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

Shihao Zhou¹,² Duosheng Chen¹ Jinshan Pan³ Jinglei Shi¹* Jufeng Yang¹,²
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 Adaptive Sparse Transformer (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.
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 Adaptive Sparse Transformer (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 reconstruction [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].
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, DRSformer [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 \( I \in \mathbb{R}^{H \times W \times C} \), AST first employs a convolution layer to produce a low-level feature representation \( F_0 \in \mathbb{R}^{H \times W \times C} \), where \( H \times W \), \( C \) are the image resolution and the number of channels, respectively. Next, the low-level representation \( F_0 \) passes through a \( N_1 \)-stage symmetric encoder-decoder network and is embedded into deep
feature \( F_d \in \mathbb{R}^{H \times W \times C} \). 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 upsampling. 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 \( R \in \mathbb{R}^{H \times W \times 3} \) from \( F_d \). The restored image is obtained by the sum of the degraded image and the residual one, i.e., \( \hat{I} = I + R \). The Charbonnier loss [3] is adopted to train AST:

\[
\ell(I', \hat{I}) = \sqrt{\|I' - \hat{I}\|^2 + \epsilon^2},
\]

where \( I' \) refers to the ground-truth image and we experimentally set \( \epsilon \) 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 \( X \in \mathbb{R}^{H \times W \times C} \), we begin by partitioning it into non-overlapping windows of size \( M \times M \), resulting in a flattened representation \( X^i \in \mathbb{R}^{M^2 \times C} \) from the \( i \)-th window. Then we generate matrices of \textit{queries} \( Q \), \textit{keys} \( K \) and \textit{values} \( V \) from \( X \):

\[
Q = XW_Q, \; K = XW_K, \; V = XW_V,
\]

where the linear projection matrices of the queries \( W_Q \), keys \( W_K \), and values \( W_V \in \mathbb{R}^{C \times d} \) that are shared among all windows. The attention computation can be defined as:

\[
A = f(QK^T/\sqrt{d} + B)V,
\]

where \( A \) denotes the estimated attention; \( B \) 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:

\[
DSA = SoftMax(QK^T/\sqrt{d} + B).
\]

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:

\[
SSA = ReLU^2(QK^T/\sqrt{d} + B).
\]

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:

\[
A = (w_1 * SSA + w_2 * DSA)V,
\]

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:

\[
w_n = \frac{e^{a_n}}{\sum_{i=1}^{N} e^{a_i}}, \; n = \{1, 2\}
\]

where \( \{a_1, a_2\} \) are learnable parameters that are initialized with 1 of the two branches. This design ensures a better
Table 1. Quantitative comparison on SPAD [78] for rain streak removal.

<table>
<thead>
<tr>
<th>Method</th>
<th>PSNR ↑</th>
<th>SSIM ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>DDN [13]</td>
<td>36.16</td>
<td>0.9463</td>
</tr>
<tr>
<td>RESCAN [33]</td>
<td>38.11</td>
<td>0.9797</td>
</tr>
<tr>
<td>PreNet [63]</td>
<td>40.16</td>
<td>0.9816</td>
</tr>
<tr>
<td>RCDNet [75]</td>
<td>43.36</td>
<td>0.9831</td>
</tr>
<tr>
<td>SPDNet [94]</td>
<td>43.55</td>
<td>0.9875</td>
</tr>
<tr>
<td>SPAIR [57]</td>
<td>44.10</td>
<td>0.9872</td>
</tr>
<tr>
<td>DualGCN [14]</td>
<td>44.18</td>
<td>0.9902</td>
</tr>
<tr>
<td>SEIDNet [39]</td>
<td>44.96</td>
<td>0.9911</td>
</tr>
<tr>
<td>MPRNet [99]</td>
<td>45.00</td>
<td>0.9897</td>
</tr>
<tr>
<td>Fu et al. [15]</td>
<td>45.03</td>
<td>0.9907</td>
</tr>
<tr>
<td>Restomer [100]</td>
<td>46.25</td>
<td>0.9911</td>
</tr>
<tr>
<td>SCF-Former [19]</td>
<td>46.89</td>
<td>0.9941</td>
</tr>
<tr>
<td>IDT [88]</td>
<td>47.34</td>
<td>0.9929</td>
</tr>
<tr>
<td>Uformer [82]</td>
<td>47.84</td>
<td>0.9925</td>
</tr>
<tr>
<td>DRSFormer [8]</td>
<td>48.53</td>
<td>0.9924</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Method</th>
<th>PSNR ↑</th>
<th>SSIM ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eigen’s [12]</td>
<td>21.31</td>
<td>0.757</td>
</tr>
<tr>
<td>Pix2pix [26]</td>
<td>27.36</td>
<td>0.836</td>
</tr>
<tr>
<td>Uformer [82]</td>
<td>29.42</td>
<td>0.906</td>
</tr>
<tr>
<td>WeatherDiff [54]</td>
<td>29.66</td>
<td>0.923</td>
</tr>
<tr>
<td>TransWeather [72]</td>
<td>30.17</td>
<td>0.916</td>
</tr>
<tr>
<td>WeatherDiff [54]</td>
<td>30.71</td>
<td>0.931</td>
</tr>
<tr>
<td>All-in-One [32]</td>
<td>31.12</td>
<td>0.927</td>
</tr>
<tr>
<td>DaRN [44]</td>
<td>31.24</td>
<td>0.926</td>
</tr>
<tr>
<td>Quan’s [62]</td>
<td>31.37</td>
<td>0.918</td>
</tr>
<tr>
<td>AttenGAN [58]</td>
<td>31.59</td>
<td>0.917</td>
</tr>
<tr>
<td>IDT [88]</td>
<td>31.87</td>
<td>0.931</td>
</tr>
<tr>
<td>MAXIM-2S [71]</td>
<td>31.87</td>
<td>0.935</td>
</tr>
<tr>
<td>AWRC [93]</td>
<td>31.93</td>
<td>0.931</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Method</th>
<th>PSNR ↑</th>
<th>SSIM ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>RIDCP [87]</td>
<td>8.09</td>
<td>0.42</td>
</tr>
<tr>
<td>DCP [20]</td>
<td>10.06</td>
<td>0.39</td>
</tr>
<tr>
<td>SGID [2]</td>
<td>13.09</td>
<td>0.52</td>
</tr>
<tr>
<td>D4 [91]</td>
<td>13.12</td>
<td>0.53</td>
</tr>
<tr>
<td>AOD-Net [30]</td>
<td>13.14</td>
<td>0.41</td>
</tr>
<tr>
<td>GridDehazeNet [43]</td>
<td>13.31</td>
<td>0.37</td>
</tr>
<tr>
<td>DA-Dehaze [65]</td>
<td>13.98</td>
<td>0.37</td>
</tr>
<tr>
<td>FFA-Net [59]</td>
<td>14.39</td>
<td>0.45</td>
</tr>
<tr>
<td>Uformer [82]</td>
<td>15.22</td>
<td>0.43</td>
</tr>
<tr>
<td>Restomer [100]</td>
<td>15.78</td>
<td>0.55</td>
</tr>
<tr>
<td>AECR-Net [86]</td>
<td>15.80</td>
<td>0.47</td>
</tr>
<tr>
<td>Fourmer [116]</td>
<td>15.95</td>
<td>0.49</td>
</tr>
<tr>
<td>DehazeFormer-S [69]</td>
<td>16.29</td>
<td>0.51</td>
</tr>
<tr>
<td>DeHammer [18]</td>
<td>16.62</td>
<td>0.56</td>
</tr>
<tr>
<td>MB-TaylorFormer-B [60]</td>
<td>16.66</td>
<td>0.56</td>
</tr>
</tbody>
</table>

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

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.
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], PRNet [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], WeatherDiff128 [54], TransWeather [72], WeatherDiff64 [54], TKLA: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 WeatherDiff128 [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], DeHammer [18] and MB-TaylorForm [60]. In Tab. 3, AST-B obtains the best values in PSNR metric among the considered 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
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:

\[
Enctropy_{\text{Att}} = - \frac{1}{H} \sum_h \frac{1}{L} \sum_{ij} Att_{ij}^{h,l} \log Att_{ij}^{h,l},
\]

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-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)}
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 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. 20JCYJJC00020), the National Natural Science Foundation of China (Nos. U22B2049, 62302240), Fundamental Research Funds for the Central Universities, and Supercomputing Center of Nankai University (NKSC).
References


[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, 7, 8


[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] Xueyao Xiao, Jiabin Huang, Delu Zeng, Yue Huang, Xing-hao Ding, and John Paisley. Removing rain from single images via a deep detail network. In CVPR, 2017. 5, 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


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


[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. TAFCC, 2024. 2


