适应或消亡：自适应稀疏Transformer与注意力特征细化用于图像恢复

周时豪[1,2] 陈多生[1] 潘金山[3] 石锦磊[1*] 杨聚峰[1,2] VCIP & TMCC & DISSec, 计算机科学学院, 南开大学[2] 南开国际先进研究中心（深圳·福田）[3] 南京科技大学计算机科学与工程学院zhoushihao96@mail.nankai.edu.cn, duoshengchen@mail.nankai.edu.cn, sdluran@gmail.com jinglei.shi@nankai.edu.cn,yangjufeng@nankai.edu.cn

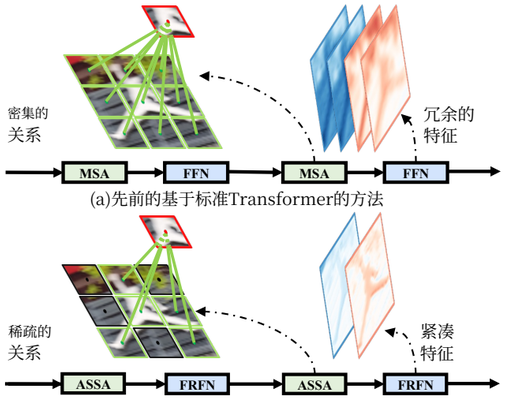
摘要

基于*Transformer*的方法在图像恢复任务中取得了显著的性能，这得益于它们建模长距离依赖的能力，这对于恢复清晰图像至关重要。尽管各种高效的注意力机制设计已经解决了与使用*Transformer*相关的大量计算，但它们通常通过考虑所有可用*token*而涉及来自无关区域的冗余信息和噪声交互。在这项工作中，我们提出了一种自适应稀疏***Transformer (****AST****)*** 来减轻无关区域的噪声交互，并在空间和通道域中消除特征冗余。***AST***包含两个核心设计，即，一个自适应稀疏自注意力块 ***(ASSA)*** 和一个特征精炼前馈网络 ***(FRFN)***。具体来说，***ASSA***采用双分支范式进行自适应计算，其中稀疏分支用于过滤低查询-键匹配分数对聚合特征的负面影响，而密集分支确保通过网络的信息流以学习判别性表示。同时，***FRFN***采用增强和缓解方案来消除通道中的特征冗余，增强清晰潜在图像的恢复。在常用基准数据集上的实验结果证明了我们方法在多个任务中的通用性和竞争性能，包括雨迹去除、真实雾霾去除和雨滴去除。代码和预训练模型可在 *https://github.com/joshyZhou/AST*。

1. 简介

图像恢复旨在从退化图像中恢复清晰图像。现有的基于CNN的方法 [6, 55, 103]取得了显著进展。然而，它们的基本单元

*通讯作者。



(b)我们的自适应稀疏Transformer(AST)模型 图1. 工作流程比较。(a)基于标准Transformer的方法将所有可用token纳入多头自注意力(MSA)计算，并将前馈网络(FFN)用于处理冗余特征。(b)提出自适应稀疏Transformer(AST)包括一个自适应稀疏自注意力(ASSA)模块来过滤无关token中的噪声交互，以及一个特征精炼前馈网络(FRFN)来减少通道中隐藏的冗余。

卷积，具有受限的感受野，在建模长距离依赖时效果较差。虽然最近的基于Transformer的 [73]架构通过结合自注意力机制来探索全局相关性，从而解决了这一限制，但在实际应用中却存在高计算复杂度的问题。

尽管尝试设计高效的注意力机制 [37, 82, 100] 来应对计算挑战，但仍然存在两个障碍：1) 标准Transformer [37, 100]采用密集的注意力关系来聚合特征，这会在无关区域无意中引入噪声交互，如图 1 所示。2) 密集聚合的特征 [77, 113]内存在冗余信息

Adapt or Perish: Adaptive Sparse Transformer with Attentive Feature Refinement for Image Restoration

Shihao Zhou[1,2] Duosheng Chen[1] Jinshan Pan[3] Jinglei Shi[1*] Jufeng Yang[1,2]
[1] VCIP & TMCC & DISSec, College of Computer Science, Nankai University
[2] Nankai International Advanced Research Institute (SHENZHEN· FUTIAN)
[3] School of Computer Science and Engineering, Nanjing University of Science and Technology
zhoushihao96@mail.nankai.edu.cn, duoshengchen@mail.nankai.edu.cn, sdluran@gmail.com
jinglei.shi@nankai.edu.cn,yangjufeng@nankai.edu.cn

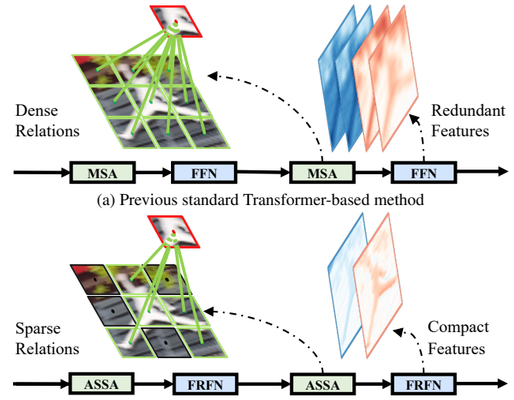
Abstract

*Transformer-based approaches have achieved promising performance in image restoration tasks, given their ability to model long-range dependencies, which is crucial for recovering clear images. Though diverse efficient attention mechanism designs have addressed the intensive computations associated with using transformers, they often involve redundant information and noisy interactions from irrelevant regions by considering all available tokens. In this work, we propose an **A**daptive **S**parse **T**ransformer (**AST**) to mitigate the noisy interactions of irrelevant areas and remove feature redundancy in both spatial and channel domains. AST comprises two core designs,* i.e.*, an Adaptive Sparse Self-Attention (ASSA) block and a Feature Refinement Feed-forward Network (FRFN). Specifically, ASSA is adaptively computed using a two-branch paradigm, where the sparse branch is introduced to filter out the negative impacts of low query-key matching scores for aggregating features, while the dense one ensures sufficient information flow through the network for learning discriminative representations. Meanwhile, FRFN employs an enhance-and-ease scheme to eliminate feature redundancy in channels, enhancing the restoration of clear latent images. Experimental results on commonly used benchmarks have demonstrated the versatility and competitive performance of our method in several tasks, including rain streak removal, real haze removal, and raindrop removal. The code and pre-trained models are available at https://github.com/joshyZhou/AST.*

1. Introduction

Image restoration aims to restore clear images from degraded ones. Existing CNN-based methods [6, 55, 103] achieve remarkable progress. However, their basic unit,

*Corresponding Author.



(b) Our Adaptive Sparse Transformer (AST) model

Figure 1. Workflows comparisons. (a) The standard Transformer-based method incorporates all available tokens into multihead self-attention (MSA) calculation, and the feed-forward network (FFN) to handle redundant features. (b) The proposed Adaptive Sparse Transformer (AST) includes an adaptive sparse self-attention (ASSA) block to filter out noisy interactions from irrelevant tokens, and a feature refinement feed-forward network (FRFN) to reduce the redundancy hidden in channels.

convolution, possesses a restricted receptive field and is less effective when modeling long-range dependencies. While recent Transformer-based [73] architecture addresses this limitation by incorporating the self-attention mechanism to explore global correlations, it suffers from high computational complexity in practical applications.

Despite attempts to design efficient attention mechanisms [37, 82, 100] to tackle the computational challenge, roadblocks persist for two reasons: 1) Standard Transformers [37, 100] adopt dense attention relations to aggregate features, which will inadvertently introduce noisy interactions in irrelevant regions as shown in Fig. 1. 2) Redundant information [77, 113] within densely aggregated feature

地图会进一步阻碍模型关注信息量大的特征。最近，人们尝试在无关区域过滤噪声交互并移除特征表示[8, 114]内的冗余信息 [8],。这些方法要么采用Top-K选择操作来选择最有用的token [8],，要么在执行自注意力计算之前将特征图投影到超像素空间[114]。由于参数K可能对特定的恢复任务敏感，并且超像素空间中的自注意力机制考虑了所有token之间的关系，它们仍然可能遇到与特征图冗余相关的挑战。

在实践中，设计一种高效机制，能够在信息流中识别最有价值特征的同时，对特定恢复任务表现出较低的敏感性。标准Transformer [82, 100] 通常会考虑所有查询-键对注意力关系来聚合特征。不幸的是，由于并非所有查询词元都与键中的相应词元密切相关，因此利用所有相似度对于清晰图像重建是无效的。直观地，开发一个稀疏Transformer来选择词元之间的最有用交互，可以增强特征聚合。为了实现注意力的稀疏性，基于平方ReLU激活 [67] 似乎是一个可行的方案。它移除了具有负相关性的相似度，而无需考虑特定参数设置，如 [8]。然而，通常需要一些特定设计 [23, 85] 来放宽稀疏性，以减轻信息损失 [66],，这与使用稀疏自注意力而不是标准密集注意力的动机相矛盾。因此，我们探索另一种范式，以确保噪声表示特征被尽可能减少，而信息特征被尽可能保留。

基于此，我们提出了一种名为 **A**daptive **S**parse **T**ransformer (**AST**) 的高效 Transformer-based 模型用于图像恢复。AST 引入了两个关键模块：自适应稀疏自注意力模块 (ASSA) 和特征细化前馈网络 (FRFN)。简而言之，ASSA 由两个分支组成：稀疏自注意力分支 (SSA) 和密集自注意力分支 (DSA)。具体来说，SSA 用于过滤掉 token 之间的无关交互，而 DSA 则用于确保必要的信息流通过整个网络。我们以自适应方式为每个分支分配权重，使模型能够适应两个分支的影响。与标准自注意力方法相比，这种设计带来了更有效的特征聚合，但计算负担有限。

另一方面，我们开发了一种简单而有效的常规前馈网络 [11], 的替代方案，即FRFN，以增强特征表示，以实现更好的潜在图像恢复。简而言之，FRFN采用增强和缓解方案进行特征转换。它增强了特征图的信息部分，然后利用门控机制减少冗余。同时，FRFN在通道维度上补充抑制冗余信息，而ASSA在空间域中减少冗余。得益于这两个互补组件的合作，AST捕获了最具有代表性的特征，同时在某种程度上抑制了不太具有信息量的特征。

总体而言，这项工作的主要贡献有三点：

- 我们提出了 AST，一种高效 Transformer-based 模型，它促进了最有用信息的流动，为清晰图像的恢复提取了更具建设性的特征。
- AST集成了一个ASSA模块，该模块包括一个密集自注意力分支和一个稀疏分支，以自适应地捕获标记之间的信息交互，同时保留重要信息。此外，我们基于一种特征变换方案开发了一种新的特征精炼前馈网络（FRFN），即，增强有价值特征的同时抑制信息量较少的特征。
- 进行了全面的实验，以去除雨迹、雾霾和雨滴等多种类型的退化，展示了我们AST设计的优越性。此外，我们提供了广泛的消融实验，以突出设计贡献。

2. 相关工作

图像恢复。 高质量的图像对于实现下游应用的满意性能至关重要，例如识别 [28, 76, 101], 分割 [97, 108, 110], 表征学习 [42, 84, 112], 和以图像 [45, 83, 115] 和视频 [107, 109, 111]形式的重建。在过去的几十年里，研究界见证了从传统的基于先验的模型 [20, 92, 103] 到基于学习的方法 [40, 50, 95], 的巨大范式转变，因为它们在去除各种退化方面表现出色，例如雨迹 [14, 39,63], 雾 [18, 60, 116], 雨滴 [54, 71, 93], 等。性能的提升可以归因于受高级视觉任务启发的多样化架构结构 [64] 和先进组件 [21, 25, 27] 。例如，U形网络设计和跳跃连接被广泛应用于获取层次多尺度表示 [9, 29, 98]和学习残差信号 [17, 44, 106]。尽管基于CNN的网络取得了令人印象深刻的结果，但它们仍然受到卷积操作感受野有限的问题的困扰。为了解决这个问题，最近的工作 [10, 53, 68]探索了注意力机制以获得更好的恢复性能。例如，SPANet [78]将IRNN模型扩展为显式生成雨迹的注意力图。RCAN [105] 设计了一种通道注意力机制来强调更多信息特征。更多的网络架构设计总结在NTIRE挑战报告 [49, 80] 和最近的综述 [31, 41, 104]中。

maps can further impede the models from attending to informative features. Recently, efforts have been made to filter out noisy interactions in irrelevant areas and remove the redundant information within feature representations [8, 114]. These methods either employ a Top-K selection operation to choose the most useful tokens [8], or project the feature map into the superpixel space before performing self-attention calculation [114]. As the parameter K can be sensitive to specific restoration tasks, and the self-attention mechanism conducted in superpixel space considers relations among all tokens, they may still encounter challenges related to feature map redundancy.

In practice, designing an efficient mechanism that identifies the most valuable features within information flows while exhibiting less sensitivity to specific restoration tasks. Standard Transformers [82, 100] usually consider all query-key pair attention relations to aggregate features. Unfortunately, since not all query tokens are closely relevant to corresponding ones in the keys, the utilization of all similarities is ineffective for clear image reconstruction. Intuitively, developing a sparse Transformer to select the most useful interactions among the tokens could enhance feature aggregation. For achieving sparsity in attention, squared ReLU-based activation [67] seems to be a feasible solution. It removes the similarities with negative relevance without considering specific parameter settings like [8]. However, some specific designs [23, 85] are often demanded to relax the sparsity for alleviating the information loss [66], which contradicts the motivation of using sparse self-attention over the standard dense one. Hence, we explore another paradigm to ensure that noisy representation features are reduced, and informative ones are retained as far as possible.

In light of this, we propose an efficient Transformer-based model named **A**daptive **S**parse **T**ransformer (**AST**) for image restoration. AST introduces two key modules: an Adaptive Sparse Self-Attention block (ASSA) and a Feature Refinement Feed-forward Network (FRFN). In brief, ASSA consists of two branches: a sparse self-attention branch (SSA) and a dense self-attention counterpart (DSA). Specifically, SSA is leveraged to filter out irrelevant interactions among tokens, while the DSA is adopted to ensure necessary information flows through the whole network. We assign weights to each branch in an adaptive fashion, allowing the model to adapt to the influence of the two branches. This design leads to a more effective feature aggregation but limited computation burdens compared to standard self-attention methods.

On the other hand, we develop a simple yet effective alternative to the regular feed-forward network [11], *i.e.*, FRFN, to enhance the feature representation for better latent image restoration. In a nutshell, FRFN performs feature transformation with an enhance-and-ease scheme. It enhances the informative part of the feature maps and then reduces redundancy using a gate mechanism. Meanwhile, FRFN complements ASSA in suppressing redundant information along channel dimensions, whereas ASSA reduces redundancy in the spatial domain. Thanks to the cooperation of the two complementary components, AST captures the most representative features, while simultaneously suppressing less informative ones to some extent.

Overall, key contributions of this work are three folds:

- We present AST, an efficient Transformer-based model, that facilitates the flow of the most useful information forward, extracting more constructive features for the recovery of clear images.
- AST incorporates an ASSA block, which includes a dense self-attention branch and a sparse one, to adaptively capture informative interactions among tokens while preserving essential information. Moreover, we develop a new feature refinement feed-forward network (FRFN) based on a feature transformation scheme, *i.e.*, enhancing the valuable features while suppressing less informative ones.
- Comprehensive experiments are performed to remove degradations of several types: rain-streaks, hazes, and raindrops, showing the superiority of our AST design. Furthermore, we provide extensive ablation studies to highlight the design contributions.

2. Related Work

Image Restoration. High-quality images are crucial to achieve satisfactory performance for downstream applications, such as recognition [28, 76, 101], segmentation [97, 108, 110], representation learning [42, 84, 112], and reconstrucion [117, 118] in forms of image [45, 83, 115] and video [107, 109, 111]. In the past decades, the research community has witnessed a great paradigm shift from traditional prior-based models [20, 92, 103] to learning-based approaches [40, 50, 95], for their impressive performance in removing diverse degradations, such as rain streak [14, 39, 63], haze [18, 60, 116], raindrop [54, 71, 93], *etc*. The performance boosts could be attributed to diverse architectural structures [64] and advanced components [21, 25, 27] inspired by high-level vision tasks. For instance, U-shaped network design and skip connection are widely applied to get hierarchical multi-scale representations [9, 29, 98] and learn residual signals [17, 44, 106]. Though CNN-based networks achieve impressive results, they still suffer from the limited receptive field issue of convolution operation. To address this limitation, recent works [10, 53, 68] have explored the attention mechanism for better restoration performance. For instance, SPANet [78] extends an IRNN model to explicitly generate the attention map of rain streaks. RCAN [105] designs a channel attention mechanism to emphasize more informative features. More network architecture designs are summarized in NTIRE challenge reports [49, 80] and recent reviews [31, 41, 104].

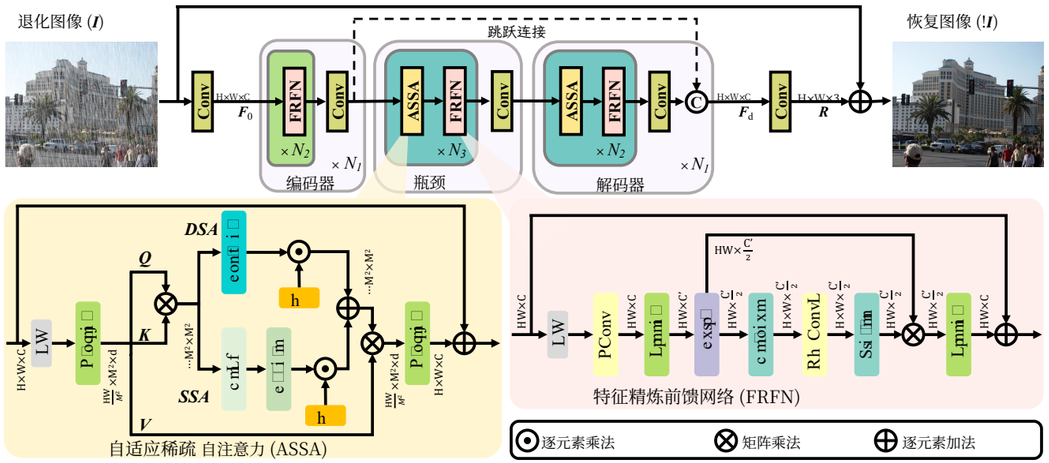


图2。我们自适应稀疏Transformer（AST）的概述。它主要由一个自适应稀疏自注意力（ASSA）和一个特征精炼前馈网络（FRFN）组成。LN指的是层归一化，Conv表示卷积操作。

视觉Transformer。由于Transformer [73] 在自然语言处理领域表现出色，基于Transformer的架构被引入计算机视觉界 [74, 79, 90]。IPT [4] 是图像恢复的先驱性基于Transformer的工作，它通过将输入图像分成小块并顺序处理来解决计算挑战。然而，原始自注意力的二次复杂度仍然阻碍Transformer应用于高分辨率图像。为了缓解这个问题，restormer [100]，开发了通道注意力，它沿通道维度执行注意力计算，从而降低计算成本。另一种潜在的补救措施是窗口注意力 [46]，如Uformer [82]，所采用的方法，它设计了一个局部增强的基于窗口的Transformer，将局部性引入Transformer架构。SwinIR [37]也利用窗口注意力，并引入了一种位移机制以实现更多的跨窗口交互。此外，GRL [35]结合窗口注意力和通道注意力形成一个强大的模型。

尽管这些高效的注意力变体有效地解决了密集计算的问题，并在去除各种退化方面表现良好，但更好的性能仍然受到特征图中无关表示或冗余的严重阻碍 [8, 114]。为此，DRS-former [8] 在注意力机制中设计了一个top-k通道选择算子，用于选择最有信息的token进行计算。类似地，CODE [114] 将特征投影到超像素空间，以减少空间和通道域中的冗余。然而，参数‘k’的具体选择可能对不同的图像恢复任务敏感。此外，在超像素空间中执行注意力机制仍然涉及所有可用的token，这可能在无关区域引入不必要交互。

总体而言，我们AST与现有方法的主要区别有两方面。一方面，我们引入了一种自适应稀疏自注意力机制，通过选择最有信息量的交互来减少冗余。采用用平方ReLU激活替换softmax层的想法来实现稀疏自注意力。我们没有像先前工作 [24, 36, 102]，那样设计复杂的组件来放宽稀疏性，而是探索了一种直接有效且有效的双分支架构来解决信息损失问题。通过这种方式，我们的模型充分利用了SSA的备用分数，而没有因为ReLU-based SSA过于稀疏的性质而难以从有限的信息中学习到令人满意的表示。另一方面，我们在AST中开发了一个关键组件，即<code style id='2'>即</code>，特征精炼前馈网络。为了简化特征图中隐藏的冗余信息，它采用了一种增强和缓解方案，即<code style id='4'>即</code>，增强最有用的特征并沿通道维度缓解不太信息量大的部分。

3. 提出方法

3.1. 整体流程

我们的AST流程概述如图2所示，给定一个图像 $\boldsymbol{I} \in \mathbb{R}^{H\times W\times 3}$，AST首先采用一个卷积层来生成低级特征表示$\boldsymbol{F_0} \in \mathbb{R}^{H\times W\times C}$，其中 $H \times W$、 C 分别为图像分辨率和通道数。接下来，低级表示 $\boldsymbol{F_0}$ 通过一个 N_1阶段对称编码器-解码器网络，并嵌入到深层

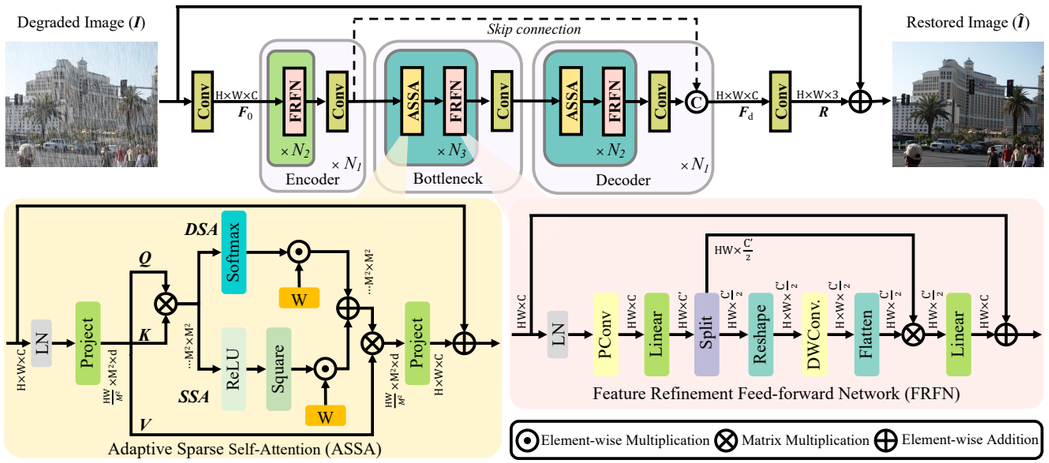


Figure 2. Overview of our Adaptive Sparse Transformer (AST). It mainly consists of an Adaptive Sparse Self-Attention (ASSA), and a Feature Refinement Feed-forward Network (FRFN). LN refers to layer normalization and Conv denotes convolution operation.

Vision Transformer. Since Transformer [73] has shown remarkable performance in the natural language processing field, Transformer-based architecture is introduced into the computer vision community [74, 79, 90]. IPT [4] is the pioneering Transformer-based work for image restoration, which addresses the computational challenge by dividing input images into small patches and processing them sequentially. Nevertheless, the quadratic complexity of vanilla self-attention still hinders Transformers from applying to high-resolution images. To alleviate this problem, channel attention is developed in restormer [100], which performs attention calculation along the channel dimension, reducing computational costs. Another potential remedy is window-based attention [46], such as the approach adopted by Uformer [82], which designs a locally-enhanced window-based Transformer to introduce locality into the Transformer architecture. SwinIR [37] also utilizes window-based attention and introduces a shift mechanism for more cross-window interactions. Furthermore, GRL [35] combines window attention and channel attention to form a powerful model.

Although these efficient attention varieties effectively address the issue of intensive computation and perform well in removing various degradations, better performance is still profoundly hindered by the irrelevant representation or redundancy within feature maps [8, 114]. To this end, DRS-former [8] designs a top-k channel selection operator in the attention mechanism to choose the most informative tokens for calculation. Similarly, CODE [114] projects feature into superpixel space to reduce redundancy in spatial and channel domains. However, the specific choice of the parameter ‘k’ can be sensitive to different image restoration tasks. Moreover, performing the attention mechanism in superpixel space still involves all available tokens, potentially introducing unwanted interactions in irrelevant areas.

Overall, the main differences between our AST and existing approaches are twofold. On the one hand, we introduce an adaptive sparse self-attention mechanism to reduce redundancy by selecting the most informative interactions. The idea of replacing the softmax layer with square ReLU activation is adopted to achieve sparse self-attention. Instead of designing complex components, like prior works [24, 36, 102], to relax sparsity, we explore a straightforward yet effective two-branch architecture to address the information loss issue. In this way, our model fully exploits the spare score of SSA without struggling to learn a satisfactory representation from limited information due to the overly sparse nature of ReLU-based SSA. On the other hand, we develop another critical component in AST, *i.e.*, the feature refinement feed-forward network. To ease the redundant information hidden in the feature map, it adopts an enhance-and-ease scheme, *i.e.*, enhancing the most useful feature and relieving the less informative part along the channel dimension.

3. Proposed Method

3.1. Overall Pipeline

The overview of our AST pipeline is shown in Fig. 2, given a image $\boldsymbol{I} \in \mathbb{R}^{H\times W\times 3}$, AST first employs a convolution layer to produce a low-level feature representation $\boldsymbol{F_0} \in \mathbb{R}^{H\times W\times C}$, where $H \times W$, C are the image resolution and the number of channels, respectively. Next, the low-level representation $\boldsymbol{F_0}$ passes through a N_1-stage symmetric encoder-decoder network and is embedded into deep

特征 $\boldsymbol{F}_d \in \mathbb{R}^{H\times W\times C}$。具体来说，编码器中的每个阶段都由 N_2 基本块和一个用于下采样的卷积层组成。编码器中的基本块包含一个FRFN。编码器部分的特征通过恒等连接与解码器部分的特征融合。在这里，由于标准Transformer块中的注意力机制具有低通滤波特性 [56] 可能会阻碍学习期望的局部模式，尤其是在早期阶段 [89]，因此我们省略了编码器中的注意力机制。在解码器侧，每个阶段都由 N_2 基本块和一个用于上采样的卷积层组成。解码器中的基本块包括一个ASSA和一个FRFN。此外，受 [82], 瓶颈阶段的影响，在解码器之前引入了一个与解码器共享相同Transformer块的瓶颈阶段，以捕获更长的依赖关系。最后，AST采用一个卷积层从 $\boldsymbol{R} \in \mathbb{R}^{H\times W\times 3}$ 生成残差图像 $\boldsymbol{F}_d$。恢复图像是通过退化图像和残差^一个的和获得的，即。 $\boldsymbol{I} = \boldsymbol{I} + \boldsymbol{R}$。采用Charbonnier损失来训练AST： [3]

$$\ell(\boldsymbol{I}', \hat{\boldsymbol{I}}) = \sqrt{\|\boldsymbol{I}' - \hat{\boldsymbol{I}}\|^2 + \epsilon^2}, \tag{1}$$

where $\boldsymbol{I}'$ refers to the 真实图像 and we experimentally set ϵ to 10^{-3}.

3.2. AST Block Design

自适应稀疏自注意力。 As the 标准Transformer [11, 73, 82] consider all tokens inside the 特征图, it may involve many of 无关区域 in the cal- culation. In this way, it not only computes uninformative areas, but also introduces 冗余的 and 无关区域 that degrade the model performance. To cope with this is- sue, we introduce squared ReLU-based self-attention for fil- tering out 特征 with negative impacts of low 查询-键匹配分数, which also ensures the 稀疏性 of the attention mechanism [102] (SSA). Meanwhile, consid- ering the oversparsity of ReLU-based self-attention [66], we introduce another 密集的 self-attention branch (DSA), which employs the softmax层, to aid in retaining crucial information. The key challenge of using this two-branch scheme is how to reduce the noisy 特征 and 冗余信息, while properly retaining the informative one as far as possible. To this end, ASSA fuses two-branch in an 自适应方式, *i.e.*, adaptively takes 特征 from branches and propagates them through the network.

给定一个归一化特征图 $\boldsymbol{X} \in \mathbb{R}^{H\times W\times C}$，我们首先将其划分为大小为 $M \times M$ 的非重叠窗口，得到从第 i 个窗口的扁平化表示 $\boldsymbol{X}^i \in \mathbb{R}^{M^2\times C}$。然后我们从 $\boldsymbol{X}$ 生成 $queries\ \boldsymbol{Q}$、$keys\ \boldsymbol{K}$ 和 $values\ \boldsymbol{V}$：

$$\boldsymbol{Q} = \boldsymbol{X}\boldsymbol{W}_Q, \boldsymbol{K} = \boldsymbol{X}\boldsymbol{W}_K, \boldsymbol{V} = \boldsymbol{X}\boldsymbol{W}_V, \tag{2}$$

其中，查询 $\boldsymbol{W}_Q$、键 $\boldsymbol{W}_K$ 和值 $\boldsymbol{W}_V \in \mathbb{R}^{C\times d}$ 在所有窗口中共享的线性投影矩阵。注意力计算可以定义为：

$$\boldsymbol{A} = f(\boldsymbol{Q}\boldsymbol{K}^T/\sqrt{d} + \boldsymbol{B})\boldsymbol{V}, \tag{3}$$

where $\boldsymbol{A}$ 表示估计的注意力； $\boldsymbol{B}$ 指代可学习的相对位置偏差，而 $f(\cdot)$是一个评分函数。值得注意的是，遵循 [46, 82], 我们并行地对不同的‘头’进行权重计算，这些‘头’被连接起来并通过线性投影进行融合。

然后我们重新审视标准密集自注意力机制（DSA），该机制被大多数现有工作采用。它使用softmax层，考虑所有查询-键对以获得注意力分数：

$$\boldsymbol{DSA} = SoftMax(\boldsymbol{Q}\boldsymbol{K}^T/\sqrt{d} + \boldsymbol{B}). \tag{4}$$

由于并非所有查询词元都与键中的相应词元密切相关，因此对所有相似度的利用对于清晰图像重建是无效的。直观地讲，开发一种稀疏自注意力（SSA）机制来选择词元之间的有用交互可以增强特征聚合。为了在注意力中实现稀疏性，基于平方ReLU的层似乎是一个合理的解决方案。它移除具有负分数的相似度，并将最有用的信息流向前传播：

$$\boldsymbol{SSA} = ReLU^2(\boldsymbol{Q}\boldsymbol{K}^T/\sqrt{d} + \boldsymbol{B}). \tag{5}$$

请注意，ReLU层SSA会导致信息损失，通常需要额外的技术来放宽稀疏性，这与使用SSA而非DSA的动机相悖。

仅应用ReLU层SSA会对管道施加过度稀疏性，即，学习到的特征表示包含不足以供后续处理的信息。相反，使用softmax层DSA会无意中在无关区域引入噪声交互，给恢复高质量图像带来挑战。因此，与其偏爱一种范式胜过另一种，我们提出了一种双分支自注意力机制，作为具有自适应注意力分数的基本组件，以利用这两种范式的优势。公式(3)中的注意力矩阵可以进一步更新为：

$$\boldsymbol{A} = (w_1 * \boldsymbol{SSA} + w_2 * \boldsymbol{DSA})\boldsymbol{V}, \tag{6}$$

其中 $w_1, w_2 \in \mathbb{R}^1$ 是用于自适应调制双分支的两个归一化权重， $*$ 表示乘法运算。更具体地说，它可以由以下方式计算：

$$w_n = \frac{e^{a_n}}{\sum_{i=1}^{N} e^{a_i}}, n = \{1, 2\} \tag{7}$$

$\{a_1, a_2\}$ 是可学习参数，它们被初始化为两个分支中的其中一个。这种设计确保了更好的

feature $\boldsymbol{F}_d \in \mathbb{R}^{H\times W\times C}$. Specifically, each stage within the encoder consists of N_2 basic blocks and a single convolution layer for down-sampling. The basic block in the encoder comprises an FRFN. The features in the encoder part are fused with those in the decoder via the identity connection. Here, we omit the attention mechanism within the standard transformer block in the encoder, due to the fact that its low-pass filter nature [56] can hinder learning desired local patterns, especially in the early stages [89]. On the decoder side, each stage is composed of N_2 basic blocks and a single convolution layer for up-sampling. The basic block in the decoder includes an ASSA and an FRFN. Additionally, inspired by [82], a bottleneck stage is introduced before the decoder that shares the same Transformer block with the decoder to capture longer dependencies. Finally, AST employs a convolution layer to produce the residual image $\boldsymbol{R} \in \mathbb{R}^{H\times W\times 3}$ from $\boldsymbol{F}_d$. The restored image is obtained by the sum of the degraded image and the residual one, *i.e.*, $\hat{\boldsymbol{I}} = \boldsymbol{I} + \boldsymbol{R}$. The Charbonnier loss [3] is adopted to train AST:

$$\ell(\boldsymbol{I}', \hat{\boldsymbol{I}}) = \sqrt{\|\boldsymbol{I}' - \hat{\boldsymbol{I}}\|^2 + \epsilon^2}, \tag{1}$$

where $\boldsymbol{I}'$ refers to the ground-truth image and we experimentally set ϵ to 10^{-3}.

3.2. AST Block Design

Adaptive Sparse Self-Attention. As the vanilla Transformers [11, 73, 82] consider all tokens inside the feature map, it may involve many of irrelevant regions in the calculation. In this way, it not only computes uninformative areas, but also introduces redundant and irrelevant features that degrade the model performance. To cope with this issue, we introduce squared ReLU-based self-attention for filtering out features with negative impacts of low query-key matching scores, which also ensures the sparse property of the attention mechanism [102] (SSA). Meanwhile, considering the oversparsity of ReLU-based self-attention [66], we introduce another dense self-attention branch (DSA), which employs the softmax layer, to aid in retaining crucial information. The key challenge of using this two-branch scheme is how to reduce the noisy features and redundant information, while properly retaining the informative one as far as possible. To this end, ASSA fuses two-branch in an adaptive fashion, *i.e.*, adaptively takes features from branches and propagates them through the network.

Given a normalized feature map $\boldsymbol{X} \in \mathbb{R}^{H\times W\times C}$, we begin by partitioning it into non-overlapping windows of size $M \times M$, resulting in a flattened representation $\boldsymbol{X}^i \in \mathbb{R}^{M^2\times C}$ from the i-th window. Then we generate matrices of $queries\ \boldsymbol{Q}$, $keys\ \boldsymbol{K}$ and $values\ \boldsymbol{V}$ from $\boldsymbol{X}$:

$$\boldsymbol{Q} = \boldsymbol{X}\boldsymbol{W}_Q, \boldsymbol{K} = \boldsymbol{X}\boldsymbol{W}_K, \boldsymbol{V} = \boldsymbol{X}\boldsymbol{W}_V, \tag{2}$$

where the linear projection matrices of the queries $\boldsymbol{W}_Q$, keys $\boldsymbol{W}_K$, and values $\boldsymbol{W}_V \in \mathbb{R}^{C\times d}$ that are shared among all windows. The attention computation can be defined as:

$$\boldsymbol{A} = f(\boldsymbol{Q}\boldsymbol{K}^T/\sqrt{d} + \boldsymbol{B})\boldsymbol{V}, \tag{3}$$

where $\boldsymbol{A}$ denotes the estimated attention; $\boldsymbol{B}$ refers to the learnable relative positional bias, and $f(\cdot)$ is a scoring function. It is worth noting that, following [46, 82], we conduct the weight calculation for different ‘heads’ in parallel, which are concatenated and then fused via linear projection.

We then revisit the standard dense self-attention mechanism (DSA), adopted in most existing works. It employs the softmax layer, considering all query-key pairs to obtain attention scores:

$$\boldsymbol{DSA} = SoftMax(\boldsymbol{Q}\boldsymbol{K}^T/\sqrt{d} + \boldsymbol{B}). \tag{4}$$

Since not all query tokens are closely relevant to corresponding ones in keys, the utilization of all similarities is ineffective for clear image reconstruction. Intuitively, developing a sparse self-attention (SSA) mechanism to pick the useful interactions among the tokens could enhance feature aggregation. For achieving sparsity in attention, a squared ReLU-based layer seems to be a plausible solution. It removes the similarities with negative scores, and propagates the most useful information flow forward:

$$\boldsymbol{SSA} = ReLU^2(\boldsymbol{Q}\boldsymbol{K}^T/\sqrt{d} + \boldsymbol{B}). \tag{5}$$

Note that ReLU-based SSA triggers information loss, additional techniques are often demanded to relax sparsity, which defies the motivation of using SSA over DSA.

Simply applying ReLU-based SSA will impose oversparsity on the pipeline, *i.e.*, the learned feature representation contains insufficient information for the following process. Conversely, using softmax-based DSA will inadvertently introduce noisy interactions in irrelevant regions, posing a challenge in recovering high-quality images. Therefore, rather than preferring one paradigm over the other, we propose a two-branch self-attention mechanism as a fundamental component with adaptive attention scores for taking advantages of both two paradigms. The attention matrix in Eq. (3) can be further updated to:

$$\boldsymbol{A} = (w_1 * \boldsymbol{SSA} + w_2 * \boldsymbol{DSA})\boldsymbol{V}, \tag{6}$$

where $w_1, w_2 \in \mathbb{R}^1$ are two normalized weights for adaptively modulating two-branch, and $*$ denotes the multiply operation. More specifically, it can be computed by:

$$w_n = \frac{e^{a_n}}{\sum_{i=1}^{N} e^{a_i}}, n = \{1, 2\} \tag{7}$$

where $\{a_1, a_2\}$ are learnable parameters that are initialed with 1 of the two branches. This design ensures a better

表1. SPAD 的定量比较 [78]用于雨迹去除。

方法	SPAD [78] PSNR ↑	SSIM ↑
DDN [13]	36.16	0.9463
RESCAN [33]	38.11	0.9797
PReNet [63]	40.16	0.9816
RCDNet [75]	43.36	0.9831
SPDNet [94]	43.55	0.9875
SPAIR [57]	44.10	0.9872
DualGCN [14]	44.18	0.9902
SEIDNet [39]	44.96	0.9911
MPRNet [99]	45.00	0.9897
Fu 等. [15]	45.03	0.9907
Restormer [100]	46.25	0.9911
SCD-Former [19]	46.89	0.9941
IDT [88]	47.34	0.9929
Uformer [82]	47.84	0.9925
DRSformer [8]	48.53	0.9924
AST-B (我们的)	49.51	0.9942
AST-B+ (我们的)	**49.72**	**0.9944**

表2。在AGAN-Data [58] 上对雨滴去除进行的模型效率分析。

方法	AGAN-Data [58] PSNR ↑	SSIM ↑
Eigen的 [12]	21.31	0.757
Pix2pix [26]	27.20	0.836
Uformer [82]	29.42	0.906
WeatherDiff$_{128}$ [54]	29.66	0.923
TransWeather [72]	30.17	0.916
WeatherDiff$_{64}$ [54]	30.71	0.931
TKL&MR [7]	30.99	0.927
All-in-One [32]	31.12	0.927
DuRN [44]	31.24	0.926
CCN [61]	31.34	0.929
Quan的 [62]	31.37	0.918
AttenGAN [58]	31.59	0.917
IDT [88]	31.87	0.931
MAXIM-2S [71]	31.87	0.935
AWRCP [93]	31.93	0.931
AST-B（我们的）	32.32	0.935
AST-B+（我们的）	**32.45**	**0.937**

表3。在密集的雾 [1]上对真实雾去除的定量比较。

方法	密集的雾 [1] PSNR ↑	SSIM ↑
RIDCP[87]	8.09	0.42
DCP [20]	10.06	0.39
SGID [2]	13.09	0.52
D4[91]	13.12	0.53
AOD-Net [30]	13.14	0.41
GridDehazeNet [43]	13.31	0.37
DA-Dehaze [65]	13.98	0.37
FFA-Net [59]	14.39	0.45
Uformer [82]	15.22	0.43
Restormer[100]	15.78	0.55
AECR-Net [86]	15.80	0.47
Fourmer[116]	15.95	0.49
DehazeFormer-S [69]	16.29	0.51
DeHamer [18]	16.62	0.56
MB-TaylorFormer-B [60]	16.66	0.56
AST-B (Ours)	17.12	0.55
AST-B+ (Ours)	**17.27**	**0.57**

where {v1}是可学习参数，它们被初始化为两个分支中的其中一个。这种设计确保了在可以过滤掉无关区域噪声交互和能够利用足够的信息特征之间取得更好的权衡。换句话说，该模型能够控制输入标记关于特定任务的稀疏程度。**特征精炼前馈网络**。常规FFN [73] 独立处理每个像素位置的信息，这作为通过自注意力机制提高特征表示的关键作用。因此，设计一个有效的FFN以增强能够提升潜在高质量图像恢复的特征至关重要。当ASSA被用作去除空间域冗余信息的基本组件时，通道中仍然存在冗余。为了克服这一点，我们开发了FRFN以在增强与简化范式下执行特征转换。具体来说，我们通过引入PConv操作 [5] 来强化特征中的信息元素，并引入门控机制来减少冗余信息的处理负担。FRFN可以表示为：

$$\begin{aligned}
&\hat{\boldsymbol{X}}' = GELU(W_1\text{PConv}(\hat{\boldsymbol{X}})), [\hat{\boldsymbol{X}}'_1, \hat{\boldsymbol{X}}'_2] = \hat{\boldsymbol{X}}', \\
&\hat{\boldsymbol{X}}'_r = \hat{\boldsymbol{X}}'_1 \otimes F(\text{DWConv}(R(\hat{\boldsymbol{X}}'_2))), \\
&\hat{\boldsymbol{X}}'_{out} = GELU(W_2\hat{\boldsymbol{X}}'_r),
\end{aligned} \tag{8}$$

其中 W_1 和 W_2 表示线性投影； $[,]$ 指通道切片操作； $R(\cdot)$ 和 $F(\cdot)$ 分别展示了重整形和展平操作，将序列输入转换为二维特征图，以及反向操作，这对于将局部性引入架构 [34] 至关重要；PConv(·) 和 DWConv(·) 分别指部分卷积 [5] 和深度卷积 [22] 操作； $\otimes$ 表示矩阵乘法。

总体而言，FRFN能够通过从信息流中提取那些代表性特征来增强特征表示，同时简化冗余的特征。它还提供了模型沿着通道维度清除无信息特征的机会。

信息流，同时简化冗余的特征。它还提供了模型沿着通道维度清除无信息特征的机会。

4. 实验

在本节中，我们评估了 AST 在各种图像恢复任务上的性能，例如雨迹、雾和雨滴去除。还进行了消融实验，以证明所提出模块的有效性。

4.1. 实验设置

实现细节。在默认设置中，AST 包含 N_1=4 个阶段用于编码器和解码器部分，并在瓶颈中开发一个阶段。我们通过改变嵌入维度 C 和 Transformer 块（编码器和解码器共享相同的 N_2 块，而瓶颈包含 N_3 块）构建了我们基础模型的两个变体，称为 AST-T 和 AST-B。对于 AST-T，我们将 C 设置为 16，N_2 和 N_3 设置为 [2,2,2,2] 和 2，而对于 AST-B，我们将 C 设置为 32，N_2 和 N_3 设置为 [1,2,8,8] 和 2。默认的分割窗口大小为 8，它们在 Transformer 块中的每个头共享相同的维度，遵循 [82] 中的方法。我们采用 AdamW 优化器 [47] 并使用默认设置来训练我们的模型。学习率最初设置为 0.0002，并使用余弦衰减策略 [48] 逐渐降低到 0.000001。我们随机使用旋转和翻转操作策略进行增强。我们使用渐进式学习策略来节省时间，类似于 [70, 100]。

评估指标。为评估恢复性能，我们采用PSNR和SSIM指标 [81]。此外，NIQE [52] 被用作非参考指标。值得注意的是，对于去雨，遵循现有工作 [75, 82]，PSNR/SSIM分数在Y通道的

Table 1. Quantitative comparison on SPAD [78] for rain streak removal.

Method	SPAD [78] PSNR ↑	SSIM ↑
DDN [13]	36.16	0.9463
RESCAN [33]	38.11	0.9797
PReNet [63]	40.16	0.9816
RCDNet [75]	43.36	0.9831
SPDNet [94]	43.55	0.9875
SPAIR [57]	44.10	0.9872
DualGCN [14]	44.18	0.9902
SEIDNet [39]	44.96	0.9911
MPRNet [99]	45.00	0.9897
Fu *et al*. [15]	45.03	0.9907
Restormer [100]	46.25	0.9911
SCD-Former [19]	46.89	0.9941
IDT [88]	47.34	0.9929
Uformer [82]	47.84	0.9925
DRSformer [8]	48.53	0.9924
AST-B (Ours)	49.51	0.9942
AST-B+ (Ours)	**49.72**	**0.9944**

Table 2. Model efficiency analysis on AGAN-Data [58] for raindrop removal.

Method	AGAN-Data [58] PSNR ↑	SSIM ↑
Eigen's [12]	21.31	0.757
Pix2pix [26]	27.20	0.836
Uformer [82]	29.42	0.906
WeatherDiff$_{128}$ [54]	29.66	0.923
TransWeather [72]	30.17	0.916
WeatherDiff$_{64}$ [54]	30.71	0.931
TKL&MR [7]	30.99	0.927
All-in-One [32]	31.12	0.927
DuRN [44]	31.24	0.926
CCN [61]	31.34	0.929
Quan's [62]	31.37	0.918
AttenGAN [58]	31.59	0.917
IDT [88]	31.87	0.931
MAXIM-2S [71]	31.87	0.935
AWRCP [93]	31.93	0.931
AST-B (Ours)	32.32	0.935
AST-B+ (Ours)	**32.45**	**0.937**

Table 3. Quantitative comparison on Dense-Haze [1] for real haze removal.

Method	Dense-Haze [1] PSNR ↑	SSIM ↑
RIDCP[87]	8.09	0.42
DCP [20]	10.06	0.39
SGID [2]	13.09	0.52
D4[91]	13.12	0.53
AOD-Net [30]	13.14	0.41
GridDehazeNet [43]	13.31	0.37
DA-Dehaze [65]	13.98	0.37
FFA-Net [59]	14.39	0.45
Uformer [82]	15.22	0.43
Restormer[100]	15.78	0.55
AECR-Net [86]	15.80	0.47
Fourmer[116]	15.95	0.49
DehazeFormer-S [69]	16.29	0.51
DeHamer [18]	16.62	0.56
MB-TaylorFormer-B [60]	16.66	0.56
AST-B (Ours)	17.12	0.55
AST-B+ (Ours)	**17.27**	**0.57**

trade-off between noisy interactions of irrelevant areas that can be filtered out, and enough informative features can be leveraged. In other words, this model is enabled to control the sparse degree of input tokens regarding the specific task. **Feature Refinement Feed-forward Network.** The regular FFN [73] processes the information at each pixel location individually, which serves as a crucial role in improving the feature representation by the self-attention mechanism. Therefore, designing an effective FFN for enhancing features that boost the latent high-quality image restoration is vital. When ASSA is adopted as a fundamental component to remove redundant information in the spatial domain, there remains redundancy in channels. To overcome this, we develop the FRFN to perform the feature transformation in an enhance-and-ease paradigm. Specifically, we construct FRFN by introducing a PConv operation [5] to reinforce the informative elements within features, and a gate mechanism to reduce the processing burden of the redundant information. The FRFN can be represented as:

$$\begin{aligned}
&\hat{\boldsymbol{X}}' = GELU(W_1\text{PConv}(\hat{\boldsymbol{X}})), [\hat{\boldsymbol{X}}'_1, \hat{\boldsymbol{X}}'_2] = \hat{\boldsymbol{X}}', \\
&\hat{\boldsymbol{X}}'_r = \hat{\boldsymbol{X}}'_1 \otimes F(\text{DWConv}(R(\hat{\boldsymbol{X}}'_2))), \\
&\hat{\boldsymbol{X}}'_{out} = GELU(W_2\hat{\boldsymbol{X}}'_r),
\end{aligned} \tag{8}$$

where W_1 and W_2 denote the linear projections; $[,]$ refers to channel-wise slice operation; $R(\cdot)$ and $F(\cdot)$ illustrate Reshape and Flatten operations that convert the sequence input to a 2D feature map and in reverse, which is crucial to introduce locality into the architecture [34]; PConv(·) and DWConv(·) refer to the partial convolution [5] and depth-wise convolution [22] operation, respectively; $\otimes$ represents matrix multiplication.

Overall, FRFN is capable of enhancing feature representations by extracting those representative features from the information flow while simplifying the redundant ones. It also provides the chance for the model to clear uninformative features along the channel dimension.

4. Experiments

In this section, we evaluate the performance of AST on various image restoration tasks, such as rain streak, haze, and raindrop removal. Ablation studies are also performed to demonstrate the effectiveness of the proposed modules.

4.1. Experiment Settings

Implementation Details. In the default setting, AST contains N_1=4 stages for both the encoder and decoder part, and develops one stage in the bottleneck. We build two variants of our vanilla model, called AST-T and AST-B, by varying the embedding dimensions C and Transformer blocks (the encoder and decoder share the same N_2 blocks, while the bottleneck includes N_3 blocks). For AST-T, we set C as 16, N_2 and N_3 as [2,2,2,2] and 2, while for AST-B, we set C as 32, N_2 and N_3 as [1,2,8,8] and 2. The default split window size is 8, and they share same dimension of each head in the Transformer block, following the approach in [82]. We adopt the AdamW optimizer [47] with the default settings to train our model. The learning rate is initially set as 0.0002 and gradually decreases to 0.000001 using the cosine decay strategy [48]. We randomly use the rotation and flipping operation strategies for augmentation. The progressive learning strategy is used to save time, similar to [70, 100].

Evaluation Metrics. To evaluate the restoration performance, we adopt PSNR and SSIM metrics [81]. Additionally, NIQE [52] is used as a non-reference metric. Notably, for deraining, following existing works [75, 82], PSNR/SSIM scores are calculated on the Y channel in the

图3. 在真实雨移除任务上对SPAD [78] 的定性比较。

YCbCr空间。当使用几何自集成策略 [38] 时，我们用’+’符号表示该方法。表格中的最佳和次佳分数被**突出显示**并加下划线。

4.2. 雨迹去除

我们在SPAD基准 [78] 上执行去雨实验，并将AST的性能与十五种最先进的算法进行比较，包括DDN [13], RES-CAN [33], PReNet [63], RCDNet [75], SPDNet [94],SPAIR [57], DualGCN [14], SEIDNet [39], MPRNet [99],Fu等人 [15], Restormer [100], SCD-Former [19],IDT [88], Uformer [82] 和 DRSformer [8]。在表1中，AST-B在PSNR指标上比之前最好的CNN方法Fu等人 [15]提高了4.48dB，比之前最好的基于Transformer的模型DRSformer [8]提高了0.98 dB。我们在图3中展示了视觉比较结果，其中AST-B可以更成功地去除真实雨迹，同时保留结构内容。

4.3. 雨滴去除

我们在AGAN-Data [58] 基准数据集上进行了雨滴去除实验，并将我们的AST与多种最先进的去雨滴方法进行了比较，包括Eigen的[12], Pix2pix[26],Uformer [82], WeatherDiff_{128} [54], TransWeather [72], WeatherDiff_{64} [54], TKL&MR [7], All-in-One [32], DuRN [44], CCN [61], Quan的 [62], AttenGAN [58],IDT [88], MAXIM-2S [71]和AWRCP [93]。在表2中，AST-B比之前的最佳方法AWRCP [93]提高了0.39 dB，在PSNR方面超过了同时的基于扩散的方法 WeatherDiff_{128} [54] 2.66 dB。

4.4. 真实雾霾去除

我们在Dense-Haze基准上进行了评估 [1]用于真实雾霾去除，并将AST与十五个最先进的去雾工作进行了比较，包括RIDCP[87],DCP [20], SGID [2], D4[91], AOD-Net [30], Grid-DehazeNet [43], DA-Dehaze[65], FFA-Net [59],Uformer [82], Restormer[100], AECR-Net[86],Fourmer[116], DehazeFormer-S [69], DeHamer [18]和 MB-TaylorForm [60]。在表3中，AST-B在PSNR指标中获得了所考虑的最佳值

表4. 不同自注意力机制的消融研究。

模型	Swin SA [37]	Top-k SA [8]	压缩SA [114]	ASSA Ours
参数	6.65	6.67	6.07	6.65
FLOPs	13.32	13.59	11.46	13.35
PSNR	44.47	44.67	44.94	45.43

表 5. 与标准自注意力机制及其对应稀疏版本的比较。

	方法	PSNR
(1)	标准局部自注意力 [82]	45.09
	稀疏局部自注意力	44.58
(2)	标准通道自注意力 [100]	44.91
	稀疏通道自注意力	44.45

最先进的方法。与之前的最佳CNN方法ARCT-Net [86]，相比，我们的AST-B的PSNR增益为1.37 dB。此外，与最近的基于Transformer的方法 [60, 69, 116] 相比，我们的AST-B实现了至少0.46 dB的改进。

4.5. 分析与讨论

探索Transformer架构中最有用的信息并减少其冗余，在多种图像恢复任务上取得了有利的结果。在这里，我们对AST进行了更深入的分析，并说明了所提出模块的有效性。对于消融实验，我们在SPAD [78] 数据集上训练了去雨模型AST-T。为了进行公平的比较，所有模型都在 128×128图像块上训练了10个epoch，并使用输入大小为 256×256计算FLOPs。

ASSA的有效性。 为了研究ASSA组件的有效性，我们将其替换为现有的有效注意力机制：(1) Swin自注意力（Swin SA） [37], (2) Top-k自注意力（Top-k SA) [8], 和 (3) 压缩自注意力（Condensed SA) [114]。我们在表。4中展示了定量结果。与Swin SA相比，ASSA提供了0.96 dB的PSNR有利增益，并且复杂度略有增加（0.03G FLOPs）。此外，与提出以清除token之间噪声交互和冗余信息的相关方法相比，我们的ASSA设计在Top-k SA [8], 上获得了0.76 dB的性能提升，在Condensed SA [114]上获得了0.49 dB的性能提升。

自适应架构设计的有效性。 所提出的自适应架构设计用于减少

Figure 3. Qualitative comparisons on SPAD [78] for real rain removal.

YCbCr space. We denote the method with the ’+’ symbol when geometric self-ensemble strategy [38] is used. The best and second-best scores in the tables are **highlighted** and underlined.

4.2. Rain Streak Removal

We perform the deraining experiments on SPAD benchmark [78] and compare the performance of AST with fifteen state-of-the-art algorithms, including DDN [13], RES-CAN [33], PReNet [63], RCDNet [75], SPDNet [94], SPAIR [57], DualGCN [14], SEIDNet [39], MPRNet [99], Fu *et al*. [15], Restormer [100], SCD-Former [19], IDT [88], Uformer [82] and DRSformer [8]. In Tab. 1, AST-B achieves a gain of 4.48 dB in terms of PSNR metric against the previous best CNN-based method Fu *et al*. [15] and 0.98 dB against the previous best Transformer-based model DRSformer [8]. We present visual comparisons in Fig. 3, where AST-B can remove the real rain streak more successfully while preserving the structural content.

4.3. RainDrop Removal

We conduct raindrop removal experiments on AGAN-Data [58] benchmark, and compare our AST with a wide range of state-of-the-art deraindrop approaches, including Eigen’s [12], Pix2pix [26], Uformer [82], WeatherDiff_{128} [54], TransWeather [72], WeatherDiff_{64} [54], TKL&MR [7], All-in-One [32], DuRN [44], CCN [61], Quan’s [62], AttenGAN [58], IDT [88], MAXIM-2S [71] and AWRCP [93]. In Tab. 2, AST-B outperforms the previous best method AWRCP [93] by a substantial 0.39 dB and surpasses the concurrent diffusion-based method WeatherDiff_{128} [54] by 2.66 dB in terms of PSNR.

4.4. Real Haze Removal

We conduct evaluation on Dense-Haze benchmark [1] for real haze removal, and compare AST with fifteen state-of-the-art dehazing works, including RIDCP[87], DCP [20], SGID [2], D4[91], AOD-Net [30], Grid-DehazeNet [43], DA-Dehaze [65], FFA-Net [59], Uformer [82], Restormer [100], AECR-Net [86], Fourmer[116], DehazeFormer-S [69], DeHamer [18] and MB-TaylorForm [60]. In Tab. 3, AST-B obtains the best values in PSNR metric among the considered

Table 4. Ablation study for different self-attention mechanisms.

Models	Swin SA [37]	Top-k SA [8]	Condensed SA [114]	ASSA Ours
Params	6.65	6.67	6.07	6.65
FLOPs	13.32	13.59	11.46	13.35
PSNR	44.47	44.67	44.94	45.43

Table 5. Comparison with standard self-attention mechanisms and corresponding sparse version.

	Method	PSNR
(1)	Standard Local Self-Attention [82]	45.09
	Sparse Local Self-Attention	44.58
(2)	Standard Channel Self-Attention [100]	44.91
	Sparse Channel Self-Attention	44.45

state-of-the-art methods. Compared to the previous best CNN-based method ARCT-Net [86], the PSNR gain of our AST-B is 1.37 dB. In addition, our AST-B achieves at least 0.46 dB improvement when compared to recent Transformer-based methods [60, 69, 116].

4.5. Analysis and Discussion

Exploring the most useful information and reducing the redundancy within Transformer architecture provides favorable results on diverse image restoration tasks. Here, we present a deeper analysis of AST and illustrate the effectiveness of the proposed modules. For ablation studies, we train the deraining models AST-T on the SPAD [78] dataset. For a fair comparison, all models are trained on 128×128 image patches for 10 epochs, and we calculate FLOPs with the input size of 256×256.

Effectiveness of ASSA. To investigate the effectiveness of the ASSA component, we replace it with existing effective attention mechanisms: (1) Swin Self-Attention (Swin SA) [37], (2) Top-k Self-Attention (Top-k SA) [8], and (3) Condensed Self-Attention (Condensed SA) [114]. We show the quantitative results in Tab. 4. ASSA provides favorable gains of 0.96 dB in PSNR, with slightly increased complexity (0.03G Flops) when compared to the Swin SA. In addition, compared to the closely related methods that proposed to clear noisy interaction among tokens and redundancy information, our ASSA design obtains a performance improvement of 0.76 dB over Top-k SA [8], and 0.49 dB over Condensed SA [114].

Effectiveness of adaptive architecture design. The proposed adaptive architecture design is used to reduce

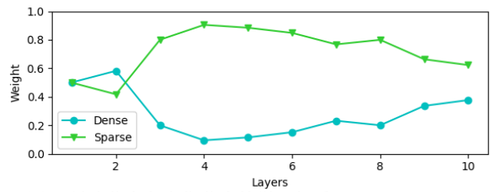


图4. 稀疏分支和密集分支的学习权重。

噪声代表性特征和冗余信息，同时正确保留信息量大的特征。为了研究配备基于ReLU的稀疏注意力的图像恢复模型在NLP领域是否会遇到类似性能下降现象，我们首先基于两种主流范式构建了两种版本的稀疏自注意力机制：(1) 局部自注意力 [82]和 (2) 通道自注意力 [100]。如表 5所示，直接用基于ReLU的稀疏注意力机制替换标准的基于softmax的密集自注意力机制，导致局部自注意力和通道自注意力分别性能下降0.51 dB和0.46 dB。

为了进一步调查性能下降是否由ReLU-based稀疏自注意力(SSA)的过于稀疏问题导致的信息损失所触发，我们计算了注意力层的熵，类似于 [16], 用于测量注意力集中度。具体来说，注意力熵定义为：

$$Enctropy_{Att} = -\frac{1}{H}\sum_{h}\frac{1}{L}\sum_{ij} Att_{ij}^{h,l} * log^{Att_{ij}^{h,l}}, \quad (9)$$

其中 $Att^{h}{}_{,ij}{}^{l}$ 表示查询词元到键词元 i 的注意力分数 j，位于层 $h \in H$ 的 $l \in L$注意力头 6。较低的熵意味着平均注意力更集中，而较高的熵则表示注意力更分散。如表 6所示，基于softmax的密集自注意力(DSA)得分最高，而SSA得分最低。换句话说，DSA从源词元中提取特征更均匀，这可能导致无关区域的噪声交互。SSA集中在少数过于稀疏以至于无法覆盖必要关系的词元上。相反，我们的方法在充分探索信息性上下文的同时忽略了冗余特征，从而显著提升了性能。

我们随后展示了使用所提出的自适应双分支架构设计的必要性和优越性，即标准密集自注意力和相应的稀疏版本，通过在表7中对训练模型变体进行实验来缓解挑战。7。与配备DSA的模型相比，直接应用SSA的性能不尽如人意。例如，ACON和Meta-

表6. 不同自注意力机制的熵分析。

结构	DSA	SSA	ASSA (我们的)
熵	3.733	1.543	3.134
PSNR	45.09	44.58	45.43

表7. 自注意力机制中各种激活选择的消融研究。

Type	密集的	稀疏的		自适应		
多样性	softmax	relu^2	StarReLu [96]	ACON [51]	Meta-ACON [51]	ASSA(我们的)
PSNR	45.09	44.58	45.30	43.23	43.67	45.43
Δ	-0.34	-0.85	-0.13	-2.20	-1.76	-

表8. 特征细化前馈网络替代方案的消融研究。

模型	FFN [11]	DFN [34]	GDFN [100]	LeFF [82]	FRFN Ours
参数	7.77	7.92	6.49	7.92	6.65
FLOPs	15.25	16.20	13.19	16.30	13.35
PSNR	44.13	43.46	44.66	44.77	45.43

ACON [51], 我们的ASSA仍然可以实现最大的性能提升（45.43 dB）。

我们最终在图 4中可视化了每个SSA和DSA分支的学习权重。如预期所示，模型最初平等对待两个分支以确保足够的信息，随着层数加深，模型更加关注稀疏分支以更好地聚合特征。我们注意到，学习权重充当软选择，从而允许模型适应两个分支的影响。

FRFN的有效性。特征图通常具有高通道维度，尤其是在深层，并非所有特征通道都包含恢复清晰图像的关键信息。简单地将相同特征变换应用于所有通道可能导致冗余信息过多。在实践中，增强信息通道以进一步推进特征表示学习是一项艰巨的任务。为了展示我们FRFN的效果，我们首先将其与四个变体进行比较，包括（1）基础前馈网络 (FFN) [11], (2) 深度卷积配备前馈网络 (DFN) [34], (3) 门控-深度卷积前馈网络 (GDFN) [100], 和（4）局部增强前馈网络 (LeFF)。定量比较列于表 8中。我们的FRFN实现了最佳的PSNR值，参数和FLOPs略多。换句话说，FRFN可以筛选更有用的信息并简化冗余特征，从而比其他考虑的方案更好地配合我们提出的ASSA设计。尽管GDFN [100] 利用了我们类似的门控机制来控制信息流，FRFN执行了精细的增强与简化特征变换以帮助选择最有信息量的特征。结果，FRFN在GDFN上实现了0.77 dB的PSNR提升。

我们还进行了消融实验，见表9以研究FRFN的影响。与基线模型（a）相比

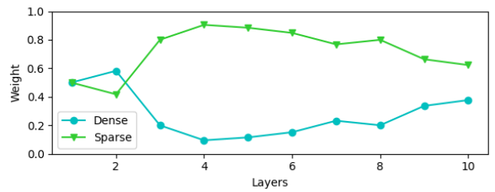


Figure 4. Learned weights for sparse and dense branches.

the noisy representative features and redundant information while properly retaining the informative one. To investigate whether models for image restoration equipped with ReLU-based sparse attention will encounter similar performance degradation phenomena in the NLP field, we first construct two versions of sparse self-attention mechanisms based on two mainstream paradigms: (1) Local Self-Attention [82] and (2) Channel Self-Attention [100]. As shown in Tab. 5, directly replacing the standard softmax-based dense self-attention with the ReLU-based sparse one leads to significant performance drops of 0.51 dB and 0.46 dB for Local Self-Attention and Channel Self-Attention, respectively.

To further investigate whether the performance drop is triggered by information loss due to the overly sparse issue of ReLU-based sparse self-attention (SSA), we calculate the entropy of the attention layer, similar to [16], to measure attention concentration. Specifically, the attention entropy is defined as:

$$Enctropy_{Att} = -\frac{1}{H}\sum_{h}\frac{1}{L}\sum_{ij} Att_{ij}^{h,l} * log^{Att_{ij}^{h,l}}, \quad (9)$$

where $Att_{ij}^{h,l}$ represents the attention score for the query token i to the key token j of head $h \in H$ at layer $l \in L$. Lower entropy means that on average the attention tends to be concentrated, while higher one indicates the attention is more distributed. As displayed in Tab. 6, softmax-based dense self-attention (DSA) achieves the highest score while SSA obtains the lowest one. In other words, DSA extracts features from source tokens more uniformly, which may introduce noisy interaction of irrelevant regions. SSA concentrates on a few tokens that are too sparse to cover necessary relations. On the contrary, our method arrived at a compromise that the informative context can be fully explored while the redundant features will be neglected, resulting in a clear performance boost.

We then show the necessity and superiority of using the proposed adaptive two-branch architecture design, *i.e.*, standard dense self-attention and the corresponding sparse version, to alleviate the challenge by conducting experiments on training model variants in Tab. 7. Directly applying SSA suffers unsatisfactory performance, compared to the model equipped with DSA. Particularly, when comparing to the adaptive activation, *e.g.*, ACON and Meta-

Table 6. Entropy analysis of different self-attention mechanisms.

Structure	DSA	SSA	ASSA (Ours)
Entropy	3.733	1.543	3.134
PSNR	45.09	44.58	45.43

Table 7. Ablation study on various activation choices in the self-attention mechanism.

Type	Dense	Sparse		Adaptive		
Variety	softmax	ReLU^2	StarReLu [96]	ACON [51]	Meta-ACON [51]	ASSA(Ours)
PSNR	45.09	44.58	45.30	43.23	43.67	45.43
Δ	-0.34	-0.85	-0.13	-2.20	-1.76	-

Table 8. Ablation study on alternatives to feature refinement feed-forward network.

Models	FFN [11]	DFN [34]	GDFN [100]	LeFF [82]	FRFN Ours
Params	7.77	7.92	6.49	7.92	6.65
FLOPs	15.25	16.20	13.19	16.30	13.35
PSNR	44.13	43.46	44.66	44.77	45.43

ACON [51], our ASSA can still achieve the largest performance gain (45.43 dB).

We finally visualize the learned weights for each SSA and DSA branch in Fig. 4. As expected, the model treats two branches equally at first to ensure sufficient information, and pays more attention to the sparse branch as layers go deeper for better aggregating features. We note that the learned weights act as a soft selection, thus allowing the model to adapt to the influence of the two branches.

Effectiveness of FRFN. Feature maps often have high channel dimensions, especially in deep layers, and not all feature channels contain the key information for recovering clear images. Simply applying the same feature transformations to all channels can result in an excess of redundant information. In practice, it is daunting to enhance the informative channels for further advances in feature representation learning. To demonstrate the effect of our FRFN, we first compare it with four variants, including (1) vanilla Feed-Forward Network (FFN) [11], (2) Depth-wise convolution equipped Feed-forward Network (DFN) [34], (3) Gated-Dconv Feed-forward Network (GDFN) [100], and (4) Locally-enhanced Feed-Forward network (LeFF). The quantitative comparisons are listed in Tab. 8. Our FRFN achieves the best PSNR value, with slightly more parameters and FLOPs. In other words, FRFN could select the more useful information and ease the redundant features, thus better cooperating with our proposed ASSA design than other considered ones. Although GDFN [100] leverages a gating mechanism like ours to control the information flows, FRFN performs a delicate enhance-and-ease feature transformation to help select the most informative features. As a result, FRFN achieves a PSNR gain of 0.77 dB over GDFN.

We also perform ablation studies in Tab. 9 to investigate the impact of FRFN. Compared to the baseline model (a)

(a) 输入 (b) DRSformer [14] (c) IDT [88] (d) Restormer [100] (e) Uformer [82] (f) AST-B

图5。在Internet-Data [78]上对真实雨移除的定性比较。

表9。FRFN在SPAD上对图像去雨的消融研究。

	增强	Ease	DWConv	参数	FLOPs	PSNR	延迟
(a)			√	8.06	17.16	45.02	15.55
(b)	√		√	8.26	17.55	45.16	17.03
(c)		√	√	6.45	12.96	45.30	19.01
(d)	√	√	√	6.65	13.35	45.43	19.75

表 10. 真实场景下去雨任务的无参考评估指标 NIQE 的结果。

方法 输入	Uformer [82]	Restormer [100]	IDT [88]	DRSformer [8]	AST-B Ours
NIQE ↓ 6.274	5.749	6.162	6.079	5.994	5.493

该模型通过深度卷积层引入局部性，遵循现有工作[82], 我们的FRFN (d) 通过设计增强和缓解方案提供了性能优势（0.41 dB）。具体而言，使用PConv算子[5]增强有价值的信息，并使用门控机制缓解特征图中隐藏的冗余，分别带来了0.28 dB和0.14 dB的提升。PConv仅卷积部分通道，可以看作是一种稀疏操作来选择有用通道。这样，它引导网络专注于重要特征，增强了提取信息特征的能力。这些结果证明了我们的FRFN设计贡献，即增强和缓解方案。

感知质量评估。 随后 [8], 我们从Internet-Data基准测试中随机选择了20张真实场景的雨景图像 [78]进行评估。如表10所示，AST-B实现了最低的NIQE值，这意味着在真实场景下，相对于其他方法，它具有更好的感知质量。此外，如图5所示的定性比较表明，AST-B消除了雨痕退化并生成了视觉上逼真的结果，这表明它能够处理未见过的真实退化。

5. 结论

这项工作的目标是通过自适应地学习最有信息量的表示并缓解特征中的噪声信息来从退化版本中恢复清晰图像。虽然我们从自然语言处理中引入了基于ReLU的稀疏自注意力（SSA）来消除无关标记之间的噪声交互，但并没有直接采用它

(a) 输入 (b)AST-B

图6. 错误恢复的示例。在具有严重退化的真实场景中，可以找到AST的典型失败案例。

这项工作的目标是通过自适应地学习最有信息量的表示并简化特征中的噪声信息，从退化版本中恢复清晰图像。虽然我们从自然语言处理中引入了基于ReLU的稀疏自注意力（SSA）来消除无关标记之间的噪声交互，但我们的目标不是直接将其作为基本组件使用，而是首先防止由于ReLU-基于SSA的熵较小而导致的信息损失。为了有效实现这一目标，我们探索了一种自适应架构设计，该设计确保必要的信息流在另一个密集分支的帮助下向前传递。此外，我们提出了一种FRFN来执行特征转换，采用增强和缓解方案，其中可以学习到判别性特征表示，以提升高质量图像重建。我们的AST优于采用选择操作（例如，Top-K选择和稀疏通道SA）或将特征投影到超像素空间（例如，压缩SA）以简化冗余的相关基线，最终在多个退化去除任务上取得了有利的结果。

局限性。 未来的工作可以关注当前的局限性（例如，为具有各种退化质量的低质量图像开发一个统一的模型），以及这项任务特定模型提供的机遇（例如，注入先验知识，如用于图像去雾的暗通道先验和用于去除低光照条件的视网膜模型先验）。一个失败案例如图6所示，其中AST难以处理具有严重退化的场景。

致谢。 本研究得到天津市自然科学基金（NO. 20JCJQJC00020）、国家自然科学基金（Nos. U22B2049, 62302240）、中央高校基本科研业务费专项资金以及南开大学超算中心（NKSC）的支持。

(a) Input (b) DRSformer [14] (c) IDT [88] (d) Restormer [100] (e) Uformer [82] (f) AST-B

Figure 5. Qualitative comparisons on Internet-Data [78] for real rain removal.

Table 9. Ablation study of FRFN on SPAD for image deraining.

	Enhance	Ease	DWConv	Params	FLOPs	PSNR	**Lantency**
(a)			√	8.06	17.16	45.02	15.55
(b)	√		√	8.26	17.55	45.16	17.03
(c)		√	√	6.45	12.96	45.30	19.01
(d)	√	√	√	6.65	13.35	45.43	19.75

Table 10. Results of no-reference assessment metric NIQE for deraining task under the real-world scenario.

Methods	Input	Uformer [82]	Restormer [100]	IDT [88]	DRSformer [8]	AST-B Ours
NIQE ↓	6.274	5.749	6.162	6.079	5.994	5.493

that introduces locality with a depth-wise convolution layer, following existing works [82], our FRFN (d) provides performance benefits (0.41 dB) by designing an enhance-and-ease scheme. Specifically, enhancing the valuable information using the PConv operator [5] and easing the redundancy hidden in the feature map with a gating mechanism yield 0.28 dB and 0.14 dB improvements, respectively. PConv convolves only part of the channels, which can be viewed as a sparse operation to select useful channels. In this way, it guides the network to concentrate on important features and enhances the ability to extract informative features. These results prove our design contributions of FRFN with the enhance-and-ease scheme.

Perceptual quality assessment. Following [8], we randomly chose 20 real-world rainy images from the Internet-Data [78] benchmark to conduct the assessment. As displayed in Tab. 10, AST-B achieves the lowest NIQE value, implying better perceptual quality over considered methods under real-world settings. Moreover, as the qualitative comparison shown in Fig. 5, AST-B clears rain-streak degradations and generates a visually faithful result, which indicates its capability to handle unseen real degradation.

5. Conclusions

The goal of this work was to recover clear images from the degraded version by adaptively learning the most informative representations and easing the noisy information within features. While we introduce the ReLU-based sparse self-attention (SSA) from NLP for removing noisy interactions among irrelevant tokens, instead of directly employing it

(a) Input (b) AST-B

Figure 6. Examples of erroneous restorations. Typical failure of AST can be found in real-world scenarios with heavy degradation.

as a fundamental component, our target is to first prevent the information loss due to the small entropy of the ReLU-based SSA. For this to be achieved effectively, we explore an adaptive architecture design, which ensures necessary information flows forward with the aid of another dense branch. Moreover, we propose an FRFN to perform the feature transformation with an enhance-and-ease scheme, where discriminative feature representation can be learned to boost high-quality image reconstruction. Our AST outperforms the relevant baselines that adopt a selection operation (*e.g.*, Top-K selection and Sparse Channel SA) or project features into superpixel space (*e.g.*, Condensed SA) for easing redundancy, while ultimately, it achieves favorable results on several degradation removal tasks.

Limitations. Future work could focus on current limitations (*e.g.*, developing a uniform model for low-quality images with various degradations), as well as opportunities that this task-specific model provides (*e.g.*, injecting priors, like dark channels prior for image dehazing and retinex model prior for removing low-light conditions). A failure case is illustrated in Fig. 6, where AST struggles to deal with scenes with heavy degradations.

Acknowledgements. This work was supported by Natural Science Foundation of Tianjin, China (NO. 20JCJQJC00020), the National Natural Science Foundation of China (Nos. U22B2049, 62302240), Fundamental Research Funds for the Central Universities, and Supercomputing Center of Nankai University (NKSC).

参考文献

[1] Codruta O Ancuti, Cosmin Ancuti, Mateu Sbert, 和 Radu Timofte. Dense-haze: 基于密集雾图和无雾图像的图像去雾基准。在ICIP，2019。5, 6[2] Haoran Bai, 潘金山, 翟兴广, 和 唐金辉. 基于渐进式特征融合的自引导图像去雾。TIP, 31:1217–1229, 2022。5, 6[3] Pierre Charbonnier, Laure Blanc-Feraud, Gilles Aubert, 和 Michel Barlaud. 计算成像的两种确定性半二次正则化算法。在ICIP，1994。4[4] Hanting Chen, 王云鹤, 郭天宇, 许昌, 邓一波, 刘振华, 马思伟, 许春静, 许超, 和高文. 预训练图像处理Transformer。在CVPR，2021。3[5] Jierun Chen, 肖杰, 郝浩, 赵伟鹏, 文松, 李柱浩, 和 S-H Gary Chan. 跑，别走：追逐更快神经网络的更高FLOPs。在CVPR，2023。5, 8[6] Liangyu Chen, 崔晓洁, 张祥宇, 和 孙剑. 图像恢复的简单基线。在ECCV，2022。1[7] Wei-Ting Chen, 黄志凯, 蔡成哲, 杨浩翔, 丁建钧, 和 郭思彦. 通过两阶段知识学习和多对比正则化学习多种恶劣天气去除：迈向统一模型。在CVPR，2022。5, 6[8] 陈翔, 李昊, 李明强, 和 潘金山. 学习稀疏Transformer网络进行有效图像去雨。在CVPR，2023。2, 3, 5, 6, 8[9] Sung-Jin Cho, Ji Seo-Won, Hong Jun-Pyo, Jung Seung-Won, 和 Ko Sung-Jea. 单图像去模糊中粗粒度到细粒度方法的重新思考。在ICCV，2021。2[10] Xin Deng 和 Pier Luigi Dragotti. 用于多模态图像恢复与融合的深度卷积神经网络。TPAMI, 43(10):3333–3348, 2021。2[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Zhai Xiaohua, Unterthiner Thomas, Dehghani Mostafa, Minderer Matthias, Heigold Georg, Gelly Sylvain, Uszkoreit Jakob, 和 Houlsby Neil. 一幅图像值16x16个词：用于大规模图像识别的Transformer。在ICLR，2021。2, 4, 7[12] David Eigen, Dilip Krishnan, 和 Rob Fergus. 恢复透过覆盖灰尘或雨水的窗户拍摄的图像。在ICCV，2013。5, 6[13] 付雪阳, 黄嘉斌, 曾德尔, 黄越, 丁兴豪, 和 Paisley John. 通过深度细节网络从单图像中去除雨水。在CVPR，2017。5, 6[14] 付雪阳, 钱琪, 扎正军, 朱宇瑞, 和 丁兴豪. 通过双图卷积网络去除雨迹。在AAAI，2021。2, 5, 6, 8[15] 付雪阳, 肖杰, 朱宇瑞, 刘爱平, 吴峰, 和 扎正军. 基于超图卷积网络的持续图像去雨。TPAMI, 45 9551, 2023 8 9534 5, 6 ( ): – . [16] Hamidreza Ghader 和 Christof Monz. 神经机器翻译中的注意力机制到底关注什么？arXiv预印本 arXiv:1710.03348, 2017。7

[17] 顾舒航, 李亚伟, 卢克·范古尔, 和 拉杜·蒂莫夫特. 自主导向网络用于快速图像去噪. ICCV, 2019. 2[18] 郭春乐, 严启新, 安萨尔·安瓦尔, 从明聪, 任文琪, 和 李崇毅. 具有传输感知3D位置嵌入的图像去雾Transformer. CVPR, 2022. 2, 5, 6[19] 郭云, 肖雪瑶, 张毅, 邓淑敏, 和 严鲁欣. 从天空到地面：面向真实雨移除的大规模基准和简单基线. ICCV, 2023. 5, 6[20] 何凯明, 孙坚, 和 戚秀峰. 使用暗通道先验的单图像雾移除. TPAMI, 33(12):2341– 2353, 2010. 2, 5, 6[21] 何凯明, 张祥宇, 任少卿, 和 孙坚. 深度残差学习用于图像识别. CVPR, 2016. 2[22] Andrew G Howard, 朱孟龙, 陈波, Dmitry Kalenichenko, 王伟军, Tobias Weyand, Marco An- dreetto, 和 Hartwig Adam. Mobilenets: 用于移动视觉应用的高效卷积神经网络. arXiv preprint arXiv:1704.04861, 2017. 5[23] Jiri Hron, Yasaman Bahri, Jascha Sohl-Dickstein, 和 Roman Novak. 无限注意力：Nngp 和 Ntk 用于深度注意力网络. ICML, 2020. 2[24] 华伟哲, 戴志航, 刘汉晓, 和 Le Quoc. 线性时间内的Transformer质量. ICML, 2022. 3[25] 黄高, 刘庄, Van Der Maaten Laurens, 和 Weinberger Kili- an Q. 密集连接卷积网络. CVPR, 2017. 2[26] Phillip Isola, 朱俊彦, 周挺辉, 和 Alexei A Efros. 基于条件对抗网络的图像到图像翻译. CVPR, 2017. 5, 6[27] Kevin Jarrett, Koray Kavukcuoglu, Marc’ Aurelio Ranzato, 和 Yann LeCun. 什么是最佳的多阶段架构用于目标识别？ ICCV, 2009 2 . 贾国利 和 杨聚[28] 2峰. S -ver: 半监督视觉情绪识别. ECCV, 2022. 2[29] Orest Kupyn, Tetiana Martyniuk, 吴俊儒, 和 王张阳. Deblurgan-v2: 更快（数量级）更好去模糊. ICCV, 2019. 2[30] 李博文, 彭秀丽, 王张阳, 许继正, 和 冯丹. Aod-net: 一体化去雾网络. ICCV, 2017. 5, 6[31] 李崇毅, 郭春乐, 韩凌浩, 蒋俊, 程明明, 顾金伟, 和 Loy Chen Change. 使用深度学习进行低光照图像和视频增强：综述. TPAMI, 44 12 9396 9416, 2022 2 ( ): – . 李若腾, [32] Tan Robby T., 和 Cheong Loong-Fah. 使用架构搜索的一体化恶劣天气移除. CVPR, 2020. 5, 6[33] 李翔, 吴建龙, 林卓辰, 刘红, 和 赵洪斌. 用于单图像去雨的循环Squeeze-and-Excitation上下文聚合网络. ECCV, 2018. 5, 6[34] 李亚伟, 张凯, 曹继章, 拉杜·蒂莫夫特, 和 卢克·范古尔. Localvit: 为视觉Transformer引入局部性. arXiv preprint arXiv:2104.05707, 2021. 5, 7

References

[1] Codruta O Ancuti, Cosmin Ancuti, Mateu Sbert, and Radu Timofte. Dense-haze: A benchmark for image dehazing with dense-haze and haze-free images. In *ICIP*, 2019. 5, 6

[2] Haoran Bai, Jinshan Pan, Xinguang Xiang, and Jinhui Tang. Self-guided image dehazing using progressive feature fusion. *TIP*, 31:1217–1229, 2022. 5, 6

[3] Pierre Charbonnier, Laure Blanc-Feraud, Gilles Aubert, and Michel Barlaud. Two deterministic half-quadratic regularization algorithms for computed imaging. In *ICIP*, 1994. 4

[4] 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 *CVPR*, 2021. 3

[5] Jierun Chen, Shiu-hong Kao, Hao He, Weipeng Zhuo, Song Wen, Chul-Ho Lee, and S-H Gary Chan. Run, don't walk: Chasing higher flops for faster neural networks. In *CVPR*, 2023. 5, 8

[6] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *ECCV*, 2022. 1

[7] Wei-Ting Chen, Zhi-Kai Huang, Cheng-Che Tsai, Hao-Hsiang Yang, Jian-Jiun Ding, and Sy-Yen Kuo. Learning multiple adverse weather removal via two-stage knowledge learning and multi-contrastive regularization: Toward a unified model. In *CVPR*, 2022. 5, 6

[8] Xiang Chen, Hao Li, Mingqiang Li, and Jinshan Pan. Learning a sparse transformer network for effective image deraining. In *CVPR*, 2023. 2, 3, 5, 6, 8

[9] Sung-Jin Cho, Seo-Won Ji, Jun-Pyo Hong, Seung-Won Jung, and Sung-Jea Ko. Rethinking coarse-to-fine approach in single image deblurring. In *ICCV*, 2021. 2

[10] Xin Deng and Pier Luigi Dragotti. Deep convolutional neural network for multi-modal image restoration and fusion. *TPAMI*, 43(10):3333–3348, 2021. 2

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 2, 4, 7

[12] David Eigen, Dilip Krishnan, and Rob Fergus. Restoring an image taken through a window covered with dirt or rain. In *ICCV*, 2013. 5, 6

[13] Xueyang Fu, Jiabin Huang, Delu Zeng, Yue Huang, Xinghao Ding, and John Paisley. Removing rain from single images via a deep detail network. In *CVPR*, 2017. 5, 6

[14] Xueyang Fu, Qi Qi, Zheng-Jun Zha, Yurui Zhu, and Xinghao Ding. Rain streak removal via dual graph convolutional network. In *AAAI*, 2021. 2, 5, 6, 8

[15] Xueyang Fu, Jie Xiao, Yurui Zhu, Aiping Liu, Feng Wu, and Zheng-Jun Zha. Continual image deraining with hypergraph convolutional networks. *TPAMI*, 45(8):9534–9551, 2023. 5, 6

[16] Hamidreza Ghader and Christof Monz. What does attention in neural machine translation pay attention to? *arXiv preprint arXiv:1710.03348*, 2017. 7

[17] Shuhang Gu, Yawei Li, Luc Van Gool, and Radu Timofte. Self-guided network for fast image denoising. In *ICCV*, 2019. 2

[18] Chun-Le Guo, Qixin Yan, Saeed Anwar, Runmin Cong, Wenqi Ren, and Chongyi Li. Image dehazing transformer with transmission-aware 3d position embedding. In *CVPR*, 2022. 2, 5, 6

[19] Yun Guo, Xueyao Xiao, Yi Chang, Shumin Deng, and Luxin Yan. From sky to the ground: A large-scale benchmark and simple baseline towards real rain removal. In *ICCV*, 2023. 5, 6

[20] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *TPAMI*, 33(12):2341–2353, 2010. 2, 5, 6

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2

[22] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 5

[23] Jiri Hron, Yasaman Bahri, Jascha Sohl-Dickstein, and Roman Novak. Infinite attention: Nngp and ntk for deep attention networks. In *ICML*, 2020. 2

[24] Weizhe Hua, Zihang Dai, Hanxiao Liu, and Quoc Le. Transformer quality in linear time. In *ICML*, 2022. 3

[25] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, 2017. 2

[26] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 5, 6

[27] Kevin Jarrett, Koray Kavukcuoglu, Marc'Aurelio Ranzato, and Yann LeCun. What is the best multi-stage architecture for object recognition? In *ICCV*, 2009. 2

[28] Guoli Jia and Jufeng Yang. S 2-ver: Semi-supervised visual emotion recognition. In *ECCV*, 2022. 2

[29] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *ICCV*, 2019. 2

[30] Boyi Li, Xiulian Peng, Zhangyang Wang, Jizheng Xu, and Dan Feng. Aod-net: All-in-one dehazing network. In *ICCV*, 2017. 5, 6

[31] Chongyi Li, Chunle Guo, Linghao Han, Jun Jiang, Ming-Ming Cheng, Jinwei Gu, and Chen Change Loy. Low-light image and video enhancement using deep learning: A survey. *TPAMI*, 44(12):9396–9416, 2022. 2

[32] Ruoteng Li, Robby T. Tan, and Loong-Fah Cheong. All in one bad weather removal using architectural search. In *CVPR*, 2020. 5, 6

[33] Xia Li, Jianlong Wu, Zhouchen Lin, Hong Liu, and Hongbin Zha. Recurrent squeeze-and-excitation context aggregation net for single image deraining. In *ECCV*, 2018. 5, 6

[34] Yawei Li, Kai Zhang, Jiezhang Cao, Radu Timofte, and Luc Van Gool. Localvit: Bringing locality to vision transformers. *arXiv preprint arXiv:2104.05707*, 2021. 5, 7

[35] 李亚伟, 范宇晨, 邢小雨, Denis Demandolx, Rakesh Ranjan, 拉杜·蒂莫夫特, 和 Van Luc Gool. 用于图像恢复的图像层次结构的有效和显式建模. CVPR, 2023. 3[36] 李志远, Bhojanapalli Srinadh, Zaheer Manzil, Reddi Sashank, 和 Kumar Sanjiv. 使用尺度不变架构的神经网络鲁棒训练. ICML, 2022. 3[37] 梁景云, 曹继章, 孙国雷, 张凯, 卢克·范古尔, 和 拉杜·蒂莫夫特. Swinir: 使用 Swin Transformer进行图像恢复. ICCV Workshops, 2021. 1, 3, 6[38] Bee Lim, Son Sanghyun, Kim Heewon, Nah Seungjun, 和 Lee Kyoung Mu. 用于单图像超分辨率的增强深度残差网络. CVPR Workshops, 2017. 6[39] 林迪, 王欣, 沈佳, 张仁杰, 刘若南, 王妙惠, 谢宇原, 郭清, 和 李平. 用于图像雨移除的生成状态估计和信息解耦. NeurIPS, 2022. 2, 5, 6[40] 刘丁, 温比安, 范宇晨, Loy Chen Change, 和 Huang Thomas S. 用于图像恢复的非局部循环网络. NeurIPS, 2018. 2[41] 刘娜, 李伟, 王银健, 陶然, 杜倩, 和 Chanussot Jocelyn. 基于低秩张量逼近的超光谱图像恢复综述. SCIS, 66(4):140302, 2023. 2[42] 刘欣 和 杨聚峰. 用于对应剪枝的渐进式邻域一致性挖掘. CVPR, 2023. 2[43] 刘晓红, 马永瑞, 石志豪, 和 陈俊. GridDehazeNet: 基于注意力的多尺度网络用于图像去雾. ICCV, 2019. 5, 6[44] 刘行, Suganuma Masanori, Sun Zhun, 和 Okatani Takayuki. 利用成对操作的潜力进行图像恢复的双残差网络. CVPR, 2019. 2, 5, 6[45] 刘欣, 肖国宝, 陈日庆 , 和 Ma Jiayi. Pgfnet: 用于双视图对应学习的偏好引导滤波网络. TIP, 32:1367–1378, 2023. 2[46] 刘泽, 林宇通, 曹越, 胡汉, 魏奕璇, 张铮, Lin Stephen, 和 Guo Baining. Swin Transformer: 使用移位窗口的层次化视觉 Transformer. ICCV, 2021. 3, 4[47] Ilya Loshchilov 和 Frank Hutter. 解耦权重衰减正则化. arXiv preprint arXiv:1711.05101, 2017. 5[48] Ilya Loshchilov 和 Frank Hutter. Sgdr: 带有热重启的随机梯度下降. ICLR, 2017. 5 [49] Andreas Lugmayr, Danelljan Martin, 拉杜·蒂莫夫特, Kim Kang- wook, Kim Younggeun, Lee Jae-young, Li Zechao, 潘金山, Shim Dongseok, Song Ki-Ung, 唐金辉, Wang Cong, 和 Zhao Zhihao. Ntire 2022挑战：学习超分辨率空间. CVPR Workshops, 2022. 2[50] 罗方舟, 吴晓林, 和 Guo Yanhui. 用于参数化图像恢复问题的功能神经网络. NeurIPS, 2021. 2[51] 马宁宁, 张祥宇, 刘明, 和 孙坚. 激活或不激活：学习定制激活. CVPR, 2021. 7

[52] Anish Mittal、Rajiv Soundararajan 和 Alan C Bovik。制作“完全盲”的图像质量分析器。SPL，20(3):209–212, 2012。 5[53] Ben Niu、Weilei Wen、Wenqi Ren、Xiangde Zhang、Lian- ping Yang、Shuzhen Wang、Kaihao Zhang、Xiaochun Cao 和 Haifeng Shen。通过整体注意力网络进行单图像超分辨率。在 ECCV，2020。 2[54] Ozan ¨Ozdenizci 和 Robert Legenstein。使用基于块的降噪扩散模型在恶劣天气条件下恢复视力。TPAMI，45(8):10346–10357，2023。 2, 5, 6[55] Jinshan Pan、Deqing Sun、Hanspeter Pfister 和 Ming- Hsuan Yang。通过暗通道先验对图像去模糊。TPAMI，40(10):2315–2328, 2017。 1[56] Namuk Park 和 Songkuk Kim。视觉 Transformer是如何工作的？在 ICLR，2022。 4[57] Kuldeep Purohit、Maitreya Suin、AN Rajagopalan 和 Vishnu Naresh Boddeti。使用失真引导网络的适应性图像恢复。在 ICCV，2021。 5, 6[58] Rui Qian、Robby T Tan、Wenhan Yang、Jiajun Su 和 Jiay- ing Liu。用于从单图像中去除雨滴的注意力生成对抗网络。在 CVPR，2018。 5, 6 [59] Xu Qin、Zhilin Wang、Yuanchao Bai、Xiaodong Xie 和 Huizhu Jia。FFA-Net：用于单图像去雾的特征融合注意力网络。在 AAAI，2020。 5, 6[60] Yuwei Qiu、Kaihao Zhang、Chenxi Wang、Wenhan Luo、Hongdong Li 和 Zhi Jin。Mb-taylorformer：使用泰勒公式扩展的多分支高效 Transformer用于图像去雾。在 ICCV，2023。 2, 5, 6[61] Ruijie Quan、Xin Yu、Yuanzhi Liang 和 Yi Yang。一次性去除雨滴和雨迹。在 CVPR，2021。 5, 6[62] Yuhui Quan、Shijie Deng、Yixin Chen 和 Hui Ji。用于透过有雨滴的窗户的深度学习。在 ICCV，2019。 5, 6[63] Dongwei Ren、Wangmeng Zuo、Qinghua Hu、Pengfei Zhu 和 Deyu Meng。渐进式图像去雨网络：更好更简单的基线。在 CVPR，2019。 2, 5, 6[64] Olaf Ronneberger、Philipp Fischer 和 Thomas Brox。U- net：用于生物医学图像分割的卷积网络。在 MICCAI，2015。 2[65] Yuanjie Shao、Lerenhan Li、Wenqi Ren、Changxin Gao 和 Nong Sang。图像去雾的领域自适应。在 CVPR，2020。 5, 6[66] Kai Shen、Junliang Guo、Xu Tan、Siliang Tang、Rui Wang 和 Jiang Bian。Transformer 中的 relu 和 softmax 研究。arXiv preprint arXiv:2302.06461，2023。 2, 4[67] David R So、Wojciech Ma´nke、Hanxiao Liu、Zihang Dai、Noam Shazeer 和 Quoc V Le。Primer：为语言建模搜索高效的 Transformer。arXiv preprint arXiv:2109.08668，2021。 2[68] Xibin Song、Dingfu Zhou、Wei Li、Yuchao Dai、Zhelun Shen、Liangjun Zhang 和 Hongdong Li。Tusr-net：具有自正则化和双重特征到像素注意力的三重展开单图像去雾。TIP，32 1244，2023 1231 2： – 。

[35] Yawei Li, Yuchen Fan, Xiaoyu Xiang, Denis Demandolx, Rakesh Ranjan, Radu Timofte, and Van Luc Gool. Efficient and explicit modelling of image hierarchies for image restoration. In *CVPR*, 2023. 3

[36] Zhiyuan Li, Srinadh Bhojanapalli, Manzil Zaheer, Sashank Reddi, and Sanjiv Kumar. Robust training of neural networks using scale invariant architectures. In *ICML*, 2022. 3

[37] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *ICCV Workshops*, 2021. 1, 3, 6

[38] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPR Workshops*, 2017. 6

[39] Di Lin, Xin Wang, Jia Shen, Renjie Zhang, Ruonan Liu, Miaohui Wang, Wuyuan Xie, Qing Guo, and Ping Li. Generative status estimation and information decoupling for image rain removal. In *NeurIPS*, 2022. 2, 5, 6

[40] Ding Liu, Bihan Wen, Yuchen Fan, Chen Change Loy, and Thomas S Huang. Non-local recurrent network for image restoration. In *NeurIPS*, 2018. 2

[41] Na Liu, Wei Li, Yinjian Wang, Ran Tao, Qian Du, and Jocelyn Chanussot. A survey on hyperspectral image restoration: From the view of low-rank tensor approximation. *SCIS*, 66(4):140302, 2023. 2

[42] Xin Liu and Jufeng Yang. Progressive neighbor consistency mining for correspondence pruning. In *CVPR*, 2023. 2

[43] Xiaohong Liu, Yongrui Ma, Zhihao Shi, and Jun Chen. GridDehazeNet: Attention-based multi-scale network for image dehazing. In *ICCV*, 2019. 5, 6

[44] Xing Liu, Masanori Suganuma, Zhun Sun, and Takayuki Okatani. Dual residual networks leveraging the potential of paired operations for image restoration. In *CVPR*, 2019. 2, 5, 6

[45] Xin Liu, Guobao Xiao, Riqing Chen, and Jiayi Ma. Pgfnet: Preference-guided filtering network for two-view correspondence learning. *TIP*, 32:1367–1378, 2023. 2

[46] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 3, 4

[47] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[48] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *ICLR*, 2017. 5

[49] Andreas Lugmayr, Martin Danelljan, Radu Timofte, Kangwook Kim, Younggeun Kim, Jae-young Lee, Zechao Li, Jinshan Pan, Dongseok Shim, Ki-Ung Song, Jinhui Tang, Cong Wang, and Zhihao Zhao. Ntire 2022 challenge on learning the super-resolution space. In *CVPR Workshops*, 2022. 2

[50] Fangzhou Luo, Xiaolin Wu, and Yanhui Guo. Functional neural networks for parametric image restoration problems. In *NeurIPS*, 2021. 2

[51] Ningning Ma, Xiangyu Zhang, Ming Liu, and Jian Sun. Activate or not: Learning customized activation. In *CVPR*, 2021. 7

[52] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a “completely blind” image quality analyzer. *SPL*, 20(3):209–212, 2012. 5

[53] Ben Niu, Weilei Wen, Wenqi Ren, Xiangde Zhang, Lianping Yang, Shuzhen Wang, Kaihao Zhang, Xiaochun Cao, and Haifeng Shen. Single image super-resolution via a holistic attention network. In *ECCV*, 2020. 2

[54] Ozan Özdenizci and Robert Legenstein. Restoring vision in adverse weather conditions with patch-based denoising diffusion models. *TPAMI*, 45(8):10346–10357, 2023. 2, 5, 6

[55] Jinshan Pan, Deqing Sun, Hanspeter Pfister, and Ming-Hsuan Yang. Deblurring images via dark channel prior. *TPAMI*, 40(10):2315–2328, 2017. 1

[56] Namuk Park and Songkuk Kim. How do vision transformers work? In *ICLR*, 2022. 4

[57] Kuldeep Purohit, Maitreya Suin, AN Rajagopalan, and Vishnu Naresh Boddeti. Spatially-adaptive image restoration using distortion-guided networks. In *ICCV*, 2021. 5, 6

[58] Rui Qian, Robby T Tan, Wenhan Yang, Jiajun Su, and Jiaying Liu. Attentive generative adversarial network for raindrop removal from a single image. In *CVPR*, 2018. 5, 6

[59] Xu Qin, Zhilin Wang, Yuanchao Bai, Xiaodong Xie, and Huizhu Jia. Ffa-net: Feature fusion attention network for single image dehazing. In *AAAI*, 2020. 5, 6

[60] Yuwei Qiu, Kaihao Zhang, Chenxi Wang, Wenhan Luo, Hongdong Li, and Zhi Jin. Mb-taylorformer: Multi-branch efficient transformer expanded by taylor formula for image dehazing. In *ICCV*, 2023. 2, 5, 6

[61] Ruijie Quan, Xin Yu, Yuanzhi Liang, and Yi Yang. Removing raindrops and rain streaks in one go. In *CVPR*, 2021. 5, 6

[62] Yuhui Quan, Shijie Deng, Yixin Chen, and Hui Ji. Deep learning for seeing through window with raindrops. In *ICCV*, 2019. 5, 6

[63] Dongwei Ren, Wangmeng Zuo, Qinghua Hu, Pengfei Zhu, and Deyu Meng. Progressive image deraining networks: A better and simpler baseline. In *CVPR*, 2019. 2, 5, 6

[64] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 2

[65] Yuanjie Shao, Lerenhan Li, Wenqi Ren, Changxin Gao, and Nong Sang. Domain adaptation for image dehazing. In *CVPR*, 2020. 5, 6

[66] Kai Shen, Junliang Guo, Xu Tan, Siliang Tang, Rui Wang, and Jiang Bian. A study on relu and softmax in transformer. *arXiv preprint arXiv:2302.06461*, 2023. 2, 4

[67] David R So, Wojciech Mańke, Hanxiao Liu, Zihang Dai, Noam Shazeer, and Quoc V Le. Primer: Searching for efficient transformers for language modeling. *arXiv preprint arXiv:2109.08668*, 2021. 2

[68] Xibin Song, Dingfu Zhou, Wei Li, Yuchao Dai, Zhelun Shen, Liangjun Zhang, and Hongdong Li. Tusr-net: Triple unfolding single image dehazing with self-regularization and dual feature to pixel attention. *TIP*, 32:1231–1244, 2023. 2

[69] Yuda Song、Zhuqing He、Hui Qian 和 Xin Du。用于单图像去雾的视觉Transformer。TIP，32:1927– 1941，2023。 5, 6[70] Fu-Jen Tsai、Yan-Tsung Peng、Yen-Yu Lin、Chung-Chi Tsai 和 Chia-Wen Lin。Stripformer：用于快速图像去模糊的条带Transformer。在 ECCV，2022。 5[71] Zhengzhong Tu、Hossein Talebi、Han Zhang、Feng Yang、Peyman Milanfar、Alan Bovik 和 Yinxiao Li。Maxim：用于图像处理的多轴MLP。在 CVPR，2022。 2, 5, 6[72] Jeya Maria Jose Valanarasu、Rajeev Yasarla 和 Vishal M. Patel。Transweather：基于 Transformer 的恶劣天气条件下图像恢复。在 CVPR，2022。 5, 6[73]Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。注意力就是全部。在 NeurIPS，2017。 1, 3, 4, 5[74]Cong Wang、Jinshan Pan、Wei Wang、Jiangxin Dong、Mengzhu Wang、Yakun J u、Junyang Chen 和 Xiao- Ming Wu。Promptrestorer：一种具有退化感知的提示图像恢复方法。在 NeurIPS，2023。 3[75] Hong Wang 、 Qi Xie 、 Qian Zhao 和 Deyu Meng。用于单图像雨移除的模型驱动深度神经网络。在 CVPR，2020。 5, 6[76] Lijuan Wang、Guoli Jia、Ning Jiang、Haiying Wu 和 Jufeng Yang。Ease：通过情绪模糊敏感合作网络进行鲁棒面部表情识别。在 ACMMM，2022。 2[77] Pichao Wang、Xue Wang、Fan Wang、Ming Lin、Shuning Chang、Hao Li 和 Rong Jin。Kvt：用于提升视觉Transformer的k-nn注意力。在 ECCV，2022。 1[78] Tianyu Wang、Xin Yang、Ke Xu、Shaozhe Chen、Qiang Zhang 和 Rynson WH Lau。具有高质量真实雨数据集的空间注意力单图像去雨。在 CVPR，2019。 2, 5, 6, 8[79] Wenhai Wang、Enze Xie、Xiang Li、Deng-Ping Fan、Kaitao Song、Ding Liang、Tong Lu、Ping Luo 和 Ling Shao。Pvt v2：使用金字塔视觉Transformer的改进基线。CVMJ，8(3):415–424，2022。 3[80] Yingqian Wang、Longguang Wang、Zhengyu Liang、Jun- gang Yang、Radu Timofte 和 Yulan Guo。Ntire 2023 挑战赛：光场图像超分辨率数据集、方法和结果。arXiv preprint arXiv:2304.10415，2023。 2[81] Zhou Wang、Alan C Bovik、Hamid R Sheikh 和 Eero P Simoncelli。图像质量评估：从错误可见性到结构相似性。TIP，13(4):600–612，2004。 5[82] Zhendong Wang、Xiaodong Cun、Jian min Bao、Wengang Zhou、Jianzhuang Liu 和 Houqiang Li。Uformer：一种通用的U形Transformer用于图像恢复。在 CVPR，2022。 1, 2, 3, 4, 5, 6, 7, 8[83] Changsong Wen、Guoli Jia 和 Jufeng Yang。Dip：用于讽刺检测的双重不一致感知网络。在 CVPR，2023。 2[84] Changsong Wen、Xin Zhang、Xingxu Yao 和 Jufeng Yang。序标签分布学习。在 ICCV，2023。 2

[85] 米切尔·沃茨曼，李载雄，贾斯汀·吉尔默和西蒙·科恩布拉特。在视觉Transformer中用relu替换softmax。arXiv预印本arXiv:2309.08586，2023。2[86] 吴海燕，曲艳云，林少辉，周健，乔瑞志，张志忠，谢元和马立庄。紧凑单图像去雾的对比学习。在CVPR，2021。5，6[87] 吴瑞琪，段正鹏，郭春乐，柴志和李崇毅。Ridcp：通过高质量码本先验重焕真实图像去雾。在CVPR，2023。5，6[88] 肖杰，付雪阳，刘爱平，吴峰和赵忠军。图像去雨Transformer。TPAMI，45(11): 12978–12995，2022。5，6，8[89] 肖铁，曼纳特·辛格，埃里克·明顿，特雷弗·达雷尔，皮奥特·多尔和罗斯·吉尔希克。早期卷积帮助Transformer看得更好。在NeurIPS，2021。4[90] 徐刚，侯启斌和程明明。双频率Transformer用于高效sdr-to-hdr转换。MIR，2024。3[91] 杨阳，王超越，刘日升，张林，郭晓杰和陶大程。通过密度和深度分解进行自增强无配对图像去雾。在CVPR，2022。5，6[92] 杨依欣，潘金山，彭忠正，杜晓宇，陶朱林和唐金辉。Bistnet：语义图像先验引导双向时间特征融合用于深度示例视频着色。TPAMI，2024。2[93] 叶天，陈思翔，白金斌，石俊，薛成浩，蒋静霞，尹俊杰，陈尔康和刘云。使用码本先验的对抗天气去除。在ICCV，2023。2，5，6[94] 易巧思，李俊成，戴庆艳，方发明，张贵旭和曾铁勇。具有残差通道先验引导的结构保持去雨。在ICCV，2021。5，6[95] 于可，王新涛，董超，唐晓鸥和刘晨辉。Path-restore：学习网络路径选择用于图像恢复。TPAMI，44 7092，2022 10 7078 2 ( ): – 。[96] 于伟豪，司晨阳，周盘，罗米，周一晨，冯嘉石，闫水成和王新超。Metaformer基线用于视觉。TPAMI，46(2):896–912，2023。7[97] 李源，刘新怡，于建南和李艳峰。基于改进点云的全牙齿分割模型++。VI，1(1):21，2023。2[98] 岳宗胜，赵倩，张雷和孟德宇。双对抗网络：面向真实世界噪声去除和噪声生成。在ECCV，2020。2[99] 赛义德·瓦卡斯·扎米尔，阿迪亚·阿罗拉，萨拉曼·汗，蒙纳瓦尔·海亚特，法哈德·沙巴兹·汗，杨明轩和邵凌。多阶段渐进图像恢复。在CVPR，2021。5，6[100] 赛义德·瓦卡斯·扎米尔，阿迪亚·阿罗拉，萨拉曼·汗，蒙纳瓦尔·海亚特，法哈德·沙巴兹·汗和杨明轩。Restormer：高效Transformer用于高分辨率图像恢复。在CVPR，2022。1，2，3，5，6，7，8

[69] Yuda Song, Zhuqing He, Hui Qian, and Xin Du. Vision transformers for single image dehazing. *TIP*, 32:1927–1941, 2023. 5, 6

[70] Fu-Jen Tsai, Yan-Tsung Peng, Yen-Yu Lin, Chung-Chi Tsai, and Chia-Wen Lin. Stripformer: Strip transformer for fast image deblurring. In *ECCV*, 2022. 5

[71] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxim: Multi-axis mlp for image processing. In *CVPR*, 2022. 2, 5, 6

[72] Jeya Maria Jose Valanarasu, Rajeev Yasarla, and Vishal M. Patel. Transweather: Transformer-based restoration of images degraded by adverse weather conditions. In *CVPR*, 2022. 5, 6

[73] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 1, 3, 4, 5

[74] Cong Wang, Jinshan Pan, Wei Wang, Jiangxin Dong, Mengzhu Wang, Yakun Ju, Junyang Chen, and Xiao-Ming Wu. Promptrestorer: A prompting image restoration method with degradation perception. In *NeurIPS*, 2023. 3

[75] Hong Wang, Qi Xie, Qian Zhao, and Deyu Meng. A model-driven deep neural network for single image rain removal. In *CVPR*, 2020. 5, 6

[76] Lijuan Wang, Guoli Jia, Ning Jiang, Haiying Wu, and Jufeng Yang. Ease: Robust facial expression recognition via emotion ambiguity-sensitive cooperative networks. In *ACMMM*, 2022. 2

[77] Pichao Wang, Xue Wang, Fan Wang, Ming Lin, Shuning Chang, Hao Li, and Rong Jin. Kvt: k-nn attention for boosting vision transformers. In *ECCV*, 2022. 1

[78] Tianyu Wang, Xin Yang, Ke Xu, Shaozhe Chen, Qiang Zhang, and Rynson WH Lau. Spatial attentive single-image deraining with a high quality real rain dataset. In *CVPR*, 2019. 2, 5, 6, 8

[79] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *CVMJ*, 8(3):415–424, 2022. 3

[80] Yingqian Wang, Longguang Wang, Zhengyu Liang, Jungang Yang, Radu Timofte, and Yulan Guo. Ntire 2023 challenge on light field image super-resolution: Dataset, methods and results. *arXiv preprint arXiv:2304.10415*, 2023. 2

[81] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 13(4):600–612, 2004. 5

[82] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *CVPR*, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[83] Changsong Wen, Guoli Jia, and Jufeng Yang. Dip: Dual incongruity perceiving network for sarcasm detection. In *CVPR*, 2023. 2

[84] Changsong Wen, Xin Zhang, Xingxu Yao, and Jufeng Yang. Ordinal label distribution learning. In *ICCV*, 2023. 2

[85] Mitchell Wortsman, Jaehoon Lee, Justin Gilmer, and Simon Kornblith. Replacing softmax with relu in vision transformers. *arXiv preprint arXiv:2309.08586*, 2023. 2

[86] Haiyan Wu, Yanyun Qu, Shaohui Lin, Jian Zhou, Ruizhi Qiao, Zhizhong Zhang, Yuan Xie, and Lizhuang Ma. Contrastive learning for compact single image dehazing. In *CVPR*, 2021. 5, 6

[87] Ruiqi Wu, Zhengpeng Duan, Chunle Guo, Zhi Chai, and Chongyi Li. Ridcp: Revitalizing real image dehazing via high-quality codebook priors. In *CVPR*, 2023. 5, 6

[88] Jie Xiao, Xueyang Fu, Aiping Liu, Feng Wu, and Zheng-Jun Zha. Image de-raining transformer. *TPAMI*, 45(11): 12978–12995, 2022. 5, 6, 8

[89] Tete Xiao, Mannat Singh, Eric Mintun, Trevor Darrell, Piotr Dollar, and Ross Girshick. Early convolutions help transformers see better. In *NeurIPS*, 2021. 4

[90] Gang Xu, Qibin Hou, and Ming-Ming Cheng. Dual frequency transformer for efficient sdr-to-hdr translation. *MIR*, 2024. 3

[91] Yang Yang, Chaoyue Wang, Risheng Liu, Lin Zhang, Xiaojie Guo, and Dacheng Tao. Self-augmented unpaired image dehazing via density and depth decomposition. In *CVPR*, 2022. 5, 6

[92] Yixin Yang, Jinshan Pan, Zhongzheng Peng, Xiaoyu Du, Zhulin Tao, and Jinhui Tang. Bistnet: Semantic image prior guided bidirectional temporal feature fusion for deep exemplar-based video colorization. *TPAMI*, 2024. 2

[93] Tian Ye, Sixiang Chen, Jinbin Bai, Jun Shi, Chenghao Xue, Jingxia Jiang, Junjie Yin, Erkang Chen, and Yun Liu. Adverse weather removal with codebook priors. In *ICCV*, 2023. 2, 5, 6

[94] Qiaosi Yi, Juncheng Li, Qinyan Dai, Faming Fang, Guixu Zhang, and Tieyong Zeng. Structure-preserving deraining with residue channel prior guidance. In *ICCV*, 2021. 5, 6

[95] Ke Yu, Xintao Wang, Chao Dong, Xiaoou Tang, and Chen Change Loy. Path-restore: Learning network path selection for image restoration. *TPAMI*, 44(10):7078–7092, 2022. 2

[96] Weihao Yu, Chenyang Si, Pan Zhou, Mi Luo, Yichen Zhou, Jiashi Feng, Shuicheng Yan, and Xinchao Wang. Metaformer baselines for vision. *TPAMI*, 46(2):896–912, 2023. 7

[97] Li Yuan, Xinyi Liu, Jiannan Yu, and Yanfeng Li. A full-set tooth segmentation model based on improved pointnet++. *VI*, 1(1):21, 2023. 2

[98] Zongsheng Yue, Qian Zhao, Lei Zhang, and Deyu Meng. Dual adversarial network: Toward real-world noise removal and noise generation. In *ECCV*, 2020. 2

[99] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *CVPR*, 2021. 5, 6

[100] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, 2022. 1, 2, 3, 5, 6, 7, 8

[101] 翟英杰，贾国利，赖宇坤，张静，杨聚峰和陶大程。通过双边姿态和运动图卷积网络从步态中感知情绪。TAFFC，2024。2[102] 张标，伊万·蒂托夫和里科·森尼奇。线性单元的稀疏注意力。在EMNLP，2021。3，4[103] 张凯，李亚伟，左旺萌，张雷，卢克·范古尔和拉杜·蒂莫夫特。即插即用的图像恢复与深度去噪先验。TPAMI，44 6376，2021 10 6360 1, 2 ( ): – 。[104] 张凯豪，任文琪，罗文涵，赖伟胜，Bj¨orn施滕格，杨明轩和李洪东。深度图像去模糊：一项调查。IJCV，130 2130，2022 9 2103 2 ( ): – 。[105] 张宇伦，李坤鹏，李凯，王立晨，钟本能和付云。使用非常深的残差通道注意力网络的图像超分辨率。在ECCV，2018。2[106] 张宇伦，李坤鹏，李凯，钟本能和付云。用于图像恢复的残差非局部注意力网络。在ICLR，2019。2[107] 张志成和杨聚峰。时序情感定位：在未修剪视频中听和看。在ACMMM，2022。2[108] 张志成，刘胜哲和杨聚峰。多平面目标跟踪。在ICCV，2023 2 。[109] 张志成，王丽娟和杨聚峰。通过跨模态时间擦除网络进行弱监督视频情感检测和预测。在CVPR，2023。2[110] 张志成，陈宋，王自川和杨聚峰。Planeseg：构建一个插件来提高平面区域分割。TNNLS，2024。2[111] 张志成，胡军耀，程文涛，帕德尔·达anda和杨聚峰。Extdm：分布外推扩散模型用于视频预测。在CVPR，2024。2[112] 张志成，赵盘成，朴恩尼尔和杨聚峰。Mart：通过掩码时间分布蒸馏进行掩码情感表征学习。在CVPR，2024。2[113] 赵广祥，林俊阳，张志远，任旭成，苏琪和孙旭。显式稀疏Transformer：通过显式选择集中注意力。arXiv预印本arXiv:1912.11637，2019。1[114] 赵海宇，勾元宝，李伯云，彭德忠，吕建城和彭希。全面而精致：一个高效的Transformer用于图像恢复。在CVPR，2023。2，3，6[115] 赵盘成，徐鹏，秦鹏达，范登平，张志成，贾国利，周 Bowen和杨聚峰。Lake-red：通过潜在背景知识检索增强扩散生成的迷彩图像。在CVPR，2024。2[116] 周曼，黄杰，郭春乐和李崇毅。Fourmer：一种用于图像恢复的高效全局建模范例。在ICML，2023。2，5，6[117] 周时豪，蒋梦溪，王奇丛和雷云琪。面向三维人体姿态估计的局部相似性保持。在ACCV，2020。2

[118] 周时豪，蒋梦溪，蔡珊珊，和雷云琪。Dc-gnet：深度网格关系捕获图卷积网络用于三维人体形状重建。在*ACMMM*，2021年。2

[101] Yingjie Zhai, Guoli Jia, Yu-Kun Lai, Jing Zhang, Jufeng Yang, and Dacheng Tao. Looking into gait for perceiving emotions via bilateral posture and movement graph convolutional networks. *TAFFC*, 2024. 2

[102] Biao Zhang, Ivan Titov, and Rico Sennrich. Sparse attention with linear units. In *EMNLP*, 2021. 3, 4

[103] Kai Zhang, Yawei Li, Wangmeng Zuo, Lei Zhang, Luc Van Gool, and Radu Timofte. Plug-and-play image restoration with deep denoiser prior. *TPAMI*, 44(10):6360–6376, 2021. 1, 2

[104] Kaihao Zhang, Wenqi Ren, Wenhan Luo, Wei-Sheng Lai, Björn Stenger, Ming-Hsuan Yang, and Hongdong Li. Deep image deblurring: A survey. *IJCV*, 130(9):2103–2130, 2022. 2

[105] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, 2018. 2

[106] Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. Residual non-local attention networks for image restoration. In *ICLR*, 2019. 2

[107] Zhicheng Zhang and Jufeng Yang. Temporal sentiment localization: Listen and look in untrimmed videos. In *ACMMM*, 2022. 2

[108] Zhicheng Zhang, Shengzhe Liu, and Jufeng Yang. Multiple planar object tracking. In *ICCV*, 2023. 2

[109] Zhicheng Zhang, Lijuan Wang, and Jufeng Yang. Weakly supervised video emotion detection and prediction via cross-modal temporal erasing network. In *CVPR*, 2023. 2

[110] Zhicheng Zhang, Song Chen, Zichuan Wang, and Jufeng Yang. Planeseg: Building a plug-in for boosting planar region segmentation. *TNNLS*, 2024. 2

[111] Zhicheng Zhang, Junyao Hu, Wentao Cheng, Danda Paudel, and Jufeng Yang. Extdm: Distribution extrapolation diffusion model for video prediction. In *CVPR*, 2024. 2

[112] Zhicheng Zhang, Pancheng Zhao, Eunil Park, and Jufeng Yang. Mart: Masked affective representation learning via masked temporal distribution distillation. In *CVPR*, 2024. 2

[113] Guangxiang Zhao, Junyang Lin, Zhiyuan Zhang, Xuancheng Ren, Qi Su, and Xu Sun. Explicit sparse transformer: Concentrated attention through explicit selection. *arXiv preprint arXiv:1912.11637*, 2019. 1

[114] Haiyu Zhao, Yuanbiao Gou, Boyun Li, Dezhong Peng, Jiancheng Lv, and Xi Peng. Comprehensive and delicate: An efficient transformer for image restoration. In *CVPR*, 2023. 2, 3, 6

[115] Pancheng Zhao, Peng Xu, Pengda Qin, Deng-Ping Fan, Zhicheng Zhang, Guoli Jia, Bowen Zhou, and Jufeng Yang. Lake-red: Camouflaged images generation by latent background knowledge retrieval-augmented diffusion. In *CVPR*, 2024. 2

[116] Man Zhou, Jie Huang, Chun-Le Guo, and Chongyi Li. Fourmer: an efficient global modeling paradigm for image restoration. In *ICML*, 2023. 2, 5, 6

[117] Shihao Zhou, Mengxi Jiang, Qicong Wang, and Yunqi Lei. Towards locality similarity preserving to 3d human pose estimation. In *ACCV*, 2020. 2

[118] Shihao Zhou, Mengxi Jiang, Shanshan Cai, and Yunqi Lei. Dc-gnet: Deep mesh relation capturing graph convolution network for 3d human shape reconstruction. In *ACMMM*, 2021. 2