

# AWRaCLE: All-Weather Image Restoration using Visual In-Context Learning

Sudarshan Rajagopalan, Vishal M. Patel

Johns Hopkins University

{sambasa2, vpatel36}@jhu.edu

## Abstract

All-Weather Image Restoration (AWIR) under adverse weather conditions is a challenging task due to the presence of different types of degradations. Prior research in this domain relies on extensive training data but lacks the utilization of additional contextual information for restoration guidance. Consequently, the performance of existing methods is limited by the degradation cues that are learnt from individual training samples. Recent advancements in visual in-context learning have introduced generalist models that are capable of addressing multiple computer vision tasks simultaneously by using the information present in the provided context as a prior. In this paper, we propose *All-Weather Image Restoration using Visual In-Context Learning* (AWRaCLE), a novel approach for AWIR that innovatively utilizes degradation-specific visual context information to steer the image restoration process. To achieve this, AWRaCLE incorporates Degradation Context Extraction (DCE) and Context Fusion (CF) to seamlessly integrate degradation-specific features from the context into an image restoration network. The proposed DCE and CF blocks leverage CLIP features and incorporate attention mechanisms to adeptly learn and fuse contextual information. These blocks are specifically designed for visual in-context learning under all-weather conditions and are crucial for effective context utilization. Through extensive experiments, we demonstrate the effectiveness of AWRaCLE for all-weather restoration and show that our method advances the state-of-the-art in AWIR.

## 1 Introduction

Unfavorable weather conditions, such as rain, snow and haze, significantly degrade the performance of computer vision systems impacting applications such as autonomous navigation, surveillance, and aerial imaging. Thus, there is a need for frameworks that mitigate weather-induced corruptions while preserving the underlying image semantics. Initial attempts at this problem (He, Sun, and Tang 2009; Roth and Black 2005; Kang, Lin, and Fu 2012) worked by modelling the physics of the degradations, and estimated physics-based parameters to obtain the restored image. These methods were not effective due to the varying degradation characteristics that occur in real-life scenarios. To overcome these limitations, many deep learning-based methods such as (Zhang, Sindagi, and Patel 2020; Zhang

and Patel 2018; Wu et al. 2021; Song et al. 2023) for dehazing, (Zhang et al. 2021; Chen et al. 2020; Liu et al. 2018) for desnowing, and (Wang et al. 2019; Wei et al. 2019; Quan et al. 2021) for derraining have been proposed. However, these methods are specific to a single degradation. For handling multiple degradations, Uformer (Wang et al. 2022), Restormer (Zamir et al. 2022), MPRNet (Zamir et al. 2021) and SwinIR (Liang et al. 2021) were proposed. These approaches can be trained to remove any one type of degradation at a time. Using them for a different degradation involves either re-training or fine-tuning the networks, rendering them impractical to use in real-life settings.

To address the above issues, there has been focus on developing All-Weather Image Restoration (AWIR) networks in which a single model is trained to restore multiple degradations. Popular works for tackling this problem include All-in-one (Li, Tan, and Cheong 2020), Transweather (Valanarasu, Yasarla, and Patel 2022), AirNet (Li et al. 2022), FAIG (Park, Lee, and Chun 2023), WeatherDiff (Özdenizci and Legenstein 2023), TSMC (Chen et al. 2022b), PromptIR (Potlapalli et al. 2024) and DiffUIR (Zheng et al. 2024). These methods attempt to learn degradation-specific representations from individual images during the training phase. However, they lack guidance that can provide additional information about the various degradations present in images. This limits their ability to effectively learn features unique to different degradations, ultimately impeding their performance for AWIR. Thus, there is a need for a framework capable of learning robust degradation-specific features that can facilitate effective restoration. This is quite a challenging task and proves to be difficult without supplementary knowledge about the nature of the corruption. A couple of works (Yan et al. 2023; Bai et al. 2023) have explored text-based guidance for all-weather restoration. However, text descriptions can only convey high-level semantic information about the degradation and fail to describe important aspects of the corruption such as its visual characteristics. We conjecture that feeding images as degradation-context to an image restoration network can overcome some of the above limitations as it will enable the network to aggregate visual information about different types of weather conditions. We accomplish this with the help of *visual in-context learning*.

In-context learning, as demonstrated by large language models, is very well-studied in Natural Language Process-

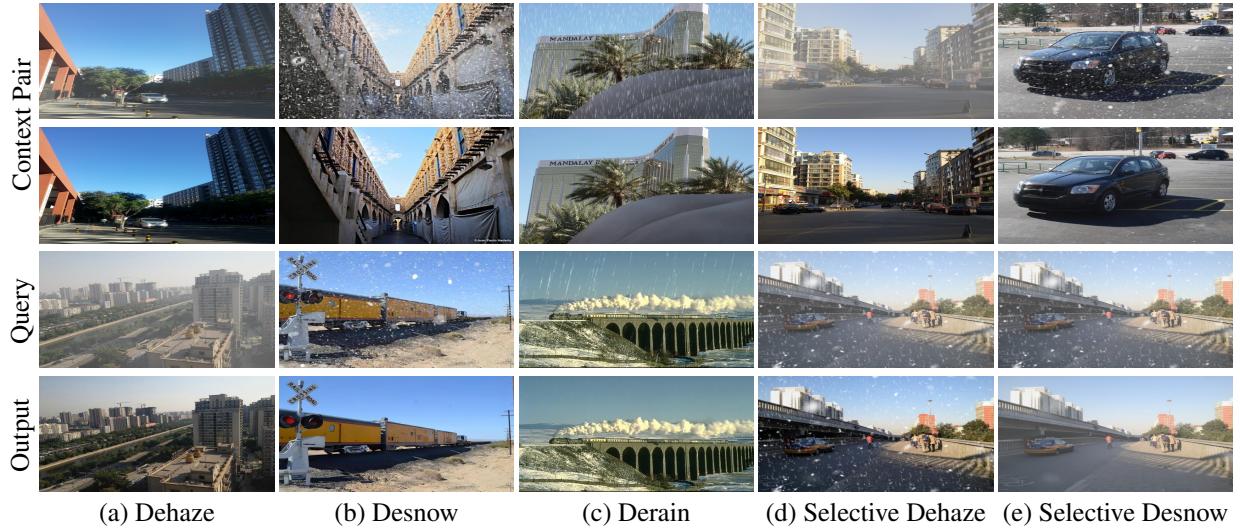


Figure 1: Illustration of AWRaCLE: Our visual in-context learning approach for all-weather image restoration. The first two rows are the context pair. The third row is the query image that needs to be restored and the fourth row is our output. (d) and (e) show results for selective removal of haze and snow, respectively, from an image containing their mixture.

ing (NLP). In comparison, visual in-context learning is an emerging area. Providing visual context has been explored by (Bar et al. 2022), Painter (Wang et al. 2023a) and Seg-GPT (Wang et al. 2023b). They propose elegant solutions that involve unifying the output space of a network to solve multiple computer vision tasks based on visual context. The core idea of these frameworks is the usage of masked image modelling to extract context information from the unmasked regions of an image (as detailed in Painter) and subsequently employing in-painting for the desired prediction. Although Painter was trained for image restoration tasks using visual in-context learning, we show that it fails to use contextual information effectively for restoration tasks, thereby significantly hindering its performance. We believe this is due to two primary reasons. Firstly, the extraction of context information relies solely on the masked image modelling framework, which lacks constraints to ensure retrieval of degradation characteristics from the context images. Secondly, the context information is provided only at the input of the encoder network in the form of unmasked tokens. Consequently, there is a risk of the context information being suppressed after the initial layers of the network and even becoming negligible at the decoder.

We propose AWRaCLE, a methodology for all-weather (rain, snow and haze) image restoration which elegantly leverages visual in-context learning. Our method aims to restore a query image by utilizing additional context (that we call context pair), which comprises of a degraded image and its corresponding clean version (see Fig. 1). The context requires paired images since without scene consistency, it is challenging to extract degradation-specific information (DSI) which will in-turn hamper restoration performance. In order to facilitate the restoration process, the degradation type in the context pair should align with the type of degradation present in the query image. During test time, the con-

text pair for a degradation can be chosen from the respective training set, thus requiring only the knowledge of the type of degradation.

We devise Degradation Context Extraction (DCE) blocks that leverage features from CLIP’s (Radford et al. 2021) image encoder and employ self-attention mechanisms to extract relevant DSI, such as the type and visual characteristics of the degradation, from the given context pair. Additionally, we introduce Context Fusion (CF) blocks designed to integrate the extracted context from the DCE blocks with the feature maps of an image restoration network. Specifically, we use the Restormer (Zamir et al. 2022) network for this purpose. The fusion process involves multi-head cross attention at each spatial level of the decoder, ensuring the propagation of context information throughout the restoration network, thus enhancing the performance. Representative results on haze, snow and rain removal are given in Fig. 1 to demonstrate the efficacy of our method. Interestingly, AWRaCLE can harness DSI from the context pair to perform selective degradation removal. For instance in Fig. 1d and e, a query image corrupted by both haze and snow is presented as input. In Fig. 1d, the context pair is given as haze and AWRaCLE returned an output that contained only snow. Similarly, in Fig. 1e, the context pair is given as snow resulting in an output that contained only haze. This reaffirms the ability of AWRaCLE in utilising degradation-context effectively. We have performed extensive experiments to demonstrate the effectiveness of AWRaCLE and show that our method achieves state-of-the-art performance for the AWIR task.

In summary, our main contributions are as follows:

1. We propose a novel approach called AWRaCLE that employs visual in-context learning for AWIR. To the best of our knowledge, this is the first work which effectively utilizes visual degradation context for AWIR.
2. We propose novel Degradation Context Extraction

- blocks and Context Fusion blocks which extract and fuse relevant degradation information from the provided visual context. Our method ensures that the extracted context is injected suitably at different stages of the restoration network to enable context information flow.
- Through comprehensive experiments, we show that AWRaCLE achieves state-of-the-art all-weather restoration performance on multiple benchmark datasets.

## 2 Related Works

In this section, we discuss relevant works on adverse weather restoration and in-context learning.

### 2.1 Adverse weather restoration

Several methods have been proposed for single weather restoration such as (Wang et al. 2019; Wei et al. 2019) for deraining, (Zhang, Sindagi, and Patel 2020; Zhang and Patel 2018) for dehazing and (Zhang et al. 2021; Chen et al. 2020) for desnowing. Recently, methods such as Restormer (Zamir et al. 2022) and MPRNet (Zamir et al. 2021) have tackled multiple degradations. However, the above methods require retraining or fine-tuning for each degradation. To overcome this limitation, AWIR methods have been actively explored. All-in-one (Li, Tan, and Cheong 2020) used neural architecture search to find the best-suited encoder for each degradation from a set of encoders. Transweather (Valanarasu, Yasarla, and Patel 2022) employed a unified network with a single encoder for multi-weather restoration. Airnet (Li et al. 2022) used Momentum Contrast (MoCo) (He et al. 2020) for improved degradation representations while TSMC (Chen et al. 2022b) proposed two-stage knowledge learning with multi-contrastive regularization for a similar objective. Recently, WeatherDiff (Özdenizci and Legenstein 2023) proposed a patch-based denoising diffusion model for adverse weather removal and WGWS (Zhu et al. 2023) extracted weather-general and weather-specific features for restoration. PromptIR (Potlapalli et al. 2024) utilized learnable prompt embeddings for AWIR while DiffUIR (Zheng et al. 2024) proposed selective hourglass mapping. Although several AWIR methods have been proposed, they do not utilize contextual guidance which limits their performance.

### 2.2 In-context Learning

Transformers have a generalized modeling capability through the use of tokens. Leveraging this, DETR (Carion et al. 2020) used transformer heads for object detection. Pix2Seq (Chen et al. 2021) discretized the output space of object detection. Unified-IO (Lu et al. 2022) and Pix2Seqv2 (Chen et al. 2022a) extended this approach to multiple vision tasks (generalist models) using task prompts. These methods use a discrete output space, which is unsuitable for continuous space of image data, making it challenging to enable visual in-context learning.

Recent advancements in in-context learning have significantly improved the zero-shot performance of large language models. GPT-3 (Brown et al. 2020) demonstrated this using text completion with prompts as context while Flamingo (Alayrac et al. 2022) used language guidance for

various image and video tasks. Visual in-context learning is an emerging area which is gaining increasing attention. VPI (Bar et al. 2022) proposed an image-based continuous output space for visual in-context segmentation. Directly using visual context for tackling multiple computer vision tasks is challenging due to the non-unified output space. To address this, Painter (Wang et al. 2023a) extended VPI for diverse computer vision tasks by unifying their output spaces. However, these approaches rely heavily on masked-image modeling to learn context (see Painter), which is ineffective for image restoration because there is no dedicated module to capture degradation-specific information. Additionally, context information provided at the network’s input may not propagate to deeper layers.

## 3 Proposed Methodology

In this section, we explain in detail our proposed approach, AWRaCLE, for performing AWIR (deraining, desnowing and dehazing) using visual in-context learning. A high-level schematic of AWRaCLE is shown in Fig. 2. The main idea of our approach involves extracting relevant degradation-context such as the type and visual characteristics of degradations from a given image-ground truth pair to effectively restore a query image with the same type of degradation. Toward this aim, we propose Degradation Context Extraction (DCE) and Context Fusion (CF) blocks that learn context information and fuse it with an image restoration network to facilitate the restoration process. Specifically, we integrate our DCE and CF blocks with a slightly modified version of the Restormer network (see supplementary for details). The DCE and CF blocks are added at each decoder level of Restormer for propagation of context information (multi-level fusion). The decoder levels are represented by  $l = 0, 1, 2$  and  $3$  in Fig. 2. AWRaCLE overcomes the limitations of Painter which solely relies on masked image modelling to extract context information. Also, they provide context information only at the input, thus lacking any mechanism to ensure its flow throughout the network.

**Terminology.** For ease of understanding, we define a few terms. We refer to the context-pair as  $\mathcal{C} = \{I_d, I_c\}$  where  $I_d$  is the degraded image and  $I_c$  is its corresponding clean image.  $I_q$  is the degraded query image which needs to be restored given  $\mathcal{C}$ . Note that  $I_q$  and  $I_d$  are affected by the same type of degradation. Additionally,  $I_d, I_c, I_q \in \mathbb{R}^{H \times W \times 3}$  where  $H, W$  indicate the spatial resolution of the images.

### 3.1 Degradation Context Extraction

The objective of the DCE blocks is to extract degradation-specific context such as the type and visual characteristics from  $\mathcal{C}$ . It is crucial for the underlying scene content in  $I_d$  and  $I_c$  to be identical so that the only distinction between them is the degradation (we call this paired context). This condition facilitates the process of extracting degradation-specific information (DSI) from  $I_d$  and  $I_c$ . In the ablations, we show that using un-paired context ( $I_d$  and  $I_c$  are from different scenes) leads to inferior performance. Furthermore, it is important to note the considerable difficulty in extracting degradation-context solely from  $I_d$ . This challenge

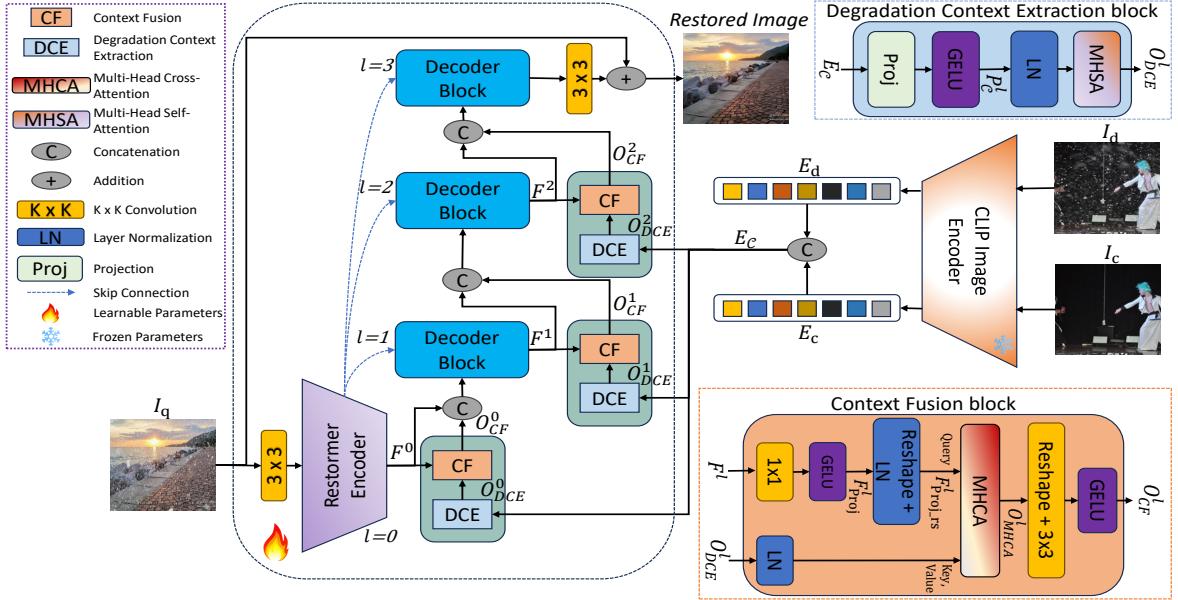


Figure 2: Block diagram of the proposed visual in-context learning approach for AWIR. CLIP features are extracted from  $I_d$  and  $I_c$  which are subsequently fed to DCE blocks at different decoder levels,  $l$ . CF blocks then fuse the degradation information obtained from the DCE blocks with decoder features,  $F^l$ , from the query image  $I_q$ . Finally, the restored image is generated.

arises since it is not straightforward to disentangle the scene content from the degradation.

We now elaborate the aforementioned process of extracting degradation-context from  $\mathcal{C}$ . Vision-Language models (VLMs) such as CLIP (Radford et al. 2021) have demonstrated the capability to learn high quality image embeddings that can be used to solve a myriad of downstream computer vision tasks (Gu et al. 2021; Zhang et al. 2022, 2018; Shen et al. 2021). Motivated by this, we obtain representations  $E_d$ ,  $E_c$  for the context pair  $\mathcal{C}$  from the final transformer block of CLIP’s image encoder (denoted by  $\text{CLIP}(\cdot)$ ). The obtained features are then fed to the DCE blocks that are present at each decoder level  $l$  of the network. The above feature extraction step using CLIP is given as follows.

$$E_d = \text{CLIP}(I_d), E_c = \text{CLIP}(I_c), \{E_d, E_c\} \in \mathbb{R}^{L \times D} \quad (1)$$

$E_d$  and  $E_c$  are then concatenated to obtain  $E_{\mathcal{C}} \in \mathbb{R}^{2L \times D}$  as the overall CLIP representation for  $\mathcal{C}$ . Here,  $L$  represents the number of tokens and  $D$  is the embedding dimension. Within a DCE block at level  $l$ ,  $E_{\mathcal{C}}$  is initially projected to a lower dimension to reduce computational complexity for forthcoming attention operations. This is followed by GELU (Hendrycks and Gimpel 2016) activation function as non-linearity. The result of these operations is  $P_{\mathcal{C}}^l \in \mathbb{R}^{2L \times C^l}$ , where  $C^l$  is the projection dimension, and these steps are summarised as below.

$$P_{\mathcal{C}}^l = \text{GELU}(\text{Proj}(E_{\mathcal{C}})), P_{\mathcal{C}}^l \in \mathbb{R}^{2L \times C^l} \quad (2)$$

The projected feature,  $P_{\mathcal{C}}^l$ , is normalized using layer normalization (LN) (Ba, Kiros, and Hinton 2016). Subsequently, Multi-Head Self-Attention (MHSA(.)) (Dosovitskiy et al. 2020) is employed to capture DSI,  $O_{\text{DCE}}^l$ , from  $P_{\mathcal{C}}^l$ . This step can be summarized as

$$O_{\text{DCE}}^l = \text{MHSA}(\text{LN}(P_{\mathcal{C}}^l)), O_{\text{DCE}}^l \in \mathbb{R}^{2L \times C^l}. \quad (3)$$

Since the scene is consistent in both  $I_d$  and  $I_c$ , the primary distinction between them is the degradation which is adeptly discerned through MHSA. This enables extraction of the necessary degradation-context from  $\mathcal{C}$  for judiciously guiding the network towards the objective of all-weather restoration. Fig. 2 shows a detailed schematic of the DCE block, highlighting the above steps. Additional details about the MHSA module are given in the supplementary document.

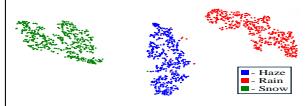
To visualize the extracted degradation-specific information, we overlay the output of the DCE block ( $O_{\text{DCE}}^l$ ) for a clean-hazy and a clean-snowy context pair, respectively. This is illustrated in Fig. 3a where activations are overlaid on the degraded image ( $I_d$ ) and its corresponding clean image ( $I_c$ ).  $A_d \in \mathbb{R}^{L \times C^l}$  and  $A_c \in \mathbb{R}^{L \times C^l}$  represent the DCE block activations obtained by splitting  $O_{\text{DCE}}^l$  for  $I_d$  and  $I_c$ , respectively. The figure shows that the DCE block captures DSI such as the spatially-varying characteristics of haze and sparseness of snow. Furthermore, to discern this information, the DCE block uses the clean image ( $I_c$ ) to identify and focus on degraded regions in  $I_d$ , evident from attention at similar locations in both  $I_d$  and  $I_c$ . Additionally, Fig. 3b provides a t-SNE plot of the DCE block output ( $O_{\text{DCE}}^l$ ), showing that DSI is clustered closely for the same type of degradation but is separated for different degradations.

### 3.2 Context Fusion

At each level  $l$ , the obtained degradation-context,  $O_{\text{DCE}}^l$ , needs to be fused with the corresponding decoder features,  $F^l \in \mathbb{R}^{K^l \times H^l \times W^l}$ , from Restomer. Here  $H^l$  and  $W^l$  are the spatial resolution of the feature map, and  $K^l$  is channel dimension. Fusion is achieved with the help of the Context Fusion (CF) blocks that utilize Multi-Head Cross Atten-



(a) DCE block activations  $A_d$  and  $A_c$  overlaid (+) on  $I_d$  and  $I_c$ , respectively, of the context pair. Yellow-High, Blue-Low



(b) t-SNE plot of DCE block outputs ( $O_{\text{DCE}}^l$ ) for hazy, rainy and snowy context pairs.

Figure 3: Analysis of DCE block outputs.

tion (MHCA ( . )) (Vaswani et al. 2017) to integrate information from  $O_{\text{DCE}}^l$  and  $F^l$ . The cross-attention mechanism is a key ingredient in the CF module as we want  $F^l$  to be enhanced by the degradation information contained in  $O_{\text{DCE}}^l$ . We achieve this by treating  $F^l$  as the query and matching it with the key and value computed from  $O_{\text{DCE}}^l$ .

The CF module which is illustrated in Fig. 2 is next described in detail. Prior to MHCA,  $F^l$  is projected to the same channel dimension ( $C^l$ ) as  $O_{\text{DCE}}^l$  using  $1 \times 1$  convolution as

$$F_{\text{Proj}}^l = \text{GELU}(\text{Conv}_{1 \times 1}(F^l)), F_{\text{Proj}}^l \in \mathbb{R}^{C^l \times H^l \times W^l}, \quad (4)$$

where we have used the GELU activation function as non-linearity. We observe that  $F_{\text{Proj}}^l$  and  $O_{\text{DCE}}^l$  have a mismatch in the number of dimensions ( $O_{\text{DCE}}^l$  is 2-D but  $F_{\text{Proj}}^l$  is 3-D), which precludes the use of standard MHCA operation. One plausible approach involves the use of reshaping operations followed by interpolation to transform  $O_{\text{DCE}}^l$  into the same dimension as  $F_{\text{Proj}}^l$ . However, interpolation causes redundancy in the degradation-context thereby hindering the performance of cross-attention. To circumvent this problem, we reshape (denoted as RH( . ))  $F_{\text{Proj}}^l$  to obtain  $F_{\text{Proj},rs}^l \in \mathbb{R}^{H^l \cdot W^l \times C^l}$  which has the same channel dimensions,  $C^l$ , as  $O_{\text{DCE}}^l$ . Notice that no interpolation operations are required as the number of dimensions are now consistent between  $O_{\text{DCE}}^l$  and  $F_{\text{Proj},rs}^l$ . Subsequently, layer normalization is applied to both  $O_{\text{DCE}}^l$  and  $F_{\text{Proj},rs}^l$ . The above steps can be summarised as

$$F_{\text{Proj},rs}^l = \text{LN}(\text{RH}(F_{\text{Proj}}^l)), O_{\text{DCE}}^l = \text{LN}(O_{\text{DCE}}^l). \quad (5)$$

Next, we employ cross-attention to integrate relevant degradation-specific information into  $F_{\text{Proj},rs}^l$  and this is achieved using MHCA. For this purpose, we use  $F_{\text{Proj},rs}^l$  to calculate the query ( $Q$ ) and  $O_{\text{DCE}}^l$  for computing the key ( $K$ ) and value ( $V$ ) as follows

$$F_{\text{Proj},rs}^l \rightarrow Q, O_{\text{DCE}}^l \rightarrow K, V, \quad (6)$$

$$O_{\text{MHCA}}^l = \text{MHCA}(Q, K, V), O_{\text{MHCA}}^l \in \mathbb{R}^{H^l \cdot W^l \times C^l}. \quad (7)$$

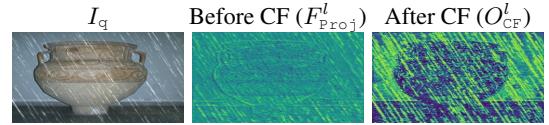


Figure 4: Comparison of activations of the restoration network prior to CF and after CF. Yellow-High, Blue-Low.

We select  $F_{\text{Proj},rs}^l$  as the query since we are looking to match the relevant DSI from  $O_{\text{DCE}}^l$  (key) to enhance the feature maps with the extracted context information.

The output of MHCA, is then reshaped back to  $\mathbb{R}^{C^l \times H^l \times W^l}$  and is projected using  $3 \times 3$  convolution ( $\text{Conv}_{3 \times 3}$ ). Again, GELU activation is applied to obtain the output of the CF block,  $O_{\text{CF}}^l$  (see Eqn. 8). More details about the workings of MHCA are provided in the supplementary document.

$$O_{\text{CF}}^l = \text{GELU}(\text{Conv}_{3 \times 3}(\text{RH}(O_{\text{MHCA}}^l))), O_{\text{CF}}^l \in \mathbb{R}^{C^l \times H^l \times W^l}. \quad (8)$$

Finally,  $O_{\text{CF}}^l$  is concatenated with  $F^l$  and propagated to the next decoder level. Fig. 4 captures the activations from the network for a rainy image ( $I_q$ ) prior to CF ( $F_{\text{Proj}}^l$ ) and after CF ( $O_{\text{CF}}^l$ ). Observe that prior to CF, not much attention is paid to degraded regions (rain streaks). However, after CF, the attention increases significantly on the rain streaks of  $I_q$ . Thus, the CF block effectively fuses the DSI ( $O_{\text{DCE}}^l$ ) into the features of the restoration network ( $F^l$ ).

The process of degradation-context extraction and context fusion is repeated at each level,  $l$ , of the decoder. This multi-scale fusion at each decoder level  $l$ , ensures that the context information is retained through the entire decoder, thereby enhancing the quality of image reconstruction.

## 4 Experimental Results

In this section, we explain our implementation, datasets used, results and ablation studies.

### 4.1 Implementation Details

Our method is trained using the AdamW optimizer with a cosine annealing Learning Rate (LR) scheduler. We train for a total of 100 epochs on 8 RTX A5000 GPUs with a batch size of 32, initial LR =  $2 \times 10^{-4}$ , weight decay = 0.01,  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$  and warm-up for 15 epochs. We use random crop size of  $128 \times 128$  pixels, and random flipping as data augmentations. The loss function used is the  $L_1$  loss. For extracting CLIP features, no augmentations are used and the images in the context pair are resized to  $224 \times 224$ . Our implementation utilized PyTorch (Paszke et al. 2019).

### 4.2 Datasets

**Training.** We use the Snow100k (Liu et al. 2018), synthetic rain (Zamir et al. 2021) datasets (SRD) and RESIDE (Li et al. 2019) to train our method for all-weather restoration. The training split of Snow100k contains 50,000 synthetic snow images along with the corresponding clean images. For deraining, we use the training split of SRD containing 13,711 clean-synthetic rainy image pairs. For dehazing,

Table 1: Quantitative comparisons of AWRaCLE with SOTA on the test sets described in Sec. 4.2. The values indicated are placeholders for PSNR/SSIM. Degradation type (S-Snow, R-Rain, H-Haze) is indicated within brackets. The best result is in bold, and second best is underlined.

Datasets	Methods							
	WeatherDiff (TPAMI'23)	WGWS (CVPR'23)	TSMC (CVPR'22)	AirNet (CVPR'22)	PromptIR (NeurIPS'23)	Painter (CVPR'23)	DiffUIR (CVPR'24)	AWRaCLE
SOTS Outdoor (H)	28.00/0.966	30.49/0.976	27.88/0.920	27.6/0.963	30.46/0.977	28.00/0.945	<u>30.97/0.977</u>	<b>31.65/0.981</b>
Rain100H (R)	<u>25.84/0.824</u>	13.91/0.410	26.48/0.822	22.96/0.692	26.32/0.821	22.50/0.792	<u>26.49/0.788</u>	<b>27.20/0.840</b>
Rain100L (R)	27.39/0.895	27.2/0.86	29.88/0.920	23.97/0.805	28.87/0.888	23.19/0.900	<u>31.83/0.932</u>	<b>35.71/0.966</b>
Snow100k (S)	31.33/0.910	32.58/0.921	32.28/0.916	29.22/0.884	<u>33.39/0.932</u>	27.87/0.871	31.77/0.915	<b>33.48/0.934</b>
Average	28.14/0.898	26.04/0.791	29.13/0.894	25.93/0.836	29.76/0.904	26.36/0.877	30.26/0.903	<b>32.01/0.930</b>

we use the Outdoor Training Set (OTS) of RESIDE which consists of 72,135 clean-synthetic hazy image pairs for training. We then split the training sets into two categories, each respectively consisting of heavy and light corruptions for better context extraction during training. More details about the splitting strategy can be found in the supplementary. In summary, we obtain 12,077 light rain images, 1,634 heavy rain images, 38,921 light haze images, 33,214 heavy haze images, 37,122 light snow images and 12,878 heavy snow images for training.

**Evaluation.** We evaluate all the methods for desnowing, de-raining and dehazing. For desnowing, we use the test split of Snow100k dataset containing 50,000 paired images. For de-raining, we evaluate the methods on Rain100H (Yang et al. 2017) for heavy rain and Rain100L (Yang et al. 2017) for light rain, each consisting of 100 paired images. For dehazing, we use RESIDE’s Synthetic Objective Testing Set (SOTS) outdoor containing 500 paired images.

### 4.3 Comparisons

We evaluate and compare the performance of AWRaCLE with seven recent AWIR approaches on the test sets described in Sec. 4.2. The methods we use for comparison are WeatherDiff (Özdenizci and Legenstein 2023), WGWS (Zhu et al. 2023), TSMC (Chen et al. 2022b), AirNet (Li et al. 2022), PromptIR (Potlapalli et al. 2024) and DiffUIR (Zheng et al. 2024). Additionally, we also compare with Painter (Wang et al. 2023a) and show that AWRaCLE uses context much more effectively. DiffUIR is a very recent SOTA method published in CVPR 2024. For a fair comparison, WeatherDiff, WGWS, TSMC, AirNet, PromptIR, DiffUIR and Painter are retrained on the training sets mentioned in Sec. 4.2. Some recent approaches such as (Patil et al. 2023; Zhang et al. 2024; Xu et al. 2024; Ai et al. 2024) have no training code, and each method is trained on different datasets. Although the authors of (Lin et al. 2024) have released training scripts, they have not provided adequate instructions for their use. Hence, we are unable to compare with these methods. We also do not compare with methods for single weather removal as our method is proposed specifically to deal with multiple degradations.

**Quantitative and Qualitative results.** We discuss the performance of all the methods on the test sets described in Sec. 4.2. Table 1 contains the Peak Signal to Noise Ra-

tio (PSNR) and Structural Similarity (SSIM) values for each method on these test sets. To evaluate AWRaCLE and Painter, we choose the context pair for each test set randomly from their respective training sets, i.e., for every test image, the context pair is chosen randomly. We resort to random selection for fairness. Moreover, the context pair is chosen from the training set, thus, requiring only the knowledge of the type of degradation during inference. From Table 1, we observe that AWRaCLE achieves excellent overall metrics. Our approach yields highest PSNR and SSIM values across all datasets. Importantly, AWRaCLE offers consistently high performance across all restoration tasks whereas competing methods perform well for some tasks but poorly for others. We also significantly outperform the in-context learning approach, Painter, highlighting the effectiveness of our in-context learning strategy. In Fig. 5, we show qualitative results for visual inspection and compare with the top-performing approaches TSMC, PromptIR and DiffUIR. It can be observed that AWRaCLE is able to handle the corruptions more effectively than the others. Additional qualitative results can be found in the supplementary along with results of AWRaCLE on real images and a discussion of the limitations of our method.

## 5 Ablation Studies

In this section, we first demonstrate the effect of the context pair provided to AWRaCLE and Painter. We show that AWRaCLE uses degradation-specific information (DSI) from the context pair to guide restoration while Painter fails to use any DSI from the context. We then show the importance of the various components of AWRaCLE.

### 5.1 Effect of Context Pairs

We first analyze the performance of our method and the in-context learning method, Painter, for correct and incorrect context on the Rain100L dataset. As shown in Table 1, when provided with correct context, AWRaCLE has a much higher PSNR (dB)/SSIM of 35.71/0.966 compared to Painter (23.19/0.900). Next, we provided incorrect context, i.e., the degradation in the query image  $I_q$  does not match the degradation present in  $I_d$  of the context pair. Providing incorrect context to AWRaCLE yields a PSNR (dB)/SSIM of 25.93/0.826 which is a  $\sim 10$  dB drop in performance with respect to correct context. However, Painter’s values with incorrect context (23.15/0.900) are nearly unchanged from

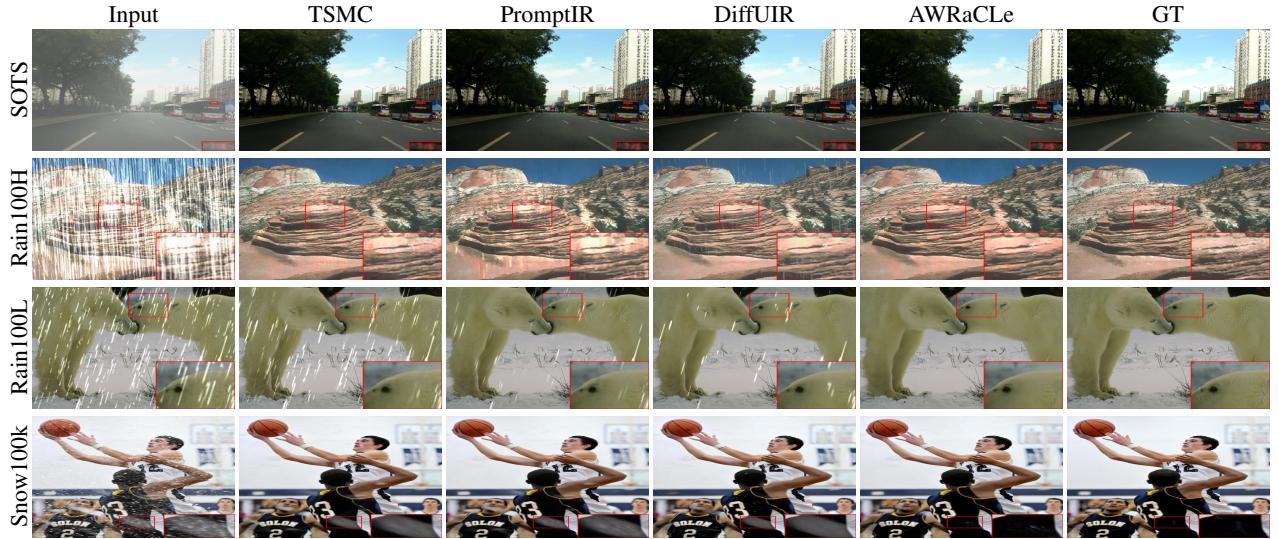


Figure 5: Qualitative comparisons of AWRaCLE with top performing approaches (TSMC, PromptIR and DiffUIR) on SOTS, Rain100L, Rain100H and Snow100k datasets. Zoomed-in patches are provided for examining fine details.

its performance with correct context, which indicates that it lacks utilization of the context for image restoration. Qualitative results for this experiment are in the supplementary.

Next, we analyze the impact of specific context pairs on the performance of AWRaCLE. In Table 2, we report mean ( $\mu$ )  $\pm$  standard deviation ( $\sigma$ ) of PSNR (row2) and SSIM (row3) obtained by randomly selecting 10 paired context images for each of deraining, dehazing and desnowing, and fixing each of these context pairs over the entire test set. This is different from the testing strategy used in our experiments, where the context pair is randomly chosen for each image of the test sets. The table shows that AWRaCLE is quite robust to different context pairs from the same degradation.

Table 2: Effect of fixing a specific context pair for the test sets for different degradations.

SOTS	Rain100L	Rain100H	Snow100k
$31.61 \pm 0.18$	$35.71 \pm 0.004$	$27.20 \pm 0.02$	$33.48 \pm 0.01$
$0.981 \pm 2e^{-4}$	$0.966 \pm 1e^{-4}$	$0.840 \pm 3e^{-4}$	$0.934 \pm 3e^{-4}$

## 5.2 Effect of Individual Components

In this section, we show the importance of the various components of AWRaCLE. Table 3 shows quantitative results for each of our ablations on the SOTS dataset. In the table, ‘‘Context’’ refers to training with either paired or unpaired context, ‘‘DCE’’ indicates if the DCE block is used, ‘‘CF’’ indicates usage of CF block and ‘‘MLF’’ refers to the incorporation of multi-level fusion. A ‘‘✓’’ in a column means that component is used, while a ‘‘-’’ means it is not used. Note that it is not possible to remove the DCE and CF blocks in their entirety. Thus, when the DCE block is removed, we remove the MHSA (.) operation, and when the CF block is removed, we replace the MHCA (.) operation with interpolation and element-wise multiplication with decoder features.

Table 3: Quantitative comparisons of different ablations conducted on AWRaCLE. All ablation settings are tested on the SOTS dataset. The best result is in bold.

Context	DCE	CF	MLF	PSNR/SSIM
-	-	-	-	29.53/0.972
Paired	-	✓	✓	30.53/0.978
Paired	✓	-	✓	31.16/0.979
Paired	✓	✓	-	30.12/0.977
Unpaired	✓	✓	✓	29.84/0.974
Paired	✓	✓	✓	<b>31.65/0.981</b>

The table shows that our proposed method, AWRaCLE (last row), demonstrates the best performance. Training with only the baseline (row 1) results in a significant drop in restoration performance. Without the DCE block (row2), there is an improvement in the model’s performance over the baseline. Omitting the CF blocks (row3) yields some performance gain but it is still quite inferior to our proposed method. The absence of MLF (row4) also leads to a drop in performance. Finally, training with unpaired context (row5) yields inferior results compared to training with paired context. Thus, these ablations highlight the importance of the components of AWRaCLE.

## 6 Conclusions

We proposed a novel approach called AWRaCLE for all-weather image restoration that leverages visual in-context learning. We showed that suitably designed degradation context extraction and fusion blocks are central to the performance of our method. Additionally, we presented multi-level fusion of context information which is key to achieving good restoration performance. AWRaCLE advances the state-of-the-art in AWIR on standard datasets for the tasks of deraining, desnowing and dehazing. We believe that our

method will be an important enabler for solving the complex AWIR task in its generality.

## 7 Acknowledgment

This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA R&D Contract No. 140D0423C0076. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

## References

- Ai, Y.; Huang, H.; Zhou, X.; Wang, J.; and He, R. 2024. Multimodal Prompt Perceiver: Empower Adaptiveness Generalizability and Fidelity for All-in-One Image Restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 25432–25444.
- Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35: 23716–23736.
- Ba, J. L.; Kiros, J. R.; and Hinton, G. E. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.
- Bai, Y.; Wang, C.; Xie, S.; Dong, C.; Yuan, C.; and Wang, Z. 2023. TextIR: A Simple Framework for Text-based Editable Image Restoration. *arXiv preprint arXiv:2302.14736*.
- Bar, A.; Gandelsman, Y.; Darrell, T.; Globerson, A.; and Efros, A. 2022. Visual prompting via image inpainting. *Advances in Neural Information Processing Systems*, 35: 25005–25017.
- Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models Are Few-Shot Learners. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781713829546.
- Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229. Springer.
- Chen, T.; Saxena, S.; Li, L.; Fleet, D. J.; and Hinton, G. 2021. Pix2seq: A language modeling framework for object detection. *arXiv preprint arXiv:2109.10852*.
- Chen, T.; Saxena, S.; Li, L.; Lin, T.-Y.; Fleet, D. J.; and Hinton, G. E. 2022a. A unified sequence interface for vision tasks. *Advances in Neural Information Processing Systems*, 35: 31333–31346.
- Chen, W.-T.; Fang, H.-Y.; Ding, J.-J.; Tsai, C.-C.; and Kuo, S.-Y. 2020. JSTASR: Joint Size and Transparency-Aware Snow Removal Algorithm Based on Modified Partial Convolution and Veiling Effect Removal. Berlin, Heidelberg: Springer-Verlag. ISBN 978-3-030-58588-4.
- Chen, W.-T.; Huang, Z.-K.; Tsai, C.-C.; Yang, H.-H.; Ding, J.-J.; and Kuo, S.-Y. 2022b. Learning Multiple Adverse Weather Removal via Two-stage Knowledge Learning and Multi-contrastive Regularization: Toward a Unified Model.
- Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- Gu, X.; Lin, T.-Y.; Kuo, W.; and Cui, Y. 2021. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*.
- He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9726–9735.
- He, K.; Sun, J.; and Tang, X. 2009. Single image haze removal using dark channel prior. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 1956–1963.
- Hendrycks, D.; and Gimpel, K. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.
- Kang, L.-W.; Lin, C.-W.; and Fu, Y.-H. 2012. Automatic Single-Image-Based Rain Streaks Removal via Image Decomposition. *IEEE Transactions on Image Processing*, 21(4): 1742–1755.
- Li, B.; Liu, X.; Hu, P.; Wu, Z.; Lv, J.; and Peng, X. 2022. All-In-One Image Restoration for Unknown Corruption. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 17431–17441.
- Li, B.; Ren, W.; Fu, D.; Tao, D.; Feng, D.; Zeng, W.; and Wang, Z. 2019. Benchmarking Single-Image Dehazing and Beyond. *IEEE Transactions on Image Processing*, 28(1): 492–505.
- Li, R.; Tan, R. T.; and Cheong, L.-F. 2020. All in One Bad Weather Removal Using Architectural Search. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3172–3182.
- Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; and Timofte, R. 2021. SwinIR: Image Restoration Using Swin Transformer. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 1833–1844.
- Lin, J.; Zhang, Z.; Wei, Y.; Ren, D.; Jiang, D.; Tian, Q.; and Zuo, W. 2024. Improving image restoration through removing degradations in textual representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2866–2878.
- Liu, Y.-F.; Jaw, D.-W.; Huang, S.-C.; and Hwang, J.-N. 2018. DesnowNet: Context-Aware Deep Network for Snow Removal. *IEEE Transactions on Image Processing*, 27(6): 3064–3073.

- Lu, J.; Clark, C.; Zellers, R.; Mottaghi, R.; and Kembhavi, A. 2022. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*.
- Özdenizci, O.; and Legenstein, R. 2023. Restoring vision in adverse weather conditions with patch-based denoising diffusion models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–12.
- Park, D.; Lee, B.; and Chun, S. 2023. All-in-One Image Restoration for Unknown Degradations Using Adaptive Discriminative Filters for Specific Degradations. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5815–5824.
- Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.
- Patil, P. W.; Gupta, S.; Rana, S.; Venkatesh, S.; and Murala, S. 2023. Multi-weather Image Restoration via Domain Translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 21696–21705.
- Potlapalli, V.; Zamir, S. W.; Khan, S. H.; and Shahbaz Khan, F. 2024. PromptIR: Prompting for All-in-One Image Restoration. *Advances in Neural Information Processing Systems*, 36.
- Quan, R.; Yu, X.; Liang, Y.; and Yang, Y. 2021. Removing Raindrops and Rain Streaks in One Go. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9143–9152.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.
- Roth, S.; and Black, M. 2005. Fields of Experts: a framework for learning image priors. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, 860–867 vol. 2.
- Shen, S.; Li, L. H.; Tan, H.; Bansal, M.; Rohrbach, A.; Chang, K.-W.; Yao, Z.; and Keutzer, K. 2021. How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*.
- Song, Y.; He, Z.; Qian, H.; and Du, X. 2023. Vision Transformers for Single Image Dehazing. *IEEE Transactions on Image Processing*, 32: 1927–1941.
- Valanarasu, J. J.; Yasarla, R.; and Patel, V. M. 2022. TransWeather: Transformer-based Restoration of Images Degraded by Adverse Weather Conditions. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2343–2353.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Wang, T.; Yang, X.; Xu, K.; Chen, S.; Zhang, Q.; and Lau, R. W. 2019. Spatial Attentive Single-Image Deraining With a High Quality Real Rain Dataset. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12262–12271.
- Wang, X.; Wang, W.; Cao, Y.; Shen, C.; and Huang, T. 2023a. Images speak in images: A generalist painter for in-context visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6830–6839.
- Wang, X.; Zhang, X.; Cao, Y.; Wang, W.; Shen, C.; and Huang, T. 2023b. SegGPT: Towards Segmenting Everything in Context. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1130–1140.
- Wang, Z.; Cun, X.; Bao, J.; Zhou, W.; Liu, J.; and Li, H. 2022. Uformer: A General U-Shaped Transformer for Image Restoration. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 17662–17672.
- Wei, W.; Meng, D.; Zhao, Q.; Xu, Z.; and Wu, Y. 2019. Semi-Supervised Transfer Learning for Image Rain Removal. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3872–3881.
- Wu, H.; Qu, Y.; Lin, S.; Zhou, J.; Qiao, R.; Zhang, Z.; Xie, Y.; and Ma, L. 2021. Contrastive Learning for Compact Single Image Dehazing. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10546–10555.
- Xu, X.; Kong, S.; Hu, T.; Liu, Z.; and Bao, H. 2024. Boosting Image Restoration via Priors from Pre-trained Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2900–2909.
- Yan, Q.; Jiang, A.; Chen, K.; Peng, L.; Yi, Q.; and Zhang, C. 2023. Textual Prompt Guided Image Restoration. *arXiv preprint arXiv:2312.06162*.
- Yang, W.; Tan, R. T.; Feng, J.; Liu, J.; Guo, Z.; and Yan, S. 2017. Deep Joint Rain Detection and Removal from a Single Image. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1685–1694.
- Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; and Yang, M.-H. 2022. Restormer: Efficient Transformer for High-Resolution Image Restoration. In *CVPR*.
- Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; Yang, M.-H.; and Shao, L. 2021. Multi-Stage Progressive Image Restoration. In *CVPR*.
- Zhang, H.; and Patel, V. M. 2018. Densely Connected Pyramid Dehazing Network. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3194–3203.
- Zhang, H.; Sindagi, V.; and Patel, V. M. 2020. Joint Transmission Map Estimation and Dehazing Using Deep Networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(7): 1975–1986.
- Zhang, K.; Li, R.; Yu, Y.; Luo, W.; and Li, C. 2021. Deep Dense Multi-Scale Network for Snow Removal Using Semantic and Depth Priors. *IEEE Transactions on Image Processing*, 30: 7419–7431.
- Zhang, Q.; Liu, X.; Li, W.; Chen, H.; Liu, J.; Hu, J.; Xiong, Z.; Yuan, C.; and Wang, Y. 2024. Distilling Semantic Priors

from SAM to Efficient Image Restoration Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 25409–25419.

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

Zhang, R.; Zhang, W.; Fang, R.; Gao, P.; Li, K.; Dai, J.; Qiao, Y.; and Li, H. 2022. Tip-adapter: Training-free adaptation of clip for few-shot classification. In *European Conference on Computer Vision*, 493–510. Springer.

Zheng, D.; Wu, X.-M.; Yang, S.; Zhang, J.; Hu, J.-F.; and Zheng, W.-s. 2024. Selective Hourglass Mapping for Universal Image Restoration Based on Diffusion Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Zhu, Y.; Wang, T.; Fu, X.; Yang, X.; Guo, X.; Dai, J.; Qiao, Y.; and Hu, X. 2023. Learning Weather-General and Weather-Specific Features for Image Restoration Under Multiple Adverse Weather Conditions. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 21747–21758.