Towards Real-World Adverse Weather Image Restoration: Enhancing Clearness and Semantics with Vision-Language Models

Jiaqi Xu[1], Mengyang Wu[1], Xiaowei Hu[2,⋆],
Chi-Wing Fu[1], Qi Dou[1], and Pheng-Ann Heng[1]

[1] The Chinese University of Hong Kong
[2] Shanghai Artificial Intelligence Laboratory

Abstract. This paper addresses the limitations of adverse weather image restoration approaches trained on synthetic data when applied to real-world scenarios. We formulate a semi-supervised learning framework employing vision-language models to enhance restoration performance across diverse adverse weather conditions in real-world settings. Our approach involves assessing image clearness and providing semantics using vision-language models on real data, serving as supervision signals for training restoration models. For clearness enhancement, we use real-world data, utilizing a dual-step strategy with pseudo-labels assessed by vision-language models and weather prompt learning. For semantic enhancement, we integrate real-world data by adjusting weather conditions in vision-language model descriptions while preserving semantic meaning. Additionally, we introduce an effective training strategy to bootstrap restoration performance. Our approach achieves superior results in real-world adverse weather image restoration, demonstrated through qualitative and quantitative comparisons with state-of-the-art works.

Keywords: Adverse weather · Deraining · Dehazing · Desnowing

1 Introduction

Images captured under challenging weather conditions, such as rain, haze, and snow, are plagued by a variety of artifacts that significantly affect the image quality. These imperfections severely impair the efficacy of outdoor vision systems. Previous research efforts [2, 7, 12, 24, 31] have primarily focused on developing specialized techniques for mitigating the effects of individual weather phenomena, tailoring their models to the unique characteristics of rain, haze, or snow. More recently, all-in-one adverse weather removal works [4, 17, 27, 40] design single model-based methods to restore images captured under multiple adverse weather conditions. Despite the encouraging outcomes demonstrated on synthetic datasets by these approaches, their applicability to real-world scenarios remains notably constrained.

⋆ Corresponding author (huxiaowei@pjlab.org.cn)

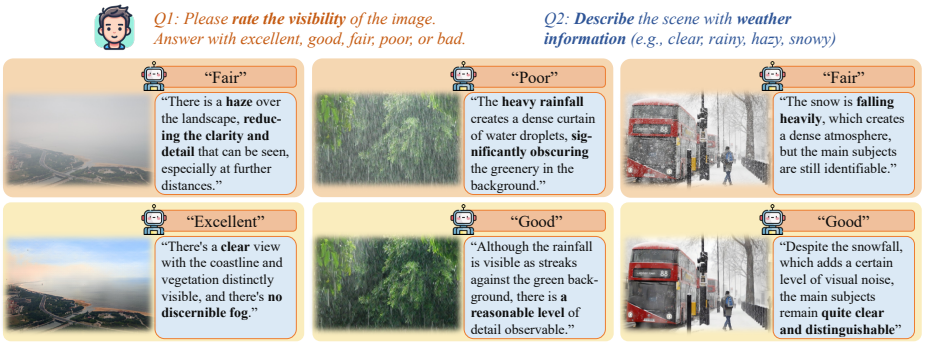

Fig. 1: The **clearness** level and the **semantics** information of real-world adverse weather images are provided by large vision-language models. This assistance is instrumental in training image restoration models to effectively utilize real-world data.

The limited generalization capability in real-world adverse weather images can be attributed to two main factors. Firstly, adverse weather removal methods are predominantly trained on synthetic datasets [3, 7, 15, 16, 24, 30], resulting in a domain gap when applied to real-world situations. Secondly, these methods primarily focus on restoring the visual clarity of images, often neglecting the semantic context of the scenes they depict. Consequently, current weather removal approaches struggle with real-world data and offer marginal enhancements to downstream high-level vision tasks under adverse weather conditions.

In response to these challenges, this work introduces a novel semi-supervised learning framework, WResVLM, that explores vision-language models (VLMs) [22, 32] to enhance image restoration in real-world scenarios across diverse adverse weather conditions. The real-world images with the weather-related artifacts are used as the unlabeled (unpaired) data to train image restoration models and the supervision signals are provided by the large vision-language models. As depicted in Fig. 1, large VLMs play a crucial role in assessing the *clearness* levels and providing *semantics* information of images under adverse weather conditions. This capability proves instrumental in training image restoration models effectively, enabling them to handle the complexities of real-world data.

To enhance the *clearness* of the restored images produced by the restoration model, we utilize real-world data for model training, evaluating image clarity with the assistance of large vision-language models. These models, exposed to a diverse array of weather conditions during training, demonstrate proficiency in recognizing and distinguishing various weather-related scenes. The approach involves two key steps: initially, the vision-language model is employed to assess images and select pseudo-labels for training the restoration model. Subsequently, weather prompt learning is introduced to tailor the VLM, ultimately utilizing it to modulate the image restoration process. This dual-step strategy enhances the restoration model's ability to address the real-world weather complexities and improve the overall clearness of the restored images.

To enhance the *semantics* of the restored images, we further integrate real-world data into the model training. This involves utilizing descriptions generated by vision-language models associated with each image, providing rich semantic information about the scene and adverse weather conditions. A unique aspect of our method involves adjusting the weather clues in the descriptions while maintaining the semantic meaning unchanged. This enables the training of the image restoration model to specifically target the removal of weather-related artifacts without altering the image's underlying semantics. In contrast to methods that might overlook semantic cues, our framework incorporates vision-language models to encompass both visual clarity and semantic context, thus presenting a more comprehensive strategy for adverse weather image restoration.

Lastly, we develop a training strategy aimed at achieving effective pseudo-label initialization and iterative updates, with the primary goal of improving restoration outcomes. We conduct experiments using real-world images captured under diverse adverse weather conditions. The results demonstrate that our method significantly surpasses both state-of-the-art adverse weather image restoration approaches and general image restoration methods. Code and data are available at GitHub.

2 Related Work

2.1 Image Restoration in Adverse Weather Conditions

Previous works focus on restoring images captured under specific weather conditions, including deraining [7, 9, 10, 48, 51, 58], dehazing [2, 6, 8, 34], and desnowing [24]. Recent works [4, 17, 27, 28, 40, 44, 53, 59] focus on all-in-one adverse weather removal, which restores images captured under various weather conditions using a single model. The pioneering All-in-One [17] achieves this by using joint training and a unified set of model weights. TransWeather [40] introduces a transformer-based architecture while Chen *et al*. [4] leverage knowledge distillation and contrastive learning. WeatherDiff [27] adapts the diffusion model for adverse weather artifact removal. Zhu *et al*. [59] learn weather-general and weather-specific features through multiple sets of model weights. AWRCP [53] enhances image restoration by exploring high-quality codebook priors. Domain adaptation technique is also utilized to handle mixed weather conditions [28]. More recent works explore prompting [29], textual information [25], and customizing pre-trained diffusion models [13]. PromptIR [29] enhances the all-in-one restoration by predicting degradation-conditioned prompts. DA-CLIP [25] learns the degradation information through image-text contrastive learning.

The prior approaches typically rely on paired synthetic data [3, 7, 15, 16, 24, 30, 49] for training and evaluation, demonstrating promising results in synthetic benchmarks. However, the trained models exhibit limited generalization capabilities toward complex real-world scenarios due to the domain gap. Additionally, WeatherStream [55] attempts to compile a dataset of real degenerated images with corresponding ground truth, yet suffering from low image quality issues, *e.g.*, compression artifacts.

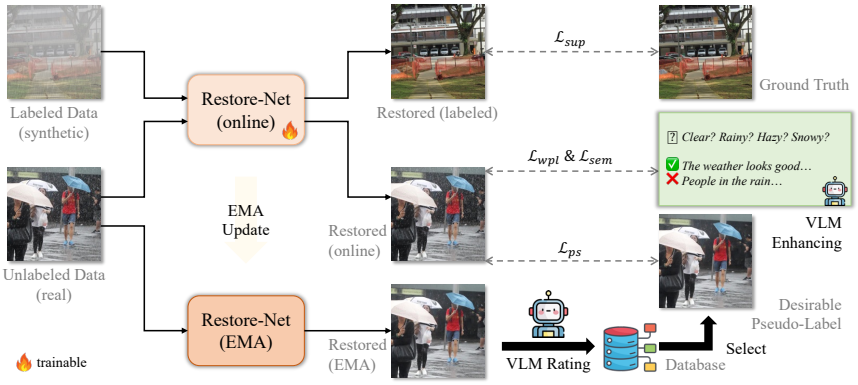

Fig. 2: The schematic illustration of the proposed semi-supervised learning framework enhancing real-world image restoration by improving clearness and semantics in varied adverse weather conditions through the utilization of vision-language models.

2.2 Vision-Language Models

Vision-language models merge computer vision and natural language processing. CLIP [32] pioneered text and image alignment through large-scale pre-training. CLIP's versatility is also demonstrated in image manipulation fields, including backlit image enhancement [18] and novel concept generation [33]. Recent advances, including GPT-4 [1] and Llama [39], demonstrate impressive conversational abilities. Large VLMs like LLaVA [22] excel in high-level multimodal visual question answering. Recent works reveal that vision-language models are also applicable to low-level applications. CLIP-IQA [41] and LIQE [56] expand upon the CLIP-like architecture for image quality assessment, showcasing VLMs' adaptability to technical evaluations of image quality. Q-Bench [46] highlights VLMs' inherent low-level perceptual capabilities. These works, however, focus primarily on general technical image quality assessment and show limited abilities to help image restoration under adverse weather conditions.

3 Methodology

In this work, we introduce a novel semi-supervised learning framework for all-in-one adverse weather image restoration, leveraging both labeled synthetic images and unlabeled real images. Our motivation emphasizes the necessity to improve image restoration in the real world. Current approaches, mostly trained on synthetic images, struggle with generalization when handling real-world adverse weather images. They frequently overlook the image context related to weather-related artifacts in real data, resulting in their limited effectiveness.

Figure 2 shows the overall pipeline of our proposed semi-supervised learning framework for all-in-one adverse weather image restoration in real-world situations. This framework adopts several VLMs to improve the images' Clearness and Semantics during the removal of weather-related artifacts.

3.1 Enhancing Image Clearness through Vision-Language Models

Restoring images in adverse weather conditions involves eliminating weather-related artifacts to generate "clean" images. Attaining *clearness* is a primary goal in adverse weather image restoration, especially in the real world. In the absence of ground truth (clean) images for real-world data, the main challenge lies in determining the quality of restored images. Moreover, limited learning objectives are designed to enhance image clearness under weather-related conditions.

Large vision-language models, trained on diverse data and vast weather imagery, exhibit strong representation abilities for image quality assessment. Additionally, with the help of prompt learning, the VLMs can better distinguish well-restored images from those degraded by rain, haze, or snow. To achieve this, we suggest two steps. First, we employ the large vision-language models to assess the images and provide pseudo-labels for training the restoration model. Then, we introduce weather prompt learning to empower the VLM's ability to identify clearness, ultimately utilizing it for modulating image restoration.

Image Assessment and Pseudo-Labeling. Our goal is to improve the restoration of real adverse weather images using unlabeled data. This involves training restoration models with pseudo-labels generated from the unlabeled images. To ensure high-quality pseudo-labels for the subsequent model training, we establish a pseudo-label database, utilizing the zero-shot capability of large vision-language models to assess adverse weather image restoration.

Image assessment. Given the real adverse weather images and the corresponding predictions from deweathering methods, a critical issue is to measure the image quality of the restored images. Existing methods [14, 36] for low-level image quality assessment focus mainly on technical distortions, including noise, blur, and compression artifacts. There, however, exists a situation where an image that suffers from adverse weather is of "good" image quality, with little common noise, yet the visibility is largely degraded due to rain, haze, and snow. Hence, it is imperative to find an effective way to automatically evaluate the image quality in the context of adverse weather artifact removal.

Inspired by recent works [46, 47] that vision-language models perform the zero-shot image quality assessment with appropriate prompting, we present to uncover the potential of VLMs for assessing the adverse weather image restoration. Technically, we prompt the VLMs with weather-related image quality questions and convert the VLMs' responses into numerical scores. In detail, we first design the conversion templates for enquiring about the VLM responses to assess the image as illustrated in Fig. 3 (a).

Then, we adopt the commonly used five-scale ratings in the mean opinion score (MOS) studies, *i.e.*, excellent, good, fair, poor, and bad, which correspond to the scores between one and five. After that, we calculate the VLM-based rating r^{vlm} by converting the VLMs' predicted probabilities over these five-word

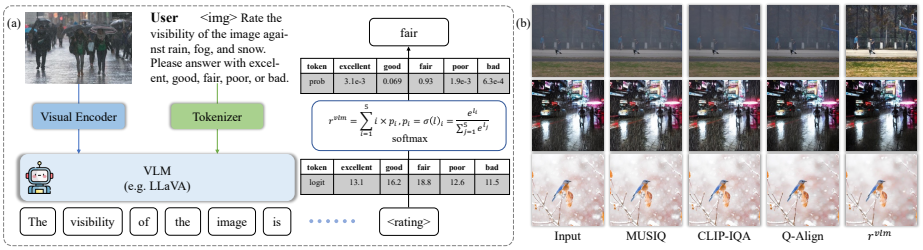

Fig. 3: VLM-based assessment of images restored from weather-related artifacts. In (a), we show the process to compute the VLM's image assessment ratings r^{vlm}. In (b), we find that r^{vlm} can select pseudo-labels with fewer weather-related artifacts.

tokens into numerical scores using *softmax*:

$$r^{vlm} = \sum_{i=1}^{5} i \times p_i, \quad p_i = \sigma(l)_i = \frac{e^{l_i}}{\sum_{j=1}^{5} e^{l_j}} \ , \tag{1}$$

where p_i denotes the probability for rating $i \in \{1, 2, 3, 4, 5\}$, l_i denotes the logit extracted from the language model for rating token i, and σ is the *softmax* operation. Thus, we obtain the visibility assessment for each restored image.

Pseudo-labeling. For the unlabeled real adverse weather image set $\mathcal{D}^u$, we assign and update the pseudo-labels $\mathcal{D}^{ps} = \{(x_i^u, y_i^{ps}) | x_i^u \in \mathcal{D}^u\}_{i=1}^{M}$ with desirable artifact-free pseudo-label images y_i^{ps} based on the VLM-based image assessment. Through investigation, we observe that r^{vlm} is able to acquire better pseudo-labels with fewer weather-related artifacts, as shown in Fig. 3 (b).

Initially, a pseudo-label database is constructed to store the current optimal pseudo-labels for the unlabeled images. Subsequently, throughout the model training process, we evaluate the VLM-based image visibility rating score for both the model's prediction and the recorded pseudo-labels. If the model achieves a superior restoration, we update the pseudo-label database accordingly [11]. In practice, we use predictions from the teacher model [37] for comparison, which is an exponential moving average of the student model. Lastly, we use the updated pseudo-labels to compute the pseudo-label loss for the online model:

$$\mathcal{L}_{ps} = \mathcal{L}_{app}(\hat{y}_i, y_i^{ps}) \ , \tag{2}$$

where $\hat{y}_i$ and y_i^{ps} are the prediction and the corresponding pseudo-label, respectively, and $\mathcal{L}_{app}$ is any kind of appearance loss, *e.g.*, $\mathcal{L}_1$ as adopted.

Weather Prompt Learning. We delve into the extensive knowledge embedded in the pre-trained vision-language model, capable of understanding the concept of images in both normal and adverse weather conditions. Specifically, we anticipate the CLIP [32] model to be indicator aware of image weather conditions, such as clear, rainy, hazy, or snowy. Subsequently, we leverage the learned

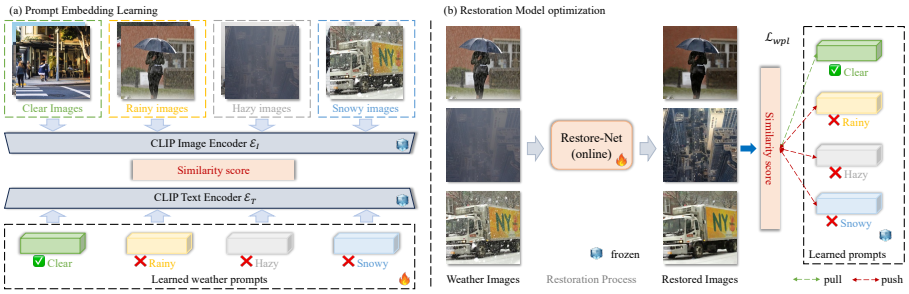

Fig. 4: The workflow of the weather prompt learning approach.

concept of "clearness" to modulate the model's learning toward achieving clear restoration. To enhance CLIP's ability to accurately differentiate weather in diverse scenarios, we employ the prompt learning approach to acquire prompt embeddings tailored to the image characteristics of each weather situation. The Weather Prompt Learning process consists of two stages: the prompt embedding learning stage and the restoration model optimization stage; see Fig. 4.

Prompt embedding learning. CLIP aligns images and text within a shared feature space. Rather than relying on fragile prompt engineering involving hand-crafted text prompts such as "rainy" or "a rainy photo", we adopt the prompt learning approach [18, 57]. Specifically, keeping the pre-trained CLIP model parameters fixed, we employ a set of four weather prompts t_c, t_r, t_h, t_s representing clear, rain, haze, and snow conditions as learnable vectors.

The weather prompts $\mathcal{E}_T(t)$ are initialized in the embedding space of the CLIP's text encoder. Meanwhile, the real images x in such clear, rainy, hazy, and snowy situations are collected, which are used to extract the reference image embeddings $\mathcal{E}_I(x)$ through the CLIP's image encoder. During the prompt embedding learning stage, the training objective is to minimize the classification loss, *i.e.*, the cross-entropy loss, by categorizing the weather prompts into their respective weather categories c: $p(c = i|x) = \sigma(z_i), z_i = cos(\mathcal{E}_I(x), \mathcal{E}_T(t_i))$, where σ denotes *softmax*, and $cos(\cdot, \cdot)$ denotes cosine similarity. Note that the learnable weather prompt embeddings are the only parameters to be optimized during this stage; see Fig. 4 (a).

Restoration model optimization. With the acquired knowledge from the learned weather prompts, we direct the training of the restoration model to generate images with enhanced clearness. Formally, during the restoration model optimization, the weather prompt learning loss $\mathcal{L}_{wpl}$ maximizes the similarity between the image embedding $\mathcal{E}_I(\hat{y})$ of the model's restored image $\hat{y}$ and the text embedding of the clear weather prompt $\mathcal{E}_T(t_c)$:

$$\mathcal{L}_{wpl} = \frac{e^{cos(\mathcal{E}_I(\hat{y}), \mathcal{E}_T(t_c))}}{\sum_{t \in \{t_c, t_r, t_h, t_s\}} e^{cos(\mathcal{E}_I(\hat{y}), \mathcal{E}_T(t))}} \,. \tag{3}$$

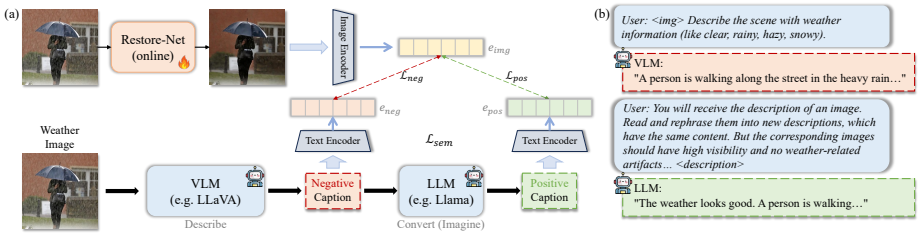

Fig. 5: Description-assisted semantic enhancement with the vision-language models.

In initial investigations, employing only $\mathcal{L}_{wpl}$ optimizes the model's prediction to reduce weather-related artifacts, yet the resulting image exhibits noticeable noise. We hypothesize that there is room for plausible solutions within the space of minimizing the weather prompt learning loss. To address this issue and regularize the model learning, a feature similarity loss is employed to align the model's prediction with both the pseudo-label y^{ps} and the input x^u:

$$\mathcal{L}_{feat} = \frac{1}{HW} \sum_{i=1}^{HW} (1 - cos(\hat{g}_i, g_i^*)) \ , \tag{4}$$

where $\hat{g}, g^*$ are image features of $\hat{y}$ as well as y^{ps}, x^u extracted from a pre-trained model, and H, W denote the spatial dimension of the feature space. In practice, we adopt the visual encoder of Depth Anything [50] for feature extraction because of its robustness against various scenarios.

3.2 Enhancing Image Semantics through Vision-Language Models

Restoring images in adverse weather conditions entails not only improving image clarity but also restoring the *semantics* distorted by weather-related artifacts. This contributes to the effectiveness of downstream vision tasks. The potential to recover image semantics is frequently disregarded by existing works trained on synthetic data. In this work, we introduce a method that leverages the image-text understanding capability of large vision-language models to enhance image semantics in the context of adverse weather image restoration.

Description-assisted semantic enhancement. Restoring degraded images is inherently challenging; describing their weather-affected appearance with natural language is straightforward. Our approach uses vision-language models to generate semantic descriptions of adverse weather images, capturing both scene context and weather conditions, including degradation levels. The comprehensive workflow is illustrated in Fig. 5. Given the input image, we employ a VLM, *e.g.*, LLaVA [22], to generate the (negative) caption with the weather description. For instance, *"A person is walking along the street in the heavy rain ..."* describes a scene with the object *person* in the weather of *rain*. The description also provides additional environment context, like *street*.

Next, we transform negative scene description d_{neg} associated with the degraded image in adverse weather conditions into a pseudo-clear representation. This transformation is achieved by prompting large language models, *e.g.*, Llama [39], to generate positive description d_{pos} corresponding to the restored image. Given the above negative description of the adverse weather, we can imagine its positive, clearly restored image, *e.g.*, *"The weather looks good. A person is walking ..."* Intuitively, d_{pos} and d_{neg} should have similar descriptions of the image content, like object and environment, but dissimilar descriptions of the weather and visibility, *i.e.*, good versus bad weather. Unlike the weather prompt in 3.1, d_{pos}, d_{neg} are tailored to a specific image; see Fig. 5.

Loss function. The model training incorporates semantic-aware regularization, promoting predictions that align with positive descriptions indicative of good weather conditions. Given the positive and negative descriptions, we formulate a description-assisted semantics regularization loss $\mathcal{L}_{sem}$:

$$\mathcal{L}_{sem} = \frac{e^{cos(\mathcal{E}_I(\hat{y}),\mathcal{E}_T(d_{pos}))}}{\sum_{d\in\{d_{pos},d_{neg}\}} e^{cos(\mathcal{E}_I(\hat{y}),\mathcal{E}_T(d))}} \ . \tag{5}$$

In initial trials, we observed that LLMs occasionally struggle with generating weather-varying descriptions that are content-invariant. To address this issue, we manually label certain negative-to-positive description conversions and introduce these examples in the in-context learning approach [26].

Finally, the overall loss is a weighted combination of the supervised appearance loss $\mathcal{L}_{sup}$, semi-supervised pseudo-label loss $\mathcal{L}_{ps}$, weather prompt loss $\mathcal{L}_{wpl}$, description-assisted semantic loss $\mathcal{L}_{sem}$, and feature similarity loss $\mathcal{L}_{feat}$:

$$\mathcal{L} = \mathcal{L}_{sup} + w_1 \times \mathcal{L}_{ps} + w_2 \times \mathcal{L}_{wpl} + w_3 \times \mathcal{L}_{sem} + w_4 \times \mathcal{L}_{feat} \ , \tag{6}$$

where w_1, w_2, w_3, w_4 are weights to balance the loss values.

3.3 Training Strategies

Training the model on the unlabeled set, particularly in the early stages, is challenging due to the domain gap between the real and synthetic data. We introduce a strategy to expedite model training by leveraging existing image restoration methods and our proposed VLM-based image assessment. Additionally, we enhance model performance through iterative updates of pseudo-labels, weather prompts, and descriptions in rounds.

Pseudo-label initialization. In the pseudo-label initialization stage, we gather the initial pseudo-labels by collecting the noisy restoration outcomes from both existing weather-specific and all-in-one image restoration methods. Subsequently, we employ the VLM-based image assessment technique to filter out the noisy samples and select the best-restored images as the pseudo-labels to initialize the pseudo-label database. To mitigate potential biases from a single vision-language model towards a specific image appearance, we utilize a diverse set of VLMs with varying architectures and parameters as experts for image assessment.

Iterative update. Leveraging expertise from multiple VLMs for image assessment to select pseudo-labels enhances the model learning process. However, due to computational constraints, it is impractical to consult every VLM during online learning. Instead, we divide the overall training into multiple rounds. In each round, only one VLM is employed for online image assessment. After a round of training, the overall assessment using the set of VLMs for pseudo-labels is conducted, incorporating new predictions from the updated model. Additionally, we update the weather prompts and augment the descriptions progressively. Note that the round number is empirically set as four during the training.

4 Experimental Results

4.1 Experimental Settings

Training & testing sets. We adopt several (pseudo-)synthetic datasets for deraining, dehazing, and desnowing, including Outdoor-Rain [16], RainDrop [30], SPA [42], OTS [15], and Snow100K [24]. Meanwhile, we leverage the unlabeled real-world adverse weather images for unsupervised learning. To achieve this, we utilize the real hazy images in URHI [15] (2,318 images) and manually collect real-world rainy and snowy images from the Internet (2,433 and 2,018 images, respectively). Besides, to train CLIP weather prompts, we employ high-quality DF2K [19,38] images for the *clear* category. We adopt real adverse weather image datasets for qualitative and quantitative evaluation, including RTTS [15] with 4,322 haze images, DDN-SIRR [45] and Real3000 [23] with 2,320 rain images, and Snow100K Realistic [24] with 1,329 snow images. Note that we remove the non-realistic images in Real3000, *e.g.*, comic and movie scenes.

Implementation details. Our semi-supervised learning framework is easily compatible with various image restoration networks. We opt for MSBDN [12] as our backbone due to its balanced performance and rapid inference speed in our main study. Each batch comprises eight labeled and unlabeled images, with a training process spanning 40,000 iterations per round. Image assessment utilizes recent VLMs [5, 20, 21, 35, 52]. Pseudo-labels are initialized through existing weather-specific and all-in-one adverse weather restoration methods. Empirical values for w_1, w_2, w_3, w_4 are set as $0.5, 0.2, 0.05, 0.2$, respectively. Implementation is based on BasicSR [43] and training is performed on two NVIDIA A40 GPUs.

4.2 Comparisons with the State-of-the-Art Methods

We benchmark our method against several state-of-the-art general and all-in-one adverse weather image restoration approaches. Our comparisons encompass recent works including Restormer [54], TransWeather [40], TKL [4], WeatherDiff [27], WGWS-Net [59], MWDT [28], PromptIR [29], and DA-CLIP [25]. We compare with the best-performing models from either retrained versions using our paired data or officially released checkpoints for fairness.

Table 1: Quantitative comparisons of deraining/dehazing/desnowing on real photos.

Method	NIMA [36] ↑ / MUSIQ [14] ↑ / CLIP-IQA [41] ↑			
	Rain	Haze	Snow	Overall
Restormer [54]	5.151 / 54.69 / 0.437	4.804 / 53.27 / 0.366	5.020 / 61.18 / 0.510	4.992 / 56.38 / 0.438
TransWeather [40]	5.068 / 51.06 / 0.358	4.716 / 46.27 / 0.292	4.928 / 59.38 / 0.416	4.904 / 52.24 / 0.355
TKL [4]	5.099 / 50.96 / 0.392	4.697 / 48.21 / 0.318	4.905 / 59.24 / 0.428	4.900 / 52.80 / 0.379
WeatherDiff [27]	5.054 / 51.82 / 0.395	4.616 / 47.70 / 0.326	4.917 / 60.52 / 0.466	4.862 / 53.35 / 0.396
WGWS-Net [59]	5.035 / 51.46 / 0.389	4.815 / 45.76 / 0.310	4.779 / 57.95 / 0.395	4.876 / 51.72 / 0.365
MWDT [28]	5.104 / 52.47 / 0.377	4.741 / 51.23 / 0.315	5.034 / 60.16 / 0.407	4.960 / 54.62 / 0.366
PromptIR [29]	5.174 / 53.48 / 0.439	4.823 / 53.88 / **0.372**	5.032 / 60.86 / 0.517	5.009 / 56.07 / 0.443
DA-CLIP [25]	5.168 / 52.98 / 0.412	4.851 / 53.23 / 0.325	5.012 / 60.57 / 0.499	5.010 / 55.59 / 0.412
Our method	**5.291** / **59.80** / **0.477**	**4.906** / **56.09** / 0.371	**5.057** / **62.12** / **0.519**	**5.084** / **59.34** / **0.456**

Method	LIQE [56] / Q-Align [47] ↑ / VLM-Vis ↑			
	Rain	Haze	Snow	Overall
Restormer [54]	2.277 / 3.795 / 0.417	1.918 / 3.068 / 0.218	3.172 / 3.646 / 0.395	2.456 / 3.503 / 0.343
TransWeather [40]	1.924 / 3.545 / 0.402	1.502 / 2.809 / 0.223	2.770 / 3.537 / 0.384	2.065 / 3.297 / 0.336
TKL [4]	2.028 / 3.588 / 0.406	1.590 / 2.908 / 0.238	2.830 / 3.557 / 0.393	2.149 / 3.351 / 0.346
WeatherDiff [27]	2.050 / 3.640 / 0.411	1.520 / 2.843 / 0.217	2.950 / 3.573 / 0.397	2.173 / 3.352 / 0.342
WGWS-Net [59]	1.965 / 3.592 / 0.411	1.506 / 2.915 / 0.238	2.619 / 3.490 / 0.383	2.030 / 3.332 / 0.344
MWDT [28]	2.068 / 3.548 / 0.426	1.720 / 2.861 / 0.273	2.903 / 3.569 / 0.412	2.230 / 3.326 / 0.370
PromptIR [29]	2.250 / 3.770 / 0.419	1.941 / 3.093 / 0.226	3.121 / 3.609 / 0.384	2.437 / 3.491 / 0.343
DA-CLIP [25]	2.250 / 3.732 / 0.412	2.014 / 3.071 / 0.230	3.050 / 3.637 / 0.395	2.438 / 3.480 / 0.346
Our method	**2.563** / **3.843** / **0.440**	**2.064** / **3.176** / **0.289**	**3.293** / **3.702** / **0.431**	**2.640** / **3.574** / **0.387**

Quantitative comparison. Note that there is no ground-truth clear image for the real adverse weather images. Therefore, we adopt several no-reference metrics for the quantitative assessment. Specifically, we use recent blind image quality evaluation metrics, including NIMA [36], MUSIQ [14], CLIP-IQA [41], LIQE [56], and Q-Align [47]. We also utilize the proposed VLM-based image visibility assessment method and report the normalized scores VLM-Vis. In detail, VLM-Vis is computed over VLM experts, standardized by the minimum and maximum statistics across the dataset for each respective VLM. The quantitative comparisons are reported in Table 1. Our proposed method is ranked first for all image quality assessment metrics on average and in almost all weather conditions. These values indicate the superior restoration quality of the images. Moreover, our method achieves the best VLM-Vis across different weather conditions. These results demonstrate the advantages of our method on real data against existing advanced adverse weather image restoration methods, which focus mainly on synthetic data evaluation.

Qualitative comparison. Our qualitative assessment is conducted on real-world evaluation datasets [15, 23, 24, 45] and the visual outcomes are presented in Fig. 6. We can observe that the compared methods are less effective in dealing with real-world adverse weather images and are limited in removing rain, haze, and snow artifacts. It is noted that MWDT [28] mitigates the haze effect but introduces severe color distortion. In comparison, our method exhibits superior visually perceptual quality, enhancing clarity and contrast, while minimizing

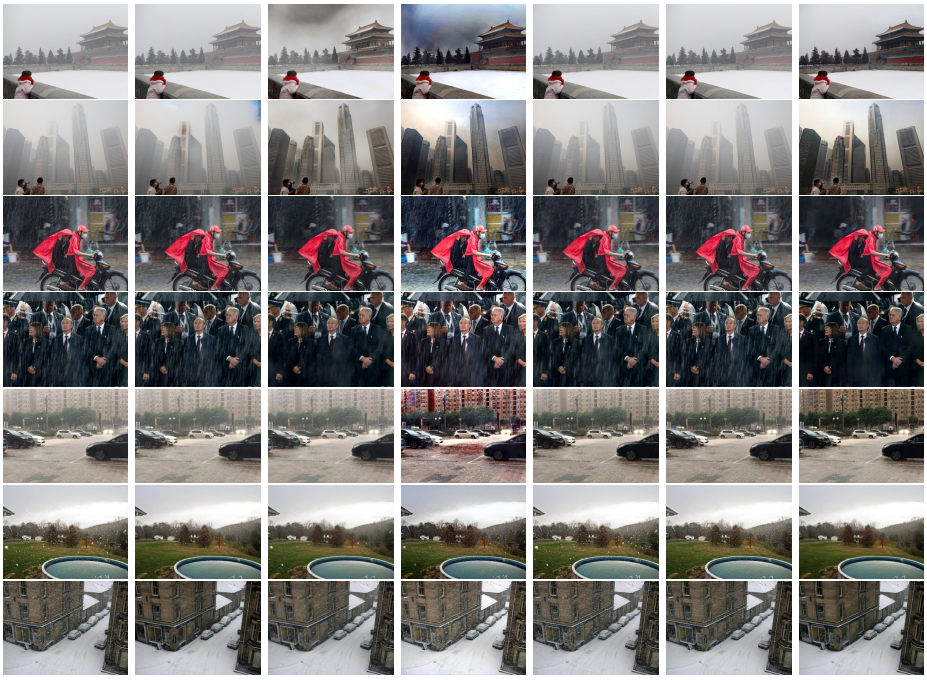

Fig. 6: Visual comparisons on real-world images [15, 23, 24, 45].

rain, haze, and snow artifacts. Notably, our approach effectively eliminates haze in rain and snow scenarios, significantly improving image visibility.

User study. We conducted a user study to evaluate the visual quality. For each weather scenario, ten real-world images are chosen. 32 participants were invited for the evaluation. Two factors are considered, *i.e.*, image visibility and quality, regarding the extent to which the weather-related artifacts are removed and the restored image is kept real. As observed in Fig. 7, MWDT obtains high image visibility scores, which aligns with our VLM-Vis metric. Overall, our method exhibits a clear advantage in visibility and quality across weather conditions.

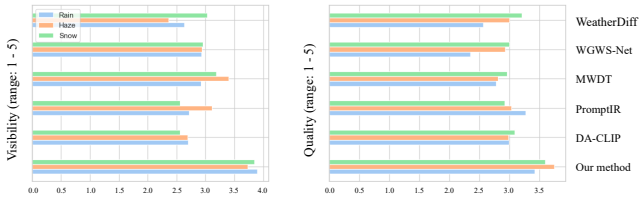

Fig. 7: User study on visibility and quality of image restoration.

Fig. 8: Ablation studies of the proposed semi-supervised learning framework.

Table 2: Ablation analysis of each component design.

$\mathcal{L}_{sup}$	$\mathcal{L}_{ps}$	r^{vlm}	*init*	$\mathcal{L}_{wpl}$	$\mathcal{L}_{sem}$	*iter*	MUSIQ ↑	CLIP-IQA ↑	VLM-Vis ↑
✓							53.41	0.388	0.343
✓	✓						54.08	0.396	0.354
✓	✓	✓					56.68	0.429	0.366
✓	✓	✓	✓				57.34	0.425	0.370
✓	✓	✓	✓	✓			58.13	0.437	0.376
✓	✓	✓	✓	✓	✓		58.91	0.445	0.381
✓	✓	✓	✓	✓	✓	✓	**59.34**	**0.456**	**0.387**

4.3 Ablation Studies

Effectiveness of semi-supervised learning framework. We start with the baseline model, trained exclusively through supervised learning ($\mathcal{L}_{sup}$) on labeled synthetic data. Subsequently, we employ the naive mean-teacher [37], a semi-supervised learning method, to explore unlabeled real data and utilize predictions from the teacher network as pseudo-labels ($\mathcal{L}_{ps}$). We investigate the effectiveness of the proposed VLM-based components and training strategies, including: (1) Incorporating VLM-based image assessment r^{vlm} for updating pseudo-labels, (2) Pseudo-label initialization (*init*), (3) Weather prompt learning ($\mathcal{L}_{wpl}$), (4) Semantics regularization ($\mathcal{L}_{sem}$), and (5) Iterative update (*iter*).

Quantitative outcomes with overall performance across different weather conditions are presented in Table 2, while visual comparisons are depicted in Fig. 8. It is evident that the baseline, trained solely on synthetic data using a straightforward semi-supervised learning approach, struggles to effectively address real rain, haze, and snow artifacts. In contrast, our proposed VLM-based image assessment progressively refines the selection of superior pseudo-labels, emphasizing higher clearness and resulting in predictions with improved visibility. This effect is further amplified with the incorporation of the pseudo-label initialization strategy. Moreover, the proposed weather prompt learning and description-assisted semantic enhancement largely improve the restoration performance. This is evidenced by the boosted image quality, the visibility metric scores, and the visual quality with reduced weather-related artifacts (Fig. 8). Lastly, our iterative training strategy further enhances the overall quantitative and qualitative outcomes.

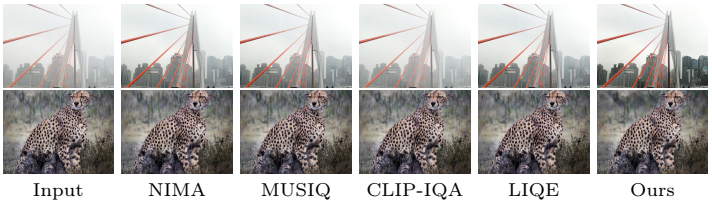

Fig. 9: Ablation studies of the proposed image assessment method.

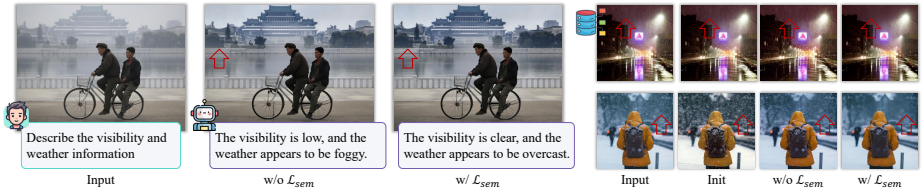

Fig. 10: Analysis of the semantics regularization.

Impact of VLM-based image assessment. We conduct experiments to investigate the VLM-based image assessment for selecting pseudo-labels. We compare our proposed method with existing image quality assessment metrics, including NIMA [36], MUSIQ [14], CLIP-IQA [41], and LIQE [56], by replacing the pseudo-label update criteria. As discussed in Sec. 3.1, our VLM-based rating approach can select pseudo-labels with less weather-related artifacts. Consequently, the trained models show superior restoration ability, as illustrated in Fig. 9.

Analysis of semantics regularization. We study the impact on the semantics regularization based on Fig. 10. The VLM [5] can detect the nuanced difference of the restored image to be foggy or overcast. By further monitoring the training process, we observe that the semantics-enhanced approach benefits learning by leading to better-stored pseudo-labels and subsequent training. Hence, the model trained with the description-assisted semantics regularization $\mathcal{L}_{sem}$ addresses the subtle weather context misalignment, improving the visual quality.

5 Conclusion

This paper advances real-world adverse weather image restoration using vision-language models, overcoming the limitations of methods trained on synthetic data. By evaluating clearness and semantics in natural images, our semi-supervised approach trains models on real, unlabeled images. Our dual-step strategy, combining image assessment and weather prompt learning, enhances clearness with real data. Further, semantics enhancement adjusts weather conditions in vision-language model descriptions, addressing context semantics in adverse weather. Experimental results show that our method outperforms state of the arts. Yet, the computational burden of using large VLMs remains a limitation.

Acknowledgements

The work was supported by the National Key R&D Program of China (Grant No. 2022ZD0160100), the Research Grants Council of the Hong Kong Special Administrative Region, China (Grant No. 14201620), and the Hong Kong Innovation and Technology Fund (Grant No. MHP/092/22).

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Cai, B., Xu, X., Jia, K., Qing, C., Tao, D.: Dehazenet: An end-to-end system for single image haze removal. TIP (2016)
3. Chen, W.T., Fang, H.Y., Hsieh, C.L., Tsai, C.C., Chen, I., Ding, J.J., Kuo, S.Y., et al.: All snow removed: Single image desnowing algorithm using hierarchical dual-tree complex wavelet representation and contradict channel loss. In: ICCV (2021)
4. Chen, W.T., Huang, Z.K., Tsai, C.C., Yang, H.H., Ding, J.J., Kuo, S.Y.: Learning multiple adverse weather removal via two-stage knowledge learning and multi-contrastive regularization: Toward a unified model. In: CVPR (2022)
5. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: CVPR (2024)
6. Deng, Z., Zhu, L., Hu, X., Fu, C.W., Xu, X., Zhang, Q., Qin, J., Heng, P.A.: Deep multi-model fusion for single-image dehazing. In: ICCV (2019)
7. Fu, X., Huang, J., Zeng, D., Huang, Y., Ding, X., Paisley, J.: Removing rain from single images via a deep detail network. In: CVPR (2017)
8. He, K., Sun, J., Tang, X.: Single image haze removal using dark channel prior. TPAMI (2010)
9. Hu, X., Fu, C.W., Zhu, L., Heng, P.A.: Depth-attentional features for single-image rain removal. In: CVPR (2019)
10. Hu, X., Zhu, L., Wang, T., Fu, C.W., Heng, P.A.: Single-image real-time rain removal based on depth-guided non-local features. TIP (2021)
11. Huang, S., Wang, K., Liu, H., Chen, J., Li, Y.: Contrastive semi-supervised learning for underwater image restoration via reliable bank. In: CVPR (2023)
12. Jiang, K., Wang, Z., Yi, P., Chen, C., Huang, B., Luo, Y., Ma, J., Jiang, J.: Multi-scale progressive fusion network for single image deraining. In: CVPR (2020)
13. Jiang, Y., Zhang, Z., Xue, T., Gu, J.: Autodir: Automatic all-in-one image restoration with latent diffusion. arXiv preprint arXiv:2310.10123 (2023)
14. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. In: ICCV (2021)
15. Li, B., Ren, W., Fu, D., Tao, D., Feng, D., Zeng, W., Wang, Z.: Benchmarking single-image dehazing and beyond. TIP (2018)
16. Li, R., Cheong, L.F., Tan, R.T.: Heavy rain image restoration: Integrating physics model and conditional adversarial learning. In: CVPR (2019)
17. Li, R., Tan, R.T., Cheong, L.F.: All in one bad weather removal using architectural search. In: CVPR (2020)
18. Liang, Z., Li, C., Zhou, S., Feng, R., Loy, C.C.: Iterative prompt learning for unsupervised backlit image enhancement. In: ICCV (2023)

19. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: CVPRW (2017)
20. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: CVPR (2024)
21. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llava-next: Improved reasoning, ocr, and world knowledge (January 2024), https://llava-vl.github.io/blog/2024-01-30-llava-next/
22. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023)
23. Liu, Y., Yue, Z., Pan, J., Su, Z.: Unpaired learning for deep image deraining with rain direction regularizer. In: ICCV (2021)
24. Liu, Y.F., Jaw, D.W., Huang, S.C., Hwang, J.N.: Desnownet: Context-aware deep network for snow removal. TIP (2018)
25. Luo, Z., Gustafsson, F.K., Zhao, Z., Sjölund, J., Schön, T.B.: Controlling vision-language models for universal image restoration. In: ICLR (2024)
26. Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., et al.: Language models are few-shot learners. In: NeurIPS (2020)
27. Özdenizci, O., Legenstein, R.: Restoring vision in adverse weather conditions with patch-based denoising diffusion models. TPAMI (2023)
28. Patil, P.W., Gupta, S., Rana, S., Venkatesh, S., Murala, S.: Multi-weather image restoration via domain translation. In: ICCV (2023)
29. Potlapalli, V., Zamir, S.W., Khan, S., Khan, F.S.: Promptir: Prompting for all-in-one blind image restoration. In: NeruIPS (2023)
30. Qian, R., Tan, R.T., Yang, W., Su, J., Liu, J.: Attentive generative adversarial network for raindrop removal from a single image. In: CVPR (2018)
31. Qin, X., Wang, Z., Bai, Y., Xie, X., Jia, H.: Ffa-net: Feature fusion attention network for single image dehazing. In: AAAI (2020)
32. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)
33. Richardson, E., Goldberg, K., Alaluf, Y., Cohen-Or, D.: Conceptlab: Creative generation using diffusion prior constraints. arXiv preprint arXiv:2308.02669 (2023)
34. Song, Y., He, Z., Qian, H., Du, X.: Vision transformers for single image dehazing. TIP (2023)
35. Sun, Q., Cui, Y., Zhang, X., Zhang, F., Yu, Q., Wang, Y., Rao, Y., Liu, J., Huang, T., Wang, X.: Generative multimodal models are in-context learners. In: CVPR (2024)
36. Talebi, H., Milanfar, P.: NIMA: Neural image assessment. TIP (2018)
37. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: NeruIPS (2017)
38. Timofte, R., Agustsson, E., Van Gool, L., Yang, M.H., Zhang, L.: Ntire 2017 challenge on single image super-resolution: Methods and results. In: CVPRW (2017)
39. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
40. Valanarasu, J.M.J., Yasarla, R., Patel, V.M.: Transweather: Transformer-based restoration of images degraded by adverse weather conditions. In: CVPR (2022)
41. Wang, J., Chan, K.C., Loy, C.C.: Exploring clip for assessing the look and feel of images. In: AAAI (2023)

42. Wang, T., Yang, X., Xu, K., Chen, S., Zhang, Q., Lau, R.W.: Spatial attentive single-image deraining with a high quality real rain dataset. In: CVPR (2019)
43. Wang, X., Xie, L., Yu, K., Chan, K.C., Loy, C.C., Dong, C.: BasicSR: Open source image and video restoration toolbox. https://github.com/XPixelGroup/BasicSR (2022)
44. Wang, Y., Ma, C., Liu, J.: Smartassign: Learning a smart knowledge assignment strategy for deraining and desnowing. In: CVPR (2023)
45. Wei, W., Meng, D., Zhao, Q., Xu, Z., Wu, Y.: Semi-supervised transfer learning for image rain removal. In: CVPR (2019)
46. Wu, H., Zhang, Z., Zhang, E., Chen, C., Liao, L., Wang, A., Li, C., Sun, W., Yan, Q., Zhai, G., et al.: Q-bench: A benchmark for general-purpose foundation models on low-level vision. In: ICLR (2024)
47. Wu, H., Zhang, Z., Zhang, W., Chen, C., Liao, L., Li, C., Gao, Y., Wang, A., Zhang, E., Sun, W., et al.: Q-align: Teaching lmms for visual scoring via discrete text-defined levels. In: ICML (2024)
48. Xiao, J., Fu, X., Liu, A., Wu, F., Zha, Z.J.: Image de-raining transformer. TPAMI (2022)
49. Xu, J., Hu, X., Zhu, L., Dou, Q., Dai, J., Qiao, Y., Heng, P.A.: Video dehazing via a multi-range temporal alignment network with physical prior. In: CVPR (2023)
50. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: CVPR (2024)
51. Yang, W., Tan, R.T., Feng, J., Liu, J., Guo, Z., Yan, S.: Deep joint rain detection and removal from a single image. In: CVPR (2017)
52. Ye, Q., Xu, H., Ye, J., Yan, M., Hu, A., Liu, H., Qian, Q., Zhang, J., Huang, F.: mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In: CVPR (2024)
53. Ye, T., Chen, S., Bai, J., Shi, J., Xue, C., Jiang, J., Yin, J., Chen, E., Liu, Y.: Adverse weather removal with codebook priors. In: ICCV (2023)
54. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: CVPR (2022)
55. Zhang, H., Ba, Y., Yang, E., Mehra, V., Gella, B., Suzuki, A., Pfahnl, A., Chandrappa, C.C., Wong, A., Kadambi, A.: Weatherstream: Light transport automation of single image deweathering. In: CVPR (2023)
56. Zhang, W., Zhai, G., Wei, Y., Yang, X., Ma, K.: Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In: CVPR (2023)
57. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. IJCV (2022)
58. Zhu, L., Deng, Z., Hu, X., Xie, H., Xu, X., Qin, J., Heng, P.A.: Learning gated non-local residual for single-image rain streak removal. TCSVT (2020)
59. Zhu, Y., Wang, T., Fu, X., Yang, X., Guo, X., Dai, J., Qiao, Y., Hu, X.: Learning weather-general and weather-specific features for image restoration under multiple adverse weather conditions. In: CVPR (2023)