Improving the Lipschitz Stability in Spectral Transformer through Nearest Neighbour Coupling

Abhishek Kumar Sinha [1]

Abstract

Statistical physics has played a pivotal role in the formulation of neural networks and understanding their behaviour. However, the effort to utilize the physical principle in the transformer architecture is still underexplored. In our work, we first show that spectral feature learning with self-attention is prone to instability. Inspired from the Ising model, we then propose a transformer based network using a adjacently coupled spectral attention and demonstrate its effectiveness in learning the spectral mapping from RGB images. We further analyse its stability using the theory of Lipschitz constant. The method is evaluated and compared with different state-of-the-art methods on multiple standard datasets.

1. Introduction

Neural networks are well known to exhibit properties that are commonly derived from statistical physics. This is clearly because of the large population of neurons in an network that certainly follows the fundamental physical laws. While some works have highlighted the underlying behaviour of neural nets in terms of these principles (Huang, 2023), some others explicitly utilize the principles from different physics domains and apply them to machine learning (Raissi et al., 2019). In this work, we apply the idea of nearest neighbour coupling from Ising model (Brush, 1967) and remodel the self-attention to learn spectral reconstruction. It is known that spectral reconstruction is an ill-posed problem Lin & Finlayson (2020). Hyperspectral to RGB projection can be thought as projecting the hyperspectral image vector along the spectral response space. This in turn 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. In recent years, transformers gained popularity for application in computer vision problems. They found applications in low level vision problems like image super-resolution Lu et al. (2022); Sinha et al. (2022), image inpainting Li et al. (2022), and so on. 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 task. 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 self-attention layer is proportional to the variance in input that results into larger sensitivity factor (Kim et al. (2021)). To alleviate this issue, we present a spectral attention layer that is relatively more stable than self-spectral attention. We further utilize the theory of Lipschitz constant to mathematically show the stability under trivial assumptions.

Notations X and Y are input and output feature maps from the proposed self-attention model. $W_f * X$ (or alternately $f(X)$) is used for convolution operation by kernel W, and $op(W_f)X$ signifies its equivalent matrix-vector multiplication using operator matrix of W. X_c and Y_c are the c^{th} input and output feature maps, respectively. Additionally, f_c^F and f_c^B are used to denote convolution operation on c^{th} channel in the forward and backward channel coupling, respectively. For studying Lipschitz stability, $J_{i,j}$ is the Lipschitz constant for i^{th} channel to j^{th} channel. Alternately, if j^{th} channel feature map is perturbed by a certain amount, the change in i^{th} channel feature map is $J_{i,j}$ times the perturbation on j^{th} channel.

2. Proposed Method

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 scales the number of channels in intermediate layers without changing

[*]Equal contribution [1]Signal and Image Processing Area, Space Applications Center. Correspondence to: Abhishek Kumar Sinha <aks@sac.isro.gov.in>.

Accepted after peer-review at the 1st workshop on Synergy of Scientific and Machine Learning Modeling, SynS & ML ICML, Honolulu, Hawaii, USA. July, 2023.

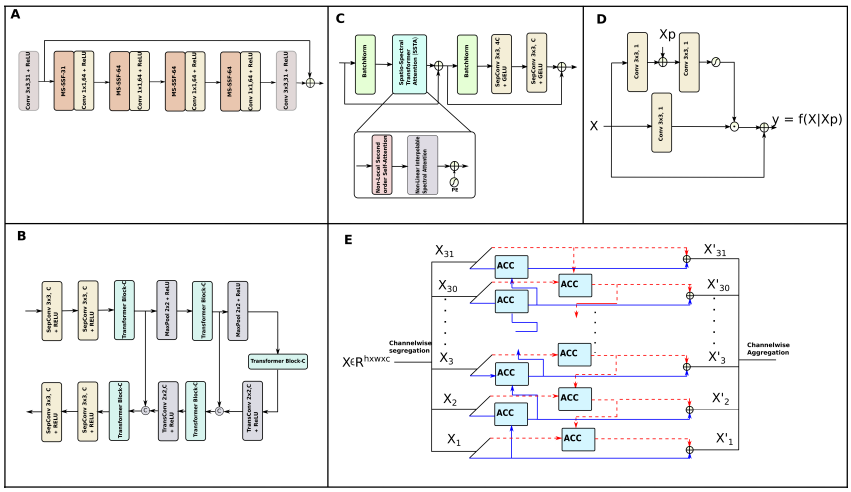
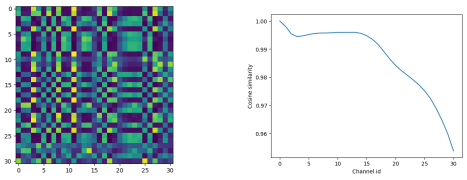


Figure 1. A: End-to-end transformer network. B: MSSSFB-C: Multi-Scale Spatio Spectral Feature Block with C number of input channels. C: Transformer block with C number of input channels. D: Adjacent Channel Coupler (ACC). E: Architecture of spectral attention.

the spatial context. Figure 1B shows the architecture of the MS-SSF block that follows U-Net Ronneberger et al. (2015) like architecture. MS-SSF block uses a separable convolution layer for feature transformation, and a transformer block to learn spatial-spectral feature dependencies. 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 Chen et al. (2021). Though BatchNorm based pure self-attention suffers from instability issues, it works reasonably well for mixed architecture. Spatio-Spectral Transformer Attention (SSTA) is the core attention module to learn inter-channel and spatial interactions using the adjacently coupled spectral self-attention and non-local second-order self-attention (Appendix A.1) and the resulting attention coefficient is governed by the spectral features in the majority.

2.1. Adjacently Coupled Spectral Attention



((a)) ((b))

Figure 2. (a) 31x31 shaped Spectral attention map of MST (Cai et al., 2022a) for 31 channels. (b) Cosine similarity of first channel to other 30 channels.

Before we discuss the adjacently coupled spectral attention, we intuitively look at the the spectral attention map of MST ((Cai et al., 2022a)) (shown in Figure 2(a)) in which every row i depicts the global dependency of channel i on other channels. Figure 2(b) shows the spectral similarity of the first channel to the others in a hyperspectral image. While the spectral dependency is obvious to decrease on the farther channels, same could not be observed in the spectral map learnt by MST (Cai et al., 2022a). This can make any given channel more susceptible to the perturbation in the farther channel, which is contradicting to the scientific correlation of the spectral channels. To mitigate the physical incoherence in learning the global spectral dependency, we came up with the idea of adjacent channel coupling that allows the spectral dependency to be modulated through propagation between the two given channels. The detailed analysis of the proposed attention is provided in the section 2.1.1.

2.1.1. Lipschitz Stability

The motivation to propose coupled spectral attention 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 feature map $X \in \mathbf{R}^{H\times W\times C}$ 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 $A_{ij} = \sum_{k=0}^{HW-1} Q_{i,k}K^T_{k,j}$. It squeezes the spatio-spectral context between two channels to a single scalar value causing the information loss. Second,the Lipschitz constant of self-attention is bounded by the variance of the input resulting in larger sensitivity Kim et al. (2021). To support the argument, Corollary 2.1 shows that the L_2 norm of diagonal elements of Jacobian is proportional to the squared dynamic range of the input.

Corollary 2.1. *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 L_2 norm of diagonal elements of Jacobian in self-attention network is given by,*

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

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

Proof. See Appendix A.2. □

To combat the sensitivity issue, we use the Ising model (Brush, 1967) as a reference taht is commonly used to describe the phenomenon of ferromagnetism. The total energy of system in Ising model is a function of spin states coupled with their nearest neighbour and an external field trying to align these spin states. In other words,

$$-E(s_i) = J \sum_{<i,j>} s_i s_j + \mu_B B \sum_i s_i \tag{2}$$

$$-E(s_i) = f_1(s_i, s_{j:j\in<i,j>}) + f_2(B, s_i) \tag{3}$$

In equation 2, J is the coupling constant, s_i is the spin variable (hidden state), $< i, j >$ indicates nearest neighbour, and B is the external magnetic field (input). Equation 3 represents the Ising model in the form of generalised functions. Following this, we formulate the adjacent coupling in spectral unit. The channel wise output can be mathematically described as,

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

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 coupling (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 nonlinear cross-attention and Value embedding, respectively. $\odot$ refers to elementwise-multiplication. For necessary derivations, we further simplify the gating vector F^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,

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

Unlike self-attention, coupled channels in equation 6 only allows information to propagate through interaction between adjacent feature maps. Additionally, we use non-linear function in the place of correlation based spin coupling to simulate the spectral interaction. Equation 6 can be rewritten as matrix operation on the vectorized mappings $X_c, Y_c \in \mathbf{R}^{HW}$ as,

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

where the convolution, being a linear operation, is replaced by the matrix-vector operation. $op(W)X$ implies that matrix form of convolution kernel W is operated on the vectorized input X.

Lemma 2.2. *Let $\omega = e^{2\pi i/HW}$ and W^f be the convolution kernel in the function f. Let J be the difference between the learned kernel and its initialization and given by $J = W^f - W_0^f$. Also, let F be a complex matrix such that $F_{ij} = \omega^{ij}$. If $\epsilon^f = \frac{1}{9}(F^T J F)_{0,0}$, then upper bound on the L_2 norm of diagonal elements of Jacobian of adjacently coupled spectral self-attention is given by,*

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

Proof. See Appendix A.3. □

Lemma 2.3. *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 $\|J_{i,j}\|_2$ after T iterations is given by,*

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

for $i > j$ and,

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

for $i < j$.

Proof. See Appendix A.4. □

Table 1. Quantitative comparison of different spectral reconstruction methods. The best ones are shown in **bold**.

Method	Params (M)	FLOPS (G)	CAVE		NTIRE2020		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
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

Lemma 2.2 shows that the upper bound on L_2 norm of the Jacobians's diagonal elements (Lipschitz constant for diagonal elements) of the model is proportional to the maximum of absolute dynamic range. Furthermore, Lemma 2.3 derives the general formulation for Lipschitz constant under mild assumptions.

3. Experiments

3.1. Datasets

Three publicly available datasets are used for training and performance assessment, including NTIRE 2020 Arad & Timofte (2020), NTIRE 2022 Arad & Timofte (2022), and CAVE Yasuma et al. (2010) 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 remaining 12 images are 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.

3.2. Performance Comparison

The proposed approach is compared with latest state-of-the-art methods, including AWAN (Li et al. (2020)), MST (Cai et al. (2022a)), HSCNN+ (Shi et al. (2018)), HRNet (Zhao et al. (2020)), and EDSR (Lim et al. (2017)) in Table 1. For NTIRE2020 and NTIRE2022, we compare them in terms of Mean Relative Absolute Error (MRAE) and Root Mean Squared Error (RMSE). Whereas, Spectral Angle Mapper (SAM) is used for CAVE dataset instead of MRAE due to high sensitivity of zero value pixels in spectral bands. 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.

Table 2. 2-Lipschitz constant for spectral self-attention from MST

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. 2-Lipschitz constant for Adjacent channel coupling.

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

3.3. Stability Analysis

To further validate the Lemmas 2.2 and 2.3, we emperically estimate the Lipschitz constants by perturbing a specific channel(denoted by j in Tables 2 and 3) in the feature map. It is to be noted that these values are estimated for RGB images with dynamic range of 255 to observe the effects more prominently. 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-multispectral self-attention and our method, 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 adjacently coupled feature maps. This clearly indicates that perturbation in the spectral channel of multihead-spectral self-attention induces instability in the other channels too, whereas this effect is suppressed to a large extent in channel coupled approach.

4. Conclusions

Though Transformer has been the emergent approach in various applications, the performance and training stability still requires to be carefully studied. Moreover, the physics based inductive bias is yet to be explored in the context of

vision based transformers. This work specifically focuses on the implications of self-attention along the spectral dimension, and therefore proposes a modified structure with theoretical Lipschitz constant to enhance the overall stability of the transformer. The derived theory is also empirically observed in the model. As a future scope, it will be worth to find out other methods to study and evaluate the stability of the machine learning models that are derived from principles of statistical physics.

References

Arad, B. and Timofte, R. Ntire 2020 challenge on spectral reconstruction from an rgb image. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1806–1822, 2020.

Arad, B. and Timofte, R. Ntire 2022 spectral recovery challenge and data set. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 862–880, 2022.

Brush, S. G. History of the Lenz-Ising Model. *Reviews of Modern Physics*, 39(4):883–893, October 1967.

Cai, Y., Lin, J., Hu, X., Wang, H., Yuan, X., Zhang, Y., Timofte, R., and Gool, L. V. Mask-guided spectral-wise transformer for efficient hyperspectral image reconstruction. 2022a.

Cai, Y., Lin, J., Lin, Z., Wang, H., Zhang, Y., Pfister, H., Timofte, R., and Gool, L. V. Mst++: Multi-stage spectral-wise transformer for efficient spectral reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 744–754, 2022b.

Chen, Z., Xie, L., Niu, J., Liu, X., Wei, L., and Tian, Q. Visformer: The vision-friendly transformer. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 569–578, 2021.

Huang, H. *Statistical Mechanics of Neural Networks*. Springer Singapore, 2023.

Kim, H., Papamakarios, G., and Mnih, A. The lipschitz constant of self-attention. In *ICLR*, 2021.

Li, J., Wu, C., Song, R., Li, Y., and Liu, F. 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)*, pp. 1894–1903, 2020.

Li, W., Lin, Z., Kun Zho and, L. Q., Wang, Y., and Jia, J. Mat: Mask-aware transformer for large hole image inpainting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Lim, B., Son, S., Kim, H., Nah, S., and Lee, K. M. Enhanced deep residual networks for single image super-resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1132–1140, 2017.

Lin, Y.-T. and Finlayson, G. D. Physically plausible spectral reconstruction from rgb images. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2257–2266, 2020.

Long, P. M. and Sedghi, H. Generalization bounds for deep convolutional neural networks. 2020.

Lu, Z., Li, J., Liu, H., Huang, C., Zhang, L., and Zeng, T. Transformer for single image super-resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022.

Raissi, M., Perdikaris, P., and Karniadakis, G. E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019.

Ronneberger, O., Fischer, P., and Brox, T. Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

Shi, Z., Chen, C., Xiong, Z., Liu, D., and Wu, F. Hscnn+: Advanced cnn-based hyperspectral recovery from rgb images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1052–10528, 2018.

Sinha, A. K., Moorthi, S. M., and Dhar, D. Nl-ffc: Non-local fast fourier convolution for image super resolution. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 466–475, 2022.

Wang, X., Girshick, R., Gupta, A., and He, K. Non-local neural networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7794–7803, 2018.

Yasuma, F., Mitsunaga, T., Iso, D., and Nayar, S. K. Generalized assorted pixel camera: Postcapture control of resolution, dynamic range, and spectrum. *IEEE Transactions on Image Processing*, 19(9):2241–2253, 2010.

Zhao, Y., Po, L.-M., Yan, Q., Liu, W., and Lin, T. Hierarchical regression network for spectral reconstruction from rgb images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1695–1704, 2020.

5. Broader impact of the work

1. **Ethical considerations:** Lipschitz stability can play a role in addressing ethical concerns related to the deployment of machine learning models. By ensuring that the model's predictions are not overly sensitive to small changes in the input, it helps reduce biases and unwarranted discrimination that may arise from minor variations in the input. This promotes fairness and accountability in the use of Transformer models.

2. **Attention mechanism:** The Ising model's concept of interactions between spins can inspire modifications to the attention mechanism, allowing it to capture more nuanced dependencies and interactions between tokens. This could potentially improve the model's ability to understand context and capture long-range dependencies.

3. **Adversarial defense:** Adversarial attacks aim to manipulate inputs to mislead machine learning models. Lipschitz stability can help in defending against such attacks by limiting the extent to which an input can be perturbed without significantly changing the model's output. By constraining the Lipschitz constant, the model becomes more resistant to adversarial manipulations, thereby enhancing its security.

4. **Interpretability and explainability:** Lipschitz stability can contribute to the interpretability and explainability of transformer models. When a model's behavior is Lipschitz stable, it implies that small changes in the input space correspond to small changes in the output space. This property enables researchers and practitioners to analyze the model's behavior more easily, interpret the effects of input features on the output, and provide more reliable explanations for model predictions.

A. Appendix

A.1. Non-local second-order self-attention

Unlike adjacently coupled spectral attention, non-local second-order self-attention (Wang et al., 2018) is used to learn global inter-dependencies for every spatial position. The general formulation of self-attention is

$$Y = softmax(X)V, \tag{8}$$

where $V = W^V X$ is squeezed to the dimension $HW \times C/2$ and X is the spatial attention-map given by,

$$X = WX\hat{I}(WX)^T, \tag{9}$$

where $\hat{I} = \frac{1}{HW}(I - \frac{1}{HW}1)$, and I and 1 are identity matrix and ones matrix, respectively. Due to quadratic complexity, the model is trained and inferred in patches to reduce memory footprint.

A.2. Proof of Corollary 2.1

From (Kim et al., 2021),

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

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 $i = j$,

$$J_{ii} = W^K W^Q X^T P^{(i)} e_{ii} X + W^K W^Q Var(X) + P_{ii} \tag{10}$$

$$\|J_{ii}\|_2 \leq A.(P_{i,i}X_i - (P_{i,i}X_i)^2) + A.Var(X) + \|P_{ii}\|_2, \tag{11}$$

where $A = \left\|W^K W^Q\right\|_2$
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 $\|J_{i,i}\|_2$,
$\|J_{ii}\|_2 \leq \frac{A}{4} + A.Var(X) + 1$

A.3. Proof of Lemma 2.2

The proof is the immediate application of operator norm for convolution kernels in Long et. al. (Long & Sedghi, 2020).

In equation 6, the functions f, g and h, being 2D convolutions, can be represented using the matrix multiplication with corresponding operator matrix (Long & Sedghi, 2020), i.e. $W * x \stackrel{\text{def}}{=} op(W)\mathbf{x}$.

$$\begin{aligned} J_{i,i} &= \frac{\partial Y_i}{\partial X_i} \\ &= \mathbb{I} + \frac{1}{2}\Big(diag(op(W_{f_i^F})X_i \odot \sigma(\alpha_1)(1-\sigma(\alpha_1)))op(W_{g_i^F}) \\ &+ diag(op(W_{f_i^B})X_i \odot \sigma(\alpha_2)(1-\sigma(\alpha_2)))op(W_{g_i^B})\Big) + \frac{1}{2}\Big((op(W_{f_i^F})) \odot diag(\sigma(\alpha_1))\Big) + \\ &\frac{1}{2}\Big((op(W_{f_i^B})) \odot diag(\sigma(\alpha_2))\Big) \end{aligned} \tag{12}$$

Here, $\alpha_1 = g^F(X_i) + h^F(Y_{i-1})$ and $\alpha_2 = g^B(X_i) + h^B(Y_{i+1})$. Applying the operator norm from Long et. al. (Long & Sedghi, 2020) and taking the L_2 norm to estimate the Euclidean Lipschitz constant,

$$\|J_{i,i}\|_2 \leq 1 + \frac{1}{8}\big((1+9\epsilon_i^{f^F})(1+9\epsilon_i^{g^F}) + (1+9\epsilon_i^{f^B})(1+9\epsilon_i^{g^B})\big)\sqrt{HW}max(|M|,|m|) + \frac{1}{2}\big((1+9\epsilon_i^{f^F}) + (1+9\epsilon_i^{f^B})\big) \tag{13}$$

A.4. Proof of Lemma 2.3

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,

$$\frac{\gamma}{2}\|w_T\|^2 \leq \rho_T \tag{14}$$

If the maximum L_2 norm of a kernel is C, then $\|w_T\|^2 \leq NC^2$. Substituting it in equation 14, we get,

$$C \leq \sqrt{\frac{2\rho_T}{\gamma N}} \tag{15}$$

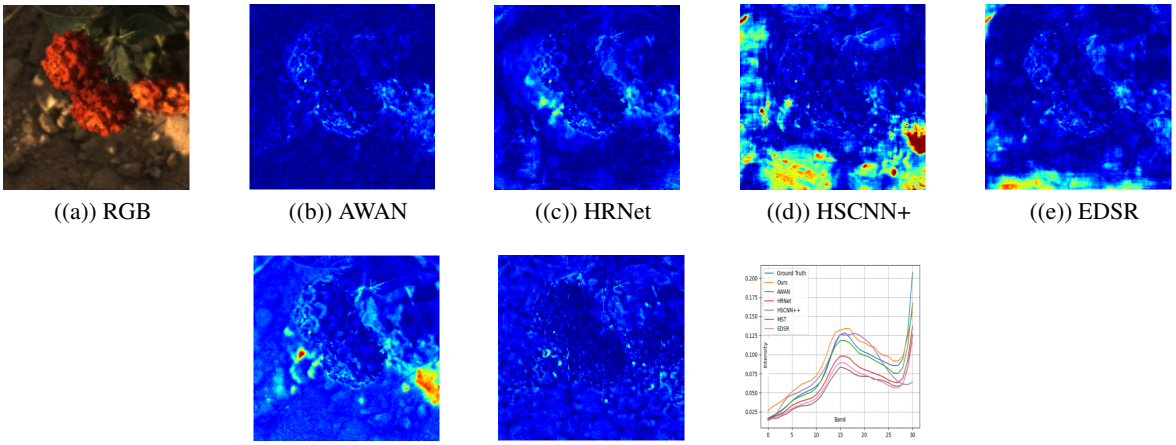
$$\begin{aligned} J_{i,j} &= \frac{\partial Y_i}{\partial X_j} \\ &= \frac{1}{2} diag(op(W_{f_i^F})X_i \odot \sigma(\alpha_1)(1-\sigma(\alpha_1)))op(W_{h_i^F})\frac{\partial Y_{i-1}}{\partial X_j} \\ &\leq \frac{1}{2} diag(V_i^F \odot \sigma(F_i^F(Q_i, K_{i-1})))op(W_{h_i^F})\frac{\partial Y_{i-1}}{\partial X_j} \\ &\leq \Big(\prod_{k=j+1}^{i} \frac{1}{2} diag(V_k^F \odot \sigma(F_k^F(Q_k, K_{k-1})))op(W_{h_k^F}) \Big) \frac{\partial Y_j}{\partial X_j} \end{aligned} \tag{16}$$

$$\|J_{i,j}\|_2 \leq \Big(\prod_{k=j+1}^{i} \frac{1}{2} \left\|V_k^F \odot \sigma(F_k^F(Q_k, K_{k-1}))\right\|_2 \left\|op(W_{h_k^F})\right\|_2 \Big) \left\|\frac{\partial Y_j}{\partial X_j}\right\|_2 \tag{17}$$

Substituting (15) in (17), we get

$$\|J_{i,j}\|_2 \leq \Big(\prod_{k=j+1}^{i} \sqrt{\frac{2\rho_T}{\gamma N}} \left\|V_k^F \odot \sigma(F_k^F(Q_k, K_{k-1}))\right\|_2 \Big) \left\|\frac{\partial Y_j}{\partial X_j}\right\|_2 \tag{18}$$

The proof of $\|J_{i,j}\|_2$, $i < j$ follows the same approach and has similar upper bound as in (18).



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

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

Figure 3. Illustration of residual map in the spectral band predicted by different methods. Spectral profile compares the spectral profiles generated by different methods.

A.5. Implementation Details

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 Cai et al. (2022b), 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, the computation requires 1.58 seconds per image on testing dataset using single A100 GPU.

A.6. Qualitative Results

Figure 3 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 multi-spectral bands.