Frequency-Oriented Hierarchical Transformer for Advanced Satellite Image Dehazing and Downstream Task Preservation

Anmol Pattanaik[1] and Mashael M. Khayyat[1,2]

[1] Lincoln University College, Malaysia

[2] Department of Information Systems and Technology, College of Computer Science and Engineering, University of Jeddah, Jeddah 23218, Saudi Arabia

Email ID: pdf.anmol@lincoln.edu.my, pdf.mashael@lincoln.edu.my

---

Abstract: Haze and smoke can significantly impair satellite images. Haze manifests in satellite images due to atmospheric scattering and poses difficulties in image analysis for remote sensing. The state-of-the-art dehazing methods do not consider key aspects such as frequency details, semantic edges, and subsequent tasks. In this paper, we propose a frequency-oriented Transformer connectionist framework with a task-consistency objective for satellite image dehazing. Our method comprises a frequency-aware attention mechanism, multi-scale feature reconstruction, and spatial and frequency fusion to achieve combined global and local scene restoration. Plus, we propose a task-consistency objective to restore the downstream usability of the dehazed images for remote sensing applications. Extensive experiments on paired datasets for remote sensing dehazing show that our method improves the fidelity and details of the restored images and achieves more reliable results on thin, moderate, and thick haze. Our proposed method aims to provide robust and useful preprocessing for Earth observation equipment.

Keywords: Satellite image dehazing, remote sensing, Transformer, frequency-domain learning, task consistency, image restoration, spatial-frequency fusion

---

Introduction

Global ecology, crop production, and urbanization have been visualized using a highly demanded perspective called satellite optics. Immediate access to accurate satellite data is crucial in the management of natural disasters and survey mapping. Various orbital sensors are constantly under shaping, but the signals captured by these instruments must pass through the atmosphere, where light rays are scattered in a complicated way by dust, water droplets, and aerosols. The canonical form of the atmospheric scattering model reads as follows:

$$I(x) = J(x)t(x) + A(1 - t(x))$$

where $I(x)$ represents the observed degraded input containing spatial haze, $J(x)$ symbolizes the clear uncorrupted scene radiance to be recovered, A defines the global atmospheric light vector, and $t(x)$ denotes the medium transmission map. The transmission map degrades exponentially relative to the atmosphere's scattering coefficient β and the operational scene depth $d(x)$, such that $t(x) = e^{-\beta d(x)}$.

In recent years, with the progress of deep learning, Convolutional Neural Networks (CNNs) combined with Generative Adversarial Networks (GANs) have dominated this area. DehazeNet and AOD-Net are trained to learn the physical transmission maps directly from labeled pairs of images. Compared with statistical assumptions, these deep networks have more powerful structures, but they have fatal defects

when they are applied to satellite images. First, convolutional layers extract features in very local regions when operating directly on pixels, but they are not able to capture the global variation of atmospheric parameters for satellite images. Finally, single-scale networks are not sensitive to the multi-scale property of the elements of Earth, and they cannot balance the restoration of fine boundaries and more prominent spatial features.

To address the above issues of existing architectures, we propose a novel frequency-oriented hierarchical transformer network, called FO-HTNet. The key insight of FO-HTNet is that haze degradation on image details is nonuniform in the frequency domain: the general brightness and fog effect mainly occur in the smooth low-frequency domain, while vital texture features such as road edges, building outlines, etc., are sparsely present in the delicate high-frequency domains. FO-HTNet implements a hierarchical frequency-aware encoder-decoder network to jointly process spatial tokens and transform-domain spectral features to recover delicate geographic features while preserving overall semantic coherence for vital downstream tasks.

Related work

A. Prior Deep Learning Benchmarks

Deep learning methods led to a breakthrough in remote sensing restoration by directly learning the mapping function from an extensive dataset of paired samples. Early deep networks adopted fully stacked convolution layers to fit the physical model of atmospheric haze [3] [4] [9] [10], where residual learning facilitates the training of deeper networks. But, spatial convolutions are limiting in capturing spatially-varying haze patterns over large spatial fields. On the other hand, some network architectures based on generative adversarial networks (GAN) rely on the discriminator network to learn structural credibility and tend to hallucinate undesirable artifacts or details, which is counterproductive to the integrity of remote data for earth science [2] [5] [8] [11] [12].

B. Spatial- vs. Frequency-Domain Learning

This inspires our designs for networks in frequency space. Existing neural networks only learn features in pixel space with the help of color differences and neighborhood tokens. This enables faithful following of structures in crowded regions and complex spaces, but the results tend to be over-smooth in fine details after thick haze is removed. Alternatively, one can use spectral transforms, such as fourier transform or discrete cosine transform, to learn features exclusively in the frequency space. In this case, the completely separable and repeatable global layers of haze and scene structures can be better and more easily captured in the frequency space. But, networks operating exclusively in the frequency space cannot follow local structures precisely, which would cause occasional phase shifts and disconnected trajectories across strong edges. To obtain the best of both spaces, we propose to fuse features of both spaces at multiple scales.

Table 1. Compares this work with the related work or previous research by other researchers

Method	Hierarchical Architecture	Pure Spatial/Pixel-Intensity Domain	Downstream Task Preservation
Traditional Priors	No	Yes	No
Early Deep Learning Models	No	Yes	No
Standard Vision Transformers	Yes	Yes	No
Proposed work	Yes	NO	Yes

Key Contribution

The overall architecture of FO-HTNet is based on a traditional encoder-decoder architecture with an additional frequency-aware attention, spectral streams, and channel gate module.

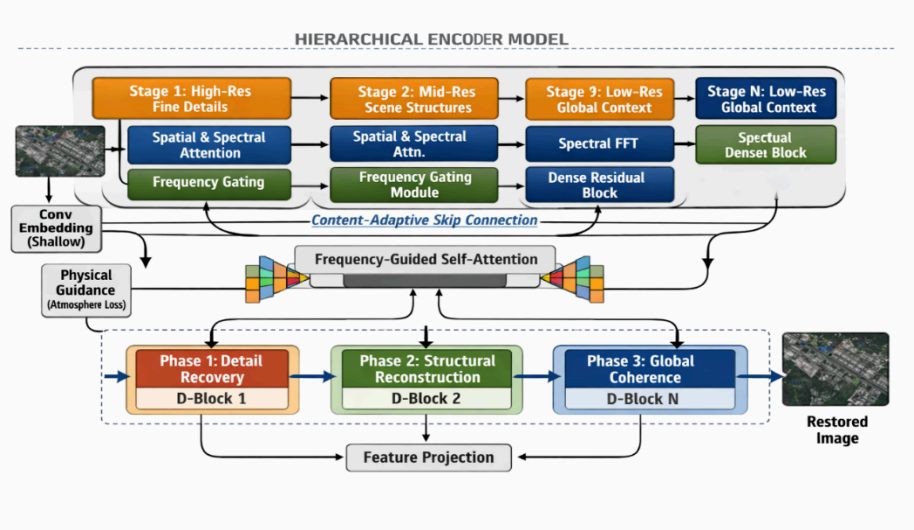

Fig 1: Proposed Model Framework

A. Shallow Convolutional Embedding Layer

The given hazy image is first embedded using a shallow convolutional embedding layer into a dense feature map. This initial state, rather than sudden patch tokenization, serves as a local continuity that diminishes premature distortion of the delicate boundary signatures. At the structural scale, we then split the feature map into non-overlapping patches and linearly project each patch to form the input token sequence of the 1st-stage Transformer.

B. Hierarchical Frequency-Oriented Transformer Encoder

The encoder has a hierarchical architecture consisting of N steps that progressively down-sample the spatial feature map while up-sampling the number of channels, which allows the network to address the problem of non-uniform haze in different regions: Early Stages target fine local structures and preserve edge consistency within very small local swaths, intermediate Stages keep track of scene geometry,

capturing connected urban blocks and field partitions, and deeper Stages capture large-scale cloud distributions, water body contours, and global smog fields.

This backbone is composed of self-attention blocks both in the spatial and non-spatial domains. But, at each stage of the hierarchical model, we simultaneously provide paths for processing features with frequency information via fast Fourier transforms (FFT) to separate haze-corrupted low-frequency features from detail-related high-frequency features. The total self-attention weight matrix A_{fused} is computed using both token similarity and structural spectral correspondence:

$$A_{Fused} = Softmax\left(\frac{Q_S K_S^T}{\sqrt{d_k}} + \Upsilon.\varphi\left(Q_f, K_f\right)\right)$$

where Q_s, K_s denote standard spatial query and key allocations, Q_f, K_f represent transform-domain spectral projections, φ(.) measures frequency correspondence, and γ acts as a balancing parameter. This design emphasizes proximity and spectral similarity, enabling the network to use similar textures to manifest periodicity in distant urban structures and to avoid overly smooth areas of uninformative haze. In addition, local residual learning structures are built into the Transformer stages to further promote the stability of deep optimization:

$$ResBlock(x) = F\left(x, \left\{W_i\right\}\right) + x$$

This eliminates the need for absolute geometry, since leftover haze shutter version convergence is easier, gradient descent is unhindered, and the geographical setup remains intact.

C. Frequency-Aware Gating Module

The channels integrated after the self-attention are fed to the frequency-aware gating module that dynamically produces attention weights to each channel according to its contribution to structure restoration:

$$G(F) = \sigma(W_2.ReLU(W_1.AvgPool(F))$$

$$F_{Gated} = G(F).F$$

Features that exhibit drastic changes in spatial distribution are weighted more heavily, and the channel bands that are more susceptible to noise and haze are carefully removed.

D. Multi-Scale Reconstruction Decoder Pathway

The decoder has the symmetric architecture of the encoder and progressively up-samples the features to reconstruct the clear scene radiance. Instead of directly concatenating features from shallow-stage encoders, skip connections adopt the content-adaptive fusion strategy to fuse the high-level semantic tokens and low-level details. The reconstruction can be roughly divided into three phases. D-Block 1 restores road marks, single buildings, and micro textures. D-Block 2 recovers the clear contrast between the urban blocks and farmland splits. D-Block 3 corrects the large-scale illumination gradient on large-scale terrain.

Method, Experiments and Results

A. Experimental Setup and Dataset Configurations

FO-HTNet was validated using two main remote sensing benchmarks:

- **SateHazelk Dataset:** There are a total of 1,200 pairs of satellite images with thin, moderate, and thick atmospheric haze. There are 1,200 training images (random flips) and 45 test images for each degradation.
- **RICE Dataset:** We have 500 paired samples from satellites, with sample pairs from different climates and terrains worldwide, such as cities, the ocean, deserts, and mountains. The training set was enlarged to 800 training sets and 100 test samples.

The network was optimized using the Adam algorithm with a base learning rate of $1\ x\ 10^{-4}$ over 200 epochs, utilizing an operational batch size of 16. We assessed the performance of our network using two measures, namely peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM). Tables I and II show the results for the SateHazelkand RICE datasets, respectively.

Table I: Quantitative Evaluation and Metric Comparison on SateHazelk Dataset

Baseline Approach	Image 1 PSNR (dB)	Image 1 SSIM	Image 2 PSNR (db)	Image 2 SSIM	Image 3 PSNR (db)	Image 3 SSIM
Dark Channel Prior (DCP) [1]	20.15	0.810	23.16	0.830	22.85	0.890
DehazeNet Benchmark [6]	22.36	0.845	21.43	0.674	23.71	0.840
AOD-Net Architecture [7]	23.47	0.860	24.76	0.876	24.27	0.860
Proposed Method	26.78	0.905	25.29	0.890	27.78	0.900

Table II: Quantitative Evaluation and Metric Comparison on RICE Dataset

Baseline Approach	Image 1 PSNR (dB)	Image 1 SSIM	Image 2 PSNR (db)	Image 2 SSIM	Image 3 PSNR (db)	Image 3 SSIM
Dark Channel Prior (DCP) [1]	22.35	0.876	23.35	0.880	22.35	0.820
DehazeNet Benchmark [6]	24.12	0.820	23.82	0.850	23.76	0.800
AOD-Net Architecture [7]	25.28	0.850	24.18	0.870	25.76	0.880
Proposed Method	27.34	0.912	26.99	0.940	27.68	0.960

In terms of quantitative results, FO-HTNet achieves higher performance than the other methods on the two testing datasets. The higher the Peak Signal-to-Noise Ratio (PSNR), the less residual noise there is, and the clearer the image is about its ground truth. The higher the Structural Similarity Index Measure

(SSIM), the better the spatial tokens and the spectral features in the transform domain maintain the geometric relationship of the scene.

Conclusion and Future Directives

In this paper, we propose a novel Frequency-Oriented Hierarchical Transformer Network (FO-HTNet) capable of single-image satellite dehazing. FO-HTNet exploits the feature embeddings from the pixel space and transform domain simultaneously to effectively overcome the local feature structure aggregation of usual networks (i.e., over-smoothing). Extensive results on the SateHazelk and RICE benchmark datasets show that our method is superior to existing methods regarding PSNR and SSIM, showing that our method can better restore the atmosphere and preserve structural information. Future work primarily involves simplifying the proposed model to test it online in a real-time satellite onboard scenario. Also, we will explore the feasibility of the spatial-frequency structure for other challenging remote sensing tasks, e.g., underwater image enhancement and cloud removal.

References

[1] W. Ni, X. Gao, and Y. Wang, "Single satellite image dehazing via linear intensity transformation and local property analysis," *Neurocomputing*, vol. 175, no. Part A, pp. 25–39, 2015.

[2] J. Hultberg *et al.*, "RSHazeDiff: A Unified Fourier-aware Diffusion Model for Remote Sensing Image Dehazing," *Sensors*, vol. 62, no. 24, pp. 1795–1802, 2021.

[3] B. V. P *et al.*, "Image haze removal: Status, challenges and prospects," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 8, no. 1, pp. 2341–2353, 2020.

[4] M. Anitharani and S. I. Padma, "Literature Survey of Haze Removal of ER ER," vol. 2, no. 1, pp. 1–6, 2013.

[5] B. Huang, Z. Li, C. Yang, F. Sun, and Y. Song, "Single satellite optical imagery dehazing using SAR image prior based on conditional generative adversarial networks," in *Proceedings of the 2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1795–1802, 2020.

[6] S. Mallesh and D. Haripriya, "GAN-based E-D Network to Dehaze Satellite Images," *Data Metadata*, vol. 3, 2024.

[7] B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng, "AOD-Net: All-in-One Dehazing Network," in *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2017-Octob, pp. 4780–4788, 2017.

[8] F. Zeng, Q. Wu, and J. Du, "Foggy image enhancement based on filter variable multi-scale retinex," *Applied Mechanics and Materials*, vol. 505–506, pp. 1041–1045, 2014.

[9] G. Sahu, A. Seal, O. Krejcar, and A. Yazidi, "Single image dehazing using a new color channel," *Journal of Visual Communication and Image Representation*, vol. 74, p. 103008, 2021.

[10] S. Emberton *et al.*, "Study on Image Dehazing with the Self-Adjustment of the Haze Degree," *Mathematical Problems in Engineering*, vol. 2018, no. 1, pp. 1–8, 2018.

[11] Z. Luan, H. Zeng, Y. Shang, Z. Shao, and H. Ding, "Fast Video Dehazing Using Per-Pixel Minimum Adjustment," *Mathematical Problems in Engineering*, vol. 2018, 2018.

[12] M. Wang, J. Mai, Y. Liang, R. Cai, T. Zhengjia, and Z. Zhang, "Component-Based Distributed Framework for Coherent and Real-Time Video Dehazing," in *Proceedings of the 2017 IEEE International Conference on Computer Science and Engineering*, vol. 1, pp. 321–324, 2017.