MWFormer: Multi-Weather Image Restoration Using Degradation-Aware Transformers

Ruoxi Zhu, *Graduate Student Member, IEEE*, Zhengzhong Tu, *Member, IEEE*, Jiaming Liu, *Graduate Student Member, IEEE*, Alan C. Bovik, *Life Fellow, IEEE*, and Yibo Fan, *Member, IEEE*

***Abstract*— Restoring images captured under adverse weather conditions is a fundamental task for many computer vision applications. However, most existing weather restoration approaches are only capable of handling a specific type of degradation, which is often insufficient in real-world scenarios, such as rainy-snowy or rainy-hazy weather. Towards being able to address these situations, we propose a multi-weather Transformer, or MWFormer for short, which is a holistic vision Transformer that aims to solve multiple weather-induced degradations using a single, unified architecture. MWFormer uses hyper-networks and feature-wise linear modulation blocks to restore images degraded by various weather types using the *same* set of learned parameters. We first employ contrastive learning to train an auxiliary network that extracts content-independent, distortion-aware feature embeddings that efficiently represent predicted weather types, of which more than one may occur. Guided by these weather-informed predictions, the image restoration Transformer adaptively modulates its parameters to conduct both local and global feature processing, in response to multiple possible weather. Moreover, MWFormer allows for a novel way of tuning, during application, to either a single type of weather restoration or to hybrid weather restoration without any retraining, offering greater controllability than existing methods. Our experimental results on multi-weather restoration benchmarks show that MWFormer achieves significant performance improvements compared to existing state-of-the-art methods, without requiring much computational cost. Moreover, we demonstrate that our methodology of using hyper-networks can be integrated into various network architectures to further boost their performance. The code is available at: https://github.com/taco-group/MWFormer.**

***Index Terms*— Image restoration, adverse weather, multi-task learning, low-level vision, transformer.**

Received 2 December 2023; revised 9 June 2024 and 14 October 2024; accepted 8 November 2024. Date of publication 25 November 2024; date of current version 27 December 2024. The work of Ruoxi Zhu, Jiaming Liu, and Yibo Fan was supported in part by the National Key Research and Development Program of China under Grant 2023YFB4502802, in part by the National Natural Science Foundation of China under Grant 62031009, in part by Fudan-ZTE Joint Laboratory, in part by Alibaba Innovative Research (AIR) Program, and in part by Alibaba Research Fellow (ARF) Program. The associate editor coordinating the review of this article and approving it for publication was Dr. Nam Ik Cho. *(Zhengzhong Tu and Yibo Fan contributed equally to this work.) (Corresponding author: Yibo Fan.)*

Ruoxi Zhu, Jiaming Liu, and Yibo Fan are with the State Key Laboratory of Integrated Chips and Systems, Fudan University, Shanghai 200433, China (e-mail: rxzhu22@m.fudan.edu.cn; liujm22@m.fudan.edu.cn; fanyibo@fudan.edu.cn).

Zhengzhong Tu was with the Department of Electrical and Computer Engineering, University of Texas at Austin, Austin, TX 78712 USA. He is now with the Department of Computer Science and Engineering, Texas A&M University, College Station, TX 77840 USA (e-mail: tzz@tamu.edu).

Alan C. Bovik is with the Laboratory for Image and Video Engineering (LIVE), The University of Texas at Austin, Austin, TX 78712 USA (e-mail: bovik@ece.utexas.edu).

Digital Object Identifier 10.1109/TIP.2024.3501855

I. Introduction

IMAGES captured in the real world are often of defective quality due to adverse capture or environmental conditions. For example, CMOS-based cameras typical in mobile devices often struggle to produce high-quality pictures in low light. The photos produced under such conditions can be noisy, blurry, and under-exposed. Other common occurrences of degradation are caused by possibly multiple coincident weather conditions, such as rain, fog, and snow, that affect human-perceived image quality. When the images are fed to automated vision systems, these distortions can severely hamper the performances of computer vision algorithms, which are often trained on datasets of pictures taken under normal weather conditions. Failing to account for and ameliorate the effects of these and other natural phenomena can often lead to catastrophic outcomes in vision-dependent applications like autonomous driving, robotics, security, and surveillance, etc.

Developing image processing algorithms that are able to analyze and subsequently restore weather-degraded pictures is an active research topic [2], [3], [4]. In recent years, deep learning-based restoration methods have been widely utilized to conduct weather-related image restoration tasks, such as deraining [3], [5], snow-removal [6], [7], [8], and dehazing [4], [9], [10]. Although these methods delivered promising results, each is designed to handle only a single type of adverse weather condition. Whereas in many real-world scenarios, the weather conditions are generally unknown to the restoration algorithm. Moreover, there are often multiple commingled conditions, which result in multiply-distorted pictures that the above-mentioned methods are unable to adequately improve.

Recently, several unified solutions have been proposed to restore images impaired by multiple coincident weather-induced degradations [1], [11], [12], [13]. For example, the authors of [1], [2], [12] train single networks on combined datasets, each representative of a single weather condition, with the expectation that the models would learn to adaptively process each weather degradation. However, these methods often deliver unsatisfying and unbalanced generalization performances across different weather types, and are unable to handle artifacts from co-occurring weather conditions. An important reason for this is that multiple coinciding distortions mutually interact, creating new and highly diverse distortions.

Towards making further progress on this important problem, we propose an efficient, degradation-aware **M**ulti-**W**eather

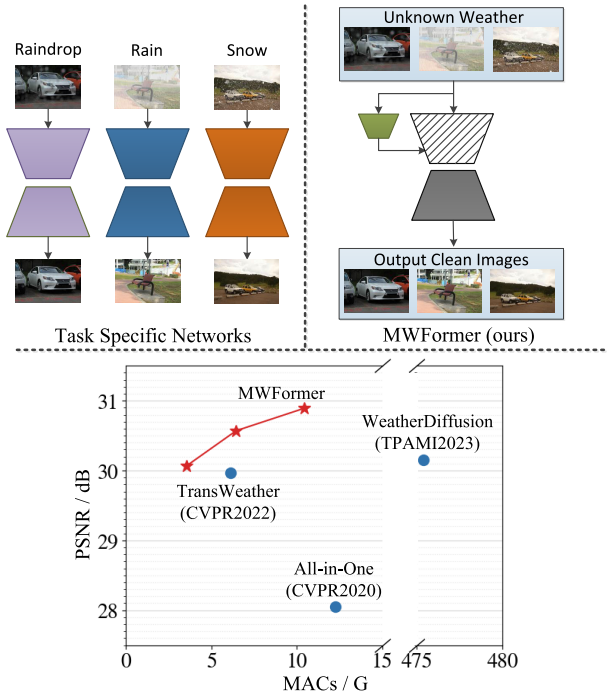


Fig. 1. Top row: Comparison of the MWFormer architecture with those of existing task-specific networks. Bottom: Restoration performance and computational cost of three versions of MWFormer having different numbers of channels against three competitive multi-weather restoration models. MWFormer achieves generally better performance with **100**× less computation than the SOTA model WeatherDiffusion [1].

Trans**Former** which we call **MWFormer**, that uses the architecture shown in Fig. 2. MWFormer is designed to provide a strong restoration backbone for conducting image restoration tasks in the presence of unknown adverse weather conditions. MWFormer is able to account for different weather-induced degradation types using a small auxiliary hyper-network that extracts degradation-informed features from an input image. These features guide the generation of the parameters of the image restoration backbone, allowing it to adaptively process the picture conditioned on the predicted weather degradation. We also show that the new hypernet-based multi-weather feature extractor enables a novel way of test-time tuning to either handle a fixed weather condition with less computation, or to handle combined, hybrid weather-induced degradations, without any retraining. This offers greater flexibility and controllability than existing multi-task methods. Notably, the proposed model is the first one capable of handling hybrid-weather degradations that were unseen during training. Some extended applications of the hyper-network have also been developed, such as identifying the adverse weather type, and guiding the pre-trained weather-specific image restoration models, which shows its versatility. Experimental results on benchmark datasets show that MWFormer is able to significantly outperform previous state-of-the-art (SOTA) models, both quantitatively and qualitatively, on a multi-weather restoration benchmark. Our methodology can also be integrated into various other network architectures to boost their performance in multi-weather restoration. To sum up, our contributions are summarized as follows:

- We introduce a novel Transformer-based architecture called MWFormer for multi-weather restoration, which can restore pictures distorted by multiple adverse weather degradations using a single, unified model.
- A hyper-network is employed to extract content-independent weather-aware features that are used to dynamically modify the parameters of the restoration backbone, allowing for degradation-dependent restoration and other related applications.
- The feature vector produced by the hyper-network is leveraged to guide the restoration backbone's behavior across all dimensions and scales (i.e., locally spatial, globally spatial, and channel-wise modulations).
- Two variants of MWFormer are created—one for lower computational cost, and the other for addressing hybrid adverse weather degradations unseen during training.
- Comprehensive experiments and ablation studies demonstrate the efficacy of the proposed blocks and the superiority of MWFormer in terms of visual and quantitative metrics. We also develop and analyze multi-weather restoration models in the context of downstream tasks.

II. Related Work

A. Image Restoration

Image restoration is a long-standing computer vision problem that aims to reconstruct a high-quality image from a degraded input. Recently, there has been a trend of employing end-to-end training of large neural networks on large-scale paired image datasets for a broad range of tasks, such as denoising [14], [15], deblurring [16], [17], [18], super-resolution [19], [20], low-light enhancement [21], [22], [23], dehazing [9], [10], deraining [12], [24], etc. The impressive advancements on these problems have been mainly driven by the development of novel network architectures. For example, encoder-decoder architectures have been widely adopted for a wide variety of restoration tasks [16], [17], [18], [22], [23], largely because of the efficacy of multi-scale feature learning. Similarly, the spatial and channel self-attention mechanisms have been used to learn spatially focused and sparser features [3], [25]. More recently, multi-stage progressive networks [26], [27], [28] have been deployed on more challenging tasks like deblurring and deraining, achieving impressive performances.

B. Image Deraining

Rain can significantly degrade the quality of captured pictures. Extensive research efforts have aimed to mitigate the adverse effects of rain on images. Restoring "rainy" images involves two sub-tasks: eliminating rain streaks and removing raindrops. For instance, Li et al. [5] leveraged a combination of dilated convolutional neural networks and recurrent neural networks to effectively expunge rain streaks from pictures. Yasarla et al. [29] utilized a Gaussian Process-based semi-supervised learning framework, demonstrating impressive generalization capabilities on real-world images. Ba et al. [30] proposed a novel deraining network trained on a new and comprehensive dataset of real-world rainy images. Beyond merely addressing rain streaks, there's an increasing emphasis on tackling the challenges posed by raindrops. Qian et al. [31] introduced a dataset specifically designed to capture

raindrop-related artifacts. They also trained an attentive GAN to effectively remove raindrops. Quan et al. [32] developed a cascaded network designed to simultaneously remove both raindrops and rain streaks. More recently, Xiao et al. [24] developed a Transformer architecture to conduct joint raindrop and rain streak removal, obtaining promising visual results.

C. Image Desnowing

Snow is a complex atmospheric phenomenon that plagues the performance of computer vision models, such as the object detectors used in autonomous vehicles. DesnowNet [7] pioneered the use of deep learning to conduct single-image desnowing, and the authors also built the first "snowy" picture dataset, called Snow-100K. Building on this foundation, Chen et al. [33] addressed the *veiling effect*—a phenomenon whereby snowflakes obscure and diminish picture clarity, by proposing a size- and transparency-aware snow removal algorithm. Recently, Lin et al. [34] designed a lightweight Laplace Mask Query Transformer for snow removal, achieving SOTA performance.

D. Multi-Weather Restoration

The existence of many different weather types in the real world poses a significant challenge to single weather restoration models, leading to growing interest in developing image restoration models that can effectively restore images affected by various complex weather conditions within a single, unified framework. Chen et al. [35] leveraged a two-stage knowledge learning mechanism to handle three different types of weather with a unified network. Li et al. [12] designed an architecture called All-in-One, equipped with multiple encoders to capture different degradations and a single decoder. While this approach is promising, its significant computational overhead poses challenges for real-world applicability. Valanarasu et al. [2] unveiled a more efficient Transformer-based architecture called TransWeather by incorporating intra-patch Transformer blocks (intra-PT blocks) and using learnable weather-type queries. The intra-PT blocks share the same architecture as the vanilla Transformer blocks, but take smaller patch embeddings as input, which are sub-patches yielded from the original patch embeddings. These smaller sub-patches facilitate the network to extract finer details which are beneficial for mitigating smaller degradations. Ozdenizci et al. [1] employed denoising diffusion models to conduct multi-weather image restoration, setting new benchmarks in performance. Yet, this approach suffers from extremely slow inference time, making it unsuitable for real-time deployments. Also, the model's design overlooks specific treatments regarding the characteristics of various weather types. Zhu et al. [36] proposed a more explainable method to extract the weather-general and weather-dependent features for multi-weather restoration. Besides, in addition to image restoration models, some researchers [37] have also proposed image segmentation models that can handle different real weather types.

E. Transformers for Image Restoration

Building on foundational works [38], [39], Transformer architectures have become popular for various computer vision tasks including image restoration, often significantly surpassing the previous CNN-based solutions. The Image Processing Transformer (IPT) [40] was the first to employ a pure Transformer architecture for image processing tasks, which was pre-trained on a large number of corrupted image pairs using contrastive learning. The pre-trained IPT could efficiently adapt to many image processing tasks after fine-tuning, outperforming state-of-the-art methods. The SwinIR [41] architecture, based upon the Swin Transformer [39], effectively handled low-level vision tasks by leveraging local-attention models. The Restormer [42] architecture deployed a novel Transformer variant able to capture long-range pixel interactions while remaining efficient using a transposed attention mechanism. Furthermore, the Uformer [43] presented a U-shaped Transformer architecture with locally enhanced windows that has been shown to perform remarkably well across diverse image restoration tasks.

III. Proposed Method

Here, we explain the technical details of the proposed MWFormer multi-weather restoration model. Our primary objective is to learn a single, unified model capable of handling multiple different weather degradations with the same set of learned parameters. This is similar to the challenge of real-world image denoising, where an algorithm is expected to deal with various noise sources, types, and levels. Denoising with prior knowledge of the noise characteristics generally outperforms denoising without prior knowledge, since additional noise information helps a denoising network to better learn to adapt its parameters. Thus, adding an extra noise estimation module can enhance the performance of the denoising network and increase its flexibility. Drawing inspiration from this, we propose to deem different weather types as analogous to varying noise sources or types. Features descriptive of weather type can be extracted beforehand, then fed to the main restoration network, which gains degradation adaptivity conditioned on the input weather types. Consequently, our proposed algorithm may be bifurcated into two phases: weather-feature extraction (by the hyper-network) followed by a weather-type-informed image restoration process.

A. Overall Architecture

An overall schematic diagram of MWFormer is illustrated in Fig. 2, showing the two major components: (i) a restoration backbone containing encoder and decoder blocks, which are responsible for recovering a high-quality image from the degraded input; (ii) a feature extraction network that yields weather-aware feature vectors. We adopt a Transformer-based architecture as the restoration backbone. Besides the vanilla Transformer blocks, our encoder network contains extra intra-PT blocks introduced in Sec. II. The decoder of the backbone is similar to the design in [2], including learnable weather-type queries that cross-attend to the key and value features from the encoders. However, this architecture is still incapable of learning to disentangle commingled weather features, arising from coexisting weather conditions, even if it is trained on multiple weather datasets. Therefore, we have designed an array of improvements that explicitly supply

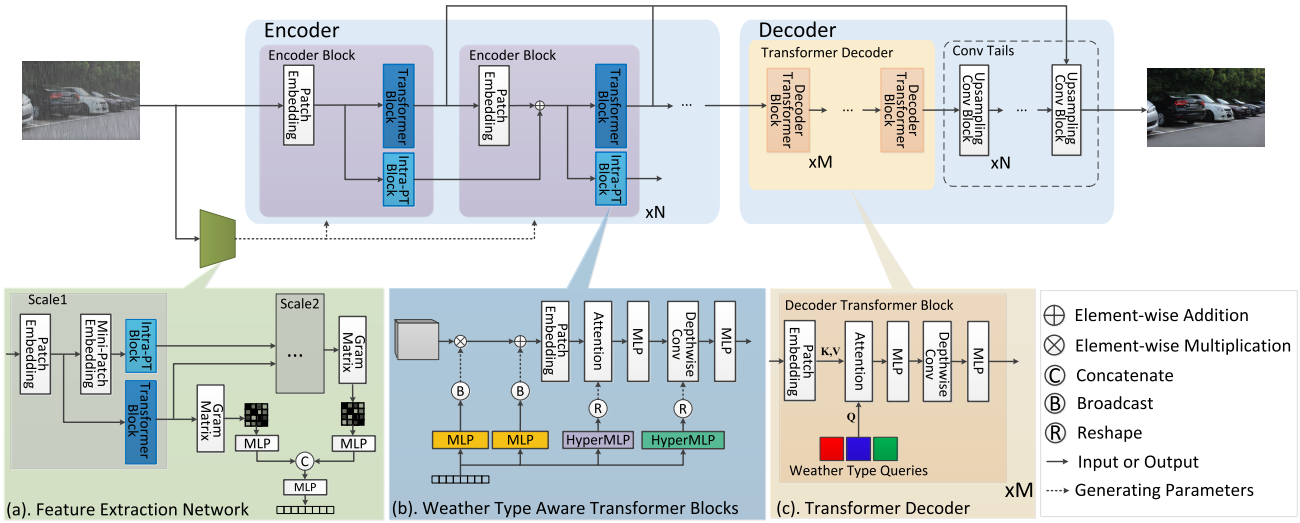


Fig. 2. The architecture of MWFormer. The main image processing network consists of a transformer encoder, a Transformer decoder, and convolution tails. (a) A feature extraction network learns to generate some of the parameters of the transformer blocks and intra-patch transformer blocks in the main network, thereby partially controlling the production of intermediate feature maps. (b) The transformer block in the encoder of the main network, which is guided by the feature vector. (c) Transformer decoder of the main network, whose queries are learnable parameters.

network flexibility in the multi-weather setting. The innovative designs we make are further explained in the following sections.

B. Feature Extraction Network

Weather variations can be viewed as distinct image "styles", which are inherently decoupled from the image content. To illustrate this idea, consider two snapshots of an identical scene, each captured under different weather conditions and manifesting distinct weather-related impairments. Each impaired (or "weather-styled") picture should be treated differently by the restoration network, but the two outputs should both faithfully recover the image content. On the other hand, pictures containing different contents, but suffering from the same weather degradation, should lead to comparable responses from the network. This is analogous to image style transfer, which emphasizes decoupling image style and content. The Gram matrix [44], which represents correlations within feature maps, is commonly used to define image styles. Yet, the original form of the Gram matrix fails in the context of multi-weather restoration, as it represents artistic styles rather than weather-relevant features. To address this, we append trainable projection layers—multi-layer perceptrons (MLPs)—on top of the vanilla Gram matrix, to learn weather-specific "style".

The architecture of our feature extraction network is shown in Fig. 2(a). We utilize the first two scales of the Transformer encoders, where a Gram matrix is computed at each scale. Since Gram matrices are symmetric, only the upper triangular parts of the two matrices are vectorized to save computation. These vectors are further fed to the two projection layers (MLPs), thereby generating two 64-dimensional embeddings. Finally, the two embeddings are concatenated and projected onto a single feature vector $\boldsymbol{v}$, which encodes the weather-degradation information from the input image.

The feature extraction network is intended to cluster images affected by similar weather degradations, hence we utilize contrastive learning [45] to train it, wherein the loss is formulated as:

$$\mathcal{L}_{con} = \sum_{(a,b)\in\mathcal{P}} \{\mathbb{I}(a,b)[m - d(\boldsymbol{v}_a, \boldsymbol{v}_b)]_+ + [1 - \mathbb{I}(a,b)]d(\boldsymbol{v}_a, \boldsymbol{v}_b)\}, \quad (1)$$

where $\mathcal{P}$ denotes every possible image pair in a batch, $d(\cdot)$ denotes cosine similarity, m is a positive margin, and $\mathbb{I}(a,b)$ is an indicator that equals 1 when the two images (a,b) contain the same weather impairments and 0 if they are captured under different weather conditions. The definition of $[\cdot]_+$ operation can be expressed as:

$$[x]_+ = \begin{cases} 0, & x \le 0, \\ x, & x > 0. \end{cases} \quad (2)$$

When calculating the contrastive loss, each possible image pair is sampled from the batch. If the two images belong to two different datasets, the term $d(\boldsymbol{v}_a, \boldsymbol{v}_b)$ enforces that their feature vectors are pushed away from each other. If the two images belong to the same dataset, the term $[m - d(\boldsymbol{v}_a, \boldsymbol{v}_b)]_+$ pulls their feature vectors closer in the embedding space. Consequently, the learned feature extraction network is able to cluster the images affected by the same weather degradation.

C. Image Restoration Network

The image restoration network contains two sets of learned parameters: fixed parameters that encode the general restoration priors relevant to all the tasks, and weather type-adaptive parameters that are generated by the feature extraction network, as shown in Fig. 2(b). More specifically, the output image $\boldsymbol{Y}$ is computed as:

$$\boldsymbol{v} = \mathcal{F}_{feat}(\boldsymbol{I}; \tau), \quad (3)$$

$$\boldsymbol{Y} = \mathcal{F}_{res}(\boldsymbol{I}; \theta_{fix}, \theta_{adap}(\boldsymbol{v})), \quad (4)$$

where $\mathcal{F}_{feat}$ is the auxiliary feature extraction network (Sec. III-B) with parameters τ, and $\mathcal{F}_{res}$ is the image restoration backbone. The parameters θ_{fix} and $\theta_{adap}(\boldsymbol{v})$

are the weather-independent and weather-adaptive weights in the encoder stages, respectively. Since different weather types require varying scales of treatments—for example, deraining mostly requires local contexts, while desnowing demands global understanding to differentiate snowflake and snowpack—we inject weather type adaptivity in multiple pillars: spatial-wise, both locally and globally in the parameter space, as well as channel-wise feature modulation, to enable better feature learning. The adaptivity is applied to both the Transformer block and the intra-PT blocks in the encoder stage. In the Transformer decoder blocks [2], the learnable weather-type queries attend to the input features, followed by standard MLP and depth-wise convolution layers, yielding restored output images $\boldsymbol{Y}$.

1) Spatially Local Adaptivity: Since vanilla Transformer architectures lack inductive biases expressive of local pixel interactions, we add a depthwise convolution layer between the two MLPs in each feed-forward network (FFN) in the Transformer blocks. Unlike previous models, however, we leverage the predicted weather type features $\boldsymbol{v}$ computed by the hyper-network $\mathcal{F}_{feat}$ to generate the parameters of the depthwise convolution layers, so that pictures degraded by different weather types will be processed by different filters adaptively. The feature vector $\boldsymbol{v}$ is fed into a 2-layer projection MLP (named HyperMLP in Fig. 2 since it is intended to generate the parameters of other modules), then reshaped to the 2D depthwise convolution kernels $\boldsymbol{w} \in \mathbb{R}^{C\times 1\times 3\times 3}$ (omitting the batch dimension) that are used to convolve the input $\boldsymbol{X}_{sl}$:

$$\boldsymbol{W}_{DWC} = \text{Reshape}(\text{Proj}(\boldsymbol{v})), \tag{5}$$

$$\text{FFN}(\boldsymbol{X}_{sl}, \boldsymbol{v}) = \text{MLP}(\sigma(\boldsymbol{W}_{DWC} * \boldsymbol{X}_{sl})), \tag{6}$$

where $\boldsymbol{W}_{DWC}$ denotes the weights of the depthwise convolution generated by reshaping the projection of the $\boldsymbol{v}$ vector, $\boldsymbol{X}_{sl}$ denotes the input of the spatially local operation (i.e., depthwise convolution), $*$ denotes depthwise convolution, and σ denotes nonlinear activation.

2) Spatially Global Adaptivity: Compared with CNN architectures, Transformers excel in capturing long-range spatial relationships using self-attention layers that scan over all the tokens. To model adaptive global interactions, we use another hyper-network to compute the critical projecting parameters used in the self-attention operations. Formally, denoting an input patch embedding of the spatially global operation (i.e., self-attention block) as $\boldsymbol{X}_{sg} \in \mathbb{R}^{N\times C_{in}}$, three linear projection matrices $\boldsymbol{W}_q$, $\boldsymbol{W}_k$ and $\boldsymbol{W}_v$ are applied to obtain the query, key, and value features $\boldsymbol{Q} = \boldsymbol{X}_{sg}\boldsymbol{W}_q, \boldsymbol{K} = \boldsymbol{X}_{sg}\boldsymbol{W}_k$ and $\boldsymbol{V} = \boldsymbol{X}_{sg}\boldsymbol{W}_v$. The matrix product of $\boldsymbol{Q}$ and $\boldsymbol{K}^T$ is then calculated, yielding a global attention map to weighted-sum $\boldsymbol{V}$. Different weather types may require different attention maps when conducting restoration, thus we employ the weather type feature $\boldsymbol{v}$ again, to generate $\boldsymbol{W}_q$, $\boldsymbol{W}_k$ and $\boldsymbol{W}_v$, using a similar projection as our design in Eq. (5). The result is then reshaped to match the dimensions of $\boldsymbol{W} \in \mathbb{R}^{d_{in}\times d_{out}}$. Mathematically,

$$\boldsymbol{W}_i = \text{Reshape}(\text{Proj}(\boldsymbol{v})),\ \ i = q, k, v, \tag{7}$$

$$\boldsymbol{Q} = \boldsymbol{X}_{sg}\boldsymbol{W}_q,\ \ \boldsymbol{K} = \boldsymbol{X}_{sg}\boldsymbol{W}_k,\ \ \boldsymbol{V} = \boldsymbol{X}_{sg}\boldsymbol{W}_v, \tag{8}$$

$$\text{MSA}(\boldsymbol{X}_{sg}) = softmax(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d}})\boldsymbol{V}. \tag{9}$$

3) Channel-Wise Feature Modulation: Except for weather-type awareness in the parameter space, we also introduce a dimension of degradation dependency in the intermediate feature space. We apply a simple affine transformation on the learned intermediate representations, which has been shown to be effective in previous work [46], [47], [48]. The feature vector $\boldsymbol{v}$ is input to the projection MLPs before each patch embedding layer to generate the weights $\gamma \in \mathbb{R}^C$ and biases $\beta \in \mathbb{R}^C$. For each channel, the weight and bias are then broadcast to all pixels of the corresponding feature map, which modulates the input features $\boldsymbol{X}$ along the channel dimension:

$$\boldsymbol{X}' = \gamma\boldsymbol{X} + \beta. \tag{10}$$

These modulating blocks may be regarded as a form of channel-attention mechanism that re-calibrates the importance of different channels, conditioned on the weather.

D. Simplified Architecture for Fixed Weather Degradation

Besides the aforementioned MWFormer architecture, we also developed a lightweight test-time variant for less computational cost. Our design for learning representations of weather type using an auxiliary hyper-network, which we use to guide the restoration backbone, also enables a computation-efficient inference scheme when the weather type is already known. Assuming that the learned weather-representation feature vectors for a given weather type lie near each other in the embedding space, then we can replace the feature extraction network with a fixed feature vector that represents the weather type, which is an approximation of the full-size model. More specifically, we pre-calculate and store the average feature vector of images affected by each weather type during training, then directly use these features while testing. This simplified architecture is shown in Fig. 3(b) and formulated as

$$\boldsymbol{Y} = \mathcal{F}_{res}(\boldsymbol{I}; \theta_{fix}, \theta_{adap}(\bar{\boldsymbol{v}}_i)), \tag{11}$$

where i denotes a specific degradation type (e.g., rainstreak, raindrop, snow), and $\bar{\boldsymbol{v}}_i$ is the average weather feature vector computed on images affected by the ith degradation type during training.

E. Multi-Stage Architecture for Hybrid Weather Degradations

We have developed another test-time variant for hybrid adverse weather removal. Due to a current lack of hybrid weather datasets, previous restoration models, whether trained to handle single or multiple types of weather, are unable to successfully restore pictures captured under multiple simultaneous adverse weather conditions, such as rain + snow. However, MWFormer can be easily modified without re-training to handle previously unseen multiple weather-degraded pictures, hence, it is more generalizable than prior models.

For example, consider a rain + snow hybrid weather condition. If a model is only trained on multiple single-weather restoration datasets, then it may be capable of restoring images degraded by any of the weather factors (in this case, rain or snow), but not a combined hybrid weather condition (in this case, rain + snow). Hence, we develop a two-stage network

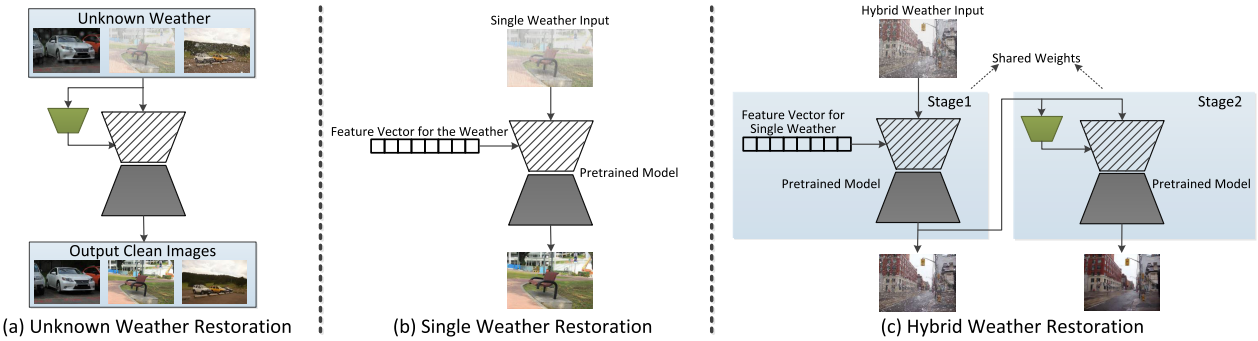


Fig. 3. Comparison of the default architecture with two test-time variants applied in special cases. To conduct a single weather-type restoration, the feature extraction network is replaced by a fixed feature vector. To conduct hybrid weather restoration, the image processing network is cascaded to remove degradations sequentially, stage by stage.

architecture as a test-time variant of MWFormer to handle such hybrid weather conditions. In the first stage of inference, the average feature vector for rainy images is used as the guidance of the image restoration backbone to produce an intermediate result that is rain-free but still contains snowflakes. Then, this intermediate output containing only a single adverse weather is processed using MWFormer again in the second stage to remove snowflakes, yielding a final clean image. The overall process can be denoted as:

$$\boldsymbol{Z} = \mathcal{F}_{res}(\boldsymbol{I}; \theta_{fix}, \theta_{adap}(\bar{\boldsymbol{v}}_i)), \tag{12}$$

$$\boldsymbol{v}_z = \mathcal{F}_{feat}(\boldsymbol{Z}; \tau), \tag{13}$$

$$\boldsymbol{Y} = \mathcal{F}_{res}(\boldsymbol{Z}; \theta_{fix}, \theta_{adap}(\boldsymbol{v}_z)). \tag{14}$$

If the image is subjected to more types of adverse weather, then further stages may be cascaded, wherein each stage restores a specific type of degradation. Note that the networks in the different stages share the same set of weights, thereby offering flexible test-time augmentation capability without requiring any re-training.

F. Extended Applications

The hyper-network that creates weather-informed feature vectors is a key aspect of our approach. Beyond generating parameters and modulating feature maps, these vectors have diverse applications due to the hyper-network's strong perception of weather features. We present two extended applications to demonstrate the versatility of the proposed hyper-network.

1) Weather-Type Identification: Our hyper-network, trained with a contrastive learning strategy on a multi-weather restoration dataset, contains rich prior information on various weather features. Leveraging this, we developed a weather-type identification method using the hyper-network without re-training.

Taking three adverse weather types as an example. Let $\bar{\boldsymbol{v}}_{rd}$, $\bar{\boldsymbol{v}}_r$ and $\bar{\boldsymbol{v}}_s$ be the average feature vectors of three weather types (raindrop, rainstreak, and snow) respectively, which is computed during training. To identify the weather type of an image $\boldsymbol{I}$ impaired by unknown adverse weather, the image's feature vector $\boldsymbol{v}$ is first computed using the feature extraction hyper-network $\mathcal{F}_{feat}$, then the cosine similarities between $\boldsymbol{v}$ and the average feature vectors of each weather type are computed:

$$\boldsymbol{v} = \mathcal{F}_{feat}(\boldsymbol{I}). \tag{15}$$

$$d_i = \frac{\boldsymbol{v} \cdot \bar{\boldsymbol{v}}_i}{\|\boldsymbol{v}\| \|\bar{\boldsymbol{v}}_i\|}, \quad i \in \{rd, r, s\}. \tag{16}$$

Finally, the scores related to each weather type are computed using Softmax function:

$$s_i = \frac{e^{d_i}}{e^{d_{rd}} + e^{d_r} + e^{d_s}}, \quad i \in \{rd, r, s\}. \tag{17}$$

The weather score s_i approximately indicates the probability that the image is degraded by the adverse weather type i. If this image is known to be affected by only one of the given adverse weather types, then it can be inferred that the highest-scoring type of weather i^* exists in the image:

$$i^* = \arg\max_i s_i, \quad i \in \{rd, r, s\}. \tag{18}$$

2) Guiding Pre-Trained Weather-Specific Models: Most existing adverse weather restoration models are trained for specific weather types, making them effective for known conditions but unable to handle unknown or even hybrid weather scenarios. This limits their real-world applicability. To make full use of these weather-specific experts, we've developed a strategy that utilizes the proposed hyper-network to guide existing pre-trained weather-specific models for restoring images affected by unknown weather conditions.

Suppose we have expert models for many different types of weather. When faced with an image affected by unknown weather conditions, our goal is to select the most suitable expert model, so that the image quality can be improved as much as possible. Without loss of generality, assume that we have three expert models for raindrop removal, rainstreak removal and desnowing respectively. We first compute the weather scores of three weather types using Eq. (15) ∼ (17). Then, the highest scoring weather type is considered to be the most typical and most impactful to image quality in this image. Therefore, the expert model corresponding to this weather type is selected to process the image. It should be noted that for images affected by hybrid weather, although the degradation may not be completely eliminated, our strategy is able to do as much as possible to improve the image quality with only one pre-trained weather-specific model, whereas other strategies cannot achieve higher image quality with the same or even more computational effort.

IV. Experiments

In this section, we first detail our experimental settings. Then, we compare the performance of MWFormer

TABLE I

QUANTITATIVE COMPARISONS OF MWFORMER AGAINST STATE-OF-THE-ART MULTI-WEATHER RESTORATION MODELS ON THREE TEST DATASETS. ALONG EACH COLUMN, THE BEST SCORE IS BOLDFACED, WHILE THE OTHER TOP THREE ARE UNDERLINED

Model	RainDrop [31]		Outdoor-Rain [49]		Snow100K [7]		Average		MACs (G) ↓
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	
AirNet [13]	24.57	0.8583	18.48	0.6719	24.41	0.8045	22.49	0.7782	301.27
Chen et al. [35]	31.83	0.9289	25.45	0.8737	28.86	0.8886	28.71	0.8971	24.56
All-in-One [12]	31.12	0.9268	24.71	0.8980	28.33	0.8820	28.05	0.9023	12.26
TransWeather [2]	30.17	0.9157	28.83	0.9000	29.31	0.8879	29.44	0.9012	6.13
WeatherDiffusion [1]	30.71	**0.9312**	29.64	**0.9312**	30.09	0.9041	30.15	**0.9222**	475.16 × 50
Zhu et al. [36]	31.31	0.93	25.31	0.90	29.71	0.89	28.78	0.91	1.36
MWFormer-S (ours)	31.09	0.9224	29.07	0.9010	30.05	0.8986	30.07	0.9073	3.57
MWFormer-M (ours)	31.56	0.9246	29.70	0.9064	30.45	0.9029	30.57	0.9113	6.45
MWFormer-L (ours)	31.73	0.9254	30.24	0.9111	30.70	0.9060	30.89	0.9142	10.41
MWFormer-real* **(ours)**	**31.91**	0.9268	**30.27**	0.9121	**30.92**	**0.9084**	**31.03**	0.9158	10.41

*MWFormer-real is trained on a larger dataset mentioned in sub-section IV-A.

against existing SOTA models both qualitatively and quantitatively. Furthermore, we also conducted comprehensive ablation studies to study the efficacy of different MWFormer model designs. Finally, we present some discussions on the effectiveness of the feature vectors in MWFormer and the generalization ability in Sec. V.

A. Training Details

For a fair comparison, we first followed the settings in [1], [2], [12] to train MWFormer on the standard benchmark for multi-weather restoration, which is a combination of three datasets: RainDrop [31], Outdoor-Rain [49], and Snow100K [7]. Similarly, we used the RainDrop test dataset [31], the Test1 dataset from Outdoor-Rain [49], and the Snow100K-L testset [7] for testing raindrop removal, draining with dehazing, and desnowing, respectively.

We first pre-trained the feature extraction network in MWFormer over 10k iterations using Eq. (1) as the loss function, with batch size 8 and learning rate $2e^{-4}$. Then, we trained the image restoration network over 200k iterations using a weighted combination of the smooth L1 loss and perceptual loss [50]. In our implementation, the difference between the feature maps extracted by a pretrained VGG16 (from the 3rd, 8th, and 15th layers) of the predicted image and that of the ground truth image was summed up as the perceptual loss. The total loss function is given as:

$$\mathcal{L}_{all} = \mathcal{L}_1 + \lambda \mathcal{L}_{perc}, \tag{19}$$

where λ was fixed at 0.04. To avoid overfitting to a specific dataset, we sampled approximately the same number of training examples from each dataset, respectively. Finally, the feature extraction network and the image restoration network were jointly fine-tuned over another 190k iterations using a reduced learning rate.

We instantiated three versions of MWFormer (Small, Medium, and Large), referred to as MWFormer-S, -M, and -L, of our proposed model by changing the number of base channels. In MWFormer-L, the number of channels for each encoder scale was 64, 128, 320, and 512, respectively, whereas the number of channels was reduced by the factors of 0.75 and 0.5 to create MWFormer-M and MWFormer-S, respectively.

Besides, it should be noted that some images in this widely adopted benchmark have different distributions from real-world scenes, which is likely to limit the model's real-world performance. For example, this dataset did not represent the veiling effect in multi-weather restoration [51]. To further improve MWFormer's applicability to real-world images, we re-trained MWFormer on a larger dataset, which is denoted MWFormer-real. Specifically, besides the previous benchmark dataset, we included another two datasets in the training set: the training set of WeatherStream [52] that contains real-world frames containing rain-fog degradations, and the training set of the CSD dataset [53] that includes images impaired by snowflakes and the veiling effects. We also re-trained TransWeather [2] on this larger dataset for a fair comparison.

B. Quantitative Comparisons

We used five state-of-the-art multi-weather restoration models as comparisons: All-in-One [12], Chen et al. [35], TransWeather [2], WeatherDiffusion [1] and Zhu et al. [36]. Another all-in-one image restoration model named AirNet [13] was also re-trained on the benchmark dataset for comparison. Table I reports the performances using PSNR and SSIM [54] as performance metrics. The computational costs of each model, evaluated by the number of multiply-accumulate operations (MACs), are also listed. As may be seen from the table, MWFormer-real performed best on all three datasets among all the compared methods in terms of PSNR, which is usually regarded as the most reliable measure of fidelity. MWFormer-L also performed better than any model trained using the benchmark dataset regarding average PSNR. Though Chen et al. [35] achieved better results on the Raindrop testset, their model performed poorly under the other two weather conditions, and the imbalance performance is not preferable in practice. In terms of the more perceptual-oriented metric SSIM, the diffusion-based WeatherDiffusion model, on average, achieved the best scores, but MWFormer yielded comparable results, performing among the top three.

Although WeatherDiffusion [1] performed well in terms of SSIM on some datasets, it requires 2000× more computation than our largest model MWFormer-L, and requires 5000× more computation than our smallest model MWFormer-S, if the iterative sampling diffusion process is considered. Overall, our MWFormer appears to deliver the best trade-off between image quality and computational cost.

Fig. 4. Qualitative comparisons on the RainDrop [31] test set. MWFormer effectively removed raindrop artifacts under various scenarios, yielding output images with either fewer shadows or less blur than the other compared models.

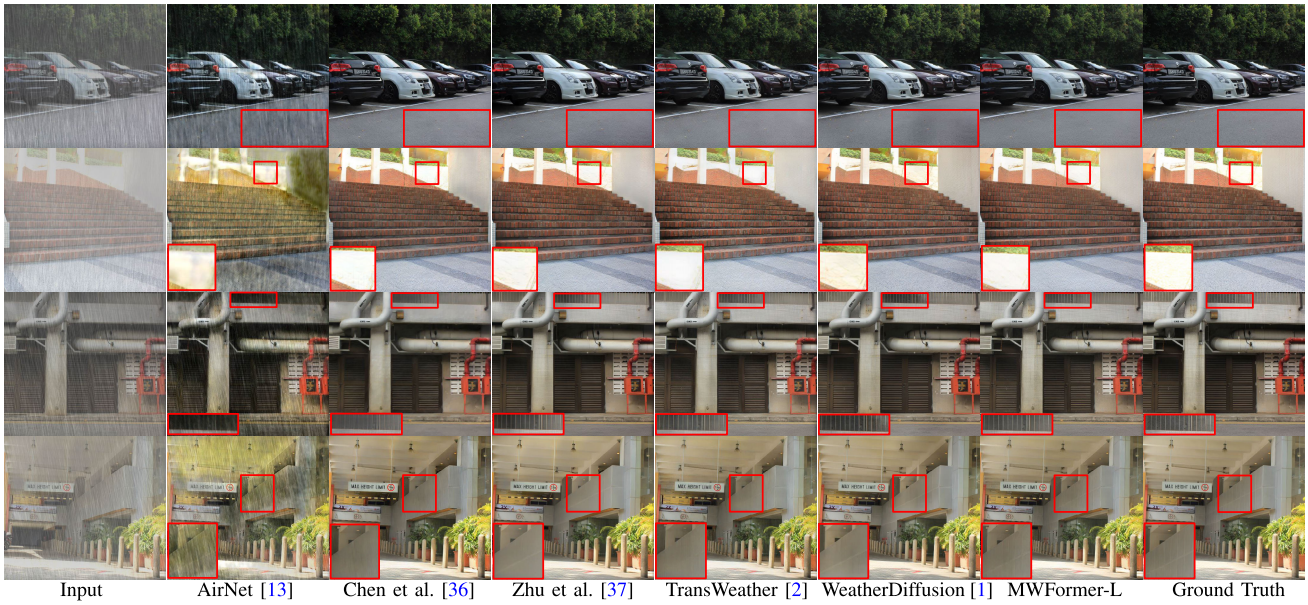


Fig. 5. Visual comparisons on the Test1 [49] (rain+fog) set. MWFormer performed the best on both detail restoration and luminance retention. AirNet failed to remove most of the degradations, TransWeather recovered fewer details, and the WeatherDiffusion introduced color distortions.

Besides, although WeatherDiffusion delivered the best SSIM results on the RainDrop and Outdoor-Rain sets, the diffusion model is occasionally prone to hallucinative artifacts. One of its failure cases is shown in the third row of Fig. 9, which exhibits unacceptable artifacts and stains that significantly alter the image contents. Since these restoration models are often employed as preprocessing modules for many downstream recognition tasks, such as object detection and semantic segmentation for autonomous vehicles, hallucinations of image content obtained from diffusion-based models could lead to hazardous outcomes in real-world scenarios.

Moreover, the comparison results of TransWeather-real and MWFormer-real are illustrated in Table. II, indicating that MWFormer still surpasses the existing leading models, such as TransWeather, if they are both trained on a larger dataset. Also, by including more images closer to the real scene, the quantity metrics on all of the test sets are boosted.

C. Qualitative Comparisons

We also obtained the visual results on each benchmark dataset as shown in Figs. 4 to 6. On the **RainDrop** test dataset, as shown in Fig. 4, AirNet failed to remove many of the raindrops. Both TransWeather and WeatherDiffusion produced artifacts such as shadows and hallucinations (see the first two rows). MWFormer, however, delivered visually pleasing results without shadows or blur. On the **Test1** (rain + fog) dataset shown in Fig. 5, MWFormer was able to restore both the luminance and detail information accurately, while the results from Chen et al. and TransWeather suffered from a loss of detail (note the texture in the last two rows), and the results produced by Zhu et al. and WeatherDiffusion included

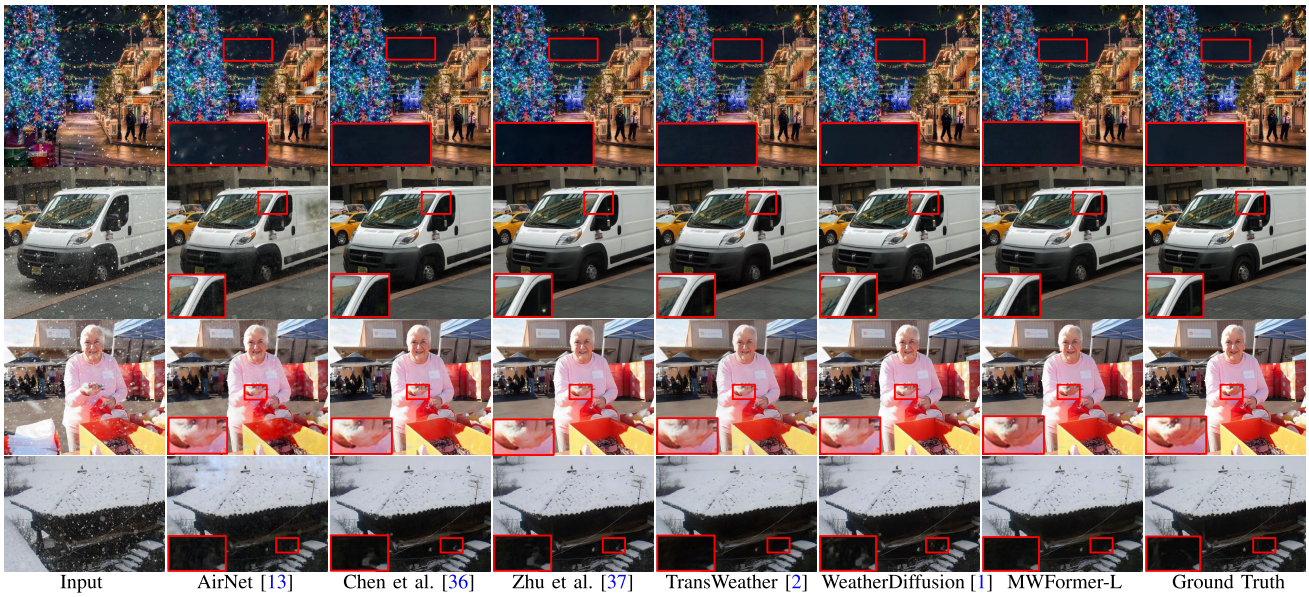


Fig. 6. Visual comparisons on the Snow100K-L [7] testset. MWFormer efficaciously removed snowflakes, delivering cleaner pictures than the other models.

TABLE II

COMPARISONS OF PERFORMANCES OF MWFORMER-REAL AND TRANSWEATHER-REAL

Model	RainDrop [31]	Outdoor-Rain [49]	Snow100K [7]	WeatherStream [52]	CSD [53]
TransWeather-real [2]	30.99 / 0.9207	28.98 / 0.9002	30.00 / 0.8996	24.29 / 0.7468	32.99 / 0.9580
MWFormer-real (ours)	31.91 / 0.9268	30.27 / 0.9121	30.92 / 0.9084	24.44 / 0.7488	34.60 / 0.9690

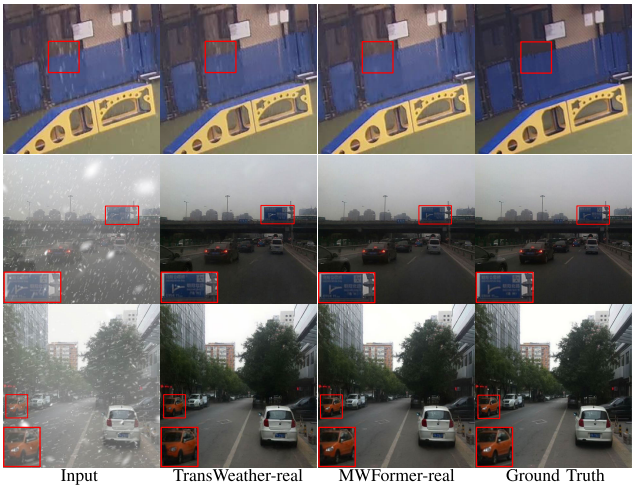


Fig. 7. Visual comparisons of TransWeather-real [2] and MWFormer-real on two additional testsets (WeatherStream testset [52] and CSD testset [53]) that are more consistent with real-world scenes.

shadows (see the first row). Additionally, WeatherDiffusion sometimes led to color distortion (see the second row). On the **Snow100K-L** dataset shown in Fig. 6, MWFormer yielded cleaner images, while AirNet, Zhu et al. and WeatherDiffusion tended to interpret some snowflakes as other image details and incorrectly preserved them, thereby reducing the image quality.

We also compared MWFormer-real and TransWeather-real on the two more realistic testsets: WeatherStream [52] and CSD [53] testsets. The visual results are shown in Fig. 7. On the WeatherStream dataset, MWFormer-real removes the rainstreaks more thoroughly than TransWeather-real, leading to more visually pleasing results. On the CSD dataset, TransWeather-real sometimes wrongly retains the snowflakes and tends to blur the small but bright objects excessively.

D. Performance on Hybrid Weather Degradations

More challenging but frequent scenarios are hybrid weather conditions. Hence, we also studied the performances of the compared models on hybrid-weather-degraded images. Using the weather synthesizing algorithms in [49], we simulated images with hybrid degradations of rain + snow using images from Snow100K. The results of restoring these degraded images are shown in Fig. 8. It can be seen that previous models failed to restore these images since obvious snowflakes, rain streaks, or fog remain in their outputs. This may be because hybrid-weather-degraded images were not part of their training data; models trained on single weather types cannot be expected to generalize to restore more complex weather degradations. However, MWFormer, which is imbued with the flexibility to conduct test-time augmentation (Fig. 3), is able to remove rain and snowflakes in two successive stages, yielding clean, degradation-free images. We also demonstrate the efficacy of multi-stage application by visualizing the effects of the stage-by-stage degradation removal process in Fig. 8.

To study the alternative approaches of our proposed multi-stage MWFormer architecture for rain + snow restoration problem, we compared four different strategies: First, we applied the simplest single-stage architecture that is intended for single-weather restoration to the rain + snow problem. Second, we applied the single-stage model to each image twice in succession. Third, using a two-stage MWFormer, we conducted desnowing first, using the average feature vector as guidance, followed by deraining. Last, we reversed the order by deraining first and then desnowing, as shown in Fig. 3(c). The performances of these models were tested on our synthetic dataset, which consists of diverse scenes, different rain levels, and different

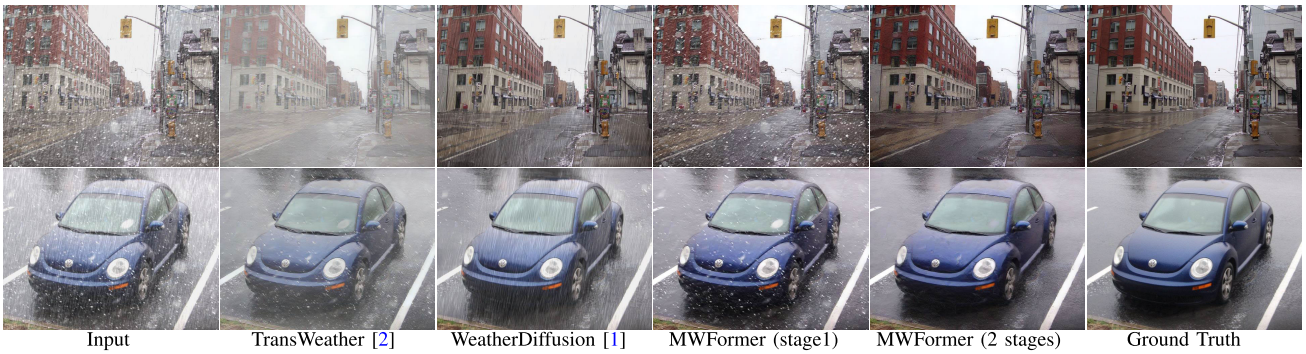


Fig. 8. Visual comparisons on hybrid-weather degradations. While most of the compared models failed to handle the complex degradations, the 2-stage MWFormer model, which sequentially removes rain streaks and snowflakes in each stage, was able to deliver more visually appealing outcomes.

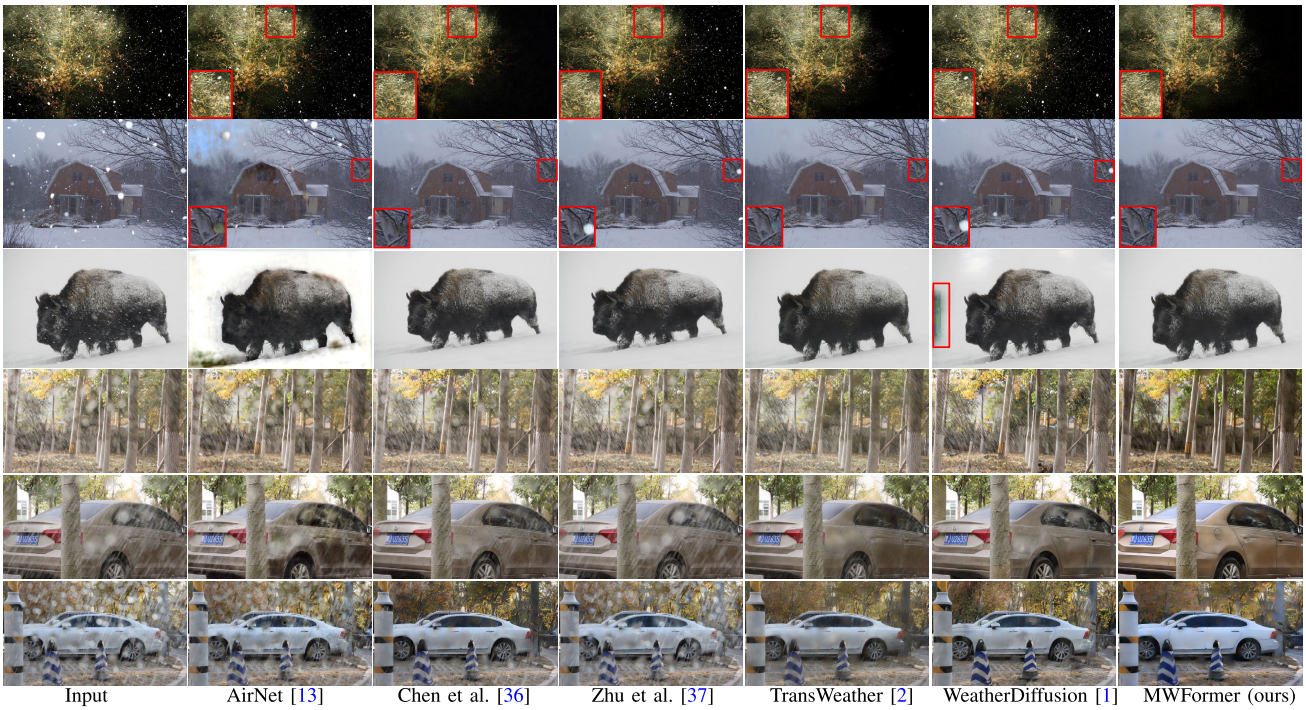


Fig. 9. Qualitative results on real images (including hybrid-weather-degraded images) from [7] and [32]. MWFormer was able to remove the snowflakes while preserving the original image structure. However, AirNet [13] and WeatherDiffusion [1] generated undesirable artifacts. Furthermore, MWFormer can capably remove the hybrid-weather degradations that were unseen during training, as shown in the last three rows.

TABLE III

COMPARISONS OF PERFORMANCES OF THREE DIFFERENT MWFORMER MODELS ON HYBRID RAIN + SNOW IMAGES

Strategy	PSNR	SSIM
Single Stage	21.24	0.7237
Two Stages with Default Settings	22.76	0.7665
Two Stages, Desnow First	21.98	0.7568
Two Stages, Derain First	24.80	0.7669

angles of rain streaks. Quantitative comparisons in Table III indicate that MWFormer performed best when deraining in the first stage, followed by desnowing in the second stage. This may be because snowflake appearance is significantly affected by rain accumulation; thus, the average feature vector for snow-degraded images may not match these images well. The intermediate results after deraining resemble the snow-degraded images in the training set, which are easier for the network to process. These results powerfully demonstrate the efficacy of the MWFormer model in dealing with multi-weather scenarios.

E. Generalization to Real Weather Degradations

We also compared MWFormer against other models on the real weather-degraded images from the Snow100K-real set [7] that contains pictures taken under real snowy conditions, and from RainDS-real dataset [32] that includes real-world images with raindrops and rainstreaks. We used the MWFormer-L to process the images in the Snow100K-real dataset, and use its variant discussed in Sec. III-E to restore the hybrid-weather-degraded images in the RainDS-real dataset. Note that no ground truth is available for these images, so we must rely on visual comparisons. As may be seen in Fig. 9, MWFormer was able to remove most of the snowflakes, yielding visually clean reconstructions as compared to other methods. As for images impaired by both raindrops and rainstreaks, MWFormer also performed the best, owing to its flexibility for hybrid-weather degradation removal. Moreover, it should be observed that WeatherDiffusion was exceedingly sensitive to domain shift—its performance varied significantly on different images, and it randomly generated unacceptable

Fig. 10. Task-driven comparisons on YOLO-V5 object detection. MWFormer helped deliver better detection performance than other compared methods. Note that AirNet [13] and WeatherDiffusion [1] were implicated in causing false positives in the detection results, likely due to inadequate restoration performance.

TABLE IV
NIQE SCORES OF MWFORMER AND PREVIOUS SOTA METHODS ON REAL-WORLD DATASETS [7], [32]

	TransWeather [2]	WeatherDiffusion [1]	MWFormer
NIQE ↓	3.2550	3.0162	**2.9469**

artifacts (the third row of Fig. 9). MWFormer, on the other hand, produced more visually consistent results in terms of real-weather generalization, which may be attributed to the smaller number of learnable parameters and the design of the weather-type feature learning. The quantitative comparisons using NIQE [55] are reported in Table IV, indicating that MWFormer outperforms the previous state-of-the-art models on this most widely adopted no-reference metric.

Fig. 11. Comparisons of the effects of visual-oriented and recognition-aware training. It may be observed that the latter strategy yielded less visually appealing outcomes, but led to better detection accuracy.

F. Task-Driven Comparisons

Image restoration results may be consumed either by humans or by machines. It is likely that weather degradation removal is more frequently used in machine vision systems, e.g., as a precursor to object detection for autonomous driving. We studied this aspect by conducting a study of task-driven image restoration performance in the context of object detection. Specifically, we evaluated the object detection performance of a pre-trained YOLO-V5 [56] object detector on images restored by the compared models. As shown in Fig. 10, on real images containing snowflakes, the pictures processed by MWFormer were able to better boost the detection performance of the YOLO-V5 as compared to applying the object detector to the original snow-degraded pictures. This suggests the potential of using MWFormer as a pre-processing component before object detectors in applications such as Autopilot [57]. The other image restoration methods, however, led to fewer detected objects and even misclassified some objects. It is worth noting that on images affected by raindrops, MWFormer only delivered slightly better detection performance than on the original image, while the other approaches had little effect or even deteriorated the detection performance. This observation is consistent with the empirical results in [3]. Lastly, we observed that AirNet and WeatherDiffusion tended to cause false positive cases ("skateboard" in the second row and "bird" in the bottom row) on some images, which could lead to unexpected and undesirable outcomes in real-world applications.

1) Recognition-Aware Training: We also studied ways to show that MWFormer can be specifically trained to benefit downstream detection models. To do this, we created a recognition-aware version of MWFormer by fine-tuning the base model of MWFormer for a few more steps, replacing the perceptual loss with a recognition loss (including both classification loss and regression loss), calculated using a MobileNetV3-SSDLite object detection network $\mathcal{R}$ with frozen weights. To calculate the recognition loss, we assigned the detection results of the clean image $\mathcal{R}(\boldsymbol{I}_{clean})$ as ground truth, and calculated its distance to the detection results on each restored image $\mathcal{R}(\mathcal{F}_{res}(\boldsymbol{I}; \theta_{fix}, \theta_{adap}(\boldsymbol{v})))$. The total loss was:

$$\mathcal{L}_{all} = \mathcal{L}_1 + \lambda(\mathcal{L}_{cls} + \mathcal{L}_{reg}), \tag{20}$$

where $\mathcal{L}_{cls} + \mathcal{L}_{reg}$ is the recognition loss consisting of two terms: $\mathcal{L}_{cls}$ is the classification loss implemented as the cross entropy between the predicted logits and the ground truth

TABLE V
ABLATION STUDIES OF COMPONENTS OF THE MWFORMER-L ARCHITECTURE

Local	Global	Channel	Fine-Tune	RainDrop [31]		Outdoor-Rain [49]		Snow100K [7]		Average	
				PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
				30.72	0.9173	29.23	0.9007	30.07	0.8992	30.01	0.9057
✓				31.11	0.9222	29.70	0.9028	30.14	0.8998	30.32	0.9083
✓	✓			31.19	0.9236	29.80	0.9055	30.18	0.9010	30.39	0.9100
✓	✓	✓		31.36	0.9235	29.89	0.9073	30.50	0.9041	30.58	0.9116
✓	✓	✓	✓	31.73	0.9254	30.24	0.9111	30.70	0.9060	30.89	0.9142

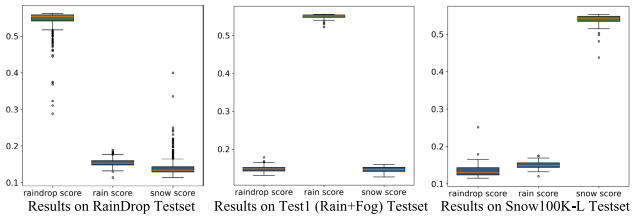


Fig. 12. Boxplots of the weather scores of different datasets.

labels, and $\mathcal{L}_{reg}$ is the regression loss implemented as smooth L1 loss between the predicted bounding box and the ground truth.

Several visualizations of the restored images with detection results overlaid are shown in Fig. 11. The detected object bounding boxes are overlaid, along with their associated detection confidence score. Some interesting observations can be drawn from this experiment. First, including the task-oriented training objective improved the performance of the downstream detection tasks, consistent with the findings in [58]. Further, optimizing for human quality perception and machine tasks led to different visual effects in the output images, indicating that deep neural network-based detectors learn different representations compared to human visual systems. Exploring more task-oriented image restoration techniques is beyond the scope of this paper, and hence, we leave it to future work.

G. Ablation Studies

To further understand and validate the efficacy of MWFormer, we conducted several comprehensive ablation studies. We used the MWFormer-L as the base model and ablated the various components trained using the same set of hyperparameters. We first trained a baseline MWFormer-L model (without the feature learning network) and gradually added 1) spatially local adaptivity, 2) spatially global adaptivity, 3) channel-wise feature modulation, and 4) joint fine-tuning, as explained in Section III-C. As may be seen in Table. V, each axis of weight adaptivity contributed to a notable performance gain on all the datasets, with Local adaptivity delivering the greatest gain on Raindrop and Outdoor-Rain datasets, while Channel adaptivity supplied the most benefit on the Snow100K. The final stage of joint fine-tuning can further boost the overall performance by aligning the separately trained feature extraction network with the image restoration backbone.

We also visualize the results of the ablation study in Fig. 13. The baseline model (without the three proposed modules) cannot thoroughly remove the artifacts, as pointed out by the up arrow in the first row and the down arrow in the second row. Some image details are also treated as artifacts and thus blurred, as indicated by the right arrow. The image quality can be largely improved after adding the local adaptivity modules to the model, owing to the adaptive local operations. Then, by adding the global adaptivity module, the model gains a better global understanding of how to differentiate snowflakes or raindrops and their background. In the first row, the model treats the content pointed by the right arrow as a light bulb rather than a snowflake. In the second row, the artifacts with the same color as the grass are suppressed. Last, by adding the channel-wise modulation module, the image details are further enhanced, as pointed out by the right arrow in the first row.

H. Results of Extended Applications

The strategy for computing weather scores was tested on RainDrop testset [31], Test1 dataset [49] (rain + fog), and Snow100K-L testset [7]. Boxplots in Fig. 12 illustrate the distribution of weather scores for each dataset, showing that each dataset scored significantly higher for its corresponding weather type than for others. Of all 17,069 test images, only 2 were misclassified. Overall, our proposed weather score aligns with the type of weather existing in the picture.

We also tested the strategy for guiding pre-trained expert models on real-world images with hybrid degradations [32]. Three SOTA pre-trained models were selected as the weather-specific experts: AST [59] for raindrop removal, ConvIR-Rain [60] for deraining, and ConvIR-Snow [60] for desnowing. Due to the absence of high-quality ground truth, we present the visual results in Fig. 14. To simulate the possible scenarios in practical use, we also implemented a comparison strategy: processing input images with each expert model separately and then averaging the outputs. This approach reflects how systems without our hyper-network cannot determine the weather characteristics of the input and cannot select a suitable expert, leading to a simple fusion of results. As shown in Fig. 14, while the simple averaging strategy required more computation, their results were far from satisfactory. In contrast, using the proposed feature extraction hyper-network, we can compute the weather scores of the input image and accordingly select the most appropriate expert model to eliminate the most visually distracting degradations in the image.

V. DISCUSSIONS

A. Detailed Comparisons With Our Baseline

1) Different Architectures: Fig. 15 compares the architecture of MWFormer and our baseline model TransWeather [2].

As for the overall framework, TransWeather only contains an image restoration backbone, whereas MWFormer additionally uses a feature extraction network to guide the

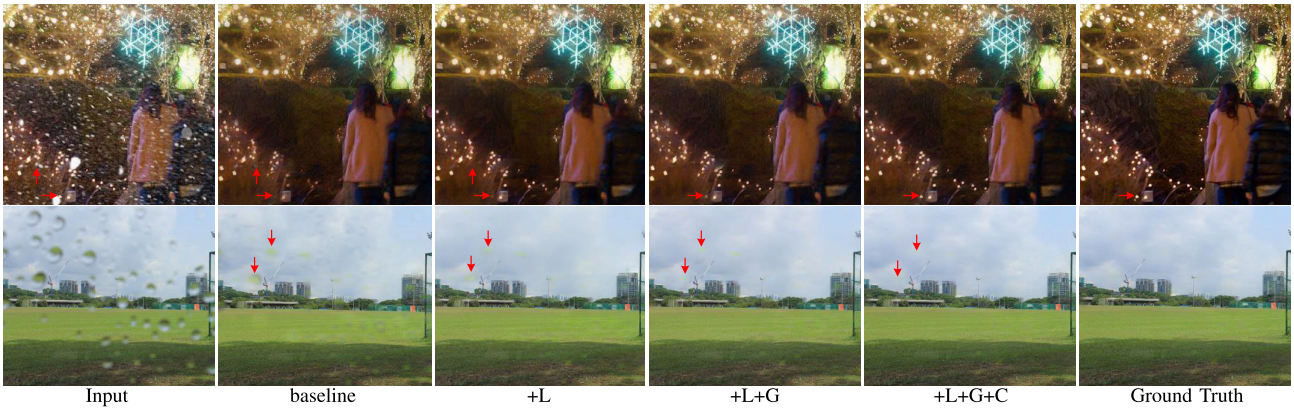


Fig. 13. Visualization of the ablation study. "L", "G" and "C" denote local adaptivity, global adaptivity and channel-wise feature modulation respectively.

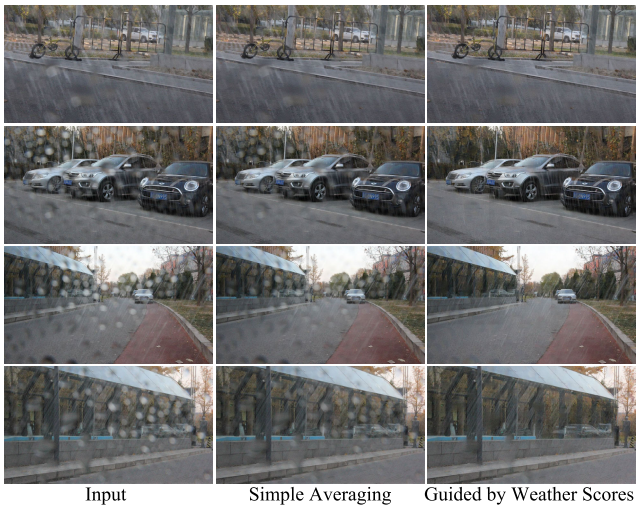


Fig. 14. Comparisons between the simple averaging strategy and our weather-score-guided strategy on real-world hybrid-weather-degraded images [32].

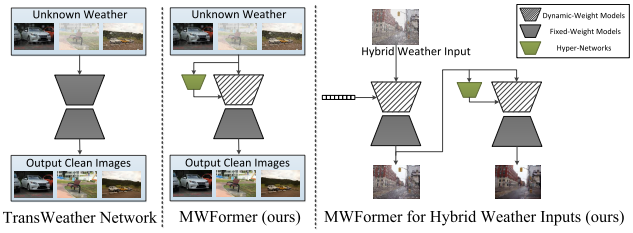


Fig. 15. Comparison between the baseline's architecture and our architectures.

operations of the image restoration backbone adaptively. With a well-designed structure and training strategy, the feature extraction network extracts information related to weather features from the Gram matrices.

As for the architecture of the image restoration backbone, TransWeather employs a common image restoration network architecture with all parameters fixed, which lacks a special design for the task of multi-weather restoration. On the contrary, our model is specifically designed for multi-weather restoration by dividing the parameters into two groups, i.e., the fixed parameters encoding the general restoration knowledge, and the weather-adaptive parameters dynamically generated using the feature vector. In addition to operations in the parameter space, the feature vector also modulates the image restoration network in the feature space.

Additionally, two test-time variants have been developed: one for reducing the computational cost, and the other for handling hybrid adverse weather types unseen during training. The proposed MWFormer is the first model capable of restoring images degraded by the unseen hybrid adverse weather.

2) Different Applications: The application of TransWeather is relatively limited, since it can only handle a few fixed weather types that have been already seen during the training phase. Our proposed MWFormer, with its flexibility, can restore the images impaired by **hybrid weather unseen during training**. This superiority over TransWeather indicates that MWFormer is more applicable to real-world scenarios, where different weather types may be commingled. Moreover, the proposed feature extraction hyper-network not only can be combined with MWFormer's image restoration backbone, but also has a wider range of application scenarios, such as identifying the weather type, and guiding pre-trained weather-specific expert models, as introduced in Sec. III-F. Besides, we have also explored ways of training the image restoration model to benefit downstream detection tasks (Sec. IV-F), which is not addressed in TransWeather [2].

B. Generalization Ability

To demonstrate the generalization ability of our methodology, we integrated our approach into three different network architectures and evaluated the results: two Transformer-based architectures (Restormer [42] and Uformer [43]) and a CNN-based architecture (UNet [61]).

For each of the architectures mentioned above, we trained two versions of the model: one using the original network structure and the other combined with our proposed adaptive method (denoted as "Ada-xxx"), both models with the same hyperparameters and number of channels. For Ada-Restormer and Ada-Uformer, we used the feature vector generated by the hyper-network to guide the restoration backbone across three dimensions and scales: locally spatial-wise, globally spatial-wise, and channel-wise. This allows part of the restoration backbone's parameters to be adaptively generated and its intermediate feature maps to be modulated based on the feature vector. Due to limited GPU memory, we reduced the encoder channels to 16 and 8 for the first scale of Ada-Restormer and Ada-Uformer, respectively, with a batch size of 16 for both. For Ada-UNet architecture, considering that CNN cannot capture long-range dependencies, we only applied the adaptivity

TABLE VI

COMPARISONS OF PERFORMANCES OF VARIOUS NETWORK ARCHITECTURES WITH OR WITHOUT OUR PROPOSED ADAPTIVE METHODOLOGY. THE PERFORMANCES OF THESE NETWORK ARCHITECTURES CAN BE SIGNIFICANTLY IMPROVED IF COMBINED WITH OUR METHODOLOGY

Model	RainDrop [31]		Outdoor-Rain [49]		Snow100K [7]		Average	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
Restormer	29.68	0.9042	28.29	0.8973	28.77	0.8768	28.91	0.8928
Ada-Restormer	29.80	0.9045	29.09	0.9035	29.21	0.8844	29.37	0.8975
Uformer	29.23	0.9266	25.41	0.8785	28.30	0.8732	27.65	0.8928
Ada-Uformer	29.88	0.9292	25.46	0.8841	28.82	0.8817	28.05	0.8983
UNet	29.19	0.9031	26.40	0.8857	28.60	0.8745	28.06	0.8878
Ada-UNet	29.70	0.9070	27.71	0.8975	29.11	0.8819	28.84	0.8955

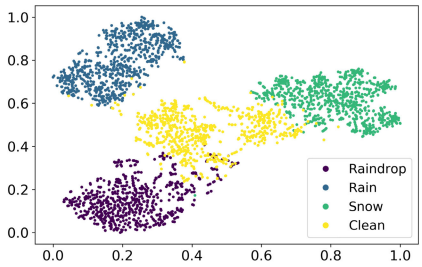


Fig. 16. A t-SNE visualization of the distributions of the extracted feature vectors from different weather datasets. The feature extraction network learns effective embedding that is able to cluster images according to their weather degradation types.

locally spatial-wise and channel-wise. In addition, we removed the batch normalization layers in Ada-UNet and the original UNet architecture, which are commonly regarded as unsuitable for image restoration tasks. The other settings are the same as those reported in Sec. IV-A.

The quantitative results are reported in Table VI, indicating that our method can significantly improve the performance of **various network architectures** on multiple datasets. These promising results show that our proposed approach can be used as **a general approach** to boost the performance of different network architectures on multi-weather restoration tasks.

C. Analysis of the Learned Weather Representation

To better illustrate how the learned feature vector improves the performance of the image restoration network, we utilized t-SNE [62] to visualize the distributions of the weather-type features learned by the feature extraction network $\mathcal{F}_{feat}$. As shown in Fig. 16, the computed feature embeddings quite effectively decoupled the weather degradations accross contents, since images degraded by the same weather type become closely clustered with little overlap. This suggests that the feature extraction network was able to learn to separate the content and degradation representations using contrastive loss.

We also examined the impact of feature vectors on image restoration using the simple version (Fig. 3(b)) of MWFormer. Using the raindrop removal as an example, we first tested the model on the Raindrop test set using the default setting for fixed weather degradation, meaning that the feature vector was the average of all the feature vectors of the raindrop images from the Raindrop training set. We then replaced the default feature vector with the feature vector computed on an arbitrary image from the Raindrop testset and the Snow100K testset, respectively. Numerical results in Table VII indicate that MWFormer performed the best when using the correct weather type embedding, demonstrating that average feature vectors

TABLE VII

RESULTS OF USING DIFFERENT FEATURE VECTORS IN THE SIMPLIFIED VERSION OF OUR MODEL

Average of Raindrop		Arbitrary Raindrop		Arbitrary Snow	
PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
29.38	0.9073	26.93	0.8961	21.76	0.8139

effectively represent their corresponding weather types. The performance slightly declined when using an arbitrary feature vector drawn from an image affected by the same weather type, and significantly dropped when using a feature vector of a different weather type. Generally, these results show that the vectors generated by our feature extraction network effectively encode weather-dependent information for guiding weather restoration tasks. Finally, owing to the design of feature guidance of our MWFormer, the users have the capability to arbitrarily control the action of the image restoration network by providing a feature vector according to their prior knowledge. This kind of flexibility during inference time is a key advantage that is unavailable in prior works.

VI. CONCLUDING REMARKS

We have introduced an efficient, all-in-one weather-aware Transformer, called MWFormer, for restoring images degraded by multiple adverse weather conditions. MWFormer consists of an encoder-decoder-based restoration backbone, augmented by an auxiliary feature extraction hyper-network that learns weather-type representations. The extracted feature vectors can be used to adaptively guide the main image restoration backbone by weight-adaptivity along the local, global, and channel axes. They can also be used for weather-type identification or guiding pre-trained expert models. Because of the availability of the auxiliary network, MWFormer can be extended to deal with fixed single-weather cases with less computation or hybrid-weather cases that were unseen during training. We conducted a spectrum of quantitative and qualitative studies on the multi-weather restoration benchmark dataset as well as on real-world datasets, and the results show that MWFormer outperforms prior known multi-weather restoration models without requiring much computational effort. Our methodology can also be integrated into a variety of network architectures to boost their performance.

ACKNOWLEDGMENT

This work was done when Zhengzhong Tu was with the Department of Electrical & Computer Engineering, University of Texas at Austin, Austin, TX 78712 USA.

REFERENCES

[1] O. Özdenizci and R. Legenstein, "Restoring vision in adverse weather conditions with patch-based denoising diffusion models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 10346–10357, May 2023.

[2] J. M. Jose Valanarasu, R. Yasarla, and V. M. Patel, "TransWeather: Transformer-based restoration of images degraded by adverse weather conditions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 2343–2353.

[3] S. Li et al., "Single image deraining: A comprehensive benchmark analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Apr. 2019, pp. 3838–3847.

[4] H. Wu et al., "Contrastive learning for compact single image dehazing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 10551–10560.

[5] X. Li, J. Wu, Z. Lin, H. Liu, and H. Zha, "Recurrent squeeze-and-excitation context aggregation net for single image deraining," in *Computer Vision—ECCV*. Cham, Switzerland: Springer, 2018, pp. 262–277.

[6] K. Zhang, R. Li, Y. Yu, W. Luo, and C. Li, "Deep dense multi-scale network for snow removal using semantic and depth priors," *IEEE Trans. Image Process.*, vol. 30, pp. 7419–7431, 2021.

[7] Y.-F. Liu, D.-W. Jaw, S.-C. Huang, and J.-N. Hwang, "DesnowNet: Context-aware deep network for snow removal," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 3064–3073, Jun. 2018.

[8] M. Li, X. Cao, Q. Zhao, L. Zhang, and D. Meng, "Online rain/snow removal from surveillance videos," *IEEE Trans. Image Process.*, vol. 30, pp. 2029–2044, 2021.

[9] S. Zhao, L. Zhang, Y. Shen, and Y. Zhou, "RefineDNet: A weakly supervised refinement framework for single image dehazing," *IEEE Trans. Image Process.*, vol. 30, pp. 3391–3404, 2021.

[10] B. Li et al., "Benchmarking single-image dehazing and beyond," *IEEE Trans. Image Process.*, vol. 28, no. 2, pp. 492–505, May 2018.

[11] L. Wang et al., "Unsupervised degradation representation learning for blind super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10576–10585.

[12] R. Li, R. T. Tan, and L.-F. Cheong, "All in one bad weather removal using architectural search," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3172–3182.

[13] B. Li, X. Liu, P. Hu, Z. Wu, J. Lv, and X. Peng, "All-in-one image restoration for unknown corruption," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17431–17441.

[14] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017.

[15] K. Zhang, W. Zuo, and L. Zhang, "FFDNet: Toward a fast and flexible solution for CNN-based image denoising," *IEEE Trans. Image Process.*, vol. 27, no. 9, pp. 4608–4622, Sep. 2018.

[16] K. Zhang et al., "Deep image deblurring: A survey," *Int. J. Comput. Vis.*, vol. 130, no. 9, pp. 2103–2130, Sep. 2022.

[17] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, and J. Matas, "DeblurGAN: Blind motion deblurring using conditional adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8183–8192.

[18] O. Kupyn, T. Martyniuk, J. Wu, and Z. Wang, "DeblurGAN-v2: Deblurring (orders-of-magnitude) faster and better," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8878–8887.

[19] C. Ledig et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4681–4690.

[20] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 136–144.

[21] X. Guo, Y. Li, and H. Ling, "LIME: Low-light image enhancement via illumination map estimation," *IEEE Trans. Image Process.*, vol. 26, no. 2, pp. 982–993, Feb. 2016.

[22] Y. Jiang et al., "EnlightenGAN: Deep light enhancement without paired supervision," *IEEE Trans. Image Process.*, vol. 30, pp. 2340–2349, 2021.

[23] Z. Meng, R. Xu, and C. M. Ho, "Gia-net: Global information aware network for low-light imaging," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 327–342.

[24] J. Xiao, X. Fu, A. Liu, F. Wu, and Z. Zha, "Image de-raining transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 2, no. 1, pp. 1–18, May 2022.

[25] X. Li, J. Wu, Z. Lin, H. Liu, and H. Zha, "Recurrent squeeze-and-excitation context aggregation net for single image deraining," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 254–269.

[26] S. W. Zamir et al., "Multi-stage progressive image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14816–14826.

[27] L. Chen, X. Lu, J. Zhang, X. Chu, and C. Chen, "HINet: Half instance normalization network for image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 182–192.

[28] Z. Tu et al., "MAXIM: Multi-axis MLP for image processing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5759–5770.

[29] R. Yasarla, V. A. Sindagi, and V. M. Patel, "Syn2Real transfer learning for image deraining using Gaussian processes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2723–2733.

[30] Y. Ba et al., "Not just streaks: Towards ground truth for single image deraining," in *Proc. IEEE Eur. Conf. Comput. Vis.*, Oct. 2022, pp. 723–740.

[31] R. Qian, R. T. Tan, W. Yang, J. Su, and J. Liu, "Attentive generative adversarial network for raindrop removal from a single image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2482–2491.

[32] R. Quan, X. Yu, Y. Liang, and Y. Yang, "Removing raindrops and rain streaks in one go," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9143–9152.

[33] W.-T. Chen, H.-Y. Fang, J.-J. Ding, C.-C. Tsai, and S.-Y. Kuo, "JSTASR: Joint size and transparency-aware snow removal algorithm based on modified partial convolution and veiling effect removal," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 754–770.

[34] J. Lin, N. Jiang, Z. Zhang, W. Chen, and T. Zhao, "LMQFormer: A laplace-prior-guided mask query transformer for lightweight snow removal," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 2, no. 1, pp. 6225–6235, Apr. 2023.

[35] W.-T. Chen, Z.-K. Huang, C.-C. Tsai, H.-H. Yang, J.-J. Ding, and S.-Y. Kuo, "Learning multiple adverse weather removal via two-stage knowledge learning and multi-contrastive regularization: Toward a unified model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17632–17641.

[36] Y. Zhu et al., "Learning weather-general and weather-specific features for image restoration under multiple adverse weather conditions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 21747–21758.

[37] S. Lee, T. Son, and S. Kwak, "FIFO: Learning fog-invariant features for foggy scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18889–18899.

[38] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[39] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, May 2021, pp. 9992–10002.

[40] H. Chen et al., "Pre-trained image processing transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12299–12310.

[41] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "SwinIR: Image restoration using Swin transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1833–1844.

[42] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5718–5729.

[43] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li, "Uformer: A general U-shaped transformer for image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17662–17672.

[44] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," 2015, *arXiv:1508.06576*.

[45] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," 2020, *arXiv:2002.05709*.

[46] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 4195–4205.

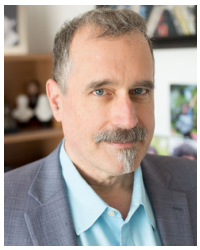
[47] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Apr. 2019, pp. 4401–4410.

[48] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. Courville, "FiLM: Visual reasoning with a general conditioning layer," 2017, *arXiv:1709.07871*.
[49] R. Li, L.-F. Cheong, and R. T. Tan, "Heavy rain image restoration: Integrating physics model and conditional adversarial learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1633–1642.
[50] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 694–711.
[51] P. W. Patil, S. Gupta, S. Rana, S. Venkatesh, and S. Murala, "Multi-weather image restoration via domain translation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 21639–21648.
[52] H. Zhang et al., "WeatherStream: Light transport automation of single image deweathering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 13499–13509.
[53] W.-T. Chen et al., "ALL snow removed: Single image desnowing algorithm using hierarchical dual-tree complex wavelet representation and contradict channel loss," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 4176–4185.
[54] W. Zhou, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 2, pp. 600–612, May 2004.
[55] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'Completely Blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2013.
[56] G. Jocher. (2020). *YOLOv5 By Ultralytics*. [Online]. Available: https://github.com/ultralytics/yolov5
[57] Tesla. *Autopilot*. Accessed: Nov. 20, 2023. [Online]. Available: https://www.tesla.com/autopilot
[58] Z. Liu et al., "Exploring simple and transferable recognition-aware image processing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3032–3046, Mar. 2023.
[59] S. Zhou, D. Chen, J. Pan, J. Shi, and J. Yang, "Adapt or perish: Adaptive sparse transformer with attentive feature refinement for image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 2952–2963.
[60] Y. Cui, W. Ren, X. Cao, and A. Knoll, "Revitalizing convolutional network for image restoration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, pp. 9423–9438, Dec. 2024.
[61] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015, *arXiv:1505.04597*.
[62] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

Ruoxi Zhu (Graduate Student Member, IEEE) received the B.E. degree from the School of Physics and Technology, Wuhan University, Wuhan, China, in 2022. He is currently working toward a Ph.D. degree with the Video & Image Processor Laboratory (VIP Lab), Fudan University, Shanghai, China. His research interests include image processing, machine learning, computational photography, and associated VLSI design.

Zhengzhong Tu (Member, IEEE) received the bachelor's and master's degrees from Fudan University, Shanghai, China, in 2016 and 2018, respectively, and the Ph.D. degree from The University of Texas at Austin, under the supervision of Prof. Alan C. Bovik. He is currently an Assistant Professor of computer science with Texas A&M University, College Station, TX, USA. Previously, he was a Research Engineer at Google Research from 2022 to 2024. His research interests include generative AI, multimodal AI, and their real-world applications, such as computational photography, autonomous driving, and robotics.

Jiaming Liu (Graduate Student Member, IEEE) received the B.E. degree in electronic science and technology from Southwest Jiaotong University, Chengdu, China, in 2019, the M.S. degree in circuit and system from Southwest Jiaotong University, in 2022. He is working toward a Ph.D. degree with the Video & Image Processor Laboratory (VIP Lab), Fudan University, Shanghai, China. His research interests include image processing, channel coding, machine learning, and associated VLSI architecture.



Alan C. Bovik (Life Fellow, IEEE) is currently the Cockrell Family Regents Endowed Chair Professor with The University of Texas at Austin. His books include *The Essential Guides to Image and Video Processing*. His research interests include image processing, digital photography, digital television, digital streaming video, social media, and visual perception. He is an Elected Member of the United States National Academy of Engineering, the Indian National Academy of Engineering, the National Academy of Inventors, and the Academy Europaea. For his work in these areas, he received the IEEE Edison Medal, the IEEE Fourier Award, the Primetime Emmy Award for Outstanding Achievement in Engineering Development from the Television Academy, the Technology and Engineering Emmy Award from the National Academy for Television Arts and Sciences, the Progress Medal from The Royal Photographic Society, the Edwin H. Land Medal from Optica, and the Norbert Wiener Society Award and the Karl Friedrich Gauss Education Award from the IEEE Signal Processing Society. He has also received about ten best journal paper awards, including the IEEE Signal Processing Society Sustained Impact Award. He co-founded and was the longest-serving Editor-in-Chief of IEEE TRANSACTIONS ON IMAGE PROCESSING and created/chaired the IEEE International Conference on Image Processing which was first held in Austin, TX, USA, in 1994.

Yibo Fan (Member, IEEE) received the B.E. degree in electronics and engineering from Zhejiang University, Hangzhou, China, in 2003, the M.S. degree in microelectronics from Fudan University, Shanghai, China, in 2006, and the Ph.D. degree in engineering from Waseda University, Tokyo, Japan, in 2009. He was an Assistant Professor with Shanghai Jiao Tong University and Fudan University from 2009 to 2014 and an Associate Professor with Fudan University from 2014 to 2019. He is currently a Full Professor with the College of Microelectronics, Fudan University. He is also the Founder of the Open Source ASIC Design Community (openasic.org) and developed the open source H.265/HEVC video encoder IP core (xk265), the H.264/AVC video encoder IP core (xk264), and the ISP IP core (xkISP). His research interests include image processing, video coding, machine learning, and associated VLSI architecture.