Efficient Deweather Mixture-of-Experts with Uncertainty-Aware Feature-Wise Linear Modulation

Rongyu Zhang[1,2], **Yulin Luo**[2], **Jiaming Liu**[2], **Huanrui Yang**[3], **Zhen Dong**[3], **Denis Gudovskiy**[4], **Tomoyuki Okuno**[4], **Yohei Nakata**[4], **Kurt Keutzer**[3], **Yuan Du**[1*], **Shanghang Zhang**[2*]

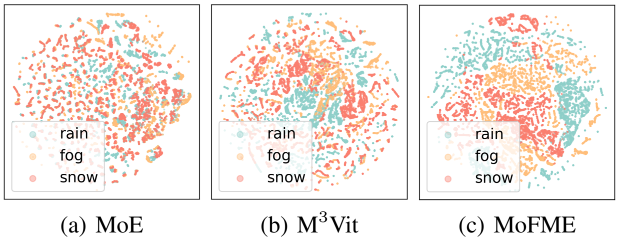
[1]Nanjing University
[2]National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University
[3]University of California, Berkeley
[4]Panasonic
yuandu@nju.edu.cn, shanghang@pku.edu.cn

Abstract

The Mixture-of-Experts (MoE) approach has demonstrated outstanding scalability in multi-task learning including low-level upstream tasks such as concurrent removal of multiple adverse weather effects. However, the conventional MoE architecture with parallel Feed Forward Network (FFN) experts leads to significant parameter and computational overheads that hinder its efficient deployment. In addition, the naive MoE linear router is suboptimal in assigning task-specific features to multiple experts which limits its further scalability. In this work, we propose an efficient MoE architecture with weight sharing across the experts. Inspired by the idea of linear feature modulation (FM), our architecture implicitly instantiates multiple experts via learnable activation modulations on a single shared expert block. The proposed Feature Modulated Expert (FME) serves as a building block for the novel Mixture-of-Feature-Modulation-Experts (MoFME) architecture, which can scale up the number of experts with low overhead. We further propose an Uncertainty-aware Router (UaR) to assign task-specific features to different FM modules with well-calibrated weights. This enables MoFME to effectively learn diverse expert functions for multiple tasks. The conducted experiments on the multi-deweather task show that our MoFME outperforms the state-of-the-art in the image restoration quality by 0.1-0.2 dB while saving more than 74% of parameters and 20% inference time over the conventional MoE counterpart. Experiments on the downstream segmentation and classification tasks further demonstrate the generalizability of MoFME to real open-world applications.

Introduction

There is a growing interest in low-level upstream tasks such as adverse weather removal (deweather) (Valanarasu et al. 2022). It intends to eliminate the impact of weather-induced noise on decision-critical downstream tasks such as detection and segmentation (Zamir et al. 2022). Previous methods (Ren et al. 2019; Chen et al. 2021) approach each type of weather effect independently, yet multiple effects can appear simultaneously in the real world. Moreover, such methods mainly focus on the deweathering performance metrics rather than an efficient deployment.



(a) MoE (b) M^{3}Vit (c) MoFME

Figure 1: t-SNE visualization of the router's outputs between different MoE architectures with adverse weather inputs.

One promising way to address several weather effects concurrently is the conditional computation paradigm (Bengio 2013), where a model can selectively activate certain parts of architecture, i.e. the task-specific experts, depending on the input. In particular, the sparse Mixture-of-Experts (MoE) (Riquelme et al. 2021) with parallel Feed Forward Network (FFN) experts rely on a router to activate a subset of FFNs for each weather-specific input image. Figure 2 shows a pipeline with an upstream MoE model to overcome a number of weather effects. For example, Ye et al. (2022) propose the DAN-Net method that estimates gated attention maps for inputs and uses them to properly dispatch images to task-specific experts. Similarly, Luo et al. (2023) develop a weather-aware router to assign an input image to a relevant expert without a weather-type label at test time.

Meanwhile, challenges exist in building a practical MoE-based model for deweather applications: ❶ *Efficient deployment*. Conventional MoE-based models with multiple parallel FFN experts require a significant amount of memory and computing. For example, MoWE (Luo et al. 2023) architecture contains up to hundreds of experts with billions of parameters. Hence, it is infeasible to apply such architectures to edge devices with limited resources for practical upstream tasks, e.g. to increase the safety of autonomous driving(Chi et al. 2023). Previous attempts to reduce memory and computation overheads inevitably sacrifice model performance (Xue et al. 2022). ❷ *Diverse feature calibration*. Existing MoE networks typically use naïve linear

[*]Corresponding author

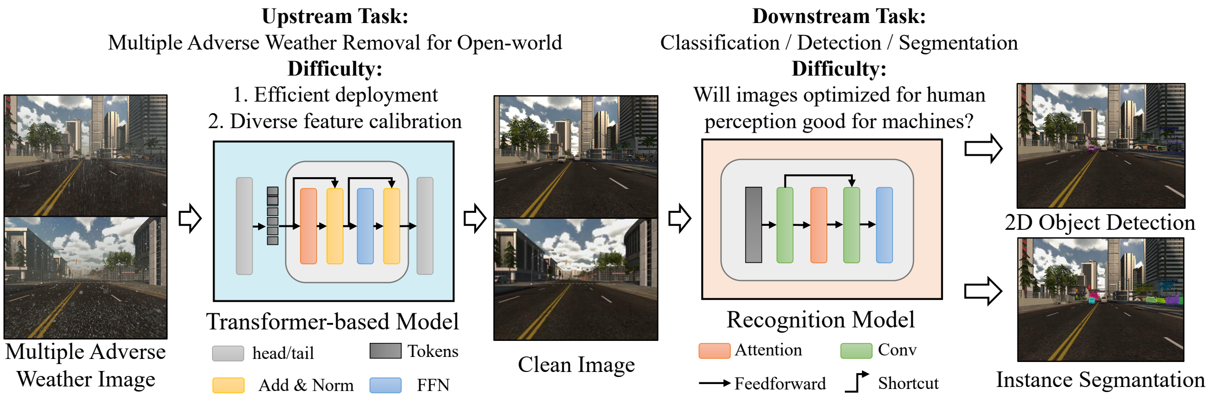


Figure 2: Deweather pipeline: upstream adverse weather removal with Mixture-of-Experts (MoE) and downstream instance segmentation and 2D object detection tasks.

routers for expert selection. This leads to poor calibration of router weights with diverse input features. Multi-gate MoE (Ma et al. 2018) overcomes this challenge by designing an additional gating network to distinguish task-specific features. However, this introduces additional computation costs. Therefore, we are motivated by the following objective: *is it possible to design a computationally-efficient MoE model while improving its deweathering metrics for real-world applications?*

To approach this objective, we start by analyzing redundancies in the conventional MoE architecture. The main one comes from multiple parallel experts containing independently learned weights. Meanwhile, previous research shows a possibility to simultaneously learn multiple objectives with diverse features using a mostly shared architecture and weights. For example, feature modulation (FM) (Perez et al. 2018; Liu et al. 2021, 2023) performs an input-dependent affine transformation of intermediate features with only two additional feature map parameters. Hence, the FM method allows decoupling multiple tasks simultaneously and implicitly represents ensemble models (Turkoglu et al. 2022) with low parameter overhead. Inspired by the FM method, we develop an efficient MoE architecture with feature-wise linear modulation for open-world scenarios. In particular, we propose **Mixture-of-Feature-Modulation-Experts** (MoFME) framework with two novel components: **Feature Modulated Expert** and **Uncertainty-aware Router**.

FME adopts FM into the MoE network via a single shared expert block. This block learns a diverse set of activation modulations with a minor overhead on the weight count. In particular, FME performs a feature-wise affine transformation on the model's intermediate features that is conditioned on the task-specific inputs. Next, it fuses task-specific modulated features with a single shared FFN expert, which allows it to efficiently learn a set of input-conditioned models. Thus, FME increases generalization to a wider range of substantially different tasks during training. As the T-SNE visualization shown in Figure 1, MoFME can better correlate the features with clearer partitions and boundaries.

The conventional MoE router adopts the top-K mechanism, which introduces non-differentiable operations into the computational graph and complicates the router optimization process. Previous research has found that such MoE router is prone to mode collapse, where it tends to direct all inputs to a limited number of experts (Riquelme et al. 2021). At the same time, Kendall, Gal, and Cipolla (2018) shows that uncertainty captures the relative confidence between tasks in the multi-task setting. Therefore, we propose our **UaR** router that estimates uncertainty using MC dropout (Gal and Ghahramani 2016). The estimated uncertainty is used to weigh modulated features and, therefore, route them to the relevant experts.

We verify the proposed MoFME method by conducting experiments on the deweather task. For instance, evaluation results with All-weather (Valanarasu et al. 2022) and RainCityscapes (Hu et al. 2019) datasets show that the proposed MoFME outperforms prior MoE-based model in the image restoration quality with less than 30% of network parameters. In addition, quantitative results on the downstream segmentation and classification tasks after applying the proposed MoFME further demonstrate the benefits of our pipeline with the upstream pre-processing. Our main contributions are summarized as:

- We introduce Mixture-of-Feature-Modulation-Experts (MoFME) framework with two novel components to improve upstream deweathering performance while saving a significant number of parameters.
- We develop Feature Modulation Expert (FME), a novel MoE layer to replace the standard FFN layers, which leads to improved performance and parameter efficiency.
- We devise an Uncertainty-aware Router (UaR) to enhance the assignment of task-specific inputs to the subset of experts in our multi-task deweathering setting.
- Experimental results demonstrate that the proposed MoFME can achieve consistent performance gains on both low-level upstream and high-level downstream tasks: our method achieves **0.1-0.2 dB** PSNR gain in image restoration compared to prior MoE-based model and outperforms SOTA baselines in segmentation and classification tasks while saving more than **72%** parameters and **39%** inference time.

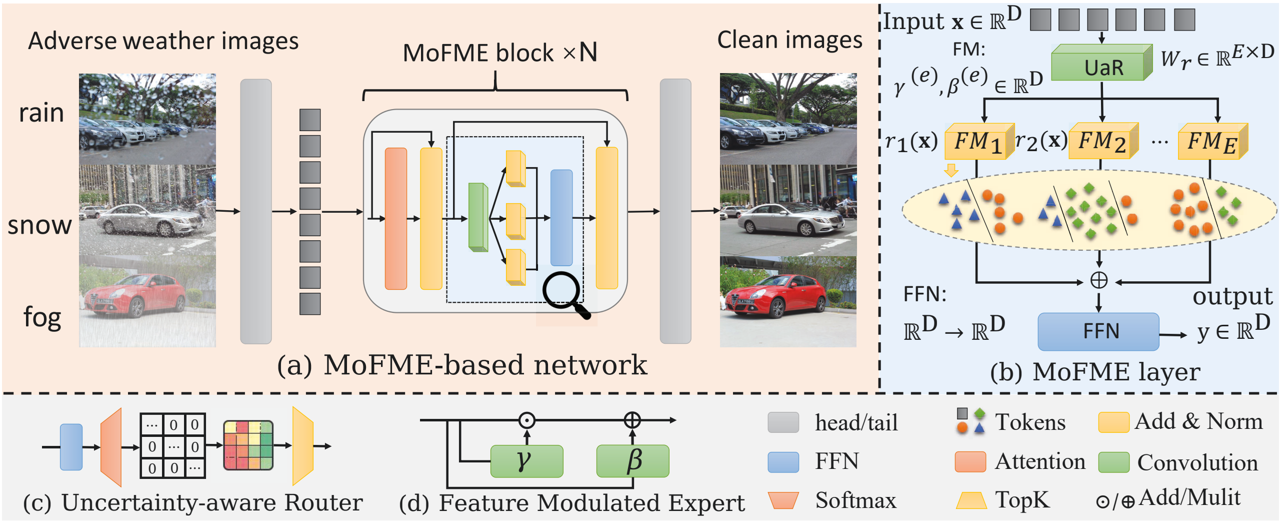


Figure 3: Schematic illustration of the proposed (a) Mixture-of-Feature-Modulation-Experts (MoFME) network, and the (b) detailed MoFME layer with two novel components (c) Uncertainty-aware Router and (d) Feature Modulated Expert.

Related Work

Mixture-of-Experts (MoE). Sub-model assembling is a typical way to scale up model size and improve performance in deep learning. MoE is a special case of assembling with a series of sub-models that are called the experts. It performs conditional computation using an input-dependent scheme to improve sub-model efficiency (Sener and Koltun 2018; Jacobs et al. 1991; Jordan and Jacobs 1994). Specifically, Eigen, Ranzato, and Sutskever (2013); Ma et al. (2018) assemble mixture-of-experts models into an architectural block that is known as the MoE layer. This enables more expressive modeling and decreases computation costs. Another solution is to sparsely activate only a few task-corresponding experts during training and inference. Liang et al. (2022) propose M^{3}ViT, which sparsely chooses the experts by using the transformer's token embeddings for router guidance. This helps the router to assign features to a selected expert during training and inference and to reduce computational costs. Our proposed MoFME is orthogonal to these MoE designs With the same goal of saving computational cost, our method instead proposes MoFME to substitute the over-parameterized parallel FFN experts with a lightweight feature modulation module followed by a single shared FFN expert.

Efficient MoE. Though MoE shows advantages in many popular tasks, its conventional architectures cannot meet requirements for practical real-world applications due to large model sizes. With many repetitive structures, pruning is the most common way to increase parameter efficiency. Wang et al. (2020); Yang et al. (2019); Chen et al. (2022) formulate channels and kernels as experts and introduce the task-specific gating network to filter out some parameters for each individual task. Several recent works (Xue et al. 2022; Rajbhandari et al. 2022) also consider applying knowledge distillation to obtain a lightweight student model for inference only. However, the above methods sacrifice model performance. Besides, Jiang et al. (2021); Liang et al. (2022) study how to efficiently adapt MoE networks to hardware devices while saving communication and computational costs. Instead, our MoFME aims to decrease computational costs and targets the redundancies in conventional over-parameterized FFN experts without a drop in performance by learning lightweight feature-modulated layers.

Adverse Weather Removal. Adverse weather removal has been explored in many aspects. For example, MPRNet (Zamir et al. 2021), SwinIR (Liang et al. 2021), and Restormer (Zamir et al. 2022) are architectures for general image restoration. Some methods can remove multiple adverse weathers at once. All-in-One (Li, Tan, and Cheong 2020) uses neural architecture search (NAS) to discriminate between different tasks. TransWeather (Valanarasu et al. 2022) uses learnable weather-type embeddings in the decoder. Transformer is also applied in this task. UFormer (Wang et al. 2022) and Restormer (Zamir et al. 2022) construct pyramidal network structures for image restoration based on locally-enhanced windows and channel-wise self-attention, respectively.

Proposed Methods

Feature Modulated Expert

We consider a common Mixture-of-Experts setting with the Vision Transformer (ViT) architecture (Dosovitskiy et al. 2021), where the dense FFN in each transformer block is replaced by a Mixture-of-Experts layer. The MoE layer inputs are N tokens $\boldsymbol{x} \in \mathbb{R}^D$ from the Multi-head Attention layer. Each token $\boldsymbol{x}$ is assigned by an input-dependent router into a set of E experts with router weight $r(\boldsymbol{x})$.

In a typical MoE design with a linear router, the functionality of the router can be formulated as

$$r(\boldsymbol{x}) = TopK(\text{softmax}(\mathbf{W}_r \boldsymbol{x})), \tag{1}$$

$$TopK(\mathrm{v}) = \begin{cases} \mathrm{v}, & \text{if v is in the top } K \text{ elements} \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

where $\mathbf{W}_r \in \mathbb{R}^{E \times D}$ is a trainable parameter, which maps input token into E router logits for experts selection. To reduce the computation cost, the experts in the model are sparsely activated, with $TopK(\cdot)$ setting all elements of the router weight to zero except the elements with the largest K values. For clarity in the rest of the paper, we denote the router weight of the i^{th} expert as $r_i(\boldsymbol{x})$.

The output of the MoE layer is therefore formulated as the weighted combination of the experts' output on the input token $\boldsymbol{x}$ (Shazeer et al. 2017) as

$$MoE(\boldsymbol{x}) = \sum_i r_i(\boldsymbol{x}) e_i(\boldsymbol{x}), \quad (3)$$

where $e_i(\cdot)$ denotes a functionality of the i^{th} expert, typically designed as a FFN in the context of vision transformers. This process is illustrated in Figure 3(a).

In this work, we employ the technique of Linear Feature Modulation (Perez et al. 2018) into the design of MoE to propose the efficient Feature Modulated Expert block, as illustrated in Figure 3(b). Specifically, the diverse task-specific features, i.e. tokens, are first modulated with a dynamic feature modulation unit, where the tokens are directed to different learned affine transformations based on an input-dependent router. The modulated features are then fused by a single shared FFN expert. In this way, we implicitly represent each expert in the MoE architecture as the cascading modules of a lightweight affine feature modulation transformation and a shared FFN, significantly reducing the parameter and computation overhead for adding additional experts.

First we formulate a single Feature Modulation (FM) block (Perez et al. 2018). We obtain input-dependent feature modulation parameters $\gamma \in \mathbb{R}^D$ and $\beta \in \mathbb{R}^D$ with two functions $g : \mathbb{R}^D \to \mathbb{R}^D$ and $b : \mathbb{R}^D \to \mathbb{R}^D$ respectively according to an input token $\boldsymbol{x}$ as

$$\gamma = g(\boldsymbol{x}) \qquad \beta = b(\boldsymbol{x}), \quad (4)$$

where g and b can take arbitrary learnable functions. In practice, those functions are implemented with lightweight 1×1 convolutions. The input token is then modulated as

$$FM(\boldsymbol{x}) = \gamma \circ \boldsymbol{x} + \beta, \quad (5)$$

where $\circ$ is the Hadamard (element-wise) product taken w.r.t. the feature dimension.

To combine the FM module with MoE, we instantiate E independent FM modules to modulate diverse task-specific features, each parameterized with $\gamma^{(i)}$ and $\beta^{(i)}$, where $i \in \{1, ..., E\}$. Adapting from the traditional MoE formulation, we let the router select which FM module to apply on the input token, rather than which FFN to be used. Specifically, our FME module is formulated as

$$\begin{aligned} &FME(\boldsymbol{x}|\gamma, \beta) \\ &= FFN\left\{\sum_i r_i(\boldsymbol{x}) \cdot [\gamma^{(i)} \circ \boldsymbol{x} + \beta^{(i)}]\right\}, \end{aligned} \quad (6)$$

where a single shared FFN module can process the mixture of multi-task features by the diverse feature modulations.

Uncertainty-aware Router

To improve the FME performance, we propose Uncertainty-aware Router (UaR), which performs implicit uncertainty estimation on the router weights according to MC dropout (Gal and Ghahramani 2016). Model uncertainty (Lakshminarayanan, Pritzel, and Blundell 2017) measures if the model *knows what it knows*. Although there exists ensemble-based uncertainty estimation methods (Ovadia et al. 2019; Ashukha et al. 2020) that often achieve the best calibration and predictive accuracy, the high computational complexity and storage cost motivates us to use the more efficient MC dropout (Rizve et al. 2021).

Specifically, we can regard the output of a certain router $r(\boldsymbol{x})$ as a Gaussian distribution to calibrate its uncertainty. The mean and covariance of such distribution can be estimated via a "router ensemble", where we pass the token representation $\boldsymbol{x}$ to get $r(\boldsymbol{x})$ with the router for M times according to MC dropout. We denote the resulted ensemble as $r^m(\boldsymbol{x}) = \{r^1(\boldsymbol{x}), r^2(\boldsymbol{x}), ..., r^M(\boldsymbol{x})\}$, and the mean and covariance of the router weights in the ensemble as $\check{\mu}$ and $\check{\Sigma}$ respectively. We calibrate and normalize the router's logits according to Al-Shedivat et al. (2020) as

$$\check{r}(\boldsymbol{x}) = \check{\Sigma}^{-1}[r(\boldsymbol{x}) - \check{\mu}]/||\check{\Sigma}^{-1}[r(\boldsymbol{x}) - \check{\mu}]||_2, \quad (7)$$

where $\check{r}(\boldsymbol{x})$ is used in the forward and backward pass during the training. The mean $\check{\mu}$ and inverse covariance $\check{\Sigma}^{-1}$ are both formulated as zero-padded diagonal matrices in the computation. The detailed structure is shown in Figure 3.

Optimization Objective

MoE-based model would suffer from performance degradation if most inputs are assigned to only a small subset of experts (Fedus, Zoph, and Shazeer 2022; Lepikhin et al. 2021). A load balance loss $\mathcal{L}_{lb}$ (Lepikhin et al. 2021) is therefore proposed for MoE to penalize the number of inputs dispatched to each router:

$$\mathcal{L}_{lb} = \frac{E}{N} \sum_{n=1}^{N} \sum_{i=1}^{E} v_i(\boldsymbol{x}_n) r_i(\boldsymbol{x}_n), \quad (8)$$

where x_n is the n-th input token, and $v_i(\boldsymbol{x}_n)$ is 1 if the i-th expert is selected for $\boldsymbol{x}_n$ by the top-k function, otherwise 0. The combined MoE training loss therefore becomes

$$\mathcal{L}_{MoE} = \mathcal{L}_{ts} + \lambda_1 \mathcal{L}_{lb}, \quad (9)$$

where λ_1 is empirically set to $1e^{-2}$ and $\mathcal{L}_{ts}$ indicates the task-specific loss computed by model outputs and corresponding labels, e.g., MSE loss for image restoration task.

Following Lepikhin et al. (2021), we further leverage the covariance $\check{\Sigma}$ of $r^m(\boldsymbol{x})$ to penalize the updating of UaR and MoFME and formulate the uncertainty loss $\mathcal{L}_{uc}$ as

$$\mathcal{L}_{uc} = \frac{E}{N} \sum_{n=1}^{N} \sum_{i=1}^{E} \check{\Sigma}_i \cdot v_i(\boldsymbol{x}_n), \quad (10)$$

where v is defined the same as in Equation (8). $\mathcal{L}_{uc}$ can further reduce the model uncertainty when optimized together with other losses, where the final MoFME objective is

$$\mathcal{L}_{MoFME} = \mathcal{L}_{ts} + \lambda_1 \mathcal{L}_{lb} + \lambda_2 \mathcal{L}_{uc}, \quad (11)$$

where λ_2 is empirically set to $5e^{-3}$.

Base model	MoFME		Param.	FLOPs	Derain		Deraindrop		Desnow		Average	
	FME	UaR	(M)	(GMAC)	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Baseline	-	-	8.71	34.93	27.64	0.9329	28.21	0.9249	28.40	0.8860	28.08	0.9146
MoE	-	-	44.19	37.06	27.91	0.9359	28.54	0.9307	28.76	0.8926	28.40	0.9197
	✓	-	18.53	36.26	27.87	0.9342	28.43	0.9290	28.65	0.8901	28.32	0.9178
	-	✓	44.19	37.17	27.96	0.9363	28.52	0.9304	28.80	0.8930	28.43	0.9199
	✓	✓	18.53	36.37	28.01	0.9368	28.55	0.9311	28.78	0.8925	28.45	0.9201
M^{3}ViT	-	-	44.22	37.06	27.67	0.9356	28.42	0.9280	28.61	0.8911	28.23	0.9182
	✓	✓	18.56	36.37	27.87	0.9344	28.51	0.9301	28.70	0.8912	28.36	0.9185
MoWE	-	-	34.15	59.99	28.05	0.9370	28.93	0.9333	28.75	0.8923	28.58	0.9209
	✓	✓	21.22	48.36	28.10	0.9376	29.03	0.9346	28.84	0.8927	28.66	0.9216

Table 1: Ablation study on All-Weather using PSNR and SSIM metrics. We set 16 experts and top2 gate.

	# Experts	PSNR	FLOPs	Param.	Infer. time
MoE	64	28.45	41.31	157.7	0.039
MoFME	64	**28.46**	**37.43**	**47.1 (70%↓)**	**0.027 (31%↓)**
MoE	128	28.56	41.31	309.1	0.075
MoFME	128	**28.59**	**37.43**	**85.2 (72.5%↓)**	**0.046 (39%↓)**

Table 2: Comparison with different numbers of experts on All-Weather. We set top2 gate in the experiments.

Experiments

We benchmark our MoFME framework on adverse weather removal, assuming a test-time scenario that necessitates a single-parameter model to eliminate diverse weather conditions. Our results also highlight MoFME's seamless integration with downstream segmentation and classification tasks. Ablation analyses validate the significance of each MoFME component, with the architecture delivering a 0.1-0.2 dB gain in PSNR and reducing parameters and inference time by over 72% and 39%, respectively.

Experimental Setup

Implementation details. We implement our method with the PyTorch framework using 4×NVIDIA A100 GPUs. We train the network for 200 epochs with a batch size of 64. The initial learning rate of the AdamW optimizer and Cosine LR scheduler is set to 0.5×10^{-4} and is gradually reduced to 10^{-6}. We use a warm-up stage with three epochs. Input images are randomly cropped to 256×256 size for training, and non-overlap crops of the same size are used at test time. We randomly flip and rotate images for data augmentation. The scaling factor for the traditional MoE model is set to 4.

Metrics, datasets, and baselines. We select widely-used PSNR and SSIM metrics as performance measures for upstream image restoration. All-weather (Valanarasu et al. 2022) and Rain/HazeCityscapes (Hu et al. 2019; Sakaridis, Dai, and Van Gool 2018) datasets are used to evaluate deweathering and downstream segmentation. The CIFAR-10 dataset is for the downstream image classification task.

The comparison baselines include three CNN-based models RESCAN (Li et al. 2018), PRNet (Ren et al. 2019), and FFA-Net (Qin et al. 2020) that employ task-specific weather removal. Also, we experiment with recent transformer-based models: Restormer (Zamir et al. 2022) with a general multi-task image restoration objective, TransWeather (Valanarasu et al. 2022) with learnable weather embeddings in the decoder to remove multiple adverse effects simultaneously, conventional MoE (Shazeer et al. 2017), MMoE (Ma et al. 2018), M^{3}ViT (Liang et al. 2022), and MoWE (Luo et al. 2023) for multi-task learning, as well as efficient MoE methods such as OneS (Xue et al. 2022), which fuses the experts' weight and adopt knowledge distillation for better performance and PR-MoE (Rajbhandari et al. 2022), which propose a pyramid residual MoE architecture to demonstrate the superiority of our proposed MoFME to handle multiple tasks in both effectiveness and efficiency. We take Vision Transformer as the backbone for MoE-based methods.

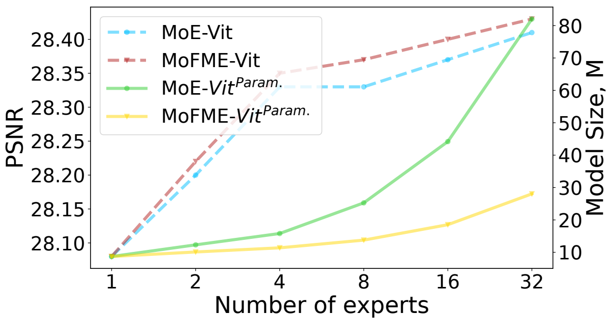


Figure 4: Number of experts v.s. PSNR and the model size. All models are trained for 100 epochs. We set top2 gate.

Ablation Study

Table 1 details our ablation studies on the MoE architecture, where we substitute FFN experts with FME and introduce MC dropout via UaR to the router. FME demonstrates considerable parameter savings with slight performance loss, while UaR enhances performance by over 0.05 dB. When applied to various MoE models, including M^{3}ViT and MoWE, our approach consistently boosts both efficiency and efficacy.

Type	Method	Derain		Dehaze		Average		Param.	FLOPs
		PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	(M)↓	(GMAC)↓
Task-specific	RESCAN	19.11	0.9118	16.96	0.9033	18.04	0.9076	0.15	32.32
	PReNet	19.95	0.8822	18.22	0.8729	19.09	0.8776	0.17	66.58
	FFA-Net	28.29	0.9411	28.96	0.9432	28.63	0.9422	4.46	288.34
Multi-task	Transweather	24.08	0.8481	22.56	0.8736	23.32	0.8609	38.05	6.12
	Restormer	28.06	0.9630	22.72	0.9167	28.11	0.9336	26.13	140.99
Multi-task MoE	MoE-Vit	32.70	0.9725	31.07	0.9623	31.89	0.9674	44.19	41.31
	MMoE-Vit	32.47	0.9698	31.08	0.9582	31.78	0.9640	44.63	42.56
	M^{3}ViT	32.56	0.9712	31.11	0.9597	31.84	0.9655	44.25	41.60
	MoWE	**32.99**	**0.9755**	*31.31*	*0.9647*	**32.15**	**0.9701**	34.15	59.99
Efficient MoE	OneS	32.40	0.9691	30.96	0.9590	31.68	0.9641	**8.71**	**34.93**
	PR-MoE	32.38	0.9700	31.03	0.9595	31.71	0.9648	27.28	37.53
	MoFME (ours)	*32.87*	*0.9721*	**31.35**	**0.9661**	*32.11*	*0.9691*	*18.53*	*37.43*

Table 3: Quantitative Comparison on Rain/HazeCityscapes using PSNR and SSIM. We set 16 experts and top4 gate.

Type	Method	mIoU		mAcc	
		Derain	Dehaze	Derain	Dehaze
Multi-task MoE	MoE	0.4652	0.4541	0.7684	0.7443
	MMoE	0.4621	0.4530	0.7643	0.7418
	M^{3}ViT	0.4634	0.4525	0.7662	0.7421
	MoWE	0.4686	0.4545	0.7701	0.7473
Efficient MoE	OneS	0.4620	0.4519	0.7665	0.7402
	PR-MoE	0.4632	0.4528	0.7660	0.7410
	MoFME	0.4650	0.4550	0.7681	0.7480

Table 4: Downstream semantic segmentation results after deweathering on Cityscapes using mIoU and mAcc. The expert number and topk settings are the same as Table 3.

Methods	MoE	Efficient	Param.	FLOPs	CIFAR-10
ViT	-	-	13.06	0.85	98.21%
MoE-ViT	✓	-	46.16	1.03	98.33%
OneS	✓	✓	13.06	0.85	98.14%
MoFME-ViT	✓	✓	18.05	0.94	98.47%

Table 5: Top-1 accuracy of image classification tasks. We set the number of experts 8 and the top2 gate.

Scalability is a hallmark of our MoFME model, as evidenced in Table 2. As expert count grows into the hundreds, MoFME retains efficiency, reducing parameters by 75% and improving performance by over 0.1 dB in adverse weather conditions compared to standard MoE. Moreover, MoFME cuts inference time by nearly 40% with 128 experts.

Quantitative Analysis

Upstream tasks In Table 3 and 6, we report the PSNR and SSIM of each type of weather and the average scores for each baseline and MoFME on All-Weather (Chen et al. 2021) and RainCityscapes (Hu et al. 2019) after training for 200 epochs. We denote the best results in bold, and the second-best results in italics. It should be noted that all the experiments are trained with a mixture of weather data and inference with a specific type of weather. The results of Table 3 and 6 reveal the advantage of MoE networks to deal with multi-task inputs compared with previous naïve transformer-based and CNN-based methods. However, as it is specifically designed for high-level tasks, M^{3}ViT fails to exert good performance on deweather tasks on both datasets. Furthermore, current efficient MoE methods like OneS and PR-MoE cannot achieve comparable performance compared with SOTA MoE networks, while MoFME can achieve 29.09 dB average PSNR score and 0.9272 average SSIM on All-Weather, and 32.11 dB PSNR and 0.9691 SSIM on RainCityscapes. While the MoWE model attains superior performance metrics, it is worth noting that both the model's size and its computational complexity, as quantified by the FLOPs, are substantially greater when compared to our traditional MoE-based approach.

We also provide the FLOPs and the number of parameters for each baseline on RainCityscapes in Table 3. The MoE-based methods can achieve very satisfying scores on PSNR and SSIM. However, the heavy network structure prevents them from practical applications. The two efficient MoE baselines exert their advantages in computational costs as PR-MoE can save about 50% parameters, and OneS merges its parameters to become a lightweight dense model. However, the certain model performance of the two methods is also sacrificed as OneS decrease almost 0.2 dB in PSNR. Our proposed MoFME takes a step forward by realizing a satisfied trade-off as it achieves compatible results compared to other SOTA baselines while saving up to 72% parameters.

Downstream task ❶ *Semantic segmentation:* Although our proposed methods exert satisfying performance with efficiency on low-level image restoration tasks, however, as has been questioned by Liu et al. (2022), *will images optimized for better human perception can be accurately recognized by machines?* We provide the quantitative comparison on Cityscapes for downstream segmentation tasks based on mIoU and mAcc in Table 4. We can find that other efficient MoE baselines fail to make satisfying predictions

Type	Method	Derain		Deraindrop		Desnow		Average	
		PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
Task-specific	RESCAN	21.57	0.7255	24.26	0.8367	24.30	0.7586	23.38	0.7736
	PReNet	23.16	0.8624	24.96	0.8629	25.19	0.8483	24.44	0.8579
	FFA-Net	27.96	0.8857	27.73	0.8894	27.21	0.8578	27.63	0.8776
Multi-task	Transweather	25.64	0.8103	27.37	0.8570	26.98	0.8305	26.66	0.8326
	Restormer	27.85	0.8802	28.32	0.8881	28.18	0.8684	28.12	0.8789
Multi-task MoE	MoE-Vit	28.47	0.9420	29.06	0.9367	29.20	0.8987	28.91	0.9258
	MMoE-Vit	28.52	0.9415	28.91	0.9368	29.13	0.8986	28.85	0.9256
	$\mathrm{M^3ViT}$	28.61	0.9428	28.75	0.9345	29.27	0.9004	28.88	0.9259
	MoWE	*28.59*	*0.9432*	**29.37**	**0.9400**	**29.37**	**0.9014**	**29.11**	**0.9282**
Efficient MoE	OneS	28.35	0.9384	28.89	0.9341	28.98	0.8976	28.74	0.9234
	PR-MoE	28.43	0.9394	28.97	0.9342	29.18	0.8980	28.86	0.9239
	MoFME(ours)	**28.66**	**0.9436**	*29.27*	*0.9385*	*29.35*	*0.8996*	*29.09*	*0.9272*

Table 6: Quantitative comparison on All-Weather using PSNR and SSIM metrics. We set 16 experts and top2 gate.

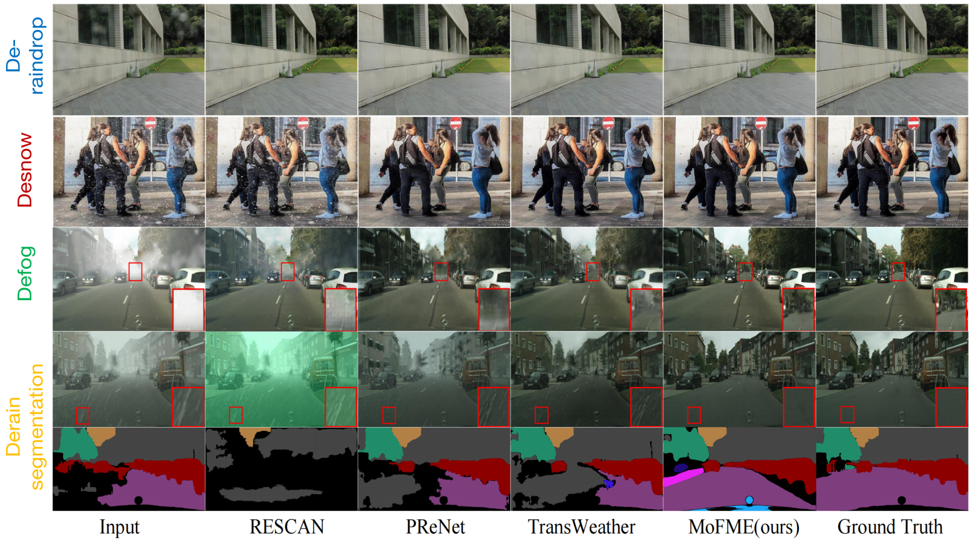


Figure 5: Examples of noisy inputs (left), noise-free ground-truth (right), and the methods' deweathering results. We show the upstream image restoration (top) and the effects on the downstream segmentation task (bottom) for RainCityscapes.

on the downstream task. On the other hand, our proposed MoFME exerts satisfying performance on both the upstream deweather task and downstream task by outperforming 2% mIoU and 2.5% mAcc compared with other efficient MoE baselines. We also provide the visualization results in Figure 5. ❷ *Image classification:* To further prove the generality of our methods, we perform image classification task on CIFAR-10 with ImageNet pre-training. The top-1 accuracy is reported in Table 5 which shows that MoE models lead to performance gain with parameter costs, while MoFME outperforms other similar size baselines by 0.2% on CIFAR-10.

Qualitative Analysis

Visual results in Figure 5 show the qualitative comparison of our method against the other methods. As shown in the top three rows, MoFME can achieve better visual results compared with previous methods, which recovers sharper information of the original image, especially in the defog setting. The visual results also demonstrate that our method can further recover downstream task-friendly images with better semantic segmentation outcomes. Our proposed MoFME can segment out clearer boundaries while maintaining consistency in color and texture.

Conclusion

In this work, we proposed MoFME with a novel FME and UaR. Experiments on deweathering tasks demonstrated that MoFME can handle multiple tasks simultaneously, as it outperformed prior MoE-based baselines by 0.1-0.2 dB while saving more than 72% of parameters and 39% inference time. Downstream classification and segmentation results proved MoFME generalization to real-world applications.

Acknowledgments

Shanghang Zhang is supported by the National Key Research and Development Project of China (No.2022ZD0117801). The authors would like to express their sincere gratitude to the Interdisciplinary Research Center for Future Intelligent Chips (Chip-X) and Yachen Foundation for their invaluable support.

References

Al-Shedivat, M.; Gillenwater, J.; Xing, E.; and Rostamizadeh, A. 2020. Federated learning via posterior averaging: A new perspective and practical algorithms. *arXiv preprint arXiv:2010.05273*.

Ashukha, A.; Lyzhov, A.; Molchanov, D.; and Vetrov, D. 2020. Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Bengio, Y. 2013. Deep learning of representations: Looking forward. In *Statistical Language and Speech Processing: First International Conference (SLSP)*.

Chen, T.; Huang, S.; Xie, Y.; Jiao, B.; Jiang, D.; Zhou, H.; Li, J.; and Wei, F. 2022. Task-Specific Expert Pruning for Sparse Mixture-of-Experts. *arXiv:2206.00277*.

Chen, W.-T.; Fang, H.-Y.; Hsieh, C.-L.; Tsai, C.-C.; Chen, I.; Ding, J.-J.; Kuo, S.-Y.; et al. 2021. All snow removed: Single image desnowing algorithm using hierarchical dual-tree complex wavelet representation and contradict channel loss. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Chi, X.; Liu, J.; Lu, M.; Zhang, R.; Wang, Z.; Guo, Y.; and Zhang, S. 2023. BEV-SAN: Accurate BEV 3D Object Detection via Slice Attention Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17461–17470.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Eigen, D.; Ranzato, M.; and Sutskever, I. 2013. Learning factored representations in a deep mixture of experts. *arXiv:1312.4314*.

Fedus, W.; Zoph, B.; and Shazeer, N. 2022. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *Journal of Machine Learning Research*.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Hu, X.; Fu, C.-W.; Zhu, L.; and Heng, P.-A. 2019. Depth-attentional features for single-image rain removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jacobs, R. A.; Jordan, M. I.; Nowlan, S. J.; and Hinton, G. E. 1991. Adaptive mixtures of local experts. *Neural computation*, 3(1): 79–87.

Jiang, H.; Zhan, K.; Qu, J.; Wu, Y.; Fei, Z.; Zhang, X.; Chen, L.; Dou, Z.; Qiu, X.; Guo, Z.; et al. 2021. Towards more effective and economic sparsely-activated model. *arXiv:2110.07431*.

Jordan, M. I.; and Jacobs, R. A. 1994. Hierarchical mixtures of experts and the EM algorithm. *Neural computation*, 6(2): 181–214.

Kendall, A.; Gal, Y.; and Cipolla, R. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems (NIPS)*.

Lepikhin, D.; Lee, H.; Xu, Y.; Chen, D.; Firat, O.; Huang, Y.; Krikun, M.; Shazeer, N.; and Chen, Z. 2021. GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Li, R.; Tan, R. T.; and Cheong, L.-F. 2020. All in one bad weather removal using architectural search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3175–3185.

Li, X.; Wu, J.; Lin, Z.; Liu, H.; and Zha, H. 2018. Recurrent squeeze-and-excitation context aggregation net for single image deraining. In *Proceedings of the European conference on computer vision (ECCV)*, 254–269.

Liang, H.; Fan, Z.; Sarkar, R.; Jiang, Z.; Chen, T.; Zou, K.; Cheng, Y.; Hao, C.; and Wang, Z. 2022. M^3 ViT: Mixture-of-Experts Vision Transformer for Efficient Multi-task Learning with Model-Accelerator Co-design. In *Advances in Neural Information Processing Systems (NIPS)*.

Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; and Timofte, R. 2021. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 1833–1844.

Liu, J.; Lu, M.; Chen, K.; Li, X.; Wang, S.; Wang, Z.; Wu, E.; Chen, Y.; Zhang, C.; and Wu, M. 2021. Overfitting the data: Compact neural video delivery via content-aware feature modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4631–4640.

Liu, J.; Yang, S.; Jia, P.; Lu, M.; Guo, Y.; Xue, W.; and Zhang, S. 2023. ViDA: Homeostatic Visual Domain Adapter for Continual Test Time Adaptation. *arXiv preprint arXiv:2306.04344*.

Liu, Z.; Wang, H.; Zhou, T.; Shen, Z.; Kang, B.; Shelhamer, E.; and Darrell, T. 2022. Exploring Simple and Transferable Recognition-Aware Image Processing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3): 3032–3046.

Luo, Y.; et al. 2023. MoWE: mixture of weather experts for multiple adverse weather removal. *arXiv:2303.13739*.

Ma, J.; Zhao, Z.; Yi, X.; Chen, J.; Hong, L.; and Chi, E. H. 2018. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.

Ovadia, Y.; Fertig, E.; Ren, J.; Nado, Z.; Sculley, D.; Nowozin, S.; Dillon, J.; Lakshminarayanan, B.; and Snoek, J. 2019. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in Neural Information Processing Systems (NIPS)*, 32.

Perez, E.; Strub, F.; De Vries, H.; Dumoulin, V.; and Courville, A. 2018. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Qin, X.; Wang, Z.; Bai, Y.; Xie, X.; and Jia, H. 2020. FFA-Net: Feature fusion attention network for single image dehazing. In *Proceedings of the AAAI conference on artificial intelligence*, 11908–11915.

Rajbhandari, S.; Li, C.; Yao, Z.; Zhang, M.; Aminabadi, R. Y.; Awan, A. A.; Rasley, J.; and He, Y. 2022. Deepspeed-moe: Advancing mixture-of-experts inference and training to power next-generation ai scale. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Ren, D.; Zuo, W.; Hu, Q.; Zhu, P.; and Meng, D. 2019. Progressive image deraining networks: A better and simpler baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Riquelme, C.; Puigcerver, J.; Mustafa, B.; Neumann, M.; Jenatton, R.; Susano Pinto, A.; Keysers, D.; and Houlsby, N. 2021. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems (NIPS)*.

Rizve, M. N.; Duarte, K.; Rawat, Y. S.; and Shah, M. 2021. In Defense of Pseudo-Labeling: An Uncertainty-Aware Pseudo-label Selection Framework for Semi-Supervised Learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Sakaridis, C.; Dai, D.; and Van Gool, L. 2018. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126: 973–992.

Sener, O.; and Koltun, V. 2018. Multi-task learning as multi-objective optimization. *Advances in Neural Information Processing Systems (NIPS)*.

Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; and Dean, J. 2017. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Turkoglu, M. O.; Becker, A.; Gündüz, H. A.; Rezaei, M.; Bischl, B.; Daudt, R. C.; D'Aronco, S.; Wegner, J. D.; and Schindler, K. 2022. FiLM-Ensemble: Probabilistic Deep Learning via Feature-wise Linear Modulation. In *Advances in Neural Information Processing Systems (NIPS)*.

Valanarasu; et al. 2022. TransWeather: transformer-based restoration of images degraded by adverse weather conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, X.; Yu, F.; Dunlap, L.; Ma, Y.-A.; Wang, R.; Mirhoseini, A.; Darrell, T.; and Gonzalez, J. E. 2020. Deep mixture of experts via shallow embedding. In *Uncertainty in artificial intelligence (UAI)*.

Wang, Z.; Cun, X.; Bao, J.; Zhou, W.; Liu, J.; and Li, H. 2022. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 17683–17693.

Xue, F.; He, X.; Ren, X.; Lou, Y.; and You, Y. 2022. One Student Knows All Experts Know: From Sparse to Dense. *arXiv:2201.10890*.

Yang, B.; Bender, G.; Le, Q. V.; and Ngiam, J. 2019. CondConv: Conditionally parameterized convolutions for efficient inference. *Advances in Neural Information Processing Systems (NIPS)*.

Ye, T.; Chen, S.; Liu, Y.; Chen, E.; and Li, Y. 2022. Towards efficient single image dehazing and desnowing. *arXiv preprint arXiv:2204.08899*.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14821–14831.