Hyperspectral Image Super-Resolution with Spatiospectral Attention in the Fourier Domain

Peter Phan

Abstract—Hyperspectral imaging provides greater spectral resolution over traditional multispectral imaging, and is useful in a wide range of tasks requiring the analysis of spectral signatures. However, physical limitations force a trade-off in lower spatial resolution for greater spectral resolution. Various deep learning-based models have been explored for the hyperspectral image super-resolution fusion task, whereby a low-resolution hyperspectral image is fused with a high-resolution multispectral image to produce a high-resolution hyperspectral image. Such models have solely focused on the spatial domain representation, neglecting intrinsic details present in frequency domain representations. In this paper, we build upon one of these models by incorporating both spatial and frequency domain representations. Our methods demonstrate greater performance in hyperspectral image reconstruction as a whole as well as high-frequency preservation and individual spectral preservation.

Index Terms—Computational Photography

1 Introduction

Hyperspectral imaging captures information across many more bands than multispectral imaging, providing greater spectral fidelity. This capability is valuable for a wide range of applications that depend on spectral analysis, such as material identification and object classification. However, the fundamental trade-off between spatial and spectral resolution remains a key limitation in HSI acquisition: sensors capable of capturing many spectral bands often yield imagery with relatively coarse spatial detail, while those offering high spatial resolution do so at the expense of spectral richness [1]. Obtaining high-resolution hyperspectral data directly from sensors is thus challenging and often impractical.

A promising way to overcome this limitation is to fuse a low-resolution HSI (LR-HSI) with a higher-resolution MSI (HR-MSI) to produce a high-resolution HSI (HR-HSI). Over the past decade, numerous techniques—ranging from traditional model-based approaches [2] to more recent deep learning methods [3] [4]—have aimed to solve this super-resolution problem. Deep learning approaches have attracted considerable attention due to their ability to automatically learn effective representations from the data itself. Networks like HSRnet have shown that it is possible to preserve spectral information while recovering fine spatial details, and the convolutional architecture of HSRnet allows it generalize across different data sets and sensor modalities. However, existing networks often focus solely on spatial domain representations, overlooking the wealth of information that can be revealed when considering the frequency domain.

To further advance the frontier of HSI super-resolution, we propose an enhanced version of HSRnet that incorporates both spatial and frequency domain information. Our method introduces a novel loss term designed to penalize reconstruction errors in the Fourier domain, thereby improving the frequency fidelity of the reconstructed HR-HSI. Furthermore, we modify the HSRnet architecture to operate in both the spatial and Fourier domains. By applying the network to the frequency-transformed version of the input images and then combining its output with the spatial-domain reconstruction through a weighted fusion, the network can better leverage complementary information from both representations. This dual-domain approach leads to improved recovery of fine spatial details and more consistent spectral signatures, while maintaining a relatively simple network structure and low computational overhead.

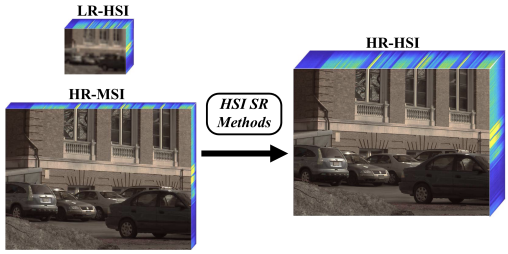

Fig. 1. Illustration of HSI super-resolution from an LR-HSI and an HR-MSI [5]

In summary, the main contributions of this work are as follows:

Dual-Domain Fusion: We extend the HSRnet framework to incorporate both spatial and frequency domain representations. This dual-domain fusion enables the network to leverage distinct and complementary features, improving spatial detail recovery and spectral integrity simultaneously.

Frequency-Aware Loss: We introduce a novel loss term defined in the Fourier domain, which penalizes frequency-domain errors. This encourages the model to maintain high-frequency components accurately, leading to sharper and more detailed reconstructions.

Enhanced Generalization and Efficiency: By combining

• *P. Phan is with the Department of Mechanical and Industrial Engineering, University of Toronto.*
E-mail: tammy.phan@utoronto.ca

spatial and frequency domain information, our proposed method increases reliance on inherent representation-based features over data-specific features.

The rest of this article is organized as follows. Section II reviews recent developments in HSI super-resolution, focusing on both model-based and deep learning-based approaches. Section III introduces the proposed dual-domain HSRnet architecture, detailing the network design and the frequency-aware loss function. In Section IV, we present extensive experimental results and discuss the effectiveness, robustness, and efficiency of our approach compared to existing methods. Finally, Section V concludes the article and outlines potential directions for future research.

2 Related Work

Existing methods for HSI–MSI fusion generally fall into two broad categories: model-based, and deep learning-based methods.

2.0.1 Model-based Methods

Traditional model-based methods formulate the fusion problem as an optimization task with prior constraints. For example, Li *et al.* [6] and Xu *et al.* [7] use the Tucker tensor decomposition to decompose the HR-HSI into compact core tensor and factor matrices, embedding sparsity or smoothness constraints to guide the reconstruction process. Classical optimization algorithms (e.g., ADMM) can then be applied to find a solution. While some of these frameworks demonstrate sound theoretical underpinnings and good interpretability, their performance can be sensitive to the choice of priors and tuning parameters. In addition, different scenes or sensors may require different parameter settings, limiting the general applicability of these methods.

2.0.2 Deep Learning Methods

Deep learning-based methods have emerged as a powerful alternative to model-based methods. The problem is formulated as a non-linear mapping that takes a LR-HSI and a corresponding HR-MSI to a HR-HSI, and deep neural networks have proven particularly effective at learning such mappings.

Early neural approaches adopted architectures inspired by image super-resolution networks designed for RGB images. Subsequently, more specialized designs have emerged, including attention modules and spectral–spatial feature fusion layers. Xie *et al.* [3] constructed a fusion model which merges the generalization models of low-resolution images and the low-rankness prior knowledge of HR-HSI images and then designed the deep network by unfolding the proximal gradient algorithm. Zhang et al. [8] designed an interpretable spatial-spectral reconstruction network (SSR-Net). FusionNet [9] approched the fusion problem using a variational probabilistic autoencoder. In this article, we focus on HSRnet, introduced by Hu *et al.* [4] HSRnet is a deep CNN incorporating attention modules to learn multi-scale spatial and spectral information.

A noteworthy trend in recent work is the integration of model-derived insights into the network design. For example, certain methods "unfold" iterative algorithms into deep architectures, ensuring that the network respects the generative models of the LR-HSI and MSI [3]. Others incorporate explicit spectral response functions, low-rank priors, or graph-based constraints directly into their layers [10]. While these data-driven strategies often yield impressive performance, they can still suffer from limitations related to training complexity, data dependence, and generalization to unseen domains.

2.0.3 Beyond the Spatial Domain

Existing HSI super-resolution methods operate primarily in the spatial domain, focusing on local patches, residual images, or semantic features learned directly from pixel neighborhoods. Comparatively less attention has been devoted to exploring the frequency domain, where global structures and periodic patterns can be more naturally captured. Some recent studies on general image restoration have highlighted the benefits of frequency-based losses or transformations [11], [12]. Such perspectives are only beginning to influence image super-resolution, suggesting that further integration of spatial and frequency representations could enhance the reconstruction quality, reduce artifacts, and improve the model's robustness.

3 Proposed Method

3.1 Overview of HSRnet

We first provide a brief overview of HSRnet. The core idea of HSRnet is to take a low-resolution hyperspectral image (LR-HSI) and combine it with a high-resolution multispectral image (HR-MSI) to produce a super-resolved HSI that retains both high spatial and high spectral fidelity.

The process begins by naively upsampling the LR-HSI to match the desired spatial resolution. Although this naive upsampling preserves most of the spectral structure, it lacks fine spatial details. To address this, HSRnet leverages the HR-MSI to learn the spatial residuals that can be added back into the upsampled HSI. At the same time, it uses the LR-HSI itself to learn spectral residuals, ensuring that the final output closely matches the ground truth hyperspectral data.

Formally, we want to find a function f that maps an LR-HSI, $Y \in \mathbb{R}^{h\times w\times S}$, and an HR-MSI, $Z \in \mathbb{R}^{H\times W\times s}$, to an HR-HSI, $X \in \mathbb{R}^{H\times W\times S}$, where h, w and H, W represent the spatial dimensions of the LR and HR images respectively, and s, S denote the number of spectral bands of the MSI and HSI. We estimate the parameters Θ of f by minimizing the loss:

$$\min_{\Theta} L = \|f_{\Theta}(Y, Z) - X\|_2^2,$$

where f_Θ is modeled by a deep convolutional neural network (CNN), and Θ represents its weights.

HSRnet's design can be understood as three parts:

Spectral Learning at Low Resolution: The LR-HSI already contains rich spectral information. By learning spectral residuals at the low-resolution scale, the model refines the spectral content of the naively upsampled HSI. As a result, the final HR-HSI aligns closely with the true spectral signatures found in X.

Spatial Learning at Multiple Scales: Spatial details are introduced at both the low-resolution and high-resolution scales using the HR-MSI. At the low-resolution scale, the HR-MSI is first downsampled and concatenated with the

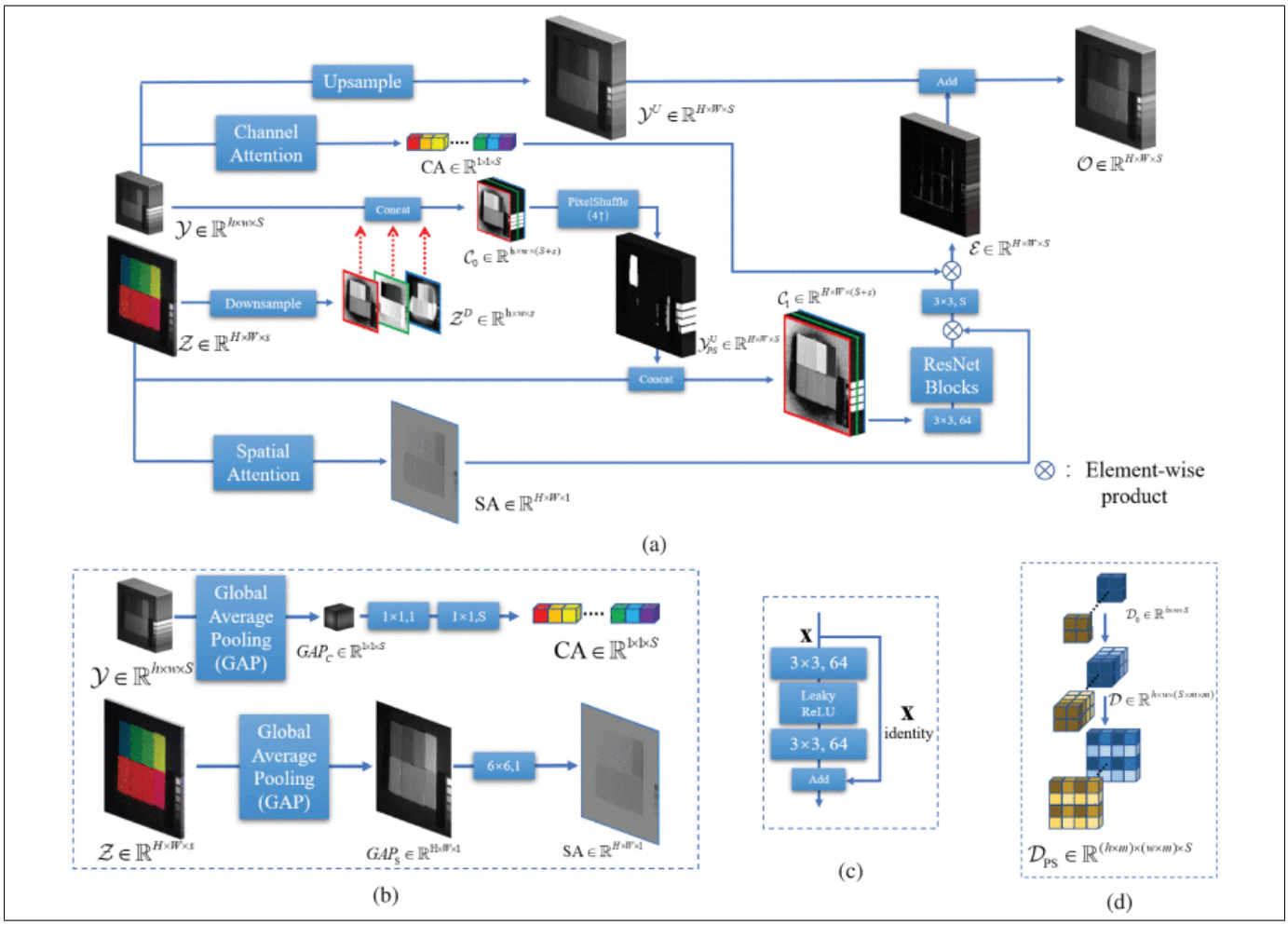

Fig. 2. HSRnet Architecture [4]

LR-HSI. A pixel shuffle layer is used to upsample it back to high resolution, and a subsequent ResNet module learns spatial residuals.

Spatial Attention at High Resolution: To enhance high-resolution spatial features, the HR-MSI passes through a spatial attention module. This module typically includes a global average pooling layer that captures global spatial context, followed by convolution with a learnable kernel.

3.2 Incorporating the Fourier Domain

While HSRnet and other image fusion models can effectively preserve low-frequency details, they do not preserve high-frequency details. This was confirmed by passing the images through a high/low pass filter in the Fourier domain. We experiment with various frequency domain-based adjustments to HSR-Net to improve frequency preservation. To punish loss of high-frequency details, we add a **High-Frequency Domain Loss Term** while training, defined by

$$\mathcal{L}_{hf} = \frac{1}{HWC} \sum_{i=1}^{HWC} \|\mathrm{HP}(X)_i - \mathrm{HP}(f_\Theta(Y,Z))_i\|_2^2$$

where HP is a high-pass filter, H, W, C is the height, width, and spectral channel dimension respectively, X is the ground truth HR-HSI, f_Θ is the network, Y is the LR-HSI and Z is the HR-MSI.

To enforce even greater high-frequency preservation, we modify the architecture directly, by learning residuals of the image in both the spatial and frequency domain, and then fuse both residuals through a weighted sum, combining the learnt spatial residuals and learnt frequency residuals. We denote this model as the **HSRnet Frequency Domain Fusion Network** (FD-HSRnet). Given the LR-HSI Y, HR-MSI Z, FD-HSRnet F_Θ is defined as

$$F_\Theta(Y,Z) = \beta f_\Theta(Y,Z) + \gamma \mathrm{IFFT}\Big(f_\Theta(\mathrm{Re}(Y_{\mathrm{FFT}}), \mathrm{Re}(Z_{\mathrm{FFT}})) + i f_\Theta(\mathrm{Im}(Y_{\mathrm{FFT}}), \mathrm{Im}(Z_{\mathrm{FFT}}))\Big)$$

where β and γ are parameters for the weighted sum, $*_{\mathrm{FFT}}$ is the Fourier transformed input, IFFT is the inverse Fourier transform, Re() takes the real part of the complex number, and Im() takes the imaginary part. In this paper, we use $\beta = 0.85$ and $\gamma = 0.15$.

3.2.1 Training

We emulate the training regime of HSRnet. All models were trained on the same dataset and hyperparameters. Training was done on the CAVE hyperspectral image dataset, which consists of 32 hyperspectral images over 31 bands. 20 images were used for training and 11 for validation. 3920 non-overlapping patches were extracted from the 20 images for

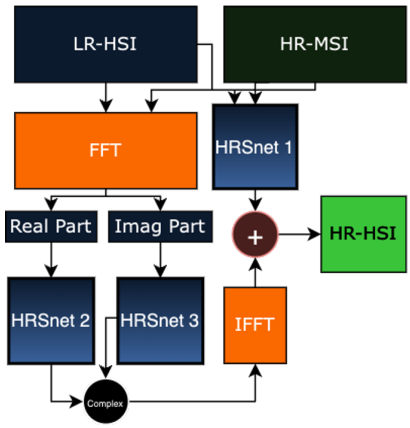

Fig. 3. FD-HSRnet

TABLE 1
Average Values over 11 Testing Images from the CAVE Dataset

Model	PSNR	SSIM	SAM	ERGAS
HSRnet	33.73	0.930	0.065	3.82
HSRnet + $0.05\mathcal{L}_{hf}$	34.17	0.941	0.061	3.66
HSRnet + $0.10\mathcal{L}_{hf}$	34.69	0.951	0.058	3.49
HSRnet + $0.15\mathcal{L}_{hf}$	35.00	0.959	0.055	3.36
HSRnet + $0.20\mathcal{L}_{hf}$	35.20	0.963	0.054	3.29
FD-HSRnet	33.73	0.93	0.065	3.82
FD-HSRnet + $0.20\mathcal{L}_{hf}$	**35.22**	**0.963**	**0.054**	**3.28**

training. To simulate low-resolution hyperspectral images, a Gaussian blur was applied with a 3×3 kernel and 0.5 standard deviation, before downsampling each patch by a factor of 4. Additionally, the spectral response function of the Nikon D700 camera was used to extract the corresponding multispectral images [13]. All models were implemented in Python 3.12.6 with PyTorch 2.4 and trained on an NVIDIA RTX 4060 GPU. The Adam optimizer was used with a learning rate of $1e-4$. All models were trained for 200 epochs, and training time ranged between 1 and 2 hours each. The original HSRnet architecture and training regime was first reimplemented in PyTorch, and all subsequent models in this paper are based on this reimplementation.

The original HSRnet was trained to minimize the mean squared error,

$$\mathcal{L}_{\text{MSE}} = \|f_\Theta(Y, Z) - X\|_2^2,$$

We denote the HSRnet models trained with our additional high-frequency domain loss term as **HSRnet+$\alpha\mathcal{L}_{hf}$** where α is the weight of the additional loss term.

$$\mathcal{L} = \mathcal{L}_{\text{MSE}} + \alpha\mathcal{L}_{hf}$$

4 Experimental Results

We train various models and compare them against the original HSRnet. For evaluation, we use the CAVE dataset. We adopt four widely used quantitative quality measures, the peak signal-to-noise ratio (PSNR), the structure similarity index, the spectral angle mapper (SAM), and the erreur relative globale adimensionnelle de synthèse (ERGAS).

TABLE 2
Average Values over 11 Testing Images from the CAVE Dataset (High Frequencies)

Model	PSNR	SSIM	SAM	ERGAS
HSRnet	18.95	0.109	1.336	38984
HSRnet + $0.05\mathcal{L}_{hf}$	20.02	0.213	1.142	35949
HSRnet + $0.10\mathcal{L}_{hf}$	20.48	0.256	1.072	34642
HSRnet + $0.15\mathcal{L}_{hf}$	21.32	0.360	0.954	31310
HSRnet + $0.20\mathcal{L}_{hf}$	21.55	0.386	0.925	30819
FD-HSRnet	18.95	0.109	1.337	38984
FD-HSRnet + $0.20\mathcal{L}_{hf}$	**21.56**	**0.386**	**0.924**	**30729**

TABLE 3
Average Values over 11 Testing Images from the CAVE Dataset (Low Frequencies)

Model	PSNR	SSIM	SAM	ERGAS
HSRnet	49.06	0.992	0.021	1.19
HSRnet + $0.05\mathcal{L}_{hf}$	49.35	0.993	0.021	1.15
HSRnet + $0.10\mathcal{L}_{hf}$	49.95	0.993	0.021	1.08
HSRnet + $0.15\mathcal{L}_{hf}$	50.05	0.993	0.021	1.07
HSRnet + $0.20\mathcal{L}_{hf}$	50.36	0.994	0.021	1.04
FD-HSRnet	49.06	0.993	0.020	1.19
FD-HSRnet + $0.20\mathcal{L}_{hf}$	**50.42**	**0.994**	**0.021**	**1.03**

The average values for each metric over the 11 testing images are shown in Table 1. To test the performance on high-frequency details, we apply a high-pass filter to each result and calculate the metrics on the output. The average values for only the high-frequencies are shown in Table 2, and for only the low-frequencies are shown in Table 3. We find that our proposed method outperforms the original HSRnet on the original images and also demonstrates improved high-frequency preservation. Notably, HSRnet + $\alpha\mathcal{L}_{hf}$ demonstrates gradually increasing performance as α increases. Furthermore, FD-HSRnet performs similarly to HSRnet, but outperforms all other methods when trained with the additional loss term. It is interesting to note that although our new methods only directly punish loss of high-frequency details, the preservation of low-frequency details also improves, as shown in Table 3.

For qualitative analysis, we display the results of a single unseen image for the 31st, 16th, and 1st bands in Figure 4. Both HSRnet + $0.2\mathcal{L}_{hf}$ and FD-HSRnet + $0.2\mathcal{L}_{hf}$ produce sharper super-resolved images than the original HSRnet.

To evaluate spectral reconstruction, we plot selected spectral vectors for another testing image, *jelly beans*, for the ground truth HR-HSI, downsampled LR-HSI, original HSRnet output, HSRnet + $0.2\mathcal{L}_{hf}$ output, FD-HSRnet output, and FD-HSRnet + $0.2\mathcal{L}_{hf}$ output in Figure 5.

For each band, we also compute the mean difference between the outputs of each model and the ground truth, for all 11 testing images, shown in Figure 6. Our analysis has so far considered the quality of the hyperspectral image as a whole. To test the possibility of increased performance due to some bands overcompensating for others, we plot for each band the *difference* between the mean squared error over all 11 testing images, as shown in Figure 7. This is additional confirmation that FD-HSRnet performs similarly

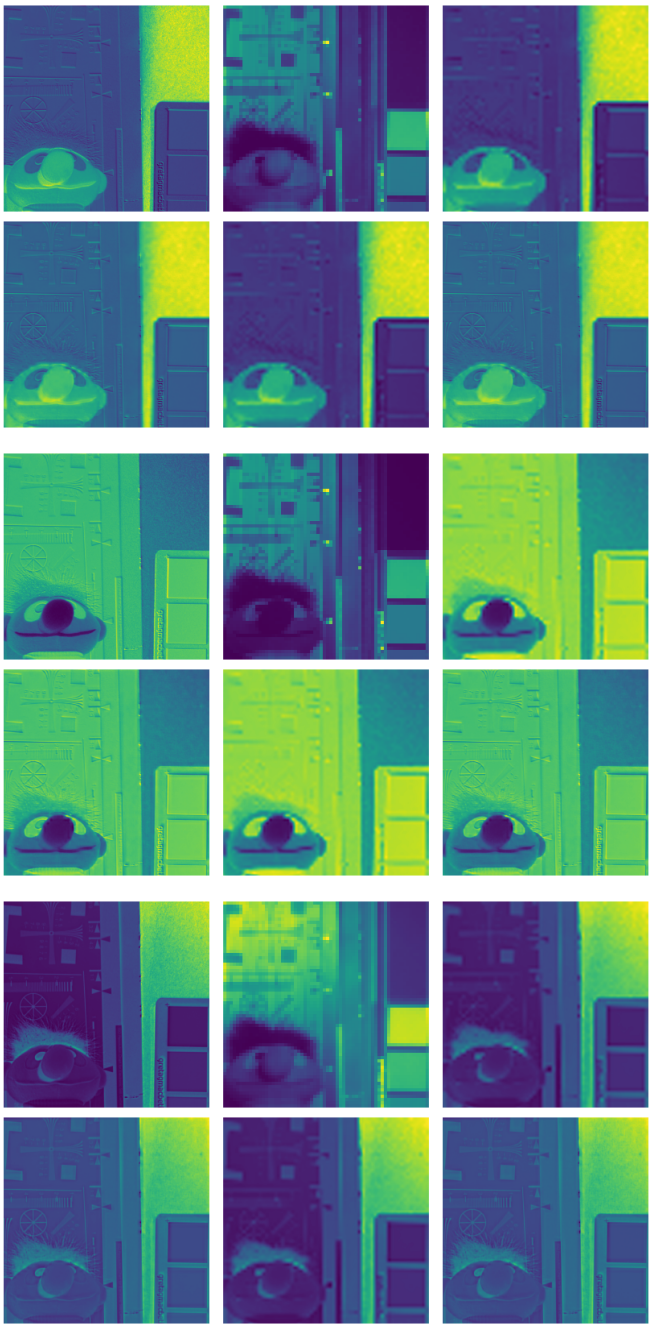

Fig. 4. Top row pair: 31st band, Middle row pair: 16th band, Bottom row pair: 1st band
For each row pair; First row: ground truth HR-HSI, downsampled LR-HSI, original HSRnet output. Second row: HSRnet + $0.2\mathcal{L}_{hf}$ output, FD-HSRnet output, FD-HSRnet + $0.2\mathcal{L}_{hf}$ output

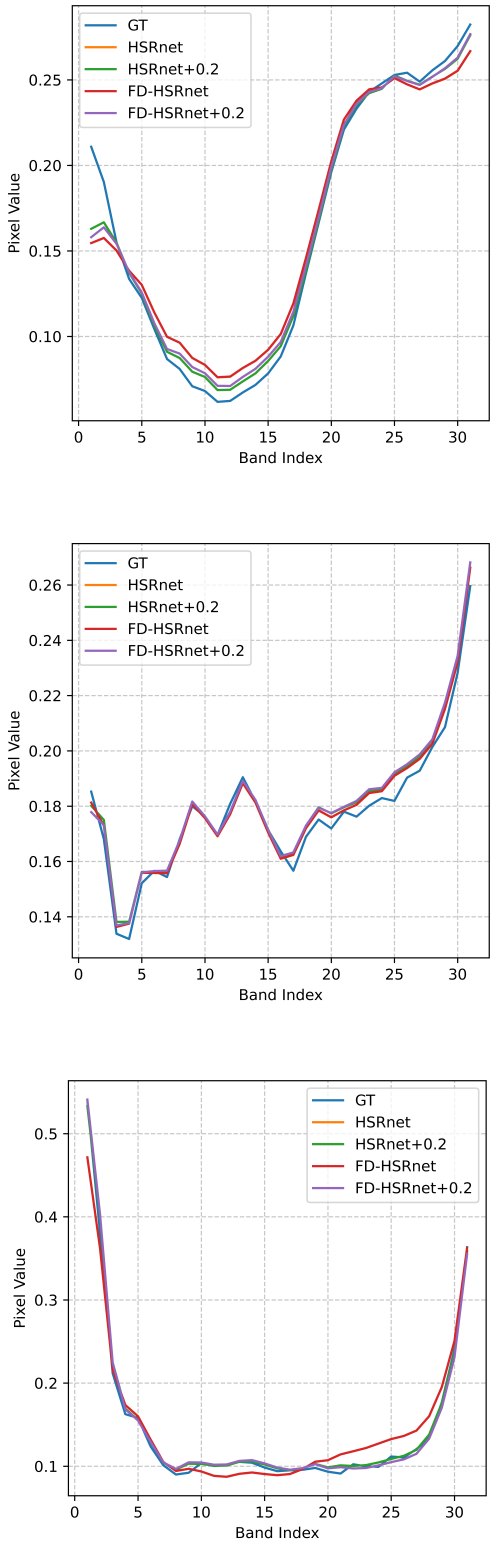

Fig. 5. Top: spectral vector at (128,480), (256, 256), (372, 464)

to HSRnet; there is almost no change in the error. However, the impact of the high-frequency loss term can be seen for both HSR-net and FD-HSRnet. In both HSRnet + $0.2\mathcal{L}_{hf}$ and FD-HSRnet + $0.2\mathcal{L}_{hf}$, the change in error is negative at every band i.e., both methods show greater performance over the original HSRnet at every single band.

5 Conclusion

The HSRnet is an architecture for the hyperspectral image super-resolution problem, where a high-resolution multi-spectral image is fused with a low-resolution hyperspectral

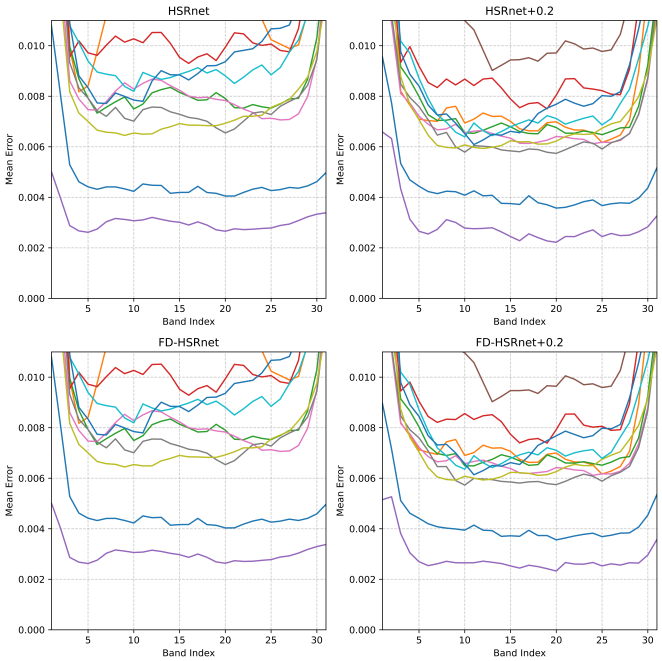

Fig. 6. Mean squared error for each testing image at each band

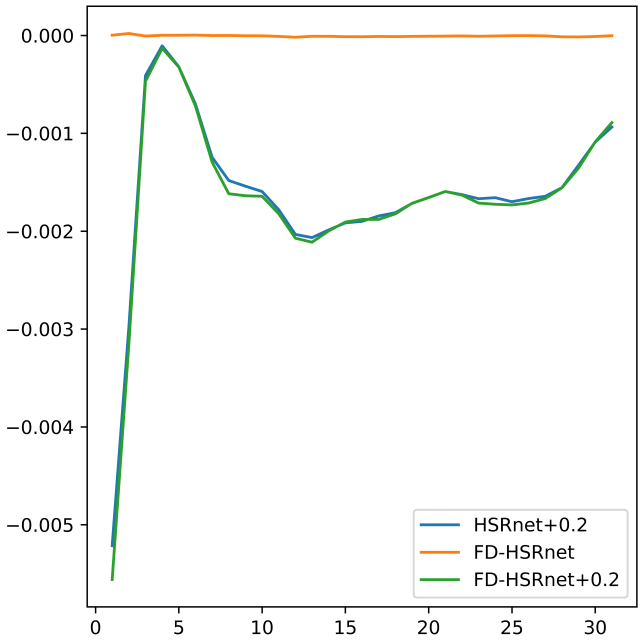

Fig. 7. Difference in mean squared error over all 11 testing images, at each band

image through a deep CNN architecture that utilizes multi-scale spatiospectral attention mechanisms. Recognizing that some intrinsic features may be better represented in the frequency domain, we build upon the HSRnet with a dual domain approach incorporating both spatial and frequency domain representations. We do this via two methods: an additional high-frequency domain loss term that punishes errors at high-frequencies, as well as a dual-domain fusion approach at the architecture level.

Experiments demonstrated that our method outperforms HSRnet with higher PSNR and SSIM values and lower SAM and ERGAS values. The improvement of the latter two values indicate reduced spectral distortion. We also demonstrate increased performance across the entire spectral range i.e., our methods improve reconstruction quality as a whole. As well, while our methods are motivated by penalizing loss of high-frequency details, both methods also improve loss of low-frequency details, demonstrating the complementary nature of the two domains.

By operating on existing network structures and incorporating frequency-domain information in a modular manner, our enhancements maintain the general applicability and low complexity of HSRnet. This enables better performance without sacrificing scalability and adaptability.

5.1 Limitations and Future Work

Due to time and computing constraints, in this paper we have only tested the high-frequency domain loss term with $\alpha \leq 0.2$, and FD-HSRnet with a single set of heuristically chosen fusion parameters $\beta = 0.85, \gamma = 0.15$. Future work could perform a parameter sweep for more optimal values. Additionally, the original HSRnet was trained for much longer ($\sim$1 hour versus $\sim$5 hours) on a similar performing GPU. Even so, our approach demonstrates that the addition of a high-frequency domain loss term may drastically decrease the training time. Future work could also explore why certain spectral bands are less affected by our approach (e.g., band number 5 in Figure 7), investigate more sophisticated frequency-domain transformations, explore frequency-aware priors inspired by physical models of imaging systems, or integrate these enhancements with recent advances in attention and transformer-based networks. Finally, future work should involve testing our approach on a broader range of datasets and adapting it to other hyperspectral super-resolution models, as well as other image restoration and enhancement tasks.

Acknowledgments

The authors would like to thank Aviad Levis for advising this project.

References

[1] R. Dian, S. Li, B. Sun, and A. Guo, "Recent advances and new guidelines on hyperspectral and multispectral image fusion," *Information Fusion*, vol. 69, pp. 40–51, 2021.

[2] R. Dian, L. Fang, and S. Li, "Hyperspectral image super-resolution via non-local sparse tensor factorization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5344–5353.

[3] Q. Xie, M. Zhou, Q. Zhao, D. Meng, W. Zuo, and Z. Xu, "Multispectral and hyperspectral image fusion by ms/hs fusion net," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 1585–1594.

[4] J.-F. Hu, T.-Z. Huang, L.-J. Deng, T.-X. Jiang, G. Vivone, and J. Chanussot, "Hyperspectral image super-resolution via deep spatiospectral attention convolutional neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 12, pp. 7251–7265, 2021.

[5] Z. Zhu, J. Hou, J. Chen, H. Zeng, and J. Zhou, "Hyperspectral image super-resolution via deep progressive zero-centric residual learning," *IEEE Transactions on Image Processing*, vol. 30, pp. 1423–1438, 2020.

[6] S. Li, R. Dian, L. Fang, and J. M. Bioucas-Dias, "Fusing hyperspectral and multispectral images via coupled sparse tensor factorization," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 4118–4130, 2018.

[7] T. Xu, T.-Z. Huang, L.-J. Deng, X.-L. Zhao, and J. Huang, "Hyperspectral image superresolution using unidirectional total variation with tucker decomposition," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 4381–4398, 2020.

[8] X. Zhang, W. Huang, Q. Wang, and X. Li, "Ssr-net: Spatial–spectral reconstruction network for hyperspectral and multispectral image fusion," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 7, pp. 5953–5965, 2020.

[9] Z. Wang, B. Chen, R. Lu, H. Zhang, H. Liu, and P. K. Varshney, "Fusionnet: An unsupervised convolutional variational network for hyperspectral and multispectral image fusion," *IEEE Transactions on Image Processing*, vol. 29, pp. 7565–7577, 2020.

[10] W. G. C. Bandara, J. M. J. Valanarasu, and V. M. Patel, "Hyperspectral pansharpening based on improved deep image prior and residual reconstruction," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2021.

[11] L. Jiang, B. Dai, W. Wu, and C. C. Loy, "Focal frequency loss for image reconstruction and synthesis," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 13 919–13 929.

[12] S. D. Sims, "Frequency domain-based perceptual loss for super resolution," in *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2020, pp. 1–6.

[13] J. Jiang, D. Liu, J. Gu, and S. Süsstrunk, "What is the space of spectral sensitivity functions for digital color cameras?" in *2013 IEEE Workshop on Applications of Computer Vision (WACV)*. IEEE, 2013, pp. 168–179.