Lips-SpecFormer: Non-Linear Interpolable Transformer for Spectral Reconstruction using Adjacent Channel Coupling

Abhishek Kumar Sinha
aks@sac.isro.gov.in
S. Manthira Moorthi
smmoorthi@sac.isro.gov.in

Signal and Image Processing Area
Space Applications Center
Ahmedabad, India

Abstract

Spectral Recovery from RGB images is a challenging yet interesting domain to learn end-to-end spectral mapping using neural nets. Transformers have recently gained popularity due to their ability to learn long range dependencies through self-attention. In our work, we show that spectral feature learning with self-attention is prone to instability. We propose a transformer based network for spectral reconstruction from RGB images with a Non-Linear Interpolable Spectral Attention (N-LISA) to learn the spectral features. We further analyse the stability of the N-LISA using the theory of Lipschitz constant. The method is evaluated and compared with different state-of-the-art methods on multiple standard datasets. In addition, ablation analysis is performed to analyse the effectiveness of the proposed spectral attention, and other modules.

1 Introduction

Hyperspectral images capture real world images in multiple narrow bands with each band describing the spectral behaviour in that bandwidth. These images have a pivotal role in scientific spectral analysis such as remote sensing, material analysis and medical computer vision as they provide information that cannot be seen by human eyes. Hyperspectral imaging, being a time consuming process, tends to utilize an algorithmic approach to reduce the acquisition time. However, it is known that spectral reconstruction is an ill-posed problem [17]. Hyperspectral to RGB projection can be thought of as projecting the hyperspectral image vector along the spectral response space that results in the loss of the image vector lying in the null space of spectral response, and therefore the exact inverse mapping cannot be performed without the unknown null space vector. The application of Deep Convolutional Networks (CNNs) has shown promising results in end-to-end mapping. In recent years, transformers gained popularity for application in computer vision problems. They found applications in low-level vision problems like image super-resolution [21, 26], image inpainting [15], and so on.

Computer vision tasks employ different methods to learn long-range spectral dependency, including self-attention, squeeze and excitation network, and non-local neural nets.

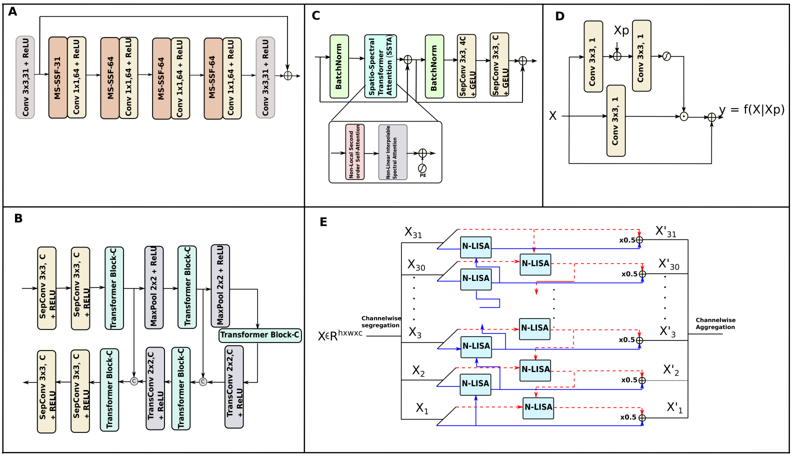


Figure 1: A: End-to-end transformer network. B: MSSSFB-C: Multi-Scale Spatio Spectral Feature Block with C number of input channels. C: Transformer block with C number of input channels. D: N-LISA: Non-Linear Interpolable Spectral Attention architecture. E: Architecture of spectral attention using N-LISA.

Self-attention is the key essence of exploiting long range dependencies in Transformers. However, this approach to estimate the spectral attention coefficients along spectral channels has serious limitations in spectral recovery tasks. Intuitively, for a feature map with C number of channels, the corresponding $C \times C$ shaped attention matrix uses a scalar value to correlate the spatial variation between two channels. Furthermore, the Lipschitz constant of the self-attention layer is proportional to the variance in input that results in a larger sensitivity factor [11]. To alleviate this issue, we propose a spectral attention layer that is relatively more stable than self-spectral attention. The major contributions of this work are:

1. We propose a transformer network with linear complexity for end-to-end RGB to spectral mapping. The proposed method uses Non-Linear Interpolable Spectral Attention (N-LISA) for long-range spectral context and non-local second-order self-attention along the spatial dimension.

2. We utilize the theory of Lipschitz constant to show that under trivial assumptions, N-LISA is relatively more stable than the spectralwise self-attention network.

3. The proposed architecture is evaluated on multiple benchmark datasets. The results show that the model outperforms the existing state-of-the-art deep learning approaches, and is adversarially more robust than other transformer architectures.

2 Related Works

2.1 Spectral Reconstruction

A wide variety of conventional hyperspectral image reconstruction [7, 8, 18] is based on handcrafted priors. These methods suffer from poor reconstruction quality and generalization issues. Recently, deep learning has played key role in overcoming these limitations.

HSCNN presented the deep learning framework to synthesize hyperspectral images from under-sampled projections. Subsequently, HSCNN+ and HSCNN-R [25] were developed using residual blocks and replacing handcrafted upsampling blocks for improved performance. Adaptive Weighted Attention Network [13](AWAN) uses dual residual attention block and patch level non-local attention block to capture the long-range spatial feature dependencies. NTIRE 2020 spectral reconstruction challenge tested the performance of multiple deep learning approaches [13, 22, 33] on a public dataset for clean and real-world tracks. Another proposed approach uses a 4-level hierarchical regression network [33] with PixelShuffle layer to learn the inter-level context information. All CNN based approaches fail to exploit the non-local information due to limited receptive field of convolution kernel. This results in degraded performance on real world dataset of NTIRE 2022. NTIRE 2022 presented new models [5, 14] that were evaluated on real world spectral data. DRCR-net [14] suggests purifying the interference and noises in the real world RGB images using a non-local purification module and captures the spatio-spectral interaction through Dense Residual Channel Re-calibration (DRCR). The performance of DRCRNet is restricted by the assumption that the images are noise corrupted in similar fashion. MST [4] proposed a transformer based approach that employs spectral wise multi-head self-attention to learn inter-channel features. It leverages the conventional self-attention for spectral attention that induces instability in the spectral profile.

2.2 Transformers beyond matrix dot-product

Transformers in computer vision have raised the bar for state-of-the-art in various tasks. The pioneer work by Kolesnikov ([12]) uses transformer in the form of 16 × 16 patches for visual recognition. Due to quadratic complexity in terms of both computational and memory usage, various methods have been explored such as patch-free methods ([30]) and product-free attention ([20, 31]). The product-free methods in [20, 31] utilize Hadamard product and replace key-query interaction by a non-linear function. Our work also proposes a dot-product free self-attention to linearize computational complexity but ensures faithful signal propagation for extract necessary spectral features. Fastformer [27] is another work that utilizes additive self-attention instead of representing pairwise interactions by applying additive attention to model global context and then subsequently transforming each token based on the global context. Overall, the main gist of replacing dot-product attention is to overcome the quadratic complexity that serves as a major bottleneck in processing high fidelity images. Another interesting work, called as PoolFormer ([29]), replaces the self-attention by pooling layer to achieve competitive performance in multiple vision related tasks. This study shows that the Metaformer structure is responsible for superiority in terms of performance instead of token mixer module. This study aids to the freedom of choosing suitable token mixer over conventional self-attention to further improve the performance in multiple aspects.

3 Proposed Methodology

Figure 1 shows the overall end-to-end architecture. It primarily consists of Multi-Scale Spatio-Spectral Feature Block (MS-SSF) followed by a pointwise convolution, and a residual connection is used to avoid the vanishing gradient problems. MS-SSF block learns spatial and spectral dependencies at different scales. The pointwise convolution layer scales the number of channels in intermediate layers without changing the spatial context. Figure 1B shows the architecture of the MS-SSF block that follows U-Net ([24]) like architecture to allow the use of multi-level context through feature concatenation. MS-SSF block uses a separable convolution layer for feature transformation, and a transformer block to learn spatial-spectral feature dependencies. Meanwhile, [9] empirically shows that depthwise convolution within the separable convolution is known to behave similarly to the local-attention of Local Vision Transformer, and therefore enhances the performance at no extra cost. The transformer block, as shown in Figure 1C, uses residual architecture and batch normalization for training stability. While the transformers in NLP tasks are inclined towards LayerNorm, the CNN based architectures for vision problems are more batch norm friendly. Many works have shown that batch norm outperforms layer norm for properly chosen batch size [6]. Though BatchNorm based pure self-attention suffers from instability issues, it works reasonably well for mixed architecture like N-LISA. Spatio-Spectral Transformer Attention (SSTA) is the core attention module to learn inter-channel and spatial interactions using N-LISA and non-local second-order self-attention and the resulting attention coefficient is governed by the spectral features in the majority.

3.1 Lipschitz Stability of Non-Linear Interpolable Self-Attention

The motivation to propose N-LISA is to overcome the limitations of using spectral wise self-attention for spectral dependencies. Firstly, To apply self-attention along the spectral dimension on the $X \in \mathbf{R}^{H\times W\times C}$ shaped feature map, the corresponding spectral attention coefficient using estimated key $K \in \mathbf{R}^{C\times HW}$ and query $Q \in \mathbf{R}^{C\times HW}$ is computed as $A_{ij} = \sum_{k=0}^{HW-1} Q_{i,k}K_{k,j}^{T}$. It squeezes the spatio-spectral context between two channels to a single scalar value causing the information loss. Second,the L_2 Lipschitz constant of self-attention, as shown in Corollary 1, is bounded by the variance of the input resulting in larger sensitivity [11].

Corollary 1. *Let m and M be the minimum and maximum values of X, and W^Q and W^K be the query and key weights in self-attention.The upper bound on the magnitude of diagonal elements of Jacobian in self-attention network is given by,*

$$\|J_{i,i}\|_2 \leq \frac{\|W^K W^Q\|_2}{4} + \|W^K W^Q\|_2 \frac{(M-m)^2}{4} + 1, \tag{1}$$

with equality if $(softmax(XW^Q(XW^K)^T)_{i,i} = 1$ and $X_i = 1$.

Proof. From [11],

$$J_{ij} = W^K W^Q X^T P^{(i)}(E_{ji}X + \delta_{ij}X) + P_{ij}I$$

where W^K and W^Q are the weights of Key and Query respectively. P is computed as $P = softmax(\frac{XW^Q(XW^K)^T}{\sqrt{HW}})$, and $P^{(i)} = diag(P_{i:}) - P_{i:}^T P_{i:}$.
For $i = j$,

$$J_{ii} = W^K W^Q X^T P^{(i)} e_{ii} X + W^K W^Q Var(X) + P_{ii} \tag{2}$$

$$\|J_{ii}\|_2 < A.(P_{i,i}X_i - (P_{i,i}X_i)^2) + A.Var(X) + \|P_{ii}\|_2, \tag{3}$$

where $A = \|W^K W^Q\|_2$
Observe that $P_{i,i}X_i - (P_{i,i}X_i)^2$ is concave in $P_{i,i}X_i$ and has maxima for $P_{i,i}X_i = \frac{1}{2}$. For $\|P_{i,i}\|_2 = 1, X_i = 0.5$. Using this in $\|J_{i,i}\|_2$,we get $\|J_{ii}\|_2 \leq \frac{A}{4} + A.Var(X) + 1$ □

To combat this sensitivity issue, N-LISA is framed in such a way that it exploits the long range dependencies without squeezing the channel wise features' relation to a scalar value. The channel wise output of N-LISA can be mathematically described as,

$$Y_c = X_c + \frac{1}{2}\Big(V_c^F \odot \sigma(F_c^F(Q_c, K_{c-1})) + V_c^B \odot \sigma(F_c^B(Q_c, K_{c+1}))\Big), \tag{4}$$

where $F_c^u(Q_c, K_p) = W_u^Q * (X_c + W_u^K * X_p)$ and $V_c^u = f_u^F(X_c)$, for $u = \{F, B\}$ and p is $c-1$ for forward regressor (F) and $c+1$ for backward (B) resulting in bidirectional global propagation. $F_c^u(Q_c, K_p)$ and V_c^u are called non-linear cross-attention and Value embedding, respectively. σ and $\odot$ refer to sigmoid and elementwise-multiplication, respectively. For necessary derivations, we further simplify the gating vector F_c^u as,

$$\begin{aligned} F_c^u(Q_c, K_p) &= W_u^Q * (X_c + W_u^K * X_p) \\ &= W_u^Q * X_c + W_u^Q * W_u^K * X_p = g^u(X_c) + h^u(X_p) \end{aligned} \tag{5}$$

Substituting 5 in 4, we get,

$$Y_c = X_c + \frac{1}{2}\Big(f_c^F(X_c) \odot (\sigma(g_c^F(X_c) + h_c^F(Y_{c-1}))) + f_c^B(X_c) \odot (\sigma(g_c^B(X_c) + h_c^B(Y_{c+1})))\Big) \tag{6}$$

As shown in equation 6, each feature map is estimated as the average of forward and backward regressors to learn the residual bidirectional spectral dependencies. The long-range dependency is exploited by using the updated adjacent feature maps in the attention map. Intuitively, this mechanism does not necessarily require estimating the correlation with every spectral channel and implicitly passes the relevant features through the adjacent channels in a bidirectional manner. In equation 6, the functions f, g and h, being 2D convolution, can be represented using the matrix multiplication with corresponding operator matrix [19], i.e. $W * x \overset{\text{def}}{=} op(W)X$. Equation 6 can be rewritten as matrix operation on the vectorized mappings $X_c, Y_c \in \mathbf{R}^{HW}$ as,

$$\begin{aligned} Y_c = X_c + \frac{1}{2}\Big(& op(W_{f_c^F})X_c \odot \sigma(op(W_{g_c^F})X_c + op(W_{h_c^F})(Y_{c-1})) \\ & + op(W_{f_c^B})X_c \odot \sigma(op(W_{g_c^B})X_c + op(W_{h_c^B})Y_{c+1})\Big) \end{aligned} \tag{7}$$

Theorem 1. *Let $\omega = e^{2\pi i/HW}$ and W^f be the convolution kernel in the function f of N-LISA. Let Λ be the difference between the learned kernel and its initialization and given by $\Lambda = W^f - W_0^f$. Also, let F be a complex matrix such that $F_{ij} = \omega^{ij}$. If $\varepsilon^f = \frac{1}{9}(F^T \Lambda F)_{0,0}$, then upper bound on the magnitude of diagonal elements of Jacobian in N-LISA is given by,*

$$\begin{aligned} \|J_{i,i}\|_2 \leq & 1 + \frac{1}{8}\big((1+9\varepsilon_i^{f^F})(1+9\varepsilon_i^{g^F}) + (1+9\varepsilon_i^{f^B})(1+9\varepsilon_i^{g^B})\big)\sqrt{HW}max(\|M\|, \|m\|) \\ & + \frac{1}{2}\big((1+9\varepsilon_i^{f^F}) + (1+9\varepsilon_i^{f^B})\big) \end{aligned}$$

Proof. The proof is the immediate application of operator norm for convolution kernels in Long et. al. [19]. The operator norm of $c \times c \times 1$ shaped kernel W is $||op(W)||_2 = 1 + \varepsilon c^2$.

$$\begin{aligned} J_{i,i} = \frac{\partial Y_i}{\partial X_i} = 1 + \frac{1}{2}\Big(&diag(op(W_{f_i^F})X_i \odot \sigma(\alpha_1)(1-\sigma(\alpha_1)))op(W_{g_i^F}) \\ &+ diag(op(W_{f_i^B})X_i \odot \sigma(\alpha_2)(1-\sigma(\alpha_2)))op(W_{g_i^B})\Big) \\ &+ \frac{1}{2}\Big((op(W_{f_i^F})) \odot diag(\sigma(\alpha_1))\Big) + \frac{1}{2}\Big((op(W_{f_i^B})) \odot diag(\sigma(\alpha_2))\Big) \end{aligned} \tag{8}$$

Here, $\alpha_1 = g_i^F(X_i) + h_i^F(Y_{i-1})$ and $\alpha_2 = g_i^B(X_i) + h_i^B(Y_{i+1})$. Applying the operator norm from Long et. al. [19] and taking the L_2 norm to estimate the Euclidean Lipschitz constant,

$$\begin{aligned} \|J_{i,i}\|_2 \leq & 1 + \frac{1}{8}((1+9\varepsilon_i^{f^F})(1+9\varepsilon_i^{g^F}) + (1+9\varepsilon_i^{f^B})(1+9\varepsilon_i^{g^B}))\sqrt{HW} max(\|M\|, \|m\|) \\ &+ \frac{1}{2}((1+9\varepsilon_i^{f^F}) + (1+9\varepsilon_i^{f^B})) \end{aligned} \tag{9}$$

□

Theorem 1 shows that, unlike self attention, the magnitude of $\|J_{i,i}\|_2$ is only bounded by the absolute maximum of the input, and therefore relatively more stable for inputs with large dynamic range.

Theorem 2. *Let the loss function for transformer network be $\mathcal{L}_t = \rho_t + \frac{\gamma}{2}\|w_t\|_2^2$ at time t, where ρ is the data fidelity term and γ is L_2 regularisation parameter. Assume that there N numbers of $k \times k$ convolution filters in the neural net. The upper bound on the magnitude of Jacobian $\left\|J_{i,j}\right\|_2$ after T iterations is given by,*

$$\|J_{i,j}\|_2 \leq \Big(\prod_{k=j+1}^{i} \sqrt{\frac{2\rho_T}{\gamma N}} \|V_k^F \odot \sigma(F_k^F(Q_k, K_{k-1}))\|_2\Big) \left\|\frac{\partial Y_j}{\partial X_j}\right\|_2, \; for \; i > j \; and,$$

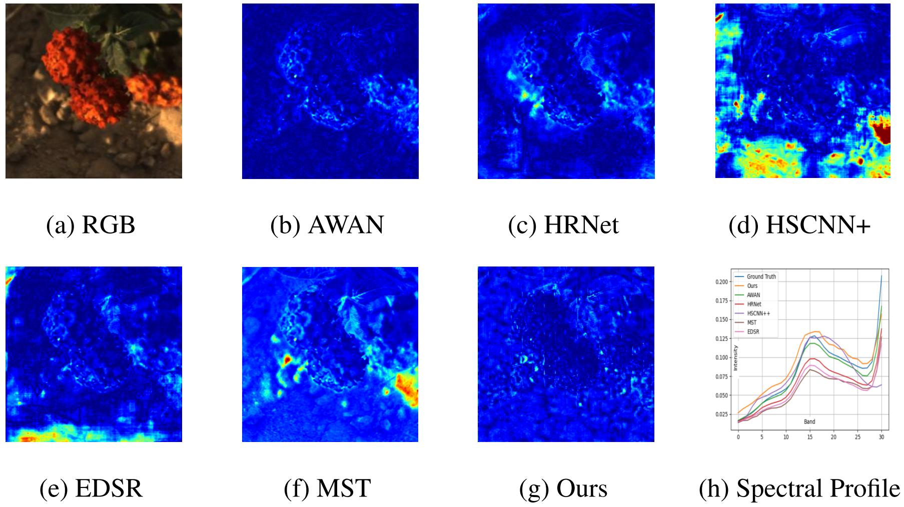
$$\|J_{i,j}\|_2 \leq \Big(\prod_{k=i}^{j-1} \sqrt{\frac{2\rho_T}{\gamma N}} \|V_k^B \odot \sigma(F_k^B(Q_k, K_{k+1}))\|_2\Big) \left\|\frac{\partial Y_j}{\partial X_j}\right\|_2, \; for \; i < j.$$

Proof. Without the loss of generality, we can assume that the choice of regularization constant is such that data fidelity term in the total loss dominates at any given step T,

$$\frac{\gamma}{2}\|w_T\|^2 \leq \rho_T \tag{10}$$

If the maximum L_2 norm of a kernel is C, then $\|w_T\|^2 \leq NC^2$. Substituting it in equation 10, we get $C \leq \sqrt{\frac{2\rho_T}{\gamma N}}$.

$$\begin{aligned} J_{i,j} = \frac{\partial Y_i}{\partial X_j} &= \frac{1}{2} diag(op(W_{f_i^F})X_i \odot \sigma(\alpha_1)(1-\sigma(\alpha_1)))op(W_{h_i^F})\frac{\partial Y_{i-1}}{\partial X_j} \\ &\leq \frac{1}{2} diag(V_i^F \odot \sigma(F_i^F(Q_i, K_{i-1})))op(W_{h_i^F})\frac{\partial Y_{i-1}}{\partial X_j} \\ &\leq \Big(\prod_{k=j+1}^{i} \frac{1}{2} diag(V_k^F \odot \sigma(F_k^F(Q_k, K_{k-1})))op(W_{h_k^F})\Big)\frac{\partial Y_j}{\partial X_j} \end{aligned} \tag{11}$$



(a) RGB (b) AWAN (c) HRNet (d) HSCNN+

(e) EDSR (f) MST (g) Ours (h) Spectral Profile

Figure 2: Illustration of residual map of the spectral band predicted by different methods. Spectral profile compares the spectral profiles generated by different methods.

$$\left\|J_{i,j}\right\|_2 \leq \left(\prod_{k=j+1}^{i} \frac{1}{2}\left\|V_k^F \odot \sigma(F_k^F(Q_k, K_{k-1}))\right\|_2 \left\|op(W_{h_k^F})\right\|_2\right) \left\|\frac{\partial Y_j}{\partial X_j}\right\|_2 \tag{12}$$

Substituting inequality for C in (12), we get

$$\left\|J_{i,j}\right\|_2 \leq \left(\prod_{k=j+1}^{i} \sqrt{\frac{2\rho_T}{\gamma N}}\left\|V_k^F \odot \sigma(F_k^F(Q_k, K_{k-1}))\right\|_2\right) \left\|\frac{\partial Y_j}{\partial X_j}\right\|_2 \tag{13}$$

The proof of $\left\|J_{i,j}\right\|_2$, $i < j$ follows the same approach. □

Further generalizing the upper bound on the magnitude of Jacobian, Theorem 2 presents the generalized version to estimate the bound on the magnitude of Jacobian. If the number of trainable convolution kernel is chosen such that $N > \frac{1}{\gamma}$ and ρ_T is closer to zero after training, then $\sqrt{\frac{2\rho_T}{\gamma N}}$ is always smaller than 1. Therefore, the perturbation in a particular channel gets attenuated for farther channels. Nevertheless, the key-query based gating controls the propagation of perturbation to the adjacent features.

4 Experiments and Results

4.1 Datasets and Implementation Details

Three publicly available datasets are used for training and performance assessment, including NTIRE 2020 [1], NTIRE 2022 [2], and CAVE [28] datasets. The network is trained on the training sets of NTIRE images, and evaluated on the provided validation sets. For CAVE images, 20 out of 32 images are randomly selected for training and reaming 12 images are

Table 1: Quantitative comparison of different spectral reconstruction methods. The best ones are shown in **bold**. P stands for parameters and Fl stands for FLOPS.

Method	P(M)	Fl(G)	CAVE		NTRIE 2020		NTIRE2022	
			RMSE	SAM	MRAE	RMSE	MRAE	RMSE
Bicubic	-	-	0.1689	34.382	0.1745	0.0506	0.2005	0.0712
HSCNN+	4.65	266.84	0.0353	12.208	0.0684	0.0182	0.3814	0.0588
HRNet	31.70	143.51	0.0298	8.150	0.0682	0.0178	0.3476	0.0550
EDSR	2.42	142.53	0.0384	8.755	0.0707	0.0162	0.3277	0.0437
AWAN	4.04	231.29	0.0375	8.654	0.0678	0.0175	0.2500	0.0367
HD-Net	2.66	173.81	0.0326	8.314	0.0722	0.0176	0.2047	0.0317
MPRNet	3.62	101.59	0.0294	7.864	0.0722	0.0168	0.1817	0.0270
MST	2.45	**26.29**	0.0289	7.812	0.0747	0.0173	0.1772	**0.0256**
Ours	**1.18**	36.84	**0.0246**	**7.661**	**0.0669**	**0.0158**	**0.1767**	0.0301

used to validate the performance. All of these datasets have 31 multi-spectral bands covering the visible spectra (400-700 nm) at an interval of 10 nm.

The RGB images are linearly scaled in the range of [0,1] and are fed as a batch of 64×64 cropped images. The batch size is set to 20, and the network is optimized using Adam optimizer with default setting of $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning is initialized to 0.0002 and subsequently reduced to 10^{-6} using cosine annealing for 300 epochs. Similar to [5], data augmentation is also performed using random flipping of the cropped images to avoid overfitting. The training is performed using Mean Relative Absolute Error (MRAE) as the loss function. The testing phase also requires linear scaling of RGB images to [0,1]. Owing to sequential estimation in N-LISA, the computation requires 1.58 seconds per image on testing dataset using single A100 GPU.

For NTIRE 2020 and 2022 datasets, we use MRAE and RMSE as the evaluation metrics for comparing the performance of different methods. For CAVE dataset images, RMSE and Spectral Angle Mapper (SAM) are used since zero value in CAVE images causes divergence of MRAE metric.

4.2 Performance Comparison

The proposed approach is compared with latest state-of-the-art methods, including AWAN [13], MST [4], HSCNN+ [25], HRNet [33], HD-Net [10], MPRNet [3] and EDSR [16]. Table 1 quantitatively compares the performances on three datasets. It is worth mentioning that our method outperforms the State-of-the-art models with fewer parameters. However, our approach requires relatively more number of FLOPS since the spectralwise attention is estimated for all spatial positions through convolution operation. Moreover, all CNN based architectures, even with relatively large number of parameters, perform worse than both transformer networks (MST and ours). Figures 2 illustrates the residual in the predicted spectral band of wavelength at 410 nm, and the spectral profile at the centre region of the image. It can be observed that other methods are sensitive to variation in brightness and contrast, and therefore incur large residual in the some of regions of predicted multispectral bands.

4.3 Complexity Analysis

We analyse the Big-$\mathcal{O}$ complexity of the proposed spectral self-attention for an input feature of dimension $H \times W \times C$. Formally, we assume that X is projected to key and value spaces. For time complexity of each channel, it involves: (1) Computation of $F(Q,K)$ for each channel: $\mathcal{O}(HW)$ (2) Element wise multiplication: $\mathcal{O}(HW)$. Adding all of them, the resulting total computation complexity for each channel is $\mathcal{O}(HW)$. The overall complexity for all channels together is $\mathcal{O}(HWC)$, which shows that N-LISA is linear spectral self-attention. Additionally, we only use the input feature map for computation and do not store any variable larger than the size of the feature map. Therefore, the memory complexity is $\mathcal{O}(HWC)$.

4.4 Ablation Studies

4.4.1 Lipschitz stability

To further validate the Theorems 1 and 2, we empirically estimate the Lipschitz constants by perturbing a specific channel(denoted by j in Tables 2 and 3) in the feature map. To estimate 2-Lipschitz constants, the inputs to the attention modules are perturbed to observe the corresponding change in the output. Tables 2 and 3 show the estimated Lipschitz constants for multihead-self-attention and N-LISA respectively. While 2-Lipschitz constant of diagonal Jacobian elements are found to be comparable, unlike multihead spectral self-attention, any perturbation in a given channel is not propagated to other channels in the feature maps of N-LISA. This clearly indicates that perturbation in the spectral channel of multihead-multispectral self-attention induces instability in the other channels too that should be suppressed for enhanced stability.

Table 2: Some of 2-Lipschitz constant for spectral self-attention of MST for different perturbed channel.

j	$\|\|J_{0,j}\|\|_2$	$\|\|J_{5,j}\|\|_2$	$\|\|J_{20,j}\|\|_2$
0	3.894	0.0011	0.0013
5	0.050	0.428	0.0011
20	0.00043	0.0008	0.426

Table 3: Some of 2-Lipschitz constant for spectral self-attention of N-LISA for different perturbed channel.

j	$\|\|J_{0,j}\|\|_2$	$\|\|J_{5,j}\|\|_2$	$\|\|J_{20,j}\|\|_2$
0	0.431	0	0
5	0	0.770	0
20	0	0	0.776

4.4.2 N-LISA vs Spectral MSA

To further establish the efficacy of the N-LISA block, we re-train the baseline architecture by replacing N-LISA with spectral multi-head self-attention (MSA) block [6]. The results from Table 4 proves that N-LISA outperforms the spectral MSA by significant margin.

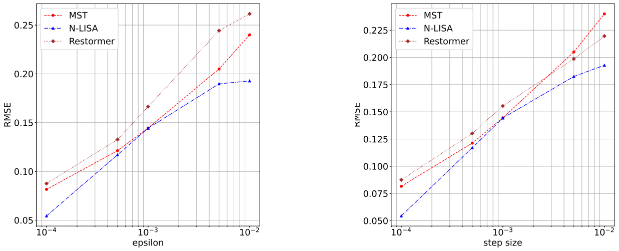
4.4.3 Adversarial robustness

Previously, few works, such as [32], have studied the connections between Lipschitz stability and the robustness under adversarial attack. In connection to this, we evaluate the

Table 4: Quantitative results for different attention layers.

	NTIRE-2020		NTIRE-2022	
Attention	**MRAE**	**RMSE**	**MRAE**	**RMSE**
MSA	0.1120	0.0420	0.2420	0.0512
N-LISA	0.0669	0.0158	0.1767	0.0301

performance of three transformer architectures for Fast Gradinet Sign Method (FGSM) and Projected Gradient Descent (PGD) attacks. For both FGSM and PGD-20 attacks, we tune the step-sizes and perturb the input RGB images from NTIRE-2022 dataset. The reason to choose NTIRE-2022 dataset is that RGB images contain camera noises, and therefore provides a realistic scenario to study adversarial robustness. From graphs in Figures 3(a) and 3(b), it is observed that the proposed method incurs lesser error in generated bands compared to MST and restormer. Hence, this supports our claim that N-LISA enhances the stability of spectral attention.



((a)) Robustness analysis under FGSM attack

((b)) Robustness analysis under PGD-20 attack

Figure 3: Comparison of adversarial robustness under FGSM and PGD-20 attacks.

5 Conclusions

In this paper, we present a dot product free attention, N-LISA, for channelwise spectral attention in transformer to overcome the limitations of pure self-attention. The proposed transformer network consists of alternating Multi Scale Spatio-Spectral Feature Blocks and Convolution layers and follows residual net like architecture. Multi Scale Spatio-Spectral Feature Block has UNet like architecture consisting of Transformer blocks and scaling layers where each transformer block utilizes a Non-local second order spatial self-attention and N-LISA for spectral attention. We thoroughly analyse the stability and long range dependency in terms of the Lipschitz constant. The experiments also show the superior performance of the proposed method on three benchmark spectral datasets. Despite this performance gain, the proposed approach requires relatively higher computation compared to other methods. Some recent works, such as Lipsformer ([23]), have already started addressing the Lipschitz stability of transformer emphasising the robustness of architecture. In conclusion, tighter Lipschitz bounds and improved computational efficiencies are the major aspects of the future scope of this work.

References

[1] Boaz Arad and Radu Timofte. Ntire 2020 challenge on spectral reconstruction from an rgb image. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1806–1822, 2020.

[2] Boaz Arad and Radu Timofte. Ntire 2022 spectral recovery challenge and data set. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 862–880, 2022.

[3] Aditya Bibitemmpr Syed Waqas Zamir, Salman Arora, Munawar Khan, Fahad Hayat, Ming-Hsuan Shahbaz Khan, and Ling Yang. Multi-Stage progressive image restoration. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14816–14826, Nashville, TN, USA, 2021.

[4] Yuanhao Cai, Jing Lin, Xiaowan Hu, Haoqian Wang, Xin Yuan, Yulun Zhang, Radu Timofte, and Luc Van Gool. Mask-guided spectral-wise transformer for efficient hyperspectral image reconstruction. 2022.

[5] Yuanhao Cai, Jing Lin, Zudi Lin, Haoqian Wang, Yulun Zhang, Hanspeter Pfister, Radu Timofte, and Luc Van Gool. Mst++: Multi-stage spectral-wise transformer for efficient spectral reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 744–754, 2022.

[6] Zhengsu Chen, Lingxi Xie, Jianwei Niu, Xuefeng Liu, Longhui Wei, and Qi Tian. Visformer: The vision-friendly transformer. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 569–578, 2021.

[7] J. Bioucas Dias and M. Figueiredo. A new twist: Two-step iterative shrinkage/thresholding algorithms for image restoration. *IEEE Transactions of Image Processing*, 16:2992–3004, 2007.

[8] MÁrio A. T. Figueiredo, Robert D. Nowak, and Stephen J. Wright. Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. *IEEE Journal of Selected Topics in Signal Processing*, 1(4):586–597, 2007. doi: 10.1109/JSTSP.2007.910281.

[9] Qi Han, Zejia Fan, Qi Dai, Lei Sun, Ming-Ming Cheng, Jiaying Liu, and Jingdong Wang. On the connection between local attention and dynamic depth-wise convolution. In *International Conference on Learning Representations*, 2022.

[10] Xiaowan Hu, Yuanhao Cai, Jing Lin, Haoqian Wang, Xin Yuan, Yulun Zhang, Radu Timofte, and Luc Van Gool. Hdnet: High-resolution dual-domain learning for spectral compressive imaging. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17521–17530, 2022. doi: 10.1109/CVPR52688.2022.01702.

[11] Hyunjik Kim, George Papamakarios, and Andriy Mnih. The lipschitz constant of self-attention. In *ICLR*, 2021.

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

[13] Jiaojiao Li, Chaoxiong Wu, Rui Song, Yunsong Li, and Fei Liu. Adaptive weighted attention network with camera spectral sensitivity prior for spectral reconstruction from rgb images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1894–1903, 2020.

[14] Jiaojiao Li, Songcheng Du, Chaoxiong Wu, Yihong Leng, Rui Song, and Yunsong Li. Drcr net: Dense residual channel re-calibration network with non-local purification for spectral super resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1258–1267, 2022.

[15] Wenbo Li, Zhe Lin, Lu Qi Kun Zho and, Yi Wang, and Jiaya Jia. Mat: Mask-aware transformer for large hole image inpainting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[16] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1132–1140, 2017.

[17] Yi-Tun Lin and Graham D. Finlayson. Physically plausible spectral reconstruction from rgb images. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2257–2266, 2020.

[18] Yang Liu, Xin Yuan, Jinli Suo, David J. Brady, and Qionghai Dai. Rank minimization for snapshot compressive imaging. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(12): 2990 – 3006, 2019.

[19] Philip M. Long and Hanie Sedghi. Generalization bounds for deep convolutional neural networks. In *ICLR*, 2020.

[20] Jiachen Lu, Jinghan Yao, Junge Zhang, Xiatian Zhu, Hang Xu, Weiguo Gao, Chunjing Xu, Tao Xiang, and Li Zhang. SOFT: Softmax-free transformer with linear complexity. In *Advances in Neural Information Processing Systems*, 2021.

[21] Zhisheng Lu, Juncheng Li, Hong Liu, Chaoyan Huang, Linlin Zhang, and Tieyong Zeng. Transformer for single image super-resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022.

[22] Hao Peng, Xiaomei Chen, and Jie Zhao. Residual pixel attention network for spectral reconstruction from rgb images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2012–2020, 2020.

[23] Xianbiao Qi, Jianan Wang, Yihao Chen, Yukai Shi, and Lei Zhang. Lipsformer: Introducing lipschitz continuity to vision transformers. In *The Eleventh International Conference on Learning Representations*, 2023.

[24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

[25] Zhan Shi, Chang Chen, Zhiwei Xiong, Dong Liu, and Feng Wu. Hscnn+: Advanced cnn-based hyperspectral recovery from rgb images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1052–10528, 2018.

[26] Abhishek Kumar Sinha, S. Manthira Moorthi, and Debajyoti Dhar. Nl-ffc: Non-local fast fourier convolution for image super resolution. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 466–475, 2022.

[27] Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. Fastformer: Additive attention can be all you need. *arXiv preprint arXiv:2108.09084*, 2021.

[28] Fumihito Yasuma, Tomoo Mitsunaga, Daisuke Iso, and Shree K. Nayar. Generalized assorted pixel camera: Postcapture control of resolution, dynamic range, and spectrum. *IEEE Transactions on Image Processing*, 19(9):2241–2253, 2010.

[29] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10819–10829, 2022.

[30] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. 2022.

[31] Shuangfei Zhai, Walter Talbott, Nitish Srivastava, Chen Huang, Hanlin Goh, Ruixiang Zhang, and Josh Susskind. An attention free transformer, 2021. URL `https://arxiv.org/pdf/2105.14103.pdf`.

[32] Bohang Zhang, Du Jiang, Di He, and Liwei Wang. Rethinking lipschitz neural networks and certified robustness: A boolean function perspective. 2022.

[33] Yuzhi Zhao, Lai-Man Po, Qiong Yan, Wei Liu, and Tingyu Lin. Hierarchical regression network for spectral reconstruction from rgb images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1695–1704, 2020.