IMPROVING THE LIPSCHITZ STABILITY IN SPECTRAL TRANSFORMER THROUGH NEAREST NEIGHBOUR COUPLING

Anonymous authors
Paper under double-blind review

ABSTRACT

Ising model has already played a pivotal role in the formulation of neural networks like Hopfield networks and Restricted Boltzmann machines. Recently, Transformers have recently gained popularity due to their ability to learn long range dependencies through self-attention. 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 to learn 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

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 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

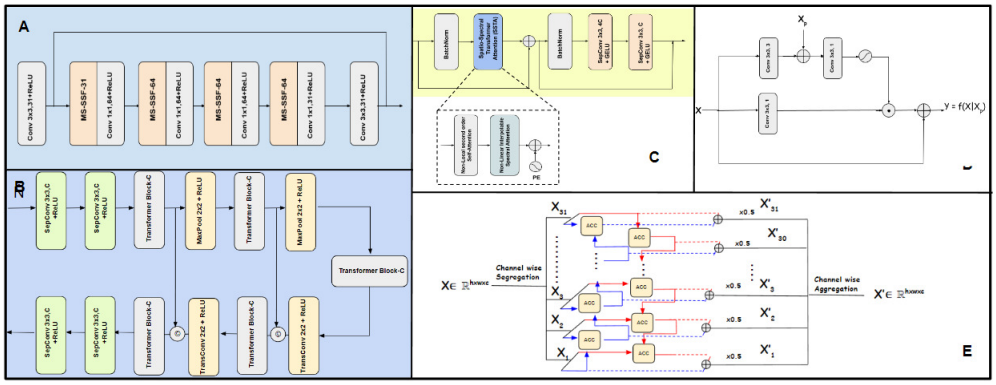


Figure 1: A: End-to-end transformer network. B: MSSSFB-C: Multi-Scale Spatio Spectral Feature Block with C number of input channels. C: Transformer block with C number of input channels. D: Adjacent Channel Coupler (ACC). E: Architecture of spectral attention.

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 and the resulting attention coefficient is governed by the spectral features in the majority.

2.1 Lipschitz Stability of Adjacently Coupled Spectral Attention

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, Lemma 1 shows that the L_2 norm of diagonal elements of Jacobian is proportional to the squared dynamic range of the input.

Lemma 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.1.

To combat the sensitivity issue, we use the Ising model Brush (1967) as a reference. 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,

$$Y_c = X_c + \frac{1}{2}\Big(f_c^F(X_c) \odot (\sigma(g_c^F(X_c) + h_c^F(Y_{c-1}))) + f_c^B(X_c) \odot (\sigma(g_c^B(X_c) + h_c^B(Y_{c+1})))\Big), \tag{4}$$

wher c is the channel, f^F, g^F, h^F and f^B, g^B, h^B are the convolution functions in forward and backward regressors respectively, and $\sigma(.)$ is the sigmoid function. As shown in equation 4, each feature map is estimated as the average of forward and backward regressors to learn the residual coupled spectral dependencies. At the same time, the long-range dependency is exploited by using the updated adjacent feature maps in the attention map. Since there is no hidden layer, the output acts as the hidden unit in the model. It is to be noted that unlike correlation based spin coupling, we use non-linear interpolation approach to explicitly model the smooth spectral profile. Equation 4 can be rewritten as matrix operation on the vectorized mappings $X_c, Y_c \in \mathbf{R}^{HW}$ as,

$$\begin{aligned} Y_c =& X_c + \frac{1}{2}\Big(op(W_{f_c^F})X_c \odot (\sigma(op(W_{g_c^F})X_c + op(W_{h_c^F})Y_{c-1})) \\ &+ op(W_{f_c^B})X_c \odot (\sigma(op(W_{g_c^B})X_c + op(W_{h_c^B})Y_{c+1}))\Big) \end{aligned} \tag{5}$$

Lemma 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.2.

Lemma 3. *Let the loss function and learning rate for transformer network be $\mathcal{L} = \rho + \frac{\gamma}{2}\|w\|_2^2$ and η respectively, where ρ is the data fidelity term and γ is regularisation parameter. Let k be the kernel size of the filters. If η and γ are chosen such that after T iterations $max\left\|\frac{\partial\rho}{\partial w}\right\|_2 \leq e^{-\eta\gamma T}\frac{k}{\eta T}$, then the upper bound on the L_2 norm of Jacobian $\|J_{i,j}\|_2$ after T iterations is given by,*

$$\|J_{i,j}\|_2 \leq e^{-\|i-j\|\eta\gamma T}\prod_{k=i}^{j} k\,\|Y_k\|_2 \left\|\frac{\partial Y_j}{\partial X_j}\right\|_2$$

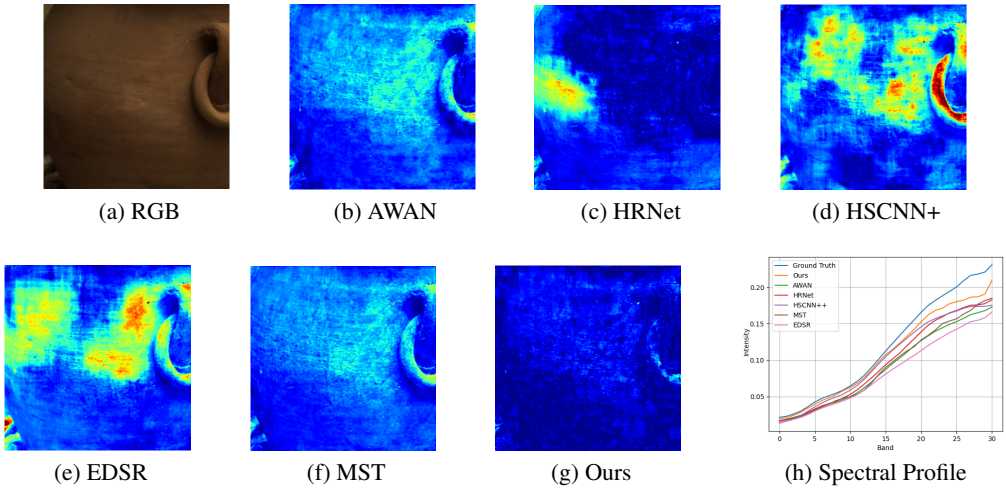
Proof: See Appendix A.3.

Lemma 2 shows that the 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 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.



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

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

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

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). 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. Figure 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.

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	MRAW	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

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. 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

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)*, pp. 1806–1822, 2020.

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)*, pp. 862–880, 2022.

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

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. In *CVPR*, 2022a.

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*, pp. 744–754, 2022b.

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)*, pp. 569–578, 2021.

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

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

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)*, pp. 1894–1903, 2020.

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.

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)*, pp. 1132–1140, 2017.

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)*, pp. 2257–2266, 2020.

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

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.

Maziar Raissi, Paris Perdikaris, and George E Karniadakis. 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.

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

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)*, pp. 1052–10528, 2018.

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)*, pp. 466–475, 2022.

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.

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)*, pp. 1695–1704, 2020.

A Appendix

A.1 Proof of Lemma 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 $i = j$,

$$J_{ii} = W^K W^Q X^T P^{(i)} e_{ii} X + W^K W^Q Var(X) + P_{ii} \tag{6}$$

$$\|J_{ii}\|_2 \leq A.(P_{i,i}X_i - (P_{i,i}X_i)^2) + A.Var(X) + \|P_{ii}\|_2 , \tag{7}$$

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.2 Proof of Lemma 2

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

In equation 4, 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. $W * x \stackrel{\text{def}}{=} op(W)\mathbf{x}$.

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

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. Long & Sedghi (2020) and taking the L_2 norm to estimate the Euclidean Lipschitz

constant,

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

A.3 Proof of Lemma 3

It can be easily shown that weights are being updated by the following equation,

$$W_t = (1-\eta\gamma)W_{t-1} - \eta\frac{\partial\rho}{\partial W} \tag{10}$$

Consequently, after T iterations,

$$W_T = (1-\eta\gamma)^T W_0 - \eta\Big(\sum_{k=0}^{T-1}(1-\eta\gamma)^k\frac{\partial\rho}{\partial W_k}\Big) \tag{11}$$

Without the loss of generality, weights for the 2D kernel of size $k \times k$ is initialized using Xavier-Glorot initialization,

$$W_0 \sim \mathcal{U}\Big(-\frac{1}{k},\frac{1}{k}\Big) \tag{12}$$

$$\|W_T\|_2 \leq \big\|(1-\eta\gamma)^T W_0\big\|_2 + \left\|\eta\Big(\sum_{k=0}^{T-1}(1-\eta\gamma)^k\frac{\partial\rho}{\partial W_k}\Big)\right\|_2 \tag{13}$$

$$\|W_T\|_2 \leq (1-\eta\gamma)^T k + \eta\sum_{k=0}^{T-1}\left\|\frac{\partial\rho}{\partial W_k}\right\|_2 \tag{14}$$

$$\|W_T\|_2 \leq (1-\eta\gamma)^T k + \eta T max\left\|\frac{\partial\rho}{\partial W}\right\|_2 \tag{15}$$

$$\|W_T\|_2 \leq e^{-\eta\gamma T}k + \eta T max\left\|\frac{\partial\rho}{\partial W}\right\|_2 \leq 2e^{-\eta\gamma T}k \tag{16}$$

In (13), $W_0 \in \mathbf{R}^{k\times k}$ and $max\,\|W_0\|_2 = k^2 max(|W_0|) = k^2.1/k = k$, since W_0 is sampled from the uniform distribution having lower and upper limits set to $-1/k$ and $1/k$ respectively.

Without the loss of generality, we show the proof for $J_{i,j}$, $i > j$,

$$\begin{aligned}J_{i,j} &= \frac{\partial Y_i}{\partial X_j} \\ &= \frac{1}{2}diag(op(W_{f_i^F})X_i \odot \sigma(\alpha_1)(1-\sigma(\alpha_1)))op(W_{h_i^F})\frac{\partial Y_{i-1}}{\partial X_j} \\ &\leq \frac{1}{2}diag(Y_i)op(W_{h_i^F})\frac{\partial Y_{i-1}}{\partial X_j} \\ &\leq \prod_{k=j+1}^{i}\frac{1}{2}diag(Y_k)op(W_{h_k^F})\frac{\partial Y_j}{\partial X_j}\end{aligned} \tag{17}$$

$$\|J_{i,j}\|_2 \leq \prod_{k=j+1}^{i}\frac{1}{2}\|Y_k\|_2\left\|op(W_{h_k^F})\right\|_2\left\|\frac{\partial Y_j}{\partial X_j}\right\|_2 \tag{18}$$

Substituting (16) in (18), we get

$$\|J_{i,j}\|_2 \leq e^{-|i-j|\eta\gamma T} \prod_{k=i}^{j} k \, \|Y_k\|_2 \, \|J_{j,j}\|_2 \tag{19}$$

The proof of $\|J_{i,j}\|_2$, $i < j$ follows the same approach and has similar upper bound as in (19).

B 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.