减少帧间与专家激活冗余的视频去雪混合特征调制专家模型

摘要

在现实场景中部署深度神经网络常常会面对需要为多个大同小异的任务分别配置一个特化模型而造成冗余的问题。本工作首先在 Allweather 数据集上复现 Zhang 等人提出的混合特征调制专家模型去完成多合一图像去除天气任务，用一个模型进行图片中多种天气效果（雨、雪、雾等）的去除。进一步地，我们观察到在视频任务中不同帧的类型和退化程度存在差异但连续帧之间也存在着时序冗余，即对不同帧的处理有异有同，因此基于原文方法增加了参考帧机制和动态路由机制，使其能够用于视频去雪任务，并通过复用上一帧结果和动态早退出进一步减少推理开销。我们在 RVSD 数据集上进行了训练和测试，实验结果表明我们的方法确实能够实现帧种类的感知并依此激活不同的专家，且在几乎不影响性能的同时使激活专家数量减半。

关键词：视频恢复；多任务学习；混合专家模型

1 引言

图像退化在影响视觉质量的同时，也往往会降低下游视觉任务如目标检测和语义分割的准确性，致力于纠正这些退化的图像恢复方法在过去几十年已经得到了长足的发展。其中，图像去除天气方法专注于去除图像中天气效果（如雨、雪、雾）的去除。尽管单张图像的恢复方法已经取得了先进的性能，视频的恢复方法仍然存在计算开销大等问题。本文主要聚焦于实现一个高效的基于深度神经网络的视频去雪方法。

现有的图像去除天气方法大多专注于去除特定类型的天气效果，这使得在面对真实世界场景中存在多种类型的天气效果时，往往需要同时部署多个特化模型以实现问题类型的覆盖。这一方面极大提高了部署成本，另一方面引入了模型之间的切换开销。有的研究试图用将模型在不同数据集上训练的方法来解决上述问题，但固定的网络架构没有利用退化种类不同的先验，与退化种类特定知识的多样性相悖。

为了解决上述问题，研究者们提出了多合一图像去除天气方法，致力于通过网络架构的特殊设计（如基于提示词的方法、混合专家模型、多模态模型），用一个模型进行图片中多种天气效果的高效率高质量去除。最近，Zhang 等人 [15] 通过使用特征调制层和不确定性感知路由分别替换混合专家模型中的前馈网络层和线性路由，提出了混合特征调制专家模型，利用稀疏激活和专家组合实现子网络的复用，在提升效果的同时减少模型参数量。其高效率高质量的特点促使我们将其作为骨干网络，以进一步实现一个用于视频任务的模型。

我们首先复现了 Zhang 等人 [15] 提出的混合特征调制专家模型，其原理是利用混合专家模型的条件计算范式，根据输入选择性地激活网络结构中的特定部分即特定任务对应的专家网络。整个网络在推理时处于稀疏激活的状态，一方面减少了计算开销，另一方面避免了不同任务之间的互相干扰。其中的专家层是特征线性调制层，可以根据输入特征先对其进行不同的仿射变换，这一过程使得后续网络具有更大的参数共享可能，进一步提升了模型的性能和效率。同时，不确定性感知路由使用蒙特卡罗随机失活算法对路由的输出进行校正，提高了路由层输出的置信度。

我们希望将混合特征调制专家模型在图像任务上的优势迁移到视频任务中，最朴素的做法是对视频中的每一帧都应用独立的推理过程。由于一段视频序列中相邻帧之间的变化通常很小，所以视频帧在时序上具有巨大的冗余，对每一帧进行独立的处理可想而知会造成极大的不必要的额外开销。因此我们决定增加参考帧机制，对于视频中第一帧之后的其他帧，在输入当前帧退化图像的同时，也输入上一帧的重建结果，以期降低当前帧的恢复难度。在观察不同类型帧（没有参考帧的第一帧和有参考帧的其他帧）激活专家的规律时，我们发现不同种类的帧激活的专家有同有异，并且其他帧分配给各专家的权重更集中于少数几个专家。这表明我们的方法实现了对帧类型的感知，同时虽然其他帧和第一帧存在有无参考帧的差别造成激活专家不同，但相同的图像处理任务保证了激活专家上仍然有相当的重叠，并且有参考帧的其他帧恢复难度较小，在分配专家时更为自信。总结来说，本文的工作主要有以下几点：

（1）在 Allweather 数据集上复现了原论文进行 All-in-One Image Deweather 任务的结果，复现结果与原文数据基本吻合；

（2）增加了参考帧机制，在 RVSD 数据集上进行了 Video Desnow 任务，使用简单的设计达到了不错的效果，进一步证明了 MoFME 模型的有效性；

（3）基于对不同种类帧激活各专家概率的统计规律的观察，增加了动态路由机制，进一步挖掘了利用混合专家模型稀疏性的潜力。

本文剩余的章节将按以下逻辑展开，在第二章介绍与本文相关的工作，第三章详细介绍本文的方法，第四章介绍复现的细节以及本文创新点，第五章进行实验并分析结果，最后总结本文的结论和展望。

2 相关工作

2.1 多合一图像恢复

现有的多合一图像恢复方法，作为多任务学习的一种形式，采用各种架构设计来处理跨多个任务的输入和输出，实现了它们之间的有效信息共享。Li 等人 [7] 提出了一种多头共享解码器的方法，用共享的解码器权重处理多种天气退化，而多头对每种退化类型独立训练。Han 等人 [3] 提出的 BIDeN 则采用了多解码器共享主干网络的设计，用一个没有任务特定指示器的统一模型接收混合的输入，使用一个共享的主干网络进行特征提取。然而，这种方法又引入了多解码器的复杂度，并且需要退化类型标签的额外监督学习。Valanarasu 等人 [11] 提出的 TransWeather 重新采用了统一的编码器解码器架构，属于基于提示词的模型，使用提示词携带退化种类相关的信息。Chen 等人 [1] 提出的 IPT 引入了具有任务特定的头尾网络的可复用的预训练 transformer 主干网络，使用先验知识显著简化了网络结构。Yu 等人 [14] 提出的 MEASNet 使用了混合专家模型，输入通过一个门控机制被路由给不同的专家网络。通过像

素级和全局特征级的考虑——包括低频和高频信息——选择恰当的专家网络进行图像恢复。Zhang 等人 [16] 提出的 Perceive-IR 利用冻结的视觉语言模型获取视觉与语言之间对齐的语义，然后辅助对高维特征嵌入的预测从而增强恢复过程。Potapalli 等人 [9] 提出的 promptIR 使用了单对编解码器的架构，但在多个解码阶段插入可学习的视觉提示词来隐式地预测退化信息。提示词引导解码器选择性地恢复多种退化图像，表现为只带来很少额外参数的轻量级的即插即用的模块。进一步扩展提示词的内涵，Yan 等人 [12] 提出的方法集成了语义和多模态的提示词，可以使用自然语音指令或结合视觉和语义信息引导恢复过程，增强了对位置退化种类的适应性。最后，promptGIP [8] 使用了问答范式，使得用户可以根据自己的偏好通过插入用户输入和据此调整恢复过程来自定义图像恢复。

2.2 视频去雪

在基于深度学习的方法取得成功前，视频去雪任务使用的是传统的计算机视觉技术。Ren 等人 [10] 利用背景的低秩假设来分离稀疏和密集的雪，并对动态场景中的大雪进行处理。Kim 等人 [4] 在他们的低秩矩阵补全雪花去除算法中考虑了全局和局部运动，以及雪花的不同尺寸。由于深度学习模型的快速发展和它们学习复杂模式的能力，一些研究者提出了用于视频去雪的深度学习方法。[13] 利用自适应雪侦测和基于分块的高斯混合模型，在去除视频中的稀疏或密集的雪上表现良好。Li 等人 [5, 6] 提出了一种用于动态背景的方法，使用在线多尺度卷积稀疏编码模型将雪编码并去除。

3 本文方法

3.1 混合专家模型

模型规模是提升模型性能的关键因素之一。在有限的计算资源预算下，用更少的训练步数训练一个更大的模型，往往比用更多的步数训练一个较小的模型效果更佳。混合专家模型（Mixture-of-Experts, MoE）的一个显著优势是它们能够在远少于稠密模型所需的计算资源下进行有效的预训练。这意味着在相同的计算预算条件下，可以显著扩大模型或数据集的规模。特别是在预训练阶段，与稠密模型相比，混合专家模型通常能够更快地达到相同的质量水平。作为一种基于 Transformer 架构的模型，混合专家模型主要由两个关键部分组成：1）稀疏 MoE 层：这些层代替了传统 Transformer 模型中的前馈网络（Feedforward Network, FFN）层。MoE 层包含若干“专家”（例如 8 个），每个专家本身是一个独立的神经网络。在实际应用中，这些专家通常是前馈网络（FFN），但它们也可以是更复杂的网络结构，甚至可以是 MoE 层本身，从而形成层级式的 MoE 结构。2）门控网络或路由：这个部分用于决定哪些令牌（token）被发送到哪个专家。例如，在下图中，“More”这个令牌可能被发送到第二个专家，而“Parameters”这个令牌被发送到第一个专家。有时，一个令牌甚至可以被发送到多个专家。令牌的路由方式是 MoE 使用中的一个关键点，因为路由器由学习的参数组成，并且与网络的其他部分一同进行预训练。总结来说，在混合专家模型（MoE）中，传统 Transformer 模型中的每个前馈网络（FFN）层被替换为 MoE 层，其中 MoE 层由两个核心部分组成：一个门控网络和若干数量的专家。

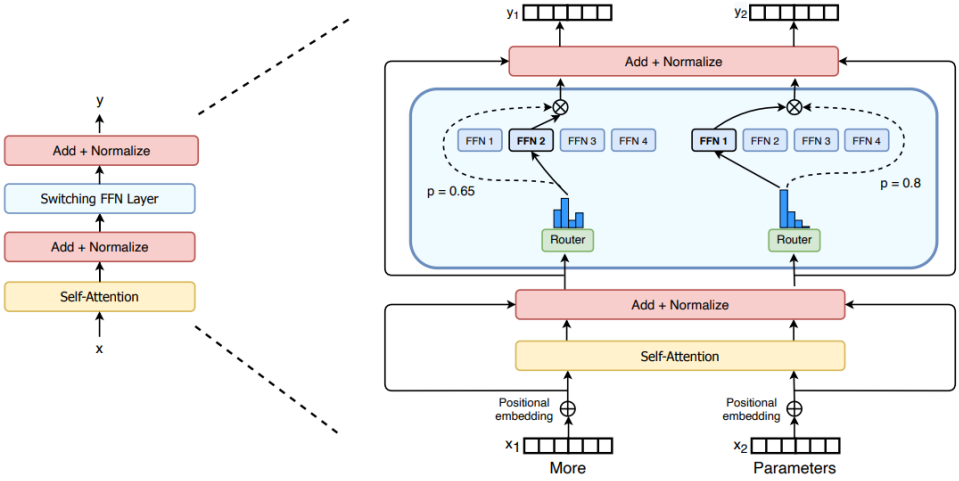

Figure 2: Illustration of a Switch Transformer encoder block. We replace the dense feed forward network (FFN) layer present in the Transformer with a sparse Switch FFN layer (light blue). The layer operates independently on the tokens in the sequence. We diagram two tokens (x_1 = "More" and x_2 = "Parameters" below) being routed (solid lines) across four FFN experts, where the router independently routes each token. The switch FFN layer returns the output of the selected FFN multiplied by the router gate value (dotted-line).

图 1. Switch Transformer paper [2] 中的 MoE Layer

3.2 特征调制专家

在典型的基于 Vision Transformer（ViT）架构的混合专家模型中，每个 transformer 块中的稠密前馈网络层被替换为 MoE 层，这些 MoE 层内部有许多平行的专家网络，一般为前馈网络。Zhang 等人 [15] 将这些作为专家网络的前馈网络替换为特征调制网络。具体来说，具有多样性的任务特定的特征，即令牌，首先被送入一个动态特征调制单元，在这里被路由根据输入导向不同的可学习的仿射变换。如此调制过后的特征再通过一个简单的共享前馈网络专家融合。通过这种方式，Zhang 等人 [15] 将 MoE 架构中的每个专家隐式地表示为一个轻量级的仿射变换和一个共享 FFN 的级联模块，从而显著减少了添加额外专家的参数和计算开销。

$$\gamma = g(x) \quad \beta = b(x),$$

依据输入 x 获得特征调制参数 γ 和 β，g 和 b 可以是任意的可学习函数

$$FM(x) = \gamma \circ x + \beta,$$

用获得的调制参数对输入进行调制得到单个特征调制专家的输出，其中 $\circ$ 为哈达玛积

$$FME(x|\gamma, \beta) = FFN\{\sum_i r_i(x) \cdot [\gamma^{(i)} \circ x + \beta^{(i)}]\}$$

对所有专家的输出进行加权求和后，得到整个特征调制专家层的输出。

通过多样化的特征调制，单个共享的前馈网络模块就能够处理多任务特征的混合。

3.3 不确定性感知路由

为了提高特征调制专家的性能，Zhang 等人 [15] 提出了不确定性感知路由，使用蒙特卡罗随机失活算法隐式地估计路由权重的不确定性。模型的不确定性反映了模型是否知道它知道什么。尽管目前存在的基于集成学习的不确定性估计方法往往能够达到最优的校准和预测准确度，但其高昂的计算复杂度和存储开销促使 Zhang 等人 [15] 使用更高效的蒙特卡罗随机失活算法。

具体来说，Zhang 等人 [15] 将一个特定的路由的输出视为符合高斯分布以校准其不确定性。这样一个分布的均值和方差可以通过路由集成来估计，即根据蒙特卡罗随机失活算法将一个令牌多次传递给路由得到对应的输出。然后根据下式对路由的输出进行校准和正则化：

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

其中均值 μ 和方差的相反数 Σ^{-1} 在计算中被构建为零填充的对角矩阵。

4 复现细节

4.1 与已有开源代码对比

本工作复现了 Zhang 等人 [15] 在论文《Efficient Deweather Mixture-of-Experts with Uncertainty-Aware Feature-Wise Linear Modulation》中所提出的混合特征调制专家模型。官方的代码实现已在 github 上开源。

因此本工作在复现原有代码的基础上增加了原创的参考帧机制和动态路由机制，并在 RVSD 数据集上重新进行了训练和测试。

表 1. 代码文件版权说明

文件/类	版权	备注
RVSD 类	自主实现	数据处理与参考帧机制
DynamicTopK 类	自主实现	动态路由机制
infer_video.py 文件	自主实现	视频推理与参考帧机制
其它	源自MoFME-Pytorch	原文官方开源代码

4.2 实验环境搭建

本次复现工作的实验平台为 8×NVIDIA RTX 4090 GPUs，使用大小为 64 的 batch size 训练 200 个 epoch，AdamW optimizer 和 Cosine LR scheduler 的初始学习率设置为 0.5×10^{-4} 并逐步衰减到 10^{-6}，使用三个 epoch 作为三个 epoch 作为热身阶段，以上超参数与原文一致。

4.3 创新点

本工作除了复现原文提出的混合特征调制专家模型在 All-in-One Image Deweather 任务上的结果外，也做了一些自己的增量工作，具体如下：

（1）增加参考帧机制，如图2所示，每一帧以其上一帧为参考帧，恢复过程中复用上一帧结果以利用视频序列在时序上的冗余；第一帧没有参考帧，为保持输入一致性，以形状相同的全零矩阵作为其参考帧。这一改进使得原本用于图片任务的模型能够用于视频任务，同时避免引入过多冗余运算，结果表明增加了参考帧机制的模型具有良好的视频去雪性能，且实现了帧种类的感知，有无参考帧的帧被分配到的专家权重有明显的区别；

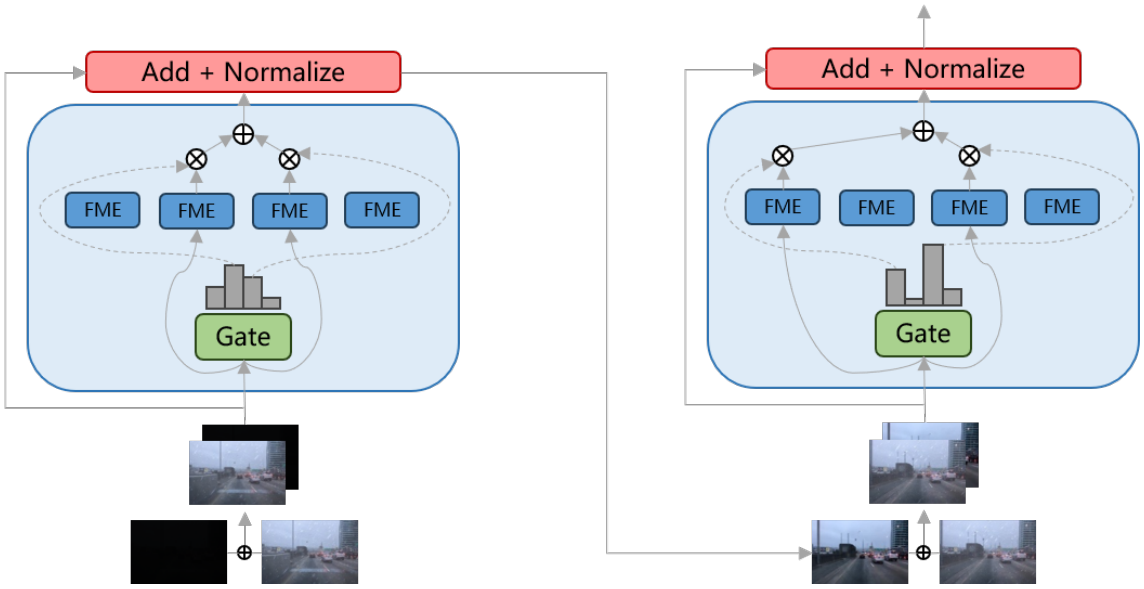

图 2. 参考帧机制

（2）增加动态路由机制（弹性专家选择机制），激活专家数量不再是固定的 k 个，而是设置一个概率阈值 p，当按路由分配的权重从高到低依次选择激活专家时概率和超过阈值后就不再激活其他专家。在本次工作中，p 被经验地设为 0.6，此时模型在取得与 top4 策略（激活权重最高的 4 个专家）相当的性能表现的同时，将平均激活专家数量降低了约一半。

5 实验结果分析

本部分对实验所得结果进行分析，详细对实验内容进行说明，实验结果进行描述并分析。

5.1 复现原论文结果

本工作首先使用上文阐述的实验平台对混合特征调制专家模型在 Allweather 数据集上进行了多合一图像去天气任务的复现，复现结果与原文结果对比如表2所示。复现结果中，模型参数量与原文一致，性能指标比原论文略差 0.03-0.1dB，与原文作者声称的 0.1-0.2dB 的性能提升相比尚在可接受范围之内，产生误差的原因可能是实验平台不同以及随机种子的设置。不过复现结果依然初步证明了混合特征调制专家模型完成 All-in-One 任务的能力。

表 2. 复现结果与原文结果对比

	参数量	去雨		去雨滴		去雪		平均	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
原文结果	21.22M	28.66	0.9436	29.27	0.9385	29.35	0.8996	29.09	0.9272
复现结果	21.22M	28.63	0.9417	28.97	0.9348	29.29	0.9005	28.96	0.9257

5.2 模型对帧种类的感知

引入参考帧机制后，在训练周期较少的训练前期阶段，第一帧的重建帧亮度偏低，原因应该是第一帧的参考帧是一个全黑的图片，而模型在训练前期尚未学习到如何对不同类型的帧区别对待，因此使得第一帧不当地过多参考了这一全黑图片。当训练轮次足够后，这一问题不再存在，第一帧和其他帧的恢复效果基本一致，这种变化让我认为模型已经习得了如何对不同种类的帧区别对待，因此我对路由分配给第一帧和其他帧的专家权重的统计规律进行观察，如表3所示，我们发现：一方面，第一帧和其他帧分配到的权重最高的几个专家存在重合；另一方面，在这些专家内部权重的分配也存在差异，第一帧分配到的权重最高的专家在其他帧那里显得 “默默无闻”。这说明，模型确实能够根据输入帧的类型分配对应的专家组合，即实现了帧种类的感知。

表 3. 路由分配给不同帧的专家权重分布

专家编号	3	4	5	6	9	10	14
第一帧	0.2581	0.0611	0.1744	0.0005	0.0839	0.0808	0.3092
其他帧	0.3894	0.0094	0.2348	0.0642	0.1203	0.0227	0.0779

5.3 动态路由实现弹性专家选择

将不同种类帧分配权重前八名的专家按权重从高到低的顺序排列后，进一步发现，有参考帧的非第一帧激活专家的概率集中在少数几个专家上，由于特征调制专家层最后的输出是各专家输出加权求和的结果，因此很多大部分专家对于输出的贡献很小。而无参考帧的第一帧激活专家的概率分布就比较平滑，因此更多专家的输出得到了利用。这个现象可以这样解释，无参考帧的帧恢复起来难度较大，因此需要更大规模的子网络。对于难度不同的任务，按照混合专家模型原来的激活固定数量专家的策略就显得不太合理，因此我们实现了动态路由机制，使得模型可以根据任务难度主要是帧种类激活数量不同的专家网络。

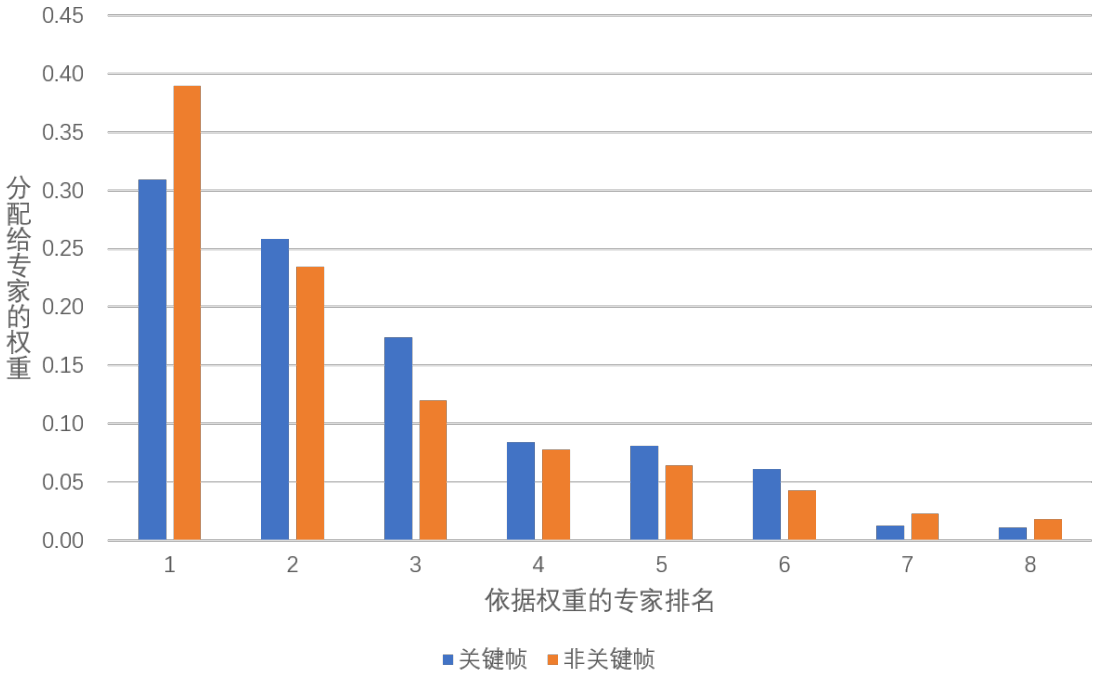

图 3. 不同种类帧分配权重前八名的专家按权重排名

6 总结与展望

在本工作中，我们复现了论文《Efficient Deweather Mixture-of-Experts with Uncertainty-Aware Feature-Wise Linear Modulation》所提出的 MoFME 模型进行多合一图像去除天气任务的结果，复现结果与原文数据基本吻合；此外，我们增加了两个自己的改进之处：1）增加了参考帧机制，在 RVSD 数据集上进行了 Video Desnow 任务，使用简单的设计达到了不错的效果，进一步证明了 MoFME 模型的有效性；2）基于对不同种类帧激活各专家概率的统计规律的观察，增加了弹性专家选择机制：进一步挖掘了利用混合专家模型稀疏性的潜力。

然而，由于时间有限，一方面参考帧机制的设计十分朴素，另一方面只在视频去雪任务上进行了实验。因此参考帧机制的设计还有很大的改进之处，如可以尝试利用光流估计网络；模型在视频去雨、去雾等其它天气条件下的性能也有待评估。

参考文献

[1] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12299–12310, June 2021.

[2] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: scaling to trillion parameter models with simple and efficient sparsity. *J. Mach. Learn. Res.*, 23(1), January 2022.

[3] Junlin Han, Weihao Li, Pengfei Fang, Chunyi Sun, Jie Hong, Mohammad Ali Armin, Lars Petersson, and Hongdong Li. Blind image decomposition. In *Computer Vision –ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVIII*, page 218–237, Berlin, Heidelberg, 2022. Springer-Verlag.

[4] Jin-Hwan Kim, Jae-Young Sim, and Chang-Su Kim. Video deraining and desnowing using temporal correlation and low-rank matrix completion. *IEEE Transactions on Image Processing*, 24(9):2658–2670, 2015.

[5] Minghan Li, Xiangyong Cao, Qian Zhao, Lei Zhang, Chenqiang Gao, and Deyu Meng. Video rain/snow removal by transformed online multiscale convolutional sparse coding, 2019.

[6] Minghan Li, Xiangyong Cao, Qian Zhao, Lei Zhang, and Deyu Meng. Online rain/snow removal from surveillance videos. *IEEE Transactions on Image Processing*, 30:2029–2044, 2021.

[7] Ruoteng Li, Robby T. Tan, and Loong-Fah Cheong. All in one bad weather removal using architectural search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[8] Yihao Liu, Xiangyu Chen, Xianzheng Ma, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Unifying image processing as visual prompting question answering, 2023.

[9] Vaishnav Potlapalli, Syed Waqas Zamir, Salman Khan, and Fahad Khan. Promptir: Prompting for all-in-one image restoration. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[10] Weihong Ren, Jiandong Tian, Zhi Han, Antoni Chan, and Yandong Tang. Video desnowing and deraining based on matrix decomposition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2838–2847, 2017.

[11] Jeya Maria Jose Valanarasu, Rajeev Yasarla, and Vishal M. Patel. Transweather: Transformer-based restoration of images degraded by adverse weather conditions, 2021.

[12] Qiuhai Yan, Aiwen Jiang, Kang Chen, Long Peng, Qiaosi Yi, and Chunjie Zhang. Textual prompt guided image restoration, 2023.

[13] Bin Yang, Zhenhong Jia, Jie Yang, and Nikola K. Kasabov. Video snow removal based on self-adaptation snow detection and patch-based gaussian mixture model. *IEEE Access*, 8:160188–160201, 2020.

[14] Xiaoyan Yu, Shen Zhou, Huafeng Li, and Liehuang Zhu. Multi-expert adaptive selection: Task-balancing for all-in-one image restoration, 2024.

[15] Rongyu Zhang, Yulin Luo, Jiaming Liu, Huanrui Yang, Zhen Dong, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, Yuan Du, and Shanghang Zhang. Efficient Deweather Mixture-of-Experts with Uncertainty-Aware Feature-Wise Linear Modulation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(15):16812–16820, March 2024.

[16] Xu Zhang, Jiaqi Ma, Guoli Wang, Qian Zhang, Huan Zhang, and Lefei Zhang. Perceive-ir: Learning to perceive degradation better for all-in-one image restoration, 2024.