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, 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

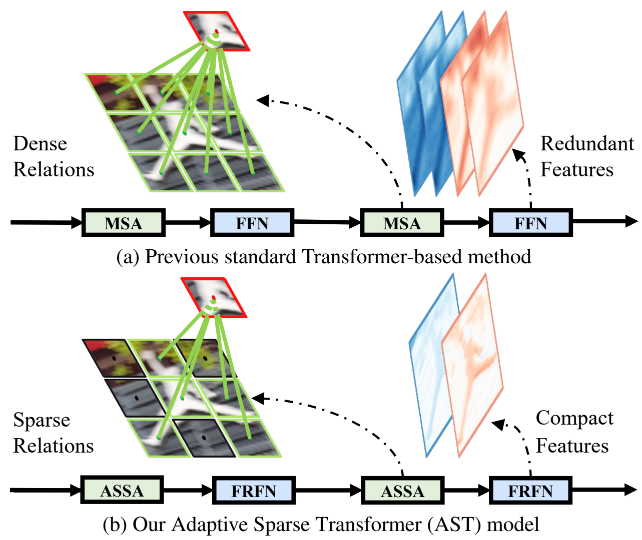


Figure 1. Workflows comparisons. (a) The standard Transformer-based method incorporates all available tokens into multihead self-attention (MSA) calculation, and the feed-forward network (FFN) to handle redundant features. (b) The proposed Adaptive Sparse Transformer (AST) includes an adaptive sparse 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.

*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 **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].

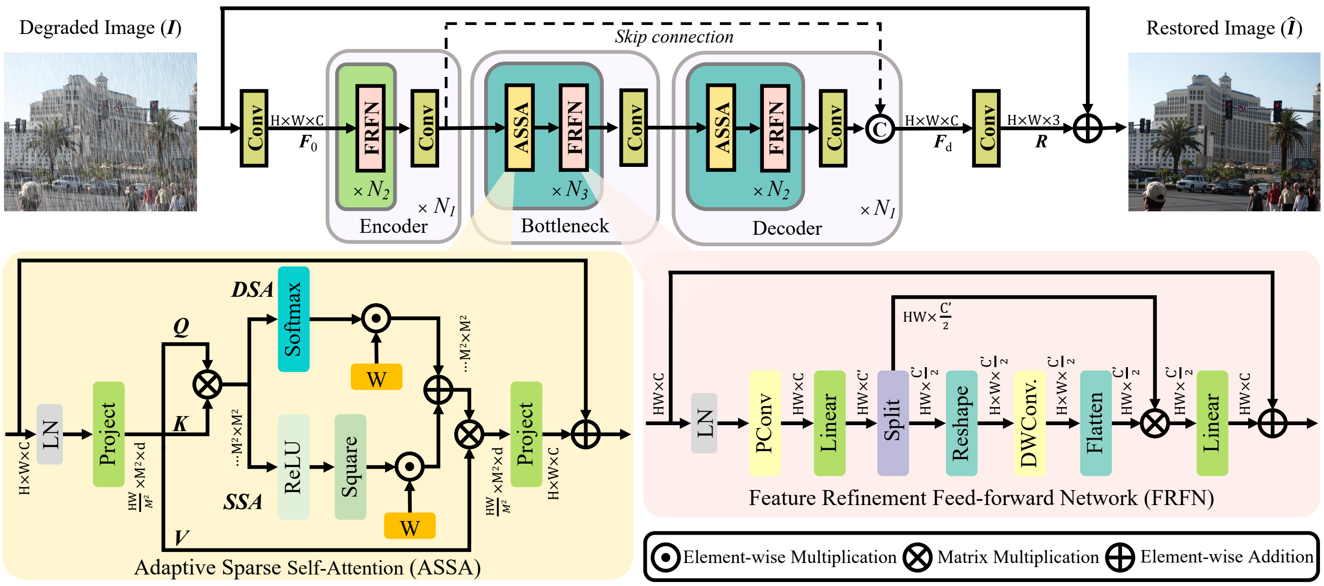


Figure 2. Overview of our Adaptive Sparse Transformer (AST). It mainly consists of an Adaptive Sparse Self-Attention (ASSA), and a Feature Refinement Feed-forward Network (FRFN). LN refers to layer normalization and Conv denotes convolution operation.

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, 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 $\boldsymbol{I} \in \mathbb{R}^{H\times W\times 3}$, 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 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

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)	<u>49.51</u>	<u>0.9942</u>
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	<u>0.935</u>
AWRCP [93]	31.93	0.931
AST-B (Ours)	<u>32.32</u>	<u>0.935</u>
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	<u>0.56</u>
MB-TaylorFormer-B [60]	16.66	<u>0.56</u>
AST-B (Ours)	<u>17.12</u>	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$(\cdot)$ and DWConv$(\cdot)$ refer to the partial convolution [5] and depthwise 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

Reference | Rainy | SPDNet [94] | Restormer [100] | DRSformer [14] | Ours

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], RESCAN [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 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].

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

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

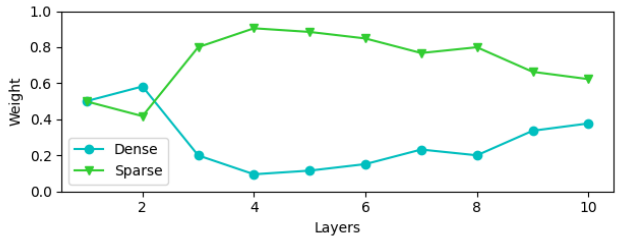


Figure 4. Learned weights for sparse and dense branches.

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

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	-

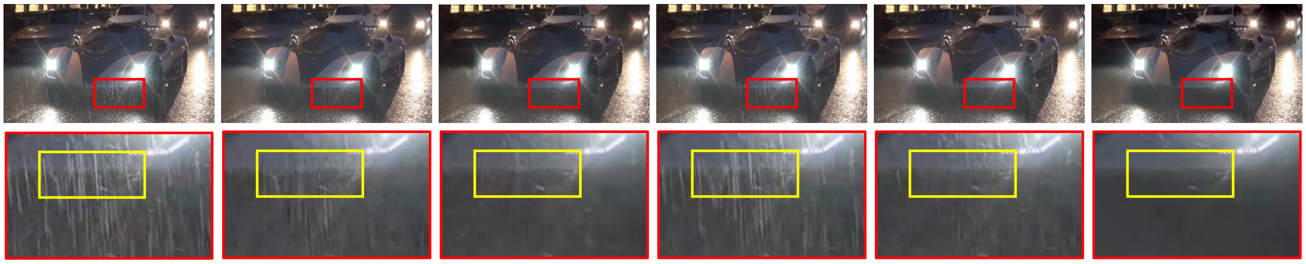
Table 8. Ablation study on alternatives to feature refinement feed-forward network.

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

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) 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

(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.

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. 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).

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] 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, 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] 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