一种用以高效图像复原的自适应稀疏 Transformer

周世豪，潘金山，杨巨峰

摘要—基于 Transformer 的方法凭借其能够建模对于恢复清晰图像至关重要的长距离依赖关系的能力，在图像复原任务中展现了巨大的潜力。尽管现有工作已提出多种高效注意力机制用以减轻 Transformer 巨大的计算负担，但是由于它们通常会考虑所有可用的令牌（tokens），因此性能往往受到冗余信息以及不相关区域噪声交互的影响。在本文中，我们提出了一种**自适应稀疏** Transformer（AST-v2），通过减少无关区域的噪声交互并消除通道维度上的特征冗余来缓解上述问题。AST-v2 包含两个核心组件：一个自适应稀疏自注意力（ASSA）模块和一个特征细化前馈网络（FRFN）。ASSA 采用双分支设计，其中的稀疏分支负责引导标准密集注意力权重的调制。该范式在保留重要交互的同时，减少了不相关令牌交互所带来的消极影响。同时，FRFN 利用一种“增强-消除”（enhance-and-ease）策略来消除通道间的特征冗余，从而增强了恢复清晰图像的能力。在常用基准数据集上的实验结果表明，我们的方法在包括雨痕去除、去雾、阴影去除、去雪、去模糊和低光照增强在内的 6 个修复任务上，均表现出优越的性能。**源代码可见**https://github.com/joshyZhou/ASTv2。

关键词—图像复原,transformer，注意力机制，底层视觉，深度学习

1 引言

图像复原任务旨在从降质图像中恢复出清晰图像。基于 CNN 的模型已经取得了显著的进展 [1], [2], [3]。然而，CNN 模型中基本模块卷积操作的感受野有限且在捕获长距离依赖方面表现低效。相比之下，最近基于 Transformer 的架构 [4], [5] 利用自注意力机制来建模全局相关性，从而克服了上述局限。然而，基于 Transformer 的方法存在计算复杂度过高的问题，这也限制了这类模型在实际场景中的应用。

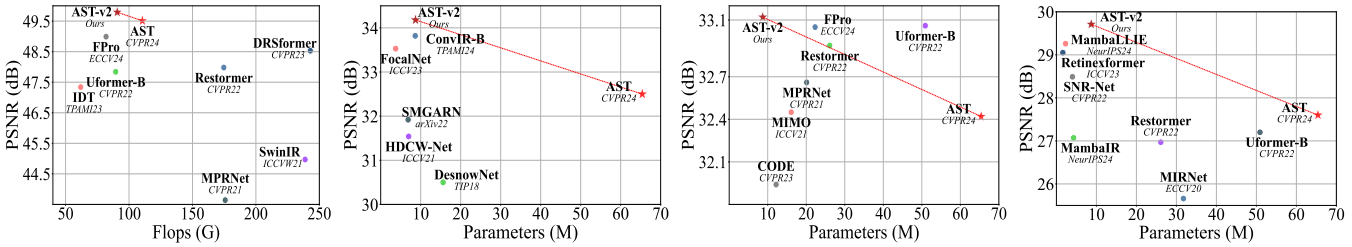
尽管一些工作尝试通过设计高效注意力机制以解决计算负担方面的挑战 [10], [11], [12]，但目前仍然存在两个关键的难题尚未解决：1) 如图 1所示，标准的 Transformer [10], [11] 通过建模密集的注意力关系来聚合特征，但这种做法会不可避免地引入无关区域的噪声交互。2) 密集聚合的特征图中所包含的冗余信息 [13], [14] 会阻碍模型关注包含关键信息的特征。最近，有研究者提出了一些方法 [15], [16]，旨在从特征中滤除噪声交互并去除冗余信息。这些方法或采用 Top-K 选择操作来保留最相关的令牌（tokens）[15] 交互，或在执行自注意力计算前将特征图投影到超像素空间 [16] 以压缩无关信息。尽管如此，这些方法仍存在局限：一方面，针对不同的复原任务，参数 K 的选择可能十分敏感；另一方面，超像素空间中的自注意力机制仍会考虑所有令牌间的关系。因此，这些方法可能依旧会面临特征图中存在冗余信息的挑战。

在实践中，如何设计一种高效的机制，使模型既能从信息流中识别出最有价值的特征，同时又能最小化对特定复原任务的敏感性，是一个社区正面临的挑战。标准的 Transformer 模型 [11], [12] 在聚合特征时，通常会计算所有“查询-键”（query-key）之间的关系。然而，由于并非所有“查询”令牌都与其对应的“键”令牌密切相关，因此利用全局所有的相似性并不利于清晰图像的重建。直观上，开发一种能够选择令牌间最相关交互的稀疏 Transformer，可以有效提升聚合特征的表征力。为了实现注意力的稀疏性，基于 ReLU 平方的激活函数 [17] 提供了一种可行的解决方案，其可以在无需考虑设置任务特定参数 [15] 的情况下，移除不相关的相似度计算结果。然而，为了缓解该方案造成的信息丢失问题 [18]，往往需要采用一些特定的设计 [19], [20] 来放宽其稀疏性，而这恰恰违背了使用稀疏自注意力替代密集自注意力的初衷。因此，我们

(a) DSA (b) SSA (c) ASSA

图 1: 不同自注意力机制的比较。(a) 标准的基于 softmax 激活函数的自注意力机制将所有可用的令牌（tokens）纳入关系计算，从而产生密集的自注意力（DSA）分数。来自无关区域的噪声交互将无可避免地被引入计算，进而阻碍模型学习到具有表征力的特征。(b) 将传统自注意力机制中的 softmax 激活函数替换为 ReLU 层，可以得到稀疏自注意力（SSA）来滤除负值。然而，这种设计会引发信息丢失问题，从而导致性能下降。(c) 本文提出的自适应稀疏自注意力（ASSA）采用双分支设计，其中的稀疏分支负责引导标准密集注意力权重的调制，从而在保留关键交互的同时，减少无关令牌（token）交互带来的负面影响。

- 周世豪，杨巨峰：南开大学计算机学院，中国，*300350*。*(邮件：zhoushihao96@mail.nankai.edu.cn; yangjufeng@nankai.edu.cn)*
- 潘金山：南京理工大学计算机科学与工程学院，中国，*210094*。*(邮件：sdluran@gmail.com)*。



(a) (b) (c) (d)

图 2: 本文提出的 AST-v2 与 SOTA 算法在四种图像修复任务上的对比。(a) 在 SPAD [6] 数据集上，图像去雨任务的 PSNR 与 FLOPs 对比；(b) 在 SNOW100K [7] 数据集上，图像去雪任务的 PSNR 与参数量对比；(c) 在 GoPro [8] 数据集上，图像运动去模糊任务的 PSNR 与参数量对比；以及 (d) 在 SMID [9] 数据集上，图像暗光增强任务的 PSNR 与参数量对比。

探索了一种替代方法，旨在确保尽可能保留有效特征表示的同时，减少噪声特征表示。

鉴于此，我们提出了一种高效的基于 Transformer 的模型用于图像复原任务，模型由两个核心模块组成：自适应稀疏自注意力（ASSA）模块和特征细化前馈网络（FRFN）。简而言之，ASSA 由两个分支组成：一个稀疏自注意力分支（SSA）和一个密集自注意力分支（DSA）。具体而言，SSA 引导调制标准的密集注意力权重，从而在保留重要交互的同时，减少不相关令牌（token）交互的负面影响。与此同时，这种方法也缓解了基于 ReLU 的 SSA 所固有的过度稀疏问题。

另一方面，我们开发了一种对常规前馈网络 [5] 的简单而有效的替代方案，即 FRFN，旨在通过增强特征表示来改善潜在的图修复效果。本质上，FRFN 是通过一种 “增强-简化” (enhance-and-ease) 的方案来执行特征转换。该方案首先会增强特征图中的有效成分，然后通过门控机制来减少冗余成分。通过抑制通道维度上的冗余信息，FRFN 使模型能够有效地学习到最具代表性的特征。

总体而言，本文的主要贡献体现在以下三个方面：

- 本文提出了一种高效的基于 Transformer 的模型，通过增强最有价值的信息在网络中的流动来提取更具意义的特征，最终实现清晰图像的修复。
- 所提出的模型包含一个自适应稀疏自注意力（ASSA）模块。该模块遵循双分支范式，旨在捕获令牌（token）间的有效特征交互，并充分保留用于学习代表性特征的关键信息。此外，我们还开发了一种特征细化前馈网络（FRFN）。该网络基于 “增强-简化”（enhance-and-ease）的特征变换方案，能够在增强有价值特征的同时，抑制信息量较少的特征。
- 我们通过在六种降质任务（包括去雨痕、去雾、去阴影、去模糊、去雪和暗光增强）上进行的大量实验，验证了所提出模型的有效性。

本文是我们前期工作 [21] 的扩展版本，主要区别如下：1) 我们通过稀疏自注意力得分对标准的密集自注意力进行像素级调制，实现了自适应稀疏注意力分数，从而带来显著的性能提升。2) 我们在 FRFN 中采用了深度可分离（depth-wise）的方式实现的部分卷积（Partial Convolution），帮助模型选择有效的信号。3) 我们在多个图像修复任务上验证了所提方法（AST-v2）的有效性。这些任务不仅包括会议版本涉及的图像去雨和去雾任务，还新增了去雪、去阴影、去模糊以及暗光增强这四项新任务。如图 2所示，我们提出的 AST-v2 在所有任务上的性能均优于当前最先进的（SOTA）算法。此外，我们还通过大量的实验证明，AST-v2 在真实世界场景中同样表现出色，尤其体现在泛化能力以及对下游高级任务的促进方面。

2 相关工作

图像复原。在过去的十年里，研究社区见证了从依赖手工特征和先验假设的传统模型 [3], [22], [23]，转向数据驱动基于学习方法 [24], [25], [26] 的范式转变。后者在解决各种图像降质问题上，例如雨痕 [27], [28], [29]、雾 [30], [31], [32] 和雨滴 [33], [34], [35] 等方面，展现出了卓越的性能。这些方案取得的性能提升又主要归功于受复杂高级视觉任务所启发的、多样化的神经网络架构 [36] 以及先进的组件 [37], [38], [39]。例如，能够获取层级化多尺度表示的 U 形网络设计 [40], [41], [42]，以及有助于学习残差特征的跳跃连接 [43], [44], [45]，都被广泛地应用在底层视觉研究领域。尽管基于 CNN 的网络在图像修复任务中取得了令人印象深刻的结果，但它们由于受到卷积操作固有的有限感受野的制约，限制了其捕获长距离依赖的能力。为了克服这一局限，近期的工作 [46], [47], [48] 探索了使用注意力机制来提升修复性能。例如，SPANet [6] 对 IRNN 模型进行了扩展，使其能够针对雨痕显式地生成注意力图，从而提升了模型去除雨痕伪影的能力。RCAN [49] 利用一种通道注意力机制来选择性地放大具有丰富信息的特征。更多关于网络架构的设计在 NTIRE 挑战赛报告 [50], [51] 和近期的综述文章 [52], [53], [54] 中有更详尽的探讨。

视觉 Transformer。受启发于 Transformer 在自然语言处理（NLP）领域取得卓越性能 [4] 的事实，基于 Transformer 的架构逐渐被引入计算机视觉领域 [55], [56]。IPT [57] 是最早将 Transformer 架构应用于图像修复的工作之一。该方法通过将输入图像划分为多个小块、并对其进行顺序处理的方式，解决了计算资源紧张的挑战，进而实现高效且有效的修复。尽管如

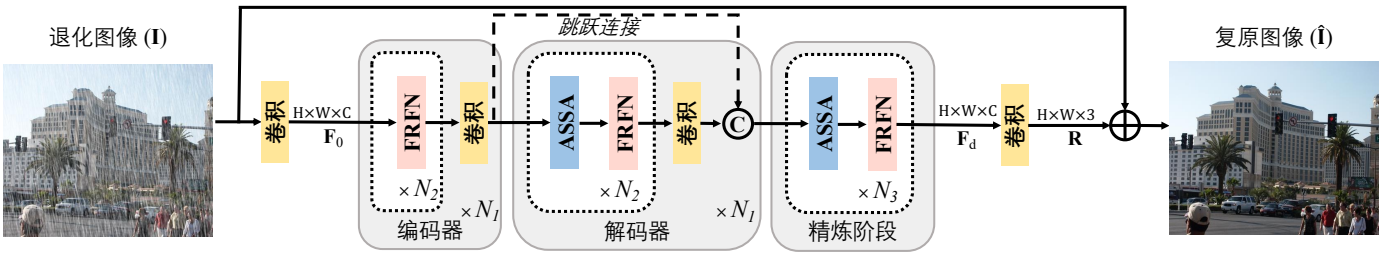


图 3: 本文提出的自适应稀疏 Transformer (AST-v2) 框架。它主要由自适应稀疏自注意力（ASSA）模块和特征细化前馈网络（FRFN）构成。

此，原生自注意力（vanilla self-attention）机制中“每个令牌（token）都与其他所有令牌进行交互”的特性，导致了其平方级别的计算复杂度，这成为将 Transformer 应用于高分辨率图像时的一个计算瓶颈。为了缓解这一问题，Restormer [11] 引入了通道注意力机制，沿通道维度计算注意力权重，从而有效降低了计算成本。另一个解决方案是基于窗口的注意力机制 [58]。例如，Uformer [12] 通过设计一个经过局部增强的、基于窗口的 Transformer，从而将“局部性”（locality）引入了 Transformer 架构之中。SwinIR [10] 同样采用了基于窗口的注意力机制，并在此基础上引入了一种移位机制（shift），从而促进了更多的跨窗口的交互。此外，GRL [59] 整合了窗口注意力和通道注意力，利用两者互补的优势来构建一个有效的图像复原模型。

尽管这些高效的注意力机制能够有效缓解计算负担，并在处理各类降质图像时展现出优越性能，但特征图中存在的无关表征或信息冗余问题，仍然是进一步提升模型性能的主要障碍 [15], [16]。为了解决这一问题，DRSformer [15] 在其注意力机制中引入了一个 top-k 通道选择算子。该算子根据相关性来选择最具信息的令牌（token），从而在保留关键信息的同时，有效降低了计算成本。类似地，CODE [16] 将特征投影到超像素空间中，同时减少空间域和通道域的冗余，进而得到更高效的特征表示。然而，在 top-k 通道选择算子中，对不同的图像修复任务而言参数 k 的具体选择十分敏感。此外，在超像素空间中执行注意力机制，依然需要计算所有可用令牌（token）间的关系，这同样引入不必要的无关区域交互。

总体而言，我们提出的 AST-v2 与现有方法的主要不同之处体现在以下两个方面：一方面，我们引入了一种自适应稀疏自注意力（ASSA）机制。该机制能够根据输入特征动态地调整稀疏模式，以此减少冗余信息，并最终选择出信息量最丰富的交互。我们首先用平方 ReLU 激活函数替换 softmax 层，以获得稀疏得分。同时，我们没有仿照先前的工作 [60], [61], [62] 选择设计复杂的组件来解决由过度稀疏引起的信息丢失问题，而是探索了一种更直接有效的方法，即利用稀疏分数来指导调整标准注意力权重。通过这种方式，我们的模型能够充分利用 SSA 分支的稀疏分数，从而避免了因“基于 ReLU 的 SSA 过度稀疏”特性所导致的信息有限、进而难以学习到令人满意表征的问题。另一方面，我们在 AST-v2 中引入了另一个关键组件，即特征细化前馈网络（FRFN）。为了处理特征图中隐藏的冗余信息，该模块采用了一种“增强-简化”（enhance-and-ease）方案，通过沿通道维度增强有效特征，同时抑制信息量较少的部分。

3 提出的方法

3.1 总体流程

本文提出的 AST-v2 的整体流程如图 3所示。对于一张给定的图像 $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$，AST-v2 首先利用一个卷积层来提取底层特征表示 $\mathbf{F_0} \in \mathbb{R}^{H\times W\times C}$，其中 $H \times W$ 代表图像分辨率，C 代表通道数。接下来，$\mathbf{F_0}$ 会通过一个包含 N_1 个阶段的对称编码器-解码器网络，从而得到深度特征表示 $\mathbf{F}_d \in \mathbb{R}^{H\times W\times C}$。具体而言，编码器内的每个阶段均由 N_2 个基本模块构成，之后再连接一个用于下采样的卷积层。编码器中的每个基本模块都包含一个 FRFN。来自编码器的特征会通过恒等连接，与解码器中的特征进行融合。考虑到注意力机制固有的低通滤波器特性 [63] 可能会阻碍模型在早期阶段 [64] 学习到必要的局部模式，我们省略了编码器中标准 Transformer 模块内的注意力计算。在解码器端，每个阶段则由 N_2 个基本模块构成，之后连接一个用于上采样的卷积层。解码器中的每个基本模块则包含一个 ASSA 和一个 FRFN。在此之后，模型还会引入一个精炼阶段（refinement stage）。该阶段在与输入降质图像相同的高空间分辨率下运行，旨在进一步增强所学习到的特征。最后，AST-v2 使用一个卷积层，从深度特征 $\mathbf{F}_d$ 中生成残差图像 $\mathbf{R} \in \mathbb{R}^{H\times W\times 3}$。修复后的图像是通过将残差图像与降质图像相加得到的，即 $\widehat{\mathbf{I}} = \mathbf{I} + \mathbf{R}$。我们采用 Charbonnier 损失函数 [65] 来训练 AST-v2：$\ell(\mathbf{I}', \widehat{\mathbf{I}}) = \sqrt{\|\mathbf{I}' - \widehat{\mathbf{I}}\|^2 + \epsilon^2}$，其中 $\mathbf{I}'$ 指的是真实图像，ϵ 经验性地设置为 10^{-3}。

3.2 AST-v2 模块设计

自适应稀疏自注意力。原始 Transformer [4], [5], [12] 会考虑特征图内的所有令牌（token），这导致计算许多无关区域的交互。该范式不仅处理了信息量低的区域，还引入了冗余且不相关的特征，从而导致了模型性能的下降。为了解决该问题，我们引入了基于平方 ReLUh 函数的稀疏自注意力（SSA）机制。该机制能够过滤掉那些因“查询-键”（query-key）匹配分数过低而带来负面影响的特征，从而保证了注意力机制的稀

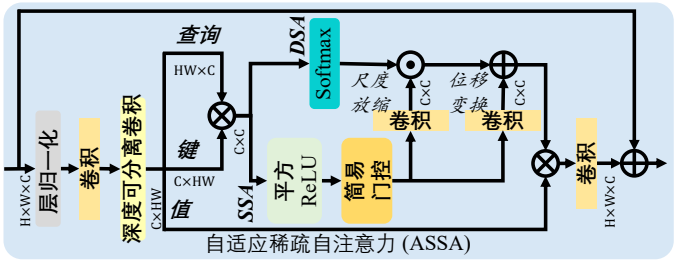


图 4: 自适应稀疏自注意力（ASSA）机制示意图。

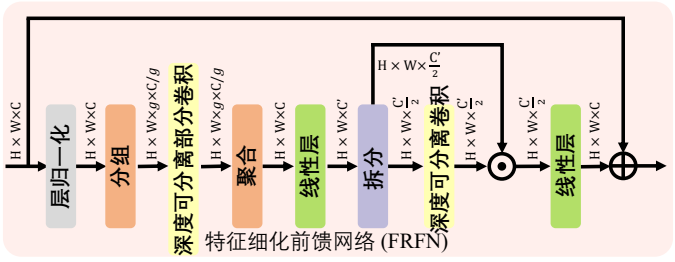


图 5: 特征细化前馈网络（FRFN）示意图。

疏性 [60]。与此同时，考虑到基于 ReLU 的自注意力存在过度稀疏的问题 [18]，我们并没有直接用稀疏注意力来取代标准的密集自注意力（DSA），而是利用稀疏分数作为引导，来调整标准的密集注意力权重。因此，关键的挑战在于如何设计一种引导机制，使其能够在尽可能多地保留有效特征的同时，有效减少噪声特征和冗余信息。为此，ASSA 通过使用尺度放缩和位移变换，对注意力分数进行像素级调制，这一做法与批归一化技术 [66] 相类似。

对于一个归一化后的特征图 $\mathbf{X} \in \mathbb{R}^{\hat{H}\times\hat{W}\times\hat{C}}$，我们首先利用 1×1 卷积和 3×3 深度可分离卷积，从 $\mathbf{X}$ 中生成查询（$\mathbf{Q}$）、键（$\mathbf{K}$）和值（$\mathbf{V}$）矩阵，其计算过程如下：$\mathbf{Q} = \mathbf{X}\mathbf{W}_Q$，$\mathbf{K} = \mathbf{X}\mathbf{W}_K$，$\mathbf{V} = \mathbf{X}\mathbf{W}_V$。接下来，为了解决原始自注意力的高复杂度问题，我们采用 [11] 中计算通道注意力的方法，对查询和键的投影进行重塑，最终得到一个大小为 $\mathbb{R}^{\hat{C}\times\hat{C}}$ 的转置注意力图 $\mathbf{A}$。其注意力计算可定义如下：

$$\mathbf{A} = f\left(\frac{\mathbf{Q}\mathbf{K}^T}{\alpha}\right)\mathbf{V}, \tag{1}$$

其中，$\mathbf{A}$ 表示估计的注意力，α 是可学习的缩放因子，而 $f(\cdot)$ 是一个评分函数。我们并行地计算不同的“头”（head），然后将这些“头”进行拼接，并最终通过一个线性投影进行融合。

接下来，我们首先回顾在多数现有工作 [10], [11] 中所采用的标准 DSA 机制。该机制通常采用一个 softmax 层，通过考虑所有“查询-键”（query-key）对来获得注意力分数：

$$\mathbf{DSA} = SoftMax\left(\frac{\mathbf{Q}\mathbf{K}^T}{\alpha}\right). \tag{2}$$

由于并非所有“查询”令牌（query token）都与其对应的“键”令牌（key token）密切相关，因此利用全部令牌间的交互来重建清晰图像的做法无疑是低效的。直观上，开发一种能够选择令牌（token）间有效交互的 SSA 机制，可以增强特征聚合的效果。采用基于平方 ReLU 的层是实现注意力稀疏性的一种可行方案，通过移除负分的相似度结果，从而将最有用的信息流向前传播：

$$\mathbf{SSA} = ReLU^2\left(\frac{\mathbf{Q}\mathbf{K}^T}{\alpha}\right). \tag{3}$$

但值得注意的是，基于 ReLU 的 SSA 通常会引发信息丢失问题，因而需要额外的技术 [19], [20] 来放宽其稀疏性，而这种做法却恰恰违背了使用 SSA 替代 DSA 的初衷。简单地应用基于 ReLU 的 SSA 会导致过度稀疏，这意味着学习到的特征表示会因缺乏足够信息而无法支持后续的处理。反之，使用基于 softmax 的 DSA 会无意间引入无关区域的噪声交互，这就给高质量图像的恢复带来了挑战。因此，我们不偏好任何一种范式，而是提出利用稀疏分数作为引导来调制 DSA 的分数，从而形成一种自适应注意力机制。为此，我们将学习到的 SSA 先通过一个简单的门控机制，再对其进行线性投影，从而得到所需的尺度和位移因子进行像素级调制：

$$\begin{aligned} \mathbf{F}_{\mathbf{SG}} &= GELU(\mathbf{SSA}) \odot \mathbf{SSA}, \\ \gamma = \mathbf{F}_{\mathbf{SG}}\mathbf{W}_\gamma, &\beta = \mathbf{F}_{\mathbf{SG}}\mathbf{W}_\beta, \\ \mathbf{A} &= (\gamma \odot \mathbf{DSA} + \beta)\mathbf{V}, \end{aligned} \tag{4}$$

其中，$GELU(\cdot)$ 表示 GELU 激活函数 [67]，$\odot$ 代表逐元素乘法，$\mathbf{F}_{\mathbf{SG}} \in \mathbb{R}^{\hat{C}\times\hat{C}}$ 是经过简易门控机制处理后的特征图，而 $\mathbf{W}_\gamma$ 和 $\mathbf{W}_\beta$ 则分别是尺度放缩项 $\gamma \in \mathbb{R}^{\hat{C}\times\hat{C}}$ 和位移变换项 $\beta \in \mathbb{R}^{\hat{C}\times\hat{C}}$ 的投影矩阵。

这种针对 DSA 的像素级调制设计，实现了自适应的稀疏自注意力分数，从而确保了模型在保留并利用充足有效特征的同时，也能滤除无关区域的噪声交互。换句话说，模型能够根据具体的任务需求，动态地调整输入令牌的稀疏程度。

特征细化前馈网络。标准的前馈网络（FFN）[4] 会独立处理每个像素位置的信息，它在增强经由自注意力机制处理后的特征表示方面，扮演着至关重要的角色。在本文中，我们开发了 FRFN，采用一种“增强-消除”（enhance-and-ease）范式来对特征进行变换，旨在实现更好的潜在高质量图像复原结果。具体来说，我们通过整合两个部分来构建 FRFN：其一，用以增强特征中的有效信息元素的优化部分卷积操作 [68]；其二，用以减轻冗余信息处理负担的门控机制。给定一个经过层归一化的输入特征图 $\mathbf{X} \in \mathbb{R}^{\hat{H}\times\hat{W}\times\hat{C}}$，我们首先沿着通道维度将其分为 G 组。令 $\mathbf{X} = [\mathbf{X}_1, \cdots, \mathbf{X}_G]$，其中 $\mathbf{X}_i \in \mathbb{R}^{\hat{H}\times\hat{W}\times g\times\hat{C}/g}$，$i \in \{1, \cdots, G\}$。然后，对于每个子特征，我们使用一种深度可分离部分卷积来选择有用元素，即 $\widehat{\mathbf{X}}_i = \text{P-DWConv}(\mathbf{X}_i)$，其中 P-DWConv(·) 指的是深度可分离实现的部分卷积 [68]。然后，我们将这些经过变换的特征进行聚合，得到：$\widehat{\mathbf{X}} = [\widehat{\mathbf{X}}_1, \cdots, \widehat{\mathbf{X}}_G]$，其中 $\widehat{\mathbf{X}} \in \mathbb{R}^{\hat{H}\times\hat{W}\times\hat{C}}$。接

下来，我们简化特征中的冗余：

$$\begin{aligned}\widehat{\mathbf{X}}' &= \mathbf{W}_1\widehat{\mathbf{X}}, \widehat{\mathbf{X}}' = [\widehat{\mathbf{X}}'_1, \widehat{\mathbf{X}}'_2],\\ \widehat{\mathbf{X}}'_r &= \widehat{\mathbf{X}}'_1 \odot \mathrm{DWConv}(\widehat{\mathbf{X}}'_2),\\ \widehat{\mathbf{X}}'_{out} &= \mathbf{W}_2\widehat{\mathbf{X}}'_r,\end{aligned} \tag{5}$$

其中 $\mathbf{W}_1$ 和 $\mathbf{W}_2$ 表示线性投影；$[,]$ 指的是通道级切片操作；$\mathrm{DWConv}(\cdot)$ 指的是深度可分离卷积 [69]；

特征图，尤其是在网络深层的特征图，通常具有较高的通道维度，但并非所有通道都包含用于恢复清晰图像的关键信息。对所有通道应用相同的特征变换可能会导致信息冗余。因此，我们采用了分组部分卷积操作，仅对部分通道进行卷积，其作用类似于一种稀疏操作，旨在筛选出有效的通道。之后，我们对增强后的特征图应用门控机制，以去除通道维度上信息量不大的特征。总的来说，FRFN 通过从信息流中提取代表性元素并简化冗余部分的方式，增强了特征表示。

4 实验

在本节中，我们评估了 AST-v2 在 6 种图像修复任务上的性能，包括图像去雨、图像去雾、图像去雪、运动去模糊、阴影去除和低光照增强。同时，我们进行了消融实验，以探究所提出模块的有效性。

4.1 详细实验设置

我们在表 1中总结了用于训练和评估的数据集。接下来，我们将针对每个任务，详细说明其所使用的数据集的具体情况。

去雨。我们在 SPAD 数据集 [6] 上进行了去雨任务的实验。该数据集包含 638,492 张训练图像和 1,000 张测试图像。我们将训练图像块的大小从 128×128 逐步增大到 256×256 进行渐进式学习。参照 [12]，我们利用测试集中的 512×512 图像块对该模型进行评估。

去雾。对于图像去雾任务，我们遵循 [70] 的设定，在 SOTS 数据集 [71] 上训练 AST-v2。该数据集包含 72,135 对用于训练的雾天/清晰成对图像，以及 500 对测试数据。我们遵循 [30] 的做法，在全分辨率图像上评估 AST-v2。我们将训练图像块的大小从 128×128 逐步增大到 384×384 进行渐进式学习。

去阴影。对于图像去阴影任务，我们在 ISTD 数据集 [72] 上训练 AST-v2。该数据集包含 1,870 个图像三元组，每个三元组都由一张阴影图像、一张阴影掩码图像以及一张对应的无阴影图像组成。我们使用 128×128 和 256×256 的训练图像块进行渐进式学习。遵照 [83] 的协议，我们对尺寸放缩为 256×256 的图像进行评估。

去雪。对于图像去雪任务，我们在 Snow100K 基准数据集 [7] 上验证 AST-v2 的性能。该数据集包含 100,000 张合成的降雪图像，并与对应的无雪图像成对提供。按照 [84] 的设置，我们在 Snow100K 的一个子集上进行训练，该子集包含用于训练的 2,500 对图像以及 2,000 张测试图像。此外，该数据集还包含 1,329 张真实的降雪图像，用于测试模型在真实场景下的性能表现。我们使用尺寸为 128×128 和 256×256 的训练图像块进行渐进式学习。按照 [85] 中的评估协议，我们在原始分辨率的完整图像上评估 AST-v2 的性能。

表 1: 图像修复任务数据集总结

Tasks	Dataset	Train	Test	Type
rain streak	SPAD [6]	638,492	1,000	Real
	Internet-Data [6]	0	147	Real
haze	SOTS [71]	72,135	500	Syn
shadow	ISTD [72]	1330	540	Real
snow	SNOW100K [7]	2,500	2,000	Syn
	Realistic [7]	0	1,329	Real
low-light	SMID [9]	15,763	5,046	Real
	LOL-v2-real [73]	689	100	Real
	LOL-v2-syn [73]	900	100	Syn
motion blur	GoPro [8]	2,130	1,111	Syn
	RealBlur [74]	0	1,960	Real

暗光增强。我们在三个代表性的基准数据集上评估 AST-v2 的性能，包括 SMID [9], LOL-v2-real [73] 和 LOL-v2-syn [73]。LOL-v2-real 数据集的训练集与测试集按照 689:100 的比例划分，LOL-v2-synthetic 数据集则按照 900:100 的比例划分。SMID 数据集由 20,809 对短曝光与长曝光的 RAW 图像组成，其中 15,763 对被用于模型训练，其余用于性能评估。

运动去模糊。在图像去模糊任务中，我们在 GoPro 数据集 [8] 上对所提出的模型进行训练，并在真实场景的 RealBlur 数据集 [74] 和 GoPro 的测试集共计两个数据集上进行评估。GoPro 基准数据集的训练集由 2,103 对模糊图像与对应的清晰图像组成，测试集包含 1,111 对图像对，用于模型性能验证与评估。RealBlur 数据集包含两个子集，每个子集都含有 980 对模糊/清晰成对图像。考虑到这些基准数据集中存在非方形的测试输入，我们遵循 [86] 的做法，应用滑动窗口策略来消除块状伪影。我们将训练图像块的大小从 128×128 逐步增大到 512×512 进行渐进式学习。我们遵循 [11], [12], [87] 的做法，在全尺寸分辨率的图像上评估了所提出的模型。

实现细节。在默认设置下，AST-v2 的编码器和解码器部分均包含 N_1=3 个阶段，细化部分则包含一个阶段。具体来说，我们将嵌入维度 C 设置为 48。编码器与解码器的 Transformer 模块数量相同，均为 N_2；而细化部分则包含 $N_3 = 2$ 个模块。我们遵循 [11] 的做法，在 Transformer 模块中采用了多头自注意力。我们通过改变 Transformer 模块的数量（即 N_2），实现了模型的两个变体，分别命名为 AST-v2 和 AST-v2 (L)。具体来说，对于 AST-v2，我们将 N_2 设置为 $[2, 4, 6]$；而对于 AST-v2 (L)，我们则将 N_2 设置为 $[4, 6, 8]$。我们采用 AdamW 优化器 [88] 及其默认设置来训练模型。学习率的初始值设置为 $3e^{-4}$，并采用余弦衰减策略 [89] 逐渐降低至 $1e^{-6}$。我们随机应用旋转和翻转操作来进行数据增强。为了节省训练时间，我们采用了渐进式学习策略 [11], [87]。

评估指标。为了评估修复性能，我们采用 PSNR 和 SSIM 两个经典指标 [90]。此外，我们还使用 NIQE [91] 作为无参考指

表 2: 在 SPAD [6] 数据集上针对雨痕去除任务的定量比较。

Methods	PReNet [27]	RCDNet [75]	SPDNet [76]	SPAIR [77]	DualGCN [28]	SEIDNet [29]	MPRNet [78]	AST [21]
PSNR	40.16	43.36	43.55	44.10	44.18	44.96	45.00	<u>49.51</u>
SSIM	0.9797	0.9816	0.9831	0.9875	0.9872	0.9902	0.9911	**0.9942**
Methods	Fu et al. [79]	Restormer [11]	SCD-Former [80]	IDT [81]	Uformer [12]	DRSformer [15]	FPro [82]	AST-v2 (本文方法)
PSNR	45.03	46.25	46.89	47.34	47.84	48.53	48.99	**49.79**
SSIM	0.9907	0.9911	<u>0.9941</u>	0.9929	0.9925	0.9924	0.9936	0.9939

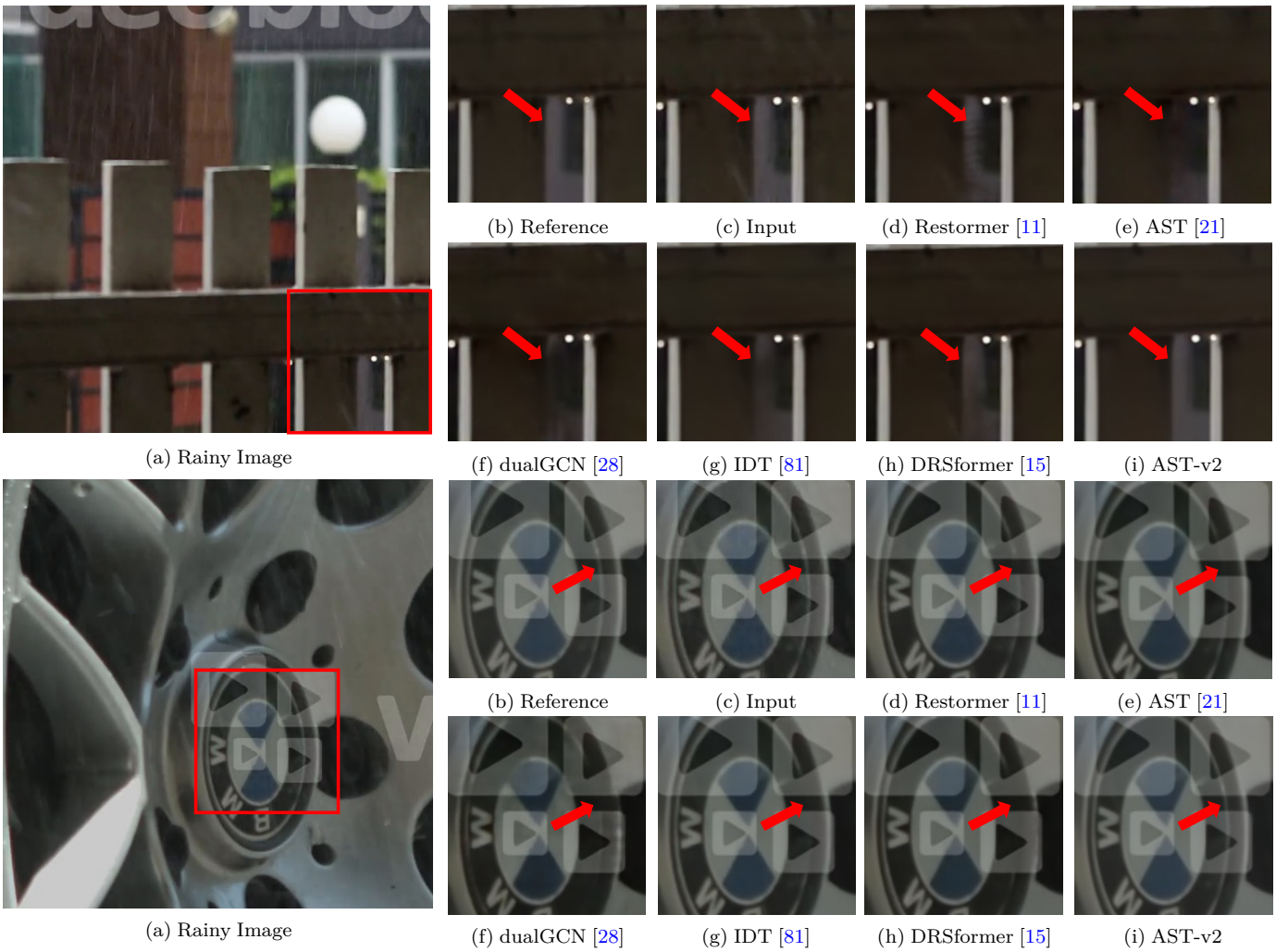


图 6: SPAD [6] 数据集上雨痕去除任务的定性结果。相比之下，其他方法都残留了雨痕，而 AST-v2 生成的结果不仅干净，也更接近参考图像。

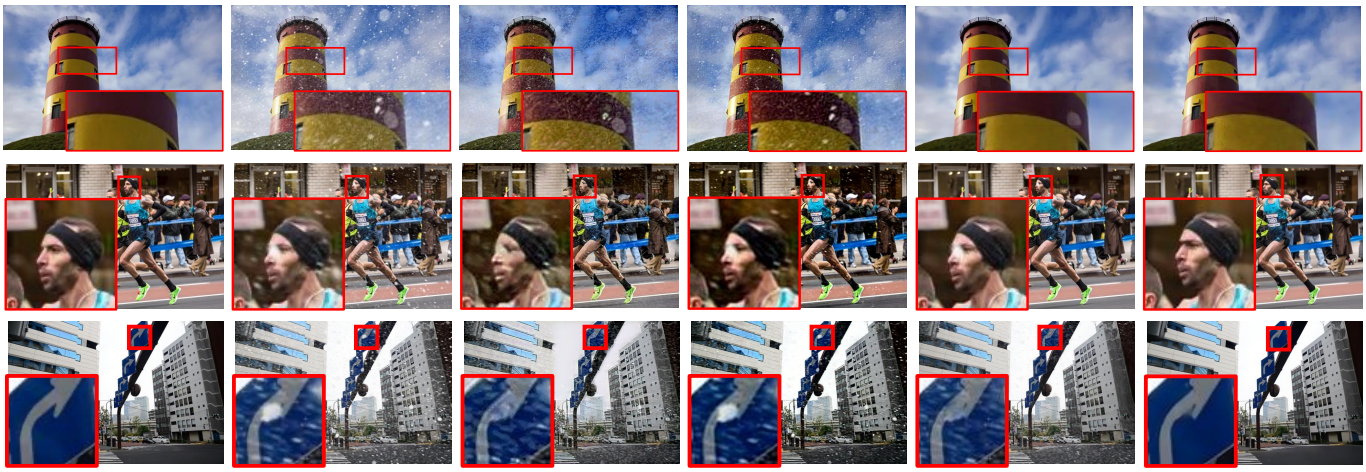
标。值得注意的是，对于去雨任务，我们遵循现有工作 [12], [75] 的做法，在 YCbCr 颜色空间的 Y 通道上计算 PSNR 和 SSIM 分数；而对于其他任务，这些指标则是在 RGB 颜色空间中计算的。在结果表格中，最佳分数用**粗体**标出，次优分数则用<u>下划线</u> 标出。

4.2 图像去雨

我们在 SPAD 基准数据集 [6] 上进行了去雨实验。AST-v2 的性能优于包括 PReNet [27]、RCDNet [75]、SPDNet [76]、SPAIR [77]、DualGCN [28]、SEIDNet [29]、MPRNet [78]、AST [21]、Fu et al. [79]、Restormer [11]、SCD-Former [80]、IDT [81]、Uformer [12]、DRSformer [15] 和 FPro [82] 在内的十五种最先进算法。如表 2所示，AST-v2 在 PSNR 指标上比之前最好的基于 CNN 的方法 [79] 高出 4.76 dB。值得注意的是，AST-v2 相较于各种基于 Transformer 的方法（包括通用的图像恢复方法 [11], [12], [21], [82] 以及专为去雨设计的模型 [15], [80], [81]）展现出一致的优势，这表明了在自注意力机制中选择有用的 token 交互对于图像去雨的有效性。图 6中的视觉对比结果表明，AST-v2 能更有效地去除真实雨痕，同时保持图像的结构内容（图 6i）。由于感受野有限，基于 CNN 的方法 [28] 未能很好地实现复原任务，并留下了明显的伪影（图 6f）。另一方面，基于 Transformer 的方法 [11], [15], [21],

表 3: 在 Snow100K [7] 数据集上针对图像去雪任务的定量比较。

Methods	JSTASR [92]	All in One [93]	CycleGAN [94]	Uformer [12]	DesnowNet [7]	DDMSNet [95]	HDCW-Net [96]	AST [21]
PSNR	23.12	26.07	26.81	29.80	30.50	30.76	31.54	32.50
SSIM	0.86	0.88	0.89	0.93	0.94	0.91	0.95	**0.96**
Methods	TransWeather [85]	SMGARN [97]	NAFNet [1]	MSP-Former [98]	FocalNet [99]	SFNet [84]	ConvIR-S [100]	AST-v2 (本文方法)
PSNR	31.82	31.92	32.41	33.43	33.53	33.79	33.79	**34.18**
SSIM	0.93	0.93	0.95	**0.96**	0.95	0.95	0.95	0.94



(a) Reference (b) Input (c) JSTASR [92] (d) HDCW [96] (e) AST [21] (f) AST-v2

图 7: 在 Snow100K [7] 数据集上进行去雪任务的定性结果。与所对比的其他技术相比，AST-v2 在有效去除雪花的同时，没有牺牲细节纹理。

[81] 能够建模长距离依赖关系，并有效去除雨纹。然而，无关区域之间存在的噪声交互使得这些方法错误地将图像内容识别为雨痕，例如，顶部示例中的白墙和底部示例中轮胎内的白圈，导致这些干净区域遭到修改（图 6d、6e、6g 和 6h）。这违背了图像恢复任务的目标，即仅仅去除退化内容。

4.3 图像去雪

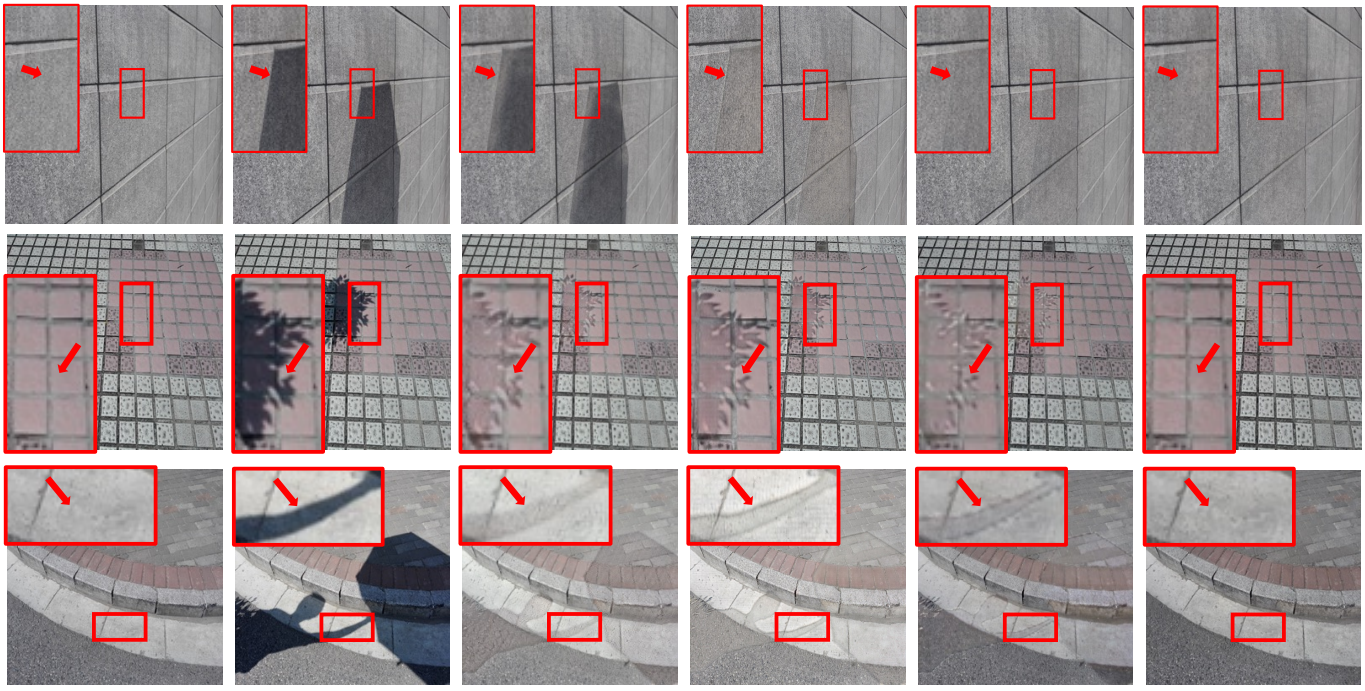
我们在表 3中将 AST-v2 与十五种代表性的去雪方法在 Snow100K 数据集 [7] 上进行了比较。可以看出，AST-v2 取得了最佳的 PSNR 指标和具有竞争力的 SSIM 分数。值得注意的是，AST-v2 超越了专为去雪任务设计的方法 [7], [92], [95], [96], [97], [98]，在 PSNR 指标上高出至少 0.75 dB。与旨在应对恶劣天气条件的图像恢复方法 [85], [93] 相比，AST-v2 在 PSNR 和 SSIM 指标上仍表现出优势。此外，与近期强大的基于 CNN 的基线模型 NAFNet [1] 相比，我们的模型在 PSNR 指标上取得了显著的 1.77 dB 提升。而且，AST-v2 相较于之前性能最佳的基于 CNN 的算法 ConvIR-S [100]，在性能上提升了 0.39 dB。AST-v2 相较于竞争对手的这些优势，充分展示了其先进的设计，证明了我们方法在去雪任务上的有效性。图 7显示，与所比较的技术相比，AST-v2 能够有效地去除雪花，同时不牺牲图像的细节纹理（图 7f）。基于先验的算法 [96] 在去雪效果上表现不佳（图 7d）。尽管基于 CNN 的方法 [92] 取得了相比基于先验的算法更好的效果，但图 7c 中仍包含明显的雪痕残留。尽管 AST [21] 与 AST-v2 同样采用了基于 Transformer 的架构，但其性能（图 7e）仍不如我们提出的模型（图 7f），这表明了规避学习冗余表征的重要性。

表 4: 在 ISTD [72] 数据集上针对图像去阴影任务的定量比较。† 表示该方法利用了额外的阴影掩码信息。

Method	Shadow		Non-Shadow		All	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
LG [102]	30.87	0.979	25.41	0.964	23.88	0.934
DC [101]	31.69	0.976	28.99	0.958	26.38	0.922
G2R [105]	31.63	0.975	26.19	0.967	24.72	0.932
DSP-FFANet [106]	33.17	-	**33.09**	-	29.94	-
Uformer [12]	34.06	0.981	30.02	0.959	28.06	0.930
DMTN [107]	35.29	**0.989**	31.25	**0.973**	29.04	0.958
†Le and Samaras [108]	31.43	0.981	26.21	0.969	24.69	0.941
†SID [109]	32.89	0.986	26.11	0.965	25.01	0.948
†Fu et al. [83]	34.71	0.975	28.61	0.880	27.19	0.846
†SG-ShadowNet [110]	32.85	0.987	26.22	0.967	25.10	0.949
AST [21]	**36.78**	**0.989**	31.91	0.969	30.22	0.954
AST-v2	36.41	**0.989**	32.19	**0.973**	**30.26**	**0.959**

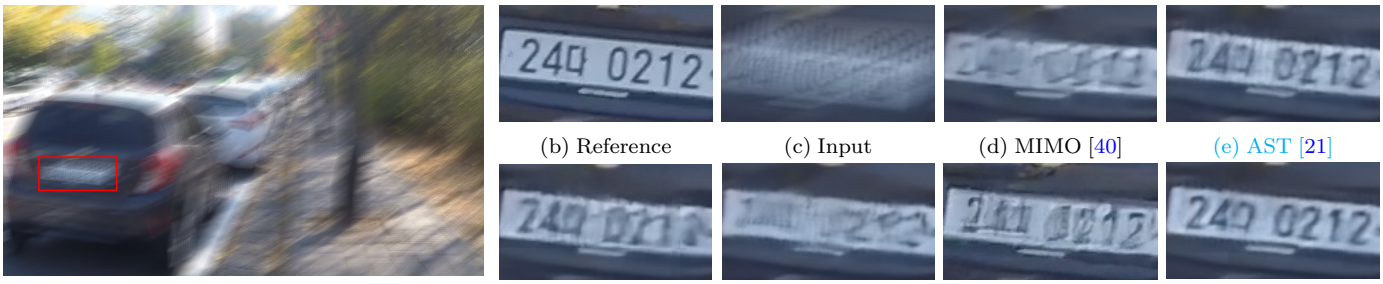
4.4 图像去阴影

对于阴影去除任务，我们在 ISTD 数据集 [72] 上进行了实验。如表 4所示，AST-v2 在 PSNR 和 SSIM 指标上和所对比的方法均表现出有竞争力。与近期提出的 DSP-FFANet 方法 [106] 相比，我们的方法在 PSNR 指标上取得了 0.32 dB 的显著性能提升。值得一提的是，即使与额外利用阴影掩码信息的算



(a) Reference (b) Input (c) DC [101] (d) LG [102] (e) AST [21] (f) AST-v2

图 8: 在 ISTD [72] 数据集上进行阴影去除任务的定性结果。相比之下，AST-v2 生成的无阴影图像不会引入明显的伪影。



(a) Blurry Image (b) Reference (c) Input (d) MIMO [40] (e) AST [21]

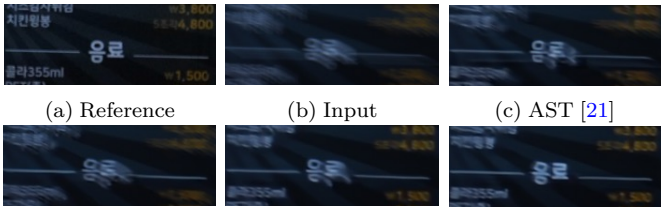
(f) CODE [16] (g) DegAE [103] (h) IR-SDE [104] (i) AST-v2 (L)

图 9: 在 GoPro [8] 数据集上针对合成运动模糊去除任务的定性结果。与其他所考虑的方法相比，AST-v2 (L) 生成的结果更清晰，模糊伪影也更少。

法 [83], [108], [109], [110] 相比，AST-v2 在全区域指标上仍至少获得 3.07 dB 的提升。图 8展示去阴影的定性比较结果。由于显著的阴影退化，部分方法 [21], [101] 未能成功生成清晰的图像（图 8c 和 8e）。我们注意到，尽管 LG [102] 能够有效去除阴影退化，但它会同时会留下残影（如图 8d 所示）。相比之下，AST-v2（图 8f）生成的无阴影图像没有引入明显的伪影。

4.5 图像去模糊

在图像去模糊实验中，我们在主流的 GoPro [8] 数据集上训练模型，并直接在 GoPro 和 RealBlur [74] 两个基准数据集上进行测试。我们在表 5中，从复原精度和模型复杂度两个方面，将所提出的方法与当前最先进的方案进行了比较。可以看出，我们的 AST-v2 (L) 在所有精度指标上都取得了最佳结果。与基于 CNN 的方法 MIMO [40]、基于 Transformer 的方法 CODE [16] 以及基于扩散模型的方法 IR-SDE [104] 相比，我们的网络 AST-v2 (L) 获得了至少 0.67 dB 的性能提升。



(a) Reference (b) Input (c) AST [21]

(d) Stripformer [87] (e) Restormer [11] (f) AST-v2 (L)

图 10: 在 RealBlur [74] 数据集上进行真实模糊伪影去除的定性结果。AST-v2 (L) 恢复的结果中，字符更为清晰。

可以注意到，与近期的修复方案 FPro [82] 相比，我们的方法在参数量更少（59.6%）的情况下，取得了更好的性能。在 GoPro 和 RealBlur 这两个基准数据集上的定性比较结果分别如图 9和图 10所示。我们的模型在合成与真实世界两种场景下，都展现了卓越的模糊模式去除能力，并且由所提方法去模糊后的结果也更忠实于参考图像。

图 11: 在 LOL-v2 [73] 数据集上进行暗光增强的定性结果。两个例子分别来自合成子集（上）和真实子集（下）。在颜色和纹理方面，经 AST-v2 增强后的结果在视觉上更接近参考图像。

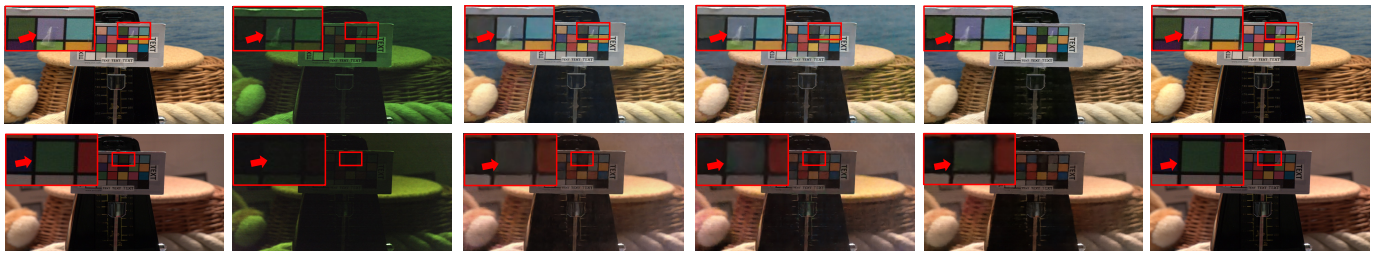


图 12: 在 SMID [9] 数据集上进行暗光增强的定性结果。与所对比的方法相比，我们方法生成的图像颜色更接近参考图像。

4.6 图像暗光增强

针对图像增强任务，我们首先在表 6中将本文提出的方法与当前最先进的方法在广泛使用的 LOL-v2 [73] 数据集上进行了比较。在 PSNR 和 SSIM 指标上，AST-v2 性能优于所对比的其他方法。在两个子集上取平均后，AST-v2 相比于首个基于 Transformer 架构用于低光照图像增强的算法 Retinexformer [118]，在 PSNR 上取得了 0.16 dB 的性能提升。与首个基于 Mamba 的图像修复方法 MambaIR [116] 相比，我们的模型获得了 1 dB 的增益。此外，我们的模型性能显著超越了近期的 QuadPrior 模型 [115]，高出 6.1 dB。所有这些结果都验证了 AST-v2 在暗光增强任务上的有效性。在 LOL-v2 [73] 数据集的合成子集和真实子集上的定性比较结果如图 11所示。在上方的案例中，以往的方法要么难以恢复真实颜色 [112], [113], [114]（如图 11d, 11f, 11h），要么存在色彩失真的问题 [11], [21]（如图 11e, 11g）；相比之下，我们的方法生成的结果则更为生动自然（如图 11i）。对于下方在真实世界场景中捕获的案例，AST-v2 生成了令人更满意的图像（如图 11i），既没有过曝/欠曝问题 [11], [112]（如图 11d, 11g），也没有引入模糊伪影 [73], [113]（如图 11f, 11h）。

此外，我们还在表 7中，进一步将 AST-v2 与当前最先进的方法在真实世界基准 SMID [9] 上进行了比较。AST-v2 在 PSNR 和 SSIM 两个指标上均取得了最佳分数。值得注意的是，与近期专为暗光增强设计的方法 MambaLLIE [119] 相比，我们的方法在 PSNR 上高出了 0.25 dB。同时，与包括基于 CNN 的 [114]、基于 Transformer 的 [11], [12], [21], [57] 以及基于 Mamba 的 [116] 通用图像修复方案相比，AST-v2 取得

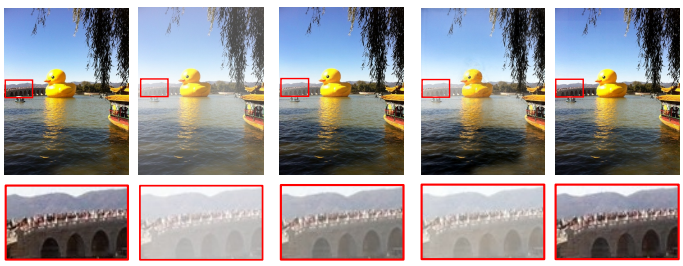
表 5: 在 GoPro [8] 和 RealBlur [74] 基准数据集上针对图像去模糊任务的定量比较。所有方法仅在 GoPro 数据集上训练。

Method	GoPro		RealBlur		Param.
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	(M)
DeblurGAN-v2 [41]	29.55	0.934	31.98	0.905	60.9
IR-SDE [104]	30.70	0.901	27.89	0.800	34.2
MT-RNN [111]	31.15	0.945	32.12	0.907	2.6
DegAE [103]	31.90	0.925	29.62	0.867	11.8
CODE [16]	31.94	0.928	30.03	0.870	12.2
MIMO [40]	32.45	0.957	31.59	0.892	16.1
MPRNet [78]	32.66	0.959	32.35	0.913	20.1
Restormer [11]	32.92	0.961	32.58	0.918	26.1
FPro [82]	33.05	0.961	30.82	0.903	22.3
Stripformer [87]	33.08	**0.962**	32.45	0.915	19.7
AST [21]	32.42	0.957	30.12	0.908	65.4
AST-v2 (L)	**33.12**	**0.962**	**32.62**	**0.921**	13.3

表 6: 在 LOL-v2 [73] 数据集上针对暗光增强任务的定量比较。

Method	LOL-v2-real		LOL-v2-syn		Average	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
KinD [113]	14.74	0.641	13.29	0.578	14.02	0.610
RUAS [112]	18.37	0.723	16.55	0.652	17.46	0.688
Uformer [12]	18.82	0.771	19.66	0.871	19.24	0.821
Restormer [11]	19.94	0.827	21.41	0.830	20.68	0.829
MIRNet [114]	20.02	0.820	21.94	0.876	20.98	0.848
Sparse [73]	20.06	0.816	22.05	0.905	21.06	0.861
QuadPrior [115]	20.48	0.811	16.11	0.758	18.30	0.785
MambaIR [116]	21.25	0.831	25.55	0.929	23.40	0.880
SNR-Net [117]	21.48	**0.849**	24.14	0.928	22.81	**0.889**
Retinexformer [118]	**22.80**	0.840	25.67	0.930	24.24	0.885
AST [21]	17.21	0.845	22.65	0.930	19.93	0.888
AST-v2	22.66	0.840	**26.13**	**0.937**	**24.40**	**0.889**

了至少 1.91 dB 的 PSNR 性能提升。在 SMID [9] 数据集上的定性比较结果如图 12所示。所对比的技术 [21], [117], [118] (如图 12c, 12d, 12e) 在恢复自然的色彩方面表现出局限性，例如上方样本中的紫色和下方样本中的绿色，并且这些问题在更暗的条件下（从上到下）会愈发严重。相比之下，AST-v2 生成的图像（图 12f）更接近参考图像（图 12a），提供了一种令人信服且视觉效果更佳的修复效果。



(a) Reference (b) Input (c) PromptIR [70] (d) AST [21] (e) AST-v2 (L)

图 13: 在 SOTS [71] 基准上进行去雾任务的定性结果。与其他被考虑的方法相比，AST-v2 (L) 恢复的结果更为清晰。

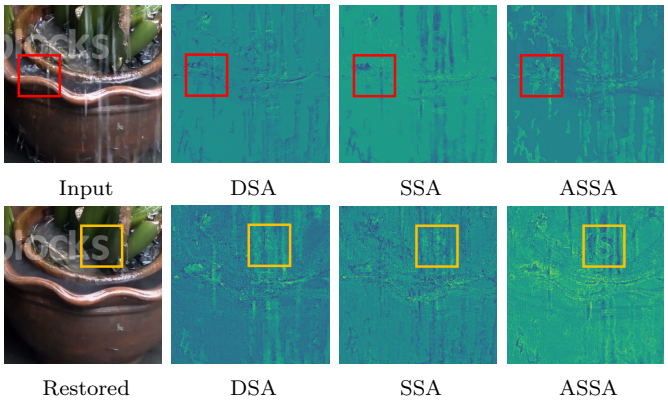


图 14: 特征图可视化。特征图提取自最后一个（顶部）和倒数第二个阶段特征（底部），用以展示采用不同自注意力机制的模型之间的差异。

4.7 图像去雾

针对图像去雾任务，我们在 SOTS [71] 数据集上进行实验。表 8表明，AST-v2 在 PSNR 和 SSIM 两个指标上都取得了最佳分数。与近期的图像修复方法 FSNet [126] 相比，我们的模型取得了 1.17 dB 的显著性能增益。尽管现有工作尝试通过引入多样化提示（prompt） [70], [125] 或探索退化的有效表征 [128] 来获得更好的修复结果，但它们的效果都不如 AST-v2。我们在图 13中提供了定性比较，其中 AST-v2 (L) 恢复出了一个更为清晰的结果（如图 13e 所示）。与其他会导致不必要伪影（图 13c）或色彩失真（图 13d）的方法相比，我们方法生成的图像更接近真实图像（图 13a）。

4.8 消融实验

前文的实验证明，探索有效特征并减少冗余特征能为不同的图像修复任务带来优越的结果。接下来，我们对 AST-v2 框架进行全面的分析，以证明所提出的模块在提升图像修复性能方面的有效性。我们通过在 SPAD [6] 基准上训练多种去雨模型进行对应的消融实验。为确保公平比较，所有模型均在 128×128 的图像块上训练了 300,000 次迭代，而 FLOPs 则是在输入尺寸为 256×256 的情况下评测。

ASSA 的有效性。为了评估所提出的 ASSA 的有效性，我们将其替换为其他同类注意力机制，包括：(1) 通道自注意力 (Channel SA) [11]，(2) Swin 自注意力 (Swin SA) [10]，(3) Top-k 自注意力 (Top-k SA) [15]，以及 (4) 压缩自注意力 (Condensed SA) [16]。定量结果总结在表 9中。与 Swin SA 和 Channel SA 相比，ASSA 在 PSNR 上分别取得了 0.72 dB 和 0.68 dB 的显著提升。此外，与那些旨在减轻令牌间噪声交互的相关方法相比，ASSA 模块在性能上比 Top-k SA 高出 0.24 dB，比 Condensed SA 高出 0.43 dB。

我们提出的自适应稀疏设计在有效保留关键特征的同时，也减轻了噪声特征并减少了冗余信息。为了评估该模块是否解决了由基于 ReLU 的注意力机制的过度稀疏特性所引发的

表 7: 在 SMID [9] 数据集上针对暗光增强任务的定量比较。

Methods	QuadPrior [115]	KinD [113]	Sparse [73]	MIRNet [114]	RUAS [112]	Restormer [11]	AST [21]
PSNR	15.50	22.18	25.48	25.66	25.88	26.97	27.60
SSIM	0.604	0.634	0.766	0.762	0.744	0.758	0.802
Methods	IPT [57]	MambaIR [116]	UFormer [12]	SNR-Net [117]	Retinexformer [118]	MambaLLIE [119]	AST-v2 (本文方法)
PSNR	27.03	27.07	27.20	28.49	29.15	<u>29.26</u>	**29.51**
SSIM	0.783	0.774	0.792	0.805	0.815	<u>0.818</u>	**0.824**

表 8: 在 SOTS [71] 数据集上针对去雾任务的定量比较。

Methods	MSCNN [120]	DehazeNet [121]	EPDN [122]	FDGAN [123]	AirNet [124]	InstructIR [125]	AST [21]
PSNR	22.06	22.46	22.57	23.15	23.18	30.22	27.94
SSIM	0.908	0.851	0.863	0.921	0.900	0.959	0.947
Methods	Restormer [11]	NAFNet [1]	FSNet [126]	PromptIR [70]	DehazeFormer [127]	NDR-Restore [128]	AST-v2 (L) (本文方法)
PSNR	30.87	30.98	31.11	31.31	31.78	<u>31.96</u>	**32.28**
SSIM	0.969	0.970	0.971	0.973	<u>0.977</u>	**0.980**	**0.980**

表 9: 分析实验：不同自注意力机制与所提 ASSA 的比较。

Models	Channel SA [11]	Swin SA [10]	Top-k SA [15]	Condensed SA [16]	ASSA
PSNR	48.20	48.16	48.64	48.45	48.82

表 10: 不同自注意力机制的熵分析。

Structure	DSA	SSA	ASSA
Entropy	3.2340	0.2993	3.0176
PSNR	48.20	48.48	48.82

信息丢失问题 [18]，我们遵循 [131] 的方法，计算了能够量化注意力集中程度的熵值。具体而言，其计算公式如下：

$$Entropy_{Att} = -\frac{1}{H}\sum_{h}\frac{1}{L}\sum_{ij}|Att_{ij}^{h,l}| \odot \log\left(|Att_{ij}^{h,l}|\right), \quad (6)$$

其中，$|Att_{ij}^{h,l}|$ 表示在第 $l \in L$ 层、第 $h \in H$ 个“头”(head) 中，查询令牌 i 和键令牌 j 之间相似度得分的绝对值。较低的熵值意味着注意力更集中，而较高的熵值则反映出注意力更为分散。如表 10所示，基于 softmax 的密集自注意力（DSA）取得了最高的熵分数，这表明它会更均匀地从所有源令牌（token）中提取特征，从而可能引入无关区域的噪声交互。相比之下，基于 ReLU 的稀疏自注意力（SSA）则表现出最低的熵分数，意味着它仅关注一个有限的令牌（token）集合，可能会导致其忽略掉一些必要的关系。我们提出的方法在两者之间达到了平衡：它既能有效捕获信息丰富的上下文，又能忽略冗余特征，从而带来了显著的性能提升。此外，我们可视化了配备不同自注意力机制的模型所生成的中间特征图，以探究所提 ASSA 的有效性。图 14分别展示了从最后一个（顶部）和倒数第二个阶段（底部）生成的特征。对比 DSA 或 SSA 的变体，ASSA 能够更精确地估计出退化成分（如红色矩形框内的雨痕），并恢复出更清晰的特征（如橙色区域内的字符），从而提升了整体的修复性能。

表 11: 关于自注意力机制中多种激活函数选择的消融实验。

Type	Dense			Sparse
Variety	Softmax	starReLu [129]	GELU [67]	$ReLU^2$
PSNR	48.20	48.68	48.42	48.48
Δ	-0.62	-0.14	-0.40	-0.34
Type	Sparse	Adaptive		
Variety	ReLU	ACON [130]	Meta-ACON [130]	ASSA
PSNR	48.57	48.63	48.64	48.82
Δ	-0.25	-0.19	-0.18	-

表 12: 分析实验：不同前馈网络与所提 FRFN 的比较。

Models	FFN [5]	DFN [132]	GDFN [11]	LeFF [12]	FRFN
PSNR	42.36	43.17	48.52	47.39	48.82

紧接着，我们通过实验来证明所提出的 ASSA 架构的必要性和优越性，其结果如表 11所示。ASSA 的核心思想是利用稀疏分数的指导来调整标准的密集注意力权重，为此我们训练了配备不同种激活函数的变体进行比较。与采用 ASSA 的模型相比，直接应用密集自注意力（例如 Softmax、starReLu [129] 和 GELU [67]）和稀疏自注意力（例如 $ReLU^2$ 和 ReLU）都会导致次优的性能。与 ACON 和 Meta-ACON [130] 等自适应激活函数相比，ASSA 获得了最高得分，48.82dB。

FRFN 的有效性。特征图，尤其是在网络深层的特征图，通常具有较高的通道维度，但并非所有通道都包含用于恢复清晰图像的关键信息。对所有通道应用统一的变换会带来引入冗余信息的风险，因此我们开发了 FRFN，来选择性地增强通道，以推动特征表示学习。为了评估提出的 FRFN 模块，我们将其与四种变体进行了比较，包括：(1) 原生前馈网络（FFN）[5]，(2) 配备了深度可分离卷积的前馈网络（DFN）[132]，(3) 门控深度可分离卷积前馈网络（GDFN）[11]，以及 (4) 局部增强前馈网络（LeFF）。定量比较结果呈现在表 12中，其中 FRFN 在 PSNR 上取得了最高分。尽管 GDFN [11] 和 FRFN

(a) Input (b) DRSformer [28] (c) IDT [81] (d) Restormer [11] (e) FPro [82] (f) AST [21] (g) AST-v2

图 15: 在 Internet-Data [6] 上进行真实去雨任务的定性比较。AST-v2 能有效去除退化，且不会像其他方案引入伪影。

表 13: 关于 FRFN 中分组数量的敏感性分析。

Groups	1	2	4	6
PSNR	48.42	48.76	48.51	48.64
Δ	-	+0.34	+0.09	+0.22
Groups	8	12	16	24
PSNR	48.56	48.82	48.53	48.29
Δ	+0.14	+0.40	+0.11	-0.13

表 14: 在真实世界场景下，针对去雨任务的无参考评估指标 NIQE 的结果。

Methods	Input	Uformer [12]	Restormer [11]	IDT [81]	DRSformer [15]	FPro [82]	AST-v2 (Ours)
NIQE ↓	5.403	5.193	5.048	5.151	5.238	5.152	4.821

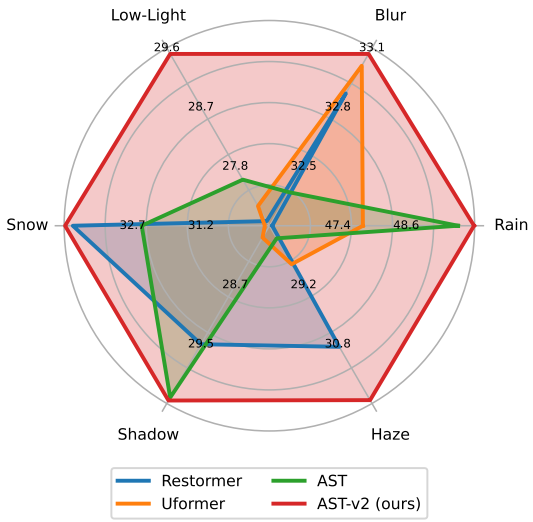


图 16: 在 6 个任务上与基线模型的 PSNR 比较。

都采用了门控机制，但我们的模块进一步采用了一种“增强-消除”设计，对有效特征进行更精细的选择，最终比 GDFN 取得了 0.3 dB 的 PSNR 增益。我们提出的 FRFN 沿通道维度采用部分深度可分离卷积，并将其划分为多个组。为了评估分组参数的影响，我们进行了相关实验，将分组数量设置从 1 到 24。如表 13所示，与基线模型（1 个组）相比，我们的模块在一定的分组数量范围（即 [2,16]）内都有良好表现。基于上述结果，我们在具体的实验中选择 12 个组以获得最佳性能。

4.9 分析与讨论

真实世界场景评估。我们遵循 [15] 的做法，从 Internet-Data [6] 基准数据集中随机选择了 20 张真实世界的雨天图像用于评估。如表 14所示，AST-v2 获得了最低的 NIQE 分数，这表明在真实世界条件下，其感知质量优于其他方法。此外，图 15中的定性比较表明，AST-v2 能有效去除雨痕退化，且不会像其他方案引入伪影，这证明了其处理复杂真实世界退化的能力。

与图像修复基线的比较。我们提供了与基线模型的比较，以证明所提出的 AST-v2 的有效性。具体来说，我们将 AST-v2 与 Uformer [12] 和 Restormer [11] 进行了比较，后两者均基于 Transformer 架构构建，并为通用图像修复任务而设计。此外，我们本文方法与这项工作的会议版本，即 AST [21]，进行了比较。如图 16所示，AST-v2 在所有任务上都超越了这些方法，并在 GoPro [8] 上的去模糊、SPAD [6] 上的雨痕去除、SOTS [71] 上的去雾、ISTD [72] 上的阴影去除、Snow100K [7] 上的去雪以及 SMID [9] 上的暗光增强任务中均取得最高分。

模型效率。为了衡量模型效率，我们针对图像去雨任务，比较了模型的性能（PSNR）、复杂度（FLOPs 和参数量）以及延迟（运行时间）。具体来说，FLOPs 和运行时间是在输入图像尺寸为 256×256 的情况下测量的，而 PSNR 分数则是在 SPAD 数据集 [6] 上评估的。如表 15所示，与 Restormer [11] 和 SwinIR [10] 相比，AST-v2 在 PSNR 方面取得了更好的分数，并且其模型复杂度更低。此外，我们的模型获得了次优的延迟，这一成绩超越了所有考虑的基于 Transformer 的方法，即 SwinIR [10]、Uformer-S [12]、Restormer [11]、IDT [81]、DRSformer [15] 和 FPro [82]。

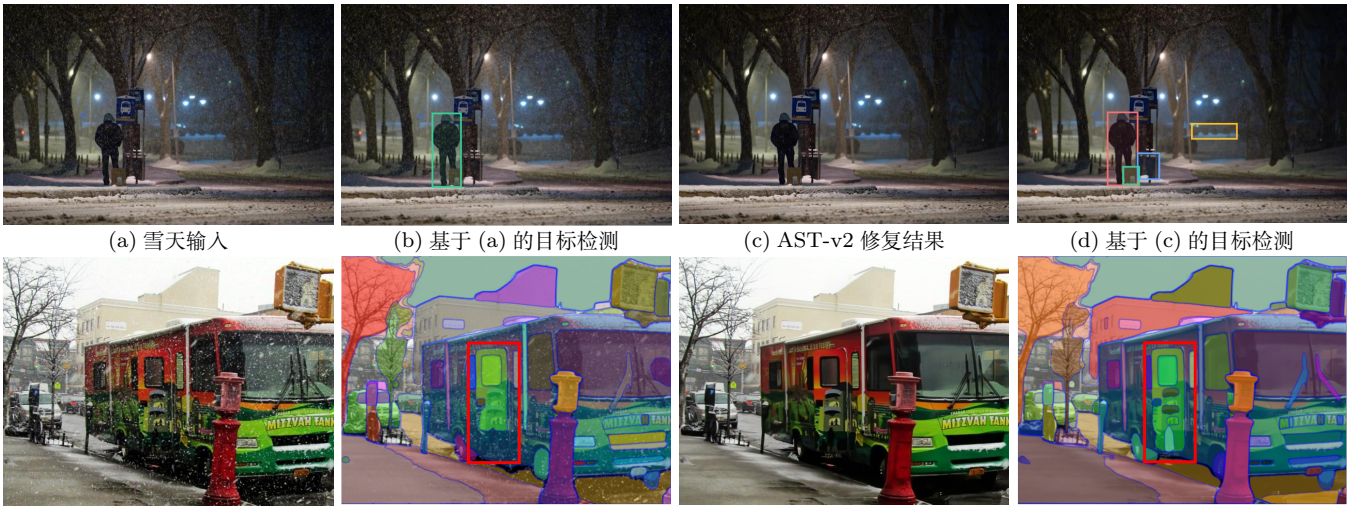
在表 16中，我们进一步比较了 AST-v2 与 AST [21] 在修复性能和模型复杂度方面的表现。结果表明，AST-v2 在准确性上优于 AST，实现了 0.28 dB 的 PSNR 提升，同时还将参

表 15: 在 SPAD [6] 上的模型效率分析。

Method	MPRNet [78]	SwinIR [10]	Uformer-S [12]	Restormer [11]	IDT [81]	DRSformer [15]	FPro [82]	AST-v2
FLOPs/G	175.8	238.0	**43.9**	174.7	61.9	242.9	81.9	90.7
Parameters/M	20.1	11.5	20.6	26.1	16.4	33.7	22.3	**8.7**
Run-times/s	**0.03**	1.83	0.12	0.14	0.28	0.08	0.08	0.05
PSNR/dB	43.64	44.97	46.13	47.98	47.34	48.53	48.99	**49.79**

表 16: 在 SPAD 基准 [6] 的图像去雨任务上，于相同实验设置下对 AST-v2 和 AST [21] 进行的比较。FLOPs 和推理时间是在 256×256 尺寸的图像上计算的。与 AST 相比，AST-v2 更准确，同时模型也显著更轻量、更快速。值得注意的是，模型复杂度上的这些改进在各种图像修复任务中是一致的。

Methods	PSNR (dB)	#Param. (M)	FLOPs (G)	Train Time (h)	Inference Time (ms)
AST [21]	49.51	65.36	110.55	215	104
AST-v2 (Ours)	49.79	8.73 (87% ↓)	90.66 (18% ↓)	121 (44% ↓)	53 (49% ↓)



(a) 雪天输入 (b) 基于 (a) 的目标检测 (c) AST-v2 修复结果 (d) 基于 (c) 的目标检测

(e) 雪天输入 (f) 基于 (e) 的语义分割 (g) AST-v2 修复结果 (h) 基于 (g) 的语义分割

图 17: 来自退化图像和修复后图像的高级计算机视觉任务结果。由 AST-v2 修复的图像能够提升下游任务（即目标检测和语义分割）的性能，其中，原先的失败样本（例如，修复后图像中的手提箱和公交车门）可以被成功地检测或分割。

表 17: ExDark 数据集上的低光照检测结果。

Class	Input	AST-v2	Δ
Bicycle	58.6	62.3	+3.7
Boat	43.0	49.6	+6.6
Bottle	43.1	50.7	+6.9
Bus	72.3	74.0	+1.7
Car	54.2	57.3	+3.1
Cat	37.6	45.9	+8.3
Chair	36.3	48.7	+12.4
Cup	39.6	47.8	+8.2
Dog	49.0	58.0	+9.0
Motorbike	35.1	37.4	+2.3
People	41.3	51.4	+10.1
Table	30.6	36.6	+6.0
Mean	45.1	51.7	+6.6

数量减少了 87%，FLOPs 减少了 18%。此外，AST-v2 的训练速度提升了 1.7 倍，推理速度提升了 1.9 倍。

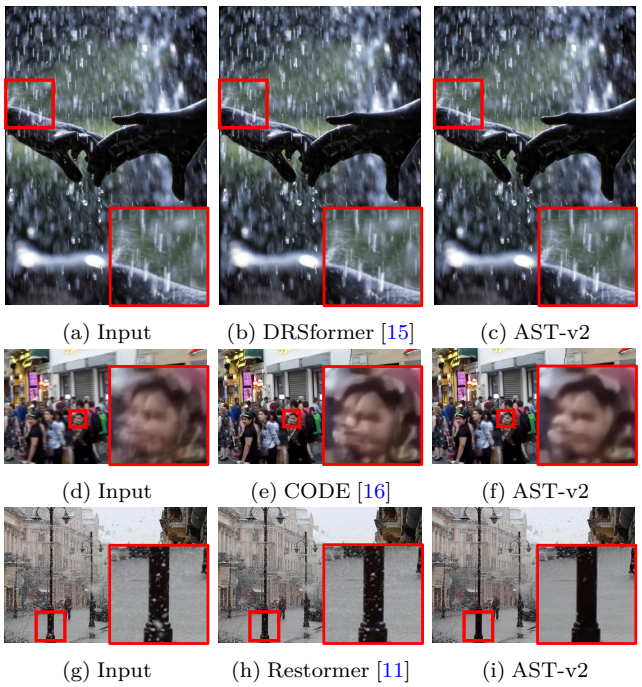
在高级视觉任务上的应用。如图 17所示，由 AST-v2 修复的图像能够提升下游高级视觉任务（例如，目标检测[1]和语义分割[2]）的性能。原本失败的样本（如顶行的手提箱和底行的公交车门）能够被成功地检测或分割。这一比较凸显了 AST-v2 的能力，不仅能提升修复图像的视觉质量，还能为后续的高级视觉任务带来增益。

我们在 ExDark [135] 基准上进行了暗光目标检测实验。AST-v2 在 LoL-v2 数据集上进行训练，并被直接用作 1473 张测试图像的预处理模块，同时使用预训练的 YOLO-v3 [136] 模型作为检测器。表 17中展示了平均精度（AP）的结果。总的来说，增强后的结果在所有类别上都取得了更好的分数，其中在“椅子”（Chair）这一类别上，性能提升最高可达 12.4。这些比较进一步凸显了 AST-v2 所取得的增益。

局限性。虽然 AST-v2 在多种图像修复任务上都展现了优异的性能，但仍存在几个未来可改进的方向。一个潜在的改进

1. 目标检测结果和置信度由 detr-resnet-50 模型 [133] 通过 Hugging Face API 支持：https://huggingface.co/facebook/detr-resnet-50。

2. 语义分割结果由 SAM（Segment Anything Model）[134] 通过在线 API 支持：https://segment-anything.com/。



(a) Input (b) DRSformer [15] (c) AST-v2

(d) Input (e) CODE [16] (f) AST-v2

(g) Input (h) Restormer [11] (i) AST-v2

图 18: AST-v2 的错误修复案例。典型的失败案例通常是由严重的退化引起。

方向是解决图 18中所示的 AST-v2 在处理严重退化时的失败案例。当输入的图像遭受了极其严重的退化时（例如，严重的模糊效应、密集的雨痕或厚重的雪花伪影），AST-v2 的输出会变得与退化输入几乎完全相同。在这些情况下，退化内容已经压倒性地破坏了语义内容，几乎没有留下任何可供模型利用的有效信号。无论模型配备了多么先进的机制（例如，标准的密集自注意力 [11]、稀疏自注意力 [15], [16]，或本文提出的自适应稀疏 ASSA），它们都难以在信息量较少的令牌之间建立可靠的关系，从而导致最终修复性能下降。值得一提的是，如果退化图像包含局部清晰的区域（例如，带有清晰背景的雪天场景），我们的模型仍然能够通过利用这些有效区域来恢复有用的信息（参见第三行的结果）。

5 总结

本文旨在通过自适应地学习最具信息量的表征、减轻特征中的噪声信息，来从其退化图像中恢复清晰图像。虽然我们引入了来自 NLP 领域的、基于 ReLU 的稀疏自注意力（SSA）来移除特征间不相关的交互，但我们的方法并未将其直接用作核心组件。相反，我们将重点放在了防止因基于 ReLU 的 SSA 的低熵特性而导致的信息丢失问题上。为了有效地实现这一目标，我们设计了一种自适应架构，通过一个互补的、基于标准 Softmax 的密集分支以像素级调制的方式提供辅助，从而确保关键信息得以保留。此外，我们提出了一种 FRFN，它通过“增强-消除”方案执行特征变换，从而能够学习到判别性特征，以改善高质量图像的重建效果。与那些使用 Top-K 选择、稀疏通道自注意力等操作来减少冗余，或将特征投影到超像素空间（如压缩自注意力）的相关基线模型相比，AST-v2 在多种退化去除任务上都取得了更优越的结果。

参考文献

[1] L. Chen, X. Chu, X. Zhang, and J. Sun, “Simple baselines for image restoration,” in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 17–33.

[2] J. Pan, D. Sun, H. Pfister, and M.-H. Yang, “Deblurring images via dark channel prior,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 10, pp. 2315–2328, 2018.

[3] K. Zhang, Y. Li, W. Zuo, L. Zhang, L. Van Gool, and R. Timofte, “Plug-and-play image restoration with deep denoiser prior,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6360–6376, 2022.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Adv. Neural Inform. Process. Syst.*, 2017, pp. 5998–6008.

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *Int. Conf. Learn. Represent.*, 2021.

[6] T. Wang, X. Yang, K. Xu, S. Chen, Q. Zhang, and R. W. Lau, “Spatial attentive single-image deraining with a high quality real rain dataset,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 12 270–12 279.

[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, 2018.

[8] S. Nah, T. Hyun Kim, and K. Mu Lee, “Deep multi-scale convolutional neural network for dynamic scene deblurring,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 3883–3891.

[9] C. Chen, Q. Chen, M. N. Do, and V. Koltun, “Seeing motion in the dark,” in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 3185–3194.

[10] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, “Swinir: Image restoration using swin transformer,” in *Proc. Int. Conf. Comput. Vis. Workshops*, 2021, pp. 1833–1844.

[11] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, “Restormer: Efficient transformer for high-resolution image restoration,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 5728–5739.

[12] 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 Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 17 683–17 693.

[13] P. Wang, X. Wang, F. Wang, M. Lin, S. Chang, H. Li, and R. Jin, “Kvt: k-nn attention for boosting vision transformers,” in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 285–302.

[14] G. Zhao, J. Lin, Z. Zhang, X. Ren, Q. Su, and X. Sun, “Explicit sparse transformer: Concentrated attention through explicit selection,” *arXiv preprint arXiv:1912.11637*, 2019.

[15] X. Chen, H. Li, M. Li, and J. Pan, “Learning a sparse transformer network for effective image deraining,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 5896–5905.

[16] H. Zhao, Y. Gou, B. Li, D. Peng, J. Lv, and X. Peng, "Comprehensive and delicate: An efficient transformer for image restoration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 14 122–14 132.

[17] D. So, W. Mańke, H. Liu, Z. Dai, N. Shazeer, and Q. V. Le, "Searching for efficient transformers for language modeling," in *Adv. Neural Inform. Process. Syst.*, 2021, pp. 6010–6022.

[18] K. Shen, J. Guo, X. Tan, S. Tang, R. Wang, and J. Bian, "A study on relu and softmax in transformer," *arXiv preprint arXiv:2302.06461*, 2023.

[19] M. Wortsman, J. Lee, J. Gilmer, and S. Kornblith, "Replacing softmax with relu in vision transformers," *arXiv preprint arXiv:2309.08586*, 2023.

[20] J. Hron, Y. Bahri, J. Sohl-Dickstein, and R. Novak, "Infinite attention: Nngp and ntk for deep attention networks," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 4376–4386.

[21] 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 Conf. Comput. Vis. Pattern Recog.*, 2024, pp. 2952–2963.

[22] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2341–2353, 2011.

[23] Y. Yang, J. Pan, Z. Peng, X. Du, Z. Tao, and J. Tang, "Bistnet: Semantic image prior guided bidirectional temporal feature fusion for deep exemplar-based video colorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 8, pp. 5612–5624, 2024.

[24] K. Yu, X. Wang, C. Dong, X. Tang, and C. C. Loy, "Path-restore: Learning network path selection for image restoration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 7078–7092, 2022.

[25] D. Liu, B. Wen, Y. Fan, C. C. Loy, and T. S. Huang, "Non-local recurrent network for image restoration," in *Adv. Neural Inform. Process. Syst.*, 2018, p. 1673—1682.

[26] F. Luo, X. Wu, and Y. Guo, "Functional neural networks for parametric image restoration problems," in *Adv. Neural Inform. Process. Syst.*, 2021, pp. 6762–6775.

[27] D. Ren, W. Zuo, Q. Hu, P. Zhu, and D. Meng, "Progressive image deraining networks: A better and simpler baseline," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 3937–3946.

[28] X. Fu, Q. Qi, Z.-J. Zha, Y. Zhu, and X. Ding, "Rain streak removal via dual graph convolutional network," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 1352–1360.

[29] D. Lin, X. Wang, J. Shen, R. Zhang, R. Liu, M. Wang, W. Xie, Q. Guo, and P. Li, "Generative status estimation and information decoupling for image rain removal," in *Adv. Neural Inform. Process. Syst.*, 2022, pp. 4612–4625.

[30] C.-L. Guo, Q. Yan, S. Anwar, R. Cong, W. Ren, and C. Li, "Image dehazing transformer with transmission-aware 3d position embedding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 5812–5820.

[31] M. Zhou, J. Huang, C.-L. Guo, and C. Li, "Fourmer: an efficient global modeling paradigm for image restoration," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 42 589–42 601.

[32] Y. Qiu, K. Zhang, C. Wang, W. Luo, H. Li, and Z. Jin, "Mb-taylorformer: Multi-branch efficient transformer expanded by taylor formula for image dehazing," in *Proc. Int. Conf. Comput. Vis.*, 2023, pp. 12 802–12 813.

[33] T. Ye, S. Chen, J. Bai, J. Shi, C. Xue, J. Jiang, J. Yin, E. Chen, and Y. Liu, "Adverse weather removal with codebook priors," in *Proc. Int. Conf. Comput. Vis.*, 2023, pp. 12 653–12 664.

[34] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, and Y. Li, "Maxim: Multi-axis mlp for image processing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 5769–5780.

[35] 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. 10 346–10 357, 2023.

[36] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Intervent.*, 2015, pp. 234–241.

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 770–778.

[38] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun, "What is the best multi-stage architecture for object recognition?" in *Proc. Int. Conf. Comput. Vis.*, 2009, pp. 2146–2153.

[39] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 4700–4708.

[40] S.-J. Cho, S.-W. Ji, J.-P. Hong, S.-W. Jung, and S.-J. Ko, "Rethinking coarse-to-fine approach in single image deblurring," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 4641–4650.

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

[42] Z. Yue, Q. Zhao, L. Zhang, and D. Meng, "Dual adversarial network: Toward real-world noise removal and noise generation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 41–58.

[43] X. Liu, M. Suganuma, Z. Sun, and T. Okatani, "Dual residual networks leveraging the potential of paired operations for image restoration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 7007–7016.

[44] Y. Zhang, K. Li, K. Li, B. Zhong, and Y. Fu, "Residual non-local attention networks for image restoration," in *Int. Conf. Learn. Represent.*, 2019.

[45] S. Gu, Y. Li, L. V. Gool, and R. Timofte, "Self-guided network for fast image denoising," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 2511–2520.

[46] X. Deng and P. L. Dragotti, "Deep convolutional neural network for multi-modal image restoration and fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3333–3348, 2021.

[47] X. Song, D. Zhou, W. Li, Y. Dai, Z. Shen, L. Zhang, and H. Li, "Tusr-net: Triple unfolding single image dehazing with self-regularization and dual feature to pixel attention," *IEEE Trans. Image Process.*, vol. 32, pp. 1231–1244, 2023.

[48] B. Niu, W. Wen, W. Ren, X. Zhang, L. Yang, S. Wang, K. Zhang, X. Cao, and H. Shen, "Single image super-resolution via a holistic attention network," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 191–207.

[49] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 286–301.

[50] A. Lugmayr, M. Danelljan, R. Timofte, K.-w. Kim, Y. Kim, J.-y. Lee, Z. Li, J. Pan, D. Shim, K.-U. Song, J. Tang, C. Wang,

and Z. Zhao, "Ntire 2022 challenge on learning the super-resolution space," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. Workshops*, 2022, pp. 786–797.

[51] Y. Wang, L. Wang, Z. Liang, J. Yang, R. Timofte, and Y. Guo, "Ntire 2023 challenge on light field image super-resolution: Dataset, methods and results," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. Workshops*, 2023, pp. 1320–1335.

[52] K. Zhang, W. Ren, W. Luo, W.-S. Lai, B. Stenger, M.-H. Yang, and H. Li, "Deep image deblurring: A survey," *Int. J. Comput. Vis.*, vol. 130, no. 9, pp. 2103–2130, 2022.

[53] C. Li, C. Guo, L. Han, J. Jiang, M.-M. Cheng, J. Gu, and C. C. Loy, "Low-light image and video enhancement using deep learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9396–9416, 2022.

[54] N. Liu, W. Li, Y. Wang, R. Tao, Q. Du, and J. Chanussot, "A survey on hyperspectral image restoration: From the view of low-rank tensor approximation," *Sci. China Inf. Sci.*, vol. 66, no. 4, p. 140302, 2023.

[55] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pvt v2: Improved baselines with pyramid vision transformer," *Comput. Vis. Media*, vol. 8, no. 3, pp. 415–424, 2022.

[56] C. Wang, J. Pan, W. Wang, J. Dong, M. Wang, Y. Ju, J. Chen, and X.-M. Wu, "Promptrestorer: A prompting image restoration method with degradation perception," in *Adv. Neural Inform. Process. Syst.*, 2023, pp. 8898–8912.

[57] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, and W. Gao, "Pre-trained image processing transformer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 12 299–12 310.

[58] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 10 012–10 022.

[59] Y. Li, Y. Fan, X. Xiang, D. Demandolx, R. Ranjan, R. Timofte, and V. L. Gool, "Efficient and explicit modelling of image hierarchies for image restoration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 18 278–18 289.

[60] B. Zhang, I. Titov, and R. Sennrich, "Sparse attention with linear units," in *Conf. Empir. Methods Nat. Lang. Process.*, 2021, pp. 6507–6520.

[61] W. Hua, Z. Dai, H. Liu, and Q. Le, "Transformer quality in linear time," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 9099–9117.

[62] Z. Li, S. Bhojanapalli, M. Zaheer, S. Reddi, and S. Kumar, "Robust training of neural networks using scale invariant architectures," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 12 656–12 684.

[63] N. Park and S. Kim, "How do vision transformers work?" in *Int. Conf. Learn. Represent.*, 2022.

[64] T. Xiao, M. Singh, E. Mintun, T. Darrell, P. Dollar, and R. Girshick, "Early convolutions help transformers see better," in *Adv. Neural Inform. Process. Syst.*, 2021, pp. 30 392–30 400.

[65] P. Charbonnier, L. Blanc-Feraud, G. Aubert, and M. Barlaud, "Two deterministic half-quadratic regularization algorithms for computed imaging," in *Proc. IEEE Int. Conf. Image Process.*, 1994, pp. 168–172.

[66] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[67] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[68] J. Chen, S.-h. Kao, H. He, W. Zhuo, S. Wen, C.-H. Lee, and S.-H. G. Chan, "Run, don't walk: Chasing higher flops for faster neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 12 021–12 031.

[69] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[70] V. Potlapalli, S. W. Zamir, S. H. Khan, and F. Shahbaz Khan, "Promptir: Prompting for all-in-one image restoration," in *Adv. Neural Inform. Process. Syst.*, 2023, pp. 71 275–71 293.

[71] B. Li, W. Ren, D. Fu, D. Tao, D. Feng, W. Zeng, and Z. Wang, "Benchmarking single-image dehazing and beyond," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 492–505, 2018.

[72] J. Wang, X. Li, and J. Yang, "Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 1788–1797.

[73] W. Yang, W. Wang, H. Huang, S. Wang, and J. Liu, "Sparse gradient regularized deep retinex network for robust low-light image enhancement," *IEEE Trans. Image Process.*, vol. 30, pp. 2072–2086, 2021.

[74] J. Rim, H. Lee, J. Won, and S. Cho, "Real-world blur dataset for learning and benchmarking deblurring algorithms," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 184–201.

[75] H. Wang, Q. Xie, Q. Zhao, and D. Meng, "A model-driven deep neural network for single image rain removal," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 3103–3112.

[76] Q. Yi, J. Li, Q. Dai, F. Fang, G. Zhang, and T. Zeng, "Structure-preserving deraining with residue channel prior guidance," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 4238–4247.

[77] K. Purohit, M. Suin, A. Rajagopalan, and V. N. Boddeti, "Spatially-adaptive image restoration using distortion-guided networks," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 2309–2319.

[78] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, "Multi-stage progressive image restoration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 14 821–14 831.

[79] X. Fu, J. Xiao, Y. Zhu, A. Liu, F. Wu, and Z.-J. Zha, "Continual image deraining with hypergraph convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 9534–9551, 2023.

[80] Y. Guo, X. Xiao, Y. Chang, S. Deng, and L. Yan, "From sky to the ground: A large-scale benchmark and simple baseline towards real rain removal," in *Proc. Int. Conf. Comput. Vis.*, 2023, pp. 12 097–12 107.

[81] J. Xiao, X. Fu, A. Liu, F. Wu, and Z.-J. Zha, "Image de-raining transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 11, pp. 12 978–12 995, 2023.

[82] S. Zhou, J. Pan, J. Shi, D. Chen, L. Qu, and J. Yang, "Seeing the unseen: A frequency prompt guided transformer for image restoration," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 246–264.

[83] L. Fu, C. Zhou, Q. Guo, F. Juefei-Xu, H. Yu, W. Feng, Y. Liu, and S. Wang, "Auto-exposure fusion for single-image shadow

removal," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 10 571–10 580.

[84] Y. Cui, Y. Tao, Z. Bing, W. Ren, X. Gao, X. Cao, K. Huang, and A. Knoll, "Selective frequency network for image restoration," in *Int. Conf. Learn. Represent.*, 2023.

[85] J. M. J. Valanarasu, R. Yasarla, and V. M. Patel, "Transweather: Transformer-based restoration of images degraded by adverse weather conditions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 2353–2363.

[86] W. Zou, M. Jiang, Y. Zhang, L. Chen, Z. Lu, and Y. Wu, "Sdwnet: A straight dilated network with wavelet transformation for image deblurring," in *Proc. Int. Conf. Comput. Vis. Workshops*, 2021, pp. 1895–1904.

[87] F.-J. Tsai, Y.-T. Peng, Y.-Y. Lin, C.-C. Tsai, and C.-W. Lin, "Stripformer: Strip transformer for fast image deblurring," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 146–162.

[88] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Int. Conf. Learn. Represent.*, 2019.

[89] L. Ilya and H. Frank, "SGDR: Stochastic gradient descent with warm restarts," in *Int. Conf. Learn. Represent.*, 2017.

[90] Z. Wang, 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. 4, pp. 600–612, 2004.

[91] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Sign. Process. Letters*, vol. 20, no. 3, pp. 209–212, 2012.

[92] 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.

[93] R. Li, R. T. Tan, and L.-F. Cheong, "All in one bad weather removal using architectural search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 3175–3185.

[94] D. Engin, A. Genc, and H. Kemal Ekenel, "Cycle-dehaze: Enhanced cyclegan for single image dehazing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. Workshops*, 2018, pp. 825–833.

[95] 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.

[96] W.-T. Chen, H.-Y. Fang, C.-L. Hsieh, C.-C. Tsai, I. Chen, J.-J. Ding, S.-Y. Kuo *et al.*, "All snow removed: Single image desnowing algorithm using hierarchical dual-tree complex wavelet representation and contradict channel loss," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 4196–4205.

[97] B. Cheng, J. Li, Y. Chen, and T. Zeng, "Snow mask guided adaptive residual network for image snow removal," *Comput. Vis. Image Underst.*, vol. 236, p. 103819, 2023.

[98] S. Chen, T. Ye, Y. Liu, T. Liao, J. Jiang, E. Chen, and P. Chen, "Msp-former: Multi-scale projection transformer for single image desnowing," in *IEEE Int. Conf. Acoust. Speech Signal Process.*, 2023, pp. 1–5.

[99] Y. Cui, W. Ren, X. Cao, and A. Knoll, "Focal network for image restoration," in *Proc. Int. Conf. Comput. Vis.*, 2023, pp. 13 001–13 011.

[100] 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, 2024.

[101] Y. Jin, A. Sharma, and R. T. Tan, "Dc-shadownet: Single-image hard and soft shadow removal using unsupervised domain-classifier guided network," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 5027–5036.

[102] Z. Liu, H. Yin, Y. Mi, M. Pu, and S. Wang, "Shadow removal by a lightness-guided network with training on unpaired data," *IEEE Trans. Image Process.*, vol. 30, pp. 1853–1865, 2021.

[103] Y. Liu, J. He, J. Gu, X. Kong, Y. Qiao, and C. Dong, "Degae: A new pretraining paradigm for low-level vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 23 292–23 303.

[104] Z. Luo, F. K. Gustafsson, Z. Zhao, J. Sjölund, and T. B. Schön, "Image restoration with mean-reverting stochastic differential equations," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 23 045–23 066.

[105] Z. Liu, H. Yin, X. Wu, Z. Wu, Y. Mi, and S. Wang, "From shadow generation to shadow removal," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 4927–4936.

[106] N. P. D. Gallego, J. Ilao, M. I. Cordel, and C. Ruiz, "Training a shadow removal network using only 3d primitive occluders," *The Visual Computer*, 2024.

[107] J. Liu, Q. Wang, H. Fan, W. Li, L. Qu, and Y. Tang, "A decoupled multi-task network for shadow removal," *IEEE Trans. Multimedia*, vol. 25, pp. 9449–9463, 2023.

[108] H. Le and D. Samaras, "From shadow segmentation to shadow removal," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 264–281.

[109] H. Le and D. Samaras, "Physics-based shadow image decomposition for shadow removal," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 9088–9101, 2022.

[110] J. Wan, H. Yin, Z. Wu, X. Wu, Y. Liu, and S. Wang, "Style-guided shadow removal," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 361–378.

[111] D. Park, D. U. Kang, J. Kim, and S. Y. Chun, "Multi-temporal recurrent neural networks for progressive non-uniform single image deblurring with incremental temporal training," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 327–343.

[112] R. Liu, L. Ma, J. Zhang, X. Fan, and Z. Luo, "Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 10 561–10 570.

[113] Y. Zhang, J. Zhang, and X. Guo, "Kindling the darkness: A practical low-light image enhancer," in *Proc. ACM Int. Conf. Multimed.*, 2019, pp. 1632–1640.

[114] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, "Learning enriched features for real image restoration and enhancement," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 492–511.

[115] W. Wang, H. Yang, J. Fu, and J. Liu, "Zero-reference low-light enhancement via physical quadruple priors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2024, pp. 26 057–26 066.

[116] H. Guo, J. Li, T. Dai, Z. Ouyang, X. Ren, and S.-T. Xia, "Mambair: A simple baseline for image restoration with state-space model," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 222–241.

[117] X. Xu, R. Wang, C.-W. Fu, and J. Jia, "Snr-aware low-light image enhancement," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 17 714–17 724.

[118] Y. Cai, H. Bian, J. Lin, H. Wang, R. Timofte, and Y. Zhang, "Retinexformer: One-stage retinex-based transformer for low-light image enhancement," in *Proc. Int. Conf. Comput. Vis.*, 2023, pp. 12 504–12 513.

[119] J. Weng, Z. Yan, Y. Tai, J. Qian, J. Yang, and J. Li, "Mambalie: Implicit retinex-aware low light enhancement with global-then-local state space," in *Adv. Neural Inform. Process. Syst.*, 2024.

[120] W. Ren, S. Liu, H. Zhang, J. Pan, X. Cao, and M.-H. Yang, "Single image dehazing via multi-scale convolutional neural networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 154–169.

[121] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao, "Dehazenet: An end-to-end system for single image haze removal," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5187–5198, 2016.

[122] Y. Qu, Y. Chen, J. Huang, and Y. Xie, "Enhanced pix2pix dehazing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 8160–8168.

[123] Y. Dong, Y. Liu, H. Zhang, S. Chen, and Y. Qiao, "Fd-gan: Generative adversarial networks with fusion-discriminator for single image dehazing," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 10 729–10 736.

[124] B. Li, X. Liu, P. Hu, Z. Wu, J. Lv, and X. Peng, "All-in-one image restoration for unknown corruption," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 17 452–17 462.

[125] M. V. Conde, G. Geigle, and R. Timofte, "Instructir: High-quality image restoration following human instructions," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 1–21.

[126] Y. Cui, W. Ren, X. Cao, and A. Knoll, "Image restoration via frequency selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 2, pp. 1093–1108, 2024.

[127] Y. Song, Z. He, H. Qian, and X. Du, "Vision transformers for single image dehazing," *IEEE Trans. Image Process.*, vol. 32, pp. 1927–1941, 2023.

[128] M. Yao, R. Xu, Y. Guan, J. Huang, and Z. Xiong, "Neural degradation representation learning for all-in-one image restoration," *IEEE Trans. Image Process.*, vol. 33, pp. 5408–5423, 2024.

[129] W. Yu, C. Si, P. Zhou, M. Luo, Y. Zhou, J. Feng, S. Yan, and X. Wang, "Metaformer baselines for vision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 2, pp. 896–912, 2024.

[130] N. Ma, X. Zhang, M. Liu, and J. Sun, "Activate or not: Learning customized activation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 8032–8042.

[131] H. Ghader and C. Monz, "What does attention in neural machine translation pay attention to?" *arXiv preprint arXiv:1710.03348*, 2017.

[132] Y. Li, K. Zhang, J. Cao, R. Timofte, and L. Van Gool, "Localvit: Bringing locality to vision transformers," *arXiv preprint arXiv:2104.05707*, 2021.

[133] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.

[134] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proc. Int. Conf. Comput. Vis.*, 2023, pp. 4015–4026.

[135] Y. P. Loh and C. S. Chan, "Getting to know low-light images with the exclusively dark dataset," *Comput. Vis. Image Underst.*, vol. 178, pp. 30–42, 2019.

[136] J. Redmon, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.