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 ***A**ll-**W**eather Image **R**esto**r**a**t**ion using Visual In-**C**ontext **Le**arning* (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.

Project Page — https://sudraj2002.github.io/awraclepage/

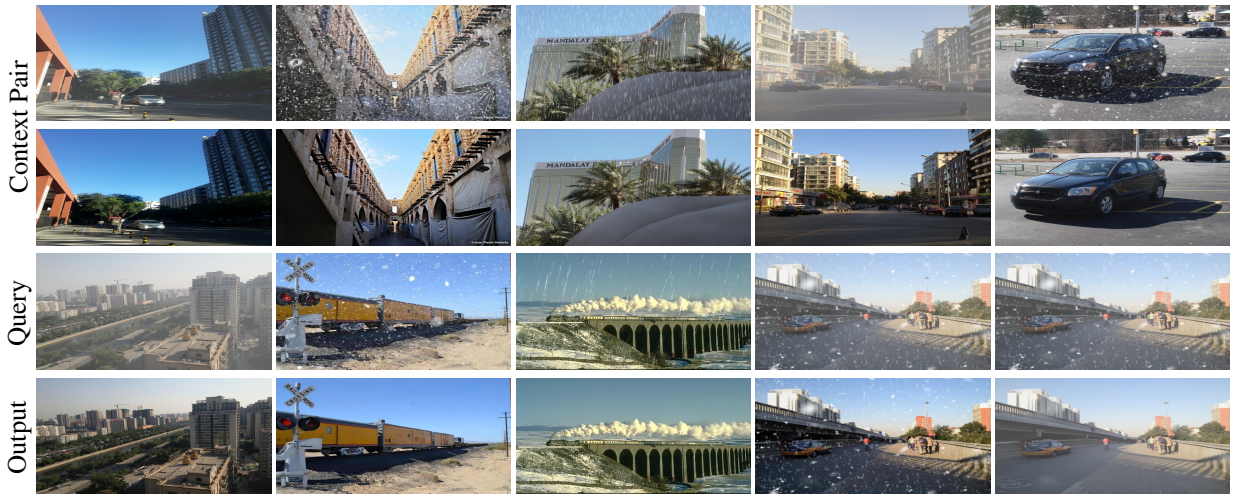
1 Introduction

Unfavorable weather conditions, such as rain, snow and haze, significantly degrade the performance of autonomous navigation, surveillance, and aerial imaging systems. Thus, there is a need for frameworks that mitigate weather-induced corruptions while preserving the underlying image semantics. Intial physics-based methods (He, Sun, and Tang 2009; Roth and Black 2005; Kang, Lin, and Fu 2012) struggled to handle real-world variability in degradations. Deep learning approaches that handle a single degradation (Song et al. 2023; Chen et al. 2020; Wei et al. 2019; Liang et al. 2021; Zamir et al. 2022, 2021; Wang et al. 2022) at a time were then proposed, but they must be retrained or fine-tuned for each new condition, reducing their practicality.

To address these issues, recent work has focused on All-Weather Image Restoration (AWIR) networks that handle multiple degradations with a single model (Li, Tan, and Cheong 2020; Valanarasu, Yasarla, and Patel 2022; Li et al. 2022; Park, Lee, and Chun 2023; Özdenizci and Legenstein 2023; Chen et al. 2022b; Potlapalli et al. 2024; Zheng et al. 2024). However, these approaches learn degradation representations only from individual images and lack guidance that provides detailed degradation-specific information (DSI). This limits their ability to effectively learn features unique to different degradations, impeding their performance. Thus, there is a need for a framework capable of learning robust degradation-specific features that can facilitate effective restoration. This is challenging without supplementary knowledge about the nature of the corruption. While some methods (Yan et al. 2023; Bai et al. 2023) have introduced text-based guidance for AWIR, 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.

To address this limitation, recent approaches such as Diff-Plugin (Liu et al. 2024) and DA-CLIP (Luo et al. 2023) attempt to extract DSI directly from images. Diff-Plugin extracts task-specific and spatial prompts from the degraded input image. However, it requires multiple independently trained task-specific plugins for each degradation. DA-CLIP extracts text-aligned DSI from the degraded image but lacks detailed visual information, as text conveys only high-level features. Moreover, both methods face the challenge of disentangling scene characteristics from DSI because they rely on extracting DSI from a single image. We conjecture that feeding both the clean and degraded images as context to an image restoration network can overcome these limitations as the consistent scene between the pair allows the network to aggregate visual information specific to the degradation. 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 Processing (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), SegGPT (Wang et al. 2023b) and PromptGIP (Liu et al. 2023).



(a) Dehaze (b) Desnow (c) Derain (d) Selective Dehaze (e) Selective Desnow

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.

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 (MIM) to extract context from the unmasked regions of an image 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, hindering its restoration performance. We believe this is due to two primary reasons. Firstly, the extraction of context relies solely on the MIM framework, which lacks constraints to ensure retrieval of degradation characteristics from the context images. Secondly, providing context only at the encoder's input can suppress contextual information after the initial network layers, making it negligible at the decoder. PromptGIP also suffers from the above issues, leading to inferior performance.

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 DSI which will in-turn hamper restoration performance. To facilitate restoration, the degradation type in the context pair should align with that of the query image. During test time, the context pair for a degradation can be chosen from the respective training set, thus requiring only the knowledge of the degradation type.

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.
3. 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. 2021a; 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. These AWIR methods do not utilize contextual guidance which limits their performance. Diff-Plugin (Liu et al. 2024) and DA-CLIP (Luo et al. 2023) attempt to extract degradation-specific information (DSI) directly from images to improve performance. However, this approach makes it challenging to disentangle scene characteristics from DSI, unlike our method.

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. PromptGIP (Liu et al. 2023) proposed a visual prompting question-answering framework for extracting context. However, these approaches rely heavily on masked-image modelling 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 $\mathcal{C} = \{I_{\mathrm{d}}, I_{\mathrm{c}}\}$ 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, $I_{\mathrm{d}}, I_{\mathrm{c}}, I_{\mathrm{q}} \in \mathbb{R}^{H \times W \times 3}$ 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 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

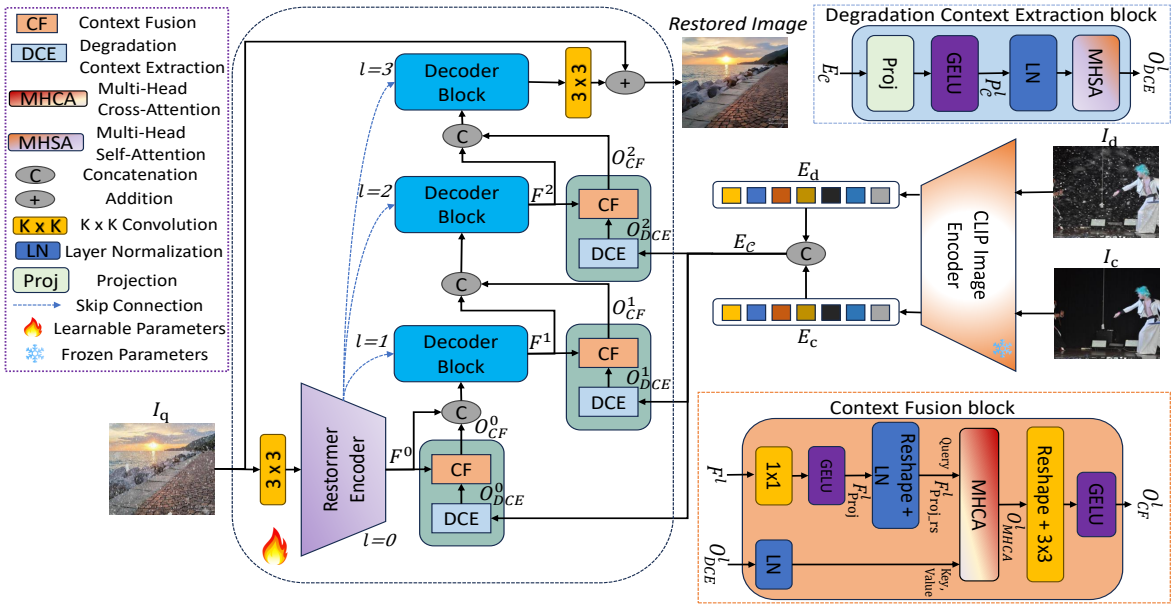


Figure 2: Block diagram of the proposed visual in-context learning approach for AWIR. CLIP features are extracted from I_{d} and I_{c} which are subsequently fed to DCE blocks at different decoder levels, l. CF blocks then fuse the degradation information obtained from the DCE blocks with decoder features,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

(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 $E_{\mathrm{d}}, E_{\mathrm{c}}$ for the context pair $\mathcal{C}$ from the final transformer block of CLIP's image encoder (denoted by $\texttt{CLIP}(.)$). 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.

$$E_{\mathrm{d}} = \texttt{CLIP}(I_{\mathrm{d}}), E_{\mathrm{c}} = \texttt{CLIP}(I_{\mathrm{c}}), \{E_{\mathrm{d}}, E_{\mathrm{c}}\} \in \mathbb{R}^{L \times D} \quad (1)$$

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 $P^l_{\mathcal{C}} \in \mathbb{R}^{2L \times C^l}$, where C^l is the projection dimension, and these steps are summarised as below.

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

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

$$O^l_{\mathrm{DCE}} = \texttt{MHSA}(\texttt{LN}(P^l_{\mathcal{C}})), O^l_{\mathrm{DCE}} \in \mathbb{R}^{2L \times C^l}. \quad (3)$$

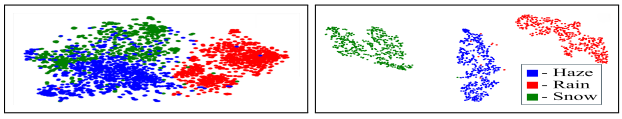
Since the scene is consistent in both I_{d} and I_{c}, the primary distinction between them is the degradation which is adeptly discerned through MHSA. This enables extraction of the necessary degradation-context from $\mathcal{C}$ for judiciously guiding the network towards the objective of all-weather 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-specifc information, we overlay the output of the DCE block (O^l_{DCE}) for a clean-hazy and a clean-snowy context pair, respectively. This is illustrated in Fig. 3a where activations are overlayed on the degraded image (I_{d}) and its corresponding clean image (I_{c}). $A_{\mathrm{d}} \in \mathbb{R}^{L \times C^l}$ and $A_{\mathrm{c}} \in \mathbb{R}^{L \times C^l}$ represent the DCE block activations obtained by splitting O^l_{DCE} 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 t-SNE plots of the CLIP embeddings, $E_{\mathcal{C}}$, and DCE block outputs (O^l_{DCE}) for hazy, snowy and rainy context pairs. Although there is separation in the t-SNE plot with $E_{\mathcal{C}}$, it is significantly enhanced after using the DCE block. Thus, the DCE block extracts DSI that 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^l_{DCE}, needs to be fused with the corresponding decoder features, $F^l \in \mathbb{R}^{K^l \times H^l \times W^l}$, from Restormer. Here H^l and W^l are the spatial resolution of the feature map, and K^l is chan-

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



(b) t-SNE plot of CLIP embeddings (E_C, right) and DCE block outputs ($O^l_{\texttt{DCE}}$, left), for hazy, rainy and snowy context pairs. Separation signficantly improves after using the DCE block.

Figure 3: Analysis of DCE block outputs.

nel dimension. Fusion is achieved with the help of the Context Fusion (CF) blocks that utilize Multi-Head Cross Attention ($\texttt{MHCA}(.)$) (Vaswani et al. 2017) to integrate information from $O^l_{\texttt{DCE}}$ 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 $O^l_{\texttt{DCE}}$. We achieve this by treating F^l as the query and matching it with the key and value computed from $O^l_{\texttt{DCE}}$.

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 $O^l_{\texttt{DCE}}$ using 1×1 convolution as

$$F^l_{\texttt{Proj}} = \texttt{GELU}(\texttt{Conv}_{1\times1}(F^l)), F^l_{\texttt{Proj}} \in \mathbb{R}^{C^l \times H^l \times W^l}, \quad (4)$$

where we have used the GELU activation function as non-linearity. We observe that $F^l_{\texttt{Proj}}$ and $O^l_{\texttt{DCE}}$ have a mismatch in the number of dimensions ($O^l_{\texttt{DCE}}$ is 2-D but $F^l_{\texttt{Proj}}$ 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 $O^l_{\texttt{DCE}}$ into the same dimension as $F^l_{\texttt{Proj}}$. However, interpolation causes redundancy in the degradation-context thereby hindering the performance of cross-attention. To circumvent this problem, we reshape (denoted as $\texttt{RH}(.)$) $F^l_{\texttt{Proj}}$ to obtain $F^l_{\texttt{Proj\_rs}} \in \mathbb{R}^{H^l \cdot W^l \times C^l}$ which has the same channel dimensions, C^l, as $O^l_{\texttt{DCE}}$. Notice that no interpolation operations are required as the number of dimensions are now consistent between $O^l_{\texttt{DCE}}$ and $F^l_{\texttt{Proj\_rs}}$. Subsequently, layer normalization is applied to both $O^l_{\texttt{DCE}}$ and $F^l_{\texttt{Proj\_rs}}$. The above steps can be summarised as

$$F^l_{\texttt{Proj\_rs}} = \texttt{LN}(\texttt{RH}(F^l_{\texttt{Proj}})), O^l_{\texttt{DCE}} = \texttt{LN}(O^l_{\texttt{DCE}}). \quad (5)$$

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

$$F^l_{\texttt{Proj\_rs}} \rightarrow \texttt{Q}, O^l_{\texttt{DCE}} \rightarrow \texttt{K}, \texttt{V}, \quad (6)$$

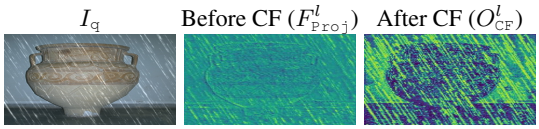


Figure 4: Comparison of activations of the restoration network prior to CF and after CF. Yellow-High, Blue-Low.

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

We select $F^l_{\texttt{Proj\_rs}}$ as the query since we are looking to match the relevant DSI from $O^l_{\texttt{DCE}}$ (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 3×3 convolution ($\texttt{Conv}_{3\times3}$). Again, GELU activation is applied to obtain the output of the CF block, $O^l_{\texttt{CF}}$ (see Eqn. 8). More details about the workings of MHCA are provided in the supplementary.

$$O^l_{\texttt{CF}} = \texttt{GELU}(\texttt{Conv}_{3\times3}(\texttt{RH}(O^l_{\texttt{MHCA}}))), O^l_{\texttt{CF}} \in \mathbb{R}^{C^l \times H^l \times W^l}. \quad (8)$$

Finally, $O^l_{\texttt{CF}}$ 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 ($F^l_{\texttt{Proj}}$) and after CF ($O^l_{\texttt{CF}}$). 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 ($O^l_{\texttt{DCE}}$) 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 128×128 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 224×224. 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

Datasets	WeatherDiff TPAMI'23	WGWS CVPR'23	TSMC CVPR'22	AirNet CVPR'22	PromptIR NeurIPS'23	Painter CVPR'23	DA-CLIP ICLR'24	DiffUIR CVPR'24	DiffPlugin CVPR'24	PromptGIP ICML'24	AWRaCLe -
SOTS	28.0/0.966	30.5/0.976	27.9/0.920	27.6/0.963	30.5/0.977	28.0/0.945	26.9/0.958	31.0/0.977	23.6/0.778	17.9/0.672	**31.7/0.981**
Rain100H	25.8/0.824	13.9/0.410	26.5/0.822	23.0/0.692	26.3/0.821	22.5/0.792	23.4/0.730	26.5/0.788	16.1/0.527	18.0/0.482	**27.2/0.840**
Rain100L	27.4/0.895	27.2/0.860	29.9/0.920	24.0/0.805	28.9/0.888	23.2/0.900	30.1/0.918	31.8/0.932	25.4/0.698	22.8/0.662	**35.7/0.966**
Snow100k	31.3/0.910	32.6/0.921	32.3/0.916	29.2/0.884	33.4/0.932	27.9/0.871	30.6/0.893	31.8/0.915	23.5/0.658	20.8/0.615	**33.5/0.934**
Average	28.1/0.898	26.0/0.791	29.1/0.894	25.9/0.836	29.8/0.904	26.4/0.877	27.8/0.874	30.3/0.903	22.2/0.665	19.9/0.608	**32.0/0.930**

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. The best result is in **bold**, and second best is underlined.

$13,711$ clean-synthetic rainy image pairs. For dehazing, 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, deraining and dehazing. For desnowing, we use the test split of Snow100k dataset containing $50,000$ paired images. For deraining, 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 ten 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), DiffUIR (Zheng et al. 2024), DiffPlugin (Liu et al. 2024) and DA-CLIP (Luo et al. 2023). Additionally, we also compare with Painter (Wang et al. 2023a) and PromptGIP (Liu et al. 2023), and show that AWRaCLe uses context much more effectively. For a fair comparison, all methods 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. 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 Ratio (PSNR) and Structural Similarity (SSIM) values for each method on these test sets. To evaluate AWRaCLe, Painter and PromptGIP, 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 approaches, Painter and PromptGIP, highlighting the effectiveness of our in-context learning strategy. Additionally, we provide quantitative comparisons with LPIPS and FID scores in the supplementary.

SOTS	Rain100L	Rain100H	Snow100k
31.61 ± 0.18	$35.71 \pm 4{\cdot}10^{-3}$	27.2 ± 0.02	33.48 ± 0.01
$0.981 \pm 2{\cdot}10^{-4}$	0.966 ± 10^{-4}	$0.84 \pm 3{\cdot}10^{-4}$	$0.934 \pm 3{\cdot}10^{-4}$

Table 2: Effect of fixing a specific context pair for the test sets described in Sec. 4.2 for different degradations.

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. More qualitative results, performance of AWRaCLE on real images along with a user study, and a discussion of the limitations of our method are provided in the supplementary.

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,

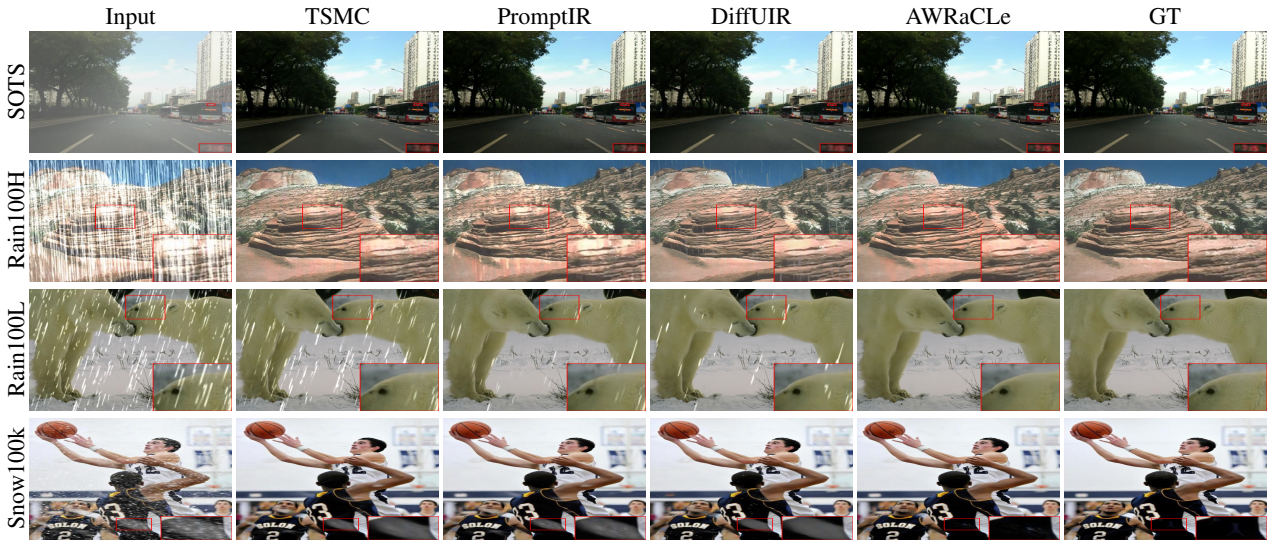


Figure 5: Qualitative comparisons of AWRaCLe with top performing approaches (TSMC, PromptIR and DiffUIR) on SOTS, Rain100L, Rain100H and Snow100k datasets. Zoomed-in patches are provided for examining fine details.

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 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 (μ) $\pm$ standard deviation (σ) 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.

Finally, we tested our model's robustness to out-of-distribution (OoD) context pairs by sampling them from the following datasets unseen during training: Foggy Cityscapes (Hahner et al. 2019) for dehazing, rain images from RainDS (Quan et al. 2021) for deraining, and SnowCityscapes (Zhang et al. 2021b) for desnowing. Our model obtained PSNR/SSIM of $31.02/0.976$ on SOTS Outdoor, $35.72/0.966$ on Rain100L, $27.19/0.840$ on Rain100H and $33.40/0.934$ on Snow100k datasets. These results show only minimal deviations from those reported in Table 1, showcasing our model's resilience to out-of-distribution context.

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. The table shows that our proposed method, AWRaCLe (last row), demonstrates the best performance.

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	✓	✓	✓	**31.65/0.981**

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.

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.

Acknowledgments

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*.

Hahner, M.; Dai, D.; Sakaridis, C.; Zaech, J.-N.; and Gool, L. V. 2019. Semantic Understanding of Foggy Scenes with Purely Synthetic Data. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, 3675–3681.

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.

Liu, Y.; Chen, X.; Ma, X.; Wang, X.; Zhou, J.; Qiao, Y.; and Dong, C. 2023. Unifying image processing as visual prompting question answering. *arXiv preprint arXiv:2310.10513*.

Liu, Y.; Ke, Z.; Liu, F.; Zhao, N.; and Lau, R. W. 2024. Diff-Plugin: Revitalizing Details for Diffusion-based Low-level Tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4197–4208.

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*.

Luo, Z.; Gustafsson, F. K.; Zhao, Z.; Sjölund, J.; and Schön, T. B. 2023. Controlling vision-language models for universal image restoration. *arXiv preprint arXiv:2310.01018*, 3(8).

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9147–9156.

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.

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. 2021a. Deep Dense Multi-Scale Network for Snow Removal Using Semantic and Depth Priors. *IEEE Transactions on Image Processing*, 30: 7419–7431.

Zhang, K.; Li, R.; Yu, Y.; Luo, W.; and Li, C. 2021b. 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 adaption 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.