zhentao Fan, Pengpeng Li, Longgang Dai, Caihua Kong, Zhuoran Zheng, Yufeng Huang, and Yufeng Li. Unpaired deep image dehazing using contrastive disentanglement learning. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 632–648, Cham, 2022. Springer Nature Switzerland. 2

[26] Yang Yang, Chaoyue Wang, Risheng Liu, Lin Zhang, Xiaojie Guo, and Dacheng Tao. Self-augmented unpaired image dehazing via density and depth decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2037–2046, 2022. 2, 7, 8, 9

[27] Yuanjie Shao, Lerenhan Li, Wenqi Ren, Changxin Gao, and Nong Sang. Domain adaptation for image dehazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2808–2817, 2020. 2, 7, 8, 9

[28] Yi Li, Yi Chang, Yan Gao, Changfeng Yu, and Luxin Yan. Physically disentangled intra- and inter-domain adaptation for varicolored haze removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5841–5850, 2022. 2

[29] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017. 2

[30] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C Courville, and Yoshua Bengio. Generative adversarial nets. In *Neural Information Processing Systems (NeruIPS)*, 2014. 3

[31] Xiaofeng Cong, Jie Gui, Jing Zhang, Junming Hou, and Hao Shen. A semi-supervised nighttime dehazing baseline with spatial-frequency aware and realistic brightness constraint, 2024. 3

[32] Ming Tong, Yongzhen Wang, Peng Cui, Xuefeng Yan, and Mingqiang Wei. Semi-uformer: Semi-supervised uncertainty-aware transformer for image dehazing, 2022. 3

[33] Ran He, Wei-Shi Zheng, Tieniu Tan, and Zhenan Sun. Half-quadratic-based iterative minimization for robust sparse representation. *IEEE transactions on pattern analysis and machine intelligence*, 36(2):261–275, 2013. 3

[34] Mingye Ju, Chunming He, and Juping Liu. Ivf-net: An infrared and visible data fusion deep network for traffic object enhancement in intelligent transportation systems. *IEEE Trans. Intell. Transp. Syst.*, 2022. 3

[35] Kai Zhang, Luc Van Gool, and Radu Timofte. Deep unfolding network for image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3217–3226, 2020. 3

[36] Di You, Jingfen Xie, and Jian Zhang. Ista-net++: Flexible deep unfolding network for compressive sensing. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021. 3

[37] Miaoyu Li, Ying Fu, Ji Liu, and Yulun Zhang. Pixel adaptive deep unfolding transformer for hyperspectral image reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12959–12968, 2023. 3

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 5

[39] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *International Conference on Learning Representations (ICLR)*, 2015. 6

[40] Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *IEEE Transactions on Image Processing*, 28(1):492–505, 2019. 7, 8, 9, 10

[41] Mu He, Le Hui, Yikai Bian, Jian Ren, Jin Xie, and Jian Yang. Ra-depth: Resolution adaptive self-supervised monocular depth estimation. In *European Conference on Computer Vision*, pages 565–581. Springer, 2022. 7

[42] Raanan Fattal. Dehazing using color-lines. *ACM transactions on graphics (TOG)*, 34(1):1–14, 2014. 7, 8, 9

[43] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 7

[44] P Kingma Diederik. Adam: A method for stochastic optimization. *(No Title)*, 2014. 8

[45] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Controlling vision-language models for universal image restoration. *arXiv preprint arXiv:2310.01018*, 2023. 8, 18

[46] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5148–5157, 2021. 8

[47] Lark Kwon Choi, Jaehee You, and Alan Conrad Bovik. Referenceless prediction of perceptual fog density and perceptual image defogging. *IEEE Transactions on Image Processing (TIP)*, 24(11):3888–3901, 2015. 8

[48] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012. 8

[49] Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE Transactions on Image Processing (TIP)*, 27(8):3998–4011, 2018. 8

[50] Chaofeng Chen and Jiadi Mo. IQA-PyTorch: Pytorch toolbox for image quality assessment. [Online]. Available: https://github.com/chaofengc/IQA-PyTorch, 2022. 8

[51] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013. 9

[52] Xiaobo Wang, Lizhen Deng, and Guoxia Xu. Image threshold segmentation based on glle histogram. In *CPSCom*, pages 410–415. IEEE, 2019. 10

[53] Wenhan Yang, Ye Yuan, Wenqi Ren, Jiaying Liu, Walter J Scheirer, Zhangyang Wang, Taiheng Zhang, Qiaoyong Zhong, Di Xie, Shiliang Pu, et al. Advancing image understanding in poor visibility environments: A collective benchmark study. *IEEE Transactions on Image Processing*, 29:5737–5752, 2020. 10

[54] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 10

[55] Yuelin Zhang, Pengyu Zheng, Wanquan Yan, Chengyu Fang, and Shing Shin Cheng. A unified framework for microscopy defocus deblur with multi-pyramid transformer and contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11125–11136, June 2024. 10

[56] Chunming He, Chengyu Fang, Yulun Zhang, Kai Li, Longxiang Tang, Chenyu You, Fengyang Xiao, Zhenhua Guo, and Xiu Li. Reti-diff: Illumination degradation image restoration with retinex-based latent diffusion model. *arXiv preprint arXiv:2311.11638*, 2023. 10

[57] Xunpeng Yi, Han Xu, Hao Zhang, Linfeng Tang, and Jiayi Ma. Diff-retinex: Rethinking low-light image enhancement with a generative diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12302–12311, 2023. 10

[58] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Siran Chen, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4117–4125, 2024. 10

[59] Jiangshan Wang, Yue Ma, Jiayi Guo, Yicheng Xiao, Gao Huang, and Xiu Li. Cove: Unleashing the diffusion feature correspondence for consistent video editing. *arXiv preprint arXiv:2406.08850*, 2024.

[60] Shenghai Yuan, Jinfa Huang, Yujun Shi, Yongqi Xu, Ruijie Zhu, Bin Lin, Xinhua Cheng, Li Yuan, and Jiebo Luo. Magictime: Time-lapse video generation models as metamorphic simulators. *arXiv preprint arXiv:2404.05014*, 2024.

[61] Shenghai Yuan, Jinfa Huang, Yongqi Xu, Yaoyang Liu, Shaofeng Zhang, Yujun Shi, Ruijie Zhu, Xinhua Cheng, Jiebo Luo, and Li Yuan. Chronomagic-bench: A benchmark for metamorphic evaluation of text-to-time-lapse video generation. *arXiv preprint arXiv:2406.18522*, 2024.

[62] Zunnan Xu, Yukang Lin, Haonan Han, Sicheng Yang, Ronghui Li, Yachao Zhang, and Xiu Li. Mambatalk: Efficient holistic gesture synthesis with selective state space models. *arXiv preprint arXiv:2403.09471*, 2024. 10

[63] Zanlin Ni, Yulin Wang, Renping Zhou, Jiayi Guo, Jinyi Hu, Zhiyuan Liu, Shiji Song, Yuan Yao, and Gao Huang. Revisiting non-autoregressive transformers for efficient image synthesis. In *CVPR*, 2024. 10

[64] Zanlin Ni, Yulin Wang, Renping Zhou, Rui Lu, Jiayi Guo, Jinyi Hu, Zhiyuan Liu, Yuan Yao, and Gao Huang. Adanat: Exploring adaptive policy for token-based image generation. In *ECCV*, 2024.

[65] Yulin Wang, Gao Huang, Shiji Song, Xuran Pan, Yitong Xia, and Cheng Wu. Regularizing deep networks with semantic data augmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[66] Chenyang Zhu, Kai Li, Yue Ma, Chunming He, and Li Xiu. Multibooth: Towards generating all your concepts in an image from text. *arXiv preprint arXiv:2404.14239*, 2024. 10

[67] Ziqing Wang, Yuetong Fang, Jiahang Cao, and Renjing Xu. Bursting spikes: Efficient and high-performance snns for event-based vision. *arXiv preprint arXiv:2311.14265*, 2023. 10

[68] Yifan Pu, Yizeng Han, Yulin Wang, Junlan Feng, Chao Deng, and Gao Huang. Fine-grained recognition with learnable semantic data augmentation. *IEEE Transactions on Image Processing*, 2024. 19

[69] Yifan Pu, Weicong Liang, Yiduo Hao, Yuhui Yuan, Yukang Yang, Chao Zhang, Han Hu, and Gao Huang. Rank-detr for high quality object detection. *Advances in Neural Information Processing Systems*, 36, 2024.

[70] Jiangshan Wang, Yifan Pu, Yizeng Han, Jiayi Guo, Yiru Wang, Xiu Li, and Gao Huang. Gra: Detecting oriented objects through group-wise rotating and attention. *arXiv preprint arXiv:2403.11127*, 2024.

[71] Yicheng Xiao, Zhuoyan Luo, Yong Liu, Yue Ma, Hengwei Bian, Yatai Ji, Yujiu Yang, and Xiu Li. Bridging the gap: A unified video comprehension framework for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18709–18719, 2024.

[72] Zhengxi Zhang, Liang Zhao, Yunan Liu, Shanshan Zhang, and Jian Yang. Unified density-aware image dehazing and object detection in real-world hazy scenes. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 19

[73] Chunming He, Kai Li, Yachao Zhang, Longxiang Tang, Yulun Zhang, Zhenhua Guo, and Xiu Li. Camouflaged object detection with feature decomposition and edge reconstruction. In *CVPR*, pages 22046–22055, 2023. 19

[74] Chunming He, Kai Li, Yachao Zhang, Guoxia Xu, and Longxiang Tang. Weakly-supervised concealed object segmentation with sam-based pseudo labeling and multi-scale feature grouping. *NeurIPS*, 2024.

[75] Chunming He, Kai Li, Yachao Zhang, Yulun Zhang, Zhenhua Guo, and Xiu Li. Strategic preys make acute predators: Enhancing camouflaged object detectors by generating camouflaged objects. In *ICLR*, 2024.

[76] Jian Hu, Jiayi Lin, Shaogang Gong, and Weitong Cai. Relax image-specific prompt requirement in sam: A single generic prompt for segmenting camouflaged objects. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 12511–12518, 2024.

[77] Jian Hu, Jiayi Lin, Junchi Yan, and Shaogang Gong. Leveraging hallucinations to reduce manual prompt dependency in promptable segmentation. *arXiv preprint arXiv:2408.15205*, 2024.

[78] Fengyang Xiao, Pan Zhang, Chunming He, Runze Hu, and Yutao Liu. Concealed object segmentation with hierarchical coherence modeling. In *CAAI*, pages 16–27. Springer, 2023.

[79] Fengyang Xiao, Sujie Hu, Yuqi Shen, Chengyu Fang, Jinfa Huang, Chunming He, Longxiang Tang, Ziyun Yang, and Xiu Li. A survey of camouflaged object detection and beyond. *arXiv preprint arXiv:2408.14562*, 2024.

[80] Zunnan Xu, Zhihong Chen, Yong Zhang, Yibing Song, Xiang Wan, and Guanbin Li. Bridging vision and language encoders: Parameter-efficient tuning for referring image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17503–17512, 2023.

[81] Zunnan Xu, Jiaqi Huang, Ting Liu, Yong Liu, Haonan Han, Kehong Yuan, and Xiu Li. Enhancing fine-grained multi-modal alignment via adapters: A parameter-efficient training framework for referring image segmentation. In *2nd Workshop on Advancing Neural Network Training: Computational Efficiency, Scalability, and Resource Optimization (WANT@ ICML 2024)*, 2024. 19

[82] Yuelin Zhang, Kim Yan, Chun Ping Lam, Chengyu Fang, Wenxuan Xie, Yufu Qiu, Raymond Shing-Yan Tang, and Shing Shin Cheng. Motion-guided dual-camera tracker for endoscope tracking and motion analysis in a mechanical gastric simulator, 2024. 19

[83] Longxiang Tang, Zhuotao Tian, Kai Li, Chunming He, Hantao Zhou, Hengshuang Zhao, Xiu Li, and Jiaya Jia. Mind the interference: Retaining pre-trained knowledge in parameter efficient continual learning of vision-language models. *arXiv preprint arXiv:2407.05342*, 2024. 19

[84] Sixiang Chen, Tian Ye, Jun Shi, Yun Liu, JingXia Jiang, Erkang Chen, and Peng Chen. Dehrformer: Real-time transformer for depth estimation and haze removal from varicolored haze scenes. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 19

[85] Lei Xu, Hui Wu, Chunming He, Jun Wang, Changqing Zhang, Feiping Nie, and Lei Chen. Multi-modal sequence learning for alzheimer's disease progression prediction with incomplete variable-length longitudinal data. *MIA*, 82:102643, 2022. 19

[86] Yicheng Xiao, Lin Song, Shaoli Huang, Jiangshan Wang, Siyu Song, Yixiao Ge, Xiu Li, and Ying Shan. Grootvl: Tree topology is all you need in state space model. *arXiv preprint arXiv:2406.02395*, 2024. 19

[87] Yang Yue, Rui Lu, Bingyi Kang, Shiji Song, and Gao Huang. Understanding, predicting and better resolving q-value divergence in offline-rl. *Advances in Neural Information Processing Systems*, 36, 2024.

[88] Yang Yue, Bingyi Kang, Zhongwen Xu, Gao Huang, and Shuicheng Yan. Value-consistent representation learning for data-efficient reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 11069–11077, 2023.

[89] Le Yang, Haojun Jiang, Ruojin Cai, Yulin Wang, Shiji Song, Gao Huang, and Qi Tian. Condensenet v2: Sparse feature reactivation for deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3569–3578, 2021.

[90] Le Yang, Yizeng Han, Xi Chen, Shiji Song, Jifeng Dai, and Gao Huang. Resolution adaptive networks for efficient inference. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2369–2378, 2020.

[91] Yizeng Han, Zeyu Liu, Zhihang Yuan, Yifan Pu, Chaofei Wang, Shiji Song, and Gao Huang. Latency-aware unified dynamic networks for efficient image recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–17, 2024.

[92] Yizeng Han, Yifan Pu, Zihang Lai, Chaofei Wang, Shiji Song, Junfeng Cao, Wenhui Huang, Chao Deng, and Gao Huang. Learning to weight samples for dynamic early-exiting networks. In *European conference on computer vision*, pages 362–378. Springer, 2022.

[93] Dongchen Han, Tianzhu Ye, Yizeng Han, Zhuofan Xia, Shiji Song, and Gao Huang. Agent attention: On the integration of softmax and linear attention. In *ECCV*, 2024.

[94] Dongchen Han, Yifan Pu, Zhuofan Xia, Yizeng Han, Xuran Pan, Xiu Li, Jiwen Lu, Shiji Song, and Gao Huang. Bridging the divide: Reconsidering softmax and linear attention. In *NeurIPS*, 2024. 19

Figure 8: Detailed Comparison of fig. 4. **Red** region reflects all past methods have had haze residues, but our method have the least in the same case. **Purple** region shows our method restores richer detail and truer colors. Please zoom in for a better view.

A Appendix

A.1 Declaration of eq. (7) and eq. (10)

Having gotten eq. (5) and eq. (6), the solution of $\hat{\mathbf{T}}$ can be formulated according to the proximal gradient algorithm:

$$\mathcal{T}(\hat{\mathbf{J}}_{k-1}, \mathbf{T}_{k-1}) = \frac{1}{2}\|\mathbf{P} - \hat{\mathbf{J}}_{k-1} \cdot \hat{\mathbf{T}} + \hat{\mathbf{T}} - \mathbf{I}\|_2^2 + \frac{\lambda_k}{2}\|\hat{\mathbf{T}} - \mathbf{T}_{k-1}\|_2^2. \tag{22}$$

Then we obtain the partial derivative:

$$\partial_{\hat{\mathbf{T}}}\mathcal{T}(\hat{\mathbf{J}}_{k-1}, \mathbf{T}_{k-1}) = (\mathbf{I} - \hat{\mathbf{J}}_{k-1})^T(\mathbf{P} - \hat{\mathbf{J}}_{k-1} \cdot \hat{\mathbf{T}} + \hat{\mathbf{T}} - \mathbf{I}) + \lambda_k(\hat{\mathbf{T}} - \mathbf{T}_{k-1}). \tag{23}$$

Let the partial derivative be equal to zero, we archieve the closed-form solution for $\hat{\mathbf{T}}$ in eq. (7).

Similarly, the solution of $\hat{\mathbf{J}}$ can be formulated as

$$\mathcal{J}(\hat{\mathbf{T}}_k, \mathbf{J}_{k-1}) = \frac{1}{2}\|\mathbf{P} - \hat{\mathbf{J}} \cdot \hat{\mathbf{T}}_k + \hat{\mathbf{T}}_k - \mathbf{I}\|_2^2 + \frac{\mu_k}{2}\|\hat{\mathbf{J}} - \mathbf{J}_{k-1}\|_2^2. \tag{24}$$

The corresponding partial derivative is

$$\partial_{\hat{\mathbf{J}}}\mathcal{J}(\hat{\mathbf{T}}_k, \mathbf{J}_{k-1}) = -\hat{\mathbf{T}}_k^T(\mathbf{P} - \hat{\mathbf{J}} \cdot \hat{\mathbf{T}}_k + \hat{\mathbf{T}}_k - \mathbf{I}) + \mu_k(\hat{\mathbf{J}} - \mathbf{J}_{k-1}). \tag{25}$$

The closed-form solution for $\hat{\mathbf{J}}$ is presented in eq. (10) when let the partial derivative be equal to zero.

A.2 Declaration of CLIP module.

The haze density evaluator $\mathcal{D}(\cdot)$ can be formulated as:

$$\mathcal{D}(\cdot) = \frac{Enc_{image}(\cdot)}{\|Enc_{image}(\cdot)\|} \cdot \left(\frac{Enc_{text}(\text{Text})}{\|Enc_{text}(\text{Text})\|}\right)^\top \tag{26}$$

The text we used is "hazy" from the DA-CLIP provided in the text list.

Table 7: Ablation of CPMM module of our CORUN.

Modules	NIMA ↑	BRISQUE ↓	FADE ↓
Hazy(Input)	4.483	36.642	2.484
w/o CPMM	4.836	38.197	1.362
w/ CPMM (CORUN+)	5.342	11.956	0.824

Table 8: Ablation of our trusted weights present as a map or value.

Methods	NIMA ↑	BRISQUE ↓	FADE ↓
Only Full	5.229	13.099	0.803
Partition+Full(CORUN+)	5.342	11.956	0.824

Table 9: Effects of more Colabator components.

Modules	NIMA ↑	BRISQUE ↓	FADE ↑
w/o Colabator	4.856	16.541	1.091
w/o Strong aug.	5.084	12.671	0.813
w/o DA-CLIP[45]	5.358	11.200	0.856
CORUN+	5.342	11.956	0.824

Table 10: Ablation of mainstream datasets setting.

RIDCP[7] Pipeline		-	Metrics		
Data.+Gen.	Aug.	OTS	NIMA ↑	BRISQUE ↓	FADE↓
✓			4.845	20.779	0.765
		✓	4.991	16.478	0.840
✓	✓		5.342	11.956	0.824

A.3 Ablation Study

Effect of the CPMM Module in CORUN. We evaluate the impact of the Cooperative Proximal Mapping Modules (CPMM) in our CORUN architecture. As shown in table 7, incorporating CPMM leads to a significant improvement in performance across all metrics. The model with CPMM (CORUN+) outperforms the variant without CPMM, highlighting the importance of CPMM in enhancing dehazing efficiency and image quality.

Impact of Trusted Weight Representations. We investigate the effect of using trusted weights as either a map or a value. As seen in table 8, the combination of partitioned and full trusted weights (CORUN+) achieves better results compared to using only full trusted weights. This emphasizes the value of our trusted weight representation in improving the accuracy of dehazing.

Effects of Additional Colabator Components. To analyze the contribution of individual components within the Colabator framework, we performed ablation studies, with the results presented in table 9. Removing essential elements, such as strong augmentation or DA-CLIP, results in performance deterioration, confirming the importance of each Colabator component in ensuring optimal dehazing outcomes.

Ablation on Dataset Configurations. We evaluate our methods with different dataset settings. As shown in table 10, the results verify our method still achieves a leading place under the three settings compared with existing methods.

Effectiveness of the Simplified ASM Formula. We assess the impact of simplifying the Atmospheric Scattering Model (ASM) formula. The results in table 11 indicate that using the simplified ASM formula will lead to a slight decrease in the dehazing ability, but it can evidently improve the image quality of the results.

Influence of Loss Functions. We compare the effect of using different loss functions (eq. (15) and eq. (16)). Table 12 shows that combining both loss functions yields better performance than using either one alone, demonstrating the advantage of this combined loss strategy in refining the dehazing process.

Table 11: Ablation of our simplified ASM formula.

ASM formula	NIMA ↑	BRISQUE ↓	FADE↓
w/o simplify	5.203	14.469	0.817
w/ simplify(CORUN+)	5.342	11.956	0.824

Table 12: Ablations of eq.15 and eq.16 loss functions. Our strategy achieves a better result.

Loss	NIMA ↑	BRISQUE ↓	FADE ↓
Eq.15 Only	5.249	13.997	1.035
Eq.16 Only	5.220	12.484	0.795
Both Loss (Ours)	5.342	11.956	0.824

Table 13: Effects of integrating our Colabator with more cutting-edge dehazing methods. The gains brought by Colabator are significant.

Methods	NIMA ↑	BRISQUE ↓	FADE↓
C2PNet[22]	4.715	34.314	2.064
C2PNet+Colabator	4.823	23.662	1.329
FFA-Net[17]	4.822	33.235	2.080
FFA-Net+Colabator	4.839	29.219	0.958
GDN[16]	5.074	33.051	2.611
GDN+Colabator	5.258	23.691	0.947

Integration with Other Dehazing Methods. To test the generalizability of Colabator, we integrated it with various state-of-the-art dehazing models, such as C2PNet [22], FFA-Net [17], and GDN [16]. As shown in table 13, incorporating Colabator leads to performance gains across all metrics, demonstrating its effectiveness as a plug-and-play module for improving dehazing in various architectures.

A.4 Broader Impacts

Real-world image dehazing is a crucial task in image restoration, aimed at removing haze degradation from images captured in real-world scenarios. In computer vision, dehazing can benefit downstream tasks such as object detection [68–72], image segmentation [73–81], tracking [82, 83], depth estimation [84, 85], and more vision related tasks [86–94], with applications ranging from autonomous driving to security monitoring. Our paper introduces a cooperative unfolding network and a plug-and-play pseudo-labeling framework, achieving state-of-the-art performance in real-world dehazing tasks. Notably, image dehazing techniques have yet to exhibit negative social impacts. Our proposed CORUN and Colabator methods also do not present any foreseeable negative societal consequences.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Relevant information is included in section 1.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Relevant information is included in section 5

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: For each theoretical result we declared in section 1 and introduced in section 3 has full set of assumptions and complete and correct proof in section 3 and section 4.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Relevant information is included in section 3 and section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide a clear algorithm description section 3, which is conducive to reproducing, and the code and data will be open.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Relevant information is included in section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The results of both our ablation experiments and the comparison experiments section 4.3 effectively demonstrate the validity of our method and claims.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the GPU model, video memory and quantity used by the computers we use for training and testing. section 4

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research conducted in the thesis complies with the NeurIPS Code of Ethics in every respect.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: There is no societal impact of the work performed as we explained in appendix A.4

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite all the work covered in this article and follow their license.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Our code, data and model will be open.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: Our experiment included a User study, and we reported our requirements and work content of the volunteers we invited. table 5

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our user study does not involve any potential risks.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.