RAM：基于掩码图像建模的一体化盲图像修复方法

Chu-Jie Qin[1,2*], Rui-Qi Wu[1,2], Zikun Liu[3], Xin Lin[5],
Chun-Le Guo[1,2], Hyun Hee Park[4], and Chongyi Li[1,2†]

[1] VCIP, CS, Nankai University
[2] NKIARI, Shenzhen Futian
{chujie.qin,wuruiqi}@mail.nankai.edu.cn
{guochunle, lichongyi}@nankai.edu.cn
[3] Samsung Research, China, Beijing (SRC-B)
[4] The Department of Camera Innovation Group, Samsung Electronics
{zikun.liu,inextg.park}@samsung.com
[5] Sichuan University
linxin@stu.scu.edu.cn

Abstract. 一体化图像复原旨在使用一个模型处理多种退化类型。本文提出了一种简单的一体化盲图像复原方法**R**estore **A**nything with **M**asks (**RAM**)，简称RAM。我们通过使用掩码图像建模来提取内在的图像信息，专注于图像内容，而不是像其他方法那样区分退化类型。我们的方法包括两个阶段：掩码图像预训练和掩码归因通量的微调。专门针对一体化图像恢复，我们设计了一种简单的掩码预训练方法。这种方法增强了网络从各种退化中优先提取图像内容先验的能力，从而在不同的恢复任务中取得更平衡的表现，并实现更好的整体效果。为了在尽可能保留已学习的图像先验的同时弥合输入完整性的差距，我们选择性地对一小部分层进行微调。具体来说，每一层的重要性由我们提出的掩码归因通量模块（**MAC**）进行排序，贡献较高的层被选中进行微调。大量实验表明，我们的方法达到了最高的性能。我们的代码和模型将开源在：https://github.com/Dragonisss/RAM.

Keywords: 图像复原 · 一体化方法 · 掩码图像建模

1 Introduction

图像复原涉及修复受到各种退化影响的低质量图像，这些退化通常源于不利的环境条件（如雨天、雾霾、低光）、与硬件相关的问题（如噪声和模糊）以及后处理伪影（如JPEG压缩）。图像复原不仅有助于增强图像的视觉效果，还在自动驾驶和监控等实际应用场景中发挥重要作用。

在图像复原领域，目前方法主要集中于学习退化过程中的固定模式，即退化先验。一些研究 [28, 29, 60]利用特定任务的先验来解决某种退化问题，而另一类研究 [3, 27, 38, 47, 55]则尝试设计一种通用的网络架构，可以有效地学习每种退化模式。然而，以上方法仅能使网络学习单一退化，当处理多种类型的退化时会导致不平衡的情况。

*A part of this work is done during Chu-Jie Qin's internship at Samsung.
†Chongyi Li is the corresponding author.

为了解决上述问题，一体化图像复原方法应运而生，旨在使用一个模型处理多种退化。大多数这些方法倾向于利用显式先验（如AirNet [20]）或引入额外的模块（如PromptIR [41]）来识别图像退化模式，从而协助模型进行复原。然而，这些方法将重点放在区分图像中的退化类型上，而不是图像内容本身，这导致了当涉及更多退化类型时，模型的可扩展性降低和决策边界模糊化。我们认为，图像复原的本质在于从损坏的图像中提取内在的图像信息，而不是消除退化模式，即学习图像先验而不是退化先验。值得注意的是，TAPE [30]同样认为通过引入自然图像先验来理解正常图像的性质，从而达到辅助图像复原的目的。然而，TAPE将模型输出作为优化目标，这会导致模型放大其自身的误差，并带有偏差地学习图像先验。

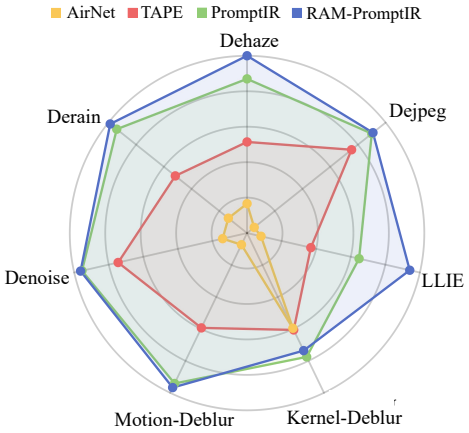


Fig. 1: 我们提出的RAM在一体化盲图像复原方面，对比最新的其他方法(AirNet, TAPE, PromptIR)具有更平衡和更强的表现。

在本文中，我们专注于解决**如何从多样化的损坏图像中提取内在的图像信息**。一些尝试 [2, 6]通过在底层视觉任务中应用掩码图像建模（MIM）引起了我们的注意。作为一种预训练策略，MIM由于其通用的图像表示形式，在高级任务中被广泛验证其有效性。同时，模型也学习了自然图像的分布，这包含了我们希望从图像中提取的内在信息。基于MIM，我们提出了一个简单的一体化盲图像复原方法，即**RAM**，本方法包括两个阶段：掩码预训练阶段和掩码归因通量（MAC）微调阶段。在预训练阶段，我们在像素级别随机掩盖损坏的图像，并强制网络预测与掩码像素对应的清晰图像，从损坏的图像中提取固有的图像信息。在微调阶段，我们专注于克服预训练期间由于改变掩码输入到整个图像推理所导致的输入完整性差距，同时尽可能多地保留已学习的先验。

具体来说，我们首先通过提出的MAC方法评估了每个网络层在解决这一差距中的重要性。接着，我们选择了最关键的前k%层进行微调，同时保持其余网络层冻结。我们证明，在经过简短的微调后（即使仅微调10%的层），模型即可达到令人满意的表现，超越了使用传统成对训练方法训练的模型。此外，我们的流程可以在任何网络中以即插即用的方式使用，不会引入额外的计算开销。

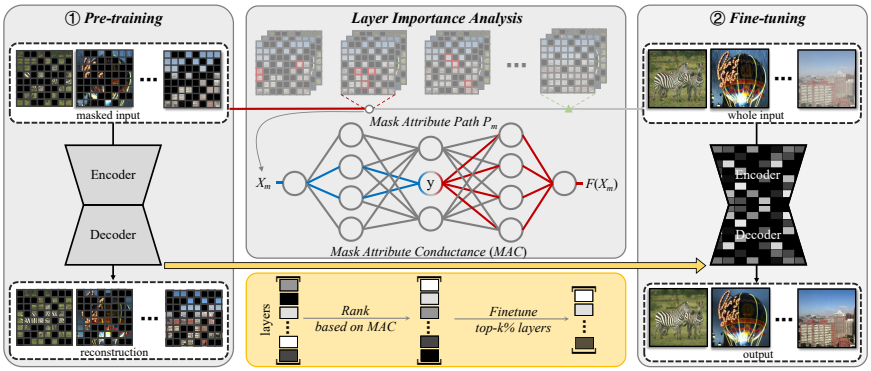


Fig. 2: 我们的方法。1) 使用适用于底层视觉的掩码图像预训练方法进行模型预训练。我们以50%的掩码比例随机掩盖退化图像的像素，并重建清晰图像。2) 接着进行微调阶段，以克服预训练期间掩码输入变为推理期间整个图像所导致的输入完整性差距。我们根据提出的MAC分析每个网络层在解决输入完整性差距中的重要性，并按降序对其进行排序。选择前k%的重要网络层在完整图像上进行微调。

本工作的贡献如下：

- 我们讨论了在底层视觉任务中采用MIM的挑战，并提出了一种基于MIM的预训练策略，该策略专为一体化盲图像复原而设计，使复原网络能够在保证重建结果的同时有效学习固有的图像信息。
- 我们提出了掩码归因通量模块（MAC），用于评估每个层在解决输入完整性差距中的重要性，从而只需微调一小部分（例如10%）关键层，就可以弥合这一差距，同时保留MIM学习到的图像先验。
- 我们提出的RAM为实现更平衡、更强大的一体化盲图像复原提供了一个全新的思路，其重点是从损坏的图像中提取固有的图像信息。我们的方法可以应用于任何图像复原网络，而不会引入额外的计算开销。

2 Related Work

2.1 多重退化图像复原

虽然神经网络在单一退化图像复原任务中表现出色 [8, 12, 13, 16, 21, 22, 28, 29, 49, 60]，但最近的工作转向了更具挑战性的多重退化图像复原领域。一些方法 [3, 27, 38, 47, 55]旨在设计一种通用架构，可以有效地学习每种退化模式。SwinIR [27]采用窗口注意力机制，将全局注意力转化为局部化的方法，有效地减少了计算开销。此外，基于U形 Transformer的方法 [47, 55]用于提取多尺度特征并减少计算开销。然而，这些方法必须在每个复原任务上单独训练。几种方法 [1, 23]利用多个输入和输出头来增强网络复原各种类型退化图像的能力，但这种方法可能导致模型的可扩展性降低。最近，一些后续方法 [4, 20, 35, 40, 41, 56, 61]提出使用统一的网络来解决多个恢复问题。这些方法大多强调学习如何区分不同类型的退化并复原损坏的图像。AirNet [20]首次提出了一体化图像复原任务，该方法最初基于对比学习预训练了一个退化分类器，

并随后利用其辅助一体化图像复原。PromptIR [41]引入了一个可学习的提示模块，不再限定退化类别，而是通过自适应提示使模型自主学习有利于算法高效运作的特征。我们的RAM采取了一个新的思路，专注于从损坏图像中提取共同的内容信息，而无需额外的设计来区分退化，这使得我们在处理涉及更多退化类型时实现了更优秀的性能。

2.2 掩码图像建模

受掩码语言建模 [17, 42]的启发，掩码图像建模（MIM） [14, 51]被引入作为一种预训练方法，用于学习高层视觉中的通用表示。MAE [14]有效利用MIM预测隐藏的tokens，展示了其在各种下游任务中的强大性能和泛化能力。SimMIM [51]提出了一种基于Swin-ViT [33]的通用掩码图像建模方法。Painter [46]将多个任务统一到图像到图像的翻译中，并利用MIM进行预训练。近年来，一些方法将MIM引入到底层视觉领域，以增强模型的泛化能力。其中， [2]和 [6]与我们的研究方向最为接近。 [2]使用MIM模型来增强去噪任务的模型泛化能力，但没有探索其在多任务场景中的潜力。 [6]使用MIM对模型编码器进行预训练，以引入生成先验，随后使用解码器进行恢复。然而，它并未充分利用MIM的潜力。我们提出的RAM使用MIM将各种图像恢复任务的优化目标统一为重建内在的图像信息。这使得网络能够更平衡和有效地学习恢复功能。此外，为了保留MIM学习到的图像先验，我们设计了一种基于MAC分析的微调策略（见 Sec. 3.3）。这使得我们仅通过微调一小部分（例如10%）层，即可实现可比的性能，充分挖掘MIM的潜力。

2.3 基于梯度的归因方法

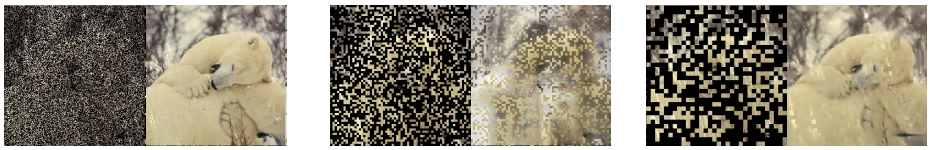
基于梯度的归因方法 [5, 11, 43–45, 50]常用于阐明隐藏单元（或输入）如何影响网络输出。一种常用的方法是集成梯度（IG） [44, 45]，它沿着从基线输入到目标输入的线性路径在像素/特征空间中累积梯度。之后，IntInf [18]和层通量 [5]修改IG以沿同一路径归因于神经元的重要性。在我们的工作中，我们希望找到能够有效克服训练数据和推理数据之间分布转移的关键层。我们基于层通量提出了掩码归因通量模块（MAC），并沿掩码特性路径（MAP）累积每层的MAC。MAC可以表示每层沿MAP的重要性。通过这种方式，我们可以微调预训练网络中最关键的$k\%$层，在很大程度上保留预训练过程中学习到的图像先验。

3 Methodology

在本节中，我们首先讨论在底层视觉任务中使用MIM的挑战（Sec. 3.1）。然后，我们介绍了一种一体化盲图像复原的流程，该流程包含两个部分：使用MIM进行预训练（Sec. 3.2）以及使用掩码归因通量（MAC）进行微调（Sec. 3.3）。

3.1 重新思考底层视觉中的MIM

MIM是一种随机遮蔽图像的某些部分并从剩余可见部分中提取特征来重建整个图像的过程。它使模型能够获得图像的通用表示，从而在许多高级任务中实现



patch size=1 patch size=4 patch size=8

Fig. 3: 具有不同块大小的掩码图像建模重建。我们使用不同的块大小进行预训练，并可视化掩码输入（左）及其对应的MIM重建（右）。

良好的预训练效果 [14, 51]。此外，模型在图像重建过程中还学习到了自然图像的分布，即MIM预训练。这种附带获得的先验知识在图像复原等任务中非常重要。

尽管有这些优势，但将MIM应用于底层视觉任务的预训练模型中仍然研究不足，这主要因为在这个过程存在需要解决的挑战。

首先，原始的MIM的主要目的不是高质量重建，而是为下游任务提取良好的特征。因此，它对更大范围的图像进行遮盖以收集语义信息，而不是像素级内容，这反映在Token级遮盖和高遮盖比例上。CSFormer [6]直接采用了这种用于底层视觉预训练的策略。然而，一些研究表明，语义信息对图像复原的重要性不如它在模式识别任务中的重要性 [32, 36]。此外，高度遮盖会产生细节缺失的结果，如Fig. 3所示，这对底层任务是不利的。

其次，MIM的训练目标是重建被遮盖的输入图像，因此它只能生成与输入图像相同域的结果。然而，我们希望模型获得跨越低质量域到高质量域的能力，即从退化的输入中复原干净的内容。因此，在使用MIM预训练图像复原模型时，有必要引入成对数据（有关详细信息，请参阅 Sec. 4.3中的实验）。Chen等人 [2]表明，成对的MIM训练增强了对不同类型噪声图像的泛化性能。在本文中，我们进一步探索了MIM在更大方差多重退化上的有效性。

3.2 MIM的预训练

基于以上分析，我们设计了一个适用于底层视觉的MIM预训练范式。

Masking 在预训练阶段，我们随机遮盖退化图像的像素（以1×1块大小遮盖图像），遮盖比例为50%。我们发现，细粒度的遮盖块和均衡的遮盖比例对图像复原有利，这在Sec. 4.3中有体现。

此外，由于我们的MIM预训练目标与后续的底层任务相似，我们不需要像MAE [14]那样更改解码器，只需微调它即可。

重建目标 按照Bert [17]和MAE [14]的思路，我们选择使用L1损失来监督被遮盖部分。训练目标可以写为：

$$\arg\min_{\theta} \mathbb{E}[||\tilde{\mathcal{M}}(I - f(\mathcal{M}(I_d), \theta))||], \tag{1}$$

其中$\{I, I_d\}$表示一对干净图像和退化图像，$f(\cdot, \theta)$表示带有参数θ的网络，$\mathcal{M}(\cdot)$是一个随机的二元遮盖操作，$\tilde{\mathcal{M}}(\cdot) = 1 - \mathcal{M}(\cdot)$。

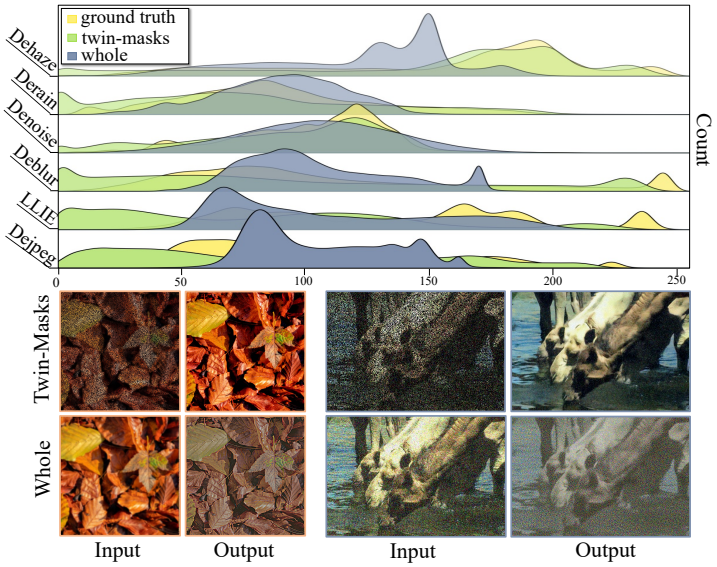


Fig. 4: 在核去模糊（橙色边框）和去噪（蓝色边框）上的不同输入完整性对MIM重建的影响。我们还可视化了各种任务中重建结果的颜色分布。结果表明，使用双重掩码方法作为输入获得的重建结果的分布比使用整体输入获得的结果更接近真实图像（GT）。

3.3 使用掩码归因通量分析的微调

观察 在预训练过程中，网络学习到丰富的内容先验。然而，掩码输入的不完整性阻止了预训练模型的直接推理使用，因为这会导致输出中的分布偏移。如Fig. 4所示，我们首先将整个图像输入到预训练模型中，导致颜色失真的结果。接着，我们使用一对互补的掩码（称为双重掩码）来单独遮盖图像。随后，我们将这两个互补遮盖的图像输入网络。通过组合每个图像预测的像素值，我们生成一个更高质量的图像。这一观察表明，阻碍直接使用掩码预训练模型进行推理的原因在于输入不完整性，而不是模型无法学习复原功能。

基于这一洞察，我们探索了通过模型微调来尽量减少数据输入格式差异的影响。为了保持所学先验，必须尽可能多地保留预训练参数，同时使用最少但最有效的层进行微调。为了解决这个问题，我们引入了掩码归因通量（MAC）的概念，用以量化每一层相对于微调目标的重要性。然后，我们识别出最关键的前k%的层进行微调。

初步 在定义掩码归因通量（MAC）之前，我们简要回顾一下积分梯度 [44]（IG）和神经元通量 [5]（Cond）的定义。考虑从基准输入x'到目标输入x的线性路径$\gamma(\alpha) = x' + \alpha(x - x')$，我们可以通过计算其积分梯度来将输出变化$F(x) - F(x')$归因于输入/特征$x_i$（例如像素）的第i个维度，形式如下：

$$\mathrm{IG}_i(x) := (x_i - x'_i) \cdot \int_0^1 \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} \, d\alpha. \tag{2}$$

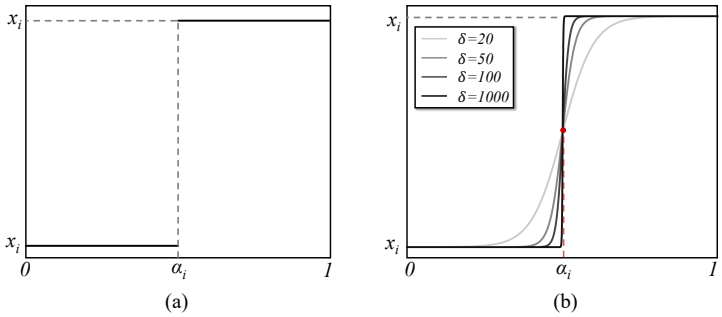


Fig. 5: 在Eq. (5)中(a) X_i^m和在Eq. (6)中(b) $\tilde{X}_i^m$的说明。

我们还可以通过改进IG来将输出变化归因于特定神经元y，其中涉及计算通量。隐藏神经元y沿$\gamma(\alpha)$的通量为:

$$\begin{aligned}\text{Cond}^y(x) &:= \sum_i (x_i - x_i') \cdot \int_0^1 \frac{\partial F(x' + \alpha(x - x'))}{\partial y} \cdot \frac{\partial y}{\partial x_i}\, d\alpha \\ &= \sum_i \int_0^1 \frac{\partial F(\gamma(\alpha))}{\partial y} \cdot \frac{\partial y}{\partial \alpha}\, d\alpha,\end{aligned} \tag{3}$$

注意，$(x_i - x_i') = \frac{\partial(x' + \alpha(x_i - x_i'))}{\partial \alpha}$。当然，我们可以将Eq. (3)扩展为沿任意给定路径$\alpha : [s, t] \to P$积分时计算的通量:

$$\text{GeneralCond}^y(x) := \sum_i \int_P \frac{\partial F(X_i(\alpha))}{\partial y} \cdot \frac{\partial y}{\partial \alpha}\, d\alpha, \tag{4}$$

其中，$X : R \to R^n$是从x'到x路径的函数，其满足$X(s) = x'$，$X(t) = x$。$[s, t]$表示路径函数X的定义域。

使用MAC进行微调。 为了找到有效的微调层，我们提出了**掩码归因通量**（**MAC**），用于评估每一层在克服输入完整性差距方面的效果。考虑一个从零输入x'到整体输入x的非线性路径$\alpha : [0, 1] \to P_m$，该路径函数X^m满足:

$$X_i^m(\alpha; \alpha_i) = \begin{cases} x_i', & \alpha < \alpha_i \\ x_i, & else \end{cases}, \tag{5}$$

其中，i指像素的索引，$\alpha_i \in (0, 1]$是一组参数，指示每个像素何时被遮盖。我们将这条路径定义为掩码特性路径（MAP）。显然，$X^m(0) = x'$ 且 $X^m(1) = x$。

然而，X^m是不可微的，这使得它成为无效的属性路径函数。为了解决这个问题，我们使用一组类似S形的函数$\tilde{X}^m$来逼近X^m:

$$\tilde{X}_i^m(\alpha; \alpha_i) = \frac{(x_i' - x_i)}{1 + e^{-\delta(x_i' - \alpha_i)}}. \tag{6}$$

Table 1: 七个具有挑战性的图像修复任务的定量比较，包括去雾、去雨、去噪、运动去模糊、低光图像增强（LLIE）、卷积核去模糊和JPEG伪影去除。**加粗**和下划线分别表示最佳和次佳结果。

Method	SOTS [19] PSNR↑/SSIM↑	Rain13k-Test [31] PSNR↑/SSIM↑	BSD68 [37] PSNR↑/SSIM↑	GoPro [39] PSNR↑/SSIM↑	LOL [48] PSNR↑/SSIM↑	LSDIR-Blur [25] PSNR↑/SSIM↑	LSDIR-Jpeg [25] PSNR↑/SSIM↑	Average PSNR↑/SSIM↑
Restormer [55]	22.89/0.9172	27.05/0.8469	**30.95**/**0.8657**	27.46/0.8497	23.65/0.8458	19.60/0.3658	**30.46**/**0.9141**	26.01/0.8007
MPRNet [38]	25.23/0.9463	25.36/0.8068	29.83/0.8317	25.90/0.7949	22.29/0.8170	25.68/0.8281	28.96/0.8865	26.18/0.8445
NAFNet [3]	25.74/0.9445	24.65/0.7877	30.37/0.8540	25.53/0.7909	21.50/0.8104	29.08/0.9130	29.09/0.8955	26.57/0.8566
DL [7]	21.16/0.9042	19.56/0.6508	16.15/0.5861	17.63/0.5862	19.26/0.7777	17.98/0.6121	19.55/0.6965	18.75/0.6877
TAPE [30]	25.14/0.9319	23.66/0.7818	30.11/0.8354	25.97/0.7962	18.95/0.7632	24.26/0.7654	29.28/0.8965	25.34/0.8243
AirNet [20]	21.66/0.8366	20.21/0.6402	27.99/0.7250	23.36/0.7503	16.65/0.6708	23.84/0.7358	24.36/0.8020	22.58/0.7372
SwinIR [27]	27.29/0.9622	25.32/0.8258	30.65/0.8540	26.61/0.8125	18.66/0.8048	27.82/0.8839	30.13/0.9071	26.64/0.8643
RAM-SwinIR	28.47/0.9689	26.31/0.8486	30.83/0.8611	26.89/0.8200	21.62/0.8291	26.66/0.8514	30.22/0.9096	27.28/0.8698
PromptIR [41]	28.70/0.9659	27.46/0.8585	30.84/0.8625	27.71/0.8565	21.19/0.8356	**31.01**/**0.9385**	30.30/0.9117	28.17/0.8899
RAM-PromptIR	**29.64**/**0.9695**	**28.47**/**0.8751**	30.86/0.8624	**28.02**/**0.8592**	**24.46**/**0.8581**	29.57/0.9179	30.33/0.9119	**28.76**/**0.8935**

我们可以看到，当δ足够大时，$\tilde{X}^m$与X^m非常接近（如 Fig. 5所示）。对于每个$\tilde{X}_i^m$，当α在α_i的邻域内时，它将急剧变化从x_i'到x_i。

在这里，我们可以给出**MAC**的定义如下：

$$\begin{aligned} \mathrm{MAC}^y(x) &:= \sum_i \int_{P_m} \frac{\partial F(X_i(\alpha))}{\partial y} \cdot \frac{\partial y}{\partial \alpha}\, d\alpha \\ &\approx \sum_i \int_0^1 \frac{\partial F(\tilde{X}_i^m(\alpha;\alpha_i))}{\partial y} \cdot \frac{\partial y}{\partial \alpha}\, d\alpha. \end{aligned} \tag{7}$$

实际上，从一个具有任意遮盖比例r的遮盖输入x_m到整体输入x的部分路径也可用于特性：

$$\mathrm{MAC}_r^y(x) \approx \sum_i \int_{1-r}^1 \frac{\partial F(\tilde{X}_i^m(\alpha;\alpha_i))}{\partial y} \cdot \frac{\partial y}{\partial \alpha}\, d\alpha. \tag{8}$$

在实际操作中，我们使用N步离散化来近似Eq. (8)的积分形式，这遵循了 [43]的做法：

$$\begin{aligned} \mathrm{MAC}_r^y(x) \approx \sum_i \sum_{j=1}^N & \frac{\partial F(\tilde{X}_i^m(\frac{jr}{N};\alpha_i))}{\partial y} \\ & \cdot (F_y(\tilde{X}_i^m(\frac{(j+1)r}{N})) - F_y(\tilde{X}_i^m(\frac{jr}{N}))). \end{aligned} \tag{9}$$

我们计算预训练网络每一层的MAC值，按MAC值从高到低进行排序，并选择前k%的层进行微调。 网络通过预训练权重进行初始化，只有前k%的层将被微调。 更多实现细节可以在补充材料中找到。

4 Experiment

4.1 实验设置

数据集和评估指标。 我们结合来自各种修复任务的数据集形成训练集，遵循 [59] 的方法。对于退化难以合成的高成本任务，我们利用现有的配对数据

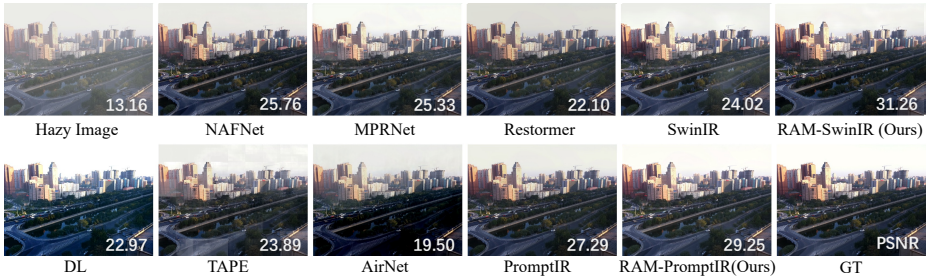


Fig. 6: SOTS数据集上的去雾视觉比较。请放大查看细节。

集，包括用于去雾的RESIDE [19]、用于去雨的Rain13k [9, 24, 26, 34, 53]、用于运动去模糊的GoPro [39]以及用于低光图像增强（LLIE）的LOL-v2 [54]。对于退化易于合成的低成本任务（如噪声、卷积核模糊和JPEG伪影），我们在训练过程中使用LSDIR数据集 [25]生成损坏的图像，这涉及生成具有随机变化 $\sigma \in (0, 50]$ 的高斯噪声，创建使用大小为 $k = 15$ 和随机 $\sigma \in [0.1, 3.1]$ 的模糊核的高斯模糊图像，并引入具有随机质量参数 $q \in [20, 90]$ 的JPEG伪影。

为了评估，我们使用SOTS-outdoor [19]进行去雾评估，使用Rain13k-Test（Rain100L [52]、Rain100H [52]、Test100 [58]、Test1200 [57]和Test2800 [10]的组合）进行去雨评估，使用GoPro进行运动去模糊评估，使用LOL [48]进行低光增强评估，使用BSD68 [37]进行去噪评估，使用LSDIR-val进行卷积核去模糊和JPEG伪影去除评估。此外，我们还进行了不同方差为15、25和50的去噪测试，不同模糊核参数 $k = 15$ 和 $\sigma = 2.0$ 的去模糊测试，以及质量参数为 $q = 50$ 的JPEG伪影去除测试。

实现细节。 我们将提出的**RAM**应用于SwinIR [27]和PromptIR [41]。对于RAM-SwinIR，输入大小为64，而对于RAM-PromptIR，输入大小为128。在预训练阶段，我们使用Adam优化器对RAM-SwinIR和RAM-PromptIR进行训练，持续300个周期，学习率从1e-4开始，按照余弦调度逐渐衰减至6e-5。在微调阶段，我们使用Adam优化器对从RAM-SwinIR和RAM-PromptIR的MAC分析中获得的网络层进行微调，持续40个周期，学习率从2e-4逐渐衰减至1e-7，仍按照余弦调度。RAM-SwinIR和RAM-PromptIR在预训练和微调阶段的批量大小分别为（12,4）和（4,4）。

4.2 对比分析

为了验证RAM的增益能力和有效性，我们将提出的RAM应用于SwinIR（一种通用的图像复原方法）和PromptIR（一种一体化图像复原方法）。我们考虑了四种基于通用架构的图像复原方法 [3,27,38,55]和四种一体化方法 [7,20,30,41]进行对比。我们确保所有其他方法在预训练阶段使用的监督像素数与我们的方法相同。

如 Tab. 1所示，我们的方法在每项任务中都取得了最佳或相当的表现。在七项不同任务的平均得分中，我们使用PromptIR [41]的方法相比第二好的算法获得了0.59dB的性能提升。此外，使用RAM 的SwinIR在PSNR上也提高了2.40%。具体来说，我们的RAM在去雾和低光增强任务中表现出显著优

Table 2: 在BSD68和Urban100数据集上，不同噪声水平下的定量高斯去噪结果（以PSNR衡量）。

Method	BSD68 [37]				Urban100 [15]			
	$\sigma = 15$	$\sigma = 25$	$\sigma = 50$	Average	$\sigma = 15$	$\sigma = 25$	$\sigma = 50$	Average
NAFNet [3]	33.22	30.59	27.30	30.37	32.67	30.21	26.97	29.92
MPRNet [38]	32.73	30.11	26.65	29.83	32.06	29.46	25.77	29.10
Restormer [55]	33.79	31.17	27.90	30.95	33.83	31.40	27.99	31.07
DL [7]	16.04	16.20	16.19	16.15	19.17	19.11	18.47	18.92
TAPE [30]	33.10	30.37	26.86	30.11	32.59	29.93	26.19	29.57
AirNet [20]	31.63	28.83	23.52	27.99	29.79	26.90	21.35	26.01
SwinIR [27]	33.53	30.89	27.54	30.65	33.50	30.99	27.37	30.62
RAM-SwinIR	33.65	31.06	27.77	30.82	33.82	31.43	27.94	31.07
performance gains	(↑0.12)	(↑0.17)	(↑0.23)	(↑0.17)	(↑0.32)	(↑0.44)	(↑0.57)	(↑0.45)
PromptIR [41]	33.67	31.06	27.80	30.84	33.56	31.08	27.64	30.76
RAM-PromptIR	33.70	31.08	27.79	30.86	33.70	31.30	27.92	30.97
performance gains	(↑0.03)	(↑0.02)	(↓0.01)	(↑0.02)	(↑0.14)	(↑0.22)	(↑0.28)	(↑0.21)

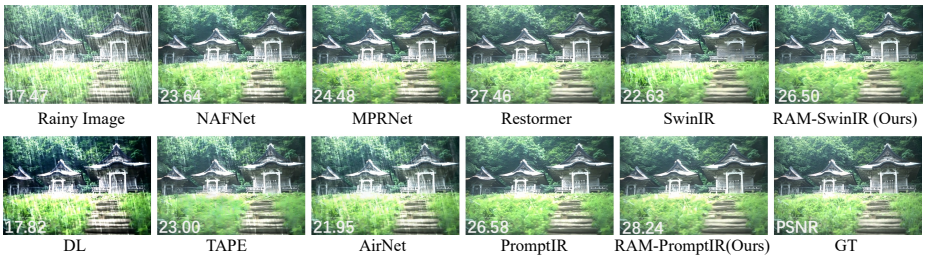


Fig. 7: 在Rain13k-Test数据集上去雨的视觉对比, 放大查看细节。

势。Tab. 2显示了不同噪声水平下的定量去噪结果。无论是RAM-SwinIR还是RAM-PromptIR，性能均优于原始版本。

Fig. 6-Fig. 10展示了各种方法在不同数据集上的定性结果对比。在Fig. 6中，我们的方法在去雾效果（右侧区域）和曝光校正（天空部分）上表现更好。在去雨任务中（Fig. 7），我们的方法更好地去除了雨水条纹并复原了被遮挡区域的纹理。在去噪（Fig. 9）和去模糊任务（Fig. 8）中，我们取得了更清晰的结果，且具有更少的伪影。在低光图像增强任务中（Fig. 10），我们也表现出了更好的颜色校正（左侧的紫色毛毯）和曝光校正效果。为了简化起见，卷积核去模糊和JPEG伪影去除的定性效果将在补充材料中展示。

4.3 消融实验

在本节中，我们对遮盖比例、遮盖块大小、预训练策略、微调策略和微调比例进行了消融实验，以证明我们的MIM预训练和微调策略的有效性。

Fig. 8: 在GoPro数据集上运动去模糊的视觉对比, 放大查看细节。

Table 4: 不同预训练策略的消融结果

RAM-SwinIR	PSNR↑	SSIM↑
pre-trained w/ gt	26.62	0.8580
pre-trained w/ paired data	27.28	0.8698

Table 5: 不同微调策略的消融结果

RAM-SwinIR	PSNR↑	SSIM↑
random	26.86	0.8535
IG [44]	26.92	0.8554
MAC (Ours)	**27.28**	**0.8698**

Table 3: 遮盖率的消融结果

Masking ratio	20%	40%	50%	60%	80%
PSNR↑	27.28	27.21	**27.28**	27.26	27.08
SSIM↑	0.8663	0.8683	**0.8698**	0.8694	0.8642

遮盖块大小和遮盖比例 是决定图像遮盖连续性和区域的两个重要超参数。在高层任务中，MAE [14]使用16 × 16的块大小遮盖图像的75%。然而，这会破坏图像的局部细节，这不适合图像复原任务。

Table 6: 在微调比例方面的PSNR消融结果。 我们比较了在去除未见噪声（分布外去噪）和已知退化图像（分布内）情况下的性能表现。在这种情况下，分布内的设置与Tab. 1相同。

Method	Out-of-Distribution Denoising				In-Distribution
	Possion	Pepper	Speckle	Average	Average
SwinIR [27]	12.83	10.00	20.86	14.56	26.64
RAM-SwinIR$_{10\%}$	**13.67**	**19.23**	**21.07**	**17.99**	27.28
RAM-SwinIR$_{20\%}$	13.27	19.09	20.68	17.68	27.35
RAM-SwinIR$_{50\%}$	12.75	16.51	20.36	16.54	27.38
RAM-SwinIR$_{100\%}$	12.47	15.31	20.01	15.93	**27.54**

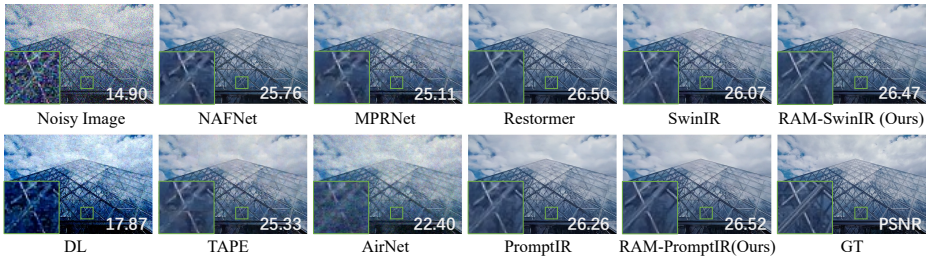


Fig. 9: CBSD68 数据集的去噪可视化对比, 放大查看细节。

我们首先通过在 1×1、4×4 和 8×8 的区域大小下对 SwinIR [27] 进行预训练，找到最佳的区域大小选择，如 Fig. 3 所示。由于 SwinIR 的注意力层将 8×8 的区域视为一个 token，因此 4×4 的预训练会产生严重的伪影。此外，8×8 的预训练生成的结果缺乏细节，例如北极熊爪子上的纹理。相比之下，使用 1×1 区域大小预训练的模型（这也是我们的最终选择）能够实现令人满意的重建效果，并去除大部分雨条纹。

然后，我们将遮盖比例从 20% 调整到 80%。如 Tab. 3 所示，使用 50% 遮盖比例预训练的模型性能最高。此外，当我们继续增加遮盖比例时，PSNR 从 27.28dB 显著下降到 27.08dB，这也证明了我们认为高遮盖比例对图像复原有害的观点。

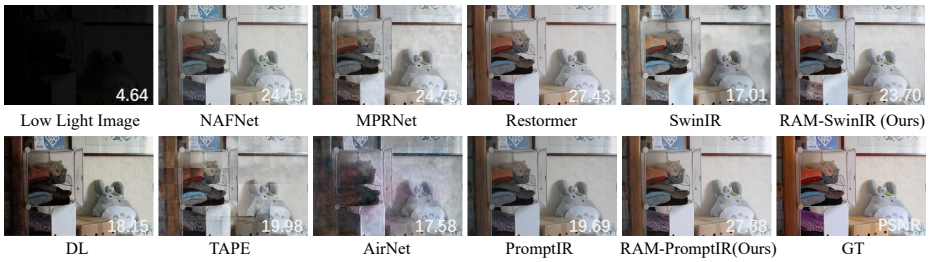


Fig. 10: LOL 数据集的LLIE可视化对比, 放大查看细节。

使用配对数据进行预训练 Tab. 4 比较了使用配对数据进行掩码图像预训练（我们的预训练策略）与仅使用真实图像进行掩码图像预训练的结果。结果表明，使用配对数据进行预训练对我们的 RAM 方法是必要的。在高质量图像上预训练模型并不能有效地支持图像复原任务的学习，它仍然需要配对数据来引导模型的学习过程。

微调策略 为了验证我们微调策略的有效性，我们分别通过 MAC 分析、IG [44] 和均匀采样选择网络层的 10% 进行微调，结果如 Tab. 5 所示。与 IG 相比，我们在 PSNR 上提升了 0.36 dB，SSIM 提升了 1.6%，这表明我们的选择策略优于 IG。

微调比例 我们在 Tab. 6 中进行了消融实验，比较了不同微调比例下网络的性能。我们发现，使用我们的微调策略，预训练网络只需微调少量层（例如 10%）即可达到相当的性能，同时，为了在给定任务上获得最佳性能，我们需要微调几乎所有的网络参数。

性能 *vs* 泛化能力 我们在 Tab. 6 中分析了分布内性能和分布外泛化之间的权衡。我们发现，微调的层数越多，处理分布外任务的泛化能力就越弱。使用我们的微调方法，模型在保持相当性能的同时，可以拥有更强的泛化能力。

5 Conclusion

本文提出了 RAM，一种使用掩码图像建模（MIM）预训练从受损图像中提取内在图像信息的方法。我们设计了一个专门用于图像复原的 MIM 预训练策略，并提出了一种从掩码图像到完整图像过渡的微调算法。通过 MAC 分析层的重要性，我们在最小化参数调整的情况下实现了更高的性能。大量实验表明，我们的 RAM 可以为各种架构带来性能提升，并实现最先进的性能，朝着一体化图像复原的统一解决方案迈进。

References

1. Chen, H., Wang, Y., Guo, T., Xu, C., Deng, Y., Liu, Z., Ma, S., Xu, C., Xu, C., Gao, W.: Pre-trained image processing transformer. In: CVPR. pp. 12299–12310 (2021)
2. Chen, H., Gu, J., Liu, Y., Magid, S.A., Dong, C., Wang, Q., Pfister, H., Zhu, L.: Masked image training for generalizable deep image denoising. In: CVPR. pp. 1692–1703 (2023)
3. Chen, L., Chu, X., Zhang, X., Sun, J.: Simple baselines for image restoration. In: ECCV. pp. 17–33. Springer (2022)
4. Chen, W.T., Huang, Z.K., Tsai, C.C., Yang, H.H., Ding, J.J., Kuo, S.Y.: Learning multiple adverse weather removal via two-stage knowledge learning and multi-contrastive regularization: Toward a unified model. In: CVPR. pp. 17653–17662 (2022)
5. Dhamdhere, K., Sundararajan, M., Yan, Q.: How important is a neuron. In: ICLR (2019)
6. Duan, H., Shen, W., Min, X., Tu, D., Teng, L., Wang, J., Zhai, G.: Masked autoencoders as image processors. arXiv preprint arXiv:2303.17316 (2023)
7. Fan, Q., Chen, D., Yuan, L., Hua, G., Yu, N., Chen, B.: A general decoupled learning framework for parameterized image operators. PAMI **43**(1), 33–47 (2019)
8. Fang, Y., Zhang, H., Wong, H.S., Zeng, T.: A robust non-blind deblurring method using deep denoiser prior. In: CVPRW. pp. 735–744 (June 2022)
9. Fu, X., Huang, J., Ding, X., Liao, Y., Paisley, J.: Clearing the skies: A deep network architecture for single-image rain removal. TIP **26**(6), 2944–2956 (2017)
10. Fu, X., Huang, J., Zeng, D., Huang, Y., Ding, X., Paisley, J.: Removing rain from single images via a deep detail network. In: CVPR. pp. 3855–3863 (2017)
11. Gu, J., Dong, C.: Interpreting super-resolution networks with local attribution maps. In: CVPR. pp. 9199–9208 (2021)
12. Guo, C.L., Yan, Q., Anwar, S., Cong, R., Ren, W., Li, C.: Image dehazing transformer with transmission-aware 3d position embedding. In: CVPR. pp. 5812–5820 (2022)

13. Guo, C., Li, C., Guo, J., Loy, C.C., Hou, J., Kwong, S., Cong, R.: Zero-reference deep curve estimation for low-light image enhancement. In: CVPR. pp. 1780–1789 (2020)
14. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick., R.: Masked autoencoders are scalable vision learners. In: CVPR. pp. 16000–16009 (2022)
15. Huang, J.B., Singh, A., Ahuja, N.: Single image super-resolution from transformed self-exemplars. In: CVPR. pp. 5197–5206 (2015)
16. Jin, X., Han, L.H., Li, Z., Guo, C.L., Chai, Z., Li, C.: Dnf: Decouple and feedback network for seeing in the dark. In: CVPR. pp. 18135–18144 (2023)
17. Kenton, J.D.M.W.C., Toutanova, L.K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: NAACL. pp. 4171–4186 (2019)
18. Leino, K., Sen, S., Datta, A., Fredrikson, M., Li, L.: Influence-directed explanations for deep convolutional networks. In: ITC. pp. 1–8. IEEE (2018)
19. Li, B., Ren, W., Fu, D., Tao, D., Feng, D., Zeng, W., Wang, Z.: Benchmarking single-image dehazing and beyond. TIP **28**(1), 492–505 (2018)
20. Li, B., Liu, X., Hu, P., Wu, Z., Lv, J., Peng, X.: All-in-one image restoration for unknown corruption. In: CVPR. pp. 17452–17462 (2022)
21. Li, C., Guo, C.L., Liang, Z., Zhou, S., Feng, R., Loy, C.C., et al.: Embedding fourier for ultra-high-definition low-light image enhancement. In: ICLR (2022)
22. Li, D., Zhang, Y., Cheung, K.C., Wang, X., Qin, H., Li, H.: Learning degradation representations for image deblurring. In: ECCV. pp. 736–753. Springer (2022)
23. Li, R., Tan, R.T., Cheong, L.F.: All in one bad weather removal using architectural search. In: CVPR. pp. 3175–3185 (2020)
24. Li, X., Wu, J., Lin, Z., Liu, H., Zha, H.: Recurrent squeeze-and-excitation context aggregation net for single image deraining. In: ECCV. pp. 254–269 (2018)
25. Li, Y., Zhang, K., Liang, J., Cao, J., Liu, C., Gong, R., Zhang, Y., Tang, H., Liu, Y., Demandolx, D., et al.: Lsdir: A large scale dataset for image restoration. In: CVPR. pp. 1775–1787 (2023)
26. Li, Y., Tan, R.T., Guo, X., Lu, J., Brown, M.S.: Rain streak removal using layer priors. In: CVPR
27. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: Image restoration using swin transformer. In: CVPR. pp. 1833–1844 (2021)
28. Lin, X., Ren, C., Liu, X., Huang, J., Lei, Y.: Unsupervised image denoising in real-world scenarios via self-collaboration parallel generative adversarial branches. In: ICCV. pp. 12642–12652 (2023)
29. Lin, X., Yue, J., Ren, C., Guo, C.L., Li, C.: Unlocking low-light-rainy image restoration by pairwise degradation feature vector guidance. arXiv preprint arXiv:2305.03997 (2023)
30. Liu, L., Xie, L., Zhang, X., Yuan, S., Chen, X., Zhou, W., Li, H., Tian, Q.: Tape: Task-agnostic prior embedding for image restoration. In: ECCV. pp. 447–464. Springer (2022)
31. Liu, Y., He, J., Gu, J., Kong, X., Qiao, Y., Dong, C.: Degae: A new pretraining paradigm for low-level vision. In: CVPR. pp. 23292–23303 (2023)
32. Liu, Y., Liu, A., Gu, J., Zhang, Z., Wu, W., Qiao, Y., Dong, C.: Discovering distinctive "semantics" in super-resolution networks. arXiv preprint arXiv:2108.00406 (2021)
33. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: CVPR. pp. 10012–10022 (2021)
34. Luo, Y., Xu, Y., Ji, H.: Removing rain from a single image via discriminative sparse coding. In: ICCV. pp. 3397–3405 (2015)

35. Luo, Y., Zhao, R., Wei, X., Chen, J., Lu, Y., Xie, S., Wang, T., Xiong, R., Lu, M., Zhang, S.: Mowe: mixture of weather experts for multiple adverse weather removal. arXiv preprint arXiv:2303.13739 (2023)
36. Magid, S.A., Lin, Z., Wei, D., Zhang, Y., Gu, J., Pfister, H.: Texture-based error analysis for image super-resolution. In: CVPR. pp. 2118–2127 (2022)
37. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: ICCV. vol. 2, pp. 416–423. IEEE (2001)
38. Mehri, A., Ardakani, P.B., Sappa, A.D.: Mprnet: Multi-path residual network for lightweight image super resolution. In: CVPR. pp. 2704–2713 (2021)
39. Nah, S., Hyun Kim, T., Mu Lee, K.: Deep multi-scale convolutional neural network for dynamic scene deblurring. In: CVPR
40. Park, D., Lee, B.H., Chun, S.Y.: All-in-one image restoration for unknown degradations using adaptive discriminative filters for specific degradations. In: CVPR. pp. 5815–5824 (2023)
41. Potlapalli, V., Zamir, S.W., Khan, S.H., Shahbaz Khan, F.: Promptir: Prompting for all-in-one image restoration. NeurIPS **36** (2024)
42. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al.: Improving language understanding by generative pre-training (2018)
43. Shrikumar, A., Su, J., Kundaje, A.: Computationally efficient measures of internal neuron importance. arXiv preprint arXiv:1807.09946 (2018)
44. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: ICML. pp. 3319–3328. PMLR (2017)
45. Sundararajan, M., Taly, A., Yan, Q.: Gradients of counterfactuals. ICLR (2017)
46. Wang, X., Wang, W., Cao, Y., Shen, C., Huang, T.: Images speak in images: A generalist painter for in-context visual learning. In: CVPR. pp. 6830–6839 (2023)
47. Wang, Z., Cun, X., Bao, J., Zhou, W., Liu, J., Li, H.: Uformer: A general u-shaped transformer for image restoration. In: CVPR. pp. 17683–17693 (2022)
48. Wei, C., Wang, W., Yang, W., Liu, J.: Deep retinex decomposition for low-light enhancement. In: BMVC. British Machine VLOLision Association (2018)
49. Wu, R.Q., Duan, Z.P., Guo, C.L., Chai, Z., Li, C.: Ridcp: Revitalizing real image dehazing via high-quality codebook priors. In: CVPR. pp. 22282–22291 (2023)
50. Xie, L., Wang, X., Dong, C., Qi, Z., Shan, Y.: Finding discriminative filters for specific degradations in blind super-resolution. NeurIPS **34**, 51–61 (2021)
51. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: Simmim: a simple framework for masked image modeling. In: CVPR. pp. 9653–9663 (2022)
52. Yang, W., Tan, R.T., Feng, J., Liu, J., Guo, Z., Yan, S.: Deep joint rain detection and removal from a single image. In: CVPR. pp. 1357–1366 (2017)
53. Yang, W., Tan, R.T., Wang, S., Fang, Y., Liu, J.: Single image deraining: From model-based to data-driven and beyond. PAMI **43**(11), 4059–4077 (2020)
54. Yang, W., Wang, W., Huang, H., Wang, S., Liu, J.: Sparse gradient regularized deep retinex network for robust low-light image enhancement. TIP **30**, 2072–2086 (2021)
55. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: CVPR. pp. 5728–5739 (2022)
56. Zhang, C., Zhu, Y., Yan, Q., Sun, J., Zhang, Y.: All-in-one multi-degradation image restoration network via hierarchical degradation representation. In: ACMMM. pp. 2285–2293 (2023)
57. Zhang, H., Patel, V.M.: Density-aware single image de-raining using a multi-stream dense network. In: CVPR. pp. 695–704 (2018)

58. Zhang, H., Sindagi, V., Patel, V.M.: Image de-raining using a conditional generative adversarial network. TCSVT **30**(11), 3943–3956 (2019)
59. Zhang, J., Huang, J., Yao, M., Yang, Z., Yu, H., Zhou, M., Zhao, F.: Ingredient-oriented multi-degradation learning for image restoration. In: CVPR. pp. 5825–5835 (2023)
60. Zheng, N., Zhou, M., Dong, Y., Rui, X., Huang, J., Li, C., Zhao, F.: Empowering low-light image enhancer through customized learnable priors. In: ICCV. pp. 12559–12569 (2023)
61. Zhu, Y., Wang, T., Fu, X., Yang, X., Guo, X., Dai, J., Qiao, Y., Hu, X.: Learning weather-general and weather-specific features for image restoration under multiple adverse weather conditions. In: CVPR. pp. 21747–21758 (2023)